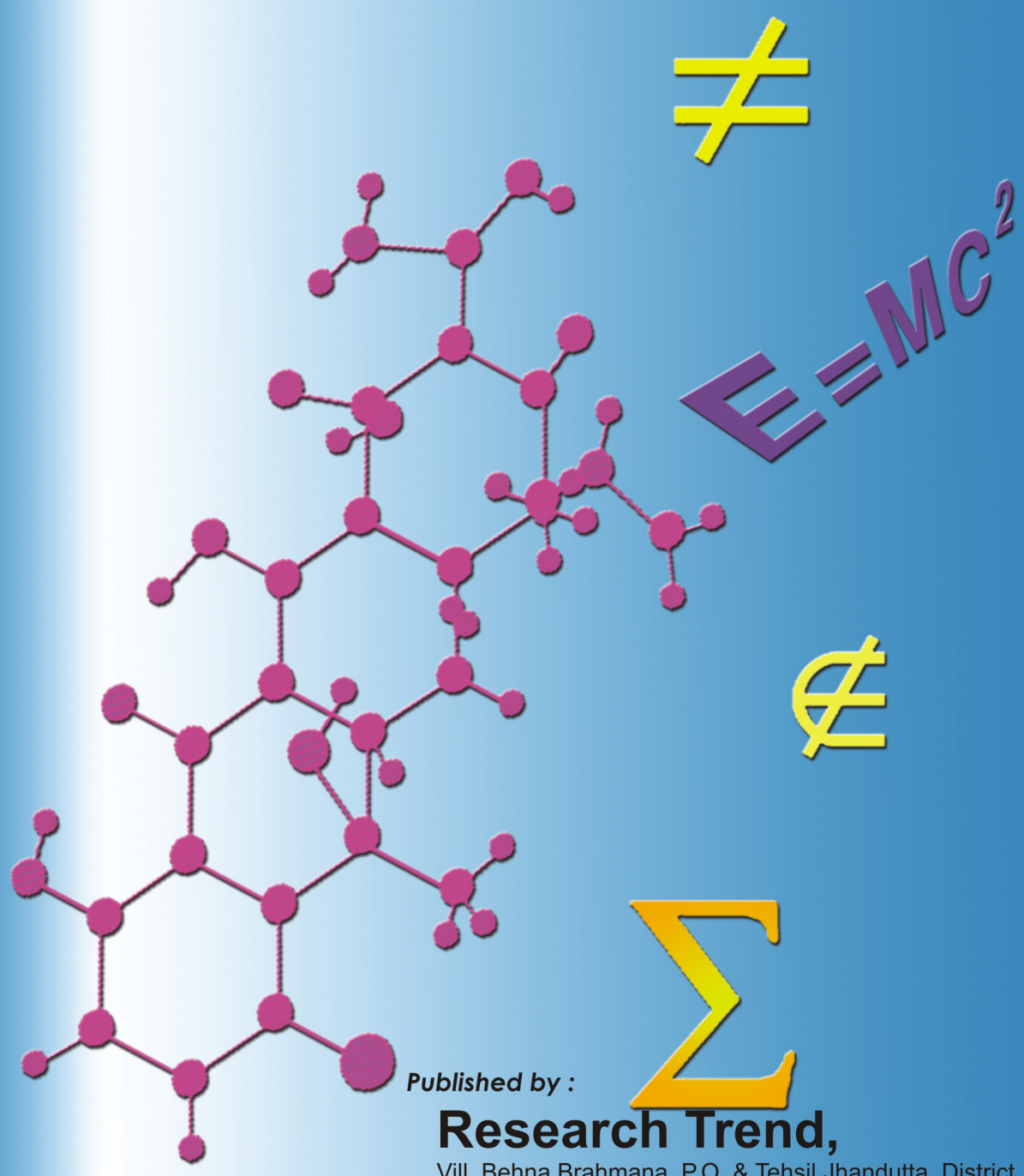
International Journal of Theoretical & Applied Sciences

VOL. 10(1): January-June, 2018

Print ISSN No.: 0975-1718

Online ISSN No.: 2249-3247



Published by :

Research Trend,

Vill. Behna Brahmana, P.O. & Tehsil Jhandutta, District. Bilaspur, Himachal Pradesh-174031 (India) Website: www.researchtrend.net

Special Issue Released
for

International Conference on SMAC

- Reshaping the Future

9th February 2018

In Commemoration of
Diamond Jubilee Year

Organised by

Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore – 641043

International Journal of Theoretical & Applied Sciences, *Special Issue* 10(1): (2018)
International Conference on SMAC - Reshaping the Future (February 2018)

January–June, 2018 **Vol. 10(1): 2018** **ISSN No. (Print): 0975-1718**
ISSN No. (Online): 2249-3247

Research Trend,

J-4/50, 2nd Floor, Khirki Extension, Malviya Nagar, New Delhi-17.
Website: www.researchtrend.net **Email**: dheeraj_vasu_72066@yahoo.co.in,
researchtrend09@gmail.com, Mobile : **9868001440**

International Journal of Theoretical & Applied Sciences, *Special Issue* 10(1): (2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Editorial

"India, a land which gave birth to civilization in ancient times and where much of the earlier tradition and wisdom guides actions even in modern times the philosophy of Vasudhaiva Kutumbakam which means that the whole universe is one family, dominates global efforts to protect the global commons".

We are profoundly privileged to bring before you the efforts of a few genius minds in the form of ***"International Journal of Theoretical and Applied Sciences".*** Its aims to encourage to make aware the human being towards scientific attitude for the betterment of ecosystem and social life.

Cheers to all those involved directly on front or indirectly behind the curtain in this noble attempt of serving ecosystem.

Science belongs to the whole world, and before it, vanish all the barriers of nationality. With the resonance of science in all the activities of our lives, we are trying to marvel this age of specialization. Standing on this verge of eternity we are trying to possess more and more of power and pelf. For it, we require a quality of mind, which should be special and should have an extreme advantage in leading to make discoveries. What matters is the power of never letting exceptions go unnoticed.

The foundation blocks of research are to do the right thing, at the right time, in the right way; to anticipate requirements; to develop resources and then to recognize no impediments and thence to master circumstances. One has to act from reason rather than rule and to be satisfied with nothing short of perfection. True researcher resides in the capacity or evaluation of uncertain, hazardous and conflicting information. Curiosity, Confidence, Courage and Constancy are the hallmarks. Research is a long road to be traded by the brave ones. One has to brave all odds, long pangs of suffering and frustration to bring the work in hand to fruition. To attain the pinnacle of success one ought to nurture research with hard work and toil.

We congratulate and wish luck to the researchers for their contributions and aspire that these works will leave a glowing trail for the generations to come. These works will act as lighthouses to the future generations and rare milestones in the fields of Science and Technology. It may also provide the courage to tread the long and weary path of research, as success and hard work has a taste beyond everything.

– Editor-in-chief

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): (2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): (2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Table of Contents Vol. 10(1): 2018

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 01-10(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Analytics on Web Logs for Exploring User Behavior Patterns using R Package

V. Sharmila, Dr. G. Sudhamathy and Dr. G. Padmavathi
Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: Web users encounter the problems such as "Finding relevant Information", "Personalization of Information" & "Learning about consumers or individual users" when interacting with the web. Assessment of user actions in a web site can offer insights causing to customization and personalization of a web site, there by promoting e-business and e-services. Website users need automated tools to track and analyze the web usage patterns. There are many tools currently existing in this field but they lack in some aspect such as in need of huge storage requirements, excessive I/O cost, scalability problems when additional information is introduced into this analysis. Hence, the task of mining useful information is a challenge nowadays given the complexity of the data and its huge volume. Web masters would prefer to have software that monitors and counts the web users, the pages they visit, their duration on the web site and other activity that can provide insight on the behavior pattern. None of the works so far has explored the use of R software for web usage pattern analysis. As R is open source software that is now widely used for data analytics, the feasibility and effectiveness of constructing a web analytics framework using R is explored in this paper. Moreover, R can support scalability, huge storage requirements and low I/O cost when compared other existing software and thereby it addresses the gap.

Keywords: Web Mining, Web Usage Mining, Analytics, Recommender Systems

I. INTRODUCTION

The web usage mining tools are evolving and the present techniques still have rooms for improvement to make them prevail in the web based information systems. New tools are needed which will not use up too much resources or process time during the web mining process. It is important to improve visualization, as much of the data is unorganized and difficult for the user to understand. There are some web usage mining tools that could provide web usage statistics. These statistics could be useful for web administrators to get a sense of the actual load on the server.

Table 1: Existing Web Usage Mining Research Based Tools.

Tool Name	Available Features
ArchCollect	It uses fast semantic interaction acquisition algorithms and form a dimensional cube of information directly which serve as input to an IR approach. [5]
i-JADE Web-Miner	To use Agent technology, together with Web mining technology, to automate a series of product search and selection activities. It is based on multi-agent development platform iJADE. [10]
i-Miner	To optimize the concurrent architecture of a fuzzy clustering algorithm (to discover data clusters) and a fuzzy inference system to analyze the trends. [12]
SEWeP	Developed as a system that makes use of both usage logs and semantics of a Web site's content in order to personalize it. [13]
AWUSA	A framework based on combination of information architecture, automated usability evaluation and web mining techniques for data gathering and analysis. [14]
Web Quilt	Web logging and visualization system that helps web design teams capture usage traces which can be aggregated and visualized in a zooming interface that shows the web pages people viewed. [15]
KOINOTITES	A system which uses data mining techniques for the construction of user communities on the Web. [18]
INSITE	To generate user profiles in real time through the use of a unique Connectivity Matrix Model (CM model), and show the Efficacy and Scalability. [19]
MiDAS	It introduces a new algorithm called MiDAS that extends traditional sequence discovery with a wide range of web-specific features. [17]
STRATDYN	Developed as add on module to the WUM, extends its capability by exploiting site semantics. [16]
WebTool	It uses sequential pattern mining which relies on PSP an algorithm developed by the authors. [20]
WebLogMiner	Use data mining and OLAP on treated and transformed web access files. [21]
SpeedTracer	Reconstructs the user transversal paths for session identification by using the referrer page and the URL of the requested page as a traversal step. [22]

However, the statistical data available from the normal web log data files or even the information

provided by web trackers could only provide the information explicitly because of the nature and limitations of the methodology itself. After browsing through some of the features of the best trackers available it is easy to conclude that rather than generating statistical data and texts, they really do not help too much in providing meaningful information.

Table 2: Commercial Web Log Analyzer Tools Comparison.

Compared Features	Log Analyzer Tools			
	AWStats	**Analog**	**Webalizer**	**Sawmill Analytics**
Software Language	Perl	C	C	C / Salang
Price	Free	Free	Free	From $99 per profile
License	GPL	GPL	GPL	Lite /Pro / Ent
	General Public License (GPL)			
Works with W3C log format	Yes	Yes	Need a patch	Yes
Works with personalized log format	Yes	Yes	No	Yes
FTP Log Files	Yes	No	No	Yes
Reports number of "human" visits	Yes	No	Yes	Yes (Sessions)
Reports Session duration	Yes	No	No	Yes
Reports most often viewed pages	Yes	Yes	Yes	Yes
Reports entry pages	Yes	No	Yes	Yes
Reports exit pages	Yes	No	Yes	Yes
Reports OS	Yes	Yes	No	Yes
Reports Browser	Yes	Yes	Yes	Yes
Reports HTTP Errors	Yes	Yes	Yes	Yes
Provides graphical statistics	Yes	Yes	Yes	Yes

Hence the list of existing Research Based Web Usage Mining Tools (Table 1) and a comparison of the features provided by the various Commercial Web Log Analyzer Tools (Table 2) are presented here. This information throws light on what the existing work has brought in to existence and what is missing and expected by the current web site users and designers.

So, in this work, we try to analyze the user preferences from the freely available web log file of the University of Saskatchewan's located in Canada. This file is available online in the link "http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html".

This file is loaded into the R Tool, parsed for getting only the required fields, few pre-processing steps are carried out to clean the data, and user identification and session identification are carried out for making the data ready for further processing. This cleansed and filtered data is then used to rank the pages and users based on few criteria. The pages are ranked based on the visit frequencies, time spent and a combination of the visit frequency and time spent. The users are also ranked based on their visit frequency, time spent and a combination of visit frequency and time spent. The combination of the pages and users are also ranked based on the visit frequency and time spent and combination of visit frequency and time spent. As discussed before the user grouping based on time spent in the various months is also found. Similarly, the page grouping based on time spent in the various months is also found.

Similar works have been discussed in the "Review of Literature" Section and the gap in exploring using R has been highlighted. The "Methodology" Section explores the step by step approach that has been followed in this data analytics work. The "Results and Discussions" Section explores the coding and the resultant graphs obtained in this work. It is also important to note that the efficiency of the results obtained by using the R Package techniques are more useful and visually makes more sense.

II. REVIEW OF LITERATURE

In the work by C. J. Aivalis, [3] a novel hybrid solution was proposed that is based on the junction of log files with operational data and page tagging, which allows even exacter measurements of customer behavior. It allows a customization of the Analysis Tool that survives the shift of the technologies. ArchCollect [5], uses fast semantic interaction acquisition algorithms and form a dimensional cube of information directly which serve as input to an IR approach. Its main function is to monitor users interactions in web media.

In the work by Nasraoui *et al.,* [6], they present a complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing an ontology of the Web content. The research by Pascual-Cid, [7] aims at proving the usefulness of a set of information visualization techniques in order to analyze Web data, using a visual Web mining tool that allows the combination, coordination and exploration of visualizations to get insight on Web data.

The i-JADE Web-Miner [10] is mainly for E-Commerce Applications, which uses Agent technology, together with Web mining technology, to automate a series of product search and selection activities. It is based on multi-agent development platform iJADE. i-

Miner [12] is mainly used for Pattern Discovery and trend analysis from web usage data. It optimizes the concurrent architecture of a fuzzy clustering algorithm (to discover data clusters) and a fuzzy inference system to analyze the trends.

SEWeP [13] is mainly used for Personalization. This is developed as a system that makes use of both usage logs and semantics of a Web site's content in order to personalize it. AWUSA [14] is an automated website usability evaluation tool. This is a framework based on combination of information architecture, automated usability evaluation and web mining techniques for data gathering and analysis.

Web Quilt [15] was developed to run usability tests and analyze the collected data from web logs. It is a web logging and visualization system that helps web design teams capture usage traces which can be aggregated and visualized in a zooming interface that shows the web pages people viewed. STRATDYN [16] was mainly developed for visualization of navigation patterns. It was developed as add on module to the WUM, extends its capability by exploiting site semantics. MiDAS (Mining Internet Data for Associative Sequences) [17] is mainly used for pattern discovery. It introduces a new algorithm called MiDAS that extends traditional sequence discovery with a wide range of web-specific features.

KOINOTITES [18] is mainly used for personalization. It is a system which uses data mining techniques for the construction of user communities on the Web. INSITE [19] was mainly developed for acquisition extracts and stores the essence of the captured information in real time and visualizes the result. It generates user profiles in real time through the use of a unique Connectivity Matrix Model (CM model), and show the Efficacy and Scalability. WebTool [20] is mainly used for usage profiling. It uses sequential pattern mining which relies on PSP an algorithm developed by the authors. WebLogMiner [21] is mainly for Mining web server log files. It uses data mining and OLAP on treated and transformed web access files. Speed Tracer [22] reconstructs the user transversal paths for session identification by using the referrer page and the URL of the requested page as a traversal step.

In the area of web usage mining, the widely used data mining algorithm is the clustering analysis [8], [9], [11]. Rui Wu, 2010 [4], in his article uses the fuzzy method to discover generalized fuzzy association rules among the web pages from web logs. These approaches have crisp boundary problems and have considerable computational periods.

Mining algorithms yield usage patterns, but finding the ones that constitute new and interesting knowledge in the domain remains a challenge. There are many commercial tools which perform analysis on web log data and they are based on statistical analysis techniques, while only a few products exploit Data Mining techniques. There are other web usage mining tools existing like the Rapid Miner and Data Miner which is mostly a pre-set up tool that only uses the existing old algorithms.

Hence there is a need found to design a framework that integrates all recent efficient web usage mining techniques and to present the results as graphical representations for better web personalization.

III. METHODOLOGY

The step by step approach of the proposed methodology is explained in the Fig. 1.

Loading Data. The web log files of the University of Saskatchewan's is taken from the link http://ita.ee.lbl.gov/html/contrib/Sask-HTTP.html and loaded as comma separated value file in the R Studio.

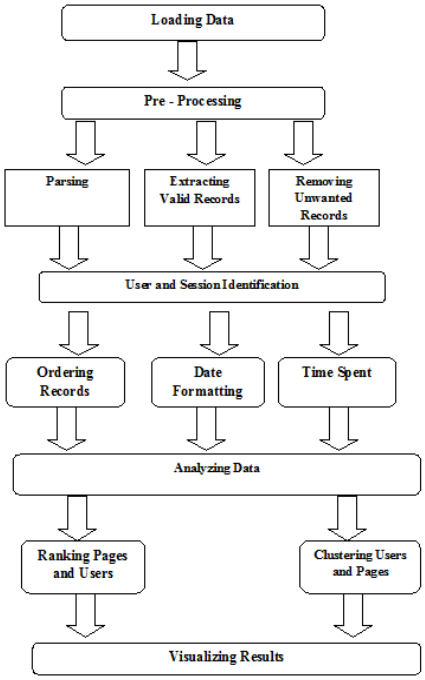


Fig. 1. Flow Diagram for the Proposed Methodology.

The unloaded dataset contains seven months of data from 01/Jun/1995 to 31/Dec/1995. This file consists of 200000 records and it does not have any header record.

Pre-Processing. Among the nine fields available in this dataset, the fields such as User Id, Timestamp and URL are extracted for further processing. The records that have only the strings “HTTP”, “GET” and “.html” in the URL field are only retained and the other records are removed. This is done as a pre-processing step to analyze only the valid page requests. Further, the records with the strings “.gif”, “~”, “search”, “?query”, “?file” and “/ HTTP/1.0” in the URL field are removed.

This is done to remove the unwanted requests that include images, search queries and some files download. In the URL field the characters from 5 to 100 are only retained as this file has some unwanted junk data in the first four characters. In the Timestamp field the characters from 2 to 25 are only retained as there is an unwanted junk character in the first field. These pre-processing steps are carried out based on the visual inspection of the input log file.

User and Session Identification. There is a package named *sqldf* in the R Tool that helps to handle the input data as we do in any DBMS system. This package is installed and loaded in order to execute the further steps. The records are ordered by based on the ascending order of User Id and Timestamp. The Timestamp field is separated into a Date and Timestamp fields for further processing. Now we calculate the time spent field for each record. The time spent is calculated based on the time difference between two subsequent records if both have the same user id. If the time difference is more than 30 minutes and the user ids are same, then for the particular record the time spent field is set to 60 seconds.

Analyzing Data. Next we run the queries in the resultant data frame for the below results. 1) Top pages visited frequently 2) Top pages visited with more time spent 3) Top users who visit frequently 4) Top users who browse for more time 5) Top user page combination based on visit frequency 6) Top user page combination based on time spent

The Month field is added to the data frame in order to cluster the users and pages within each month. For each month the minimum time spent and the maximum time spent are calculated. Based on this the users within each month are grouped into five clusters which can be termed as "Most Loyal Users", "Loyal Users", "Most Frequent Users", "Frequent Users" and "Least Frequent Users". Also the Pages are grouped into five clusters which can be termed as "Most Popular Pages", "Popular Pages", "Most Favorite Page", "Favorite Page" and "Least Favorite Page". The results are then plotted as bar graphs and pie charts. The Users and the Pages are grouped within each month to view the change in the user preferences over the period across several months.

Visualizing Results. Finally, the obtained analysis results are viewed in the form of bar graphs and pie charts for better visualization and interpretation by the web masters.

IV. RESULTS

The input we log file "UofS_access_log.txt" is loaded into the R Tool, it is preprocessed, user sessions are identified, complex SQL queries are used to analyze the data (Ranking and Clustering of Users and Pages) and finally the results are presented as graphs.

The parameters collected in this experiment are:

a. Top Pages Visited Frequently
b. Top Pages with more Time Spent
c. Top Users who Visited Frequently
d. Top Users who Browse for more Time
e. Top User Page Combination based on Visit Frequency
f. Top user Page Combination based on Time Spent
g. Clustering of Users Month Wise
h. Clustering of Pages Month Wise

The bar graphs and the pie charts produced by the experiments are as presented in the below figures.

Results. (a) Top Pages Visited Frequently

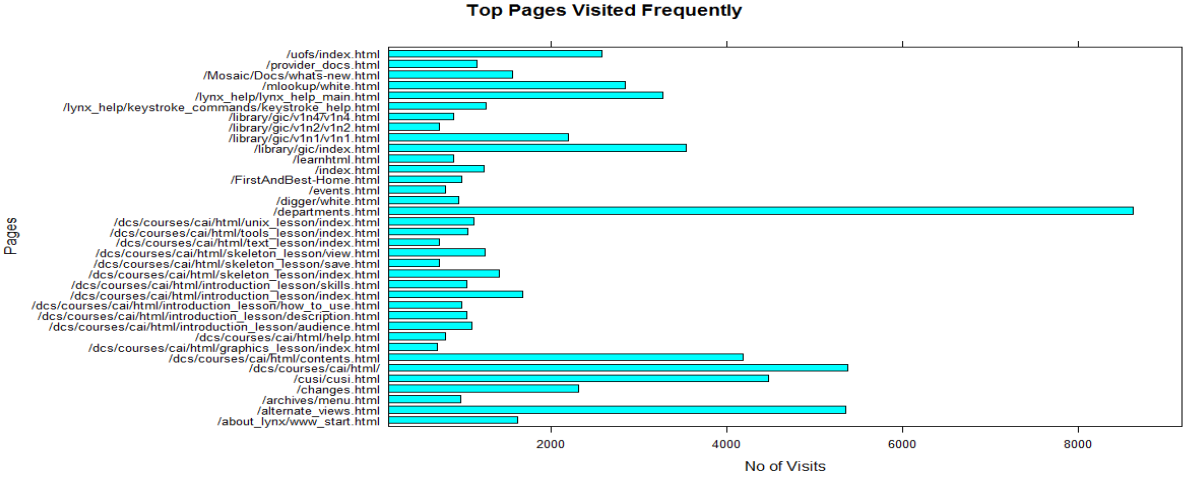


Fig. 2. Top Pages Visited Frequently.

b. Top Pages with more Time Spent

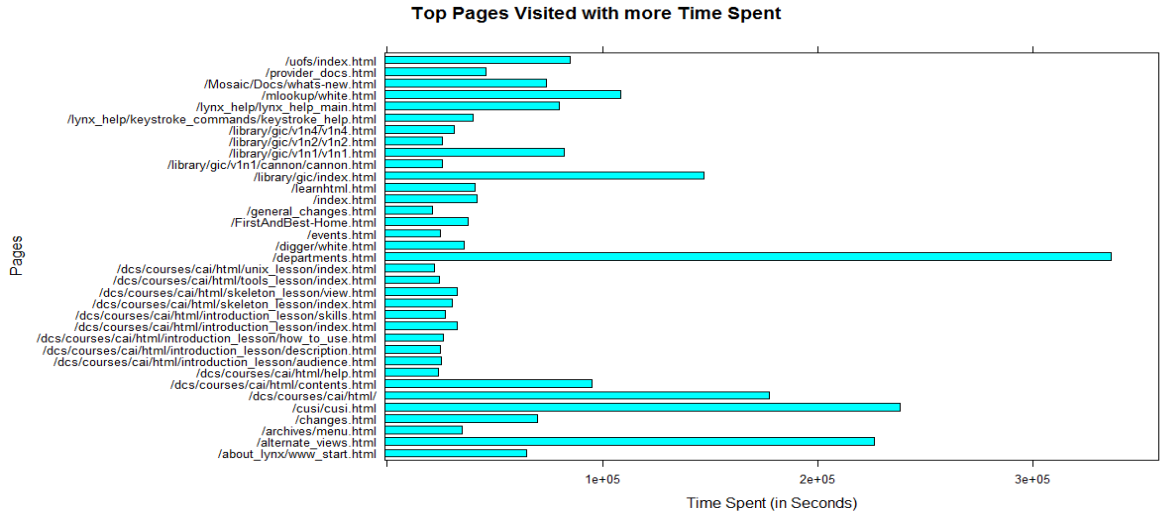


Fig. 3. Top Pages Visited with more Time Spent.

c. Top Users who Visited Frequently

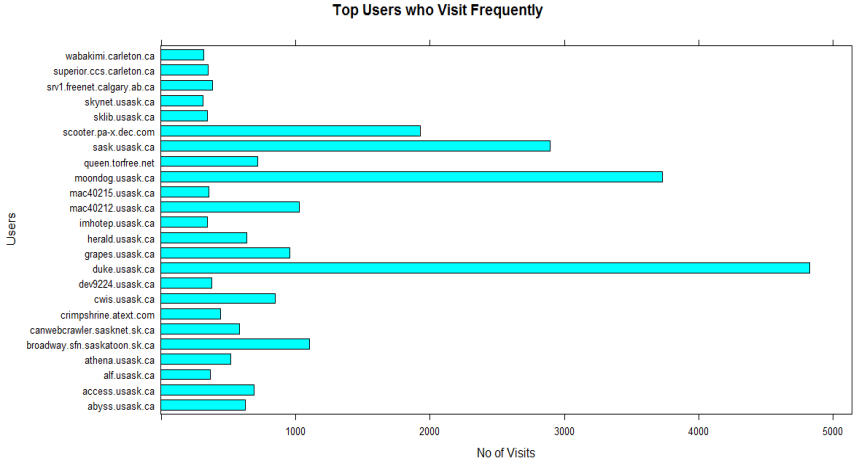


Fig. 4. Top Users Visited Frequently.

d. Top Users who Browse for more Time

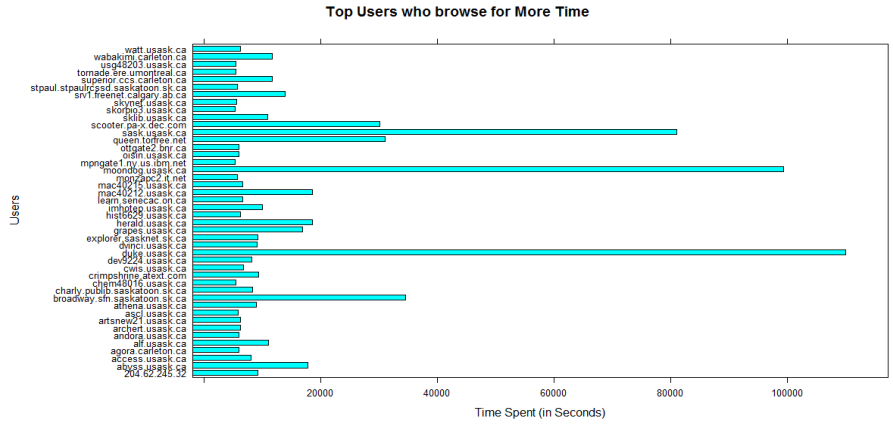


Fig. 5. Top Users Who Browse for More Time.

e. Top User Page Combination based on Visit Frequency

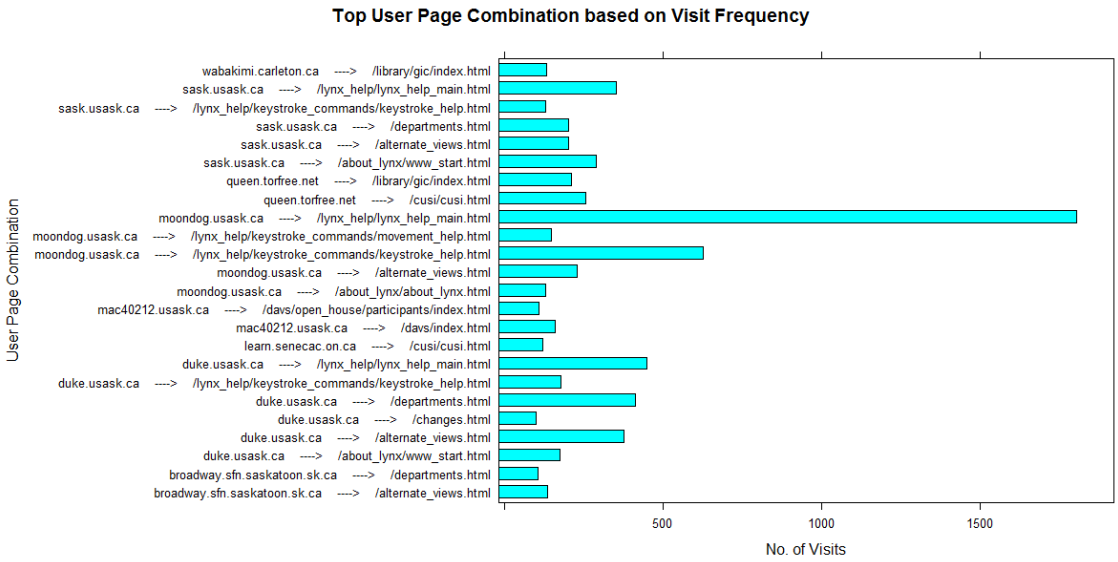


Fig. 6. Top User Page Combination based on Visit Frequency.

f. Top user Page Combination based on Time Spent

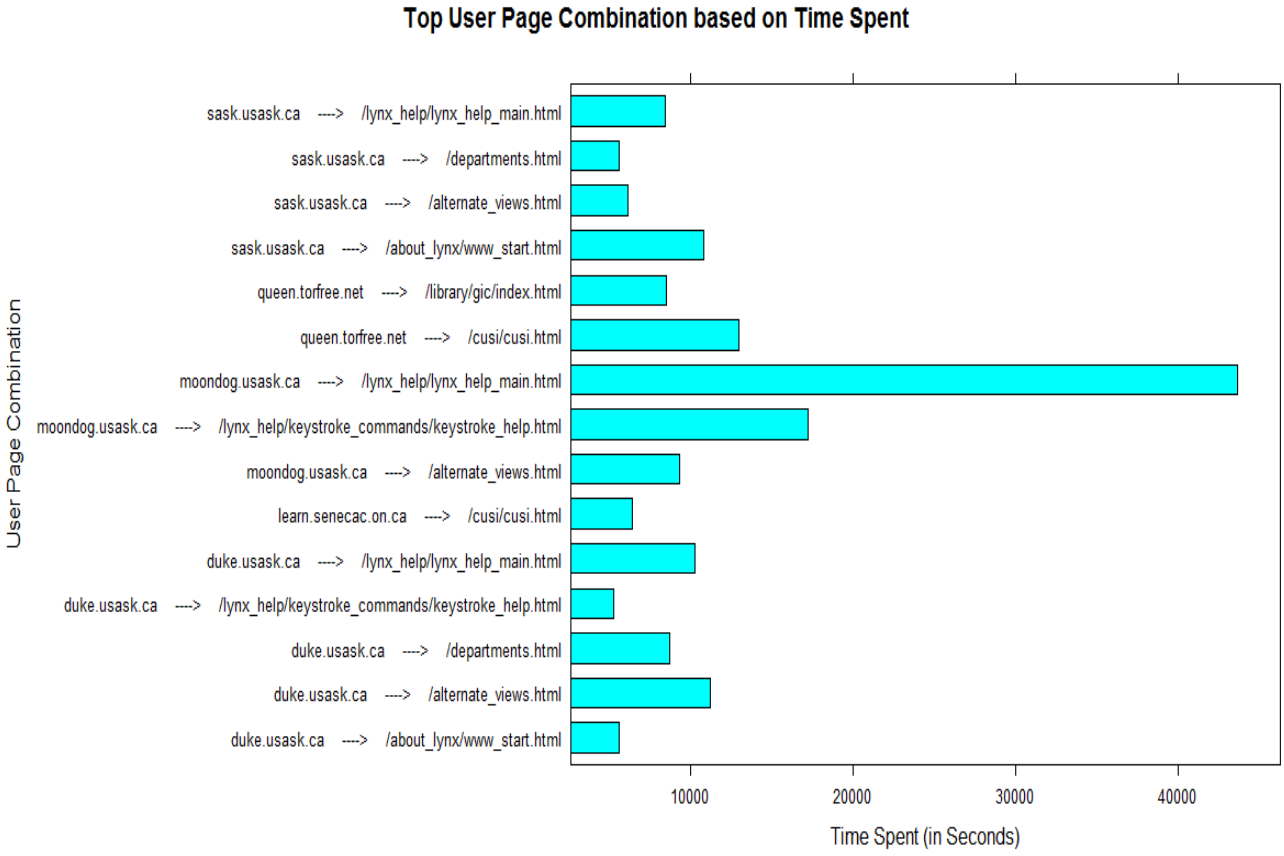


Fig. 7. Top user page combination based on time spent.

g. Clustering of Users Month Wise

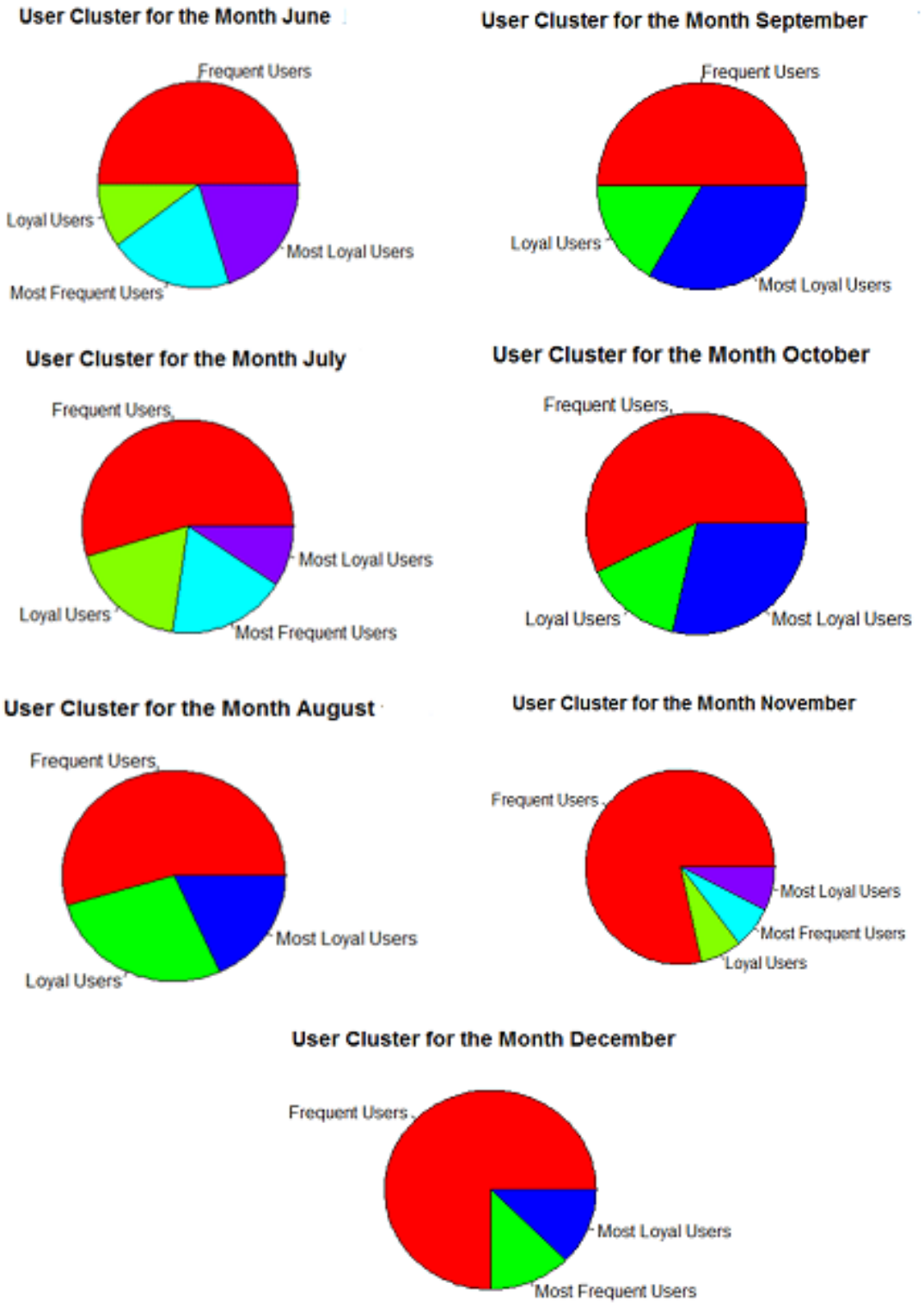


Fig. 8. Month wise Clustering of Users.

h. Clustering of Pages Month Wise

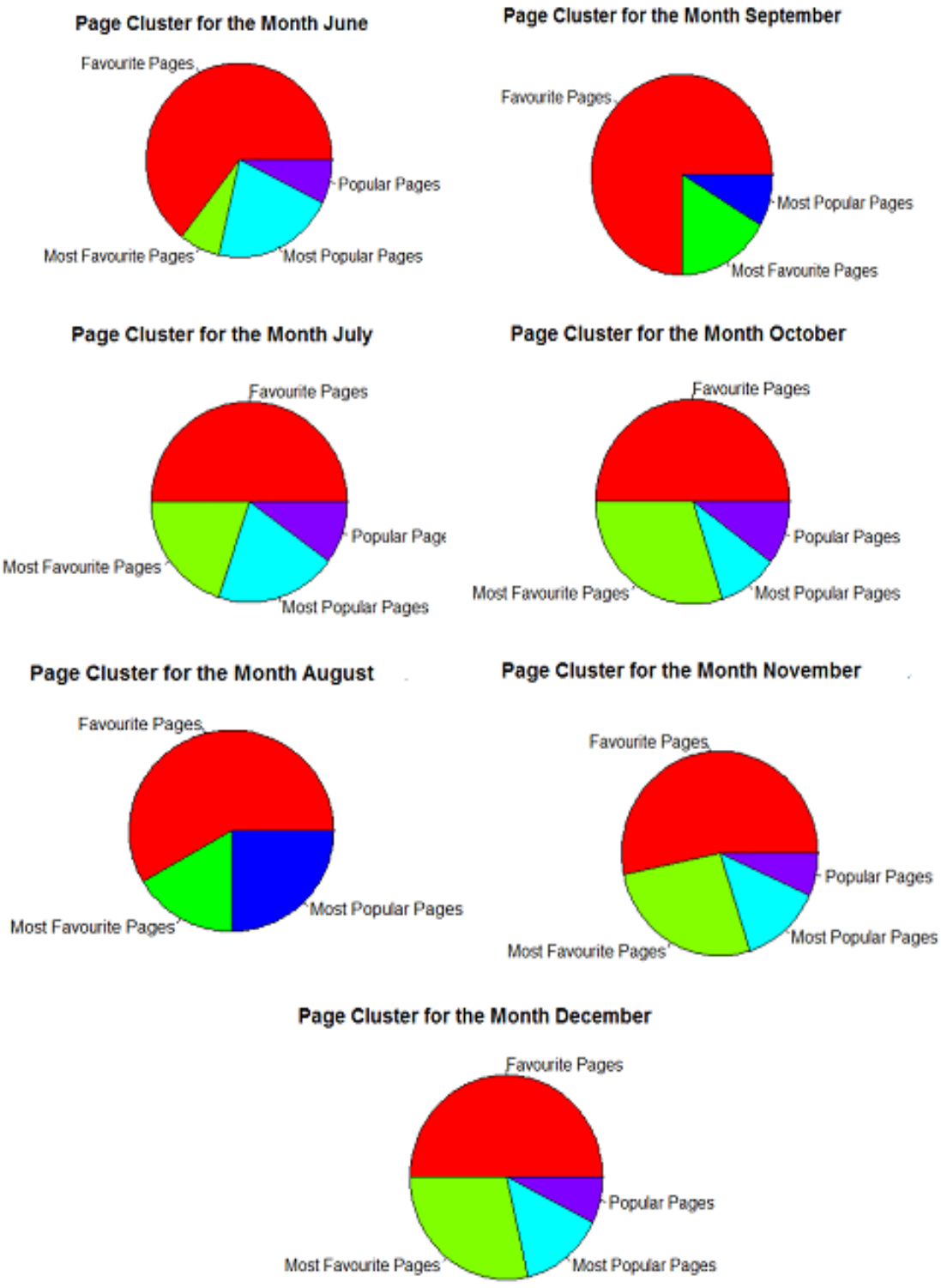


Fig. 9. Month wise Clustering of Pages.

V. DISCUSSION

a. From the bar graphs, it is evident that the web page "/departments.html" in the university web site is the one that is mostly visited by the various users.
b. From the bar graphs, it is evident that the web page "/departments.html" in the university web site is the one that is in which more time is spent by the various users.
c. The user "duke.usask.ca" is the one who has visited the various pages of the university web site frequently.
d. The user "duke.usask.ca" is the one who has spent more time in the web pages.
e. The user / page combination that has the most t visit frequency is "moondog.usask.ca → /lynx_helplynx_help_main.html". That is the user "moondog.usask.ca" has visited the page "/lynx_helplynx_help_main.html".
f. The user / page combination that has the most time spent is "moondog.usask.ca → /lynx_helplynx_help_main.html". That is the user "moondog.usask.ca" has spent most of the time in the page "/lynx_helplynx_help_main.html".
g. From the User Cluster pie charts we can see how the different groups of users are migrating from one group to other over the period of time. The percentage of the Most Loyal Users and the Loyal Users are seen to differ over the months. In the months of August, September and October the Most Frequent Users group is missing and in the month of December, the Loyal Users group is not found in the data.
h. From the Page Cluster pie charts we can see how the different groups of pages are migrating from one group to other over the period of time. Also, the percentage of the Most Popular Pages and the Popular Pages are seen to differ over the months. In the months of August and September, the Popular Pages group is missing.
The reason for the above observations can be further explored by applying data mining techniques.
1) "In the months of August and September, the Popular Pages group is missing"
2) "In the months of August, September and October the Most Frequent Users group is missing"
3) "In the month of December the Loyal Users group is missing".

V. CONCLUSIONS

In this paper we have done basic analytics to analyze the user preferences from the freely available web log file of the University of Saskatchewan's located in Canada. The R Package which is an open source data analytics tool is used to explore the user behaviors in this web site. The techniques applied in this work are basically advanced query processing using the *sqldf* package and visualization using the *lattice* package. Thus after this analysis a web site owner can know about the most preferred web page by the users and how long do they spend in the various web pages. Moreover the study of the change in the user preferences can help the web site manager to understand the reason behind the shift in preferences over time. As a future scope of this work, more such insights can be explored by applying the data mining techniques such as clustering, association rule mining and outlier detection from the R software and the same can be applied on several real time web log files from any web server.

REFERENCES

[1] C. Jothi Venkateswaran, G. Sudhamathy, "Pre Processing of Web Logs – An Improved Approach For E-Commerce Websites", *International Journal of Engineering and Technology (IJET),* vol. **7**, no. 1, pp. 234-244, 2015.

[2] G. Sudhamathy, C. Jothi Venkateswaran, "Fuzzy Temporal Clustering Approach for E-Commerce Websites", *International Journal of Engineering and Technology (IJET)*, vol. **4**, no. 3, pp. 119-132, 2012.

[3] C. J. Aivalis, "Log File Analysis of E-commerce Systems in Rich Internet Web 2.0 Applications", Informatics (PCI), pp. 222-226, 2011.

[4] Rui Wu, "Mining generalized fuzzy association rules from Web logs", Fuzzy Systems and Knowledge Discovery (FSKD), Seventh International Conference, Yantai, Shandong, 5, pp. 2474 – 2477, 2010.

[5] Esmin, A., Lima, J., Yano, Tiago, E. T., Carneiro, G.S., "ArchCollect - A Tool for WEB Usage Knowledge Acquisition from User's Interactions", Proceedings of the Tenth International Conference on Enterprise Information Systems, Barcelona, Spain, pp. 375-380, 2008.

[6] O. Nasraoui, M. Soliman, E. Saka, A. Badia, R. Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", Knowledge and Data Engineering, IEEE Transactions, pp. 202-215, 2008.

[7] V. Pascual-Cid, "An information visualisation system for the understanding of web data", Visual Analytics Science and Technology, pp. 183-184, 2008.

[8] Qingtian Han, Xiaoyan Gao, Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", 9th International Conference on Computer-Aided Industrial Design and Conceptual Design, CAID/CD, IEEE, 2008.

[9] Shafiq Alam, Gillian Dobbie, Patricia Riddle, "Particle Swarm Optimization Based Clustering of Web Usage Data", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WIIAT, pp. 451–454, 2008.

[10] Lee, R. S. T., and Liu, J. N. K., "iJADE Web-Miner: An Intelligent Agent Framework for Internet Shopping", *IEEE Transactions on Knowledge and Data Engineering*, **16**(4), pp. 461- 473, 2004.

[11] Pawan Lingras, Mofreh Hogo, Miroslav Snorek, "Temporal Cluster Migration Matrices for Web Usage Mining", Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, WI'04, 2004.

[12] Abraham, A., "i-Miner: A Web Usage Mining Framework Using Hierarchical Intelligent Systems",

IEEE International Conference on Fuzzy Systems, FUZZ-IEEE '03, IEEE Press, pp. 1129-1134, 2003.

[13] Eirinaki, M., Vazirgiannis, M., and Varlamis, I., "SEWeP: using site semantics and a taxonomy to enhance the Web personalization process", Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 99-108, 2003.

[14] Tiedtke, T., Märtin, C., and Gerth, N., "AWUSA – A Tool for Automated Website Usability Analysis", Proceedings of the 9th International Workshop on the Design, Specification and Verification of Interactive Systems, 2002.

[15] Hong, J. I., Heer, J., Waterson, S., and Landay, J. A., "WebQuilt: A proxy-based approach to remote web usability testing", *ACM Transactions on Information Systems*, **19**(3), pp. 263-285, 2001.

[16] Berendt, B., "Web usage mining, site semantics, and the support of navigation", KDD Workshop on Web Mining for E-Commerce Challenges and Opportunities", pp. 83–93, 2000.

[17] Buchner, A. G., Baumgarten, M., Anand, S. S., Mulvenna, M. D., and Hughes J. G., "Navigation Pattern Discovery from Internet Data", Proceedings of Web Usage Analysis and User Profiling at the International WEBKDD99 Workshop, pp. 74-91, 2000.

[18] Pierrakos, D., Paliouras, G., Papatheodorou, C., and Spyropoulos, C. D., "KOINOTITES: A Web Usage Mining Tool for Personalization", *Proceedings of Panhellenic Conference on Human Computer Interaction, Greece, Patras,* pp. 231-236, 2000.

[19] Shahabi, C., Faisal, A., Kashani, F. B., and Faruque, J., "INSITE: A Tool for Real-Time Knowledge Discovery from Users Web Navigation", *Proceedings of the 26th International Conference on Very Large Databases (VLDB), Cairo, Egypt*, pp. 635-638, 2000.

[20] Masseglia, F., Poncelet, P., and Cicchetti, R., "WebTool: An Integrated Framework for Data Mining", Proceedings of the 10th International Conference on Database and Expert Systems Applications (DEXA '99), Trevor J. M. Bench-Capon, Giovanni Soda, and A. Min Tjoa (Eds.). Springer-Verlag, London, UK, pp. 892-901, 1999.

[21] Zaiane, O.R., Xin, M., and Han, J., "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", Proceedings of Advances in Digital Libraries Conference, (ADL"98), Santa Barbara, CA, pp. 19–29, 1998.

[22] Wu, K. L., Yu, P. S., Ballman, A., "SpeedTracer: A Web Usage Mining and Analysis Tool", IBM Systems Journal, 37(1), pp. 89-105, 1997.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 11-15(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Perspectives of Big Data and Analytics in the Higher Education Sector and an Overview of the Opportunities and Challenges in its Implementation

Manonmani. M[1] and Sarojini. B[2]
[1]Reasearch Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India.
[2]Assistant Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: Over the past few decades, big data and analytics have started to positively encroach on the huge amount of data present in the public education system. There has been a major change in the field of education and its management due to the evolving nature of ICT and its impact in the field of higher education sector. This substantial change has created a great impact on the development of higher education taking into consideration the vast amount of data that need to be integrated into the existing system and the delivery of viable information required for the overall development of the educational institutions. This paper focuses on the concepts of Big Data and Analytics and an understanding of its implementation within higher education. It discusses the opportunities in this growing research area as well as major challenges associated with its exploration and implementation.

Keywords: Information and Communication Technology (ICT), Big Data, Analytics.

I. INTRODUCTION

Big Data and learning analytics are used in the field of teaching- learning systems and produced models that might be of great use to the overall development of the educational institutions.

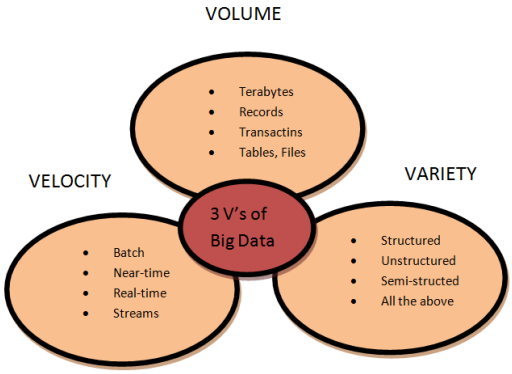
In the recent years, Big Data is placing an important foot in the field of educational research as it uses data analysis to perform decisions. It is currently being explored mostly in business, government and health care due to the growing plethora of data collected and stored in these environments. The main aim of Big Data research is the examination of the vast amount of voluminous data that it stored in organizational database and to identify recurring behavioral patterns and meaningful trends as the need be.

Ben Daniel (2014) has envisaged that Educational Institutions have started to use analytics for improving the services they provide and for increasing student grades and retention. With analytics and data mining experiments in education starting to increase manifold, sorting out the required data and identifying research possibilities and practical applications are not an easy task. This issue brief is intended to help policymakers and administrators understand how big data and analytics have created reasonable opportunities surpassing the challenges and drawbacks to achieve the goals for educational improvement.

Big data are found in all segments of society. The digital revolution has brought about a tremendous increase in the volume of data that is available for processing and if harnessed, potential outcomes can be arrived which will form a valuable resource for the data analyst and end users. The dependence on ICT cannot be ignored in any sector, same is the case with the higher higher education sector too. Student information systems, the online learning environment and the library generate a wealth of potentially interesting data. When all systems are interconnected, we can gain insight into the students' learning behavior, the quality of teaching and the institution's effectiveness. It obviously is not so easy to gather, analyze and report data from learning environments as suggested here. Before arriving at the desired results, it is important that the educational and technical challenges be taken into account and it is required that these challenges be overcome to produce better results.

II. BIG DATA – MEANING

The term "big data" is generally used to describe data sets so large they must be analyzed by computers. These huge voluminous data may be analyzed to produce unique patterns, trends, and associations, especially relating to human behavior and interactions.
Big data consists of 3 Vs,
1. Extreme **Volume** of data
2. Variety of data types
3. Velocity at which the data must be analyzed and processed.



In addition to the 3 Vs of Big data, there are additional Vs that IT, business, educational institutions and data scientists need to be concerned with, most importantly big data Veracity.

An overview of the 6 V's of Big Data is given below:

Volume. Big data is concerned with enormous volumes of data. Earlier the data used to be created by employees. But now the same data is generated by machines, networks and from social media. This implies that the volume of data to be analyzed in Big Data is massive.

Variety. Variety refers to both structured and unstructured types of data and also the various sources from which data is collected. The data stored in spreadsheets and databases are now available in the form of audio, video, images, messages and the like. This wide variety of data from a variety of sources are of concern in terms of storage, mining and analysis.

Kevin Normandeau (2013) has analyzed other features of Big Data as explained below:

Velocity. Big Data Velocity implies the rate at which data flows in from sources like business processes, machines, networks, mobile devices, social media and the like. The flow of data from these sources is massive and continuous. This real-time data will be of use to the researchers and business organizations who can take decisions that contribute towards the overall development of the organization.

Veracity. Big Data Veracity refers to the bias if any, noise and abnormality in the data that is taken into consideration. Big Data ensures that at each stage of analysis that data that is being stored and mined is meaningful to the problem at hand. Veracity in big data is the biggest challenge in data analysis when compared to the huge voluminous data and velocity. Data that is stored and mined should pass through the data cleaning step to avoid noisy data and unwanted data accumulated in the systems.

Validity. Validity refers to data that is kept for use must be correct and accurate. Valid data is very important in proceeding with data analysis and then only viable information can be produced with the available data source.

Volatility. Big data volatility refers to the time period of the valid data that needs to be stored and made available. In real time data, it is important to determine the time frame of relevant data that needs to be stored for current analysis.

III. GOALS AND PURPOSES OF BIG DATA IN HIGHER EDUCATION

Big Data essentially consists of three major stages in the higher education sector as depicted in fig.1.

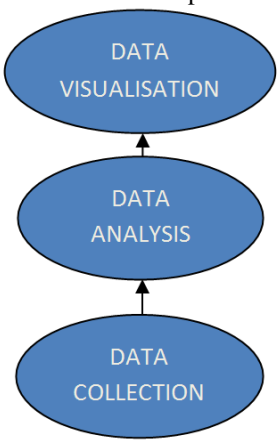


Fig.1. Three major stages of Big Data in the higher education sector.

The Value of Big Data in Education. Athanasios S. Drigas and Panagiotis Leliopoulus (2014) has investigated into the use of Big Data in the higher education sector. Big Data has the future to change not just research, but also education. A late accurate significant similarity of many approaches taken by 35 charter schools in NYC has discovered that one of the top five policies connected with significant academic effects was the use of data to guide instruction.

Big Data can support the classic educational system, helping teachers to analyze what students know and what techniques are most effective for each pupil. In this way, teachers are also able to learn new techniques and methods about their education work.

Hence Technologies such as Data mining and Data analytics can provide a fast feedback to students and teachers about their academic performance as depicted in the first stage of Fig1. These methods can provide a deep analysis of some education patterns and extract valuable knowledge from them. In this way, collective and big scale data can predict who student needs more help from the education system, avoiding the danger of failure or drop out. This has as a result to find pedagogic approaches that seem most effective with particular students and special needs.

On the other hand, as Siemens and Gasevic say "Big Data can easily find apply at online education." Online education has played a major role in the field of higher

education. Furthermore, digital learning is actually a collection of data and analytics, which can contribute to teaching and learning. In this way many students participate in online or mobile learning, where are crated new data. These new data, also with the help of social networks, are helping the students with the different background to correlate between them and help them to understand core course concepts.

IV. ANALYTICS

Analytics is a field of data analysis. Analytics involves study of the past historical data to research potential data, to analyze the results of certain decisions or events, or to analyze the performance of a given tool or scenario. The aim of analytics is to improve the business by gaining knowledge which can be used to make improvements or changes in the existing system. Analytics combined with Educational Big Data generally refers to the process of collecting the data from the institution, conducting detailed analyses, generating corresponding insights, and using the generated new information to make viable decisions which will reflect a positive change in the entire system of data information system used by all the stakeholders of the institution.

Learning analytics. B.R. Prakash *et al*. (2014) has commented that as higher educational institutions adopt to the integration of ICT in all levels of teaching – learning scenarios, learning is taking place at a faster rate with the availability of online learning tools. The educational data mining community and learning modelling communities have already explored ways to track student behaviors, recognize patterns to improve the overall academic performance of the students, track the previous records of the institution to determine the admission to particular course and provide predictive models which enable the students in the selection of their career etc. Inclusion of behavior-specific data add to an ever-growing repository of student-related information. Learning analytics is an emerging research area that intends to access and understands these data and adds a new dimension to the learning process.

Learning analytics consists of the measurement, collection, and analysis and final report of data about students and their contexts, which is mostly done with the main objective of understanding and optimizing learning and the environments in which it occurs. Learning analytics software and techniques aim towards increasing the utility of processes and workflows, analyzing academic and institutional data and help in the overall improvement of the organizational capability.

Learning analytics is used more in the teaching and learning level of an institution and is largely concerned with improving learner success. Another aspect of analytics is teaching analytics, which is an extended form of learning analytics. In teaching analytics, teacher's online behaviors are analyzed in the context of utilizing digital library and online resources. Educational data mining techniques are used to identify different groups of instructional architectures to determine diverse online behaviors.

Opportunities. Large volumes of student information are present in the higher educational institution, including student enrollment, academic and disciplinary records and other important data sets needed to benefit from targeted analytics. Big Data and analytics in higher education can help in the transformation of the existing processes of administration, teaching, learning, academic work contributing to policy and practice outcomes and helping address contemporary challenges facing higher education.

Big Data ensures the provision of predictive tools needed by the institutions of higher education to improve the learning outcomes of the individual students as well as ensuring the delivery of high-quality education standards. By developing programmes that collect data at every step of the students learning processes, universities can address the needs of the students with customized modules, assignments, and feedback in the curriculum that will promote better and richer learning.

One of the ways higher education can utilize Big Data tools is to analyze the performance and skill level of individual students and create a personalized learning experiences that meet their specific learning pathways. Effective use of Big Data can help institutions enhance the learning experience and improving student performance, reduce the rate of dropout and increase the number of graduates.

V. HIGHER EDUCATION - BIG DATA ANALYTICAL MODELS

Big Data Analytics envisages three major models in higher education as discussed below:

Descriptive analytics. Descriptive analytics aims at describing and analyzing historical data collected on students, teaching, research, policies and other administrative processes. The goal is to identify patterns from samples to report on current trends—such as student enrollment, graduation rates and progressions into the higher degrees.

Descriptive analytics also provide institutions of higher education with an opportunity to analyze transactions and interactional data about teaching, learning and research to identify discernible trends and patterns that are likely to trigger important dialogue on current and future issues. More importantly, descriptive analytics enable institutions to investigate data within learning management systems by looking into the frequency of logins, page views, course completion rates for a particular programme over time, student's

attributes compared with those who have finished and those who find it difficult to compete, repeated access to a particular material, etc.

Predictive analytics. Predictive analytics can provide institutions with better decisions and actionable insights based on data. Predictive analytics aims at estimating likelihood of future events by looking into trends and identifying associations about related issues and identifying any risks or opportunities in the future. Predictive analytics could reveal hidden relationships in data that might not be apparent with descriptive models. It can also be used help to look at students who are exhibiting risk behaviors early in the semester that might result to dropping out or failing a course. It can help teachers look at predicting course completion rate for a particular tool and content in the course are directly correlated to student success.

Prescriptive analytics. Prescriptive analytics help institutions of higher education assess their current situation and make informed choices on the alternative course of events based on valid and consistent predictions. It combines analytical outcomes from both descriptive and predictive models to look at assessing and determining new ways to operate to achieve desirable outcomes while balancing constraints. Basu (2013) indicated that prescriptive analytics enable decision makers to look into the future of their mission critical processes and see the opportunities (and issues) as well as presents the best course of action to take advantage of that foresight in a timely manner.

Comprehensively, Big Data Analytics provides institutions of higher education to leverage existing data and collect missing data to help make better decisions with various outcomes.

VI. BIG DATA ANALYTICS – OUTCOMES IN THE HIGHER HIGHER EDUCATION SECTOR

.M. West (2012) has outlined the major outcomes in the higher education with regard to the use of Big Data Analytics. Big Data finds use in the institutions with the following performance and process outcomes:

Performance outcomes

- Vast amount of institutional data can be understood more clearly.
- Understand academic problems early and take remedial action
- Encourage better teaching-learning techniques to improve performance
- An understanding of the data required for analytics.
- Standardized and streamlined data processes can be generated.
- Analyze enormous experimental data sets in near real-time using predictive models.
- Eliminate errors through automated data acquisition.
- Differentiate research capabilities versus other institutions.
- Data-driven decision making and practice can be achieved.
- Basis for hypothesis testing, web experimenting, scenario modelling, simulation, sensibility and data mining.

Process outcomes

- Increased use of historical, institutional data to make viable decisions.
- Tools that are required for collecting, processing, analyzing, and interpretation of data can be developed as per the requirement.
- Improved database management and system interconnectivity.
- Increased data analytics and predictive modeling.
- Results of instructor performances and understanding of learning materials can be analyzed to better decide upon further enhancements.
- Within the departments, performance indicators and metrics can be retrieved and analyzed.
- Predicting possible outcomes from the institutional data that are explored for future enhancements.

Challenges of implementation. There are a number of anticipated challenges associated with the implementation of analytic techniques for Big Data in higher education.

Some of these challenges are associated with getting users to accept Big Data as a channel for adopting new processes and change management. Apart from the high cost involved in collecting, storing and developing an algorithm to derive the required data, there tends to be a high rate of complexity involved in mining these voluminous data not neglecting the time factor associated with data collection. Furthermore, many times the institutional database management systems are not interoperable and hence there lies a great challenge in aggregating administrative data.

Furthermore, data integration challenges are eminent, especially where data come in both structured and unstructured formats and needed to be integrated from varied sources, most of which are stored in the systems that are managed by different departments. Additionally, data cleansing when performing integration of structured and unstructured data is likely to result in loss of data.

There are also challenges associated with quality of data collected and reported. Lack of standardized measures and indicators make international comparison

difficult, as the quality of information generated from Big Data is totally dependent on the quality of data collected and the robustness of the measures or indicators used.

The successful implementation of Big Data in higher institution would depend on collaborative initiatives between various departments in a given institution. However, there is still a gap between those who know how to extract data and what data are available, and those who know what data are required and how it would best be used, all which makes collaboration difficult.

VII. CONCLUSION

In the present era of dependence on ICT and related tools and use of internet on a daily basis, higher education institutions are increasingly connecting with their students that in turn imply that there are vast opportunities for the collection and utilization of big data in university recruitment and far reaching innovations in the higher education sector.

The paper enables the higher education institutions to understand how Big Data can be utilized to address the implementation challenges when the major factor of implementation is the amount of data that has to be explored. The paper will help policymakers and information technology analyst to make viable choices during the implementation of Big Data programmes in their institutions.

Big Data is reshaping the way students involve in the teaching-learning process given the provision for inclusion of new tools and digital classrooms. The increasing involvement of analytics in business and government sectors agrees to the significant role that big data analytics plays in higher higher education sector too.

Education technology companies are leveraging big data to improve classroom engagement and make learning more effective.

The increasingly competitive nature of institutions recruiting a limited number of students will assure that those that take advantage of these new data sources to augment what they already know will be an added advantage to the existing system. They will continue to invent new and better business processes and efficiencies and they will do so by evolving their Information Architecture in an impactful manner.

The only concern lies in the acquisition of the right skills and resources in order to manage the increased volume of data and how best the data can be analyzed and made effective for the end user.

REFERENCE

[1]. Ben Daniel, "Big Data and analytics in higher education: Opportunities and Challenges", British Journal of Educational Technology (2014).

[2]. B.R. Prakash *et al.,* "Big Data in Educational Data Mining and Learning Analytics", IJIRCCE, Vol. **2**, Issue 12, Dec 2014.

[3]. Kevin Normandeau, "Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity" Sep 12, 2013, inside BIGDATA.

[4]. Athanasios S. Drigas and Panagiotis Leliopoulus, "The Use of Big Data in Education", IJCSI International Journal of Computer Science Issues, Vol. **11**, Issue 5, No 1, September 2014.

[5]. J. Bughin, M. Chiu, and J. Manyika, "Clouds, big data, and smart assets", Ten tech-enabled business trends to watch", Mc.Kinsy Q., Vol. **56**, 2000.

[6]. D.M. West, "Big Data for Education: Data Mining, Data Analytics, and Web Dashboards", Gov. Stud. Brook. US Reuters, 2012.

[7]. Jay Liebowitz, "Thoughts on Recent Trends and Future Research Perspectives in Big Data and Analytics in Higher Education".

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 16-21(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Online Shopping – Attitude, Intention and Behaviour

J. Jenica[1] and Dr. P. Chitramani[2]
[1]Research Scholar, Avinashilingam School of Management Technology,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India
[2]Professor, Avinashilingam School of Management Technology,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India

ABSTRACT: Online retailing is the trend in towards marketing space and e-retailing is an emerged as the most important marketing and sales tool for corporate entities. The research paper presents the, perceived ease of use and usefulness, risk, trust, attitude online shopping behaviour and shopping intention. An online survey was carried out among online shoppers and the paper presents their perceptions on online shopping. The results indicate that convenience of the internet is one of the primary reasons for consumers' willingness to buy online while fear of identity and financial theft, product genuineness, lack of touch and feel, delivery time and fixed price format holds them back. It's imperative for online vendors to understand the factors that influence the formation of consumer's behavioural intention toward online shopping and implement them as part of their marketing strategy.

Keywords: Online shopping, Perceived use and usefulness, Trust and Risk, Online shopping attitude, intention and behavior.

I. INTRODUCTION

The emergence and rapid growth of internet based electronic commerce (e-com) has created a paradigm shift in the way consumers shop since the last decade [1], [3], [4], [14], [15], [16], [18], [19]. Customers are no longer tied to the opening hours or specific locations of retail stores, but are active virtually [6], [8], [10], [12], [17].

Online stores have reduced operating costs compared to traditional retail outlets, cutting on labour and store rental costs which allows for lower prices offered to consumers [7]. E-commerce has become an important marketing and sales channel, and it is important for retailers to understand the determinants of online purchasing to meet consumer's needs and target consumers effectively [5], [20].

Among the BRIC nations, India has the fastest growing e-commerce market adding over 18 million internet users and growing at an annual rate of 41% [12] and is expected to touch $56 billion by 2023 [12]. Consumer e-commerce is perceived to have wider and stronger impact on the retail economy [11]**.** Given the scenario, understanding Indian consumer's online shopping behaviour is found to be apt from a retailer as well as researcher's point of view.

II. REVIEW OF LITERATURE

The Internet offers more interactivities between consumers and product/service providers and has greater transparency of information sharing about products/services (Chen, 2009). For e-commerce managers, a better understanding of online consumer behaviour is critical to effectively attract and retain online consumers (Kacen, Hess and Kevin – Chiang, 2013; Kamari and Kamari, 2012; Rose, Clark, Samouel and Hair, 2012; Dawn and Kar, 2011; Tseng, Kao, Lee and Wu, 2011; Wu and Cheng, 2011; Ling, Chai and Piew, 2010; Constantinides, 2004, Li, Jiang and Wu, 2014; Zhang, Shi and Lu, 2014; Kim, Han and Lee, 2014; Yoon and Steege, 2013; Gao, Zhang, Wang and Ba, 2012; Lee, Shi, Cheung, Lim and Sia, 2011).

Researchers have relied on the Theory of Reasoned Action (TRA), Theory of Planned Behaviour (TPB), Technology Acceptance Model (TAM), Expectation – Confirmation Theory (ECT), Innovation Diffusion Theory (IDT) and Transaction Cost Theory (TCT) to understand the determinants of online purchasing. (Laroche and Richard, 2014; Dennis, Merrilees, Jayawardhena and Wright, 2009; Cheung, Chan and Limayem, 2005).

Many studies have extended TAM and TPB by identifying major antecedents or mediating factors that have improved the understanding of the determinants of

online purchase intentions (Ho and Chen, 2014; Safeena, Date and Kammani, 2013; Al – Ajam and Nor, 2013; Lee, Eze and Ndubisi, 2011).

Prior research findings advocate trust as an important construct that has great influence on online purchase intentions. The three trusting beliefs that are used most often in literature namely competence, integrity and benevolence (Ho and Chen, 2014; Bhattacherjee, 2000; McKnight, Choudhury and Kacmar, 2002; Pavlou, 2003, Kaur and Madan, 2014; Hsu, Chuang and Hsu, 2014; Chang, Cheung and Tang, 2013; McCole, Ramsey and Williams, 2010; Wu and Chen, 2005; Gefen and Straub, 2004). Lack of trust in the online transactions and the web vendors is an important obstacle in e-shopping (Lan and Yizeng, 2014; Khare, Mishra and Singh, 2012; Becerra and Korgaonkar, 2011; Liao, Liu and Chen, 2011; Palvia, 2009; Chen and Barnes, 2007; Liu, Marchewka, Lu and Yu, 2004).

Risk, in general, means the perceived probability of loss or harm (Rousseau *et al.,* 1998). Perceived web risk means the extent to which a user believes it is unsafe to use the web or that negative consequences are possible (Grazioli and Jarvenpaa, 2000). Identity theft has risen greatly over the past few years (Sharma and Lijuan, 2014; Karakaya and Stark, 2013; Tajpour, Ibrahim and Zamani, 2013; Wang and Nyshadham, 2011; Chen, Xie and Jing, 2013). The five evaluation criteria for the model of perceived risk: understandability, predictability, reliability and effectiveness, practicality and availability (Mitchell, 1999,. Perceived risk in e-commerce has a negative effect on shopping behaviour on the Internet, attitude toward usage behaviour and intention to adopt E-commerce (Nepomuceno, Laroche and Richard, 2014; Chang and Fang, 2013; Chang and Wu, 2012; Almousa, 2011; Glover and Benbasat, 2010; Featherman and Wells, 2010; Crespo *et al.,* 2009; Verhagen, Meents and Tan, 2006). Pavlou (2003) identified them as economic risk, seller performance risk, privacy risk and security risk while. Bhatnagar, Misra and Rao (2000) differentiated two types of risks: product category risk and financial risk.

III. METHODOLOGY

The purpose of the research is to identify the variables influencing the online shopping intention and ascertaining the perceived risk and trust through an online survey. The final questionnaire employed pre-validated measurement scales the theoretical constructs from the Technology Acceptance Model (TAM) was measured using a 5-point scale. The responses delivered were in the form of a range of agreement, where 5 meant strongly agree and 1 meant strongly disagree. The construct perceived risk was measured using reverse scaling. A pilot study was carried out with 50 online shoppers residing in Coimbatore city to collect preliminary data. The reliability value for all the variables were found to be above 0.70. The high Alpha values (close to or greater than 0.7), as per Nunnally and Bernstein (1994), indicate that internal consistency of the factors is high.

Table 1: Source of Questionnaire Items.

Source	Variable (Items)	Alpha values
Lim and Ting, 2012	Perceived ease of use (7)	.70
Lim and Ting, 2012	Perceived usefulness (6)	.78
Javadi *et al.,* 2012 Forsythe *et al.,* 2006	Perceived Risk (20)	.84
Chen and Barnes, 2007	Trust (9)	.73
Lim and Ting, 2012; Velarde, 2012	Attitude (7)	.73
Lim and Ting, 2012	Online purchase intention (7)	.76
Javadi *et al.,* 2012	Online shopping behavior (2)	.79

Online survey is suitable for collecting data from individuals who had purchased online in the last six months at the moment of data collection and are residing in Coimbatore city. The sample size and adequacy was determined using G-Power. A sample with alpha value of 5% and beta of 5% and the largest predictors of 20 with a medium effect size ($f^2 = 0.15$) measured to 222 samples. The present study was conducted for a total number of 303 online with an effective sample size of 9.9%. Snowball sampling technique was adopted, whereby the link to the web-based questionnaire was sent to acquaintances residing in Coimbatore city.

IV. DISCUSSIONS

The empirical data based on Descriptive Statistics and the discussions are presented is as follows:

A. Perception of Perceived Ease of Use and Usefulness

Theory of Reasoned Action (TRA) model postulates perceived usefulness and perceived ease of use as the two external variables that determine attitude towards intention to use and actual use. Perceived Ease of Use (PEOU) refers to the degree to which a person believes that using a particular system would be free from effort (Davis, 1989, Wixom and Todd, (2005), Davis, (1989). PEOU is the customer's perception that online shopping will involve a minimum of effort. (Chen *et al.,* 2002; Chen and Tan, 2004; Kim and Forsythe, 2007; Koufaris, 2002; O'cass and Fenech, 2003; Vijayasarathy, 2004; Davis *et al.,* 1989; Kleijnen *et al.,* 2004; Wang *et al.,* 2003).

Table 3. Perceived Ease of Use and Usefulness.

Variables	Mean
Shopping sites easy to use	4.42
Easy learning to use	4.20
Easy to find what I want	4.21
Easy to become skilful	4.12
Easy to compare products	4.42
Flexible to interact with	3.86
Browse with ease	4.30
Accomplish shopping goals quickly	4.48
Improve shopping performance	4.48
Increase shopping productivity	4.31
Increase shopping effectiveness	4.14
Website useful in aiding purchase decisions	4.12
Easier to satisfy needs	4.36
Overall	**4.26**

The prospect of accomplishing shopping goals quickly and improved shopping performance, convenience and time saving; thereby enabling the consumer's to achieve their shopping objectives without any hassle is cited positive. Easy site usage and easy product comparison , is in line with the work of Heijden (2000) who suggested that the easier it is for consumers to use online shopping sites, the more useful online shopping will be perceived by consumers. Ramayah and Ignatius (2005) contested that ease of use of the technology and degree to which the shopper is satisfied with the online shopping experience is imperative in predicting the potential e-shopper's intent.

Ease of browsing online shopping sites is consistent with the work of Childers *et al.* (2001) who argued that online shopping sites which are easy to browse and operate, with less mental effort requirement, and allows consumers to shop the way they want results in favourable attitudinal attachment to online retailers. Henderson and Divett (2003) indicate that the easier the customer finds items for purchase in a web store, the stronger will be the user-performance relationship. Easy learning to use online shopping and increase in shopping productivity and easy to find products of choice are consistent with previous research findings that consumers can improve their shopping productivity when they find it easy to learn and use new technology (Bertrand and Bouchard, 2008).

Shopping effectiveness and flexibility of websites indicate perceived ease of use is posited to have a direct impact on perceived usefulness. The flexibility of the web interface will have a positive impact on the consumer's ability to use the websites thereby increasing their shopping effectiveness (Kurnia and Chien, 2003).

B. Online Trust and Risk

Trust and risk are mirror images in which an inverse relationship exists. Trust is o influences perceived risk that, in turn, influences behaviour (Joubert and Belle, 2013), the paper highlights consumer's propensity towards vendor attributes.

Table 4. Perceived Trust.

Variables	Mean
Shopping sites trustworthy and honest	4.34
Sufficient information	4.44
Keeps promises and obligations	4.59
Infrastructure dependable	4.56
Provides secure personal privacy	4.60
Keeps best interests in mind	4.63
Shopping sites are secure and reliable	4.58
Does not misuse personal information	4.50
Performance meets expectation	4.29
Overall	**4.50**

Online shoppers express that the e-retailers keep the best interest of consumers by providing personal privacy and keeping their delivery promises and obligations and are secure and reliable with integrity and dependable infrastructure and indicates high level of trust on online vendors

The risk of credit card details compromised and personal details compromised indicate. Security/ privacy risk. Consumer's fear that the open internet network is not secure and their personal information may be compromised when transmitting sensitive information through online transactions (Bhatnagar *et al.,* 2000; Bhatnagar and Ghose, 2004; Kim *et al.,* 2009; Crespo *et al.,* 2009; Forsythe *et al.,* 2006). The perceived risk among consumers translates into their reluctance to use credit card information over the Internet resulting in their disengagement from electronic transactions (Hoffman *et al.,* 1999). The fear of being overcharged indicates financial risk. Crespo *et al.* (2009) found financial risk to have the most influence on overall perceived risk, which in turn was found to predict purchase intention. The product risk indices are: Difficult to find right products, not receiving products ordered , risk of receiving malfunctioning merchandise , problem returning products and not able to examine the products. In online shopping, consumers cannot accurately evaluate the quality of a product prior to purchase, making product risk an important consideration (Bhatnagar, Misra, and Rao 2000, Jarvenpaa *et al.,* 2000). The risk of credit card details compromised and personal details compromised indicate. Security/ privacy risk. Consumer's fear that the open internet network is not secure and their personal information may be compromised when transmitting sensitive information through online transactions (Bhatnagar *et al.,* 2000; Bhatnagar and Ghose, 2004; Kim *et al.,* 2009; Crespo *et al.,* 2009; Forsythe *et al.,* 2006). The perceived risk among consumers translates into their reluctance to use credit card information over

the Internet resulting in their disengagement from electronic transactions (Hoffman *et al.,* 1999).

Table 5. Perceived Risk.

Variables	Mean
Credit card details compromised	3.89
Personal information compromised	2.89
Might get overcharged	2.93
Might receive malfunctioning merchandise	2.53
Might not get what is ordered	2.82
Hard to judge quality	1.91
Difficult to find right products	2.72
Cannot examine the products	1.86
Not easy to cancel orders	2.65
Difficult settling disputes	2.46
Wait for merchandise	2.02
Not receive products ordered	2.63
Problem returning products	2.41
No reliable and well equipped shippers	2.79
No free shipment service	2.38
Return products without frills and strings	2.07
No money back guarantee	1.68
Too complicated	2.69
Difficult to find appropriate websites	3.00
Images take too long to load	2.87
Overall	**2.56**

The fear of being overcharged indicates financial risk. Crespo *et al.* (2009) found financial risk to have the most influence on overall perceived risk, which in turn was found to predict purchase intention. The product risk indices are: Difficult to find right products, not receiving products ordered, risk of receiving malfunctioning merchandise, problem returning products and not able to examine the products. In online shopping, consumers cannot accurately evaluate the quality of a product prior to purchase, making product risk an important consideration (Bhatnagar, Misra, and Rao 2000, Jarvenpaa *et al.,* 2000).

Perceived risk is also influenced by delivery risk factors like delayed product delivery, replacement of defective products, money back guarantee and settling disputes (Forsythe *et al.,* 2006). The delivery risk factors are: Unreliable and ill equipped shippers, not easy to cancel orders, difficult settling disputes, unavailability of free shipment service, returning products without frills and strings and waiting for merchandise.

Consumers tend to assume that a trusted e-retailer will not engage in opportunistic behaviour. Thus trust reduces perceived risk. When an e-retailer can be trusted to show competence, integrity, and benevolence, there is much less risk involved in purchasing and transaction (Kesharwani and Bisht, 2012). Online trust plays a key role in creating satisfied and expected outcomes in online transactions (Pavlou, 2003; Yousafzai *et al.,* 2003; Gefen and Straub, 2004)., and perceived risk is an important trust antecedent which can affect consumers' decision to take part in an online transaction. Even though consumers perceive the Internet as offering a number of benefits, consumers perceive a higher level of risk when purchasing on the Internet compared with traditional retail formats (Lee and Tan, 2003).

C. Perception of Attitude towards Online Shopping

An individual's attitude towards behaviour is determined by his affective beliefs about behavioural consequences and the evaluations (Fishbein and Ajzen, 1980). The attitude towards internet shopping has a significant impact on the intention to web purchase (Liao and Shi, 2009, Shim *et al.* 2001). The empirical data reinforce the assumption that attitude has a positive direct influence on intention to shop online.

Table 6: Attitude.

Variables	Mean
Comfortable to shop	4.46
Purchase what I need	4.63
Seek product information	4.66
Feel happy shopping online	4.73
Shopping online good idea	4.66
Shopping online wise choice	4.51
Positive evaluation	4.57
Overall	**4.60**

Feeling happy to shop online imply that if consumers are exposed to pleasant and happy stimuli during their Internet shopping experience, they are then more likely to engage in subsequent shopping behaviour (Li and Zhang, 2002). Online shoppers always want to seek information within few clicks and reach to the most relevant information according to their requirements (Gao, 2005). Convenience to purchase what is needed, positive evaluation of the e-retailer, wide choice and comfort is indicative of the fact that Customers can satisfy their shopping needs in less time duration than traditional shopping which eventually results in positive evaluation towards online purchase (Lim and Ting, 2012).

Attitude influences the action to purchase and by managing the antecedents of attitude, consumers can be induced to purchase from a virtual store. Favourable attitudes towards shopping online significantly influences consumer's repurchasing behaviour and impulse buying to some extent. This is consistent with research findings that utilitarian and hedonic factors ultimately affect consumers' attitude toward shopping on the Internet (Monsuwe *et al.,* 2004).

D. Perception of Online Shopping Intention

Behavioural intentions are motivational factors that capture how hard people are willing to try to perform a behaviour (Ajzen 1991). A goal intention to purchase a

product from a Web vendor activates an intention to get information about that product from the vendor's website which might eventually lead to final purchase (Salisbury *et al.,* 2001).

Online shopping intention was assessed using seven items probing respondent's inclination in online shopping.

Table 7: Online Shopping Intention.

Variables	Mean
Continue to purchase products	4.50
Intend to continue to purchase	4.50
Visit online shopping sites for need	4.34
Plan to do more shopping online	4.16
Search for an online retailer	4.19
Purchase same product	3.94
Purchase different products	4.18
Overall	**4.25**

Online shoppers expressed positive intention to continue to purchase online, propensity in visiting websites to fulfil their shopping needs, consumers' willingness to buy and to return for additional purchases. Repurchase intention contributes to customer loyalty (Li and Zhang, 2002). This reinforces the assumption that consumers' intention to shop online is positively associated with attitude towards Internet buying, and influences their decision-making and purchasing behavior (Jarvenpaa *et al.,* 2000).

E. Perception of Online Shopping Behaviour

Online Shopping Behaviour depends on factors such as Website visibility, online retailers' credibility, information comparison, payment security, privacy, website interface, product characteristics, convenient time (Wang, 2008).

Table 8: Online Shopping Behaviour.

Variables	Mean
Shop in privacy	4.59
Do not have to leave home	4.57
Shop whenever we want	4.69
Save from traffic chaos	4.57
Save from market crowd	4.46
Detailed product information	4.38
Broader selection of products	4.23
Easy price comparison	4.65
Get user/expert reviews	3.93
No embarrassment	4.55
Can take time to decide	4.47
Buy products not available in nearby market	4.54
Makes shopping easy	4.54
Better control of expenses	3.99
Compatible with lifestyle	4.50
Overall	**4.44**

The comfort of shopping at convenience, easy price comparison, shopping in privacy and less embarrassment in abandoning purchase decisions, affirmation to buy products that are not available in nearby market indicate that online shopping is compatible with their lifestyle.. Online shopping is available for customers around the clock compared to traditional stores (Hofacker, 2001; Wang *et al.,* 2005). Some customers prefer online channels just to escape from face-to-face interaction and controlled marketplace (Goldsmith and Flynn, 2005; Parks, 2008, Lim and Dubinsky, 2004, Childers *et al.,* 2001, Amin, 2009).

V. CONCLUSION

A. Managerial implications

In marketing, a consumer can be satisfies and delighted. The satisfactory factors are hygiene factors and the delight factors are motivational factors. The empirical data presented in the research paper has the managerial implications as follows for e–retailers.

B. Strategies for enhancing online shopping experience

Perceived ease of use and Perceived Usefulness

- **Delight:** Improved shopping performance, Easy site usage, Easy comparison of products.
- **Satisfactory:** Not flexible to interact with, Website not aiding in purchase decisions
- **Recommendation:** Improve website design features, User friendly interface, Personalised website content

Perceived risk and trust

- **Delight:** Provide personal privacy, Keeps promises and obligations and provides sufficient information
- **Satisfactory:** Credit card and personal information not compromised, accurate examination of merchandize, ease of cancelling orders, Problem returning products, Performance meeting expectations, Dependable Infrastructure
- **Recommendation:** Enhance security features, system, information and service quality, make shopping sites secure and reliable, Improve website infrastructure and vendor's integrity.

Online shopping attitude, intention and behaviour

- **Delight:** Happy shopping experience and detailed product information, Shopping in privacy, easy price comparison
- **Satisfactory:** Wide choice with dependable user reviews and less control on expenses
- **Recommendation:** Improve shopping attitude with ease of use features, Increase trust among online shoppers, Offer good payment plans and options for customers to have better control of expenses

Online shopping has made significant contributions to our lifestyles on account of its abundance and diversity of information and products and consumers look at

online shopping as an entertainment activity.

REFERENCES

[1]. AlGhamdi, R., Nguyen, A., & Jones, V. (2013). A Study of Influential Factors in the Adoption and Diffusion of B2C E-Commerce. *International Journal of Advanced Computer Science and Applications,* **4**(1), 89-94.

[2]. Online shopping hits record high in 2013: ASSOCHAM; A real threat to small shopkeepers. - http://www.assocham.org/prels/shownews.php?id=4315

[3]. Chiu, C. M., Wang, E. T., Fang, Y. H., & Huang, H. Y. (2014). Understanding customers' repeat purchase intentions in B2C e-commerce: the roles of utilitarian value, hedonic value and perceived risk. *Information Systems Journal,* **24**(1), 85-114.

[4]. Corbitt, B. J., Thanasankit, T., & Yi, H. (2003). Trust and e-commerce: a study of consumer perceptions. *Electronic Commerce Research and Applications*, **2**(3), 203-215.

[5]. Cyr, D., & Bonanni, C., & Ilsever, J. (2004). Design and e-loyalty across cultures in electronic commerce. In *Proceedings of the 6th International conference on Electronic Commerce*, 351-360.

[6]. Darley, W. K., Blankson, C., & Luethge, D. J. (2010). Toward an integrated framework for online consumer behavior and decision making process: A review. *Psychology & Marketing*, **27**(2), 94-116.

[7]. Retailing value sales to show moderate growth – http://go.euromonitor.com/rs/ euromonitorinternational/images/ Retailing%20Industry%20Overview.pdf

[8]. Ghosh, A., & Ghosh, A. (2014). Traditional Buying to Online Buying: A Study of the Paradigm Shift in Consumer Buying Behavior. *Asian Journal of Management,* **5**(2), 113-116.

[9]. Global B2C Ecommerce Sales to Hit $1.5 Trillion This Year Driven by Growth in Emerging Markets - http://www.emarketer.com/Article/Global-B2C-Ecommerce-Sales-Hit-15-Trillion-This-Year-Driven-by-Growth-Emerging Markets/1010575

[10]. Jiang, L. A., Yang, Z., & Jun, M. (2013). Measuring consumer perceptions of online shopping convenience. *Journal of Service Management*, **24**(2), 191-214.

[11]. e-Commerce Rhetoric, Reality and Opportunity - https://www.kpmg.com/IN/en/ IssuesAndInsights/ArticlesPublications/Documents/KPMG-IAMAI-ES.pdf

[12]. Khare, A., & Rakesh, S. (2011). Antecedents of online shopping behavior in India: An examination. *Journal of Internet Commerce,* **10**(4), 227-244.

[13]. Khare, A., Mishra, A., & Singh, A. B. (2012). Indian customers' attitude towards trust and convenience dimensions of internet banking. *International Journal of Services and Operations Management*, **11**(1), 107-122.

[14]. Lu, Y., Cao, Y., Wang, B., & Yang, S. (2011). A study on factors that affect users' behavioral intention to transfer usage from the offline to the online channel. *Computers in Human Behavior,* **27**(1), 355-364.

[15]. Meinert, D. B., Peterson, D. K., Criswell, J. R., & Crossland, M. D. (2006). Privacy policy statements and consumer willingness to provide personal information. *Journal of Electronic Commerce in Organizations*, **4**(1), 1-17.

[16]. Rewatkar, S. (2014). Factors Influencing Consumer Buying Behavior: A Review (with reference to Online Shopping). *International Journal of Science and Research,* **3**(5), 1347 – 1349.

[17]. Scarpi, D. (2012). Work and fun on the internet: the effects of utilitarianism and hedonism online. *Journal of Interactive Marketing*, **26**(1), 53-67.

[18]. Shah, M. H., Okeke, R., & Ahmed, R. (2013). Issues of Privacy and Trust in E-Commerce: Exploring Customers' Perspective. *Journal of Basic and Applied Scientific Research*, **3**(3), 571-577.

[19]. Shiau, W. L., & Luo, M. M. (2012). Factors affecting online group buying intention and satisfaction: A social exchange theory perspective. *Computers in Human Behavior,* **28**(6), 2431-2444.

[20]. Van Slyke, C, Belanger, F., & Comunale, C. (2004). Factors influencing the adoption of web-based shopping: The impact of trust. *Database for Advances in Information Systems,* **35**(2), 32-49.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 22-26(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Associating Document Object Model with Hierarchy-Cutting and Association Semantics for Analyzing Web Documents

Nivedhita.V and D. Kavitha
Department of Computer Science,
KGiSL Institute of Information Management, Coimbatore (TN), India.

ABSTRACT: The method for realizing the layer division by the hierarchy analysis of ALN through the association semantic is a significant research topic. This work presents an automated approach to extracting domain topic from the commercial web pages. The domain information like the current core topic, core topics and basic topics belonging to the model are determined. An efficient theoretical support for knowledge recommendation along with the different particle sizes on ALNs could be provided through the multilayer theory of Association Semantic with Bat Algorithm. The third step is based on the discovered power-law distribution, up-cutting and down-cutting. The domain informations like the current core topic, core topics and basic topics belonging to the model are determined. An efficient theoretical support for knowledge recommendation along with the different particle sizes on ALNs could be provided through the multilayer theory of Association Semantic with Bat Algorithm.

Keywords: Document Object Model; semantic tree; Bat algorithm; Power-law distribution; Association Semantic Concentricity.

I. INTRODUCTION

A. Web Mining

Web mining is the latest application of data mining techniques to discover patterns from the web. Three different types as illustrated in the fig 1 as,

Fig. 1. Taxonomy of Web Mining.

B. Web Content Mining

Web content mining is the renowned process of mining, extraction and integration of useful data, information and knowledge from web.

Web Content Mining Applications

-Identify the topics represented by a web Document.
-Categorize the web Documents. Find web Pages across different servers that are alike.
-Queries: Enhance standard Query Relevance with User, Role, and/or Task Based Relevance.
-Filters: Show/Hide the documents based upon relevance score.

II. PROPOSED SYSTEM

This work presents an automated approach to extracting domain topic from the commercial web pages. It includes three phases such as, first, it analyzes the domain data information located at the leaf node of DOM tree structure of the web page, and generates the semantic information vector for other nodes of the DOM tree and find maximum repeat semantic vector pattern. In the second step Association Semantic Concentricity (ASC) and frequency analysis are computed by using the Bat Algorithm. Third, based on the discovered down-cutting points, up-cutting points, and power-law distribution are presented to divide the association semantic into three layers.

A data object template by using semantic tree matching technique is being built and uses it to extract all data from the web page. The theories of hierarchy-cutting model are presented along with this model. The hierarchy-cutting points have high accuracy is examined through the experiments [4]. The multilayer theory of association semantic with bat algorithm can provide a theoretical support for knowledge recommendation with different particle sizes on ALNs.

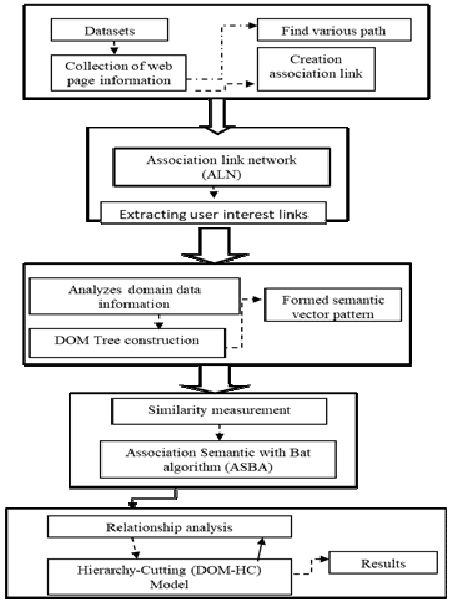


Fig. 2. Architecture Diagram.

III. ASSOCIATION LINK NETWORK (ALN)

Association Link Network (ALN) is composed of associated links between nodes. It can be denoted by ALN= (L, N), where N is a set of web resources (e.g., keywords, web pages, and web topics). L is a set of weighted semantic links belonging to documents. As a data model, ALN has the following characteristics [5].

A. ALN and Its Relevant Mechanisms

ALN can be formalized into a loosely coupled semantic model for managing various resources. As a data model, ALN consists of the following parts, as shown in Fig. 3.

-Resources representation mechanism
-Association rules generation mechanism
-Resources storage mechanism
-ALN generation mechanism
-Application mechanism

B. Two Basic Semantic Features

Before formation of the ALN first two basic semantic features are presented to analyze the semantic association tendency of keywords. Further, all keywords in ALN are divided into four types of keywords with different association roles [1].

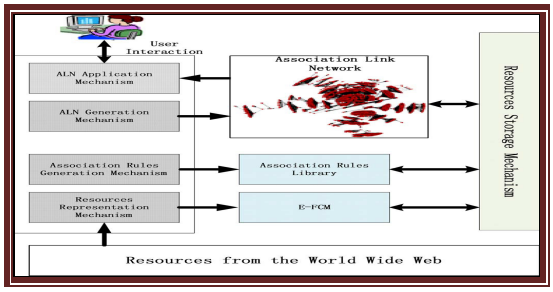


Fig. 3. ALN and Its Relevant Mechanisms.

1) Active Traction Feature of Keyword (ATF). Active Traction Feature of Keyword (ATF) is a kind of semantic feature, which is owned by antecedent keyword in a keyword-level ASL. Fig. 4, illustrates that the keyword "woman" is an object, and are listed with four attributes such as "fertility", "supplement", "pregnancy", "menopause".

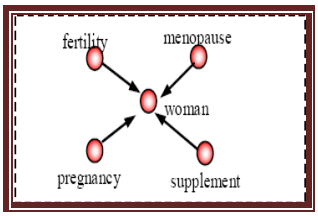


Fig. 4. Two Basic Association Semantic Features.

2) Passive Traction Feature of Keyword (PTF). A different kind of semantic feature, owned through a descendant keyword in a keyword-level ASL is often termed to be the Passive Traction Feature of Keyword (PTF). Fig. 4, exhibits the association link {"pregnancy"->"woman"} in human health domain and the keyword "woman" has passive traction feature [3].

C. Four Kinds of Keywords

All the extracted keywords that should be used as a representation of the domain knowledge from web resources are divided into the following four types on the basis of two basic semantic features, active traction and passive traction [1].

1) Active Traction Keyword (ATK). A keyword k_m, belongs to active traction keyword, only when it satisfies the following two conditions.

2) Passive Traction Keyword (PTK). Analogous, to the definition of ATK, if a keyword k_m belongs to the passive traction keyword, satisfy the following two conditions is considered to be mandatory [6].

1. It must be one of the domain keywords extracted from a domain web resource.

2. The other is that it only has the semantic feature of passive traction.

3) Bridging Traction Keyword (BTK). Bridging Traction Keyword (BTK), it consists of two types of semantic features. In an ASL, it occurs as the antecedent keyword, which includes the semantic feature of active traction.

4) Non-Traction Keyword (NTK). It does not possess the two semantic features of passive and active traction. It is accepted as one of the domain keywords extracted from a domain web resource in order to represent the semantic of web resources that are illustrated in Fig. 5.

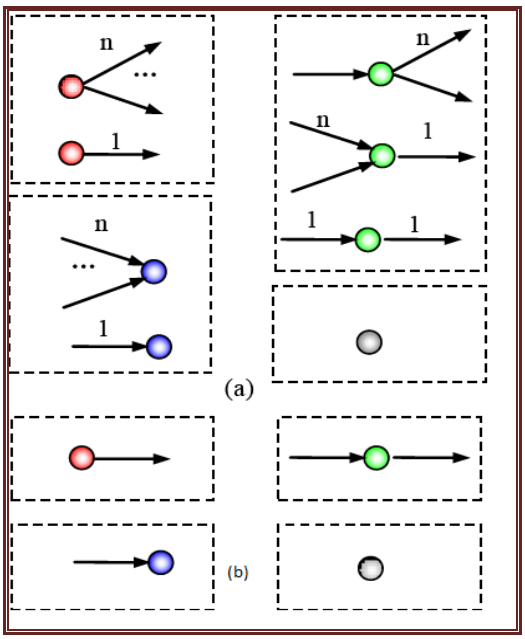


Fig. 5. Four Kinds of Keywords Based on Two Basic Association Semantic Features.

IV. DOM TREE CONSTRUCTION

Document Object Model (DOM) is in fact a tree and its nodes represent elements, attributes and other XML data. Each element node can reach only its parent, siblings, children and attributes directly (in constant time).

A. *Left-Right-Depth (LRD)*

In order to determine the relation of two nodes in constant time the Left-Right-Depth (LRD) is often used in XML algorithms.

B. *Distribution Of Four Kinds Of Keywords*

This section aims to explore the distributions of four kinds of keywords on two conditions: a given/fixed support and different/variable supports of ASLs.

1) Distribution of Keywords at a Given/Fixed Support and Confidence

The web resources are produced in a month according to a series of steps: extracting domain keywords, semantic representation, and mining ASLs [2].

The Distribution of Four Kinds of Keywords at the Changing Support

Next, the different supports help in the analysis process. Initially the calculation procedure of the distribution through the adjustments in the support is given.

ALGORITHM 1 ANALYZING THE DISTRIBUTION OF FOUR KINDS OF KEYWORDS AND THEIR RELATED ASLs

INPUT: The variable support, the semantic representation of web resources on a month.

OUTPUT: The number of different keywords and related ASLs

1. Set the support of ASLs as 2
2. Get the ASLs from the given semantic representation of web resources by the method of ASL
3. Count the number of four kinds of keywords occurring in these web resources, and their related ASLs
4. If the number of obtained ASLs is 0, then go to step 6
5. Else add the support of ASL, and go to step 2
6. End

V. HIERARCHY-CUTTING MODEL

In this section, the basic idea of the layered theory is being discussed [7]. Then, by using Bat Algorithm (BA) the layered theory based on semantic concentricity degree is been presented. Finally, the computing step of hierarchy-cutting model is presented. Association semantic concentricity (ASC) is often termed to be the repeatability score of association semantic contained in two k-ALNs ,G_i, G_j can be achieved by Luo's method [3]. $ASC\ (G_i, G_j)$ can be defined as

$$ASC(G_i, G_j) = \frac{R_{G_j}}{R_{G_i}}, \tag{4.1}$$

$$R_{G_i} = \sqrt{Count_i/\pi} \tag{4.2}$$

Where R_{G_i} and R_{G_j} denote the semantic radius of $G_i \& G_j$, $Count_i$denotes the number of ASLs

in G_i. The equation (4.1) and (4.2) is optimized using Bat Algorithm (BA).

- *Changing trend of ASC*

For m ALNs 1 2 G,G , m G , if the number of their ASLs follows up the power-law distribution, then have the sequence of $ASC(G_1, G_2)ASC(G_2, G_3), \dots, ASC(G_{m-1}, G_m)$ which is an increasing sequence.

$$ASC(G_{i-1}, G_i) = \frac{R_{G_i}}{R_{G_{i-1}}} \quad (4.10)$$

$$= \frac{\sqrt{a * \frac{fre_i^{-b}}{\pi}}}{\sqrt{a * \frac{fre_i^{-b}}{\pi}}} = \left[\frac{(i-1)}{i}\right]^{\frac{b}{2}} \quad (4.11)$$

$$ASC(G_{i-1}, G_i) = \left[\frac{i}{(i+1)}\right]^{\frac{b}{2}} \quad (4.12)$$

Similarly,

Therefore

$$\frac{ASC(G_{i-1}, G_i)}{ASC(G_i, G_{i+1})} = \left[\frac{(i^2-1)}{i^2}\right]^{\frac{b}{2}} < 1 \quad (4.13)$$

That is,

$$ASC(G_{i-1}, G_i) < ASC(G_i, G_{i+1}) \quad (4.14)$$

Association Semantic Concentricity (ASC) is a strictly increasing sequence in the case of power-law distribution.

- *Computing the Number OF ASLs*

On the m supports $fre = \{fre_1, fre_2, \dots, fre_m\}$ the keyword-ALNs. $\{G_1, G_2, \dots, G_m\}$ can be generated according to the extracted ASLs.

- *Completing the Hierarchy-Cutting*

This step aims to find the down-cutting point and up-cutting point. It can be described as finding two obvious changing/increasing values of ASC of any two adjacent layers, i.e.

Suppose ΔASC_i and ASC_j are the maximum and secondary values by equation (4.16), then the up-cutting point of support frequency is I, and down-cutting point of support frequency isj.

VI. RESULTS AND DISCUSSION

In this section, to verify the multilayer theory of association semantic the experimental data and result analysis is been presented. Three domain news data are being taken from the website http://www.reuters.com that includes the internet, health and environment, in order to build keyword-level ALNs on different supports.

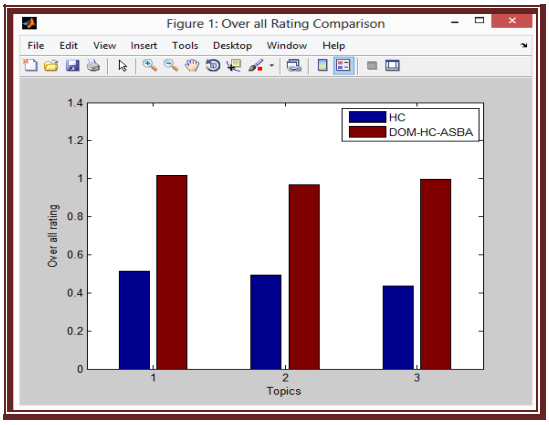


Fig. 6. Three Different Topics vs. Overall Rating.

Fig. 6 shows the overall rating comparison results of the three different topics. These three concepts are measured between the existing (HC) system and proposed DOM-HC-ASBA system, it concludes that the proposed system produces higher overall rating results for all topics and the values.

Fig. 7 shows the confidence comparison results of the three different topics 1, 2, 3. It concludes that the proposed system produces higher confidence results for all topics.

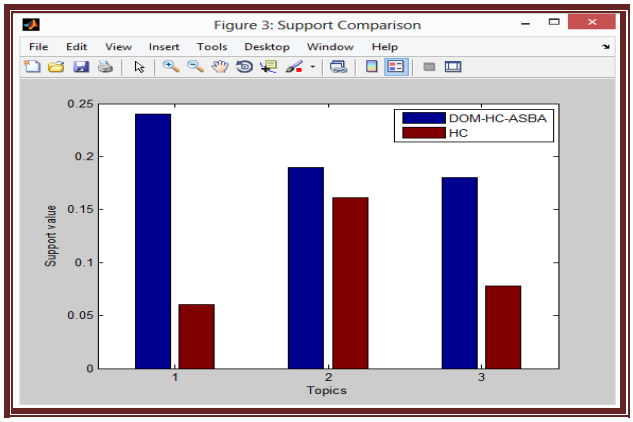


Fig. 7. Three Different Topics vs. Confidence.

Fig. 9 shows the time comparison results between the existing (HC) system and proposed DOM-HC-ASBA system, it concludes that the proposed system takes lesser time results for all topics.

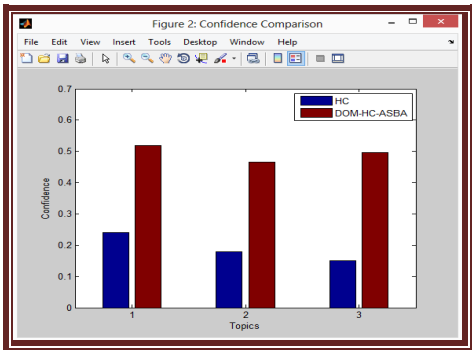


Fig. 8. Three Different Topics vs. Support.

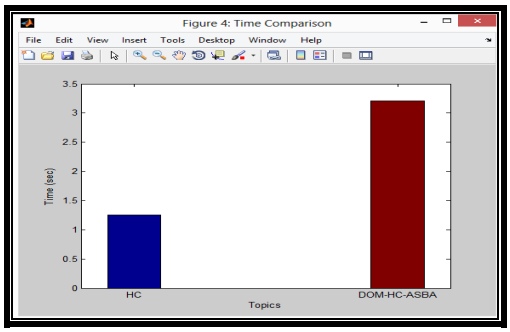


Fig. 9. Three Different Topics vs. Time Taken.

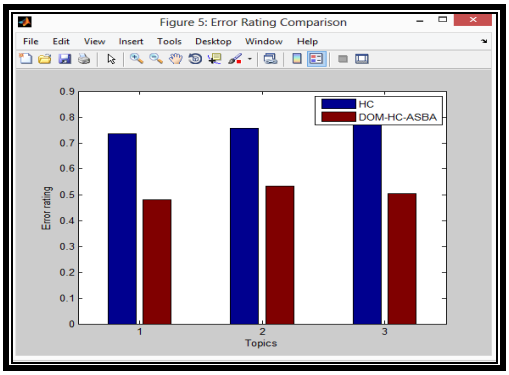


Fig. 10. Three Different Topics vs. Error Rating.

Fig. 10 shows the error rating comparison results of the three different topics 1, 2, 3 relates to the Internet domains. These three concepts are measured between the existing (HC) system and proposed DOM-HC-ASBA system, it concludes that the proposed system produces lesser error rating results for all topics.

VII. CHAPTER SUMMARY

The DOM-HC of association semantic has been built based on the Bat Algorithm so it is named as ASBA.
The proposed DOM-HC model has completed the division of three-layer association semantic, that is, "basic semantic," "main semantic," and "core semantic." "Core semantic" can provide the support for discovering current core topic. The association semantic is divided into three layers only on the basis of the discovered power-law distribution, down-cutting and up-cutting points. At the same time, the theories of DOM-HC are presented. The experiments show that Document Object Model with Hierarchy-Cutting (DOM-HC) model points has high accuracy. In future, we need further efforts to explore the organization and recommendation of the analyzed topic based on the hierarchy-cutting model. In addition, need to evaluate this research work proposal further by applying them to other multiple domains such as fashion and movies.

REFERENCES

[1]. Frias-Martinez, E.; Chen, S.Y.; Liu, X. (2007). "Automatic cognitive style identification of digital library users for personalization.". *JASIST*. **58**(2): 237–251.
[2]. Xu, Z., Zhang, S., Choo, K.K., Mei, L., Wei, X., Luo, X., Hu, C. and Liu, Y., (2017). "Hierarchy-cutting model based association semantic for analyzing domain topic on the web. IEEE Transactions on Industrial Informatics", pp.1-10.
[3]. Teevan, J., Dumais, S.T. and Horvitz, E., (2005). "Personalizing search via automated analysis of interests and activities. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval" , pp. 449-456.
[4]. S.X. Zhang, K. Lu, W. Liu, X. Yin, and G. Zhu, (2015). Generating associated knowledge flow in large-scale web pages based on user interaction. *Computer Systems Science and Engineering,* **30**(5): 377-389,
[5]. Yang, X.S., (20100. "A new metaheuristic bat-inspired algorithm. Nature inspired cooperative strategies for optimization (NICSO 2010)", pp.65-74.
[6]. Yang, X.S., (2011). Bat algorithm for multi-objective optimisation. *International Journal of Bio-Inspired Computation,* **3**(5), pp.267-274.
[7]. Yang, X.S. and Hossein Gandomi, A., (2012). "Bat algorithm: a novel approach for global engineering optimization". *Engineering Computations*, **29**(5), pp. 464-483.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 27-31(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Modified Density Based Clustering for Ranked User Preferred Patterns in Web Usage Mining

D. Kavitha[1] and Dr. B. Kalpana[2]
[1]Department of Computer Science,
KGiSL Institute of Information Management, Coimbatore (TN), India.
[2]Department of Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: The heterogeneous nature of the web combined with the rapid diffusion of web based applications has made web browsing an intricate activity for users. Web Usage Mining (WUM) refers to the application of Data Mining techniques for the automatic discovery of meaningful usage patterns characterizing the browsing behavior of users, starting from access data collected from interactions of users with websites. The preprocessing, pattern discovery, and pattern analysis are the three main phases of web usage mining. In order to implement functionalities the discovered patterns may be conveniently exploited to offer useful assistance to users. Analysis of ranked user preferred query pattern and clustering of those patterns becomes a challenging task. Modified Density Based Clustering (MDBC) is proposed for carrying out clustering of ranked user preferred patterns (datapoints). To overcome the problem of border objects encountered in DBSCAN, here the MDBC solves this problem that usually considers the datapoints from patterns dynamically during the clustering process. A comparison of MDBC with the existing K-means clustering and FCM methods reveals that there is significant improvement in performance in terms of ARI, MI, homogeneity and completeness.

Keywords: Datapoints; Modified Density Based Clustering;, ranked user preferred query pattern; Web Usage Mining;

I. INTRODUCTION

Pattern explosion is a common problem that addresses a large collection of generated patterns as a result of pattern discovery.

It also expects a domain expert to render substantial effort to identify relevant patterns which are of user's interest. Pattern analysis is a final phase of web usage mining. The goal is to eliminate the irrelevant patterns and to extract the interesting patterns, generated from the pattern discovery process. The output of web mining algorithm is not suitable for human understanding. It needs to be transformed and interpreted to a human understandable format. Methodologies and tools available for analysis help to perform the above task. Pattern analysis consists of two common approaches. They are the knowledge query mechanism such as SQL, and multi dimensional data cube construction before performing OLAP operations [1].

In pattern analysis all the methods execute, based on the assumption that the pattern discovery phase yields a structured output. Similar to other KDD processes, the result analysis of web usage mining allows the analyst to extract interesting results from uninteresting ones.

The issue with WUM results is that it is extremely hard to capture and define the notion of interestingness. This varies according to the analyst's beliefs, their needs, the web site type, its structure, the user sessions analyzed, etc. In this step of the WUM process, the analyst is interested in protecting the patterns discovered on the web site structure or on its content.

To discover user preferences through log mining, web server logs serves as a significant resource. Log mining is extensively used in web personalization, recommender systems, and web site design and evaluation [2]-[4]. Information like IP addresses, time stamps, and requested pages can be extracted from web logs, which are in turn applied to the web application in order to deduce hidden user feedbacks such as motivations, goals and preferences. Considerable research to study hidden feedback to improvise ranking is done in log mining and machine learning technologies [5]. Some of the ranking methods taken for discussion are as follows.

A personalized Frequency Based Ranking Method (FBRM) is used for web log mining results of a large scale research team. This work proposes a personalized FBRM to improve the accuracy in predicting the user preferences. The ranked user preferred patterns obtained from FBRM method are considered as datapoints for clustering [6].

The proposed Modified Density Based Clustering (MDBC) is used to carry out clustering of user preferred pattern data to overcome the problem of border points.

II. CLUSTERING ALGORITHMS

Clustering is perhaps the most important and the widely used method for unsupervised learning. It is the problem of identifying groupings of similar points that are relatively isolated from each other or in other words to partition the data into dissimilar groups of similar items. Cluster analysis [7] aims to discover the internal organization of a dataset by finding structure within the data in the form of clusters. Intuitively, the division into clusters should be characterized by within cluster similarity and between cluster dissimilarity (external). Hence, the different groups containing distinctive elements are formed by breaking the data into a number of groups composed of similar objects. In both multivariate statistical analysis and machine learning this technique is widely used.

The Clustering data is useful in the following ways [8]. It abets understanding of the data by breaking it into subsets that are significantly more uniform than the overall dataset.

In general, the clustering makes the data simpler to describe, since a new data item can be specified by indicating its cluster and then its relation to the cluster centre. Each application might suggest its own criterion for assessing the quality of the clustering obtained. Typically expecting the quality to involve some measure of fitness between a data item and the cluster to which it is assigned. This can be viewed as the pattern function of the cluster analysis. Hence, a stable clustering algorithm will give assurances about the expected value of this fitness. As with other pattern analysis algorithms this will imply that the pattern of clusters identified in the training set is not a chance of occurrence, but characterizes some underlying property of the distribution generating the data. Perhaps the most common choice for the measure assumes that each cluster has a centre and assesses the fitness of a point by its squared distance from the centre of the cluster to which it is assigned. Clearly, this will be minimized if new points are assigned to the cluster whose centre is nearest.

The issues in creating the cluster algorithms are as follows:

- To achieve an efficient algorithm the challenge lies in reducing the number of attributes.
- The type of the attributes can be diverse, and not only numeric.
- Defining a similarity function between the objects is a trivial task. Many features and types of attributes have to be handled efficiently.
- The processing time or the memory requirement of the algorithm can be huge, that has to be reduced using some heuristics.
- Validating the resulting clusters is also a difficult task. In case of having voluminous objects with high dimensionality, statistical methods have to be used and indices have to be defined which can be computationally expensive.

Many algorithms in literature, deal with the problem of clustering the large number of objects. The Density-Based Spatial Clustering of Application with Noise (DBSCAN) is a popular density based algorithm in which clusters are defined as areas of higher density of data set. Objects that lie in the low density areas are considered to be noisy [9].

A cluster of an arbitrary form consists of all density connected objects. The complexity of DBSCAN is very low as it requires a linear number of range queries on the database and discovers same results eventually. Executing multiple number of times is avoided as it is deterministic for core and noise points, but not for the border points.

III. PROPOSED METHODOLOGY

The impetuous growth of information on the internet has captivated a plethora of users. Though the search engines present a well organized way to search the relevant information from the web, the results acquired might not always be helpful to the users, as it fails to identify the user intention behind the query. A query means different things in varying context and the user alone can interpret the anticipated context. The web users are not satisfied with the search results in spite of recent development on web search technologies. Therefore, the requirement arises to explore personalized web search system which could produce output as highly ranked pages suitable to the users. A personalized web search has various levels of competence for different users, queries and search contexts.

To solve the problem of search result personalization, various web mining techniques are discussed [10]. The proposed methodology for pattern analysis is illustrated in Fig. 1.

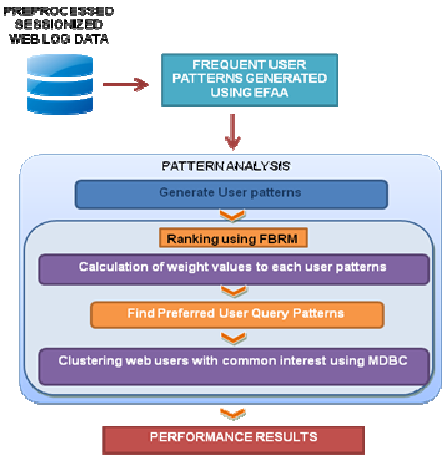


Fig. 1. Flow Diagram for Proposed Methodology.

A. Modified Density Based Clustering (MDBC)

The proposed Modified Density Based Clustering (MDBC) is used to carry out clustering of user preferred pattern data. The ranked user preferred patterns are considered as datapoints for clustering.

Given a set of datapoints from patterns denoted as $(DP = (dp_1, \ldots, dp_n))$. It groups together similar user pattern datapoints that are closely packed together i.e., points with many nearby neighbors.

DBSCAN requires two parameters: ε (eps) and the minimum number of datapoints from patterns required to form a dense region (minPts). It starts with an arbitrary starting datapoints from patterns that have not been visited, for which ε-neighborhood is retrieved, and if it contains sufficiently many datapoints from patterns, a cluster is started. Otherwise, the point is labeled as noise.

If datapoints from patterns is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all datapoints from patterns that are found within the ε-neighborhood are added, as in their own ε-neighborhood when they are also dense. This process continues until the density connected cluster is completely found. To discover further cluster or noise, a new unvisited datapoints from patterns is retrieved and processed.

The main demerit of common DBSCAN clustering techniques is the problem of border points. Here, it treats all border points as noise points. Since the border points at times don't take significant datapoints into consideration, here the Modified Density Based Clustering (MDBC) is proposed, that usually considers all the datapoints from patterns during the clustering process.

Initially in MDBC, a group is created for every datapoints from patterns by taking a set of points '*a*'. A point *a* is considered to be a core point when it comprises of at least MinPts points, which in turn, contains three or four patterns in a distance eps of it, and also, these datapoints from patterns can be directly reached from *a*. Now the remaining of the datapoints from patterns is mapped onto the group based on the distance eps. This procedure continues until every datapoints in pattern (DP) gets mapped into the group. MDBC requires two parameters eps and MinPts are required to form a dense region. It starts with random patterns (unclustered), which have not yet been visited. This attribute eps-neighbourhood is subsequently regained to every pattern, and if it has sufficiently multiple numbers of points, a cluster is generated.

Modified Density Based Clustering (MDBC) Algorithm

```
MDBC (DP, eps, MinPts)
{
  C = 0
    C_id:=Nxt_id(outlier)
     for (i=1;i<n;i++) dataset DP
      {
       Point:=a.get(i)
       if a. C_id = visited
            mark a as visited
            NeighborPts = regionQuery(a, eps)
            if sizeof(NeighborPts)< MinPts
            mark a as NOISE
            else
            {
            C = next cluster
            expandCluster(a, NeighborPts, C, eps, MinPts)
            then
            NeighborPts:= next point (NeighborPts)
            End if
         End If
            For DP_border in the border list
              result.size(DP_border) =
              Retrieve Neighbours(DP_border, Eps)
              Assign DP_border  to C_Id
              }
            End for
          }
    }
expandCluster(a, NeighborPts, C, eps, MinPts)
{
    add a to cluster C
    for each point a' in NeighborPts
      {
      if a' is not visited
      {
         mark a' as visited
         NeighborPts' = regionQuery(a', eps)
         if sizeof(NeighborPts') >= MinPts
             NeighborPts=NeighborPts joined with NeighborPts'
       }
      if a' is not yet member of any cluster
      add a' to cluster C
      }
}
regionQuery(a, eps)
return all points within a's eps-neighborhood (including a)
end
```

Otherwise, the patterns are labelled to be outliers to datapoints. This process continues until every datapoints gets mapped onto the group. Finally, every core datapoints is now allocated to its best density reachable chain. If, in case any datapoints among the clusters has just one single attribute, the clusters are combined using a linkage or amalgamation rule that decides the time when the two clusters are identical enough to be linked together.

IV. RESULTS AND DISCUSSIONS

The evaluation of proposed MDBC is compared to existing clustering methods such as K-means clustering and Fuzzy Clustering Methods.

A. Clustering Results of MDBC

In this section, the performance of Modified Density Based Clustering (MDBC) is evaluated using the different performance metrics. Each of the metrics (ARI, MI, Homogeneity and Completeness) results and their tabulated values on benchmark datasets are discussed in detail.

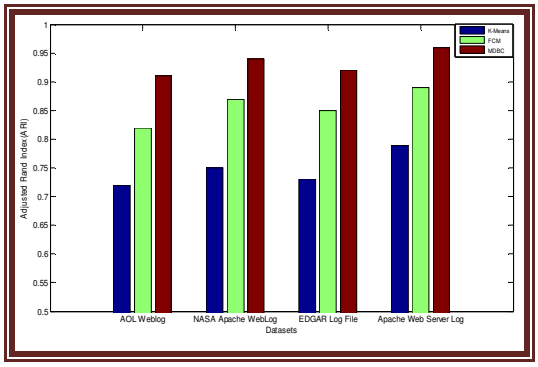


Fig. 2. ARI Versus Clustering Methods.

Fig. 2 illustrates that the proposed MDBC algorithm achieves 0.9100 ARI which is 0.19 and 0.09 higher when compared to K-means clustering and FCM methods respectively for AOL weblog dataset samples. In case of MDBC algorithm which achieves 0.9400 ARI, there is an approximate increase of 0.19 and 0.07 ARI when compared to the existing methods for NASA Apache weblog dataset samples. Proposed MDBC algorithm achieves 0.9200 ARI which is 0.19 and 0.07 higher when compared to K-means clustering and FCM methods respectively for EDGAR log dataset samples. The MDBC algorithm achieves 0.9600 ARI which is 0.17 and 0.07 higher when compared to the existing methods for Apache web server log dataset samples.

Fig. 3 shows that the proposed MDBC algorithm achieves 0.92 MI which is 0.11 and 0.09 higher when compared to K-means clustering and FCM methods respectively for AOL weblog dataset samples. In case of MDBC algorithm which achieves 0.95 MI, there is an approximate increase of 0.10 and 0.06 MI when compared to the existing methods for NASA Apache weblog dataset samples.

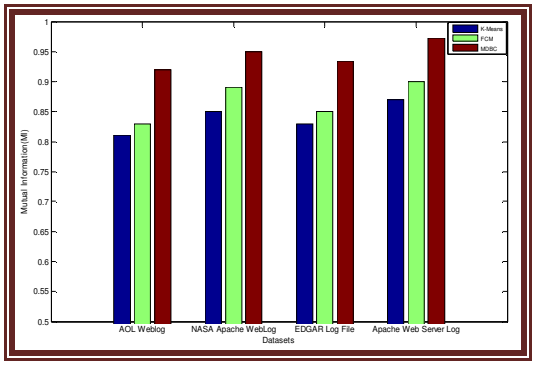


Fig. 3. MI Versus Clustering Methods.

Proposed MDBC algorithm achieves 0.935 MI which is 0.105 and 0.085 higher when compared to K-means clustering and FCM methods respectively for EDGAR log dataset samples. The MDBC algorithm achieves 0.972 MI which is 0.102 and 0.072 higher when compared to the existing methods for Apache web server log dataset samples.

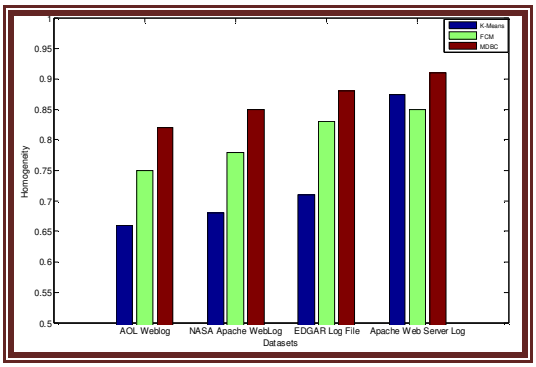


Fig. 4. Homogeneity Versus Clustering Methods.

Fig 4 shows that the proposed MDBC algorithm achieves a homogeneity score of 0.82, which is 0.16 and 0.07 higher when compared to K-means clustering and FCM methods respectively for AOL weblog dataset. In case of MDBC algorithm the homogeneity score is 0.85, which is 0.17 and 0.07 higher when compared to existing methods on NASA Apache weblog dataset samples. The homogeneity score shows an improvement of 0.88 0.17 and 0.05 when compared to K-means clustering and FCM methods respectively on EDGAR log dataset. For Apache web server log dataset, the MDBC achieves 0.91 homogeneity score which is 0.035 and 0.06 higher when compared to the existing methods.

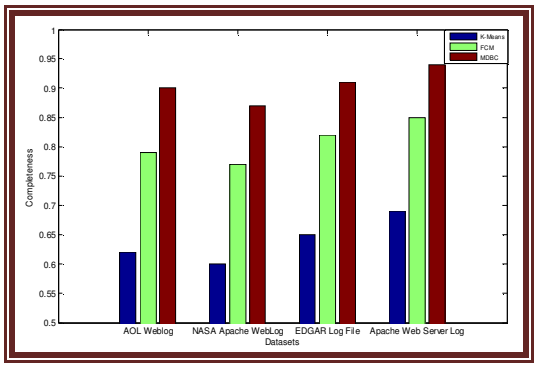


Fig. 5. Completeness Versus Clustering Methods.

Fig. 5 shows that the proposed MDBC achieves a completeness of 0.90 which shows an improvement of 0.28 and 0.11 when compared to K-means clustering and FCM methods respectively on AOL weblog dataset samples. In case of NASA Apache weblog dataset, MDBC achieves 0.87 completeness and there is an approximate increase of 0.27 and 0.10 when compared to the existing methods. For EDGAR log dataset, MDBC achieves a completeness of 0.91 which shows an improvement of 0.26 and 0.09 when compared to K-means clustering and FCM methods respectively. The MDBC achieves 0.94 completeness which is 0.25 and 0.09 higher when compared to the existing methods on Apache web server log dataset.

A comparison of MDBC with existing methods reveals that there is a significant improvement in performance in terms of ARI, MI, homogeneity and completeness.

V. CHAPTER SUMMARY

Modified Density Based Clustering (MDBC) is used for carrying out clustering user preferred pattern data. The ranked user preferred patterns obtained from FBRM method are considered as datapoints for clustering.

The clustering results of the proposed MDBC prove to be efficient when compared to existing ranking methods such as K-means and FCM. Directions for future work include investigating the effect of coarse grained or noisy feedback on learning performance, learning preferences over sets of patterns instead of individual patterns, and shifting from the pool based active learning to query synthesis, i.e. directly mining patterns for queries.

REFERENCES

[1]. Wu, Q., Ding, G., Xu, Y., Feng, S., Du, Z., Wang, J. and Long, K., (2014). "Cognitive internet of things: a new paradigm beyond connection". *IEEE Internet of Things Journal*, **1**(2), pp.129-143.

[2]. Das S, Datar M, Garg A, and Rajaram S., (2007). "Google news personalization: scalable online collaborative filtering," *In Proceedings of the 16th international WWW conference*, pp. 271-280, ACM Press.

[3]. Spiliopoulou M., (2000). "Web usage mining for web site evaluation", *Communication* , Vol. **43**, No.8, pp.127-134.

[4]. Srivastava J, Cooley R, Deshpande M, and Tan PN, (2000). "Web usage mining: Discovery and applications of usage patterns from web data", *ACM SIGKDD Explorations Newsletter*, vol. **1**, *no.* 2, pp.12-23.

[5]. Richardson M, Prakash A, and Brill E, (2006). "Beyond PageRank: machine learning for static ranking", *In Proceedings of the 15th international conference on World Wide Web*, pp. 707-715.

[6]. Kavitha, D. and Kalpana, B., (2017). "A Personalized Web Search System using Frequency Based Ranking Method (FBRM) for Web Log Analysis", *International Journal of Control Theory and Applications*, Vol. **10**.

[7]. Aggarwal, C.C. and Reddy, C.K. eds., (2013). "Data clustering: algorithms and applications". *CRC press.*

[8]. Abbas, O.A., (2008). "Comparisons Between Data Clustering Algorithms". *International Arab Journal of Information Technology (IAJIT),* **5**(3): 320-325.

[9]. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei, (1996). "A density-based algorithm for discovering clusters in large spatial databases with noise".

[10]. Ratnakumar, A.J., (2010). "An implementation of web personalization using web mining techniques", *Journal of Theoretical and applied information technology*, **18**, 1, 67-73.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 32-35(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Evaluating the Effectiveness of Modified Particle Swarm Optimization in Classification

Balasaraswathi M[1] and Kalpana B[2]

[1]Associate Professor, Department of Information Technology,
Sri Ramakrishna College of Arts and Science, Coimbatore (TN), India.
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: Particle Swarm Optimization (PSO) is a metaheuristic technique which is a swarm based algorithm. The power of PSO is due to its computational efficiency and does not require substantial intermediate computations. In this paper, we have analysed the results of experiments by implementing the modified PSO in the classification domain using three different fitness functions.

Keywords: Particle Swarm Optimization; Euclidean, Manhattan, Mahalanobis Distance function.

I. INTRODUCTION

The rise of Meta-Heuristics can mainly be attributed to the increase in data generation. This is due to the increase in the information leveraging devices such as sensors, high resolution cameras and video recorders, satellites and user information generated from the internet[1]. Hence a huge amount of data came to be available for the analyst [2,3]. But the methods for accessing them were ancient, hence a huge increase in the processing time was observed. While the scenario for analysis depicted this observation, the requirement scenario portrayed an inversely proportional view. A fast and efficient processing system that provides accurate results online (in real time) was the requirement of the user. Hence the legacy data mining algorithms lost their scope and Meta-Heuristic algorithms came into view[4].Meta heuristic algorithms perform effective optimization. And it always promises to provide near optimal results. This method was built into them in order to provide the much needed time efficiency. This proves to be the catalyst for constructing conventional data mining techniques using meta heuristic algorithms. Further, the legacy techniques provided poor performance when dealing with noisy or incomplete data. Hence pre-processing became a mandatory first phase while using them. Even though the meta heuristics will work effectively while using preprocessed data, they also provide considerably good results even without data preprocessing. They also work incredibly well when dealing with intractable data mining problems [5].

Efficiency of an algorithm is usually measured by the accuracy of the results and the time taken to provide the results. Independence from the data size, when computing the results is achieved by the heuristic approaches to a rather considerable extent [6,7].

Particle Swarm Optimization (PSO) is a metaheuristic technique which is a swarm based algorithm. This contribution discusses the possibility of using modified PSO for the process of Classification and Feature Selection using three different fitness function.

II. PSO FOR CLASSIFICATION: AN ANALYSIS

A major advantage of using PSO over Classification problems is that it is highly flexible in terms of its applicability [8]. The Objective Function that is to be used for analysis is user defined, unlike other methods where the objective function is already defined. Shared memory is very less, and to be precise, PSO has only one shared location, the *gbest* (global best value) except for that, all other values are associated with particles and not shared. This also increases the computational efficiency by eliminating the synchronizing mechanism.

Hybridization: PSO as a Part of Classification Process

PSO not only lends itself for pure computational analysis, it also provides excellent results, when hybridized with other techniques [10]. Various approaches use the PSO algorithm for certain phases in their optimization process. Some of the most commonly used approaches that incorporate PSO to offer hybrid optimization techniques include; Artificial Neural Network [9] and Support Vector Machines [11]. These approaches tend to blend well with PSO to provide effective results.

III. MODIFIED PSO WITH ATTRIBUTE ELIMINATION TECHNIQUE

This contribution presented a modified PSO algorithm (MPSO) [12] that has embedded attribute elimination techniques.Massive information created in the current scenario has led to a major bottleneck in terms of processing. The vast data that is available is not completely usable, in the sense, it does not entirely contain data that guides to the final results. The data tends to contain missing or redundant information, or information that is irrelevant to the study. Removing these data will not only reduce the processing time, it also enhances the accuracy of the processing algorithm. The MPSO algorithm eliminates unnecessary data by effectively applying feature selection techniques and hence improving the classification accuracy (Fig. 1). Analysis proves that MPSO consumes less time and provide better accuracy when compared to PSO.

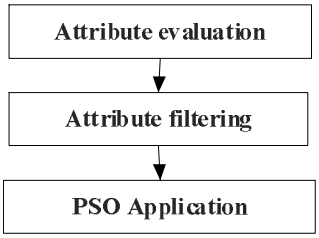


Fig. 1. Modified PSO based Classification.

Incorporating attribute elimination techniques in the regular PSO enhances the conventional classification performed by PSO. The method of attribute elimination is embedded in PSO itself; hence it works side by side with the algorithm, which proves to be advantageous. The technique of embedded PSO is performed in two major phases. The process begins with the evaluation of attributes using the CFS subset evaluator [15]. The Greedy hill climbing method is used to filter attributes and the final pruned dataset is created. PSO is applied on the pruned dataset to produce efficient results.

The CFS based feature selection is used for data preprocessing. This method evaluates the accuracy of the subset of attributes by considering the individual predictive ability of each of the feature and the degree of redundancy existing between them. Subsets containing attributes that have high correlation with the class attribute and low correlation between themselves are preferred. A feature is said to be relevant iff there exists some v_i and c for which $p(V_i = v_i) > 0$ such that

$$p(C = c | V_i = v_i) \neq p(C = c) \quad (1)$$

CFS only measures the correlation between nominal features, so numeric features are first discretized. CFS is a completely automatic algorithm, which does not require any supervision in terms of threshold limits. It operates on the original feature space, hence it can be interpreted in terms of the original features. Hence the CFS filtering technique does not incur high computational cost, due to the repeated invoking of the learning algorithm. The evaluator method used here is the Best first, which Searches the space of attribute subsets by greedy hill climbing augmented with a backtracking facility.

PSO begins by first initializing the number of particles. These particles are distributed in a uniform manner in the search space. Initial velocities are set to random values, and the particle acceleration in triggered. Velocities are set to all the dimensions of the data, hence movement is triggered relative to all the dimensions. After every displacement, the velocity of the particles are altered based on the global and the local best values determined by the fitness function. The velocity of the particles is calculated using the equation

$$V_{i,d} \leftarrow \omega V_{i,d} + \varphi_p r_p (P_{i,d} - X_{i,d}) + \varphi_g r_g (g_d - X_{i,d}) \quad (2)$$

Where r_p and r_g are the random numbers, $P_{i,d}$ and g_d are the parameter best and the global best values, $x_{i,d}$ is the value current particle position, and the parameters ω, φ_p, and φ_g are selected by the practitioner.

If the current known position of the particle is better than the particle best (*pbest*), then the *pbest* value for the particle is updated. Similarly, if the current *pbest* value is found to be greater than the global best (*gbest*), then the *gbest* value is updated to the current *pbest*. This process is continued until the application reaches termination. The termination condition is determined by either the maximum time set or on reaching the maximum required accuracy limit. In this paper we examine the ability of Modified Particle Swarm Optimization(MPSO), metaheuristic technique to efficiently face classification using three distance function for accurate and faster results along with the comparative study of results of basic PSO with the same three distance function.

A. Choice of Fitness Function

To evaluate the results of PSO and MPSO algorithm three fitness functions are taken into account based on Mahalanobis, Manhattan and Euclidean distance function [13,14].

Euclidean distance is the ordinary straight-line distance between two points x and yin Euclidean space . With this distance, Euclidean space becomes a metric space

$$\psi_1(x_i, y_i) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (3)$$

Mahalanobis distance is defined as a dissimilarity measure between two random vectors x and y of the same distribution with the covariance matrix S.

$$\psi_1(x_i, y_i) = \sqrt{(x_i - y_i)^T S^{-1}(x_i - y_i)} \quad (4)$$

The Manhattan distance function computes the distance that would be traveled to get from one data point to the other if a grid-like path is followed. The Manhattan distance between two items is the sum of the differences of their corresponding components.

$$\psi_2(x_i, y_i) = \sum_{i=1}^{N} |x_i - y_i| \quad (4)$$

IV. PERFORMANCE EVALUATION

The novel Modified Particle Swarm Optimization (MPSO) is experimented and compared with the existing algorithms for accuracy and computational time. The experimental work is conducted on seven benchmark datasets from KEEL repository.

A. Datasets

KEEL repository aims at providing to the machine learning researchers a set of benchmark datasets to analyze the behavior of the learning methods. Concretely, it is possible to find benchmarks already formatted in KEEL format for classification (such as standard, multi instance or imbalanced data), semi-supervised classification, regression, time series and unsupervised learning. Also, a set of low quality data benchmarks is maintained in the repository. Details about the dataset taken for study are provided in Table 1.

Table 1: Dataset Details.

Name	Attributes	Instances	Classes
Iris	4	150	3
Bupa	6	345	2
Heart	13	280	2
Sonar	60	208	2
Ionosphere	33	351	2
Libras-movement	90	360	15
Shuttle	9	58000	7

B. Comparison Results of PSO and MPSO Classifiers

Fig. 2. illustrates the accuracy rate of PSO with respect to the different datasets. The performance of the PSO algorithm is evaluated on six datasets with three fitness function such as Mahalanobis, Manhattan and Euclidean distance function.

The results reveal that the PSO-F3 (Euclidean distance) produces higher accuracy when compared to other fitness function.

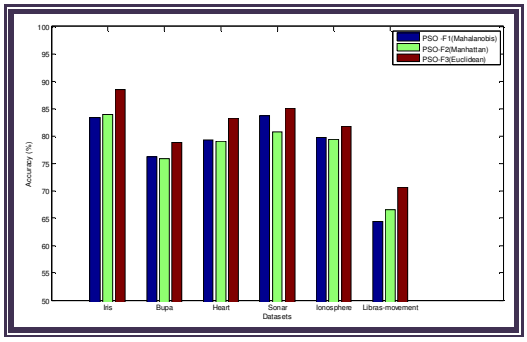


Fig. 2. Accuracy – PSO with Different Fitness Functions.

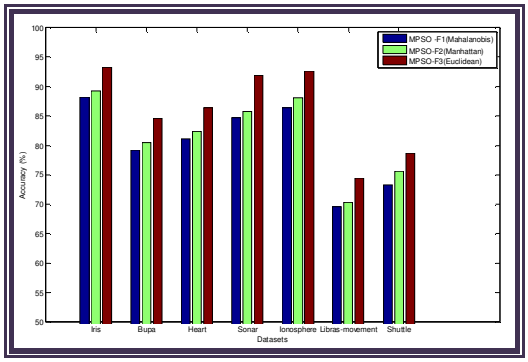


Fig. 3. Accuracy - MPSO on Benchmark Datasets.

Fig. 3 represents the accuracy of MPSO with respect to three fitness functions such as Mahalanobis, Manhattan and Euclidean distance function and it is recorded for seven datasets. From the results it concludes that the proposed MPSO-F3(Euclidean distance) produces higher accuracy results when compared to other fitness function. It is clearly evident that in the case of both PSO and MPSO, significant improvements were observed when Euclidean distance was used as the fitness function. Therefore the same has been adopted for further experiments.

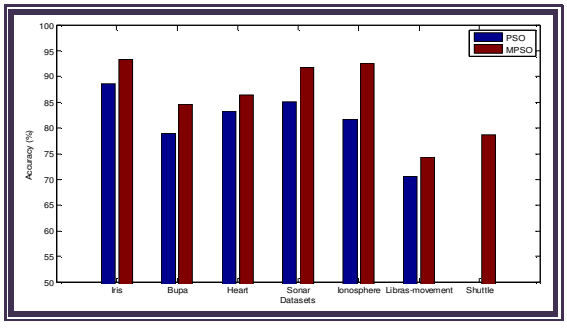


Fig. 4. Accuracy of MPSO Vs PSO.

Fig. 4 shows the comparison of accuracy while using MPSO and PSO algorithm with respect to Euclidean distance function and it is recorded on seven datasets. On an average, the overall accuracy of MPSO is 87.20% which is 5.84% higher when compared to PSO.

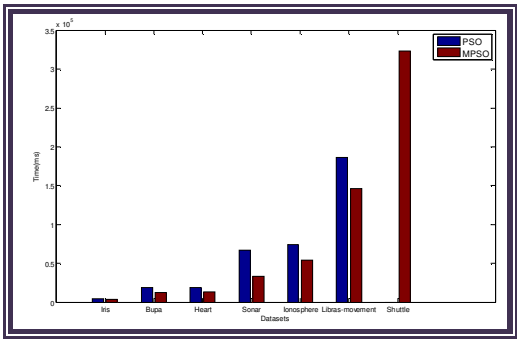


Fig. 5. Execution Time - PSO Vs MPSO.

Figure 5 shows that the proposed MPSO benefits from lesser execution time when compared to PSO. The proposed MPSO converges quickly for Shuttle dataset but it is not the case for PSO. Considering the Execution time of PSO Vs MPSO on all the six datasets, the average reduction in time is around 28.81%.

V. CONCLUSION

PSO is a metaheuristic technique based on the concept of swarm, The novel Modified Particle Swarm Optimization (MPSO) is experimented and compared with the basic PSO algorithm for accuracy and computational time. The experiment work is conducted on seven benchmark datasets from KEEL repository, it can be concluded that modified PSO is suitable for applications in classification. During the implementation of modified PSO with three distance function. It is observed that the selection of distance function plays a very important role in classification. As a conclusion, the modified PSO, which is implemented using Euclidean distance function gives best results than other two Manhattan, Mahalanobis distance function.

REFERENCES

[1]. M. Balasaraswathi, and Dr. B. Kalpana," Metaheuristics for Mining Massive Datasets: A Comprehensive Study of PSO for Classification," Advances in Natural and Applied Sciences, 9(5) May 2015, Pages: 27-38, 2015.

[2]. Han, J., Kamber, M., & Pei, J., "Data Mining: Concepts and Techniques," Elsevier, 2011.

[3]. Hand, D. J., Mannila, H., & Smyth, P., "Principles of Data Mining," MIT press, 2001.

[4]. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & Coello, C. A. C., "Survey of multiobjective evolutionary algorithms for data mining: Part II," Evolutionary Computation, *IEEE Transactions on*, **18**(1), 20-35, 2014.

[5]. Mukhopadhyay, A., Maulik, U., Bandyopadhyay, S., & CoelloCoello, C., "A Survey Of Multiobjective Evolutionary Algorithms For Data Mining: Part I," *IEEE Transactions on Evolutionary Computation*, **18**(1), 4-19, 2014.

[6]. Bottou, L., "Large-Scale Machine Learning With Stochastic Gradient Descent," In Proceedings of COMPSTAT'2010, pp. 177-186, Physica-Verlag HD,2010.

[7]. Bousquet, O., & Bottou, L., "The Tradeoffs Of Large Scale Learning," In Advances in Neural Information Processing Systems , pp. 161-168 , 2008.

[8]. De Falco, I., Della Cioppa, A., & Tarantino, E. "Facing Classification Problems With Particle Swarm Optimization," *Applied Soft Computing, 7*(3), 652-658, 2007.

[9]. McCulloch, W. S., & Pitts, W.,"A Logical Calculus of The Ideas Immanent In Nervous Activity. The Bulletin Of Mathematical Biophysics," 5(4), 115-133, 1943.

[10]. Shi, Y., &Eberhart, R. , "A Modified Particle Swarm Optimizer," In Evolutionary Computation Proceedings, May 1998. *IEEE World Congress on Computational Intelligence*, pp. 69-73.

[11]. Cortes, C., & Vapnik, V. "Support-vector networks. Machine Learning," **20**(3), 273-297, 1995.

[12]. Balasaraswathi, M., Kalpana, B., "Enhanced Classification using PSO with Embedded Attribute Elimination Techniques," ARPN :: *Journal of Engineering and Applied Sciences,* ISSN 1819-6608, Volume **10**, Number 20, November 2015, PP 9650-9658

[13]. Deepak Sinwar, Rahul Kaushik. "Study of Euclidean and Manhattan Distance Metrics using Simple K-Means Clustering," *International Journal for Research in Applied Science and Engineering Technology(IJRASET),* ISSN 2321-9653, Volume **2**, Issue V, May 2014, pp 270-274

[14]. Archana Singh, AvantikaYadav, Ajay Rana. "K-Means With Three Different Distance Metrics," *International Journal of Computer Applications*, ISSN 0975-8887, Volume 67-No.10, April 2013, pp 13-17.

[15]. Hall, M. A., "Correlation-Based Feature Selection for Machine Learning," (Doctoral dissertation, The University of Waikato), 1999.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 36-40(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Survey of Various Methods for Payload based Intrusion Detection System

T.S. Urmila[1] and Dr. R. Balasubramanian[2]
[1]Research Scholar, Mother Teresa Women's University,
[2]Dean, Karpaga Vinayaga College of Engg. & Tech.

ABSTRACT: The purpose of IDS (Intrusion Detection System) is to observe the intrusion in real time. However it is tedious to develop such a system with low false alarm rate and high detection accuracy. There are two kinds of Intrusion detection obtainable such as Anomaly and Misuse based Intrusion Detection. The anomaly detection system observes the actions of the client and analyzes the deviation of the behavior to detect intrusions. The misuse intrusion detection system evaluates the intrusion based on identified patterns. The intruders carefully craft the packet in order that it will evade the intrusion detector. If IDS solely contemplate the packet header data, these attacks contain no malicious activity because the header fields do not violate any protocol, and they do not continuously generate abnormal network traffic. Thus we have to depend upon the packet payload to defend against such attacks. This paper illustrates various methods of the anomalous and misused based techniques used for detecting the intrusions.

Keywords: Intrusion Detection System (IDS), anomaly, Payload Anomaly Intrusion detection, misuse based hybrid methods.

I. INTRODUCTION

Intrusion Detection is that the methodology by which intrusions is detected. This technique can be divided into two categories: "anomaly" intrusion detection and "misuse" intrusion detection. The first refers to intrusions that can be detected based on abnormal behavior and use of computer resources. This system of detecting intrusions makes an attempt to quantify the good or acceptable behavior and alternative irregular behavior as intrusive. A main premise of anomaly intrusion detection is that intrusive activity is a subset of anomalous activity. This might seem affordable, considering that if an outsider breaks into a computer account, with no notion of the legitimate user's pattern of resource usage, there is a good chance that his behavior will be anomalous. In contrast, the second, misuse intrusion detection, refers to intrusions that follow well outlined patterns of attack that exploit weaknesses in system and application software. Such patterns are often precisely written in advance.

Some attacks exploit the vulnerabilities of a protocol; alternative attacks obtain to survey a website by scanning and probing. These attacks will usually be detected by analyzing the network packet headers, or observance the network traffic affiliation attempts and session behavior (eg., DOS, DDOS, Probe attack). If we solely consider the packet header info, these attacks contain no malicious activity because the header fields don't violate any protocol, and they don't perpetually generate abnormal network traffic. So we've got to depend upon the packet payload to defend against such attacks. Other attacks, like worms, (R2L,U2R) involve the delivery of dangerous payload to a vulnerable service or application. These could also be detected by inspecting the packet payload. State of the art systems designed to discover and defend systems from these malicious and intrusive events depend upon "signatures" or "thumbprints" that are developed by human experts or by semi-automated means from best-known prior dangerous worms or viruses. They do not solve the "zero-day" worm problem, 0068 owever; the primary occurrence of a new unleashed worm or exploit. Payload based IDSs commonly suffer from 3 major problems, particularly higher false alarm rates, complexness of high dimensional information and dynamic structuring of protocols. This is because payload-based IDS use massive number of features to discriminate traditional packets and anomalous packets in the network traffic data.

II. ANOMALY INTRUSION DETECTION

Anomaly intrusion detection system develops a profile that consists of obviously normal packet behavior and generally attack behavior [1]. After making a profile it will be checked against a real time packet captures to investigate the deviation and report the abnormal behavior. while planning a anomaly based mostly intrusion detection system the subsequent criteria ought to be considered; 1) should handle immense volume of information 2) incremental change of profile 3) Low false alarm rate 4) expeditiously detect the mimicry attacks 5) should operate in high bandwidth environment 6) unsupervised learning. Today, the

importance of defensive against zero day attacks is becoming progressively important [1,3].

A. PAYL (Payload Anomaly Intrusion detection)

Wang and Stolfo proposed PAYL which used 1-gram to build a byte-frequency distribution model of network traffic payloads. The pre-processing of packet payload using 1-byte sliding window creates a feature vector containing the relative frequency count of each of the 256 possible 1-grams (bytes) in the payload. The profile of byte frequency distribution and standard deviation of the payload were built during the training phase [4]. Then in the detecting phase, the Mahalanobis distance was used to compute the difference between the incoming data and the profile. This scheme proves to work well at identifying new application level activities including malicious executable files or Internet worms.

B. Payload Content based Network Anomaly Detection (PCNAD)

In PCNAD, profile was created for normal payloads for a particular service on a host and makes a set of payload profiles that are expected for that service [2]. The payload's profiles are specific to the host and the communication behavior of the service; hence same profiles are not applicable across the different network environments. The system calculates byte frequency distribution using 1-gram based approach to make payload profile. Multiple profiles are created for different payload lengths. Due to this, number of profiles for a particular service becomes very large. To minimize the complexity of profile comparisons, profiles are clustered together. The clustering techniques used are lengthwise clustering and profile-wise clustering. PCNAD initially applies lengthwise clustering which combines profiles where difference between two lengths is less than the threshold. The resultant profile has byte frequency distribution and length equal to the average payload length of combined profiles. The value of threshold is kept user configurable. Once lengthwise clustering is done profile-wise clustering is applied. In profile-wise clustering the sparse profiles are combined using Manhattan distance. The system is trained in an unsupervised way for profile creation. In the testing phase the system captures incoming payloads and compares the payload with stored normal profiles. If the new payload profile does not match with any stored profile for the same service, then an alert is generated indicating a suspicious packet. If the arriving payload was found to be normal with comparison to some stored profile, then that stored profile is combined with arriving payload profile using technique used in the lengthwise clustering.

C. ALAD (Application Level Anomaly Detection)

ALAD focus the main plan is to extract the primary word of every line within the payload as a keyword. during the training phase, a keyword set should be constructed by assembling all potential keywords. These keywords are then related to different corresponding properties to construct the profile. For instance, in ALAD, keywords are connected with destination port in the packet such as "21:220" and "80: GET". Every port typically corresponds to a particular application or protocol, so it should have a limited set of keywords. In the recognition phase, if a new keyword is found, ALAD increases the anomaly score. When the anomaly score reaches the threshold, an alarm will be generated.

D. ANAGRAM

To model the structure of payload, Wang et al. proposed ANAGRAM. supervised learning was utilized to model normal traffic and attack traffic by storing n-grams of normal packets and attack packets into two separate Bloom Filters (BFs) [5]. However, consistent with Perdisci *et al.* BFs would not work in high bandwidth or high data rate networks, because ANAGRAM stores n-grams within the BFs and generates a score supported the quantity of unobserved and maliciousn-grams throughout detection phase. Unfortunately, it was more troublesome to construct an correct model due to the curse of dimensionality and attainable computational problem.

McPAD. McPAD use multiple one class Support Vector Machines (SVMs) to discover abnormal packets by majority voting. Taking motivation from Anagram which randomizes the length of short sequences, McPAD uses changed kind of n-grams known as n vgrams throughout training [6]. Then vgrams are substrings of a string in which every substring of length ' n ' is separated from alternative substring of length ' n ' by length ' v '. By varied parameter v, every payload gets depicted in numerous feature space. for every price of v, features formed are passed into a different one class SVM for training.

Table 1: Comparison of Various Anomaly based Methods.

Method	Pros	Cons
PAYL	Simple and has fixed number of features	Higher false alarms rate, can be evaded with mimicry attack, considers entire payload which presents a major problem in high-speed bandwidth network

Anagram	Higher order n-grams and resistant of mimicry attack	False positives of n-grams
McPAD	Resistant to mimicry attacks	Multiple classifiers training required.
PCNAD	Safe against mimicry attacks Good result for port 80 and 21	poor results at ports like 22, 23, 25. false positive rate is more
ALAD	Work well for application level payload combined with keyword based approach	High false alarm rate

III. SIGNATURE BASED INTRUSION DETECTION

Network intrusion detection systems (NIDS) remains the signature based approach, that relies on signatures of already best-known attacks or vulnerabilities. This technique works well if the precise patterns of certain attacks may be found, thus it is able to observe such activities by matching the pattern. It's far more reliable than anomaly based strategies on the condition that the attack signature or fingerprint might be known [7]. However, for new attacks, or mutations of notable attacks, whose fingerprints have not been discovered, the signature based approach might miss detecting the attack. For the signature based NIDS, finding the distinctive fingerprint of a particular attack is the key issue, which is typically done manually or semi-automatically. The target is to border the intrusion detection drawback as a pattern matching one, and devise efficient algorithms for such matching. In order to seek out the suspicious packets, most of IDSs use a pattern matching algorithmic rule. The formula checks the presence of a signature in the incoming packet sequence and outputs the situation of the string among the packet. The formula should be quick enough to detect the malicious behavior, and it should be scalable in order to satisfy the rise in each the quantity of signatures and the link speed. String matching algorithms are often classified into single and multiple pattern matching algorithms. In the single pattern matching, one pattern is matched against the complete text at a time. In distinction, the multiple patterns matching approach compares the text sequence against all signatures all directly. Obviously, the multiple matching approaches are a far better alternative for intrusion detection to avoid sweeping the packet persistently. However, it consumes additional memory and requires a pre-processing phase to program the patterns before matching will commence.

A. Myersalgorithm

The Myers algorithmic rule is an approximate string matching algorithmic rule The Myers algorithmic rule depends on an easy dynamic programming (DP) concept. It uses algorithmic formulas and simple bit operations to compute the edit distance between the text and patterns to seek out the equalities or differences [10]. The edit distance between two strings is expressed because the minimum edit operations needed to remodel a text t1 to another text t2 or vice versa. Commonly, there are three typical variations of edit distance. The primary form is called the hamming distance. It computes the quantity of positions in the text that has different characters, i.e., how many characters are required to convert a text t1 to a different text t2. The compared text strings should be of the same length. These condform is called the Levenshtein distance, which does not have any restriction over the text size. The edit distance is the minimum range of edit operations: insertion, deletion, and substitution, that are required to convert two strings into one another. The third one is the Damerauedit distance. It permits the transposition of two adjacent characters to complete the conversion between the two strings.

B. Exclusion-based Signature Matching (ExB)

The basic plan is to determine if the input (e.g., each packet received) contains all fixed-size bit-strings of the signature string, without considering if the bit-strings seem in-sequence, as done by existing algorithms [7]. If a minimum of one bit-string of the signature doesn't seem within the packet, then ExB determines that the signature does not match. The little size of the input ensures that ExB matches correlate well with actual matches. This approach also allows for a straightforward and efficient implementation: for every packet, ExB creates an occurrence bitmap marking every fixed-size bit-string that exists within the packet. The bit-strings for every signature are then matched against the occurrence bitmap. As packets seldom expected to match any signature, ExB performs higher within the common case compared to existing algorithms. Within the case of false matches (e.g., when all fixed-size bit-strings show up, however in arbitrary positions within the input), ExB falls back to standard algorithms.

C. Piranha

Piranha relies on the concept that if the rarest substring of a pattern does not seem, then the complete pattern will certainly not match. Every pattern is depicted by its least common 4-byte sequence, wherever standard reflects the amount of times that a particular substring

exists in all patterns. Piranha treats each byte-aligned pattern as a collection of 32-bit sub-patterns to quicker operations [11]. Pattern matching will then be developed in terms of an AND operation. Each pattern is represented by a gate. The gate has as several inputs because the range of its 32-bit sub-patterns. Every input represents whether the 32-bit sub-pattern has appeared within the payload or not. The rarest sequence is chosen as representative. It is defined as the sequence found in the least range of rules and might be found through the index table by counting the number of rules that is contained in. All alternative inputs are removed from the gate as well as the corresponding nodes from the index table.

D. E^2xB

E^2xB is designed for providing quick negatives when the search pattern does not exist in the packet payload, assuming a relatively small input size [9]. As mismatches are by far more common than matches, the pattern matching can be enhanced by first testing the input for missing fixed-size substrings of the original signature pattern called elements. The collisions induced by E^2xB, with all fixed-size substrings of the signature pattern showing up in arbitrary positions within the input, can then be separated from the actual matches. The small input assumption ensures that the rate of collisions is reasonably small.

E. Wu-Manber Algorithm

The MWM algorithm is predicated on dangerous character heuristic with one or two-byte bad shift table created by pre-processing all the patterns rather than just one. MWM performs a has on the two-character prefix of this input to index into a bunch of patterns, that are then checked ranging from the last character. It's recently enforced in Snort.

F. Exscind (exclude from Union)

It introduces an exclusion-inclusion filter that excludes clean traffic while not the requirement to perform expensive pattern matching. For that purpose a Bloom filter is changed to produce probable matches to more reduce the amount of pattern matching operations needed for suspicious packets [9]. The filter is programmed with solely the prefix of all Snort signatures so as to stay the filter process overhead to minimum and speed up matching by skipping clean packets. The incoming packet is hashed and queried for those prefixes. IF the query is negative then the packet is clean and may safely be skipped. If the query is positive then the packet most likely contains an attack signature prefix and needs more matching. The suspicious packets are explore for simply a set of signatures as against all snort signatures. additionally, the filter indicates the position among the packet wherever the primary probable match was found, to look a part of the packet ranging from that position as against looking the entire packet. Exscind works at packet level exploitation one window, one buffer for storing signature IDs and one signatures bloom vector.

Table 2: Comparison of Various String Matching Algorithms.

Method	Pros	Cons
Myers algorithm	Approximate string matching	High preprocessing cost incurred
Wu-Manber	Perform well on large sets	Performance degrades when short patterns occurs.
E^2xB	Quick decisions with small number of collisions	Additional preprocessing cost per packet
Piranha	Quick decisions on which patterns may match , compact memory footprint	Generate collision in most of the payload offset.
ExB	Performs better in common case.	If false match occurs it degrades to standard algorithms
Exscind	Scales very well with increasing number of signatures	Both hardware and software based and depends on the efficiency of Bloom filter

IV. CONCLUSION

Packet payload needs more attention for malicious packet contents. In order to detect it both anomalous and signature based methods are used. Anomalous based methods create profile of normal behavior and calculate the deviation using various methods. Similarly signature based payload detection methods formulate the problem as a pattern matching problem and proposes various string matching algorithms. Future intrusion detection systems may consider both anomaly and misuse based hybrid methods for detecting zero day attacks and known attacks.

REFERENCES

[1]. Evangelos P. Markatos, Spyros Antonatos, Michalis Polychronakis, Kostas G. Anagnostakis Exclusion-based Signature Matching for Intrusion Detection.
[2]. Sandeeo Kumar Eugene H. Spalford A Pattern matching model for Misuse intrusion detection.

[3]. Like Zhang, Gregory B. White Analysis of payload based application level network anomaly detection Proceedings of the 40th Hawaii International Conference on System Sciences – 2007
[4]. Sandeep A. ThoratAmit K. Khandelwal Bezawada Bruha deshwar K. Kishore Payload Contentbased Network Anomaly Detection
[5]. Sandeep Kumar Eugene H. SpaffordAn Application of Pattern Matching in Intrusion Detection.
[6]. Mayank Swarnkar, Neminath Hubballi OCPAD: One class Naïve Bayes classifier for payload based anomaly detection Expert Systems With Application.
[7]. Monther Aldiwairi Duaa Alansar Exscind: Fast pattern matching algorithm using Exclusion and Inclusion features.
[8]. Monther Aldwairi, Ansam M. Abu-Dalo1 and Moath Jarrah Pattern matching of signature-based IDS using Myers algorithm under MapReduce framework
[9]. S Antonatos, M. Polychronakis, P. Akritidis, k.G. Anagnostak is and E.P. Markatos PIRANHA: Fast and Memory-Efficient pattern matching for Intrusion detection
[10]. Frank S. Rietta Application Layer Intrusion Detection for SQL Injection.
[11]. Ke Wang and Salvatore J. Stolfo Anomalous Payload-Based Network Intrusion Detection

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 41-46(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

An Ontology Based Sentiment Analysis Using Protege Software

K.H. Rizwana[1] and Dr. B. Kalpana[2]
[1]M. Phil. Research Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India.
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: Product review data analysis and prediction in terms of sentiments is great challenge in the big data research. Social media provides a platform for users to share data on any topic. The knowledge could accommodate user's emotions, feedbacks, reviews and private experiences. In this research an ontology based sentiment analysis strategy for social media content (OSAPS) with negative and positive sentiment is presented. An ontology based method is designed to retrieve and analyze the customer's tweets with their emotions. Sentiment analysis done with feature extraction fails to give a deep insight about the users opinion so in our proposed approach, we reduced the feature set into clusters through Principal Component Analysis(PCA)with utilization of domain knowledge. Cluster data is classified using ensemble classification algorithm for fuzzy, neural network and support vector machine.

Keywords: Sentiment analysis, ontology, Protege, Support vector machine.

I. INTRODUCTION

The sentiment analysis of customer's social media data is very important in the present day business scenario. Customers share information about products, services and their experiences on social media. This information can be used for market research, product feedback and analyzing customer service effectiveness. Opinions describe people's sentiments or feeling towards entities, events or their properties. The sentiment analysis of the opinions could lead to many interesting results. The dynamically expanding web and social media are generating huge amount of opinion data. People's opinion about any product or service on social media is a very valuable asset for any organization. The organizations can generate information on customer's response, or its behavior for any product or service, by doing the sentiment analysis of these social media data. Sentiment or opinion mining is nothing but analyzing whether the given input is positive, negative or neutral in other words, we determine the polarity of the input. The input here can be any data source like blogs, review sites like Amazon reviews, restaurant reviews and micro-blogging like Twitter. The particular domain features can be extracted by building ontology for the interested domain. For ex., consider the sample tweet S: "The battery of Lenovo laptop was wonderful, although the screen size was bad". Scoring of the tweets can be done in two ways, either qualitative or quantitative. Qualitative is nothing but identifying whether positive, negative or neutral and quantitative is overall scoring of the tweets. In this paper, we score the tweets individually and also return how many positive, negative and neutral tweets are there. Here, in this example the opinion words are wonderful and bad, based on the opinion words, we score the tweets. Opinion words are contained in the opinion lexicon dictionary which is very important and the domain is Laptop which can be easily identified from the tweet and the features are: battery and screen size. From this sample tweet we build the ontology for the domain Laptop.

In this research, a process for sentiment analysis of customer's social media data using an ontology model is presented. This process would help in identifying the problem area associated with the customer's social media data that contains negative or positive sentiments. An ontology model and the use of ontology model for sentiment analysis, Section II provides the background research used for building the ontology model, and Section III provides information on building on discussion area of different attempts taken in this research, their limitation and usefulness.

A. Opinion Mining

Social media is one of the biggest forums to express opinions. Sentiment analysis involves the extracted information from the opinions, appraisal and emotions of people with regards to entities, events and their attributes [3].

Sentiment analysis is also known as opinion mining. Opinion mining analyses and clusters the user generated data like reviews, blogs, comments, articles etc. These data find its way on social networking sites like twitter and face book. Twitter has provided a very gigantic space for prediction of consumer brands, movie reviews, democratic electoral events, stock market, and popularity of celebrities.

Opinion mining or sentiment analysis is related with mining and analysis of natural language for tracking the mood or feedback of people about a particular product. It can be treated in short as a system to collect and classify different opinions about a particular product or service. For example, a review on a website might be overall positive about a digital camera, but can be specifically negative about how heavy it is.

B. Sentiment Analysis

Opinion analysis is performed to identify the opinion of people or groups. It can be performed using Lexicon Based approach or Machine Learning based approach. Some methods are still not efficient in extracting the sentiment features from the given content of text. Naive Bayes, Support Vector Machine are the machine learning algorithms used for sentiment analysis which has only a limited sentiment classification category ranging between positive and negative.

Sentiment analysis makes use of three terms in order to fetch the sentiment. That is object and feature, opinion holder, opinion and orientation. The technical challenges are object identification, opinion orientation classification, and feature extraction. Usually sentiment analysis can be performed using supervised and unsupervised learning such as naive Bayes, Neural Networks, and Support Vector Machines.

Among these three techniques SVM is considered to be more suitable for sentiment analysis. Sentiment classification can be performed in 3 stages such as [19].

• Document level
• Sentence level
• Feature level

In document and sentence level the sentiment analysis makes use of only a single object and extracts only a single opinion from the single opinion holder. But these types of assumptions are not suitable in many situations. Extracting sentiment for entire document/blog is not be as efficient as extracting sentiment by considering aspects of each subject in the particular sentence.

Challenges in Sentiment Analysis

Some of the challenges in Sentiment analysis are:

a) Polarity Shift

Polarity Shift is the most important issue to be addressed in Sentiment analysis. Polarity Shift means that Polarity (Sentiment) of the sentence is calculated in different way from the polarity actually expressed in the Sentence. This problem mainly arises due to polarity shifters such as negation (e.g."I don't like this bike") and contrast (e.g. "good, but it's not my style"). In the above mentioned example the sentence "I don't like this bike" is very similar to "I like this bike". Here the polarity shifter is "Don't".

b) Binary Classification

Binary Classification is another important problem to be addressed in which the given review's Polarity is classified only by using "Positive", "Negative" by ignoring the "Neutral". This type of problem mainly arises when the sentiment classification is purely based on machine learning algorithms. Opinion mining that only considers positive and Negative will not have good accuracy. Now-a-days the classification is extended by considering 5 possibilities such as "Positive", "Strong Positive"," Negative", Strong Negative" and "Neutral". By increasing the classification category it is possible to improve the accuracy of the opinion mining.

c) Data Sparsity problem

Third issue to be addressed is Data sparsity problem which is caused due to the imposed character limit in micro blog/social media websites. For instance the maximum character limit in twitter is 140.Due to this limitation people will not express their opinion in clear manner. All these three issues can attain the accuracy of the sentiment analysis [16].

Machine Learning Techniques. Machine learning, may be a discipline involved with the planning and development of algorithms that permit computers to evolve behaviors supported empirical knowledge or databases. Machine learning focuses on prediction, supported glorious properties learned from the trained data; data processing focuses on the invention of unknown properties within the knowledge. Data processing uses several machine learning ways, however with totally different goals. Machine learning employs data processing ways as "unsupervised learning" or as a preprocessing step to enhance learner accuracy. An important feature of machine learning analysis is to mechanically learn to acknowledge advanced patterns and build intelligent choices supported by data. Machine learning is useful in minimizing the loss on unseen samples.

Supervised Learning

Supervised learning is the machine learning task of inferring a function from labeled training data [8]. This is called supervised because the machine is told what is what, a significant number of times, and then is expected to predict outcome of its own.

Below are few examples

- Identifying if a news article belongs to a sports news or politics.

- Classify an animal in one of the predefined classes like mammal, bird etc.
- Classify a person as male or female based on the products bought by the user.

Some of the widely used supervised algorithms are
1) Naïve Bayes
2) Support Vector Machines
3) Random Forests
4) Decision Tree

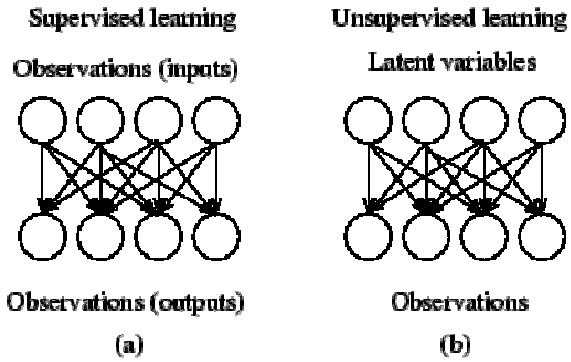
Unsupervised Learning

This technique is used when the groups (categories) of data are not known [8]. This is called unsupervised as it is left on the learning algorithm to figure out patterns in the data provided. Clustering is an example of unsupervised learning in which different data sets are clustered into groups of closely items. Some of the use cases of unsupervised learning are as follows:

- Given a set of news reports, cluster related news items together. (Used by particular websites).
- Given a set of users and movie preferences, cluster users who have similar taste.

Some of the widely used unsupervised algorithms are

1) K-Means
2) Fuzzy clustering
3) Hierarchical clustering



II. LITERATURE REVIEW

A summary of the review of sentiment analysis techniques is given in Table I

Table I: Sentiment Analysis- Approaches.

AUTHOR	PUBLICATION/YEAR	TITLE	TECHNIQUE/ALGORITHMS	LIMITATIONS
K. M. Sam and C. R. Chatwin	International Journal of e-Education, e-Business, e-Management and e-Learning Vol. 3, No.6, December 2013	Ontology-Based Sentiment Analysis Model of Customer Reviews for Electronic Products	Ontology management module, the user query processing module, information foundation module and query analysis engine module.	The extracted keyword issues are occurring due to consumer model.
Ankita Gupta, et al	International journal of computer science and mobile computing,2017	Sentiment analysis of tweets using machine learning approach	KNN + SVM	-
Khin Phyu Shein	proceedings of the 3rd international conference on communications and information technology	Ontology based combined approach for Sentiment Classification	Support vector machine	The mining issues are classified on the sentiment classification.
Mohammad Mustafa Taye	The research bulletin of Jordan A C M , I S S N : 2 0 7 8 - 7 9 52 , v o l u m e I I (I I)	Web-Based Ontology Languages and its Based Description Logics	DAML+OIL	Protege is supported only in the owl in the w3c recommendation.
Sheng Li, Lingling Liu, Zenggang Xiong	International Journal of Digital Content Technology and its Applications(JDCTA)Volume6,Number23,December 2012vol6.issue23.4	Ontology-Based Sentiment analysis of Network Public Opinions	Manual, Domain ontology, frequent pattern mining	Protégé cannot be extended with pluggable Components to add new functionalities and services.

Kim schouten, et al	IEEE xplore digital library,2017	Supervised and unsupervised aspect category detection for sentiment analysis with co-occurrence data	Association rule mining	. Limited number of data sets.
Subhabrata Mukherjee and Sachindra Joshi	IBM India research lab.	Sentiment Aggregation using Concept Net Ontology	Concept Net , Corpus Sentiment Aggregation	The ontology method is used to translate the concept in the mapping method.

Ontology based sentiment analysis

- Ontology is a knowledge base of structured list of concepts, relations and individuals.
- Hierarchical relationship between the product attributes can be best captured by an Ontology Tree.
- Ontology creation is expensive, highly domain-specific. In this work, we use Concept Net (Hugo *et al.*, 2004) to automatically construct a domain-specific ontology tree for product reviews.
- Concept Net is a very large semantic network of common sense knowledge Largest, machine-usable common sense resource consisting of more than 250,000 propositions [18].

Sentiment Annotated Ontology Tree:

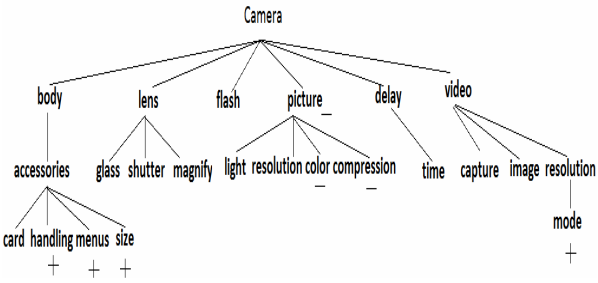


Fig. 1. Sentimented ontology Tree.

Domain Ontology Tree Creation

We construct a domain-specific ontology tree for product reviews. Concept Net is a very large semantic network of common sense knowledge which can be used to make various inferences from text. It is the largest, machine of common sense resource consisting of more than 250,000 propositions. Mining information from Concept Net can be difficult as one to- many relations, noisy data and redundancy undetermined its performance for applications requiring higher accuracy [18].

However, we use Concept Net for the following reasons:

1. The relational predicates in Concept Net have an inherent structure suitable for building ontology.
2. Concept Net has a closed class of defined relations. The relations can be suitably weighted and used for various purposes.
3. The knowledge resource through crowd-sourcing incorporates new data and enriches the ontology.
4. Ontology creation using Concept Net does not require any labeling of product reviews [18].

Ensemble Fuzzy Domain Sentiment Ontology Tree (EFDSO):

Online product reviews became a crucial opinion resource that several researchers pay their attention to a completely unique technique of opinion mining is projected and evaluated by a group of real on-line product reviews [20]. A hierarchical fuzzy domain sentiment ontology-FDSO has been introduced by this approach, which defines a space of product features and corresponding opinions, thus making it possible for a product to be classified and scored by commonly accepted features. This will enhance the user experience to search a product and compare it with other products feature by feature. The evaluation is based on the Chinese product reviews collected from 360buy.com. The experimental results show that the approach is able to automatically identify the polarity for a large of sentiment expression.

III. PROPOSED METHODOLOGY

The proposed methodology consists of two processes. The first process is to build the ontology model using the data extracted from the social media platform. The second process is to retrieve the problem area from the negative sentiments associated with a tweet using a previously built ontology model. These processes have tasks such as social media data extraction and data cleaning, identifying negative sentiment in tweets text, subjective analysis, building ontology model, query building, and retrieving information from the ontology model.

Building an Ontology model

The process of building an ontology model is shown in Fig. 1. The data is to be extracted from the social media platform, Twitter. The postal service domain with its class, object and object properties are used to build ontology model. H. Cunningham, et al. showed a

methodology to do text parsing using GATE software [13]. The subjective analysis of tweet data was done to identify the objects and their object properties for the postal service domain. A combination of NLP-based language parsing plugins in GATE was used for annotating the nouns and verbs in the tweet. The output of the GATE software was in a tag form with nouns and verbs enclosed in tags. Data cleaning was done to get the nouns and verbs from the results. The script written in Python was used for data cleaning. The data after cleaning had redundancies of noun and verb. These redundancies were removed by using Excel Macros. The final results had only nouns and verbs from tweet texts.

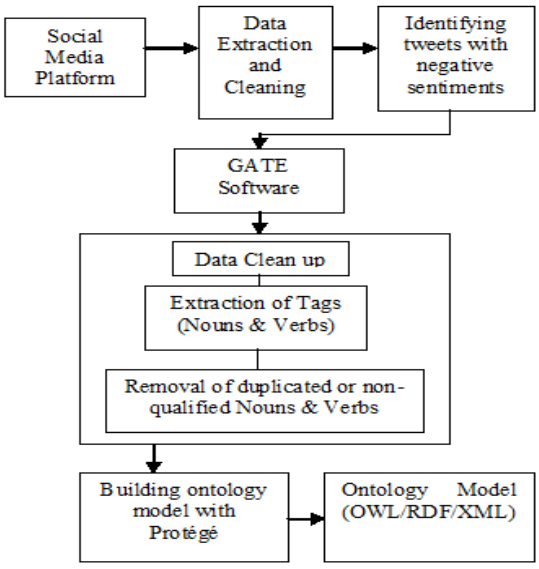


Fig. 2. Ontology model building Process.

Protégé software is used to build the ontology model. Class, object and object property are identified as entity, individual, and object property in the ontology model respectively. The relations between classes, objects and object properties were derived manually as per the human understanding of a sentence.

IV. CONCLUSION AND FUTURE WORK

This paper presents a method of ontology-based sentiment classification to classify and analyze the online product reviews. We implement and experiment our assumption with Support Vector Machine based on the lexical variation ontology. This research will focus on building an ontology model to identify the problem areas associated with customer's dissatisfaction. This is done by analyzing their social media content. A partially automatic process is developed to find out the problem area associated with the content shared on social media with the help of an ontology model.

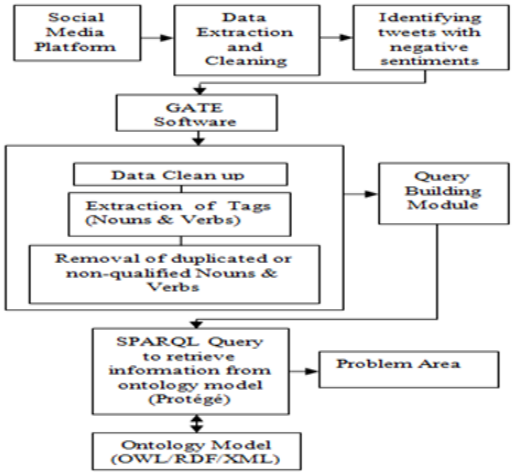


Fig. 3. Sentiment analysis using ontology model.

Data entry in Protégé software to build an ontology model, and data cleaning modules will be executed manually. Few modules such as subjective analysis with GATE software and SPARQL query building need further work in order to optimize the process and to get accurate results. The uniqueness of relation between a class, an object and an object property needs to be maintained while refining the ontology model. The identification of the tweets with the negative sentiments needs to be done in a more optimized and efficient way. The accuracy of result also depends on the uniqueness of interrelations and the knowledge built in the ontology model shown in fig. 1.

REFERENCES

[1]. Ankita Gupta, *et al.* "Sentiment analysis of tweets using machine learning approach" *International journal of computer science and mobile computing*, 2017.

[2]. Bilan V. *et al.* "An Ontology Based Approach to Opinion Mining"

[3]. Bing Liu,Lei zhang " A survey of opinion mining and sentiment analysis" Springer + Buisness media, LLC 2012.

[4]. Prof. Durgesh M. Sharma, Prof. Moiz M. Baig "Sentiment Analysis on Social Networking: A Literature Review" *International Journal on Recent and Innovation Trends in Computing and Communication*, Volume **3** Issue: 2 022– 027.

[5]. Efstratios Kontopoulos, Christos Berberidis, Theologos Dergiades, Nick Bassiliades "Ontology-based sentiment analysis of twitter posts" Expert Systems with Applications (2013).

[6]. Jeevanandam Jotheeswaran, Dr. S. Koteeswaran Sentiment Analysis "A Survey of Current Research and Techniques" *International Journal of Innovative Research in Computer.*
and Communication EngineeringVol. 3, Issue 5, May 2015

[7]. Khin Phyu Shein "Ontology based combined approach for Sentiment Classification" proceedings of the 3rd international conference on communications and information technology.

[8]. Kim Schouten, Onne van der Weijde, Flavius Frasincar, and Rommert Dekker
"Supervised and unsupervised aspect category detection for sentiment analysis with co-occurance data" IEEE transactions on cybernetics.
[9]. Lirong Qiu, "An Opinion Analysis Model for Implicit Aspect Expressions based on Semantic Ontology" *International Journal of Grid Distribution Computing*, Vol. **8**, No.5, 2015.
[10]. Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu and Chenggen Sun "Opinion Mining and Sentiment Analysis in Social Networks A Retweeting Structure-aware Approach" *IEEE/ACM 7th International Conference on Utility and Cloud Computing 2014.*
[11]. G. Parthasarathy and D. C. Tomar "A Survey of Sentiment Analysis for Journal Citation", *Indian Journal of Science and Technology,* Vol. **8**(35), December 2015.
[12]. Mohammad Mustafa Taye "Web-Based Ontology Languages and its Based Description Logics" The research bulletin of Jordan A C M, I S S N : 2 0 7 8 - 7 9 52 , V o l u m e I I (I I) .
[13]. Pratik Thakor,Dr.Sreela sasi "Ontology based sentiment analysis process for social media content "Procedia computer science, Elsevier vol. **53,** 2015.
[14]. K. M. Sam and C. R. Chatwin "Ontology-Based Sentiment Analysis Model of Customer Reviews for Electronic Products" International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. **3**, No.6, December 2013.
[15]. R. Sangeetha., Dr. B. Kalpana. "Optimizing the Kernel Selection for Support Vector Machines using Performance Measures". A2CWIC 2010, Sept 16-17, 2010, ISSN 978-1-4503-0914-7, ACM Digital Library.
[16]. Saurabh Dorle et al. "*International journal of innovative computer science and Engineering*, 2017"
[17]. Sheng Li, Lingling Liu, Zenggang Xiong "Ontology-Based Sentiment Analysis of Network Public Opinions" International Journal of Digital Content Technology and its Applications(JDCTA)Volume6,Number23,December 2012vol6.issue23.4
[18]. Subhabrata Mukherjee and Sachindra Joshi "Sentiment Aggregation using ConceptNet Ontology" IBM India research lab.
[19]. Walaa Medhat, Ahmed Hassan, Hoda Korashy "Sentiment analysis algorithms and applications" *Ain Shams Engineering Journal* (2014).
[20]. Xinhui Nie, Lizhen Liu, Hanshi Wang, Wei Song "The Opinion Mining Based on Fuzzy Domain Sentiment Ontology Tree for Product Reviews" *journal of software,* vol. **8**, no. 11, November 2013
[21]. https://www.researchgate.net/publication/266204394 "Sentiment Mining using principal component analysis [accessed Dec 12 2017].

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 47-52(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Review on Fuzzy Based Packet Dropping and Collaborative Attack Detection in MANET Using DSR Protocol

D. Nethra Pingala Suthishni and Dr. G. P. Ramesh Kumar
Department of Computer Science,
Sri Ramakrishna College of Arts & Science, Coimbatore (TN), India.

ABSTRACT: Network security tends to be a major challenge in computing as various attacks are emerging day by day. When compared to wired networks, MANETs are more vulnerable to security attacks due to lack of trusted centralized authority and limited resources. Due to the ad hoc nature of MANETs, they are used in various commercial and military applications. Its dynamic nature, infrastructure less environment, wireless links, multi-hoping feature, autonomous node movements and lack of centralized control make them more vulnerable to attacks and misbehavioral than traditional networks. In order to attain an effectual security paradigm, the following requirements such as accessibility, legitimacy, data confidentiality, reliability and non-repudiation should be ensured in MANETs. This paper emphasizes on fuzzy based packet dropping and collaborative attack detection in MANET using DSR protocol. The paper is a simulation based study on DSR routing protocol in MANET via Network Simulator tool (NS2) in the existence of the above mentioned routing attacks. Fuzzy Logic prediction rules are used here in classification and detection of attacks. Performance of the overall system is measured in terms of various performance metrics such as throughput, packet delivery ratio, end-to-end delay, jitter, routing overload and control overhead.

Keywords: Attacks, DSR Protocol, Fuzzy Logic, MANET, NS2.

I. INTRODUCTION

In MANETs, mobile nodes communicate liberally among each other without the need of definite structure. This efficiency and suppleness makes these types of networks appropriate for applications such as military zone, business zone, resident environments, omnipresent computing and sensor networks. Adhoc Network connecting nodes in the network plays the task of a host and a router in forward packets. With the help of routing protocols such as DSR, mobile nodes communicate among each other within the network. Except MANET is responsible for different kind of attack due to its dynamic wireless topology and exclusive of any pre-defined communications. Fuzzy systems have manifest their ability to resolve different type of problems in a variety of application domain. Fuzzy systems based on fuzzy if-rules have been successfully used in detection systems to classify and detect attacks. The attacks in MANET are generally classified as internal and external attacks. The internal attacks tend to be more hazardous as they are initiate from inside the network and they are hard to notice. The main intent of this research is to classify and detect both packet dropping and collaborative nodes in Mobile ad hoc networks and establish secure communication between the nodes. Routing protocol DSR is used because of its source routing and route caching aspect [6].

II. STRUCTURE OF MANET

Fig. 1 shows the structure of Mobile Ad hoc Network. A Mobile ad hoc network is an autonomous or heterogeneous collection of nodes and mobile devices that interact with each other among a wireless media. The nodes in the network collaborate in a disseminate way to exchange data in an efficient way in the lack of predetermined infrastructure.

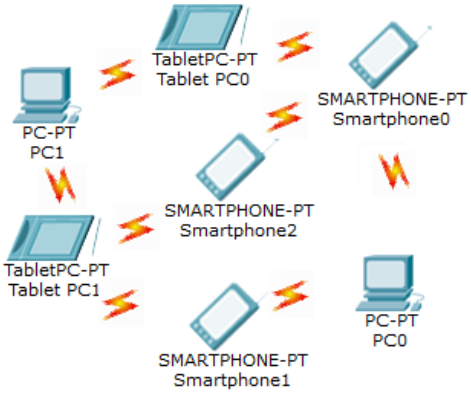


Fig. 1. Structure of MANET.

Hence, these ad hoc networks floors means for many real-time applications such as in disaster management and other such scenarios. The nodes in the network act

both as a host and as a router in forwarding packets over the network with the absence of access points. The transmission link between the nodes can be broken down easily as they are dynamic in nature. The comprehensive number of nodes on the network depends upon the application and environment it is deployed. Although, MANETs are used in real time systems, they lack in safe confinement. Therefore, the main objective of these ad hoc networks is to extend mobility into the sphere of autonomous nodes with wireless realm. MANET is more exposed to malicious attacks or intrusions due to the various vulnerabilities in the environment. They are more susceptible to attacks than wired networks due to the threats from compromised nodes inside the network, dynamic topological scenario, lack of administrative control and physical security [7].

III. ROUTING PROTOCOLS

MANET routing protocols are classified into two categories namely proactive routing protocols and reactive routing protocols[8]. Proactive routing protocols called as the table-driven ad hoc routing protocols preserves routing information of all the nodes in the network. Some of the proactive routing protocols include Destination Sequenced Distance Vector (DSDV) and Optimized Link State Routing Protocol (OLSR). Reactive routing protocols called as on-demand ad hoc routing protocol does not maintain state of the art information about the network layout and are effective and optimal for communication over ad hoc networks. Some of these protocols include Ad hoc On Demand Distance Vector Routing (AODV) and Dynamic Source Routing (DSR).

A. Dsr Protocol

DSR Protocol, as the name implies uses source routing for routing packets over the network. All the information about the routes is maintained in the source node whereas, in AODV protocol, the route information is maintained in intermediate nodes. DSR protocol is mainly used in multi-hop ad hoc networks. Consider when a source node desires to send a data packet to destination node. The source node initiates the transfer by sending RREQ packet to its neighboring nodes. This RREQ packet contains list of hops that is collected by the packet as it is disseminated through the network. It also holds the Unique ID, source and destination address of the packet sent. Once the RREQ packet reaches the destination node, it responds back with a RREP packet to the source node based on the shortest path and maintains all other information in the route cache.

The working nature of DSR protocol uses two phases of transmission. They are Route Discovery and Route maintenance. Consider the above scenario. When a source node S wants to send a packet to destination node D, the source node initiates the route discovery mechanism by broadcasting its RREQ packet to all its neighbors. Each node after receiving the RREQ packet rebroadcasts to its neighboring nodes. If the node receiving the packet has already received the packet, it discards the packet. Initially, the RREQ packet is empty holding only the unique packet id, source and destination address. When the packet is rebroadcasted each time, it holds the address of intermediate nodes. This rebroadcasting is done until the RREQ packet reaches the destination node D. When the RREQ packet reaches the destination, it replies back with a RREP packet to the source node with the reverse path of RREQ packet has travelled.

IV. CATEGORIZATION OF ATTACKS

Widely, security attacks in MANET are classified into two distinct categories [Revathi *et al,* 2012] as depicted in Fig. 2. They are:

- Internal attacks
- External attacks

A. External Attacks

External attacks are caused by nodes that do not belong to the domain of the network. They cause congestion, propagate false routing information or disturb nodes from providing services. These attacks prevent the network from healthy communication and generate additional overhead. These attacks are further classified into passive and active attacks. Some important passive attacks are snooping attacks, traffic analysis attacks and traffic monitoring attacks. Some important active attacks are Blackmail, Denial of service attack Fabrication, Gray hole Attacks, Disclosure Attacks, Routing Attacks and Recourse Consumption Attacks.

B. Internal Attacks

Internal attacks are attacks that are caused by the nodes within the network. They directly lead to attacks on nodes in the network and links interface between them[9]. These attacks are from compromised nodes that broadcast false routing information to other nodes on the network.

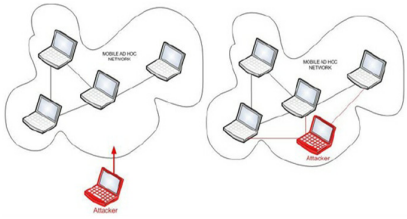


Fig. 2. External & Internal attacks.

Internal attacks are challenging than external attacks. Some of the internal attacks are packet dropping attack, collaborative attack and many more. This study imposes on packet dropping and collaborative attacks.

Packet dropping attack. A malicious node involved in a routing path may deliberately drop packets at network

layer so as to collapse the performance of network. An attacker or malicious node(s) drop the data packets not destined for disturbing the services or operations of the network[10]. For attaining the objective, the attacker requires to be on the routing path or take a part of routing operations so that the attacker drops RREQ, RREP AND RERR packets.

A packet drop may lead to packet loss. Packet loss takes place when packet(s) of data traversing across the network fail to reach the destination. A packet may be dropped under the following circumstances.

- **Selfishness of a node:** Nodes behave selfishly and fail to forward the received packets in order to conserve their limited resources battery power[11]
- **Validity of a node:** A packet may drop because of network congestion or channel conditions (free path loss, interference, noise etc) or shortage of energy]12]
- **Maliciousness of a node:** A node may behave maliciously and drops the packets under attack caused by an attacker.
- **Packet Loss:** Generally caused by network congestion.

Challenges of Packet Dropper Attacks. Following are the challenges of packet dropping attacks in MANET[13].

- The requirement that not only detect the node where the packet is dropped in the network, but also identifies whether the drop is an intentional or unintentional one.
- MAC level is limited due to which, whenever if the buffer is full any new packet coming from higher layers will be dropped. In this layer, the size of packet's transmission buffer level is limited.
- Rules of IEEE 802.11 protocol's: If the retransmission tries or the one of its corresponding RTS (Request to Send) frame has reached the maximum permitted number, due to node's movement or collision, a data packet will be dropped.
- If it is degraded during transmission due to some occurrence specific to radio transmissions such as interference, hidden nodes and high bit error rate a data packet may be dropped or lost.
- A selfish node may refuse to transmit a packet aiming to save its energetic resources in order to spread its lifetime or simply because its battery power is weak.

Some of the packet dropping detection techniques [Saira Aziz *et al,* 2016] proposed to classify and detect malicious nodes in mobile ad hoc networks are watchdog, pathrater, confidant twoack, aack and side channel monitoring [14]. Comparing various detection techniques with their shortcomings, all these methods require a new system for enhancement.

Collaborative attack. A collaborative attack in MANET is a homogeneous attack (i.e. blackhole or wormhole attack), concerning with two or more colluding nodes; classified as internal active attack that can be processed using wired or wireless link and caused by single or multiple attackers. It can also be referred to as the first level of attack, in which the adversary only interests in disturbing the foundation mechanism of the ad hoc network, for instance routing protocol, which is crucial for proper MANET operation [15].

Some of the characteristics of collaborative attacks include:

- Attackers can reveal subsequent interruptions in quick delays so the target network is unable to respond timely
- Attackers can also focus on a group of nodes or extend to different groups of nodes just for baffling the detection /prevention system in place †
- Attacks may be long-lived or short-lived ones †
- Internal and external users can collaborate to launch attacks

Challenges of Collaborative attacks. Following are the challenges of collaborative attacks in MANET.

- Synchronized attack induced by more than one attacker or combination of more than one attack that are also consistent to each other.
- Beguile normal routing information into false information.
- General understanding of the synchronization among attacks and/or the collaboration among various attackers …
- Representation and classification of Collaborative Attackers(CAs)
- Intrusion Detection Systems (IDS) capable of correlating CAs …
- Synchronized prevention or defense mechanisms

V. NEED FOR FUZZY LOGIC

Fuzzy Logic is essentially used in detection systems because of its ability to deal with uncertainty, imprecision, distorted and the ability to give acceptable reasoning rather than accurate reasoning. There are two main reasons to introduce fuzzy logic for attack detection. First, many quantitative features, both ordinal and categorical, are involved in intrusion detection and can potentially be viewed as fuzzy variables. For instance, the CPU usage time and the connection duration are two examples of ordinal measurements. An example of a linear categorical measurement is the number of different TCP/UDP services initiated by the same source host. The second reason to introduce fuzzy logic for intrusion detection is that security itself includes fuzziness. Given a quantitative measurement, a range value or an interval can be used to denote a normal value. Then, any values falling outside the interval will be considered anomalous to the same degree regardless of their different distances to the interval.

The same applies to values inside the interval, i.e., all will be viewed as normal to the same degree. Unfortunately, this causes an abrupt separation between normality and anomaly [3][4]. In MANETs, if-then

based fuzzy rules are used in all scenarios to identify attacks. The knowledge base or the rule base holds the if-then rules set by the user. This fuzzy rule based system is called as the fuzzy inference system (FIS) that is liable to take decisions in the attack detection process.

Fuzzy Logic can be used in intrusion detection systems to detect attacks due to its various modules that allow the user to identify and detect attacks over the mobile ad hoc network environment [16]. Following are the four main modules of fuzzy logic system.

- **Extraction of fuzzy based parameters**
  The system obtains desired parameters for analysis from network traffic and then passes these parameters to the next module.
- **Fuzzy inference system**
  Fuzzy rules and membership functions are executed on these parameters to find the fidelity level of each node in the network.
- **Fuzzy decision system**
  The fidelity of each node is compared to a threshold value for finding out the behavior of each node in this module.
- **Response module**
  If the calculated fidelity level is less than the chosen threshold value, the node will be considered as malicious and the response module is activated.

VI. RELATED WORKS

Sujatha *et al*. [5] proposed a new fuzzy based response model (FBRM) for the detection of internal attacks in mobile ad hoc network. In this type of attack detection, they have considered false route request (FRR) attack due to this attack flooding, congestion, DoS attack, exhaustion of resources and exhaustion of bandwidth could happen at nodes in the MANETs. In this method, Fuzzy logic controller monitors various feature such as route request rate, sequence number, Acknowledgement time and load pattern which can detect FFR attack.

Mohammed Abdel-Azim *et al*. [16] proposed an optimization of a fuzzy based intrusion detection system which automate the process of producing a fuzzy system by using an Adaptive Neuro-Fuzzy Inference System (ANFIS) for the initialization of the FIS and then optimize this initialized system by using Genetic Algorithm (GA). In addition, a normal estimated fuzzy based IDS is introduced to see the effect of the optimization on the system. From this study, it is proven that the optimized proposed IDS perform better that the normal estimated systems.

Aishwarya Sagar Anand Ukey *et al*. [17] proposed a new reputation based approach that deals with such routing misbehavior and consists of detection and isolation of misbehaving nodes. Proposed approach can be integrated on top of any source routing protocol and based on sending acknowledgement packets and counting the number of data packets of active path.

Sonal *et al*. [18] proposed a solution against black hole attack which is based on fuzzy rule .fuzzy rule based solution identify the infected node as well as provide the solution to reduce data loss over network.

Kulbhushan *et al.* [19] proposed an intrusion detection system for MANETs against blackhole attack using fuzzy logic. The system successfully detects the blackhole in the network and this information is passed to other nodes also. A detailed performance evaluation is provided based on various network parameters. Our results show that the proposed system not only detects the blackhole node, but improves the performance of AODV under the blackhole attack.

Alka Chaudhary *et al*. [20] proposed a novel intrusion detection system based on soft computing techniques for mobile ad hoc networks. The proposed system is based on neuro-fuzzy classifer in binary form to detect, one of vey possible attack, i.e. packet dropping attack in mobile ad hoc networks. Simulation results show that the proposed soft computing based approach efficiently detect the packet dropping attack with high true positive rate and low false positive rate.

A. Sharma *et al*. [21] proposed Fuzzy logic as trusted tool for mitigating the Collaborative Blackhole attack in MANET. The system suggests a trusted-fuzzy-ad-hoc routing protocol to upgrade the trust between the nodes in MANET using AODV routing protocol. Mischievous behavior of nodes is predicted on the basis of mobility based constraints that confirms the reliability to the network establishes the trust that avoids the malicious node generation. The result analysis between the proposed technique with the pre-existent technique regarding the routing overhead, throughput, packet delivery ratio shows the effectiveness of trusted-fuzzy-ad-hoc routing protocol in the secure MANET environment.

Y. Harold Robinson *et al.* [22] proposed a Fuzzy Logic Based Collaborative watchdog approach to reduce the detection time of misbehaved nodes and increase the overall truthfulness. This methodology increases the secure efficient routing by detecting the Black Holes attacks. The simulation results proved that this method improved the energy, reduced the delay and also improved the overall performance of the detecting black hole attacks in MANET.

VII. PERFORMANCE METRICS

In this study, the following performance metrics are considered in evaluating the performance of the ad hoc network.

- **Packet Delivery Ratio (PDR)**
  Ratio of total number of data packets received at destination node to the total number of packets sent by the source node. It measures the loss rate. This metric reflects network throughput. PDR always desired to have an increasing value.
- **Throughput**
  Total number of data packets delivered at the destination node within the stipulated time. It is measured in bytes or bits per second. Throughput is

generally desired to be higher in network with fewer nodes and lower in network with more nodes.

- **End-to-end Delay**
  Total time taken for a packet to travel from source node to destination node and it is measured in seconds. With DSR protocol, the system outperforms with lowest value of delay in the adhoc network. This metric characterizes the reliability of DSR routing protocol.
- **Routing overhead**
  Ratio of routing control packets such as RREQ, RREP, and RERR transmitted over the network to the total routing and data transmissions i.e. sent or forwarded packets. Sources of routing overhead are mainly caused due to network congestion and RERR packets. It is measured in bits per second or packets per second.

VIII. CONCLUSION

Due to the mobility and open media nature, the mobile ad hoc networks are more prone to security threats compared to the wired network. Therefore, security needs are higher in MANETs compared to traditional networks. This comprehensive study provides an efficient security solution that can deal with the two types of attacks. In this paper, the study of an optimized fuzzy logic based intrusion detection system is considered to perceive the effect of optimization on strength of the system. This paper also studies the effect of packet droppers and collaborative nodes on the flow of DSR routing protocol. The performance metrics considered tend to characterize both completeness and correctness of the protocol. Henceforth this study concludes Fuzzy is also an alternative for detecting attacks over Manet.

REFERENCES

[1]. A. Chaudhary, V. N. Tiwari and A. Kumar, "Analysis of Fuzzy Logic Based Intrusion Detection Systems in Mobile Ad Hoc Networks", *BIJIT - BVICAM's International Journal of Information Technology,* Vol. **6** Issue 1, pp. 690-696, January – June, 2014.

[2]. Aaditya Jain," Performance Analysis of DSR Routing Protocol With and Without the Presence of Various Attacks in MANET", *International Journal of Engineering Research and General Science*, Vol. **4** Issue 1, pp. 454-461, January-February, 2016.

[3]. Shailesh P. Thakare and Dr.M.S.Ali, "Introducing Fuzzy Logic in Network Intrusion Detection System", *International Journal of Advanced Research in Computer Science",* Vol. **3**, Issue 3, pp. 810-815, May-June 2012.

[4]. J. Luo, and S. M. Bridges, "Mining Fuzzy Association Rules and Fuzzy Frequency Episodes for Intrusion Detection", *International Journal of Intelligent Systems,* Vol. **15** Issue 8, pp. 687-704, 2000.

[5]. S. Sujatha, P. Vivekanandan, A. Kannan, "Fuzzy logic controller based intrusion handling system for mobile ad hoc networks", *Asian Journal of Information Technology,* Vol. **7** Issue 5, pp. 175-182, 2008.

[6]. Norouzi A, Berk Ustundag B, "Improvement of DSR protocol using group broadcasting", International *Journal of Computer Science & Network Security,* Vol. **10** Issue 6, pp. 319-324, 2010.

[7]. Mr. L Raja, Capt. Dr. S Santhosh Baboo, "An Overview of MANET: Applications, Attacks and Challenges", *International Journal of Computer Science and Mobile Computing*, Vol. **3** Issue 1, pp. 408-417, January 2014.

[8]. Charles E. Perkins, Ad Hoc Networking, Addison-Wesley, March 2005.

[9]. Amandeep Kaur, Dr. Amardeep Singh, "A Review on Security Attacks in Mobile Ad-hoc Networks", *International Journal of Science and Research*, Vol. **3** Issue 5, pp. 1295-1299, May 2014.

[10]. S. Sen, J. A. Clark, and J. E. Tapiador, "Security Threats in Mobile Ad Hoc Networks", 2010.
https://www.researchgate.net/publication/275829646_Security_Threats_in_Mobile_Ad_Hoc_Networks

[11]. Kennedy Edemacu, Martin Euku & Richard Ssekibuule, "Packet Drop Attack Detection Techniques in Wireless Ad Hoc Networks: A Review", *International Journal of Network Security & Its Applications*, Vol. **6** Issue 5, pp. 75-86, September 2014.

[12]. Venkatesan Balakrishnan and Vijay Varadharajan, "Packet Drop Attack: A Serious Threat to Operational Mobile Ad Hoc Networks" Proceedings of the International Association of Science and Technology for Development (IASTED) International Conference on Networks and Communication Systems, USA, pp 89-95, 2005.

[13]. Sonali Gaikwad , Dr. D. S. Adane, "Mitigating Packet Dropping Attack in Mobile Ad Hoc Networks using 2-ACK scheme and Novel routing Algorithm", *International Journal of Engineering Research & Technology*, Vol. **2** Issue 9, pp. 2299-2304, September - 2013.

[14]. Saira Aziz, Rohit Sethi & Varun Dogra, "Packet Dropping Attack Detection Techniques In Manets: A Review", *International Journal of Computer Science and Engineering*, Vol. **5**, Issue 3, pp. 1-6, Apr - May 2016.

[15]. Cong Hoan Vu, Adeyinka Soneye, An Analysis of Collaborative Attacks on Mobile Ad hoc networks", School of Computing, Blekinge Institute of Technology, Soft Center SE – 37225 RONNEBY SWEDEN, 2009.

[16]. Mohammed Abdel-Azim, Hossam El-Din Salah, Menas Ibrahim, "Black Hole attack Detection using Fuzzy based IDS", *International Journal of Communication Networks and Information Security*, Vol. 9 Issue 2, pp. 187-195, August 2017.

[17]. Aishwarya Sagar Anand Ukey, Meenu Chawla, "Detection of Packet Dropping Attack Using Improved Acknowledgement Based Scheme in MANET", *International Journal of Computer Science,* Issues, Vol. **7** Issue 4 No. 1, pp. 12-17, July 2010.

[18]. Sonal, Kiran Narang, "Black Hole Attack Detection using Fuzzy Logic", *International Journal of Science and Research*, Vol. 2 Issue 8, pp. 222-225, August 2013.

[19]. Kulbhushan & Jagpreet Singh, "Fuzzy Logic based Intrusion Detection System against Blackhole Attack on AODV in MANET", IJCA Special Issue on "Network Security and Cryptography" pp. 28-35, 2011.

[20]. Alka Chaudhary, V. N. Tiwari, and Anil Kumar, "A New Intrusion Detection System Based on Soft Computing Techniques Using Neuro-Fuzzy Classifer for Packet Dropping Attack in MANETs, "*International Journal of Network Security,* Vol.**18**, Issue 3, pp. 514-522, May 2016.

[21]. A. Sharma, P.K. Johari, "Eliminating Collaborative Black-hole Attack by Using Fuzzy Logic in Mobile Ad-hoc Network", *International Journal of Computer Sciences and Engineering*, Vol. **5** Issue 5,pp. 34-41, May 2017.
[22]. Y. Harold Robinson, M. Rajaram, E. Golden Julie, S. Balaji, "Detection of Black Holes in MANET Using Collaborative Watchdog with Fuzzy Logic", *International Journal of Computer, Electrical, Automation, Control and Information Engineering,* Vol. **10** Issue 3, pp. 622-628, 2016.
[23]. Miss Ashwini S. Barote , Dr. P. M. Jawandhiya, "An approach for defending against collaborative attacks by malicious nodes in MANETs", *International Journal of Engineering Development and Research",* Vol. **4**, Issue 3, pp. 1010-1024, 2016.
[24]. O.O.Omitola, "Performance Evaluation of Routing Protocols in MANETs using Varying Number of Nodes and Different Metrics", *African Journal of Computing* & ICT, Vol 8. No. 2, pp. 83-90, June 2015.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 53-57(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Survey on Detection and Prediction of Dengue Fever using Data Mining Techniques

Griizma K R[1] and Dr. N. Tajunisha[2]
[1]M.Phil Research Scholar, Department of Computer science
Sri Ramakrishna College of Arts and Science for Women, Coimbatore (TN), India.
[2]Associate Professor, Department of Computer science
Sri Ramakrishna College of Arts and Science for Women Coimbatore (TN), India.

ABSTRACT: Data mining is mainly used to derive the knowledge using the data analytics techniques from an enormous amount of data. Detection and prediction are commonly performed in data mining. Data mining is applied widely in healthcare industries. The detection and prediction of the dengue. Fever is classified and predicted using data mining techniques. The classification algorithms concentrated are Naive Bayes, J48, REP Tree, ZeroR, Random Tree and SOM. The performances of Various Classification Techniques applied to the Dengue dataset are evaluated and the performances are compared to evaluate the accuracy level of the result.

Keywords: Data mining, Classification- Naïve Bayesian, J48, SOM, Dengue Dataset, Technique, Prediction.

I. INTRODUCTION

Data mining is the process of analysing and extracting knowledgeable patterns from an unknown enormous dataset. Dengue Fever is broadly categorized as Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) by World Health Organization (WHO). Dengue Hemorrhagic Fever is further classified into four groups as DHF 1, DHF 2, DHF 3 and DHF 4[1]. The illness will evidence with the indications such as joint-pain, headache and retro orbital pain Life threatening disease is caused by female Aedes mosquitoes. Dengue Fever can be detected in 2-3 days. Dengue Hemorrhagic Fever is difficult to detect even after 2-7 days having prolonged symptoms and signs. Dengue Fever and Dengue Hemorrhagic Fever cannot be differentiated at the earliest [2]. Data mining provides advantages in the medical field broadly. Data mining analysis of the medical services on Dengue can be cost effective and effective which would drastically change the death rates [3]. Dengue Fever is a rapidly spreading epidemic disease usually found in hot regions. Data mining techniques are used to categorize the dengue patients correctly, offer treatments accordingly with faster and accurate results using different algorithms [4]. Weka is used to classify data and can also predict dengue with accuracy. Bioinformatics uses Weka for diagnosing and analysing dengue datasets. Weka algorithms can generate useful predictive model for extracting knowledge from dengue dataset [5]. Knowledge Discovery in Databases including Data mining techniques is a popular research tool to exploit patterns, diagnosis and prognosis Dengue in Medical research [6].

II. DENGUE FEVER

Dengue Fever also known as Arbovirus disease transmits the infections through the nip of aedes albopictus. Arthropod-borne viral disease was first infected at Philippines in 1953. Later Dengue breaks through more than 100 nations infecting 2.5 billion individual increasing in the mortality rate. World Health Organisation distinguished Dengue fever into two types as Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF). Dengue Hemorrhagic Fever (DHF) is further classified into DHF I, DHF II, DHF III, DHF IV. Every year, it is roughly assessed 100 million instances of Dengue Fever (DF) and 250,000 instances of Dengue Hemorrhagic Fever (DHF). Dengue Fever is spread by the single stranded RNA flavivirus. Dengue Hemorrhagic Fever can be detected only if fever persists for 2-7 days with symptoms such as spillage of plasma, stun and frail heartbeat. It is difficult to differentiate Dengue Hemorrhagic Fever from Dengue Fever [3].

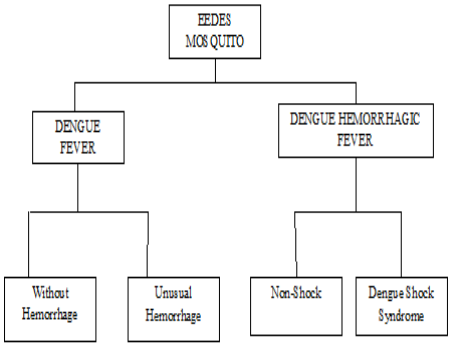


Fig. 1. Symptomatic Dengue Infection.

Table 1: Attributes.

Attribute	Values
EPID	Alpha Numeric values
Period of Fever	2/3/4/5/6/7 days
Fever temperature	100°C, 102°C, 104°C
Rashes or Red spots	Y/N
Myalgia	Y/N
Flu	Y/N
Joint/Muscle pain	Y/N
Nausea	Y/N
Low Heart Rate	Y/N
Fatigue	Y/N
Pain behind Eyes	Y/N
Head ache	Y/N
Metallic Taste	Y/N
Result	Positive/Negative

III. REVIEW OF LITERATURE

The dengue disease is analyzed, classified and predicted using various classification techniques and the performance are evaluated. In order to find the best Classification technique to classify and predict Dengue disease, the performances evaluated using different classification technique is compared.

M. Bhavani *et al*. [1] in the paper uses REP Tree, J48, SMO, ZeroR and Random Tree techniques in WEKA Data Mining Tool to classify the dengue dataset. The dengue dataset is classified with the defined symptoms such as mild bleeding, skin rash, nausea, Pain behind the eyes, severe joint and muscle pain and Fatigue. The prediction of the dengue diseases using Data Mining Tool is defined as the main objective. The dengue fever is predicted using five classification technique. The best algorithm for the prediction of dengue is obtained by comparing the classification techniques. The measure of the classification algorithm having attributes such as correctly classified, incorrectly classified, TP Rate, FP Rate, ROC Area, Precision, Recall, Accuracy and F-measure are graphically represented using Bar Charts. The objective of the paper is concluded with the prediction of Dengue fever with high accuracy and excellent performance is generated by SMO and J48 algorithms.

Kamran Shaukat *et al*. [2] in the paper, the main objective is to determine the persons infected by Dengue which is classified using Naive Bayes, J48, SMO, REP Algorithm in WEKA Data Mining Tool. Based on the Dengue Dataset, the performance of different classification techniques are compared with the help of graphs. The attributes used for testing Dengue are Fever, Bleeding, Myalgia, Flu, Fatigue and the result is class labelled as Positive and Negative. After the analysis and comparing the performance of the classification techniques, the efficiency and probability of Naive Bayes is superior rather than Random Tree and REP Tree. Thus the author concluded Naive Bayes as the best classification algorithm using Weka tool in Dengue Fever Prediction.

Vandana Rajput *et al*. [3] propose a review paper on Dengue Disease forecasting using Association rule mining. Distinctive Association algorithms such as Apriori and FP-Growth are used for the prediction of Dengue fever. The paper briefly elucidates on the characterization of the dengue fever, viruses and its side effects. The symptoms of Dengue fever included bleeding, low levels of blood platelets, low circulatory strain and metallic taste in mouth, headache, muscle joint torment and rashes. It also portrays entire information on the medicinal data mining. This paper provides an overview of the literature survey on the association rule mining which is used to find frequent patterns for the prediction of Dengue Fever.

Dave Kaveri Athulbhai *et al.* [4] in the paper conducts Prediction of dengue disease using classification techniques in data mining. The GEO dataset NCBI is used for dengue dataset. The survey paper is generic and represents the comparative analysis of data mining techniques for the classification and prediction of the Dengue disease. The performance analysis of Dengue dataset using decision Tree, Naive Bayes and Neural network is generated for 859, 1060 and 1800 records respectively.

Kashish Ara Shakil, *et al*. [5] in the paper classifies the dengue dataset using different data mining techniques. The Classified dataset is compared using three interfaces such as Explorer, Experimenter and Knowledge flow Interfaces in Weka tool. Dengue dataset is used in the prediction of Dengue has 108 instances. The dataset is classified with the attributes including P.I.D, date of dengue fever, days, current temperature, WBC, Joint muscles, metallic taste in mouth, appetite, abdomen pain, Nausea and heamoglobin. The dataset file format is CSV. The analysis of performance is carried out by Naive bayes, J48, SMO, REP tree and Random tree algorithms using Explorer, Experimenter and Knowledge flow Interface of Weka tool. The main objective of the paper is to predict dengue infection and also to find the efficient performing algorithm. The best accuracy and performance is achieved by Naive Bayes and J48 algorithm through Explorer and knowledge flow Interfaces.

A.Shameem Fathima, *et al.* [6] elucidates the survey on the data mining classification methods which helps in diagnosis and prognosis of the Dengue infection. The survey is proposed in order to find whether it is possible to analyse and predict dengue infections using data mining algorithms in R Tool. The different algorithms generated will be evaluated and compared. It provides a comprehensive view on applying machine learning in medical field.

Mantena Krishnar Satya Varma, *et al*. [7] has major objective to create a predictive model of Dengue disease using decision tree. Decision tree in data mining has its advantage which helps in discovering rules in

mining applications. The decision tree uses top-down approach to classify the dengue dataset.

IV. CLASSIFICATION

The classification technique in data mining on Weka tool classifies the dengue fever as Dengue Fever (DF) and Dengue Hemorrhagic Fever (DHF) by observing the characteristics of dengue. The performance of the classification technique using different algorithm can be evaluated using the statistical results obtained. The classification technique performed on different algorithms like Naïve Bayes, J48, SMO, REP Tree and Random Tree in order to find the best performing classification algorithm providing accurate results for detection and prediction. The accuracy of the classification is given by the following [5].

A. Correctly Classified Accuracy
The correctly classified percentage of accuracy test is intended.

B. Incorrectly Classified Accuracy
The incorrectly classified percentage of accuracy test is calculated.

C. Mean Absolute Error
The classification accuracy of algorithm is analyzed by the number of errors.

D. Time
The time required to build model in order to predict disease.

E. ROC Area
Receiver Operating Characteristic represent the classification for test performance guide of diagnostic test based as excellent (0.90-1), good (0.80-0.90), fair (0.60-0.70), poor (0.60-0.70), fail (0.50 – 0.60).

V. DATA MINING TECHNIQUES

A. Naïve bayes technique
Naive Bayes is based on Bayes formula which performs arithmetical predictions. Naive Bayes classifier provides demonstrable performance over ID3 and neural system classifiers. The outcomes of the Naive Bayes technique for given dengue dataset are noted to evaluate their performance [2].

B. REP tree
Several trees are created and reiterated different times in Rep tree. It uses regression tree reason. After different reiterations, Rep tree picks the best tree. The outcomes of the Rep tree technique for given dengue dataset are noted to evaluate their performance [1].

C. J48
C4.5 is the used to build decision tree. C4.5 is a technique used to classify the data sets with some given decisions. The outcomes of the J48 technique for given dengue dataset are noted to evaluate their performance [1].

D. SMO
Sequential Minimal Optimization is also known as Platt's SMO algorithm. It is a well organized and provides a good computational efficiency .It is a technique used for answering problem raised by the SVM during the dataset training. The outcomes of the SMO technique for given dengue dataset are noted to evaluate their performance [1].

E. RT
Random Tree is the supervised Classifier. Random tree choose random attributes at each node to classify the dengue dataset. The outcomes of the Random Tree technique for given dengue dataset are noted to evaluate their performance [2].

F. ZeroR
ZeroR is the simplest of all classification methods. It predicts only the category class and all the predictors are ignored. The outcomes of the ZeroR technique for given dengue dataset are noted to evaluate their performance [1].

VI. ANALYSIS

Title	Author
"A Data mining approach for precise diagnosis of Dengue Fever" [1]	M. Bhavani and S. Vinod Kumar
"Dengue Fever Prediction : A Data Mining Problem" [2]	Kamran Shaukat, Nayyer Masood, Sundas Mahreen and Ulya Azmeen
"A review paper on Dengue Disease forecasting using Data Mining Techniques" [3]	Vandana Rajput Prof.Amit Manjhvar
"A Survey : Prediction and Detection of Dengue – Mining methods and techniques" [4].	Dave Kaveri Atulbhai and Shilpa Serasiya
"Dengue Disease Prediction Using WEKA Data Mining Tool " [5].	Kashish Ara Shakil, Shadma Anis and Mansaf Alam
" A Review of Data Mining Classification Techniques Applied for Diagnosis and Prognosis of the Arbovirus-Dengue" [6].	A.Shameem Fathima, D.Manimegalai and Nisar Hundewale
"Dengue Data Analysis Using Decision Tree Model" [7].	Mantena Krishna Satya Varma and Konadala Kameswara Rao Nynalasetti

Dataset	Methodology	Accuracy
Dengue Dataset [1]	SMO J48 REP Tree ZeroR Random Tree	84% 76% 72% 72% 68%
Dengue Dataset collected from DHQ, Jhelum [2]	Bayesian REP Tree Random Tree J48 SMO	92.% 76% 76% 76% 76%
Dengue Dataset [3]	Apriori Algorithm FP-Growth Algorithm	-
Geo Dataset NCBI [4]	Decision Tree Naive Bayes Neural Network	96% 94% 92%
Dengue Dataset [5]	**Explorer Interface** Naïve Bayes J48 SMO REP Tree Random Tree **Experimenter Interface** Naïve Bayes J48 SMO REP Tree Random Tree **Knowledge Flow Interface** Naïve Bayes J48 SMO REP Tree Random Tree	 99% 99% 99% 74% 87% 99% 94% - - - 99% 99% 99% 74% 87%
Clinical Dengue Dataset [6]	Decision Tree, SVM, Evolutionary programming, Fuzzy sets, Neural Networks, Rough Sets	-
Raw Data collected from health department, hospital and urban local body [7].	Decision Tree Model	90%

VII. CONCLUSION

Data Mining is widely used for the data analysis of healthcare industries. Now days, Data Mining is used in prediction of Dengue Fever. Using different algorithms and techniques in data mining, the efficient results of the data can be obtained .In the review, we discussed on performance analysis of different classification algorithms in the prediction of the Dengue disease. It is concluded that high accuracy on predicting Dengue Fever is generated by Naïve Bayes, SMO and j48 algorithms.

REFERENCES

[1]. M. Bhavani, S. Vinod Kumar "A Data Mining Approach for Precise Diagnosis of Dengue Fever" *International Journal of latest trends in Engineering and Technology,* Vol. **7**, Issue 4, pp. 352-359.

[2]. Kamran Shaukat, Nayyer Masood, Sundas Mehreen, Ulya Azmeen, (2015). "Dengue Fever Prediction: A Data Mining Problem" *Data mining in Genomics and proteomics*, Vol. **6** Issue 3.

[3]. Vandana Rajput, Prof. Amit Manjhvar, April (2017). "A Review Paper on Dengue Disease Forecasting Using Data Mining Techniques" in *IJSART*, Vol. **3** Issue 4.
[4]. Dave Kaveri Atulbhai, Shilpa Serasiya, (2017). "A Survey: Prediction and Detection of Dengue-Mining Methods and Techniques" in *IJARIIE*-ISSN (O)-2395-4396, Vol. **3,** pp. 48-52.
[5]. Kashish Ara Shakil, Shadma Anis, Mansaf Alam "Dengue Disease Prediction Using Weka Data Mining Tool.
[6]. A. Shameem Fathima, D. Manimegalai, Nisar Hundewale, November (2011). "A Review of data Mining Classification Techniques Applied for Diagnosis and prognosis of the Arbovirus-Dengue " *IJCSI International journal of computer science* pp. 322-328.
[7]. M. Krishna Satya Varma, N.K. Kameswara Rao, (2015). "Dengue Data Analysis using Decision Tree Model" *International Conference on Emerging Trends in Science Technology Engineering and Management*.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 58-61(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Fabric defect detection techniques: A Review

Soumya Haridas[1] and Prof. S.N. Geethalakshmi[2]
[1]Research Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher education for Women, Coimbatore (TN), India.
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher education for Women, Coimbatore (TN), India.

ABSTRACT: Textiles are an inevitable part of human life. Defect detection in fabrics plays a vital role in the textile manufacturing industry. Defective fabrics can cause decrease in cloth exporting, due to low quality. Fabric defects can be identified either through manual inspection or through the automated inspection process. Manual inspection for defect detection is a time consuming task. We can overcome the problems associated with manual fabric defect detection by using automated fabric defect detection techniques. In this paper, various methods of automatic fabric defect detection techniques are reviewed.

Keywords: Fabrics defect, defect detection techniques.

I. INTRODUCTION

A textile means woven fabric. Textile fabrics can be produced from fibers by using either weaving, knitting or sewing. Woven fabric uses two sets of yarns by interlacing one among the other. Out of these two, one is warp yarn and the other is weft yarn. These can be produced from hand loom or power loom. Shirts, Trousers, denims, and Jeans are examples for woven fabrics. Knitted fabric uses one set of yarn by interloping. Circular or flat knitting machine is used for knitting. T-shirts, inner wears and leggings are examples of knitted fabrics. Sewing is the process of combining or joining fabrics with the help of machines fitted with needles. Quality control is a very big concern in the textile manufacturing industry. Well examined quality clothes are in great demand in the market. Defects in clothes can reduce the goodwill of any company and thus can cause big reduction in the production or turnover in business. It is therefore important to find out the defects in fabrics as early as possible. Defects can be occurred in clothes during any time of manufacturing such as weaving, dyeing, stitching etc. There can be many defects in fabrics including yarn defects, weaving defects, printing defects, embroidered defects. Fabric defects can be caused by machine malfunction, faulty yarns or machine spoils [15]. There are mainly eight types of fabric faults [5], [17]. They are as follows: (Fig.1) Float, Weft curling, Slub, Hole, Stitching, Stains, Broken ends, Miss pick

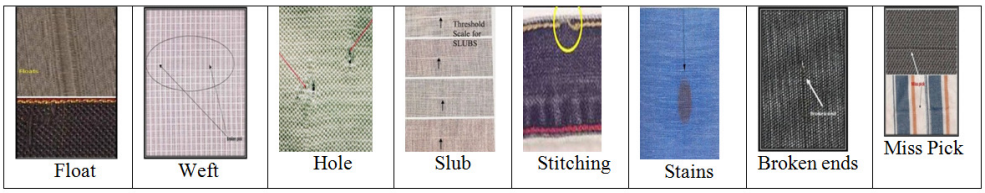


Fig. 1. Sample photos of fabric faults float, weft, hole, slub, stitching, stain.

II. FABRIC FAULT DETECTION

Fig. 2 shows the steps in fabric fault detection. Fabric fault detection is the process of finding out the defects on fabric surfaces. Using automation defects can be identified and the production of faulty fabrics can be controlled. If faults are identified at an early stage, the quality of manufactured clothes can be improved to a large extent. All these defects can be detected either manually or automatically. Manual defect detection is a time consuming task. In manual process an expert in defect detection stands in front of inspection table and pulls out woven fabric onto the table. He checks the cloth for any kind of defects and the defective parts are marked accordingly. This is done after cloth manufacturing and the defective parts can be avoided by removing the defective portion.

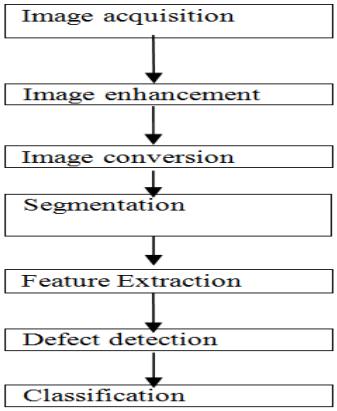


Fig. 2. Fabric fault detection process.

It is also time consuming and can cause manual errors due to accidental ignorance, unavailability of expertise, tiredness etc. Automatic defect detection can overcome all the disadvantages of manual inspection. In automatic inspection a camera is used for taking pictures of woven fabrics during manufacturing. Video camera can also be used to take images of the fabrics while moving. After collecting the images, image enhancement is carried out by removing the noises with filtering. Edge detection is performed using canny edge detection, fuzzy algorithms etc. The image is converted into binary image and segmentation is performed for feature extraction. Using any of the classification algorithms the defects can be classified by comparing it with the images in the database. The database already trained, contains defective as well as non-defective images. The process can be carried out during manufacturing and the production of defective clothes can be controlled suddenly. Since the inspection is controlled by the machines, labor cost as well as time can be reduced. Accurate results can also be produced. Table 1 shows the various fabric faults and their reasons.

Table 1: Fabric faults and reasons.

Fault	Type	Reason
Float	Weaving	Breaking of needles
Slub	Weaving	Caused by inserting a highly twisted weft thread
Weft curling	Weaving	Caused by thick places in the yarn or by fly waste being spun in yarn during the spinning process
Hole	Knitting	Mechanical fault caused by a broken machine part
Stitching	Stitching	Caused by undesired motion of loom or stitching machines
Stains	Weaving	Caused by lubricants or rust
Broken ends	Weaving	Caused by broken warp yarn.
Miss pick	Machine	Caused by restarting the stopped machine without removing the fabric.

III. REVIEW

According to Kazim Hanbay *et.al* [5] fabric faults detection approach can be classified as structural, statistical, spectral, model-based, learning based and hybrid approaches. Structural approach can be used for detecting defects from texture analysis. The statistical approach uses first order and second order statistics to extract textural features. First order statistics calculate the characteristics of each pixel values, and the interaction between pixels is not considered. Second order statistics calculates the characteristics of more pixel values which are at some specific locations. In spectral approach both spatial and frequency domain information are necessary for fabric defect detection. Spatial domain is required for locating the defect and frequency domain is required for finding the defect. In model-based approach an image model is created. The texture is identified and texture synthesis is carried out with this model. Learning based approach uses Artificial Neural Networks for image classification. Hybrid approach detects fabric defects using a combination of two or more techniques. According to Swapnil *et.al.* [9] fabric inspection can be classified as visual fabric inspection and automatic fabric inspection. In visual fabric inspection the woven fabric is rolled over inspection table and checked manually for any kind of defects. This is done by well-trained human inspectors. In automatic fabric inspection the defect detection process is much faster which causes more productivity with high quality and lowered labor cost. According to Prof. K. Y. Kumbhar *et.al.* [15] fabric defects can be caused by machine malfunction, faulty yarns or machine spoils. The defects caused by machine malfunction include broken yarn or missing yarn which can be too long or too narrow. Slubs caused by faulty yarns are pointed, but oil spots seems to be wide and irregular.

Image classification techniques such as neural network, SVM, clustering and statistical inference are considered as the most useful ones. After acquiring the images they are classified using SVM classifier considering 3 features color, oil spot and threading. The system is trained with sample image database with defective and non-defective images. When high quality images are given as input to the system they are classified based on defects as defective or non-defective. Gagandeep *et.al.* [4] used Artificial Neural Network algorithm for defect detection. Both normal and defective images are selected and classified using neural network. The

system is properly trained with faulty images with defects such as hole, oil and weft defects. After extracting features of all the samples the images are classified accordingly. Only three types of defects can only be identified using this method. According to Farida *et. al.* [13] they considered histogram based fabric fault detection. Canny edge detection was used to detect wide range of edges in images. They consider it as less time consuming and gives more accuracy. According to them main limitation is, it can only be implemented in industries. According to Harinath *et.al.* [6] they considered wavelet transform as the most suitable algorithm for quality inspection and extraction of fabric features. All the coefficients of a complete fabric were extracted using wavelet transform and coefficient subset was obtained. Using genetic algorithm suitable subsets were selected. The subsets were compared with other images and defects in the images were detected using genetic algorithm. Minal Patil *et.al.* [16] in their paper referred about 56 papers and reviewed the methods for automatic defect detection. According to them it is important to find out the defective material and also area of defect. They considered algorithms such as Auto-correlation function, Eigen filters, Histogram, Co-occurance matrix etc for defect detection. It will be better to combine some of these methods to get more accurate results. Naik Ganapati *et al*. [9] proposed a method of defect detection using golden image subtraction. The original and defective images were decomposed using wavelet decomposition. The defective image component was subtracted from non- defective image component. Defect detection was carried out using thresholding and filtering. Bangare S.L. *et al*. [1] used the technique of segmentation to detect the fault on fabrics. Video images of the fabric moving through a conveyor belt are taken and the fault detected was signaled with the help of an alarm. Image noise was removed using Gaussian blur and it was segmented and by thresholding the defect was detected. Yapi Daniel *et. al.* [2] used machine learning approach to detect the defects on fabrics. They proposed an algorithm to detect faults which consist of two phases. In the first phase the features of defective and defect-free images were obtained and classified using Bayer's classifier. The defects were detected in the second phase using the trained image set. Mahure J. *et.al*. [7] converted the acquired images into gray scale and noise removal was carried out. After filtering the output was obtained as a histogram which was used for reaching at a conclusion about the defect type and thus classifies the images. A. Shams Nateri *et. al*. [19] in their paper considered images taken by scanner. For image modification they used Otsu's method. The defects are identified considering the size and shape. Conversion of the image to binary is carried out and noise removal is performed. The no noise images are then subtracted from the ideal images in the database and the defective pixel values are obtained. The defective area is located by taking into account the no. of pixels and the area where the defect is located. According to them the quality of image is a big concern in detecting the defects. Considering various methods for defect detection and classification Articial Neural Network shows 93.3% accuracy [4], filtering based approach gives 96.7% accuracy [8], and histogram based approach gives 85% accuracy [11].

IV. CONCLUSION

This paper reviews various techniques for automatic defect detection. There are different methods which come under these categories; Statistical, structural, model-based and hybrid. Many classification algorithms are available such as Artificial Neural Network, Bayer's classification, Histogram based, wavelet based etc. Instead of using a single method a combination of various methods can be used to get better results. In the future these methods may be used for finding out defects in finished goods from the apparel industry.

REFERENCES

[1]. S.L. Bangare, N.B. Dhawas, V.S. Taware, S.K. Dighe, P.S. Bagmare. "Fabric Fault Detection using Image Processing Method." *International Journal of Advanced Research in Computer and Communication Engineering*, vol**. 6**, issue 4, pp. 405-409, Apr. 2017.

[2]. Y. Daniel, M. Marouene, S.A. Mohand, B. Nadia B. "A learning-Based Approach for Automatic Defect Detection in Textile Images." *International Federation of Automatic Control*, no.3, pp. 2423-2428, 2015.

[3]. R. C. Gonzalez, R. E. Woods, S.L. Eddins. *Digital Image Processing Using MATLAB.* New Delhi: McGraw Hill, 2016.

[4]. S. Gagandeep, S. Gurpadam , K. Mandeep. "Performance Evaluation of Fabric Defect Detection using Series of Image Processing Algorithm & Ann Operation." *International Journal of Recent Trends in Engineering & Research*, vol. **02**, issue 05,pp. 1-7, May 2016.

[5]. K. Hanbay, M. Talu, Ö.Özgüven. "Fabric defect detection systems and methods — A systematic literature review." *Optik - International Journal for Light and Electron Optics*, vol. **127**, no.24, pp.11960-11973, Sep. 2016.

[6]. D. Harinath, B. K. Ramesh, P. Satyanarayana, M.V.R. Murthy. "Defect Detection in Fabric using Wavelet Transform and Genetic Algorithm." *Transactions on Machine Learning and Artificial Intelligence*, vol.3, issue. 6, pp. 10-18, Nov.2015.

[7]. M. Jagruthi, Y.C. Kulkarni. "Fabric Faults Processing: Perfections and Imperfections." *International Journal of Computer Networking, Wireless and Mobile Communications*, vol.4, issue 2, pp. 101-106, Apr. 2014.

[8]. S. Jayaraman, S. Esakkirajan, T. Veerakumar. *Digital Image Processing.* New Delhi: Tata McGraw Hill, 2009.

[9]. R.S. Kurkute., S.P. Sonar, A.S. Shevgekar, B.D. Gosavi. "DIP Based Automatic Fabric Fault Detection." *International Research Journal of Engineering and Technology*, vol. 04, issue 04, pp. 3356-3360, April 2017.

[10]. Kumar. A. "Computer vision based fabric defect detection techniques: A Survey.", pp. 1-28, 2004.

[11]. B. Karunamoorthy, D. Somasundari, S.P. Sethu. "Automated Patterned Fabric Fault Detection using Image Processing Technique in MATLAB." *International Journal of Advanced Research in Computer Engineering & Technology*, vol.4, issue 1, pp. 63-69, Jan. 2015.

[12]. P.M. Mahajan, S.R. Kolhe, P.M. Patil. "A review of automatic fabric defect detection techniques." *Adv. Comput. Res.,* vol.**1**, issue 2, pp. 18-29, 2009.
[13]. F.S. Nadaf, N.P. Kamble, R.B. Gadekar. "Fabric Fault Detection Using Digital Image Processing." *International Journal on Recent and innovation trends in computing and communication*, vol. **5**, issue 2, pp. 128-130, Feb. 2017.
[14]. S.G. Naik, M.S. Biradar, K.B. Bhangale. "Patterned fabric defect detection using wavelet golden image subtraction method." *International Journal of Advance Research, Ideas and Innovations in Technology*, vol.**3**, issue 3,pp. 767-771, 2017.
[15]. P.Y. Kumbhar, T. Mathpati, R. Kamaraddi, N. Kshirasagar. "Textile Fabric Defects Detection and Sorting UsingImage Processing." *International Journal for Research in Emerging Science and Technology*, vol.3, issue 3, pp. 19-24, March 2016.
[16]. M. Patil, V. Sarita, J. Wakode. "A review on Fabric Defect Detection Techniques." *International Research Journal of Engineering and Technology*, vol. 04, issue 09, pp. 131-136, September 2017.
[17]. R. Singh, "Common fabric defects", Internet: https://textilelearner.blogspot.in/2013/07/common-fabric-defects-with-images.html.
[18]. M. Islam "20 woven fabric defects with pictures", Internet: http://www.garmentsmerchandising.com/20-woven-fabric-defects-with-pictures/, Mar. 26, 2016.
[19]. A. S. Nateri, F Ebrahimi, N. Sadeghzade. " Evaluation of yarn defects by image processing technique." *Optik - International Journal for Light and Electron Optics*, vol. **125**, pp.5998- 6002, Jun. 2014.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 62-67(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Applying Machine Learning Techniques in Agricultureto Forecast Crop Yield – A Survey

M.C.S. Geetha[1] and Dr. I. Elizabeth Shanthi[2]
[1]Professor in Computer Science, Kumaraguru College of Technology, Coimbatore (TN), India
[2]Professor in Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India

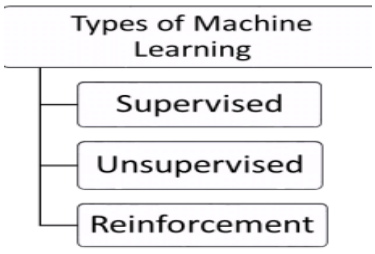
ABSTRACT: In India, agriculture plays a vital role for generating income. By combining all the factors such as biological and seasonal controls the crop production, but random changes causes a great loss to the farmers. By applying appropriate mathematical or statistical methodologies in the soil and weather data, these risks can be computed. Machine learning helps to forecast the crop yield by obtaining the valuable information from the agricultural data to make a decision on the crop for farmers, so that they can plant for the future which leads to huge profit. In this paper, wepresent a detailed study on the various machine learning algorithms that facilitate to forecast the crop yield.

Keywords: Machine Learning, Data Mining, Crop yield, Forecasting, Agriculture

I. INTRODUCTION

Agriculture forms the important source for food security. It assists human beings to grow the best food crops. Rice and wheat is the primary food in India. Indian farmers grows the following foods such as rubber, cotton, potatoes, pulses sugarcane, oilseeds. Agriculture is depended by 70 per cent of the rural family. Total GDP of 17% is contributed in agriculture. 60% employment is provided over the population. A machine learning technique helps to make decisions automatically by detecting pattern from the past data and generalizing it on the future data. 25% of the jobs can be substituted by the use ofmachine learning algorithms in the next forthcoming years.

The following are the 3 major categories of Machine Learning algorithms.



In Supervised Learning, we have input and output variables and the algorithm createa function that calculates the output based on given input variables.

Regression and Classification are the two parts: Some examples include Linear Regression, Decision Trees, Random Forest, *k* nearest neighbours, SVM, Gradient Boosting Machines (GBM), Neural Network etc.

In Unsupervised learning, only input data is present and there is no corresponding output variable. It can also be classified into two groups, namely Cluster analysis and Association. Some examples would be k-means clustering, hierarchical clustering, PCA, Apriori algorithm, etc.

In Reinforcement learning, the machine is given training to make accurate decisions from these actions and tries to capture the best possible knowledge. Some examples are Weather forecast, Speech Recognition, Game playing, face detection/Face recognition, Genetics and agriculture.

The rest of the paper is organized as follows: Chapter II explains the methods of Machine Learning. Chapter III describes about the applications of Machine Learning used in agriculture domain. Chapter IV analyses the outcomes. Chapter V discusses the conclusion.

II. MACHINE LEARNING METHODS

A. Naïve Bayes Classifier Algorithm

Naïve Bayes Classifier algorithmic rule candowellonce the input variables are categorical.

It unites preferably, demanding comparatively tiny coaching information data than alternative discriminative illustration such as logistic regression, once the Naïve Bayes restricted liberation assumption holds. It's terribly simple to estimate the category of the check information set by victimization this algorithmic rule. It is employed in the applications like Sentiment Analysis, Document Categorization, Sports, Politics and Email Spam Filtering.

B. K- Means Clustering Algorithm

K-means is an unsupervised machine learning rule for cluster analysis. K-Means could be non-deterministic and unvaried methodology. This rule works on a specified knowledge set through predefined range of clusters, k. The output of this rule suggests that there are k clusters with input file divided among the clusters. In case of globular clusters, K-Means yieldsconstricted clusters than gradable agglomeration. When we provide a lesser value of K, K-Means agglomeration calculates quicker than gradable agglomeration for big range of values.

C. Support Vector Machine

Support Vector Machine is the method used for grouping or regression issues. It supports for categorizing information into several types by ruling a hyper plane which divides the coaching data set into groups. As there are numerous such linear hyper planes, this procedure goals is to exploit the space between the various categories that are concerned and this is often called margin maximization. When categorizing the category to exploit the space between the categories, the link is provided to support the likelihood. It is classified into two categories: Linear and Non-Linear SVM.

D. Apriori Algorithm

This algorithmic program is called as unsupervised machine learning algorithm as it produces relationship procedures from a provided knowledge set. Relationship procedures implies that if secondary item A happens, then item B conjointly happens with a precise likelihood. Best of the relationship procedures created within the IF_THEN format. It will be utilized in the applications likeDetection Adverse Drug Reactions**,** Market Basket Analysis, etc. The fundamental principle bythat the algorithmic program works is as follows. If associate item set happens often then all the subgroups of the item set conjointly occur often. If associate item set happens occasionally occurs rarely then all the supersets of the item set have rareincidence.

E. Linear Regression Algorithm

This algorithmic programmed is plays the link between 2 variables and its ever-changing have an effecton the opposite. This algorithmic program describes the effect on the variable quantity on ever-changing the experimental variable. The freelance variables referred as informative variables, as they justify the factors that impact the variable quantity. Variable quantity is commonly cited as cause of interest or analyst. It will be utilized in the applications like estimating sales and helps to assess risk concerned in insurance or monetary domain.

F. Logistic Regression

This regression technique applies a logistical task to a linear mixture of choices to calculate the result of a categorical variable which supports predictor variables. The chances that define the result of a one trial shapely performs that of instructive variables. It helps to evaluate the likelyhood of decreasing into a selective level of the explicit variable which supports the provided predictor variables. It's classified into 3 sorts**–** Binary Logistical Regression, Multi-nominal Logistical Regression and Ordinal Logistical Regression**.** These procedures doesn'taccept a linear association between the dependent and freelance variables and so also can make of non-linear effects.

G. Decision Tree Algorithm

In order to make a decisionby creating a sure conditions, splitting practice is done. In a decision tree, the internal node signifies testing on the quality, each division of the tree signifies the result of the check and also the leaf node signifies a exact group tag i.e. the choice created once computing all of the characteristics. The sorting procedures pictured through the trial from root to the leaf node. These form 2 types of trees like **c**lassification and regression Trees. These procedures helps to measure information examination. Decision trees indirectly accomplish feature choice that is extremely necessary in prophetic analytics. When a decision tree is acceptable a coaching dataset, the nodes at the highest on that the decision tree is divided, thought of as necessary variables inside a specified dataset and have choice completed by default.

Decision trees supports to avoid wasting the time for getting ready information, since they are not tuned intolost values and outliers. Missing values won't stop you from rending the information for constructing a decision tree.

H. Random Forest Machine Learning Algorithm

Random Forest uses a capturingmethod to make a cluster of call trees with arbitraryset of the information. This rule achieves sensible prediction performance by coaching a modelmany times on random sample of the dataset. During this collaborative learning methodology, the yield of all the choice trees within the random forest, is shared to make the final calculation. The estimation is final once it is resulted by voting the end result of everycall tree or simply by creating a calculation that seems the foremost times within the decision trees.

It is tough to create a nasty random forest. By employing the Random Forest Machine Learning algorithms, it's easy to determine the factors to practice as they are not complex to the limitations that are accustomed to run the rule. One can construct an honest model simply without a lot of standardization. Random Forest algorithms are employed in agriculture to predict the soil.

I. Artificial Neural Networks

An Artificial Neural Networks is nothing but a group of connected units which is known as artificial neurons. Each connection unit which is called as synapse between neurons can transmit a signal to another neuron. The receiving unit which is called as postsynaptic neuron can process the signal(s) and then signal down stream neurons connected to it. Neurons may have a state, generally represented by real numbers, typically between 0 and 1.

Neural networks can be been used on a varioustasks such as speech recognition, computer vision, social network filtering, machine translation, medical diagnosis, playing board and video games and in many other domains.

J. K-Nearest Neighbour

It is a non-parametric technique helps for both regression and classification. Both jointly provide the input that consists of the *k* closest training example in the feature space. The output are depending on *k*-NN which is used for classification or regression:

The output may be a class membership in the *k-NN classification technique*. An object is divided by a better partof vote by its neighbors, in which the objects areallocated to the class well-known among its *k* nearest neighbors (*k* is a positive integer, typically small).

When the value of *k* is 1, the object is being assigned to the single nearest neighbor of the class. In the *k-NN regression technique,* the output is the property value for the object. The value is used by the average of the values from its *k* nearest neighbors.

III. APPLICATION OF MACHINE LEARNING TECHNIQUES IN AGRICULTURE

This work discusses the various techniques of Data Mining for the purpose of mining the text and the Dashboards. The tools outline helps to deploy the different situation [1]. The data has been decided with the help of Knowledge Discovery Process and also from the different Data Mining Techniques such as Classification, Association, Clustering and Regression [2].

This paper collects data from the different sources such as the Global Information System, Remote sensing and Global Positioning System. Map Reduce and linear regression algorithm is useful for weather forecasting. Developing the correctness of rainfall prediction is the main reason for this model [3]. This paper presents an application which helps the farmers to cultivate based on the climatic conditions. For building this application, the main aim is to use the open source tools. All the information is made by clicking once by selecting location from the map [4].

Our work explains about the yield prediction by relative study of the regression models.There are different algorithms which are being discussed such as Regression, Support Vector Machine and Multilayer perception Model [5]. This paper explains the model which is best suited to forecast the yield. This papergoal is to forecast the crop yield with the advent of the soil attributes such as Sulphur, Zinc, potassium, etc. Soil classification is made by the Naive Bayes algorithm.The yield is increased by relating the soil with the crops by using the Apriori algorithm. An accurate relationship is attained by using the Naïve Bayes and other algorithm [6].

The method that is used for classifying is explained to forecast the crop yield. Some of the techniques that are used in machine learning are support vector machines, regression, neural networks and random forests. With the help of climatic data and the attributes of crop, the development of the crop is analyzed [7].

In this paper, finding the knowledge by using machine learning is done on the real time data set. With the support of crop attributes, the farmers are grouped together by means of K-means clustering algorithm. The selection of 2 crops as a frequent item set is done by means of Apriori algorithm. The main aim is on the policies that the government practices on the crop of the cultivator [8]. This work explains about the crop yield for the situation of the various crops and thereby increasing the quality with the help of weather related data sets. [9] Precision farming aims at constructing the decision support model for the analyzing purpose and thus forecasting the future.

The crop yield is controlled by the diverseattributes of the environment such as developing area, Yearly rain and index of the price for food. Machine learning techniques like regression is used for the prediction of the yield of the crop [10].

This work describes the machine learning methods that are used for analyzing in the agricultural field. Some of the Machine learning methods are discussed like the classification, association, clustering and regression algorithms [11].

Explains the different machine learning techniques for developing the crop. Some of the classification techniques like regression, neural network, K-means algorithm and other algorithm are used [12].

This paper handles vast data by using the Intenet of Things. With the advent of sensors, data are collected and passed to the particular area.Pictures related to agriculture are taken with the help of global positioning system to exactly provide the data along with their current location [13].

This paper presents a relative survey on the clustering techniques on the bases of different algorithm. Regression techniques are used for forecasting the crop [14].

Table 1: Machine learning techniques used in agriculture.

Author and publication	Techniques used	Output	Limitations
Bendre *et al* , 2015 [3]	Predicting weather conditions is done by Map Reduce and Linear Regression algorithm.	The method to improve the correctness of rainfall forecasting .	Only weather data is considered.
Grajales *et al,* 2015 [4]	Open source tools are used for the structuring of web application.	All the information is made by clicking once by selecting location from the map	Nil
Rub , 2009 [5]	Regression, Support Vector Machine and Multilayer perception Modelare used.	Explained relative survey of different algorithms	Nil
Hemageethaa, 2016 [6]	Classification techniques used for forecasting yield.	Forecast the crop yield with the advent of the soil attributes such as Sulphur, Zinc, potassium, etc	Precision is low.
Kushwala , 2015 [9]	Scala is used.	Constructing the decision support model for the analyzing purpose and thus forecasting the future.	Do not explain about the crop yield for the various crops
Sujatha , 2016 [7]	Implementing support vector machines, regression, neural networksand random forests	With the help of climatic data and the attributes of crop, the development of the crop is analyzed	Soilattributes are not measured.
Veenadhari , 2014[11]	Used the various techniques such as classification, association, clustering and regression algorithms	Machine learning methods that areused for analyzing in the agricultural field.	

Author and publication	Techniques used	Output	Limitations
Veenadhari , 2014[11]	Used the various techniques such as classification, association, clustering and regression algorithms	Machine learning methods that areused for analyzing in the agricultural field.	Nil
Fathima , 2014 [8]	k means clustering algorithm are explained.	Crop attributes are taken.	Main aim is on the policies that the government practices on the crop of the cultivator.
Raorane , 2012 [12]	Classification techniques like regression, neural network, K-means algorithm and other algorithm are used.	The techniques used for crop production is discussed.	Nil
Sellam , 2016 [10]	Regression Techniques are used.	Crop yield is controlled by the diverse attributes of the environment such as developing area, Yearly rain and index of the price for food.	Nil
Ankalaki , 2016 [14]	Multiple Linear Regression are used.	Relative survey on the clustering techniques	Nil

Agriculture is widely used segment of Tamil Nadu. This paper confers about the Clustering, Classification, Association rule mining, and regression. It also talk about the application of the data mining techniques in agriculture such as Decision tree, K-means, Fuzzy set, Bayesion classification, naïve bayes, K nearest neighbor, neural networks and support vector machine [15].

IV. INFERENCES

In this survey, the machine learning techniques in agriculture such as weather forecasts, prediction of rainfall, classifying soil, yield prediction, soil moisture prediction methods and crop growth monitoring techniques are studied and compared for analyzing the performance.

From this study, it is clear that the regression technique can be used for forecasting the crop yield and rainfall. Naïve bayes can be used for soil profiling. K-means can be used to find out the yield. My future work will be to find out the optimization techniques for monitoring the soil conditions, weather forecasts and crop yield.

V. CONCLUSION

There are varied systems that uses numerous data processing technologies to control knowledge to derive intuitions and facilitates to decide the yield of the farmers. However the foremost downside is that they focus either on one crop and forecast any one parameter like harvest or amount. This research helps to predict the harvest and amount of main crops of Tamil Nadu based on past data. The information and forecasted result are manageable for the farmers through an internet application. This aids agriculturalist to make a decision on which crop they like to harvest for the forthcoming year. Additionally the internet application conjointly provides an opportunity to help the farmer's products without middlemen that facilitates them to get highest price for their crops. The discussion opportunity allows users to explain their doubts relating to the practice of the internet application.

REFERENCES

[1]. Agrawal, H., Agrawal, P., "Review on Data Mining Tools", *International Journal of Innovative Science, Engineering & Technology,* Vol. **1**, Issue 2, pp.52-56, 2014.

[2]. Kalyani, M.R., "Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, Vol. **2**, Issue 10, pp.439-442, 2012.
[3]. Bendre, M. R., Thool, R.C., Thool, V. R., "Big Data in Precision Agriculture: Weather Forecasting for Future Farming", *1st International Conference on Next Generation Computing Technologies*, pp.744-750, 2015.
[4]. Grajales, D.F.P., Mosquera, G.J.A, Mejia, F., Piedrahita, L.C., Basurto, C., "Crop-Planning, Making Smarter Agriculture With Climate Data", *Fourth International Conference on Agro-GeoInformatics,* pp.240-244, 2015.
[5]. Rub, G., "Data Mining of Agricultural Yield Data:A Comparison of Regression Models", *9th Industrial Confrence,* Vol. **5633**, pp.24-37, 2009.
[6]. Hemageetha, N., "A survey on application of data mining techniques to analyze the soil for agricultural purpose", *3rd International Conference on Computing for Sustainable Global Development (INDIACom),* pp.3112-3117, 2016.
[7]. Sujatha, R., Isakki, P., "A study on crop yield forecasting using classification techniques", *International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE),* pp.1-4, 2016.
[8]. Fathima, G.N., Geetha, R., "Agriculture Crop Pattern Using Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Engineering*, Vol. **4**, Issue 5, pp.781-786, 2014.
[9]. Kushwaha, A.K., Sweta Bhattachrya, "Crop yield prediction using AgroAlgorithm in Hadoop", *International Journal of Computer Science and Information Technology & Security (IJCSITS),* Vol. **5** No. 2, pp.271-274, 2015.
[10]. Sellam,V., Poovammal, E., "Prediction of Crop Yield using Regression Analysis", *Indian Journal of Science and Technology,* Vol. **9**(38), pp.1- 5, 2016.
[11]. Veenadhari, S., Misra, B., Singh, C.D., "Machine learning approach for forecasting crop yield based on climatic parameters", *International Conference on Computer Communication and Informatics*, pp.1-5, 2014.
[12]. Raorane, A.A., Kulkarni R.V., "Data Mining: An effective tool for yield estimation in the agricultural sector", *International Journal of Emerging Trends & Technology in Computer Science(IJETTCS),* Vol. **1**, Issue 2, pp.75-79, 2012.
[13]. Gayatri, M.K., Jayasakthi, J., Anandha Mala, G.S., "Providing Smart Agricultural Solutions to Farmers for better yielding using IoT", *IEEE Technological Innovation in ICT for Agriculture and Rural Development* (TIAR), pp.40-43, 2015.
[14]. Ankalaki, S., Chandra, N., Majumdar, J., "Applying Data Mining Approach and Regression Model to Forecast Annual Yield of Major Crops in Different District of Karnataka", *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. **5**, Special Issue 2, pp.25-29, 2016.
[15]. M.C.S. Geetha., A survey on data mining techniques in agriculture, *International Journal of innovative research in computer and communication engineering (IJIRCCE),* Volume **3**, Issue 2, Feb 2015.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 68-71(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Review on Emotional Intelligence and its Impact

Saranya Vijayan[1] and Dr. S. N. Geethalakshmi[2]
[1]M.Phil Research Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore (TN), India
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore (TN), India

ABSTRACT: Affective computing is the capability of machines to predict, interpret and detect responses and emotions of humans. The main goal of the affective computing is to detect and process the emotional information with the motive of improving the communication between the person and the machine. Emotional intelligence has got the ability to monitor our personal emotions as well as others' emotions and feelings, to discriminate amongst them, and to use this to guide our thinking and actions. The one who is emotionally intelligent is skilled in four regions such as identifying, using, expertising and regulating emotions. Emotional intelligence possesses characteristics such as understanding, coping with them, motivating, recognising emotions and coping with relationships. The objective of emotional intelligence in machines is to give them the ability to sense and recognize expressions of human emotions. It also enables the recognition that such communication is important for helping machines to choose more helpful and less aggravating behavior.

Keywords: Emotional intelligence, Affective computing

I. INTRODUCTION

Edelman *et al.* (1987) has endorsed that emotional incompetence often results from behavior discovered early in life. That automatic behavior is set in place as a regular part of living, as experience shapes the mind. As humans acquire their recurring repertoire of feeling, idea, and movement, the neural connections that assist these are strengthened, becoming major pathways for the impulses of nerves. Connections that are unused emerge as weakened, whilst those that humans use very frequently grow increasingly sturdy. When those habits have been so heavily found out, the underlying neural circuitry will become the mind's default alternative at any second. What a person does automatically and spontaneously results in little awareness of choosing to do so. Hence, for the shy people, diffidence is a habit that must be overcome. Emotional capacities like empathy or flexibility differ from cognitive abilities due to the fact they draw on different brain areas. Cognitive abilities are based on the neocortex.

Affective computing is currently one of the most important research topics and it has got increasingly intensive attention. This robust interest is driven by a huge spectrum of promising applications in lots of regions including virtual reality, clever surveillance, perceptual interface, etc.

Affective computing concerns with multidisciplinary information such as cognitive, physiology and computer sciences**.**

Cognitive learning involves fitting new data and insights into current frameworks of affiliation and information, extending and enriching the corresponding neural circuitry. However emotional learning involves cognitive learning and it requires interaction with the neural circuitry where our social and emotional dependency repertoire is saved. Changing habits which includes learning to approach people positively instead of fending off them, to listen better, or to give feedback skillfully, is a challenging task than clearly adding new information to old.

Motivational elements additionally make social and emotional learning greater difficult and complex than cognitive learning. Emotional learning often includes approaches of thinking and acting that can be more valuable to a person's identification. A person who's advised, for instance, that he must research a brand new word processing program normally turns into less disappointed than if he's instructed that he need to learn how to better manage his temper or turn out to be a higher listener. The prospect of desiring to develop greater emotional competence is a bitter pill for lots of us to swallow. It is therefore more likely to generate a resistance to change.

II. TOOLS TO DETERMINE EMOTIONAL QUOTIENT

a) **BarOn EQ-i® self-report.** Andre *et al.* (2004) have suggested that it is a tool that could be used to measure self report of emotional intelligence.The assessment of

this tool can be used when management or employee development initiatives are being considered or to assist in the recruiting or selection process. Research indicates that there is a correlation between emotional intelligence and job performance, making the assessment of this tool the ideal screening tool to aid in selecting potentially successful employees. This tool could also enable to create a profile of the top performers in any organization to determine what skills are the most valuable to your company in general or for specific job functions.

b) Self Report of Emotional Intelligence (EQ-360™) Andre *et al.* (2004) have recommended that the above mentioned tool identifies the extent of an individual's interpersonal functioning based totally on his or her responses. The tool assesses those who work closely with the client to provide information as well. Integrate external impressions of a client's emotional functioning with the client's self-report for a whole picture. The baron eq-360™ assessment also can be used to observe up and measure development in which formal coaching has been hired as a development method.

c) BarOn Emotional Quotient-Inventory: Youth Version (EQ-i:YV). Andre *et al.* (2004) has proposed that the tool could be used to guide children from 7 to 18 years old in the direction of feeling extra positive about themselves. Kids who are better able to cope with strain and stress, get in conjunction with others, and enjoy their lives emerge as less impulsive and more successful problem solvers and adapters. Development of these aspects of emotional and interpersonal abilities can significantly help optimize academic potential, interpersonal skills and ultimately life success.

III. COMPUTERS WITH EMOTIONAL INTELLIGENCE

Norman *et al.* (2014) has investigated that the intelligent machines must relate on an emotional basis with the users. With an established need for research in this area they have presented ideas in which researchers are making efforts towards building emotionally intelligent machines. All the models or ideas of emotional intelligence were referred to their ability to connect with others by expressing, managing, detecting and understanding the emotions of oneself and others. Sloman *et al.* (1981) has recommended that efforts in building machines which are emotionally sensible focuses round on a few key efforts such as empowering the machine to stumble on emotion, permitting the machine to express emotion, and finally, embodying the machine in a physical or virtual way.

IV. REVIEW: INFLUENCE OF EMOTIONAL INTELLIGENCE

Rosemary *et al.* (2003) has suggested that Teachers' emotions may also influence their *categorizing, thinking, and problem solving*. Experimentally induced positive effect often influences the way people categorize material (Isen, 1993). In positive affect conditions, people were more likely to categorize individuals favorably (e.g., classifying bartenders as "nurturant") than in neutral conditions (Isen *et al.*, 1992). Perhaps happy teachers categorize more students in positive categories such as hard working, or well behaved, or trying hard than do unhappy teachers. Teachers often treat students differently if they are classified as "trying hard but slow" rather than "lazy."

Michal *et al.* (2005) has introduced a method for verifying contextual appropriateness of emotions conveyed in conversations. Most of the popular methods focus only on simple emotion recognition, ignoring the complexity and the context dependence of emotions. However, to create a machine capable to communicate with a user on a human level, there is a need to equip it with Emotional Intelligence Framework [Mayer and Salovey, 1997]. The method described in their paper makes a step towards practical implementation of their framework, by providing machine computable means for verifying whether an emotion conveyed in a conversation is contextually appropriate. Their method has used affect analysis system to recognize user's emotions and a Web mining technique to verify their contextual appropriateness. In a rigorous evaluation, the affect analysis method was evaluated at 75% of accuracy in determining both valence and the specific emotion types. The accuracy of determining contextual appropriateness of emotions was 45% for specific emotion types and 50% for valence.

The result, although not perfect, was very encouraging, since the same evaluation performed with lenient conditions used popularly in the field gave the results of 80-85%. An agent equipped with their system could be provided with hints about what communication strategy would be the most desirable at any point. For instance, an agent can choose to either sympathize or to take precautions and help them to control their emotions. They were able to prove that computing emotions in a more sophisticated manner than simple division of positive and negative is a feasible task. Although the system as a whole was still not perfect and its components (ML-Ask and the Web mining technique) need improvement, it defines a new set of goals for Affective Computing. The computation of contextual appropriateness of emotional states is a key task on the way to full implementation of emotional intelligence in machines and as such is valuable to the research of Artificial Intelligence in general.

Maizatul *et al.* (2012) have investigated the impact of emotional intelligence on academic achievement among students of a university called universiti teknologi mara (uitm). The data of his research had

been acquired through the use of a questionnaire which elicited information on the students emotional intelligence stage as well as their academic performance. The study reveals that the students have high level of emotional intelligence. Domain names (self-emotion appraisal and understanding of emotion) of the emotional intelligence investigated was determined to be extensively and definitely associated with the respondents' academic achievement. The findings of the study hold crucial implications on the value of emotional intelligence and their relationships to students' educational performance specially among pre-service teachers.

Praveen *et al.* (2016) has proved after reviewing the literature that emotional intelligence is positively correlated with the work performance. A healthful relationship among employees and management also lead to an increase in worker's overall performance and thereby leading to improving business enterprise commitment. With the intention to enhance overall performance of administrative and practices it is necessary to develop emotional intelligence competencies in persons. The emotional intelligence constraints also are crucial for organisation productivity, social consciousness, self-management and self-cognizance. It was finally found that, effective personal competencies played a vital role in emotional intelligence, results in job satisfaction and organizational commitment that further ends in reduction in turnover intention, thereby enhancing the value and performance of the human resources.

V. APPICATION OF EMOTIONAL INTELLIGENCE IN COMPUTING ENVIRONMENT

Picard R.W *et al.* (1997) suggested that the human interaction consists of emotional information of the interlocutors that is transmitted with the aid of the explicit channel through the language and by means of the implicit channel by using the nonverbal verbal communication. The primary objective was to design affective systems that allows them to unite them with other systems and to create really intelligent and personal systems .Though, it has been evaluated that those concepts which might be additionally linked to the interpersonal relations emerge in the communication with the computers. Some of the areas that could help from the affective computing are distance learning, robotics, and psychotherapy. Furthermore, the affective computing research organization of the massachusetts institute of technology has commenced up an investigation, with the goal to help in the discipline of the distance learning, and to decide the emotional needs of the students.

Picard R.W *et al.* (1997) has proposed a help for the autistic people. This hassle affects around 1.5 in 2 people in each 1,000. The autism is a developmental disorder which continues throughout the life. This disease affects in different degrees. Some autistic people has got incredible memorizing and can discover ways to recognize facial expressions, but most of the people have problems with the compression of the emotions, and consequently they lack emotional intelligence. According to Picard, a way to aid autistic people is to rehearse them with different consequences repeatedly and by this way they will be able to respond. The plan was to design computers which had the capacity to teach those people different social scenes.

James *et al.* (2000) has recommended that another application is "driving safety". Taking into account that the lack of ability to manage one's emotions whilst driving is identified as one of the fundamental reasons for accidents, while the system identifies the driver is in a state of anger, for instance, the system could change the music depending on the driver's preferred style.

Lisetti *et al*. (2000) has suggested different application from automatic recognition of facial expression like user training, automobile driving alertness/drowsiness monitor, stress detector ,entertainment and computer games.

Nasoz *et al.* (2003) has suggested that affective computing is also beneficial in the field of telemedicine as it provides communication between medical professionals and patients where hands-on care is not required, but regular supervising is wanted. As an example tele-HHC involvements are employed in collecting crucial signal data remotely (for eg blood pressure, oxygen saturation, ecg etc)

VI. CONCLUSION

Affective computing can be also called as artificial emotional intelligence. It deals with the study and development of devices and systems that can recognize, simulate, interpret and process human affects. It is an interdisciplinary field that combines psychology,cognitive science and computer science. The computer should interpret the emotional state of humans and adapt its behavior to them, giving an appropriate response to their emotions.

Emotional Intelligence is a field that still requires research and it plays an important role in the field of artificial intelligence. It is vital to explore various methods as well as techniques in order to model emotions. The major challenge in the field of emotional intelligence is to enhance techonologies that can identify emotions. Emotional intelligence has got a set of abilities. Some of the ways to measure emotional intelligence is by finding the emotional quotient and through tests of certain abilities. The future work can be done by building machines that have emotional

intelligence as well as to build tools that help people boost their own abilities at managing emotions, both in themselves and in others.

REFERENCES

[1]. Andre *et al.,* 2004 Elisabeth Andre, Matthias Rehm, Wolfgang Minker and Dirk Buhler. Endowing Spoken Language Dialogue Systems with Emotional Intelligence, LNCS 3068:178-187, 2004.

[2]. Baba, 2003 Junko Baba. Pragmatic function of Japanese mimeticsin the spoken discourse of varying emotive intensity levels. *Journal of Pragmatics*, **35**(12):1861-1889, Elsevier. 2003.

[3]. Beijer, 2002 F. Beijer. The syntax and pragmatics of exclamations and other expressive/emotional utterances. Working Papers in Linguistics 2, The Department of English in Lund. 2002

[4]. Hager *et al.,* 2002 Joseph C. Hager, Paul Ekman, Wallace V. Friesen. Facial action coding system. Salt Lake City, UT: A Human Face, 2002.

[5]. Higuchi *et al*., 2008 Shinsuke Higuchi, Rafal Rzepka and Kenji Araki. A Casual Conversation System Using Modality and Word Associations Retrieved from the Web. In Proceedings of the EMNLP 2008, pages 382-390, 2008.

[6]. Picard *et al.,* 2001 Rosalind W. Picard, E. Vyzas, J. Healey, "Toward Machine Emotional Intelligence: Analysis of Affective Physiological State," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **23**(10): 1175-1191, 2001.

[7]. Norman, D. A. (2004). Emotional Design: Why We Love (or Hate) Everyday Things. New York: Basic Books.

[8]. Thorndike, E.L. (1920).Intelligence and its use. *Harper's Magazine*, **140**, 227-235.

[9]. Kihlstrom, J.F., & Cantor, N. (2000). Social intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence*, 2nd ed. (pp. 359-379). Cambridge, U.K.: Cambridge University Press.

[10]. Goleman, D. (1995). Emotional Intelligence. New York: Bantam Books.

[11]. Russell, S.J., Norvig, P. Artificial Intelligence: A Modern Approach. Prentice Hall, New Jersey: 1995.

[12]. P. Ekman. *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life.* Henry Holt and Company, New York, NY: 2003.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 72-76(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Survey on Classification Techniques in Data Mining

M. Jaithoon Bibi[1] and Dr. C. Yamini[2]
[1]M. Phil. Scholar, Department of Computer Science
Sri Ramakrishna College of Arts and Science for Women Coimbatore (TN), India.
[2]Assistant professor, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women Coimbatore (TN), India.

ABSTRACT: **Classification is a model finding process in data mining based on machine learning algorithms which uses probability artificial intelligence, distributions, statistics, also mathematics. It is frequently raised to supervised learning because the classes are determined before exploratory the data. Classification problem examples are text categorization, bio-informatics, optical character recognition, market segmentation and natural language processing and detecting faults in industry applications. The goal of this survey to deliver a complete review of various classification techniques in data mining based on Decision tree algorithm, Naïve bayes algorithm, K-nearest neighbour algorithm, Neural Networks algorithm, support vector machine algorithm.**

Keywords: Classification, Decision Tree, K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Artificial Neural Networks (ANN).

I. INTRODUCTION

Data mining is a group of methods for well-organized automated discovery of novel, dissimilar, valid, valuable and explicable patterns in huge databases. The present development of data mining products and functions is the results of manipulate from many disciplines including information retrieval, Databases, algorithms, Machine Learning and Statistics [2].

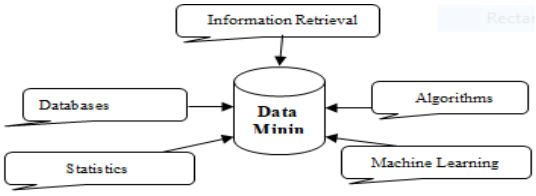


Fig. 1. Historical view of data mining.

Data mining is a method of discovering knowledge from huge databases or data warehouses several methods are used in data mining to remove patterns from enormous amount of database. Data mining algorithms to mine the information's and patterns derived by the Knowledge discovery in Databases (KDD). The KDD Process Consists of Five Steps [2].

- Selection
- Preprocessing
- Transformation
- Data Mining
- Interpretation / Evaluation

Classification is a data mining (Machine Learning) method used to predict assembly membership for data instances. Classification is a two-step process in data mining. First step is supervised learning for training data set and the second step is classification accuracy evaluation objects using cross-validation and multiple cross-validations. Classification involves supervised learning that is the training data has to require the classes. The main objective of the classification algorithm is to exploit the predictive accuracy achieved by the classification model when classifying examples in the test set unseen throughout training [3-4].

Classification algorithm finds interaction among the values of the target and the values of the predictors. For finding the relationships to apply the different classification techniques are used. Binary classification is applied for simplest type of classification problem.

- Example for two values: Yes or No.
- Examples of More than two values that is multiclass target: high, medium, low and unknown credit rating.

Classification project is usually divided into two data sets are first one is learning the model and the second one is testing the model [1].

II. LITERATURE REVIEW

Data mining is a process of analyzing data from different perceptions and assembly the knowledge from it.One of the most effective, easy to implement and

effective classification method to source the data from large database. Different classification algorithms applied for various datasets are measured in this section and explained.

Anju Rathee *et al* in [14] have described and functional ID3, C4.5 and CART decision tree algorithms on students' data to predict their performance. Comparison and evaluation of all these algorithms based on the performance and results on already existing datasets is done.

Chaitrali S. Dangare *et al* [15] in their paper have analyzed prediction systems for heart disease using a variety of input attributes which account to 15 medical attributes to predict the possibility of the patient getting a heart disease. The researchers use data mining classification techniques Decision Trees, Naive Bayes, and Neural Networks on Heart disease database and compare their performance based on accuracy of predicting heart disease.

Elakia *et al* in [16] have designed a system to justify that various data mining classification algorithms can be used on educational databases to suggest career options for high school students and predict potentially violent behavior among students by including additional parameters with academic details using a data mining tool called rapid miner.

Gilbert Ritschard *et al* in [17] have discussed the origin of tree methods and surveyed the earlier methods that led to CHAID decision tree classification algorithm. The authors have explained functioning of CHAID and briefed about the differences between the original method and the proposed extension method of CHAID.

S. Koyuncugil *et al* in [18] have presented a data mining model for detecting financial and operational risk indicators by CHAID decision tree algorithm.

D. Lavanya *et al* in [19] their paper have studied a hybrid approach wherein with CART bagging techniques and classifier feature selection have been considered to evaluate the performance based on accuracy and time for various breast cancer datasets.

Juan-Carlos Cano, Carlos T. Calafate, And Pietro Manzoni *et al* in [20], the researchers proposed a novel intelligent system which would be able to detect the road accidents automatically, notify them by using vehicular networks and estimate the severity of the accident based on data mining tools and knowledge interference. Various variables such as the vehicle speed, the type of vehicles involved, the impact speed, and the status of the airbag, etc. are used for measuring the severity of the accident. A prototype based on off-the-shelf devices was developed and validated it at the Apply us + IDIADA Automotive Research Corporation facilities, showing that this system can reduce the time needed to alert and deploy emergency services notably after an accident takes place. Three classification algorithms were used such as Decision Trees, Support Vector Machines and Bayesian networks and were compared for best results. It was found that Bayesian model for classification is the best-suited model.

III. DECISION TREE INDUCTION

Classification uses a decision tree algorithm as predictive modeling techniques. The highest task completed in these systems is consuming inductive methods to the certain values of attributes of an undefined object to determine suitable classification according to decision tree rules. The decision tree algorithms application areas are energy modeling, E-Commerce, Image processing, business, medicine, Industry, Web applications. There is various decision tree algorithm are named by C4.5, ID3, CHAID, QUEST, CARD, CRUISE, CTREE, GUIDE. ID3, C4.5, CART decision tree algorithms in data mining which practice various splitting criteria. Some of the splitting criteria characteristics which is used for impurity measures includes entropy, Classification error, Information Gain, Gain Ratio, Towing Criteria. Some of the decision tree learning software is WEKA, GATree, Alice d'Soft, See 5/C5.0 [1,8].

An attractive Inductive learning is a decision tree algorithm method for the reason is

- Classification process is self-evident so resulting decision tree provides a representation of the model that appeal to human.
- The methods are capable in computation that is comparative to the number of experiential training instances.
- Decision tree is a superior overview used for ignored instance, only if the instances are described in terms of character that are concurrent with the target class [8]

IV. NAIVE BAYES CLASSIFIER

Naive Bayes Classifier is a simple probabilistic classifier based on applying bayes theorem with strong (naïve) independence assumptions [1]. Bayesian Classification delivers a useful perception for accepting also estimating several learning algorithms. Bayesian Classification evaluates explicit possibilities for assumptions technique of naïve bayes for constructing classifier. Naive bayes needs only a lesser number of training data to appraisal the parameters for classification is a main advantage. Naive bayes also be trained by supervised learning. Simple bayes and independence bayes are as well as known as naïve bayes model. Event model is known as distribution of features in naïve bayes classifier. Some of the event model in naive bayes is Gaussian Naïve Bayes, multinomial naïve bayes, Bernoulli naïve bayes. Software are used for naive bayes classification techniques in data mining are Apache Mahout, Mallet,

Orange, Scikitlearn and WEKA, IMSL, JBNC, NClassifier [1,9].

Naive Bayes method is also known as idiot's Bayes, simple Bayes, and independence Bayes. This method is important for several reasons. This method is extremely easy to assemble, not needing any difficult iterative parameter estimation schemes. This way it may be eagerly applied to massive data sets. It is easy to understand, so users untrained in classifier technology can also be understand why it is building the classification it makes [6].

V. K- NEAREST NEIGHBOUR CLASSIFIER

K-Nearest Neighbour is based on the usage of distance measures in classification schemes.KNN techniques include the data in the set also with desired classification for each item. The following applications are applied in KNN algorithms are event recognition, Pattern recognition and object recognition. Face recognition using KNN containing features extractions are Haar face detection, mean-shirt tracking analysis, and PCA or Fisher LDA projection into features space [1, 2,10].

The neighbors are in use from a set of objects for which the correct classification (or, the value of the property, in the case of regression) is known. This can be consideration of the training set for the algorithm, while no precise training step is necessary. In order to classify neighbors, the objects are represented by position vectors in a multidimensional trait space. It is typical to apply the Euclidian distance, though other distance measures, such as the Manhanttan distance principle be used as an alternative [6].

VI. NEURAL NETWORKS

Neural Networks also referred as "Artificial Neural Networks". Group of joined units or node in a ANN is called Artificial Neurons. Neural Networks is a group of joined input/output units also every joining connection has a weight present with it.ANN widely applied on a various tasks consisting social networks, machine learning, computer vision, speech recognition, video games, playing board and medical diagnosis. The main three major paradigms are supervised learning and reinforcement learning. ANN Capabilities of broad categories are function approximation, classification, control, pattern, data processing, blind source detection, novelty detection and regression analysis [11, 12].

ANN are composed of interrelated, simple processing elements called artificial neurons.ANN possess some attractive qualities similar to those of biological neural networks, such as the fault tolerance, self-organization and capabilities of learning [5].

VII. SUPPORT VECTOR MACHINE (SVM)

Support vector machine is measured a upright classifier for it can act without the essential to enlarge a priori knowledge even once the dimension of the input space is very extraordinary, but its generalization performance is very high. Difference between the members of the two classes in the training data is used to find the superlative classification function. Support vector machine is used both non-linear and linear data for classify. Pattern classification is based on Support vector machine, so this technique widely used for classifying the different types of patterns. Applications can be used in real world problems by Support vector machine techniques are image segmentation, permutation, classification of image, text and hypertext categorization. SVM light and SVMJS live demo is a GUI for JavaScript implementations are the software tool in SVM [1, 6,13].

VIII. CLASSIFICATION ALGORITHMS : PROS AND CONS

This table contains the advantages and disadvantages of the various classification algorithms as below

Table 1: Advantages and Disadvantages of classification Algorithms.

Algorithm	Applications	Available implementation
Decision Tree	Energy modeling, E-Commerce, Image processing, business, medicine, Industry, Web applications	WEKA, GATree, Alice d'Soft, See 5/C5.0.
Naive Bayes	Categorizing news, email spam detection, face recognition, digit recognition and weather prediction, centimeter analysis	Apache Mahout, Mallet, Orange, Scikitlearn WEKA, IMSL, JBNC, NClassifier
K-Nearest Neighbour	Haar face detection, mean-shirt tracking analysis, and PCA or Fisher LDA projection into features space	CRAN, kknn, RWeka
Neural Networks	Character Recognition Image Compression Stock Market Prediction Traveling Saleman's Problem, Medicine, Electronic Nose, Security, and Loan Applications	Neural Designer, GMDH Shell, DNNGraph, HNN, Neuroph, Keras, Lasagne, DeepPy, neon, Tflearn, ConvNetJS, Synaptic
Support Vector Machine	Protein fold and remote homology detection, generalized predictive control, handwriting recognition, classification of image, text and hypertext categorization.	SVMlight SVMJS live demo, LIBSVM

IX. CLASSIFICATION ALGORITHMS : APPLICATIONS AND SOFTWARE TOOLS

This table contains the Classification techniques used Applications and Implementation tool.

Table: 2 Applications and Software tools.

Algorithm	Advantages	Disadvantages
Decision Tree	Easy to interpret and Understand	Does not handle continuous data
Naive Bayes	Eliminating irrelevant features of performance its improves	Needed huge amounts of records to get good results
K-Nearest Neighbour	Implementation is very easily and suitable for multiclass classification	It has memory limitation and slowly runs.
Neural Networks	Can Estimated any function, inspite of of its linearity, enormous for complex,/ abstract problems like image recognition	Rising precision by a few percent can bump up the scale by numerous magnitudes
Support Vector Machine	Precise accurate results will be provides	Computational expensive is very high. It runs slowly

X. CLASSIFICATION ALGORITHMS: MEASURES / VARIANTS

This table contains the Classification techniques used applications and implementation tools

Table 3: Classification Algorithm Measures / Variants.

Algorithm	Measures / Variants
Decision Tree	Entropy, Classification error, Information Gain, Gain Ratio, Twoing Criteria
Naive Bayes	Gaussian Naïve Bayes, multinomial naïve bayes, Bernoulli naïve bayes
K-Nearest Neighbour	Euclidian distance Manhanttan distance .
Neural Networks	Multi-layer kernel, Deep Predictive coding, Stacked Auto encoders
Support Vector Machine	Primal, Dual, Kernal trick,

XI. CONCLUSION

The paper provides a survey on some of the efficient classification algorithms such as Decision tree, Naïve Bayes, K- Nearest Neighbour, Neural Networks, and Support Vector Machine. This survey may motivate more researchers to use the following algorithms to clarify many research problems put-forth by the existing huge amounts of data for knowledge discovery. These algorithms are some of the most significant among the data mining classification techniques. The algorithms are reviewed in literature survey and a description of each algorithm is provides and some of the advantages and disadvantages are discussed on the algorithms. These paper also discussed about the applications, software tools, measure and variants. These algorithms can be used to solve problematic topics in data mining classification research and development.

REFERENCES

[1]. N. Satyanarayana, CH. Ramalingaswamy and Dr. Y. Ramadevi, Survey of Classification Techniques in Data Mining, *IJISET - International Journal of Innovative Science, Engineering & Technology*, Vol. **1** Issue 9, November 2014, ISSN 2348 – 7968.

[2]. Margaret H. Dunham, Data Mining - Introductory and advanced topics, eleventh impression, published by Pearson Education, ISBN: 0130888923.

[3]. S. Neelamegam, Dr. E. Ramaraj, Classification algorithm in Data mining: An Overview, *International Journal of P2P Network Trends and Technology* (IJPTT) – Volume **4** Issue 8-Sep 2013, ISSN: 2249-2615.

[4]. Mega Gupta, Naveen Aggarwal, Classification Techniques Analysis, NCCI 2010 -National Conference on Computational Instrumentation CSIO Chandigarh, INDIA, 19-20 March 2010.

[5]. Dr. A. Bharathi, E. Deepankumar, Survey on Classification Techniques in Data Mining, *International Journal on Recent and Innovation Trends in Computing and Communication,* Volume: **2** Issue: 7, ISSN: 2321-8169.

[6]. Raj Kumar, Dr. Rajesh Verma, Classification Algorithms for Data Mining: A Survey, *International Journal of Innovations in Engineering and Technology (IJIET).*

[7]. https://en.m.wikipedia.org/wiki/Aprioro_algorithm

[8]. Sonia Singh, Manoj Giri, Comparative Study Id3, Cart And C4.5 Decision Tree Algorithm: A Survey, *International Journal of Advanced Information Science and Technology (IJAIST),* Vol. **3**, No.7, July 2014, ISSN: 2319:2682.

[9]. https://en.m.wikipedia.org/wiki/Naive_Bayes_Classifier.

[10]. https://en.m.wikipedia.org/wiki/K-nearest_neighbour_classifier

[11]. Bharati M. Ramageri / Data Mining Techniques And Applications, *Indian Journal of Computer Science and Engineering* Vol. **1** No. 4 301-305.

[12]. https://en.m.wikipedia.org/wiki/K-nearest_neighbour_classifier

[13]. https://en.m.wikipedia.org/wiki/Support_Vector_Machine

[14]. Anju Rathee and Robin Prakash Mathur, Survey on Decision Tree Classification algorithms for the Evaluation of Student Performance, *International Journal of Computers & Technology*, Volume **4** No. 2, March-April, 2013, ISSN 2277-3061.

[15]. Chaitrali S. Dangare and Sulabha S. Apte ,Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques, *International Journal of Computer Applications* (0975 – 888) Volume **47**, No.10, June 2012.

[16]. Elakia, 2Gayathri, 3Aarthi, 4Naren J ,Application of Data Mining in Educational Database for Predicting Behavioural Patterns of the Students, *International Journal of Computer Science and Information Technologies*, Vol. **5** (3) , 2014, 4649-4652 with ISSN: 0975-9646.

[17]. Geneva, Switzerland, Juillet, CHAID and earlier supervised tree methods by Gilbert Ritschard, Dept of Econometrics, University.

[18]. A.S. Koyuncugil and N. Ozgulbas, Risk modeling by CHAID decision tree algorithm, vol. **11**, no.2, pp.39-46, Copyright© 2009 ICCES.

[19]. D. Lavanya and Dr. K. Usha Rani, Ensemble Decision Tree Classifier For Breast Cancer Data, *International Journal of Information Technology Convergence and Services (IJITCS*) Vol. **2**, No.1, February 2012.

[20]. Juan-Carlos Cano, Carlos T. Calafate, And Pietro Manzoni, "A System For Automatic Notification And Severity Estimation Of Automotive Accidents", Ieee, Francisco J. Martinez, Member, *IEEE Member, IEEE Transactions On Mobile Computing,* Vol. **13**, No. 5, May 2014.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 77-80(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Garbage Reporting and Monitoring App for Clean Society

Dr. R. Vijayabhanu[1] and G. Shobika[2]
[1]Assistant Professor, Department of Computer Science,
Avinashilingam Institute for Home science and Higher Education for Women, Coimbatore (TN), India.
[2]M.Sc. Computer Science Student, Department of Computer Science, Avinashilingam Institute for Home science and Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: Mobile technologies are powerful, fast-gripping and effective means of sharing information, which have brought indispensably positive changes in various management departments. Android, being a creative domain as well as an easily distributable flexible technology, it would be very effective to use its applications along with public cooperation. The cleanliness of the surroundings is relied on the responsibility of the government and also on the active cooperation and participation of the public. In that way, this proposal is aimed at assisting the government authorities (Municipality/Corporation) by developing a mobile application to monitor/report about the overflowing dustbins and receive feedback/complaints from public, thereby ensuring cleanliness, health and hygiene of our surroundings.

Keywords: Solid Waste Management (SWM), Android, Software Development Kit (SDK), Municipality.

I. INTRODUCTION

Health is wealth - The proverb highlighting a vigorous and hygienic environment leads to the prosperity of the individuals and the inhabitants. The urban population has been creating a highly polluted awful smelling atmosphere by littering the trash out of the bins. This hectic garbage chokes the drainages, spoils the beauty of the city, makes a filthy look and causes deadly diseases such as Dengue, Malaria, etc [1]. The uncontrolled hazardous waste creates various potentials risks to human health as mentioned elaborately in the report published by the *Royal Commission on Environmental Pollution of London* [2]. Solid Waste Management being a hot issue of Environmental Ministries had been undertaking numerous ways to discard the waste and trying to ensure the safety of our green environment. According to the assessment made in 2016, the annual solid waste has been approximately estimated to 1.3 billion tons and by 2025, the productivity may shoot up to about 4.3 billion tons [3]. The impact of such colossal quantity of trash would not explicitly enhance the worthy human livelihood.

Though the government takes obligatory steps to remove the trash from the dustbins, the collection of the garbage are uncertain [4]. The habitual cleaning of garbage bins at discriminatory intervals undertaken by the city municipal corporation facilitates only a partial fulfillment of our motto towards clean city. People should be aware of the consequences of littering garbage. Hence, the revolution of green and healthy situation can be achieved by the bonding of people and the regime.

II. ANDROID APPLICATIONS

Invented in 1992, the smart phones called as the hand-held computers, have become fast gripping technologies with Android OS, an Open Source platform [5]. The Software Development Kit (SDK) facilitates us to develop an application easily and is considered to be a highly secure platform as mended by Google in 2012. The application framework, JAVA class libraries and multimedia makes Android as one among the most affluent user interfaces.

Android also comes with built-in applications features such as Short Messaging Service functionality, phone capabilities and an address book (contacts). The integration of communication tools, data management functions, and multimedia tools with easy updates makes Android to be the best development kit for the SWM proposal.

III. PROPOSED SYSTEM

This proposal can be simply attained with the help of the most motivating mobile expertise which reduce physical and material efforts. The android application can be fashioned to help the municipal cooperation by posting the complaints of the overflowing waste by the public. Thus, the solid wastes are effectively managed at precise timings. However, convinced online applications are already shaped based upon the distributed sensor technology wherever the community anxiety is lacking [6].

The project of Solid Waste Management has been completed by monitoring our neighborhood trash bins and reporting their status to the municipal commission.

The main concept of this paper is to rectify the complaints given by the people within a short period of time. The system works with certain entities to complete the assignment as portrayed by following:-

Admin Section: Accomplishes the task of **adding locations** of the garbage bins and its **location**. The required **employees** are also recruited under this module.

User Section: The users can register their details thereby generating their profiles and register **complaints** about overflowing garbage and checking the status of their complaints.

Manager Section: Allocates the work to the employees based on the complaints placed by the users.

Employee Section: Realized by their **status updates** this can be inspected by the manager as well as by the users themselves.

IV. SYSTEM WORK FLOW

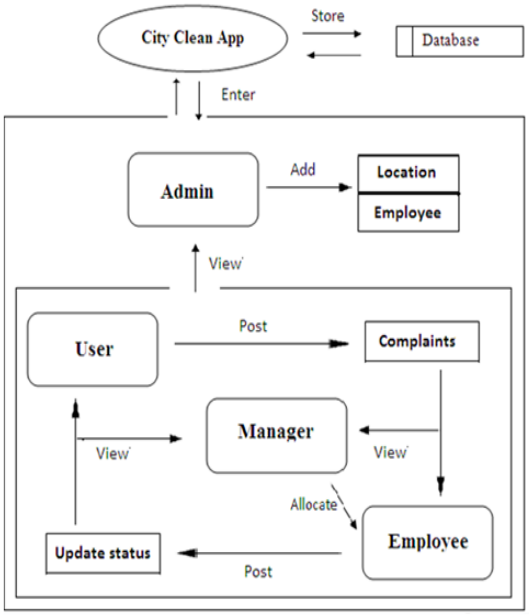


Fig 1 Systematic flowchart

V. SYSTEM DEVELOPMENT

This Android application called 'Clean City' has been developed using an open score software Eclipse, the integrated development environment as a front end and the Microsoft SQL Server as the backend. The SQL Server supports the system by effectively storing and handling large amount of records generated while using the following four modules-

Admin Section. This section has been used to add their precise details about the location of the garbage bins and their areas in the particular municipality. It also generates I.Ds and passwords for the employees such as managers and the others.

The municipal influence are about to access this module to monitor the cleanliness of our location.

User Section. This section has been created for the public who are willing to register their names and particulars as users. By generating their user IDs and passwords, they post their complaints by selecting the location of the overflowing dustbins. Their complaints can be verified with the update status sent by the employees after the completion of the work.

Manager Section. The manager being enlisted by the admin logs in to check the complaints posted by the users and assign employees to collect the garbage. The activities of the employees can be monitored and ensured by receiving the update message.

Employee Section. This section has been designed to view the complaints placed by the manager and to update the condition of cleanliness which can be checked by the manager as well as by the user.

VI. RESULTS AND DISCUSSION

The android application "Garbage Waste Reporting and Monitoring System" generated with the user-friendly software Eclipse and maintenance of the related database by the MS SQL has produced the following upshots.

The subsequent Fig. 2 invites us for the admittance of the application, through which correlated modules can be accessed.

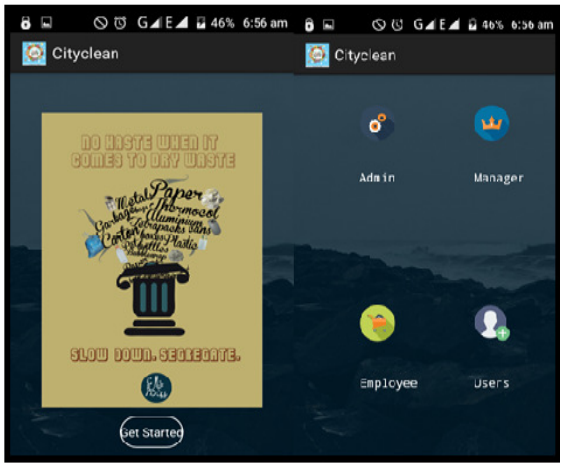


Fig 2 Home Page

1. Admin Section. Logging in to this module by the municipality authority gives him an advantage of adding the locations of the garbage bins and their areas in that particular municipality. Fig.1.a also shows us the ways of recruitment of the employees by the admin.

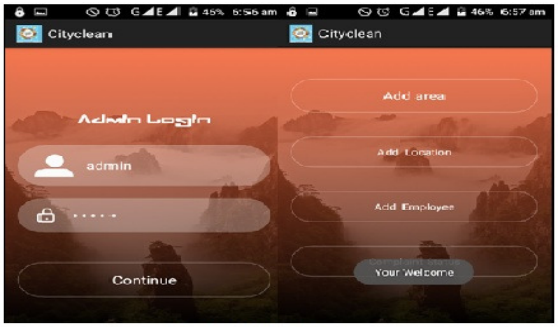


Fig 1a Admin login page

Fig 1.b further explains the process of employment by adding their details and acknowledging the employee with their login and password details through a text message.

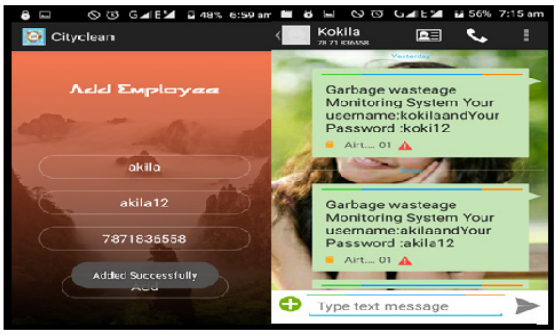


Fig 1b Adding employee

2. User Section. The registration of the user with inquired details helps them to generate their account as shown in the Fig 2.a. Later, the account can be signed in to register their complaints.

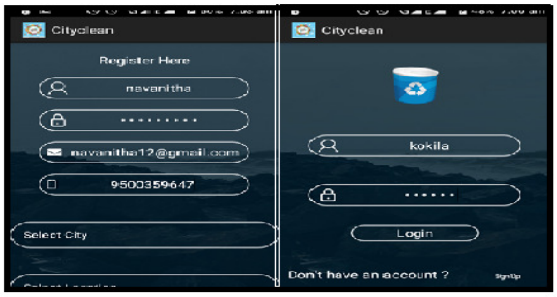


Fig 2a User registration and Login

Demonstrated Fig. 2b explains the ways in which various preferences can be accessed by the user and the complaints being cataloged.

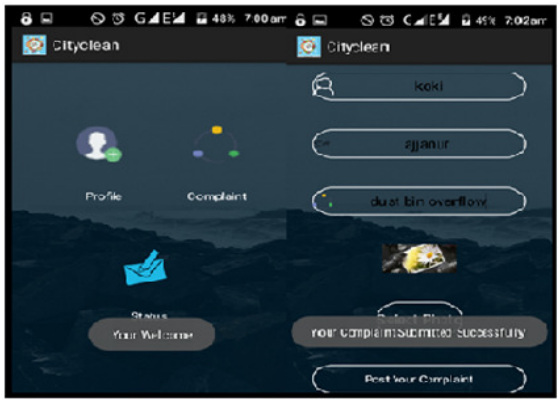


Fig 2b Registering the complaints

3. Manager Section. The presented Fig 3a gives us an idea about the login details of the manager and his inspection on the user complaints and allocation of the employee and reviewing their updates.

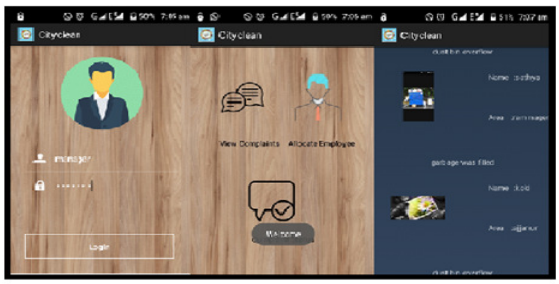


Fig 3a Logging in and viewing complaints

Fig 3b describes the assignment of the employee by the manager, according to the registered complaints.

Fig 3b Allocating work to the Employee

4. Employee Section. Fig 4a points out the login details of the employee, aids with the utility of viewing complaints and updates the status

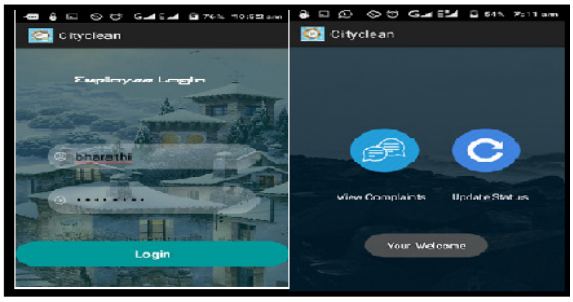


Fig 4a Employee Login Page

Viewing of the registered complaints by the employee has been expounded in the pursuing Fig 4b.

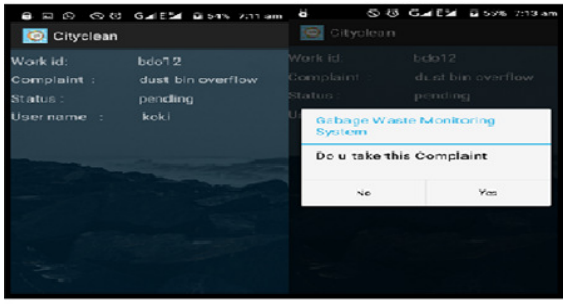


Fig 4b Employee views the complaints

Accomplishing the given task, the employee updates the status of the trash bins as in Fig 4c.

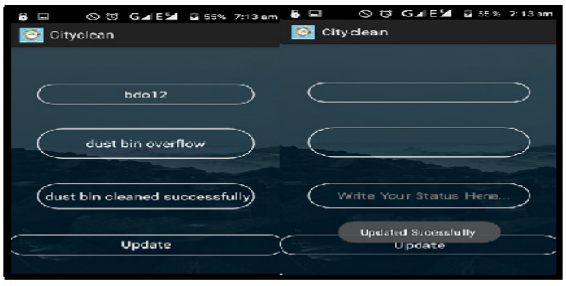


Fig 4c Updating the status

Fig. 4d indicates the updated status of the employee which has been notified to the manager as well as to the user through a text message.

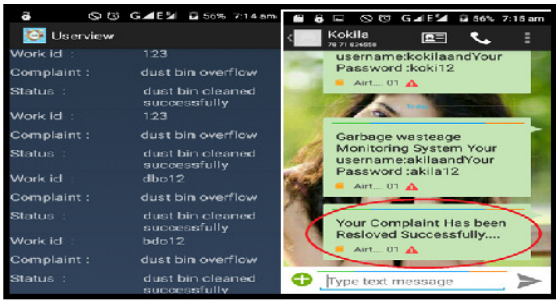


Fig 4d Updating the complaints status

VII. CONCLUSION

Thus, a solution can be provided for unsanitary environmental circumstances in a city and also by preventing many diseases caused due to the toxic gases emanating from the overflowing litters. It helps to maintain a clean and healthy environment throughout the country. Databases of these applications can be maintained in the municipalities. The communication between users and the municipalities can also be effortlessly maintained. Futhermore, the concept can be extended to maintenance of the water tank overflowing issues, Street Light Complaints, Road repair complaints, etc.

REFERENCES

[1]. Rosenbaum J, Nathan MB, Ragoonanansingh R, Rawlins S, Gayle C, Chadee DD and Lloyd LS., "*Community participation in dengue prevention and control: a survey of knowledge, attitudes and practice in Trinidad and Tobago*" *American Journal of Tropical Medicine and Hygiene*, 1995, **53**(2): 111-117.

[2]. Royal Commission on Environmental Pollution. 10th Report tackling pollution—experience and prospects London: HMSO; 1984. Feb.

[3]. Hoornweg, Daniel; Bhada-Tata, Perinaz. 2012. *What a Waste : A Global Review of Solid Waste Management. Urban development series; knowledge*, papers no. 15. World Bank, Washington, DC. © World Bank.

[4]. Tarandeep Singh, Rita Mahajan and Deepak Bagai "*Smart Waste Management using Wireless Sensor Network*" *IJIRCCE*, Vol. **4**, Issue 6, June 2016.

[5]. Hall S. P. and Anderson E., "*Operating Systems for Mobile Computing*", *Journal of Computing Sciences in Colleges,* December 2009, ISSN:1937---4771

[6]. Rovetta, A., Xiumin, F, Vicentini, F, Minghua, Z, Giusti, A, and Qichang, H., "*Early Detection and Evaluation of Waste through Sensorized Containers for a Collection Monitoring Application in Waste Management*", Vol. **29**, Issue 12, 2009, pp. 2939-2949.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 81-85(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Routing Schemes and Protocols for Internet of Things: A Review

M. Girija[1] , Dr. S. Sivagurunathan[2] and Dr. P. Manickam[2]
[1]Assistant Professor, Department of Computer Science, The American College, Madurai.
[2]Assistant Professor, Department of Computer Science and Applications, Gandhigram Rural Institute.

ABSTRACT: The Rapid development of the wireless networks such as Internet, sensor networks and wireless technology like Radio Frequency Identification (RFID) systems has directed to the idea of Internet of Things (IoT). Internet of Things (IoT) is a technological revolution in computing and communication like internet and it is the new dimension for research over the last few years. IoT is extension of internet and collection of inter connected everyday objects of different types such as digital and mechanical devices, animals, people or other objects. This emerging technology facilitates the communication between people and things and among things themselves.

Keywords: Radio Frequency IDentification (RFID), Wireless Sensor Networks (WSN)

I. INTRODUCTION

Recently, Internet of Things (IoT) [1-3][9] devices are highly utilized in diverse fields such as environmental monitoring, industries, smart home etc. Internet of Things (IoT) ensures that to establish the connection among smart devices and objects [4-5] such as computers, mobile phones, Laptops, sensor devices and other smart objects at anytime and anywhere as shown in Fig. 1. Smart objects are an important part of Internet of Things and operated by limited battery energy. IoT forms the dynamic and active network using ubiquitous and uniquely addressable objects in all domains such as civil, electrical, agriculture, mechanical fields. IoT devices are built with advanced processing such as collecting information, sends information and processing the information using RFID devices[6-8][17] with protocols. The Elements of IoT are

- Hardware - Sensors, actuators and embedded communication hardware
- Middleware - Data storage and computing tools for data analytics
- Presentation - Visualization and interpretation tools for different platforms and different applications

IoT devices have sensors and actuators to observe and collect the domain information and sends processed information to remote server with the support of wireless networks. Internet of Things (IoT) are applicable in diverse fields such as Agriculture (Monitors Humidity condition , Temperature, Growth of Plant), Industries, Traffic system, Environmental monitoring system (Temperature, Climate, monsoon), Hospital (Doctors can monitor patient's health conditions such as Body Temperature, Blood Pressure, ECG , Smart home (Control the Heat/Cool conditions of Room), City (Monitor the parking vehicles, Road layout).

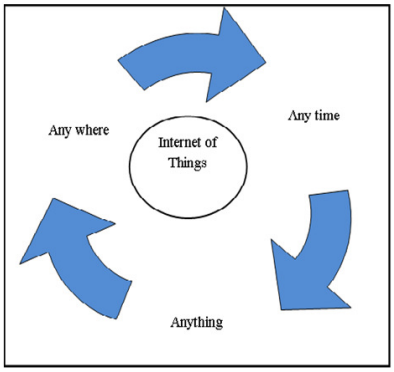


Fig. 1. Internet of things (IoT).

II. APPLICATIONS OF IoT

IoT offers various services to the world in diverse applications ranging from Agriculture to Health care system. Let us discuss few of them in this paper.

- Smart Agriculture
- Smart Traffic Management
- Smart Home
- Smart Industry
- Smart Life

A. Smart Agriculture

Agriculture is an important area to focus because people changed their life to modern life. Government

should focus agriculture to support the human life free from artificial fertilizers in addition to improve its commercial status. IoT provides services to farmers in three stages such as pre-harvest (information of soil status), harvest (monitoring crops) and post harvest (storage of crops).

B. Smart Traffic Management

IoT monitors and effectively manage the traffic in order to provide various services to humans in general and to avoid the accident in particular. IoT ensures that to monitor the traffic and make the dynamic decision at the peak hours which in turn reduce the road accident. This is more helpful for humans beings such as to drive the vehicle happily, Ambulance can reach the hospital in time, school children feel happy to travel in school bus using sensors [10][11][14], GPS.

C. Smart Hospital and Life Management

Doctors could collect and know their patient health conditions and make the right decision. *i.e.* IoT provides freedom to medical professionals like need not sit in one place and closely monitoring patient. IoT helpful for doctors to concentrate in other services in addition to provide medical services not only one patient. Similarly, IoT provides various medical services to aging people in order to satisfy their medical help.

III. ARCHITECTURE OF IoT

IoT has 3-layers such as Perception Layer, Network Layer, and Application Layers [12] shown in Table 1.

-Perception Layer observes and collects physical properties and atmosphere conditions and converts them as signals using sensors and actuators. The converted digital signal will send over the network.

-Network layer has two sub layers such as Routing Layer and Encapsulation Layer. Former is find out the optimal route using various routing protocols such as RPL, CoRPL and sends the packets. Later is frames the packet and send over the new route.

Application Layer	
Network Layer	• Encapsulation Layer • Routing Layer
Perception Layer	• Coordinator nodes • Actuator nodes • Sensing nodes

Table 1- Architecture of IoT

-Application Layer Application layer [19] provides various services irrespective of applications and acts as interface between user and networks. This layer has various management techniques to control the applications. Application layer receives data from underlying layers and provides as services to the users

IV. ISSUES AND CHALLENGES OF IoT

There are many issues and challenges [13][15] exist in IoT because of its architecture and characteristics. i.e. IoT establishes the network using wireless sensor nodes using wireless networks as the transmission medium. The Wireless Sensor Networks (WSN) are operated with limited battery power and proper plan is needed to use the WSN nodes' energy at optimized level.

Security: IoT can setup the network with any nodes anywhere in wireless [16] [18] medium. This is the major avenue for unauthorized people to hack the information. So, Security is an avenue to researchers to do research in order to address the various threats and attacks.

Battery Lifetime: As we discussed in introduction section, sensor nodes are operated by limited battery energy. In any sensor enabled network or wireless network, energy is an important factor to consider. As we know, nodes are depleting their energy by share, send or receive information. As we know, IoT consists of sensor nodes, actuators and other smart devices. Since, these devices are operated by battery energy. If nodes are in continuous transmission, nodes are may be down and the network will be partitioned which in turn packets will be dropped. This is a major challenging issue of wireless networks.

Data management: IoT collects data and information from its many of the devices per second. Also, it spends lot of efforts to process the data in order to interpret the result and make the decision at the critical situation. Moreover, IoT concentrates humans' life in particular. So, data management is another important issue in IOT.

V. ROUTING PROTOCOLS OF IoT

Routing is one of the major research areas in IoT[19] [20][21] and designing routing protocol for IoT is a really challenging task for research community. The routing protocol should minimise Energy consumption and maximise Network lifetime, Quality of services. Security is also another major research area in IoT because the transmission medium is wireless.

Routing is a process which identifies the path from a source to destination node to deliver a data packet. Path to the destination node may either be direct or indirect by including intermediate nodes. Routing algorithm is a procedure to identify the optimal route based on routing metrics and to switch the packets. Routing algorithm helps the hosts to determine the optimal route, sharing the routing information with the neighboring nodes. Routing algorithms are classified based on features such as optimality, simplicity, stability, and flexibility. Routing has two primary tasks such as

-The first task is path determination. Path determination is to choose the best path based on routing metrics. Routing metrics are hop count, bandwidth, delay, load and cost. A best path introduces minimum delay with a minimum cost to deliver a packet. Also, the best path should include minimum hop counts and load, but a maximum bandwidth.

-The second task is switching. Switching is to transmit the data packets from the source node to the destination node once the path is determined.

The chief role of the routing protocol includes finding the path to the destination, sharing the routing information among the nodes about the network and informing link break status to the nodes in the network. A routing protocol is a set of algorithms which instruct the routers to select the routes for data transmission.

In ad hoc wireless networks, an efficient routing protocol should possess the following qualities:

Adaptability-Routing protocols should be adaptable to network topology changes.

Efficiency-Routing protocols should reduce consumption of bandwidth by way of limiting control messages.

Security-Routing protocols should address the various threats. Since, IoT setup the network anywhere with any nodes in wireless medium. This is the major avenue for unauthorized people to hack the information.

Power consumption-Routing protocols should select a path from a source to some destination which minimizes the consumption of total energy for transmitting a message in that path.

Load distribution-Routing protocols should extend the life of network by balancing the traffic load and distributing the load to the multiple different paths.

We will discuss few IoT Routing Protocols in this section such as Routing Protocol for Low Power and Lossy Networks (RPL), Cognizant Routing Protocol for Low Power and Lossy Networks (CORPL).

RPL: Routing Protocol for Low Power and Lossy Networks Protocol [22] is a distance-vector routing protocol and route the packets based on graph construction (DAG-Direct Acyclic Graph, DODAG-Destination Oriented DAG).

This protocol uses three messages such as DODAG Information Solicitation (DIS), DODAG Information Object (DIO), Destination Advertisement Object (DAO). It is a source routing protocol and designed for devices with limited memory and low power processor. This protocol periodically sends information to collection point. RPL does not support multipath routing. It does not balance energy balancing as well as load balancing into consideration.

CoRPL: CoRPL Routing Protocol [23] protocol is the new version of RPL suited for Cognitive Radio environments. It works based on the Directed Acyclic Graph (DAG) approach. Based on results, CORPL achieves high throughput with minimum delay. Also, it ensures that consumes the energy at the average level.

LOADng: Lightweight On-demand Ad hoc Distance-vector Routing Protocol - Next Generation [24] is a routing protocol that emerged as an alternative to RPL. LOADng is a reactive routing protocol designed for Low Power and Lossy Networks (LLNs). It uses four control messages such as Route Request (RREQ), Route Reply (RREP), Route Reply Acknowledgment (RREP ACK), and Route Error (RERR). It supports all lengths of addresses such as IPv6 or IPv4) and does not use periodical control messages such as HELLO or Beacon messages. If source has data to send LOADng protocol ensure source to select the shortest route based on hop count. So, LOADng protocol focuses hop count to optimize the route rather than other metrics.

PAIR: Pruned Adaptive IoT Routing [25] introduces a pricing model to solve the incompatible issues. PAIR protocol supports multi hop and context aware routing protocol. PAIR protocol works based on metrics such as Residual energy, power consumption, Current load, buffer space and Distance to neighbor. PAIR protocol works in two phases such as forward and backward. This protocol has to concentrate security issues. This protocol needs large amount of memory to establish the new route when active route down due link break.

AOMDV-IoT: Ad-hoc on demand Multipath Distance Vector routing protocol for IoT [26]. AOMDV-IoT establishes the connection among regular nodes and the internet nodes. It introduces needs extra memory at every nodes in order to maintain two tables such as Internet connecting table (ICT) and the routing table. This protocol uses the following messages in order to send the packets over the route such as Route Request (RREQ), Route Reply (RREP), ACK message and Route Error (RERR). It is an on demand routing protocol i.e. When the source node has data to send, it broadcasts the route request message to establish the route to destination using RREQ message. AOMOV-IoT does not have security mechanism in the data routing. Also, it fails to optimize the route when source has data to send to destination.

NUD: Neighbor Unreachable Detection (NUD)[27] is similar to RPL. NUD has two messages such as hello packet and DIO messages. It ensures parent node to form DODAG or DIO message to all nodes in a network. DIO Message has Node Rank, Energy Status, Parent etc . Rank is the position with respect to sink node. If a node receives many different DIO messages from neighbour nodes then NUD ensure node that select the minimum rank and forwards DIO message in the network.

Also, every node maintains information about Rank of all the nodes in the networks. Either source or intermediate node sends packets over the new route if receives ACK packet. Otherwise, node will initiate alternative route to send the packet.

OF-FL: Objective Function based on Fuzzy Logic [28]. OF-FL has included Fuzzy concepts in objective function. Objective Function ensures nodes that to select the parent to forward the packet. OF-FL has included routing metrics in objective function such as point-to-point delay, Expected no of Transmissions (ETX), hop count, and battery energy level. OF-FL selected these metrics as input to Fuzzy Inference System while select the parent. Also, OF-FL maximizes network lifetime with minimization of point-to-point delay.

CAOF: Context-Aware Objective Function [29]. CAOF proposed CAOF for sensor networks to establish the route in order to send the packet. CAOF scheme concentrated and focused on nodes' battery energy due to sensor nodes are operated by limited battery energy. CAOF includes three metrics Node connectivity degree, battery energy level, and node position in objective function. Context-Aware Objective function helps nodes to select the parent to route the packets in order to achieve maximum packet delivery ratio.

ERGID: Emergency Response IoT based on Global Information Decision [30] improves the reliable data transmission in IoT. ERGID has two methods called Delay Iterative Method (DIM) and Residual Energy Probability Choice (REPC). DIM designed based on delay estimation to solve the problem of ignoring valid paths and REPC works based on residual energy of nodes. ERGID extends the network life time with minimum energy consumption. Also, Experiments done on STM32W108 sensor nodes and ERGID improved the Quality of Service of network.

EEPR: Energy Efficient Probabilistic Routing Algorithm [31] designed to extend the network lifetime. As discussed, sensor nodes are operated by limited battery energy. In any sensor enabled network or ad hoc network, energy is an important factor to consider. Nodes deplete their energy by share, send or receive information as well as control messages. Also, nodes are depleting most of their energy by processing RREQ messages or overhearing messages which in turn network will be partitioned. Due to this, packets will be dropped and EEPR has designed to address this problem. Based on results, it is evident that EEPR algorithm reduces congestion in the network and packet loss. EEPR algorithm ensures nodes that to consume the energy at minimum level.

CA-RPL: Congestion Avoidance Multipath Routing Protocol [32] designed based on Routing Protocol for low power and lossy network (RPL). Congestion is a major issue both in wired and wireless networks. But chances are more in wireless network. Nodes spent extra energy to play the various roles such as address the congestion issue, inform to either source node or destination node by sending RERR messages, source node establish the new route by sending RREQ and receiving RREP, send the packet over the new route. CA-RPL can avoid congestion by construction of Directed Acyclic Graph and minimizes packet loss ratio which minimizes energy consumption.

VI. CONCLUSION

Internet of Things (IoT) helps to setup the connection among smart devices at anytime, anywhere and moves us to next generation of networks. This paper highlights the importance of IoT concepts and focused the certain applications of IoT and explained the use of IoT in Agriculture and Hospital system in particular. In this paper, various routing protocols such as Routing Protocol for Low Power and Lossy Networks (RPL), Cognizant Routing Protocol for Low Power and Lossy Networks (CoRPL), Lightweight On-demand Ad hoc Distance-vector Routing Protocol - Next Generation (LOADng) and other protocols are discussed in detail with their working principles. As IoT is an emerging and developing technology, this paper provides directions for research community to focus in this area and initiate the research.

REFERENCES

[1]. Eleonora Borgia,"The Internet of Things vision: Key features, applications and open issues", *Elsevier, Computer Communications,* vol. **54,** 2014,1–31.

[2]. Daniele Miorandi et al.," Internet of things: Vision, applications and research challenges" , *Elsevier, Ad hoc Networks,* vol.**10**, 2012,1497-1516.

[3]. M. Buzzi, M. Conti, C. Senette, D. Vannozzi, "Measuring UHF RFID tag reading for document localization", *Proceedings of IEEE International Conference on RFID-Technologies and Applications (RFID-TA),* 2011, pp. 115–122.

[4]. M. Zorzi, A. Gluhak, S. Lange, A. Bassi, From today's INTRAnet of things to a future INTERnet of things: a wireless- and mobility related view, *IEEE Wireless Communications*, vol.**17**, No.6, 2010,pp. 44–51.

[5]. A. Krause, A. Smailagic, D.P. Siewiorek, Context-aware mobile computing: learning context dependent personal preferences from wearable sensor array, *IEEE Trans. Mob. Computers*, vol. **5**, 2006,pp.113–127.

[6]. Omar Said, Mehedi Masud, "Towards Internet of Things: Survey and Future Vision", *International Journal of Computer Networks (IJCN),* Volume (**5**), Issue (1), 2013.

[7]. Quynh, Thu Ngo, Nien LeManh, Khoi Nguyen Nguyen, "Multipath RPL protocols for greenhouse environment monitoring system based on Internet of Things", Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference on. IEEE, 2015.

[8]. R. Gadh, G. Roussos, K. Michael, G.Q. Huang, B.S. Prabhu, C.C.P. Chu, RFID – a unique radio innovation for the 21st century, *Proc. IEEE* 98 (**9**) (2010) 1546–1549.

[9]. Omar Said, Mehedi Masud, "Towards Internet of Things: Survey and Future Vision", *International Journal of Computer Networks (IJCN),* Volume (**5**) : Issue (1) : 2013.

[10]. K Ashton, "Internet of Things", *RFID Journal*. 2009, **22**(7), 97-114.

[11]. G Kortuem, F Kawsar, V Sundramoorthy, D Fitton, "Smart objects as building blocks for the internet of things", *IEEE Internet Computing*, 2010, **14**(1), 44-51.
[12]. WB Heinzelman, AP Chandrakasan, H Balakrishnan, "An application-specific protocol architecture for wireless microsensor Networks", *IEEE Transactions on wireless communications, 2002*, **1**(4), 660-70.
[13]. Suk Kyu Lee , Mungyu Bae and Hwangnam Kim, " Future of IoT Networks: A Survey", *Appl. Science,* 2017, **7**, 1072.
[14]. Vural, S.; Wang, N.; Navaratnam, P. Tafazolli, R. Caching ," Transient Data in Internet Content Routers", IEEE/ACM Trans. Netw. 2017, **25,** 1048–1061.
[15]. Hail, M.A.M.; Amadeo, M.; Molinaro, A.; Fischer, S. On the Performance of Caching and Forwarding in Information-Centric Networking for the IoT. In Wired/Wireless Internet Communications; Springer, 2015; pp. 313–326.
[16]. Ian F. Akyildiz, (2002). W. Su, Yogesh Sankarasubramaniam, and Erdal Cayirci, "Wireless sensor networks: a survey", *Computer Networks*, **38**(4), 393–422.
[17]. Kevin Ashton, (2009). "Internet of Things" Thing, RFID Journal website, http://www.rfidjournal.com/articles/view?4986, retrieved 28 April 2014
[18]. Karagiannis, V., Chatzimisios, P., Vazquez-Gallego, F., & Alonso-Zarate, J. (2015). A Survey on Application Layer Protocols for the Internet of Things. *Transaction on IoT and Cloud Computing*, **3**(1), 9-18 .
[19]. Ravi Kumar Poluru and Shaik Naseera, "A Literature Review on Routing Strategy in the Internet of Things", *Journal of Engineering Science and Technology Review*, vol. **10,** No 5 (2017) 50 – 60.
[20]. K Rose, S Eldridge, L Chapin, " The internet of things: An overview", The Internet Society (ISOC), 2015, pp.1-50.
[21]. P Sethia, SR Sarangi, "Internet of Things: Architectures, Protocols, and Applications", *Journal of Electrical and Computer Engineering, 2017*, DOI: org/10.1155/2017/9324035.
[22]. T. Winter, "RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks", March 2012, https://tools.ietf.org/html/rfc6550.
[23]. Adnan Aijaz and A. Hamid Aghvami, "Cognitive Machine-to-Machine Communications for Internet-of-Things: A Protocol Stack Perspective", *IEEE Internet of Things Journal,* Vol. **2**, No. 2, APRIL 2015.
[24]. T. Clausen, J. Yi, A. Niktash, Y. Igarashi, H. Satoh, U. Herberg, C. Lavenu, T. Lys, and J. Dean, "The lightweight on-demand adhoc distance-vector routing protocol-next generation (loadng)," Working Draft, IETF Secretariat, Internet-Draft draft-clausen-lln-loadng-14.txt, 2016.
[25]. Sharief M. A. Oteafy, Fadi M. Al-Turjman and Hossam S. Hassanein, "Pruned Adaptive Routing in the Heterogeneous Internet of Things", Global Communications Conference (GLOBECOM), 2012 IEEE, pp. 214 – 219, ISSN 1930-529X.
[26]. Yicong Tian, Rui HOU, "An Improved AOMDV Routing Protocol for Internet of Things", *International Conference on Computational Intelligence and Software Engineering (CiSE),* 2010, pp.1-4.
[27]. L Pradittasnee, "A study of the neighbor unreachability detection mechanism to improve performance of RPL protocol", *Information Technology and Electrical Engineering (ICITEE).* 2016, pp 1-6.
[28]. O. Gaddour, A. Koubˆaa, N. Baccour, and M. Abid, "OF-FL: QOS- aware fuzzy logic objective function for the rpl routing protocol," *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt),* May 2014, pp. 365–372.
[29]. B. Sharkawy, A. Khattab, and K. M. F. Elsayed, "Fault-tolerant rpl through context awareness," in 2014 *IEEE World Forum on Internet of Things (WF-IoT), March 2014*, pp. 437–441.
[30]. Tie Qiu, Amr Tolba, "An efficient routing protocol for emergency response Internet of Things", *Journal of Network and Computer Applications,* Vol. **72**, Sep. 2016, pp 104-112.
[31]. Sang-Hyun Park, Seungryong Cho, Jung-Ryun Lee "Energy-Efficient Probabilistic Routing Algorithm for Internet of Things", Hindawi Publishing Corporation, *Journal of Applied Mathematics*, 2014, Article ID 213106.
[32]. Weisheng Tang, Xiaoyuan Ma, Jun Huang, Jianming Wei, "Toward Improved RPL: A Congestion Avoidance Multipath Routing Protocol with Time Factor for Wireless Sensor Networks", *Journal of Sensors,* 2016, Article ID 264982.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 86-90(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Vehicle and Speed Detection using Image Processing Techniques

K. Mirunalini and Dr. Vasantha Kalyani David
[1]Ph.D., Research Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India

ABSTRACT: Recently vehicle and speed receives greater attention for minimizing the road accidents and controlling traffic by restricting the speed of vehicles. Using vehicle speed detection is, estimating the velocity of the moving vehicle by applying image and video processing techniques. Without any camera calibrations, a video is captured and analyzed for speed in real time. The speed is determined using distance traveled by vehicle over the number of frames and frame rate.

Speed of a moving car is assessed using radar, photo radar, drone radar, LIDAR, speedometer clocks, common speed computer systems, aircraft or video frame methods. These systems are significantly helpful to observe, monitor and manage various traffic conditions such as traffic management, prevention of accident, and secure transportation. One of the aims of these systems is to determine the speed of vehicle on the road. Vehicle speed determination has become a tough task recently. While determining the speed the two main steps involved are vehicle detection and vehicle tracking.

In order to overcome the drawbacks of the traditional method vehicle speed determination is done using image processing and MATLAB. Most of those techniques are very expensive and additionally, their accuracy is not up to the expectation. In this paper a video-based vehicle speed calculation method is followed, which has the ability of calculating the speed with higher accuracy with relatively low cost. The proposed method includes primarily 4 steps namely Preprocessing, Morphological operation, Vehicle detection and Vehicle tracking.

I. INTRODUCTION

In an intelligent traffic system, vehicle detection and speed measurement play an important role in imposing speed limits. They also provide relevant data for traffic control. These systems use intrusive and non-intrusive sensors. Intrusive sensors, based on inductive loop detectors, are widely used today. However they involve complicated installation and maintenance, accelerate asphalt deterioration, and can be damaged by wear and tear. Non-intrusive sensors, which include laser meters and Doppler radars help to avoid these problems, but are usually more expensive and frequent maintenance are required.

To detect the speed of the vehicle the most popular techniques used include RADAR (Radio Detection and Ranging) and LIDAR (Light Detection and Ranging) devices. A RADAR device bounces a radio signal of a moving vehicle, and the reflected signal is processed by a receiver. The traffic radar receiver then measures the frequency difference between the original and reflected signals, and converts it into the speed of the moving vehicle.

A LIDAR device records the time taken for a light pulse to travel from the LIDAR gun to the vehicle and return back. Based on this information, LIDAR can quickly evaluate the distance between the gun and the vehicle. LIDAR can calculate the vehicle's speed with high accuracy.

II. EXISTING SPEED DETECTION

The Existing speed detection methods advantage, Disadvantages are shown in Table 1

III. LITERATURE SURVEY

Numerous researches had been done on the topic of estimating vehicle speed based on image processing. In [5] Euclidean distance was used to measure vehicle speed and acquires the highest accuracy 90.5%. In 2016, In [6] New vehicle enhancement filters that consisted of multi scale Hessian analysis was used. After thresholding, they refine the candidate automobile detections based on evaluation of bilateral symmetry. They indicates that the proposed technique affords stepped forward detection accuracy as compared with existing automobile detection algorithms for various low-resolution aerial photos.

Table 1: Existing Methods.

Monitoring System	Information that can be collected	Advantages	Disadvantages
Computer vision	1.Vehicle detection 2.Count 3.Speed 4.Category 5.Path 6.Flowrate 7.Queue length 8.Route travel times 9.Lane changes	1)Many information can be collected 2)More than one lane can be monitored 3)Easy to install Low maintenance 4)Autonomous 5)Flexibility in changing the traffic scene 6)Information can be relayed back to authorities in real-time 7)Easy to be networked	1)Not reliable in varying lighting condition and weather 2)Field of view must be reasonably clear, free from occlusions 3)Vehicles must be separated from each other so they do not appear connected
Radar gun	Speed	Accurate speed measurement	Limited information
LIDAR	Speed	Higher Accuracy	High Operating Cost Unsuccessful during heavy rain and/or low cloud/mist.

In [1] Euclidean distance to estimate vehicle speed was used and acquired the highest accuracy of 98.12%. In [3] first the, video data is converted to frames and preprocessing is carried out to minimize shadow effect. Then, using Gaussian Mixture Model (GMM) foreground image is extracted, and it is filtered using median filter, morphology operation process and shadow removing are executed right here .The detected vehicle objects are tracked to determine the location of the vehicle in each frame so as to estimate the speed based on its distance between frames. From the analysis of results gathered, this system is capable on estimating the speed of moving vehicle within the accuracy ranging between 87.01% and 99.38%. In [2] calculated the speed by computing the movement of centroid in sequences of frames. In the calculation of speed, they considered the frames of centroid which are inside the predefined region of interest (ROI). Finally the pixel displacement is converted to a time unit of km/hour. Validation of the system is carried out by comparing the speed calculated manually and the speed obtained by the system. Here the highest accuracy is 97.52% and the lowest accuracy is 77.41%. In [4] Motion detection begins with a rough foreground / background segmentation. Vertical projection profile analyses are used to separate vehicles horizontally and region of Interest has been obtained. Vehicle speed measurements are based on the motion vectors obtained by the feature selection and tracking method.

III. PROPOSED METHODOLOGY

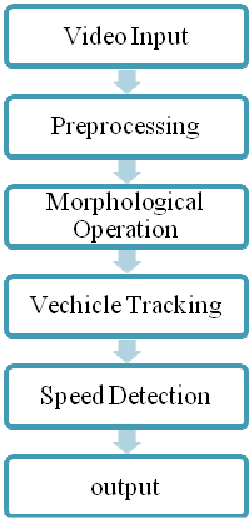


Fig. 1. Proposed Methodology.

IV. PROPOSED APPROACH

A. Video Input

The captured video is converted into frames. The video dataset are taken from the [7][8].

B. Preprocessing

The traffic images have a noise component from several interference sources.

a) Salt-and-pepper noise: It occurs when image is transmitted over a noisy channel then pixel intensity get lost at that location. Preprocessing noise removal has has been done in fig. 2.

b) RGB to BINARY Image Conversion: Each frame of the video is converted into binary image from the RGB image. It will decrease the complexity at the time of processing each frame.

C. Background Subtraction

Reference frame is a frame that does not consist of moving objects and is used to remove the background of the image which is not of our interest. Background subtraction by XOR operation or by Gaussian mixture mode is done according to our video resolution.

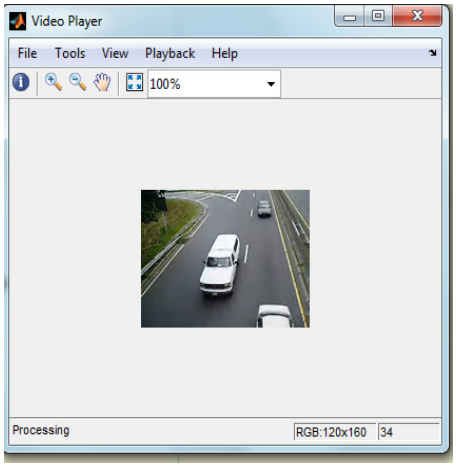


Fig. 2. Preprocessing.

In Gaussian mixture model cluster of moving pixel is formed that help us to find moving objects in the video frames. The Gaussian mixture model implementation on each frame is set to get moving objects. The Background Subtraction has been done in Fig. 3.

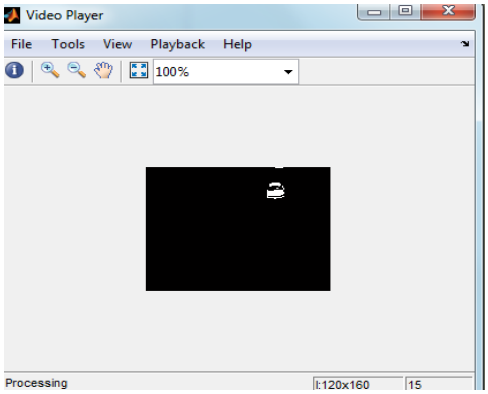


Fig. 3. Background.

Subtraction

3) Morphological Operation: Morphological image processing relies on the ordering of pixels in an image and several times is applied to binary and gray scale images. In order to remove gaps obtained along the edges, the moving edges are also enhanced. This enhancement uses the morphological operator's ***dilation*** and ***erosion***. Dilation add on the pixels to the boundaries of objects in an image, while erosion removes the pixels to the object boundaries.The number of pixels added or removed from the objects in an image relies upon the size and shape of the structuring element used to process the image. The Fig. 4 represent an image before Background Subtraction .The Background Subtraction has been done in Fig 5

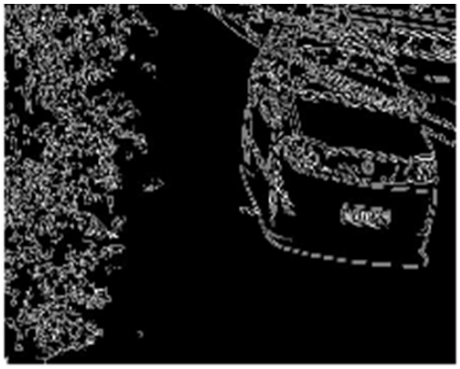


Fig. 4. Before morphing.

Fig. 5. After morphing.

4) Vehicle Tracking: For the vehicle tracking cluster of the moving pixels are found. After getting cluster of moving pixels it surround that moving cluster into bounding box. By this each moving cluster gets bounding box to each moving vehicle. After getting bounding box for each vehicle in each frame the centroid is generated for every bounding box. Centroid will help us to refer each vehicle. The Vehicle Tracking has been done in Fig. 6

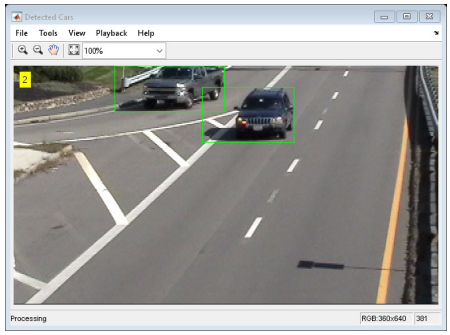


Fig. 6. Vehicle Tracking.

5) Speed Estimation: Here, the speed is calculated by measuring the distance covered by the cars from one frame to another by using the formula

$$Speed = \frac{distance\ covered\ in\ unit\ of\ pixels}{\left(\frac{1}{15}\right)}$$

As the captured video has frame rate of 15 frames per second, rate of change of one frame to another consecutive frame is 1/15 seconds.

$$Speed = \frac{distance\ covered\ in\ unit\ of\ pixels}{\left(\frac{1}{15}\right)}$$

And the unit of speed is number of pixels transferred per second. The Speed calculation has been done displayed in Fig. 7.

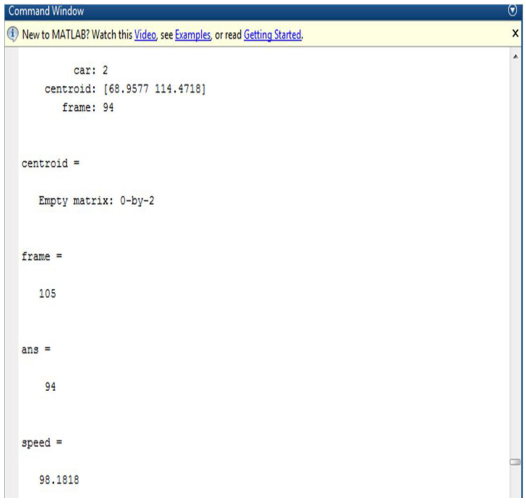


Fig. 7. Speed Estimation.

6) Output: The calculated speed will be classified into 3 types low, medium and fast according to the distance covered. If the distance covered is less than 10, the speed will be displayed as slow. If the distance covered is between 10 and 20 the speed will be classified as medium. If it is greater than 20, the speed will be displayed as fast.

Table 2: Vehicle speed Classification.

Vehicle No	Classification
1	Fast
2	Slow
3	Medium
4	Slow
5	Fast

The classification is shown in Table 1. The Figure 8 classify the speed of the vehicle.

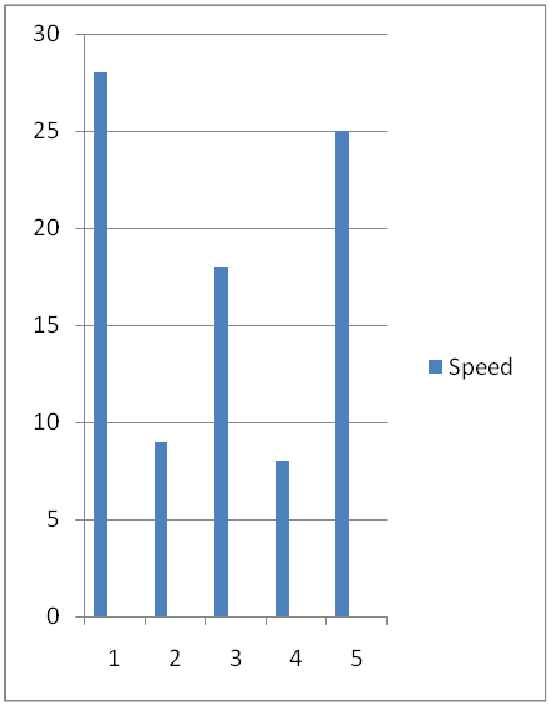


Fig. 8. Classify the Speed of the Vehicle.

V. LIMITATIONS

The main problem of the these systems is under bad weather such as heavy fog, susceptible illumination& night scenes front mirror glare, it produces inaccurate detection of vehicles . As a result bounding box will not be produced for consecutive frames and if vehicle is not predictable by their bounding box then it is not possible to calculate their speed.

VI. CONCLUSION

The moving vehicle detection, tracking and its speed measurement system had gained additional importance at present transport system. Considering the restrictions of the present systems like noise and illumination sensitivity, we have used background subtraction for noise removal in this paper. After subtracting the background image further the enhanced thinning and dilation based morphological process has made in the proposed system to get robust and accurate results. Later the vehicle is detected by a bounding box. This paper is focusing on vehicle detection and tracing on a single lane .In future it could be advanced for multi–lane machine. in addition automobile category can also be done

REFERENCES

[1]. Asif Khan, Imran Ansari" Speed Estimation of Vehicle in Intelligent Traffic Surveillance System Using Video Image Processing", *In International Journal of Scientific and Engineering Research,* Volume **5**, Issue 12, 1384 ISSN 2229-5518December 2014,

[2]. Budi Setiyono, Dwi Ratna Sulistyaningrum, Soetrisno, Farah Fajriyah, and Danang Wahyu Wicaksono" Vehicle speed detection based on Gaussian mixture model using sequential of images" *IOP Conf. Series: Journal of Physics: Conf. Series* 890 (2017) 012144 doi :10.1088/1742-6596/890/1/012144

[3]. Danang Wahyu Wicaksono and Budi Setiyono "Speed Estimation on Moving Vehicle Based on Digital Image Processing**.** *"International journal of Computing and Applied Mathematics,* VOL. **3**, NO. 1, FEBRUARY 2017.

[4]. Diogo Carbonera Luvizon, Bogdan Tomoyuki Nassu, and Rodrigo Minetto**"** A Video-Based System for Vehicle Speed Measurement in Urban Roadways" *IEEE transactions on intelligent transportation systems,* VOL. **18**, NO. 6, JUNE 2017

[5] Hardy Santosa and Agus Harjoko" Vehicle Counting and Vehicle Speed Measurement Based On Video Processing", *Journal of Theoretical and Applied Information Technology 20th February 2016.* Vol. **84**. No.2, ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195.

[6] Sundaresh Ram, Jeffrey J. Rodriguez "automated method for detecting vehicles of varying sizes in low-resolution aerial imagery", *IEEE International Conference on Image Processing, Electronic* ISBN**:** 978-1-4673-9961-6,201

Website Reference:

[7]https://github.com/gustavovelascoh/traffic-surveillance-dataset

[8].

http://www.polymtl.ca/wikitransport/index.php?title=Videobased_transportation_data_collection

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 91-94(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Study on Energy Optimization through Bio Inspired Algorithms in Wireless Sensor Networks

Mrs. E.S. Rajarajeswari[1] and Dr. B. Kalpana[2]
[1]Ph.D. Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India

ABSTRACT: The day today development of the Wireless Sensor Networks (WSNs) which are the networks of tiny nodes used to analyze the target area. Usually the WSNS meet the challenges that arise from communication link failures, energy consumption and the memory constraints. After the researches meet the saturation point in the classical routing protocol the researchers are now looking into the Bio Inspired Algorithms to optimize the arising issues in WSN. In this paper some of the biologically inspired optimization algorithms like Particle swarm optimization, Bee Colony Optimization, Ant Colony Optimization and Magnetotactic Algorithms are discussed. Recent research techniques show the efficient use of the Bio Inspires algorithms in the field of Wireless Sensor Networks. Among the various parameters applied Batteries which stand for long term became the main issue and the concentrated field.

Keywords: Wireless Sensor Network, Optimization, Bio Inspired Algorithm, PSO, ACO, Bee Sensor, MBOA.

I. INTRODUCTION

The Wireless sensor network (WSN) is termed as a numerous arrangement of sensor nodes that are deployed in equal distance in the regions and they are capable of organizing on their own. This WSN is not only self organized but also they are capable to complete the work for performing the general functionality.

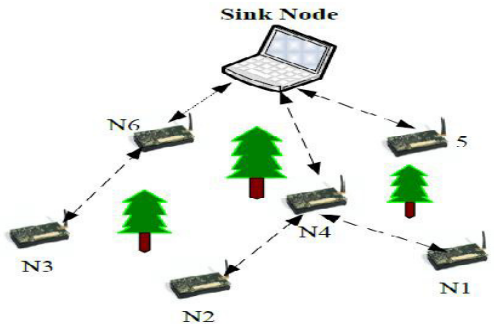


Fig. 1. Architecture of a general wireless sensor network.

Wireless Sensor Network are built with a hundreds of nodes and each and every node is comprise with a radio transceiver system with an antenna, a microcontroller, an electronic circuit which can be used to interface with the sensors and the sink node. The applications of WSN can be classified as Monitoring and Tracking. The monitoring is used for the various applications like animal monitoring, industrial monitoring, environmental monitoring and patient monitoring. The tracking application includes the tracking of the enemies in the army applications and also security applications [1]. The recent development of the wireless sensor networks has extended the applications in the field of Civil and the Disaster Management. The transmission of data, controlling of devices and the gathering of data are embedded inside the sensors and designed as built-in devices.

A. Challenging Design for WSN

The Wireless Sensor Networks are distinguished from the traditional AdHoc networks through many features. The specific data flow patterns such as single casting and multicasting are used wherever necessary [1]. The next point is WSNs are comprised of limited energy consuming battery, lesser bandwidth links, tiny memory and minimum processors. The last point to be considered is the WSNS can be deployed in large amount wherever they are necessary [6]. The communication of data between the source node and the sink node uses a large amount of energy, and the bandwidth required.

There are two methods to be followed for Energy conservation techniques in WSNs. First point is to lessen the communication frequency between the sensor nodes or to lessen the measure of data communication between the various nodes. The most appropriate method must be selected to use the lesser energy for the communication between the source and the sink nodes [3]. The most

significant issue in the WSNs is to control the utilization and lifetime of the system. This problem can be handled by the cluster head selection which affects energy utilization. A random research algorithm is proposed to limit the energy of the system and hence to improve the life time of the battery. This need has led researchers to limit energy consumption [8].

More number of researches in the field of energy conservation led are classification, routing methods including energy efficiency, protocol overhead decrease, data aggregation and cross layering components.

B. Need for Optimization in WSN

The Optimization process is defined as the methodology which is chosen to obtain the best results under a given condition considering the number of methods. In the field of networking, number of optimization methods is proposed to achieve the desired goals of the author. The designing of Wireless Sensor Network should consider the parameters like application requirement, energy efficient batteries and cost. Both hardware and the software requirements are needed to meet out the challenges. From the different techniques available in the optimization problems it is highly risk to choose and implement the proper Optimization technique.

Most number of the sensor networks is deployed either inside the target area or the area very close to the target. Each and every time the batteries cannot be taken out for the purpose of recharging or replacing. This will lead to the failure of the network or to the task which has to be accomplished. Apart from the Energy efficient nodes of the WSNs, the other requirements which are mostly needed are high quality QoS, low bandwidth, limited processing and storages in sensor nodes [9].

Due to the changes in the environmental conditions the WSNs must be designed with the ability to work autonomously, and with the scalability which provides the changing operations for the long life of the sensor network [7].

II. METHODS TO BIO INSPIRED COMPUTING

Sensor networks along with the bio inspired systems are in need to change over themselves for the changing environmental condition which includes the ability to organize on their own, the scalability and for the longer life period of the sensor network [2].

The survey papers included in this paper give a brief review on the current bio inspired thinking about the networking and also focus on the main issue of WSN which is so called Energy Efficiency.

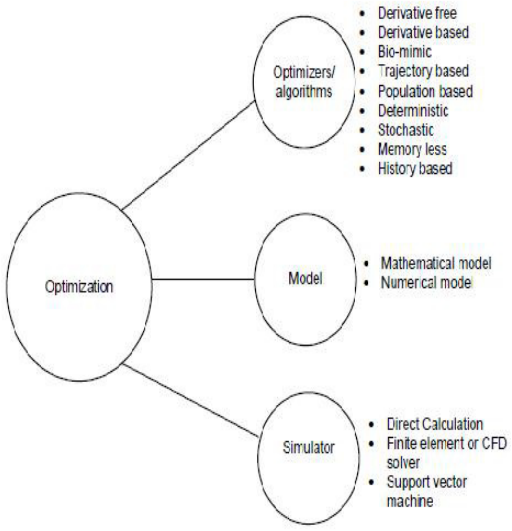


Fig. 2. A simple optimization process.

A. Particle Swarm Optimization

Vimalarani, C. *et al.,* proposed an Enhanced PSO-Based Clustering Energy Optimization (EPSO-CEO) [14] algorithm for Wireless Sensor Network. To decrease the power consumption of the WSN the clustering and clustering head selection are selected using the Particle Swarm Optimization. The results and the performance of the metrics selected are compared with the latest clustering algorithms and it is proved that the energy consumption is efficient.

Network Energy Model. Vimalarani, C. *et al.* proposed the work which simulated the WSN consisting of "*n*" number of sensor nodes which are arranged for the applications where the temperature are monitored using rectangular sensor network. The arrangements of the nodes [4] are made based on the following assumptions given in the following,

(i) The nodes are considered to be static after the arrangements are done.

(ii) The types of sensor nodes are of two: first type of node is used to sense the temperature monitoring environment and the second type is used as the sink or base station which is centered in the sensor network.

(iii) A significant identification (ID) and an appropriate primary energy are initially assigned to the sensor nodes.

(iv) If the target node is placed in a very remote place the other nodes are used for different levels of the transmission power.

(v) The cluster head gathers the sampling data from sensors as request messages in the form of packets. These request messages are sent from the BT.

(vi) Links are symmetric.

The energy model which is designed in the physical layer [4] of WSN is used for the calculation of loss of energy loss in the sensor nodes while passing the messages to the other senor nodes. Two channel propagation models used are the free space ($d2$ power loss) for the purpose of one-hop or direct transmission and the multipath fading channel (d4 power loss) for packet transmission via multihop. Thus the energy exhausted for this kind of transmission of an l-bit packet over distance d is calculated as

$$E\text{TX}(l, d) = \begin{cases} lE\text{elec} + l\varepsilon\text{fs}d^2, & d<d_0, \\ lE\text{elec} + l\varepsilon\text{mp}d^4, & d\geq d_0, \end{cases}$$

where εfs is free space energy loss, εmp is multipath energy loss, d is distance between source node and destination node, and d_0 is crossover distance:

$$d_0 = \frac{\sqrt{\varepsilon\text{fs}}}{\varepsilon\text{mp}}$$

The energy spent for the radio to receive this message is $E_{\text{RX}}(l) = lE_{\text{elec}}$.

Thus the transmission power and receiving power energy levels are designed in physical and Mac layer of the WSN[11].

B. Ant Colony Optimization

Peng Li *et, al,* proposed the ant colony algorithm [12]. He discussed that the current node finds the next node according to the pheromone concentration of the path and the node's energy. The pheromone concentration of a certain path increases significantly with more ants passing this path. However, when the pheromone is at a high level, the energy of the node in the path will decrease rapidly. Suppose that there are two paths, one with high pheromone concentration and low residual energy of node, while the second one with low pheromone concentration and high residual energy, according to the traditional ant colony algorithm, ants will continue choosing the path with high pheromone Concentration and the energy of the node on the path will become very low.

The ants will choose the path with high node residual energy even if the pheromone concentration of that path is relatively low. Therefore, this is not only conducive to ensure the stability of the WSN but also guarantees the energy consumption equilibrium of the node in the WSN. The author proposes that when the node energy decreases to a certain value, the pheromone volatilization rate of the current path should be increased. When other ants pass this path, the pheromone concentration will increase. Additionally, the node notifies its neighbor node of its residual energy when it is lower than a threshold. So, this node has no chance to be selected by its neighbor nodes as the next hop node.

C. Bee Sensor

Saleem M. Ullah *et al,* the bee-inspired Bee-Sensor protocol that is energy-aware, scalable and efficient is proposed. The important contribution of this work is a three phase protocol design strategy: (1) The first step is to take inspiration from biological systems to develop a distributed, decentralized and simple routing protocol, (2) Then it is formally modeled with the important performance metrics of the protocol to get an analytic insight into its behavior, and (3) the improvement of the protocol on the basis of our analysis in phase 2. The results of our experiments demonstrate the utility of this three phase protocol engineering, which helped BeeSensor in achieving the best performance with the least communication and processing costs – two main sources of energy consumption in sensor networks – as compared to other SI based WSN routing protocols.

D. BeeSensor: the concept-mapping

In this section, the concept-mapping of BeeSensor algorithm is described in the following steps.

1. Each sensor node has a software module – called hive – that houses different types of bee agents: packers, scouts, foragers and swarms. Packers reside inside a hive and process incoming data packets from upper layers. Packers in the hive of a source node launch scouts to find paths to a new sink node. Scouts also evaluate the quality of paths between a source node and a sink node. Once the scouts return to the source node, they recruit foragers for their paths. Foragers undertake two tasks: (1) transporting data packets, and (2) evaluating the quality of the visited paths. The quality of a given path is a function of the path length and the remaining battery levels of the sensor nodes. A scout/forager simulates ''recruiting through waggle dance'' by cloning itself a certain number of times depending on the quality of a path.
2. The number of cloned foragers would be large in two scenarios: (1) the path is shorter and the nodes that constitute it has a good amount of spare battery capacity, which means that this is a route eligible to be exploited, and (2) a large number of packers are waiting for the forager, so that the route needs to be exploited even though it might contain nodes with marginal battery capacity. On the other hand, if no data packets are waiting to be transported, then a forager with a very good route might even abstain from cloning itself because its fellow foragers are doing a good job in transporting data packets.
3. The majority of scouts in BeeSensor explore the network in the neighborhood (Hl hops) of their source nodes and only few of them are allowed to do exploration beyond Hl hops. As a consequence of this condition, we achieve three benefits: (1) substantial reduction in the routing overhead, (2) energy-efficient route discovery and sampling, and (3) relatively small routing and forwarding tables.
4. A packer stochastically selects a forager – in case foragers of different paths are available – at the source

node depending on the quality (eligibility, relative goodness) of the paths. Consequently, data packets are distributed on multiple paths. This not only helps in maximizing the performance by avoiding congestion on high quality paths but also enhances the lifetime of the network.

D. Magnetotactic Bacteria Algorithm

To add one more step in the field of research towards the Bio Inspired based optimization techniques for the energy efficiency of Wireless Sensor Networks, is the Magnetotactic Bacteria Optimization Algorithm which is now on the hook. MBOA is a new intelligent optimization Algorithm [10]. The algorithm inspired by the biological behavior of magnetic bacteria, which can move along the magnetic field lines. It matches the optimization problem with magnetic bacterial biological characteristics, finding the optimal solution. MBOA shows better performance and good potential ability in solving energy optimization in WSN.

III. CONCLUSION AND FUTURE WORK

The design of Wireless Sensor Networks should be considered with the improvement of the utilization of the limited resources of energy, bandwidth and the computational power. The development of highly effective optimization algorithm must focus on the above parameters. On the focus a huge number of bio inspired optimization techniques for the energy efficient routing of the wireless sensor networks have been addressed.

This paper focuses on the survey of the Bio Inspired Energy Efficient Routing algorithms for Wireless Sensor Networks like Particle Swarm Optimization, Ant Colony Optimization, Bee Colony and the hint of Magnetotactic Bacteria Algorithm which is on the way.

The Future Enhancement of the WSN can be focused by integrating various possible parameters QoS with Energy Efficiency and also the Security issues.

REFERENCES

[1]. C. Raghavendra, K. Krishna, T. Znati, "Wireless Sensor Networks", *Springer-Verlag*, 2004.

[2]. Dorigo M, Colorni A, V Maniezzo "Positive feedback as a search strategy", Technical Report, pp.91-016, 1991.

[3]. I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "A survey on sensor networks", *IEEE Communications Magazine*, pp.102–114, 2002.

[4]. J. Wang, X. Yang, T. Ma, M. Wu, and J.-U. Kim, "An energy efficient competitive clustering algorithm for wireless sensor networks using mobile sink," *International Journal of Grid and Distributed Computing*, vol. **5**, no. 4, pp. 79–92, 2012.

[5]. Jennifer Yick, Biswanth Mukherjee, Dipak Ghosal, "Wireless Sensor network survey", *International Journal of Computer networks*, Vol. **52**, pp. 2292- 2330, 2008.

[6]. Kemal Akkaya, Mohamed Younis "A survey on routing protocols for wireless sensor networks", *IEEE Communication Magazine on AdHoc Networks*, pp.325–349, 2005.

[7]. Kennedy, J.; Eberhart, R.C.; Shi, Y, "Swarm Intelligence", *Morgan Kaufmann Publishers,*San Francisco, 2001.

[8]. Lu, G., Krishnamachari, B. and Raghavendra, C.S., 2004, April. An adaptive energy-efficient and low-latency MAC for data gathering in wireless sensor networks. In Parallel and Distributed Processing Symposium, 2004. *Proceedings. 18th International (p. 224). IEEE.*

[9]. Md. Akhtaruzzaman Adnan, Mohammd Abdur Razzaque, Ishtiaque Ahmed, Ismail Fauzi Is nin, "Bio-mimic Optimization Strategies in Wireless Sensor Networks: A Survey", *Sensors*, Vol. **14**, pp. 299-235, 2014.

[10]. Mo, H. and Xu, L., 2012. Magnetotactic bacteria algorithm for function optimization. Journal of Software Engineering and Applications, **5**, p.66.

[11]. N. A. B. A. Aziz, A. W. Mohemmed, and B. S. Daya Sagar, "Particle swarm optimization and Voronoi diagram for wireless sensor networks coverage optimization," in *Proceedings of the International Conference on Intelligent and Advanced Systems (ICIAS '07)*, pp. 961–965, *IEEE, Kuala Lumpur, Malaysia*, November 2007.

[12]. Peng Li, Huqing Nie, Lingfing Qiu and Ruchwan Wang "Energy Optimization of Ant Colony Algorithm in Wireless Sensor Network", *International Journal of Distributed Sensor Networks, 2017*, Vol. **13**(4).

[13]. Saleem, M., Ullah, I., and Farooq, M. (2012). Bee Sensor: An energy-efficient and scalable routing protocol for wireless sensor networks. *Information Sciences,* **200**, 38-56.

[14]. Vimalarani, C., Subramanian, R. and Sivanandam, S.N., 2016. An enhanced PSO-based clustering energy optimization algorithm for wireless sensor network. *The Scientific World Journal, 2016.*

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 95-97(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Survey on Ovarian Cancer Detection using Data Mining Techniques

Pillai Honey Nagarajan[1] and Dr. N. Tajunisha[2]
[1]Ph.D. Research Scholar, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore (TN), India.
[2]Associate Professor, Department of Computer Science,
Sri Ramakrishna College of Arts and Science for Women, Coimbatore (TN), India.

ABSTRACT: The Ovarian cancer is one of the most critical cancer types that found in various ages of women. Significant number of researches is on process to predict and classify the stages of the ovarian cancer. The primary focus of this paper is to survey the cancer symptoms recognition based on prediction strategies at initial stage and to classify disease based stages. To handle the implication, lots of techniques has been proposed for the cancer detection and prediction. This Classification models enables the different kind of treatment procedures either with surgery, radiotherapy or chemotherapy. The detection of disease and classification of stage eases the treatment of patient with minimum risk and fewer side effects.

Keywords: Ovarian Cancer, Metastatic Tumor Detection, CT images, Data Mining.

I. INTRODUCTION

Ovarian cancer is the 2nd familiar cancer amongst gynecologic malignant and has the largest mortality rate [1]. Treatments for ovarian cancer are becoming better, and the best outcomes are always seen when the cancer is found earlier. Due to the deficit of an effective early cancer screening and detection method, more than 70% of ovarian cancers are diagnosed at an adduced stage (III or IV) with tumor metastasis to other organs. The current 5-year survival rate for the adduced ovarian cancer patients are less than 30%. [2]. Cancer progression can be quite different between patients, yet the patients are often treated with the same therapeutic regimen to improve the efficiency of medical diagnosis, Due to the heterogeneity of the ovarian cancer cases, identifying and applying effective therapeutic drugs to the individual patients becomes an important issue to increase patients' progression. Thus there are different data mining techniques used to detect as well as predict the cancer at earlier stage.

II. OVARIAN CANCER

There are five main subtypes of ovarian cancer, of which high-grade serous carcinoma is the most familiar. These tumors are believed to start in the cells covering the ovaries, though some may form at the Fallopian tubes. Less familiar types of ovarian cancer include germ cell tumors and sex cord stromal tumors.

A. Staging of ovarian cancer

It refers to the reach to which it has spread to other organs or tissues. This is typically estimated during surgery. Stages of ovarian cancer are as follows:

Stage I: The cancer is confined to the ovaries
Stage II: The cancer reach to the uterus or other pelvic organs
Stage III: The cancer reach to lymph nodes or lining tissues of the abdomen
Stage IV: The cancer reaches to distant sites, like the liver or lungs.

B. Diagnosing Ovarian Cancer

Image testing like Computed Tomography, Magnetic Resonance Imaging, or UTZ can impart an ovarian mass, but only a sampling of the tissue can resolve whether the mass is cancerous. A biopsy is analyzed in a laboratory to resolve whether the ovarian mass biopsied is due to cancer or not.

C. Screening Tests

Screening test for ovarian cancer in its initial stages is one is the UTZ scan of the ovaries and second is the measuring levels of protein in the blood. Else of these techniques are used to save lives to test female of average risk. Therefore, screening is currently recommended only for female at larger risk [2].

III. REVIEW OF LITERATURES

There are different techniques and methods proposed by various authors for the prediction and detection of cancer regions. Every technique has its own merits and demerits. Some of the existing techniques and methods are given. L.R. Folio *et al.* [3] In this paper the author have defined the problem as to evaluate a new software capability that integrates registration, segmentation and tumors measurement across serial exams within a picture archiving communication system (PACS) to expedite tumors measurement using Bland-Altman plots using the lung and Liver datasets, demonstrated 95% confidence interval of ±0.7 cm when comparing with the software-generated and manual RECIST measurements.

M. F. McNitt-Gray, *et al,* [4], In this paper the author has defined the problem as Issues and methods that are specific to the measurement of change in tumor volume as measured from computed topographic (CT) images and how these would relate to the establishment of CT tumor volumetric as a biomarker of patient response to therapy using the Bias and Variance a regression model with the dataset of lung. One limitation of this work is that to understand the sources of bias and variance to be able to measure their effects on tumor volumes, this work is not sufficient in itself to establish CT tumor volumetric as a biomarker.

Lin Zhang, *et al,* [5]. In this paper the author compares the gene expression profiles between healthy and diseased tissues and between patients having different responses to the same drug therapy using GIREN Method (Gene Interaction Regularized Elastic Net) even this is an Regression model and he have used the Ovarian and Breast cancer data sets. The limitation of the proposed methods is mainly the computation load. The algorithm adopts an improved "elastic net penalty" regularized by gene-gene interactions, which cannot be solved by the classic fast algorithm of elastic net. The newly developed iterative gradient descent algorithm suffers from a relative heavy computation load (almost 40 times of that of elastic net), and cannot directly apply to large genome-scale dataset, thus a feature selection step is required.

Andrew Janowczyk, *et al,* [6]. In this paper they define the problem as to quantify the extent of the vascular staining on ovarian cancer tissue microarrays using Mean-shift algorithm (MS). Normalized cuts algorithm, Segmentation, Tissue Microarray (TMA) methods with the Ovarian Cancer datasets. Since HNCut operates in the color space, and is thus largely efficient, the only limitation in this paper is the size of the image that can be analyzed by HNCut is the amount of computer memory available to read in the image data. In future work, the author has intended to explore the applicability of HN Cut to other biomarker quantification and digital pathology problems.

R. E. Bristow, *et al,.[7]* In this paper the author analysis Forty-one women with a preoperative Computed Tomography (CT) scan of the abdomen and pelvis and a histological diagnosis of Stage III or IV epithelial ovarian carcinoma following primary surgery performed by one of nine gynecologic oncologists were identified from tumor registry databases using the Receiver operating characteristic (ROC) curve analysis with the ovarian cancer datasets. Maxine Tan*, *et al*, [8] In this paper author describes the Response Evaluation Criteria in the Solid tumor is the current guideline to access the size change after Radiotherapy and Chemotherapy procedures. Current Clinical trials related to tumor results in the progression free survival of the patient. Progression Free Survival which has less association Response Evaluation Criteria with current clinical trials. It acts as major problem in accessing the response of the treatment using the B-Spline (Regression) method with Ovarian and Breast at asets. This paper presented a new approach based on a multire solution B-Spline deformable image registration method to automatically track and predict the response of the metastatic tumors in ovarian cancer patients to chemotherapy using two sets of CT images acquired pre- and post-treatment. The new method has a number of advantages comparing to using the RECIST guidelines.

Table 1: Analysis.

Reference	Author	Methodology	Datasets
[5]	Lin Zhang *et al.,*	GIREN Method (Gene Interaction Regularized Elastic Net)	Ovarian and Breast cancer data sets
[6]	Andrew Janowczyk *et al.,*	1. Mean-shift algorithm (MS). 2. Normalized cuts algorithm. 3. Segmentation, Tissue Microarray (TMA).	Ovarian Cancer
[7]	R. E. Bristow *et al.,*	Receiver operating characteristic (ROC) curve analysis	Ovarian dataset
[2]	L.R. Folio, *et al.,*	RECIST measurements using Bland-Altman plots.	Lung and Liver
[8]	Maxine Tan *et al.,*	B-Spline Curve (Regression)	Ovarian and Breast cancer datasets

IV. CONCLUSION

In this survey, a systematic review of the ovarian cancer detection, prediction and progression is presented with data mining techniques used, in order to predict the cancer at earlier stage. Also helps researchers to gain knowledge on this and by applying different techniques the future works can be done.

REFERENCES

[1]. American Cancer Society, “What are the key statistics about ovarian cancer?" http://www.cancer.org/cancer/ovariancancer/detailedguide/ovarian-cancer-key-statistics, 2014.

[2]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2916206/

[3]. L.R. Folio, M.M. Choi, J.M. Solomon and N.P. Schaub "Automated, Registration Segmentations and Measurement of Metastatic Melanoma Tumors in Serial CT scans".

[4]. M. F. McNitt-Gray, L. M. Bidaut, S. G. Armato, C. R. Meyer, M. A. Gavrielides, C. Fenimore, G. McLennan, N. Petrick, B. Zhao, A. P. Reeves, R. Beichel, H. J. Kim, and L. Kinnard, "Computed tomography assessment of response to therapy tumor volume change measurement, truth data, and error," Volume **2** Number 4 December 2009.

[5]. Lin Zhang, Hui Liu, Yufei Huang, Xuesong Wang, Yidong Chen and Jia Meng*, "Cancer Progression Prediction Using Gene Interaction Regularized Elastic Net" 1545-5963 (c) 2015 IEEE.

[6]. Andrew Janowczyk, Sharat Chandran, Rajendra Singh, Dimitra Sasaroli, George Coukos, Michael D. Feldman, and Anant Madabhushi "High-Throughput Biomarker Segmentation on Ovarian Cancer Tissue Microarrays via Hierarchical Normalized Cuts" *IEEE Transactions On Biomedical Engineering,* Vol. **59**, No. 5, May 2012.

[7]. R. E. Bristow, L.R. Duska, N. C. Lambrou, E. K. Fishman, M. J. O'Neill, E. L. Trimble, and F. J. Montz, “"A model for predicting surgical outcome in patients with advanced ovarian carcinoma using computed tomography".

[8]. Maxine Tan, Zheng Li, Yuchen Qiu, Scott D. McMeekin, Theresa C. Thai, Kai Ding, Kathleen N. Moore, Hong Liu, and Bin, “A New Approach to Evaluate Drug Treatment Response of Ovarian Cancer Patients Based on Deformable Image Registration” 10.1109/TMI.2015.2473823, *IEEE Transactions on Medical Imaging*.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 98-102(2018)
International Conference on SMAC - Reshaping the Future (February 2018)
ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Study and Analysis of Water Audit for Domestic Household an IoT Based Prototype Model

M. Gracelin[1], M. Lissa[1] and V. Bhuvaneswari[2]
[1]MCA Student, Department of Computer Applications, Bharathiar University (TN), India
[2]Asst. Professor, Department of Computer Applications Bharathiar University (TN), India

ABSTRACT: Water has become a precious resource in the current scenario which has to be preserved where various water preservation methods are discussed globally. Water audit usage is used to analyse the usage of water for prservation. In this work water audit is carried out for analysis. by collecting data from real time water usage of domestic households. The data related to water usage is collected from hostels, individual household, apartments. A detailed analysis of water usage of kitchen and toilets are analysed with respect to water flow equipments. The experimental results found that more usage of water can be preserved when replaced with low flow equipment which accounts to 571 gallons of water per household. Based on this study, an IoT based water audit framework is presented to monitor the flow from pipes and alert household when maximum water usage exceeds and also mechanism is devised to provide data on accountability of each household usage.

Keywords: Water Audit, IoT, Low cost devices, data analysis

I. INTRODUCTION

Water is the most important natural resource for living organisms in the globe. In the current scenario water has become a precious resource and evolved as important commodity. Scarcity of water is discussed globally and nationally and also predicted to be severe in the coming years, due to the various reasons such as global warming, climate change, pollution and wastage. In order to preserve water it becomes important to measure the quantity of water used and reused. Various water preserving methods are available but the audit of water usage is extremely important to preserve and save water.

It is stated that the average family waste 180 gallons water per week and 9,400 gallons of water annually from household usage leaks which is equivalent to the amount of water needed to wash more than 300 loads of laundry. Household leaks results in water waste approximating nearly 900 billion gallons of water annually which is equal to annual household water use of nearly 11 million homes [1].

A. Water Usage Statistics

Water usage statistic from literature related to leakage of households equipments are discussed. The running of the dishwasher with full load will eliminate and save the average family nearly 320 gallons of water annually [1]. The pipes when running for five minutes for washing dishes result in wastage of 10 gallons of water [2]. Turning off the tap when brushing teeth saves 8 gallons of water per day. Brushing of teeth daily and cleaning toilets 5 times per week, save nearly 5700 gallons of water per year. It is also stated in literature that the outdoor areas for gardening and other usage requires 30 percent of total water of household. Mostly 50 percent of the water used in outdoors are lost due to wind, evaporation, and runoff caused by inefficient irrigation methods. A household with an automatic landscape irrigation system that is not properly maintained and operated results in waste up to 25,000 gallons of water annually [3]. From the usage analysis of water, we have proposed smart water audit mechanism using an IoT model.

B. Internet of Things Vision

Internet of Things (IoT) is a new technology which is used to connect physical objects to digital gadgets such as cloud, mobile, through communiction medium with active sensors for creating smarter objects around the globe. Various applications are created with specific to IoT. This section provides with an overview of IoT definitions and its features [5].

Internet of things are viewed in three paradigms: Internet oriented vision (middleware), Things oriented (sensors) vision and Semantic oriented(knowledge) vision Internet of thing has become as one of the popular technologies in ICT. Wireless sensor networks and ubiquitous computing plays an important role in IoT. Physical objects are enabled with smartness using sensors , which are made to communicate and connect through devices to connect digitally.

Table 1: IoT Elements.

Name	Features
RFID	Wireless data communications
WSN	Low power integrated circuits
Data storage and analytics	Unprecedented amount of data
Visualization	2D to 3D

Table 1 presents with the components presented in IoT. Internet-of-Things (IoT) is the convergence of

Internet with RFID, Sensor and smart objects. IoT can be defined as "things belonging to the Internet" to supply and access all of real-world information. Millions of devices and technologies are expected to be associated into the system and that shall require huge distribution of networks as well as the process of transforming raw data into meaningful inferences. IoT is the biggest promise of the technology today, but still lacking a novel mechanism and some issues which can be perceived through the lenses of things and internet vision. It is a building up of IP protocol, enabled with an internet connection to create smarter objects.

The objective of this paper is to design an IoT based architecture to minimize water usage through alerts system to reduce water usage in domestic households. The IoT model is designed based on a detailed study and analysis of water usage from domestic households of hostels, apartments and individual household of Indian Scenario. The analysis is carried out based on a detailed water audit carried out for a period of 30 days for the scenarios discussed. The analysis is taken in to consideration of the water pipes and used to accurately measure and analyse the water wasted to replace with smart automated sensors. The paper is organized as follows: Section 2 provides with a detailed discussion of water audit measures. Section 3 discusses the methodology and IoT framework for prevention of water leakages and utilization of the water resource effectively. Section 5 details the analysis of water audit use cases followed by conclusion in section 6.

II. WATER AUDIT OVERVIEW

Water audit measures are tested with a detailed study in real time scenario. The water usage is collected from various households with respect to domestic usage related to toilets, kitchen from individual households, hostels and apartments. The water equipments pipe models replaced will reduce less water leakages which is presented in the below section.

Generally water pipes are classified in two types as faucets and non faucets. Faucets have an inner valve that controls the flow of water through the spout. The valve quality, with or without a washer, determines the reliability and durability of the faucets. The categories of pipes are categorized as stainless pipe and plastic pipe. Plastic pipe is a tubular section, or hollow cylinder, made of plastic.

It is usually, but not necessarily, of circular cross-section, used mainly to convey substances which can flow liquids and it can also be used for structural applications. Steel pipes are long, hollow tubes that are used for a variety of purposes. They are produced by two distinct methods which result in either a welded or seamless pipe. In both methods, raw steel is first cast into a more workable starting form. Faucets consume less water because it pushes the air, the water level and pressure is low, and the water wastage can be controlled by the faucets.

The traditional water pipes models when replaced with water sense labeled models can reduce the average family water and electricity costs by $70 and can save the average family more than 2700 gallons of water per year, equal to the amount of water needed to wash 88 loads of laundry. The average family can save 13000 gallons of water and $130 in water costs per year by replacing all old, inefficient toilets in their home with water models. Replacing old, inefficient bathroom faucets and aerators with Water models can save the average family $250 in water and electricity costs over the faucets' lifetime [3]. Water audit helps to save money on utility bill and prevent water pollution in nearby household and apartments. An average of about 14 percent of residential water is lost through leaking fixtures or pipe. Leaking toilets are common and can be large source of water loss and that leaking toilet can waste anywhere from several gallons to more than 100 gallons per day. Water loss audits and water loss programs are effective methods of accounting for all water usage by a utility within its service area of household and in apartments. Performing a reliable water audit is the foundation of water resource management and loss control for water preservation. This paper focuses on identifying water usage in various environments and performs a water audit analysis.

A. IOT Overview

The Internet of Things refers to the ever-growing network of physical objects that feature an IP address for internet connectivity, and the communication that occurs between these objects and other Internet-enabled devices and systems. Internet of Things (IoT) is a sprawling set of technologies and use cases that has no clear, single definition. One workable view frames IoT as the use of network-connected devices, embedded in the physical environment, to improve some existing process or to enable a new scenario not previously possible.

B. IOT Communication Medium

The communication technology of IoT rapidly depends on technical innovation in various fields. Technology used to connect everyday objects and devices to large databases and networks and technology used for data collection with ability to detect changes in the physical status of objects, technology to take action through embedded intelligence in objects and system, and finally to make smaller things will have the ability to interact and connect. The combination of all these developments made the effective communications on IoT applications.

Table 2: IoT Communication Topology.

Technology	Long Range	Topology	Current Consumption (3V)	TX Current Consumption
2G	Yes	P2P	2.3 mA	200-500 mA
3G	Yes	P2P	3.5 mA	500-1000 mA
LTE	100km	P2P	5.5 mA	600-1100 mA
Wifi	No	P2P/MESH	1.1 mA	19-400 mA
Zigbee	No	MESH	0.003 mA	35 mA
Wireless Hart	No	MESH	0.008 mA	48 mA
RFID	<3	MESH	~0.005 mA	20-70 mA
NFC	<0.1	MESH	0.005 mA	35 mA

III. METHODOLOGY

Water loss audit help to find the essential and to allow utilities to identify "economically recoverable" water losses *(i.e.,* losses for which investments in corrective actions have a reasonable payback period). The household water consumption has a large potential to be reduced and their benefits of reducing domestic water consumption include lower water bills or less time spent collecting water, reduced pressure of water level. Water usage can be analyzed based on the life of water pipes.

The objective of this paper is to undergo a detailed study on usage of water in household based on the equipment used for flow of water. The data for this use case is collected directly by observing various households needs in individual houses, apartment and hostel. The framework for the proposed objective is given in Fig. 1.

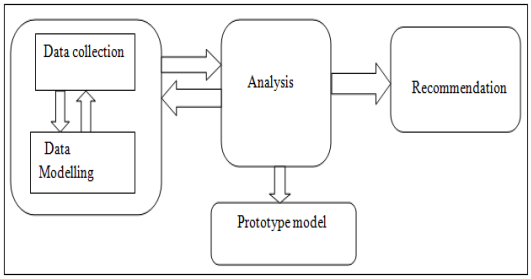


Fig. 1. Framework for Water Audit.

A. Data collection

The first phase of the framework is the data collection and data modeling. The data collected for this study is given in Table 1 and Table 2.

Table 3: Domestic Data Categories.

Type	Water Audit Usage Measure	Usage
Apartments	100 – 500 Houses	Bathing, Kitchen, Toilets
Individual Households	4 – 6	
Hostel	300 Students	

Table 4: Data: Classified Hardware Categories.

Data	Classification	Pipe	Modeling	Usage
Toilet	Indian Western Bathing	Stainless pipe	Faucets	number of usage and measure
Kitchen	Dish washing	Plastic pipe	Non faucets	number of times

B. Analysis

The data collected from various household and apartments are analyzed for period of weeks and months and water usage is based on number of member consumption. For hostels the data collected are analyzed according to the number of members. The water audit calculation measures are calculated based on the hardware pipes flow of water based on time and count. The amount of water used by faucets for each household is calculated using equation 1.

(Number of Minutes Running) x (Flow in Gallons per Minute) = Number of Gallons ……. eq (1)

The amount of water used in toilets, kitchen, and bathing for each household is calculated based on the number of users in each household which is calculated using equation 2.

(Number of Uses) x (Flow in Gallons per Use) = Number of Gallons ……. eq(2)

The detailed analysis of water audit use cases are discussed in results and discussion.

IV. IOT PROTOTYPE - WATER AUDIT

Based on the water usage analysis results , the water leakage can be prevented through automated smart approaches**.** The water usage recommendations for preserving and saving the water a technical IoT based prototype is given in Fig. 2.

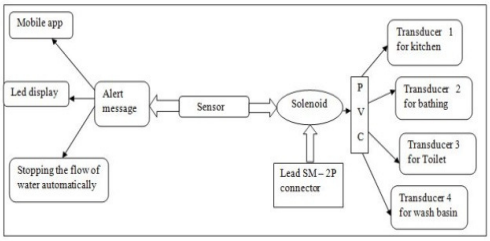


Fig. 2. IoT Water Audit Framework.

A. Working Methodology

A Technical prototype model is proposed to preserve water and study of water audit based on Apartment, Households and hostels. This prototype model is implemented with the help of Digiten sensor.

Digiten Sensor. It is easy to install and its size is small which connects to flow sensor. The flow of water is monitored by the digiten it is a flow water sensor meter with a digital LCD display of quantitative control of gallons liters per minute. The digiten controls the quantitative flow of water it monitors the real time flow rate. It senses the rate with the ultrasonic detector. A digiten is installed near the tank of water which is directly connected to the solenoid valve it is an electric flux contains inlet flow switch that passes the data to sensor meter.

Solenoid. Electric solenoid valve normally closed when it reaches the limit this is achieved through the adapter power with wire lead SM-2P connector. Each pipe of all households is connected to the value of solenoid that controls the flow rate and sends the current status of data to digital meter.

E-Water Audit Usage Data Model. The real time flow rate of water data is stored into the cloud to give automatic alert messages to mobile phones, LED display of sensor meter and automatically stops the flow when it exceeds the limit of preset value. All flow rate data stored into the cloud is taken for data analytics to take the decision on preserving the water systematically. Alerts can also be sent through an water audit app to individual household when the water usage exceeds the threshold.

V. RESULTS AND DISCUSSIONS

A. Water Usage Analysis: Toilets

Water audit analysis for toilets is calculated considering five flush as an average for every individual member during week days. The flush of water unit varies with respect to Indian and western toilets. The flow of water is dependent on pipe categories and dimensions. In this work the study of water usage is calculated based on the criteria as given in Formula 1 for household, apartments and hostel. The water usage audit based on the number of members and flushes for the toilets given in the table 5 can be inferred that the pipe categories faucets consume less water flow compared to the normal non faucet pipes.

Table 5: Water Flow Pipe, Toilet Categories: Individual household, Hostel, Apartments.

Type	Members	Western(faucets)	Indian(non faucets)
Individual household	4	13.7777	8.74125
	5	15.8966	12.5241
	6	17.8933	14.5633
Hostel	100	1155.89	1355.10
	200	1345.56	1455.11
	300	1544.84	1647.05
Apartment	100	1877.45	2578.12
	200	2500.74	3758.41
	500	3,990.14	4,585.14

There is an average of 5.28 % loss of flow of water when non faucets pipe types are used. This counts for 528 gallons of water per week for 200 members for an average of 45 household in apartments. For a month it accounts to 2112 gallons of water which measures 576 liters.

B. Water Usage Analysis: Bathing

Water audit analysis for bathing is calculated considering number of usage time as 10, 15 and 20 minutes and they use as an average of every individual during week days. The bucket of water unit varies with respect to pipe inches in bathroom.

Table 6: Water Flow Pipe, Bathing Categories: Individual household, Hostel, Apartments.

Type	Members	Western(faucets)	Indian(non faucets)
Apartments	100	10,500.74	8,500
	200	12,897.12	11,438.50
	300	14,236.23	15.123.80
	400	18,145.78	20,456.78
	500	20,500.45	22,636.20
Hostel	100	1155.89	1355.10
	200	1345.56	1455.11
	300	1544.84	1647.05
Individual household	4	1245.96	1577.52
	5	1445.78	1788.77
	6	1654.86	1966.85

The number of water flow in bucket also dependent on categories. Water usage calculation is based on the criteria using formula 2 for Apartments, household and hostels. The water usage audit based on the number of members and times during for kitchen is based on the pipe categories. From the table 6 and it can be found that the pipe categories faucets consume less water flow compared to normal non faucet pipes. There is an average of 4 % loss flow of water when non faucets pipe types are used. This counts for 850 gallons of water per week for 200 members for an average of 45 household in an apartment. For a month it accounts to 3400 gallons of water which measures

898 liters. Overall analysis it is found that more water flow occurs in western toilets. The pipe categories faucet reduces the flow of water in every minute which saves water for apartments, household, and kitchen.

Low Flow Scenario For Household

In each household, the changing of existing fixtures for low-flow alternatives with smart sensor prototype given will reduce significant water [3].

-Household 1 currently uses 531 gallons per week which can be reduced to 467 gallons per week which measures 123.28 liters when an automated sensor model is proposed.

Total Water Saving Prediction

(Household 1) = 208 gallons/week - 354 gallons/ per week = 146 gallons which measures 38.55 liters.

Water Saved in Percentage

(Household 1) = (146 gallons/ per week) / (208 gallons/week) = 70% water saved/ per week.

-Household 2 currently uses 574 gallons per week, when replaced with low-flow alternatives this number could be reduced to 129 gallons per week which measures 34 liters. This can be further reduced to 5 percentage as an average when the propose IoT prototype is fixed in the households.

(Household 2) = 195 gallons/week - 256 gallons/week = 61 gallons per minutes which measures 16.10 liters

Water Saved in Percentage

(Household 2) = (61 gallons/week) / (195 gallons/week) =30% water will be saved per week which measures to 4488 gallons per minutes. The benefit of each house varies as some depending on the pipes and the cost for each household to convert to all low flow devices with IoT monitoring prototype reducing maximum water wastage.

VI. CONCLUSION

The paper water audit based on the analysis of domestic usage households identifies the wastage in toilet and kitchen. The data are collected from various households, hostels of Bharathiar University, apartment based on the analysis it is inferred that pipe categories faucet reduce the consumption of water. In household for an average of 571 gallons of water is wasted. The paper also provides with cost estimation in saving water when replaced with proposed IoT model for preserving and saving water. The IoT framework also provides automatic alerts to households when the water usage exceeds their maximum limits. Smart water audit mechanisms will help the households in monitoring the accurate usage and excess outflow. Hence this proposed model when implemented will help to monitor and save the precious water similar to data packets transferred in internet.

REFERENCES

[1]. American water works association, cited at http://www.awwa.org/advocacy/learn/.

[2]. Water audit for residential area, 2016 *International journal of current engineering and technology.*

[3]. Study of individual household water consumption - maize Borg, Orion Edwards Sarah kimpel.

[4]. California urban water conservation council march 2003, h2ouse water saver home, cited at http://www.h2ouse.org/.

[5]. The Internet of Things(IoT) Applications and communications Enabling Technology Standards: *An Overview 2014-International conference on Intelligent Computing Applications*-V. Bhuvaneswari, R. Porkodi.

International Journal of Theoretical & Applied Sciences, *Special Issue* 10(1): 103-107(2018)
International Conference on SMAC - Reshaping the Future (February 2018)
ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

An Improved Similarity Measurement on Web Document Clustering

Dr. M. Reka
Assistant Professor, BCA,
K.S.R College of Arts and Science for Women, Tiruchengode (Tamil Nadu), India

ABSTRACT: The World Wide Web grows at each fraction with number of documents. Such growth introduces challenge in clustering the documents. There are number of clustering algorithms has been discussed earlier but suffers to achieve clustering efficiency. To overcome the deficiency, the proposed algorithm introduced an efficient clustering algorithm which considers the relevancy of documents to be precise with internal and external documents. The method first computes the topical similarity measure with all clusters and selects a higher one. In the second stage, the method computes the internal topical similarity and external topical similarity to compute the topical weight. Based on computed topical weight the method assigns the class label for the web document. It produces higher classification accuracy with less time complexity.

Keywords: Clustering, Web Documents, Interior-Exterior Similarity, Accuracy

I. INTRODUCTION

The web is the large medium which contains huge number of documents of different topics and category. The numbers of documents in the web are growing all the time and lookup process facing many challenges due to the dimension. The web document contains many information concerning many topics and for each topic there will be number of documents present in the web. So identifying the related documents at the requirement is becoming a challenging task. The people use the web for many purpose and they forever search for some information about any topic in the web. In order to provide them in efficient manner the documents of the web must be organized in proper manner. Also it is not necessary that the web document should speak about a specific topic but it can speak about many. So assemblage the document of the web is highly required one.

The clustering is the process of grouping the web documents into different category or classes. For example, there will be documents relate to the domain of "data mining", "networking" and so on. identify the documents speak about data mining and other topic then grouping them under the class name is called clustering. The document may contain many requisites relate to different topic but identifying the topic of the document is very important. The popular K-means algorithm computes the reserve between the documents of the class to perform clustering. There are number of other clustering algorithms available to group the papers under different classes.

The problem with the earlier approach is the dimensionality, the k-means algorithm computes reserve between the points and it cannot handle high dimensions. Similarly each algorithm has different issues in grouping the documents. The good organization of clustering is highly depending on the measure being used.

The topical measure is the major measure being used to perform document clustering. The topical measure represents how depth a meticulous topic is covered in the document. For example, if you compute the topical measure for the topic "Network", the topical measure Tm can be computed as follows:

Tm = Number of terms covered in the document/ Total number of terms belongs to the topic.

This is very much similar to the term frequency measure used in text cluster. More than that the topical measure can be computed using the taxonomy of words related to many topics. Using the pure topical measure will not help in improving the clustering performance. To look up the clustering performance a new measure is introduced in this paper.

Interior topical measure (ITM) is the value which is computed based on the taxonomy of words extracted from the list of documents of the class. If there exists N number of documents in the class C, then the interior topical measure is computed based on the number of terms from the documents set of class C, and the number of terms from the input document.

These characterize how depth the document is describing the topic and how it is close to the document set of the class C. Similarly the exterior topical measure (ETM) is computed based on the taxonomy of words being extract from the document set of the class C and the terms set being extracted from the document given as input. The exterior topical measure is the value which represent the Document closure to the document set of the other class C. By combing both the measure the document clustering can be computed in efficient manner.

II. RELATED WORKS

There are number of clustering approaches has been discussed for the problem of web document clustering and this section discuss about some of the methods relate to web document clustering. Agent-based document clustering [1] uses the ontology tool for preprocessing phase. The method extracts the features from the document and the process is performed with the help of synset values obtained from the tool of wordnet. This algorithm obtains the related terms for each of the semantic class and also improves the performance of clustering . Self-Organizing Map[SOM] based file Clustering Using WordNet Ontologies [3], proposed a semantic text document clustering which identifies the significance of the concepts in the document. SOM approach takes the advantages of the semantics available in knowledge base and the relationship between the words in the input documents. Some experiments are done to compare efficiency of the proposed approach with the lately reported approaches. On ontology-driven document clustering using core semantic features [4], discusses that, an ontology tool can be utilized to greatly reduce the number of features needed to do document clustering.In the preprocessing process nouns can be efficiently recognized in documents and that this alone provides improved clustering. This algorithm shows the importance of the polysemous and synonymous nouns in clustering.

An Efficient Semantic VSM based Email Categorization Method [5], select related semantic features that will increase the global information, and use them to enrich the semantic feature of an email. The proposed categorization method based on sVSM creates the sementic feature of an email category by both extracting terms of training email and enriching these terms with their concept-chains in WordNet. Next, tf*iwf*iwf algorithm is used to adjust the weight of the semantic feature vector. Experimental evaluations show that the proposed categorization method categorizing emails better than other email categorization methods based on traditional VSM, Baysian and KNN.

On Document symbol and Term Weights in Text Classification [6], explore the potential of enriching the document representation with the semantic information systematically discovered at the document sentence level. The salient semantic information is searched using a frequent word sequence method. Different from the classic tfidf weighting scheme, a prospect based term weighting scheme which directly reflect the term's strength in representing a specific category has been proposed.

Topic map based document clustering [9] predicts the meaingful terms by stemming process .The method maintains a semantic graph, which is generated with the help of semantic ontology. At the testing phase, the method computes similarity measure at each level. Based on these measures, a sub space or class of the document is identified.

A clustering approach for news group document [10] uses correlated concepts. It discusses the frequency based maximum resemblance. Correlated concept based maximum resemblance document clustering method works utilized correlation terms in a proper manner.. The algorithm has been evaluated for F-measure and purity.

Semantic similarity histogram based incremental document clustering (SHC) algorithm [11], introduced an incremental grouping algorithm grounded on Phrase-Semantic Similarity Histogram (PSSM). Here clusters are represented by histogram formats.. The PSSM analyze the term weight (word/phrase) founded on its relations with semantically similar footings that occur together in the text. And, the new text is incrementally added to the bunch, the semantic histogram ratio is intended and the insertion order problematic is spoken by making bad documents that reduce the cluster cohesiveness to leave, and recast them to a more appropriate cluster.

All these approaches identifies the merits and demerits of clustering process.

III. METHODOLOGY

Interior Exterior Topical Similarity Based Document Clustering: The problem of web document clustering has been approached using the ITM and ETM measures. The method compute the interior and exterior topical similarity measure between different class of documents and based on that the method computes the topical weight towards different class. The method use the open directory project taxonomy and wordnet synsets to compute the topical measures.

This section briefs the method in detail.

Preprocessing: In this stage, the method extracts the term from the input document. Over the extracted term set, the stemming and tagging is applied. The stemming is to remove the tense feature and to identify the pure term. The tagging is to identify the pure nouns from the word identified. For tagging the part of speech tagger is used which is given by the stanford university.

Algorithm:

Input: Document Di
Output: Term Set Ts.
Start
Read Document Di.
Split text into term set Ts.
Ts = Split(T, " ");
For each term Ti
Ti=$\int_{i=1}^{size(Ts)} Stemming(Ti,ed,Ing);$
End
For each term Ti
Ti = $\int POSTagging(Ti)$
End
Stop.

The preprocessing algorithm applies the tagging and stemming process to identify the pure nouns from the document given.

Topical Similarity Measure (TSM): The topical similarity measure is the value computed based on the taxonomy of words extracted from the document and wordnet. First the method extract the terms present in the input document Di, and extract the term set from each Class of Document set (CDs). For each class of document the method extract the term set and selects the pure nouns by applying the preprocessing technique. Once the term set has been extracted, the method identifies the list of related terms by using the wordnet and odp taxonomy. All the terms obtained from the both taxonomy is used to compute the topical similarity measure for the input document.

Input: Document Di, Class C
Output: Topical Similarity Measure Tsm.
Start
Read document Di
Term set Ts = Preprocessing(Di).
Document Set Ds = $\sum Documents \in C$
Initialize document term set DTs.
For each document Di
DTs = $\sum(Terms \in DTs) \cup Preprocessing(Di)$
End
Compute TSM.
TSM = $\frac{\sum Ti(Ts)\vdash DTs}{size(DTs)} \times \frac{Size(Ts)}{size(DTs)}$
Stop.

above discussed algorithm compute the topical similarity measure using the wordnet taxonomy and the open directory project taxonomy. This will be used to perform clustering in the final stage.

Interior Topical Measure: The interior topical measure is the value of closure which is computed between the documents of any specific class. The given document may come closure to any class but when you think about the closeness between the document of the class it may be scatter. So in order to measure the closeness between the documents of the class the ITM is computed. First the terms set are extracted and using the taxonomy the terms of other documents of the class, the ITM measure is computed.

Input: Document Di, Class C.
Output: ITM.
Start
Read Document Di.
Term set Ts = Preprocessing(Di).
For each document Dk of C
DTS = $\sum(Terms \in DTs) \cup Preprocessing(Dk)$
End
For each term Ti from DTs
DTs = $\sum(Terms \in DTs) \cup Wordnet(Ti) \cup ODP(Ti)$
End
Compute interior topical similarity Isim.
ITM = $\frac{\sum Terms(Ts)\vdash DTs(Dk)}{size(Ts(Di))} \times \frac{\sum Terms(Ts)\vdash DTs(Dk)}{number\ of\ documents\ of\ Ci}$
Stop.

The above discussed algorithm computes the interior topical measure to be used in clustering the web document.

Exterior Topical Measure: The exterior topical similarity measure represent the topical measure for the input document Di which belongs to the class Ci, towards other class of documents. This is computed by measuring the closeness of the terms and their presence the documents of other classes. The method first identifies the terms and identifies the nouns using the word net taxonomy. Using the terms identified and the nouns identified, the method compute the exterior topical similarity measure.

Input: Input Document D, Target Class TC.
Output: ETM.
Start
Read Document D.
Term set ITs = Preprocessing(D).

For each document Dk of TC

Target term set TTS = $\sum(Terms \in TTs) \cup Preprocessing(Dk)$

End

For each term Ti from TTs

TTs = $\sum(Terms \in TTs) \cup Wordnet(Ti) \cup ODP(Ti)$

End

Compute exterior topical similarity Esim.

ETM = $\frac{\sum Terms(ITs) \in ITs(Dk)}{size(ITs(D))} \times \frac{\sum Terms(ITs) \in TTs(Dk)}{number\ of\ documents\ of\ TC}$

Stop.

The above discussed algorithm computes the exterior topical measure to be used in clustering the web document.

ITM-ETM Clustering: In this method, the topical similarity measure for each class is computed using the taxonomy. Based on computed measure, the method selects a single class. Then for each class, the method computes the interior and exterior topical similarity measure.

Finally a topical weight is computed based on which the document is assigned to a class.

Input: Document D, Classes C

Output: Highest Cumulative Topical Similarity Value.

Begin

Identify list of terms and pure nouns.

For each class

Compute topical similarity.

Tsim = $\frac{\sum Terms(T) \vdash Ts}{size(C)}$

End

Choose the class with higher topical similarity.

For the class selected

Compute Interior topical similarity Isim.

For each document Di from class Ci

Compute interior topical similarity Isim.

ITM = $\frac{\sum Terms(Ts) \vdash Ts(Di)}{size(Ts(Di))} \times \frac{\sum Terms(Ts) \vdash Ts(Dk)}{number\ of\ documents\ of\ Ci}$

End

Compute exterior topical similarity

ETM = $\frac{\sum Terms(Ts) \in Ts(Di(Ck))}{size(Ts(Di(Ck)))}$

End

Compute cumulative topical similarity Cts = ITM×ETM

End

Choose the class with higher Cts value.

End

The above discussed clustering algorithm computes the ITM and ETM measures to identify the class of the document.

IV. RESULTS AND DISCUSSION

The proposed method have been evaluated using different data set to test the performance and effectiveness of the web document clustering. Table 1 shows the details of the data set used. The outcome of the experimentation demonstrates that Interior-Exterior Topical Similarity Based Web Document Clustering method has achieved 99.37% of accuracy in clustering which is a drastic improvement compared to concept based document indexing (Fatiha *et al* 2010). Data set is requlary updated according to the number of pages. The result proves that the proposed approach has produced an efficient result in web document clustering. The clustering accuracy has been predicted and compared among various methods and the number of documents has been assigned with correct class labels. The comparative result on clustering accuracy has been presented in the Figure 1. The result proves that the proposed method has improved the clustering accuracy.

Table 1 : Details of the data set used for the evaluation.

Data set	No. of Pages	Avg. Session length(in minutes)
YELP	7891556	9.8
UCI	98563	4.5
DATAGOV	156393	5.9

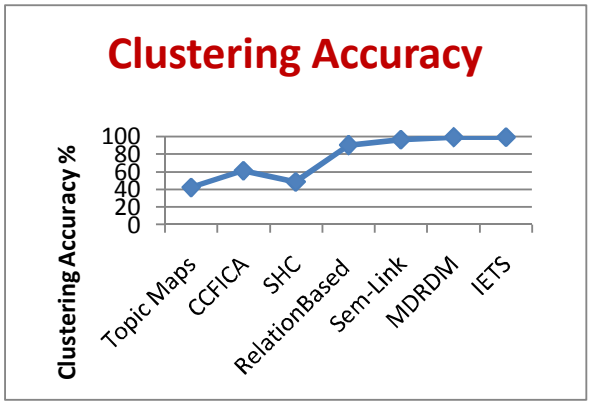


Fig. 1. Result of clustering accuracy.

Table 2 : Comparison result of clustering.

Algorithm	No. of Pages	Clustering Accuracy %	False identification Ratio %	Time Complexity in seconds
Topic Maps	20000	42.2	19	4000
CCFICA	20000	61.4	12	2400
SHC	20000	48.8	10	2000
Relation Based	20000	89.8	5	400
Sem-Link	20000	96.4	3	200
MDRDM	20000	98.97	1.5	148
IETS	20000	99.37	1.2	105

From table 2, it is clear that the proposed algorithm produces more efficient and accurate indexing where the other methods produce less indexing accuracy.

IV. CONCLUSION AND FUTURE ENHANCEMENT

In this paper, an efficient interior and exterior topical similarity based web document clustering is presented. The method preprocess the document to identify the pure terms using the stemming process. In the second stage, the topical similarity measure is computed towards each category of documents. Based on computed topical similarity measure a single class is selected. Then the clustering is evaluated by computing the interior and exterior topical similarity measure. The method has produced higher clustering accuracy than others and achieves the efficiency up to 99.37 %. Further the performance of document clustering can be improved by computing topical measure for subclasses of each category using multi level topical measure estimation technique.

REFERENCES

[1]. Khaled M Fouad and Moataz O Hassan,' Agent for Documents Clustering using Semantic-based Model and Fuzzy', *International Journal of Computer Applications* January 2013.

[2]. Fellbaum, C.,' WordNet. Theory and Applications of Ontology: Computer Applications, PP. 231-243, Springer Science, 2010.

[3]. Gharib, T., Fouad, M., Mashat, A. & Bidawi, I,'Self-Organizing Map-based Document Clustering Using WordNet Ontologies', *IJCSI International Journal of Computer Science* Issues, Vol. **9**, Issue 1, No 2., 2012.

[4]. Fodeh, S., Punch, B. & Tan P. On ontology-driven document clustering using core Springer-Verlag London Limited., 2011.

[5]. Zhao, L., Jianguo, D,'An Efficient Semantic VSM based Email Categorization Method. International Conference on Computer Application and System Modeling ,ICCASM 2010.

[6]. Liu, Y. On Document Representation and Term Weights in Text Classification, 'Handbook of Research on Text and Web Mining Technologies', PP: 1-22, 2009.

[7]. B. Fatiha, B. Mohand, T. Lynda, D. Mariam. Using WordNet for Concept-Based Document Indexing in Information Retrieval, SEMAPRO: The Fourth International Conference on Advances in Semantic Processing, Pages: 151 to 157, 2010.

[8]. Dragoni, M., Pereira, C. & Tettamanzi, A. ,An Ontological Representation of Documents and Queries for Information Retrieval Systems, pp. 555–564, Springer-Verlag Berlin Heidelberg, 2010.

[9]. Muhammad Rafi, Shahid M Shaikh and Amir Farooq. Article: Document Clustering based on Topic Maps. *International Journal of Computer Applications*, 2010.

[10]. Jayaraj Jayabharathy and **S**elvadurai Kanmani' Correlated concept based dynamic document clustering algorithms for newsgroups and scientific literature, Springer, Decision Analytics, 2013.

[11]. Daniel Muller, Modern hierarchical, agglomerative clustering algorithms, SAO/NSA strophysics Data Systems, PP.2378, 2011.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 108-111(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Comparative Study on the Performance of Clustering Algorithms using Validation Measures on Diabetes Dataset

R. Yasotha[1] and Sarojini. B[2]
[1]II M.Sc (Computer Science), Department of Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India
[2]Assistant Professor (SS), Department of Computer Science,
Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore (TN), India

ABSTRACT: **Medical data mining extracts some interesting facts from huge medical data sets. Cluster analysis or clustering is the mission of grouping a set of substances in such a way that objects in the same group are more similar to each other than to those in other groups. This research paper compares the performance of three clustering algorithms such as Hierarchical clustering, PAM clustering and K-Means clustering algorithms on Pima Indian diabetes dataset. The cluster validation measures Internal Validation and Stability Validation are used to evaluate the performance of clustering algorithms and the results are compared. Internal Validation are evaluated in terms of Connectivity, Dunn Index, Silhouette width and Stability Validation are evaluated in terms of Average Proportion of Non-overlap, Average Distance,Average Distance Means, Figure of Merit. The experimental results showthat Hierarchical Clustering gives better performance compared to other two clustering algorithms.**

Keywords*:* Clustering, Pima diabetes dataset, Hierarchical, k-Means, PAM, cluster validation measures.

I. INTRODUCTION

Data mining is the process of analyzing the given data in order to provide useful information contained in that data that cannot be retrieved through queries. It helps to view data from different angles and group it into information that may be useful in many perspectives.A huge amount of data gets accumulated in the hospitals, most of them just get stored in some form of files which are never touched back. If these data are analyzed properly they help in deriving some interesting facts. [1] Data mining helps in generating interesting facts which may remainunrevealed otherwise. Medical data mining analyse medical data and derives knowledge from medical data.Clustering is applied on the PIMA data to group similar data. The efficiency of clustering is validated by evaluating the performance of clustering

The objective of the research work is to study the performance of clustering algorithms on Pima Indian Diabetes Dataset using cluster validation measures. The performance of three clustering algorithms, hierarchical clustering algorithm, k-means clustering algorithm and PAM (K-medoids) clustering algorithm are analyzed in terms of internal validation measures and stability validation measures and results are compared to find the clustering algorithm that works best on the Pima Indian Diabetes Dataset.

II. PROPOSED WORK

The Clustering technique is used to place data rudiments into related clutches without advance knowledge of the group description. The clustering technique groups the data instances in to subsets in such a manner that similar instances are grouped together while different instances belong to different groups. The objective of this research work is to study the performance of three clustering algorithms such as Hierarchical clustering, PAM clustering and K-Means clustering algorithm on Pima Indians Diabetes Dataset. [2] The performance of the clustering algorithms dataset are evaluated in terms of internal validation measuresand Stability validation measures.

This research work is designed with three major process. They are

- Data preprocessing
- Clustering
- Validation

A. Data Preprocessing

Data preprocessing mechanisms are applied step before applying data mining algorithm to improve the quality of

the dataset. Data is preprocessed to remove missing values.

B. Applying cluster techniques

The performance of three clustering algorithms, Hierarchical clustering algorithm, K-Means clustering algorithm and Partitioning AroundMedoids (K-medoids) clustering algorithm are analyzed on Pima Indian diabetes dataset.[3] Clustering technique is used for finding hidden patterns in data mining. In this project three clustering algorithms are used for grouping the Diabetes dataset and they are as follows.

- Hierarchical algorithm
- Partitioning algorithm
  - ✓ K-means
  - ✓ PAM(k-medoids)

Hierarchical clustering: Hierarchical clustering can be subdivided into two types:

- Agglomerative hierarchical clustering algorithms start with each object being a separate cluster itself, and successively merge groups according to a distance measure. These methods generally follow a greedy-like bottom-up merging.
- Divisive hierarchical clustering algorithms follow the opposite strategy. Divisive approaches similar to the approach followed by divide-and-conquer algorithms.

Cluster and sub-cluster relationship is recurrently displayed graphically using a tree-like diagram called a dendrogram [4]. Hierarchical clustering is an agglomerative clustering algorithm that yields a dendrogram which can be cut at a chosen height to produce the desired number of clusters

Partitioning algorithm: Partitioning algorithms are clustering approaches that split the datasets, containing n number of observations, into a set of k groups (i.e. clusters). The algorithms require the analyst to specify the number of clusters to be generated.

The most common partition algorithms are:

- K-Means clustering
- K-Medoids clustering

K-Means Clustering. The popular clustering algorithm that minimizes the clustering error is the K-means algorithm. [5] K-means algorithm is first applied to an N-dimensional population for clustering them into *k* sets on the basis of a sample by MacQueen in 1967. The algorithm is established on the input parameter *k*. First of all, *k* centroid point is selected randomly. These *k* centroids are the means of *k* clusters. Then, each item in the dataset is assigned to a cluster which is nearest to them. Then, means of all clusters are computed again with new points added to them, until values of means do not modify.

K-Medoids Clustering. K-Medoids clustering or PAM Partitioning Around Medoids, Kaufman & Rousseeuw, 1990, in which, each cluster is represented by one of the objects in the cluster. [6] Its a "non-parametric" robust alternative to k-means clustering, less sensitive to outliers. Partitioning around medoids (PAM) is similar to K-means, but is considered more robust because it admits the use of other dissimilarities besides Euclidean distance.

C. Cluster validation

Clustering validation [7] includes three main tasks:

1. Cluster tendency
2. Cluster evaluation
3. Cluster stability

Clustering tendency, this process is defined as the assessing of clustering tendency or the feasibility of the clustering analysis. This process must be checked before applying clustering techniques.

Clustering evaluation measures the purity or quality of the clustering.

Clustering stability seeks to understand the sensitivity of the clustering result to various algorithmic parameters, for example, the number of clusters.

The aim of this part is to:

1. describe the different methods for clustering validation
2. compare the quality of clustering results obtained with different clustering algorithms
3. provide R lab section for validating clustering results

Clustering validation measures in clValid package

-Internal validation measures

-Stability validation measures

CLVALID is technique used to evaluatethe performance of unsupervised clustering techniques.[8] CLVALID uses the R package "clValid" to compare the relative properties of three different clustering methods over a several numbers of clusters. This technique aims to help choose a method that is most compact, well-separated, connected, and stable.

Internal Validation: This validation evaluates the purity and quality of the clustering based solely on the dataset and the clustering separation. This valuation is established by the measures cluster Connectivity, Silhouette Width and Dunn Index, which were chosen to explicate the compactness, connectedness and separation of the cluster partitions. It uses core information in the data to assess the quality of the clustering.

The cluster internal measures include in clvalid package are:

- **Connectivity:** This measure reflects the level to which items that are placed in the same cluster are also considered their nearest neighbors in the data space and the degree of connectedness of the clusters. It should be minimized**.**

$$Conn(C) = \sum_{i=1}^{N} \sum_{j=1}^{L} xi, nn(j)$$

- **Average Silhouette width:** This index defines density based on the pairwise distances between all rudiments in the cluster, and split based on pairwise distances between all points in the cluster and all points in the nearest other cluster. We used silhouette function to asses the optimal number of clusters in the previous post - andthe values as close to (+) 1 as possible are most suitable.

$$S(i) = \frac{bi - ai}{max\ (bi, ai)}$$

Where,

$$ai = \frac{1}{n(C(i))} \sum_{j \in C(i)} dist(i, j)$$

$$bi = min\ ck \in C \setminus C(i) \sum_{j \in Ck} \frac{dist(i,j)}{n\ (Ck)}$$

- **Dunn index:** Dunn Index represents the ratio of the minimum distance between the observations not in the identical cluster to the largest intra-cluster distance.

$$D = \frac{min\ i\ min\ j(min\ x \in Ci, y \in Cjd(x, y))}{max\ k(max\ x, y \in Ckd(x, y))}$$

StabilityValidation: This validation uses an iterative approach of removing one column (or row, if by genes) from the dataset and comparing the results. It is a special version of internal validation. It evaluates the steadiness of a clustering outcome by paralleling it with the clusters attained after each column is removed, one at a time.

The cluster stability measures included in clvalid package:

Average Proportion of Non-overlap (APN): The APN assessesthe average amount of observations not placed in the same cluster by clustering based on the whole data and clustering based on the data with a solitary column removed.

$$APN(C) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{i=1}^{M} \left(1 - \frac{n(C^{i,l} \cap C^{i,0})}{n\ (C^{i,0})}\right)$$

Average Distance (AD): The AD measures the middling distance between observations placed in the same cluster underneath both full dataset and removal of one column.

$$AD(C) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{i=1}^{M} \frac{1}{n(C^{i,0}) n(C^{i,l})} \left(\sum_{i \in C^{i,0}} \sum_{j \in C^{i,l}} dist(i, j) \right)$$

Average Distance Means (ADM): The ADM measures the average distance between cluster centers for observations placed in the same cluster under both cases.

$$ADM(C) = \frac{1}{MN} \sum_{i=1}^{N} \sum_{i=1}^{M} dist(x^i c^{i,l}, x^i c^{i,0})$$

Figure Of Merit (FOM): The FOM measures the average intra-cluster inconsistency of the removed column, where the clustering is based on the remaining (undeleted) columns. It has a value between 0 and 1, and minimum values are preferred.

$$FOM(l, C) = \sqrt{\frac{1}{N} \sum_{k=1}^{K} \sum_{i \in C(k)} dist(xi, l, x^i k(l))}$$

III. EXPERIMENTAL RESULTS AND DISCUSSION

The experiment is performed in R environment. The Table 1 and 1.1 shows the experimental results of clustering internal validation measures and its optimal scores. Table 2 and 2.1 stability validation measures and its optimal scores.

Table 1. Internal Validation.

Method	Measure	Cluster Sizes				
		2	3	4	5	6
Hierarchical	Connectivity	2.93	12.21	27.22	30.17	30.15
	Dunn	0.30	0.22	0.17	0.19	0.19
	Silhouette	0.43	0.35	0.32	0.29	0.26
K-Means	Connectivity	156.80	145.77	234.36	344.93	286.33
	Dunn	0.05	0.06	0.06	0.05	0.06
	Silhouette	0.22	0.23	0.23	0.19	0.18
PAM	Connectivity	91.60	183.04	292.89	290.52	346.15
	Dunn	0.06	0.06	0.04	0.04	0.04
	Silhouette	0.22	0.17	0.146	0.16	0.14

TABLE 1.1: OPTIMAL SCORES OF INTERNAL VALIDATION.

Measures	Score	Method	Cluster Size
Connectivity	2.93	Hierarchical	2
Dunn	0.30	Hierarchical	2
Silhouette	0.43	Hierarchical	2

TABLE 2. STABILITY VALIDATION.

Clustering Method	Measure	Cluster Sizes				
		2	3	4	5	6
Hierarchical	APN	0.0089	0.0110	0.0157	0.0196	0.0425
	AD	4.0027	3.9653	3.9532	3.9369	3.9230
	ADM	0.0582	0.0987	0.1088	0.1158	0.2041
	FOM	0.9799	0.9966	0.9959	0.9950	0.9896
K-Means	APN	0.1618	0.1866	0.2661	0.3059	0.3904
	AD	3.6962	3.5409	3.4652	3.4303	3.3116
	ADM	0.5364	0.7457	0.8807	0.8943	1.1285
	FOM	0.9741	0.9524	0.9453	0.9466	0.9443
PAM	APN	0.1151	0.2893	0.2815	0.3380	0.3761
	AD	3.6790	3.6110	3.4438	0.3983	3.2861
	ADM	0.3740	0.9202	0.8158	1.0609	1.0833
	FOM	0.9686	0.9620	0.9505	0.9469	0.9383

Table 2.1: OPTIMAL SCORES OF STABILITY VALIDATION.

Measures	Score	Method	Clusters
APN	0.0089	Hierarchical	2
AD	3.2861	PAM	6
ADM	0.0582	Hierarchical	2
FOM	0.9383	PAM	6

The experimental result shows that the better result has been obtained for Hierarchical clustering algorithm measured in terms of internal validation measures (connectivity, Dunn and silhouette) and stability validation measures (average propagation of non-overlap, average distance, average distance between means and figure of merits) compared to K-Means and PAM clustering algorithms.

IV. CONCLUSION AND FUTURE WORK

The performance of clustering algorithms on diabetes dataset is taken up for study.

Three clustering algorithms such as k-means clustering algorithm, PAM clustering algorithm and hierarchical clustering algorithm are applied on the diabetes dataset. The performance is analyzed based on the internal validation measure. The results shows that hierarchical clustering is the best algorithm compared to other two algorithms. The research work is developed in RStudio. The results are best visualized graphs also. The future work will be to compare other cluster algorithms that can be applied on the data set and to identify the best clustering algorithm for the dataset.

REFERENCES

[1]. Han, J. and Kamber, M., Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, 2006.

[2]. Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining", *International Journal of Engineering Research and Applications (IJERA)* Vol. **2**, Issue 3, May-Jun 2012.

[3]. FASULO, D. 1999. An analysis of recent work on clustering algorithms. Technical Report UW-CSE01 -03-02, University of Washington.

[4]. Ming-chuan hung et al., "An Efficient k-Means Clustering Algorithm using simple partitioning" *Journal of information science and engineering.*

[5]. HARTIGAN, J. 1975. Clustering Algorithms. John Wiley & Sons, New York, NY.

[6]. JAIN, A. and DUBES, R. 1988. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, NJ.

[7]. S. Datta and S. Datta, "Comparisons and validation of statistical clustering techniques for microarray gene expression data", *Bioinformatics,* Vol. **19**, No. 4, 2003.

[8]. G. Brock, V. Pihur, S. Datta, and S. Datta, "clValid: An R Package for Cluster Validation", *Journal of Statistical Software*, Vol. **25**, Iss. 4, 2008, pp. 1-20.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 112-120(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Multiobjective Firefly Optimization Based Similarity Measure for Content Based Image Retrieval

Dr. K. Haridas
Associate professor and Head, Dept of Computer Applications,
NGM College, Pollachi, Coimbatore (DT), India.

ABSTRACT: The purpose of content based image retrieval (CBIR) systems is to allow users to retrieve pictures from large image repositories. In a CBIR system, an image is usually represented as a set of low level descriptors from which a series of underlying similarity or distance functions are used to conveniently drive the different types of queries. Recent work deals with combination of distances with K means clustering, Fuzzy K means clustering from different and usually independent representations in an attempt to induce high level semantics from the low level descriptors of the images. The conventional methods generally focuses on the initial centroid selection while measuring dissimilarity measure, but still choosing the best feature similarity is a major problem. In order to overcome these problems, a multiobjective firefly optimization based framework to measure similarity among the feature vector with low level and high level features is proposed. Each firefly move from one feature vector to another feature vector to find the best global firefly based on the Average Euclidean distance (AED). Combining the low-level visual features and high-level concepts using multi objective firefly optimization methods, the proposed approach fully explores the similarities among images in database and optimizes best similarity measure result from MOFA the relevance results from traditional image retrieval system. The results demonstrate that the proposed MOFA based similarity measures improves retrieval accuracy when compared with the conventional clustering approaches like K means clustering, Fuzzy K means clustering approaches because of the best similarity distance results.

Keywords: Content based image retrieval, k means clustering, fuzzy k means clustering, multi objective firefly algorithm

I. INTRODUCTION

There are huge amount of data spread around us and that data are in various form like text, graphics, images, audio, video etc. Main problem in front of us is how to manage that data and get it at the appropriate time. Content Based Image Retrieval (CBIR) is a set of techniques for retrieving semantically-relevant images from an image database based on automatically-derived image features [1]. CBIR is extremely useful in a plethora of applications such as publishing and advertising, historical research, fashion and graphic design, architectural and engineering design, crime prevention, medical diagnosis, geographical information and remote sensing systems, etc. The main goal of CBIR is efficiency during image indexing and retrieval, thereby reducing the need for human intervention in the indexing process.

Some of the existing CBIR systems extract features from the whole image not from certain regions of it; so, they are global features. Histogram search algorithms [2] characterize an image by its color distribution or histogram. Many distances have been used to define the similarity of two color histogram representations. Euclidean distance and its variants are the most commonly used. The drawback of a global histogram representation is that information about object location, shape and texture is discarded. Color histogram search is sensitive to intensity variations, color distortions, and cropping. The color layout approach attempts to overcome the drawback of histogram search. In simple color layout indexing [2], images are partitioned into blocks and the average color of each block is stored. Thus, the color layout is essentially a low resolution representation of the original image.

Due to the high dimensionality of the images, researchers use the similarity measure to measure the degree of similarity between images along with the features from query and input image. But, there still exists a semantic gap, which just reflects the discrepancy between the relatively limited descriptive power of low-level visual features and high-level concepts. The system is based on the similarities

between the query image and images in database while ignoring the similarities among images in database.

There have been some attempts in theoretically summarizing existing dissimilarity measures [3], evaluating dissimilarity measures for texture [4], and shape based image search [5]. Our previous work [3] gives a description of dissimilarity measures on six feature spaces, but only single-image queries is conducted on one image collection (Corel), which makes the conclusions difficult to generalize to other collections. There is still a lack of a systematic investigation of dissimilarity measures on different feature spaces, with largescale real-world image collections.

Most retrieval systems including CBIR ones explicitly rely on distance, similarity or score functions aiming at relating descriptors to perceptual or subjective resemblance to some extent [6,7]. The Minkowski-form distance, the Manhattan distance, the Euclidean distance, the Hausdorff distance, the Quadratic Form (QF) distance, the Mahalanobis' distance, the Kullback-Leibler divergence [8] and the Jeffrey divergence are some of the most commonly used functions to estimate the similarity between pictures.

Clustering becomes one of the important methods to measure the similarity between the features in the CBIR system. Clustering involves dividing a set of data points into two level of the features and clusters whose similarity measure are equal to both images, k means clustering becomes one of the important clustering methods to group similar images features .Compare the two clustering techniques: K- mean and Fuzzy C-mean clustering for image retrieval. In both techniques the distance metric concept is used for the analysis, both algorithms find out the distance between the centroid of the cluster and seed point. Fuzzy k means clustering algorithm also important clustering methods to measure similarity based on fuzzy membership function initial centroid selection problem is overcome and best feature similarity measure is performed in Fuzzy k means clustering methods. In this Fuzzy k means clustering algorithm still the initial Centroid selection problem occurs ,so the similarity measure results also reduced .To overcome the problems similarity measures between the data objects with feature vector is measured using optimization methods . In this work to overcome the problem of similarity measures among the feature proposed a multiobjective firefly optimization framework to infer which images in the databases would be of most interest to the user so the low level and high level features are found between the query and input images. Three visual features, color, texture, and shape, of an image are utilized in our approach. MOFA provides an interactive mechanism to better capture user's intention than K means and Fuzzy K Means clustering algorithm.

II. BACKGROUND STUDY

There are some literatures that survey the most important CBIR systems [9], [10]. Also, there are some papers that overview and compare the current techniques in this area [11], [12]. Since the early studies on CBIR, various color descriptors have been adopted. Yoo *et al.* [13] proposed a signature-based color-spatial image retrieval system. Color and its spatial distribution within the image are used for the features. In [14], a CBIR scheme based on the global and local color distributions in an image is presented. Vadivel et al. [15] have introduced an integrated approach for capturing spatial variation of both color and intensity levels and shown its usefulness in image retrieval applications

Chin-Chin Lai *et.al.* [16] have proposed an interactive genetic algorithm (IGA) to reduce the gap between the retrieval results and the users' expectation .They have used color attributes like the mean value, standard deviation, and image bitmap .They have also used texture features like the entropy based on the gray level co-occurrence matrix and the edge histogram. They compared this methods with others approaches and achieved better results.

Gwenole Quellec *et.al.* [17] have presented a novel method to adapt a multidimensional wavelet filter bank to any specific problem .They have applied this method for content based image retrieval. The performances of the adapted wavelet filter bank over the nonadapted wavelet filter bank are higher for every database.

Region based image retrieval of [11] uses low-level features including color, texture, and edge density. For color, the histograms of image regions are computed. For texture, co-occurrence matrix based entropy, energy, etc., are calculated, and for edge density it is Edge Histogram Descriptor (EHD) that is used. To decrease the retrieval time of images, an idea is developed based on greedy strategy to reduce the computational complexity. In this strategy, the query image is compared to each of the target images in the database based on region matching in term of Euclidian distance between them. In [18] a simple linear normalization so that features are mapped to the range [0, 1] was applied; and in a genetic algorithm [19] was used to derive a set of weights for each descriptor. Evidence combination has also been used in complete CBIR systems usually to fuse possibly multimodal information taken from user feedback [20].

Content-Based Image Retrieval (CBIR) is commonly used system to handle these datasets. Basis

on the image substance CBIR extracts the images that are relevant to the user given query image from large image databases. Many of the CBIR systems retrieval of the result are corresponding to feature similarities for user given query, ignoring the similarities among images in database. These existing CBIR system measures the feature similarities by using k means algorithm, but the traditional k-means algorithm mostly depends on the selection of initial centers values, the algorithm normally uses random procedures to get them and it degrades the performance of the CBIR retrieval results. To overcome the problem of initial centroid random selection process in K means clustering algorithm use the fuzzy logic based feature similarities information with K means clustering algorithm [21] to image retrieval system. Combining both low-level and high-level visual features, the fuzzy k means algorithm entirely measures the features similarities information between the images in larger dataset. Fuzzy k means clustering algorithm optimizes the relevance results from conventional image retrieval system by firstly clustering the related images in the images database to improve the effectiveness of images retrieval system.

III. PROPOSED METHODOLOGY

The system could be any real value symmetric image retrieval system. First, extract the image features of each image in image database and apply the optimization framework to find the similar image results to analysis them, then, input the query image, extracting its features and comparing the similarities between features of it and those of images in image clustering database, and output the best matching results.

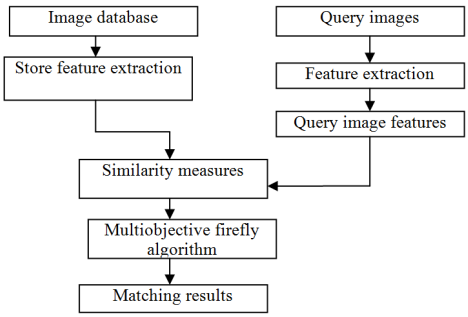


Fig. 1. Overall Proposed Framework of Content-Based Image Retrieval.

In the image retrieval, feature extraction is the most import issue of the first step. The features extracted from the images directly lead to the results. Some preprocessing is needed to avoid retrieval noise. Steps such as removing the background, highlights the objects. All the preprocessing steps help in feature extraction. Figure 1 illustrates the overall proposed framework of content-based image retrieval.

A. Preprocessing

Pre-processing is the name used for operations on images at the lowest level of abstraction. The aim of the pre-processing is an improvement of the image that suppresses unwilling distortions or enhances some image features, which is important for future processing of the images. The noise in the images is filtered using linear and non-linear filtering techniques. Median filtering is used here to reduce the noise.

Median filter. Median filter is applied to the image to eliminate noises. Median Filter is a non-linear smoothing method that reduces the blurring of edges, in which the idea is to replace the current point in the image by the median of the brightness in its neighborhood. Individual noise spikes do not affect the median of the brightness in the neighborhood and so median smoothing eliminates impulse noise quite well. Median filter is a better filtering technique according to performance and takes less computational time. It smoothes salt and pepper noises [22]. During the process of median filtering, each pixel is replaced by the median of the pixels contained in a window around it. It can be expressed as:

$$IM(m,n) = Median\,[x(m-k,n-l) \in w \quad (1)$$

B. Feature Extraction

Color feature extraction is done by HSV color histogram and for texture feature extraction use the gray scale texture moment. Color feature extraction techniques only retrieves and form the group of images on the basis of color only. All red color images occur at the single group but it does not consider the texture of an images. But when texture feature extraction is also used with that of color extraction that time it considers the frequently occurring patterns in that images.

Color Feature Extraction. Color feature is extracted by Color Histogram and Color Descriptor. The Color histogram specifies the color pixel distribution in an image. Color histogram uses two types of color space that are RGB, HSV [23]. Color Histogram (CH), contains occurrences of each color obtained by counting all image pixels having that color. Each pixel is associated to a specific histogram bin only on the basis of its own color, and color similarity across different bins or color dissimilarity in the same bin is not taken into account. Since any pixel in the image can be described by three components in a certain color space (for instance, red, green and blue components in RGB space or hue, saturation and value in HSV space), a histogram, i.e., the distribution of the number of pixels for each quantized bin, can be defined for each component. Color descriptor consists the color

expectancy, color variance and color skewness. Color expectancy is the average or mean of intensity in image. Color variance is the square root of the standard deviation. Color skewness is a measure of the asymmetry of the probability distribution of a real valued random variable. Two types of skewness are Positive skewness and Negative skewness.

Texture Feature Extraction. An image texture is a set of metrics calculated in image processing designed to quantify the perceived texture of an image. Image Texture gives us information about the spatial arrangement of color or intensities in an image or selected region of an image. Texture analysis attempts to quantify intuitive qualities described by terms such as rough, smooth, silky, or bumpy as a function of the spatial variation in pixel intensities. In this sense, the roughness or bumpiness refers to variations in the intensity values, or gray levels.

Gray Level Co-Occurrence Matrix. A statistical method for examining the texture that considers the spatial relationship of pixels is the gray-level co-occurrence matrix (GLCM), also known as the gray-level spatial dependence matrix. GLCM is a matrix that describes the frequency of one gray level appearing in a specified spatial linear relationship with another gray level within the area of investigation [24]. Here, the co-occurrence matrix is computed based on two parameters, which are the relative distance between the pixel pair d measured in pixel number and their relative orientation φ .Normally, φ is quantized in four directions $0^0, 45^0, 90^0$ and 1 [25]. In practice, for each , the resulting values for the four directions are averaged out. To show how the computation is done, for image , let represent the gray level of pixels $(x,$ and represent the gray level of pixels $(x \pm d\phi_0, y \pm d($ with level of gray tones where $0 \leq x \leq M-1, 0 \leq y \leq N$ -and $0 \leq m,n \leq L$ -. From these representations, the gray level cooccurrence matrix C,m n C , for distance d and direction can be defined as (2),

$$C_{m,n,\phi} = \sum_{x} \sum_{y} P\{I(x,y) = m \,\&\, I(x \pm d\phi_0, y \mp d\phi_2) = n\} \quad (2)$$

where $P\{.\} =$ if the argument is true and otherwise, $P\{.\} =$. For each value, its and values. One of the characteristic of the GLCM is, it is diagonally symmetry where $C_{m,n} = C$. Thus, the computation of the GLCM can be simplified as in Equation 2. Now, the ± and ∓ signs are replaced with single operation of + and – accordingly. As compensation, the resulting C, is added with C, to obtain a complete GLCM. For the rest of this paper, GLCM computation will be referred to the method as in Eq.3.

$$C_{m,n,\phi} = \sum_{x} \sum_{y} P\{I(x,y) = m \,\&\, I(x + d\phi_0, y - d\phi_2) = \cdot \quad (3)$$

Similarity measure using Multiobjective Firefly Algorithm (MOFA)

Firefly Algorithm was developed by Yang for continuous optimization [26], which was subsequently applied into and image processing [26]. In this work firefly algorithm is applied to measure the feature similarity between query image and input images based on the flashing patterns of feature vectors is considered as firefly and behaviour of fireflies measure the one feature vector input image to feature vector of the query image. In essence, FA uses the following three idealized rules to measure similarity of features:

(1) Fireflies are unisex so that one feature vectors (firefly) will be attracted to other feaure vector (firefly) regardless of their sex; (2) The attractiveness of a feature vector firefly is proportional to its brightness and they both decrease with distance $d(x_{i,}$. Thus for any two flashing fireflies (feature vectors), the less bright one will move towards the brighter one. If there is no brighter one than a particular feature vectors, it will move randomly; (3) The brightness of a feature vector (firefly) is determined by the landscape of the objective function of the $d(x_{i,}$.

For a maximization problem, the brightness can simply be proportional to the value of the feature similarity measure with distance objective function $d(x_{i,}$. As both light intensity and attractiveness affect the movement of fireflies (feature vectors) in the firefly algorithm, have to define their variations. For simplicity, can always assume that the attractiveness of a firefly is determined by its brightness which in turn is associated with the encoded distance similarity function $d(x_{i,}$. In the simplest case for maximum optimization problems, the brightness of a feature vector at a particular location can be chosen as (4),

$$I(x) \propto f(x) \quad (4)$$

However, the attractiveness is relative to most feature similarity function, it should be seen in the eyes of the judged by the other fireflies. Thus, it will vary with the distance between feature vector i and feature vector j. Therefore, can now define the attractiveness of a firefly by Eq.5,

$$\beta = \beta_0 e^{-\gamma r^2} \quad (5)$$

Where is the attractiveness at $r =$. In fact, equation defines a characteristic distance $\Gamma =$ over which the attractiveness changes significantly from to $\beta_0 \epsilon$. The distance between any two

feature vector i and j at and , respectively two - dimensional feature vectors of an image from a database and the query image, An effective Average Euclidean distance (AED) based similarity measurement is used and it is defined as Eq.6,

$$r_{ij} = \|x_i - x_j\| = d(x_i, x_j) = \sqrt{\sum_{i=1}^{n} \left(\frac{x_i - x_j}{|x_i| + |x_j|} \right)} \quad (6)$$

The proposed distance similarity measurement not only contains some relations between objects, but also comprehensively considers all dimensional feature parameters. For any given two feature vectors and , the movement of firefly i is attracted to another more attractive (brighter) firefly j is determined by Eq.8,

$$x_i^{t+1} = x_i^t + \beta_0 e^{-\gamma r_{ij}^2} \left(x_j^t - x_i^t \right) + \alpha_t \epsilon_i^t \quad (7)$$

where the second term is due to the attraction. The third term is randomization with being the randomization parameter for each feature similarity, and is a vector of random numbers drawn from a uniform distribution. The location of feature vector can be updated sequentially, by comparing and updating each pair of them in every iteration cycle. For most implementations, can take β_0 =and $\alpha_t = 0$ (1though found that it is better to use a time-dependent so that randomness can be reduced gradually as iterations proceed. It is worth pointing out that is a random walk biased towards the brighter fireflies (feature vectors) . If β_0 =, it becomes a simple random walk. The parameter now characterizes the variation of the attractiveness of the best feature similarity measure results, and its value is crucially important in determining the speed of the convergence for each feature similarity measure and how the FA algorithm behaves. In theory, $\gamma \in [0,$ but in practice, $\gamma = O($ is determined by the characteristic distance of the system to be optimized. Thus, for most applications, it typically varies from 10^{-2} to1(. To consider the scale variations of each problem, now use rescaled, vectorized parameters, $\alpha = 0.01L, \gamma =$ With $L = (U_b - l$ where and are the upper and lower bounds values of feature vectors respectively. Here the factor 0.01 is to make sure the random walks is not too aggressive, and this value has been obtained by a parametric study. For multiobjective optimization, one way is to combine all objectives into a single objective of feature vectors to measure the similarity so that algorithms for single objective optimization can be used without many modifications. For example, FA can be used directly to solve many feature similarity based problems in this manner, and a detailed study was carried out by Apostolopoulos and Vlachos [26]. Another way is to extend the firefly algorithm to produce Pareto optimal front directly. By extending the basic ideas of FA, can develop the following Multi-objective Firefly Algorithm (MOFA),

Algorithm 1 : Multiobjective feature similarity measure Firefly algorithm

Define objective functions $f_1(x), \ldots\ldots f_K(x)$ where $x = (x_1, \ldots\ldots x_d)^T$
Initialize a population of n feature vectors $x_i (i = 1, \ldots\ldots n)$
while $(t < MaxGeneration)$
for $i, j = 1, \ldots\ldots\ldots n$ (all n feature vectors)
Evaluate their approximations PF_i and PF_j to the Pareto front for feature vectors
if $i \neq j$ and when all the constraints are satisfied
if PF_j dominates PF_i,
Move firefly(feature vectors) i towards j using (7)
Generate new feature similarity measures ones if the moves do not satisfy all the constraints
end if
if no non-dominated feature vector based similarity matching solutions can be found
Generate random weights $w_k (k = 1, \ldots\ldots K)$
Find the best solution g_*^t (among all feature vectors) to minimize in (8)
Random walk around g_*^t using (9)
end if
Update the feature vector results and pass the non-dominated feature vector solutions to next iterations
end
Sort and find the current best approximation feature similarity resutl to the Pareto front
Update $t \leftarrow t + 1$
end while

Postprocess results and visualisation. The procedure starts with an appropriate definition of objective functions with SMD distance similarity values. In this work first initialize a population of n feature vectors so that they should distribute among the feature similarity search space as uniformly as possible. This can be achieved by using sampling techniques via uniform distributions. Once the tolerance or a fixed number of iterations is completed in the feature similarity measures defined, the iterations start with the evaluation of brightness or objective values of all the fireflies(feature vectors) and compare each pair of fireflies(feature vectors) . Then, a random weight vector is generated, so that a combined best solution feature similarity measure can be obtained. The non-dominated feature simliarity solutions are then passed onto the

next iteration. At the end of a fixed number of iterations, in general n non-dominated feature similarity solution points can be obtained to approximate the true Pareto front. In order to do random walks more efficiently, can find the current feature similarity best which minimizes a combined objective via the weighted sum

$$\psi(x) = \sum_{k=1}^{K} w_k f_k, \sum_{k=1}^{K} w_k = 1 \quad (8)$$

Here w_k = where are the random numbers drawn from a uniform distributed $Unif[0$ In order to ensure that $\sum_k w_k =$, a rescaling operation is performed after generating K uniformly distributed numbers. It is worth pointing out that the weights should be chosen randomly at each iteration, so that the non-dominated feature similarity solution can sample diversely along the Pareto front. If a firefly (feature vector) is not dominated by others in the sense of Pareto front, the firefly moves

$$x_i^{t+1} = g_*^t + \alpha_t \epsilon_i^t \quad \ldots (9)$$

where is the best feature similarity solution found so far for a given set of random weights. Furthermore, the randomness can be reduced as the iterations proceed, and this can be achieved in a similar manner as that for simulated annealing and other random reduction techniques.

$$\alpha_t = \alpha_0 0.9^t \quad (10)$$

where is the initial randomness factor for a minimization problem a feature similarity solution vector $u = (u_1, \ldots, u_i$ is said to determine another $v = (v_1, \ldots, v_i$ if and only if $u_i \leq$ for all $\forall i \in \{1, \ldots \ldots$ and $\exists i \in \{1, \ldots \ldots n\}: u_i \prec$ .In other words no component of u is larger than corresponding component of v and atleast one component is smaller .Similarly define another dominance relationship by,

$$u \preccurlyeq v \Leftrightarrow u \prec v \vee u = v \quad \ldots(11)$$

It is worth pointing out that for maximization problems, the dominance can be defined by replacing $\prec with$. Therefore, a point is called a non-dominated solution if no solution can be found that dominates it .The Pareto front PF of a multiobjective (feature similarity measures) can be defined as the set of non-dominated feature similarity solutions so that,

$$PF = \{s \in S | \exists\, s' \in S: s' \prec s \quad \ldots(12)$$

where S is the solution set. Then query retrieval results are obtained and the result of each method is measured using experimental results evaluation.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

To show the effectiveness of the proposed system, some experiments will be reported. Selecting a suitable image database is a critical and important step in designing an image retrieval system. At the present time, there is not a standard image database for this purpose. Also, there is no agreement on the type and the number of images in the database. Since most image retrieval systems are intended for general databases, it is reasonable to include various semantic groups of images in the database. In this work experimental results were compared by using the WANG database. WANG database is a division of 1,000 images of the Corel stock photo database which have been physically preferred and which outline 10 classes of 100 images every. The WANG database can be well thought-out comparable to universal stock photo retrieval tasks by means of numerous images from every category and a possible user having an image from a specific type and looking for similar images which have e.g. cheaper royalties or which have not been used by other media. The 10 classes are used for relevance estimation for a given query image, it is unspecified that the user is searching images for the same class, and consequently the remaining 99 images from the similar class are considered relevant and the images from all other classes are considered inappropriate.

To evaluate the effectiveness of the proposed approach, examined how many relevant images to the query were retrieved. The retrieval effectiveness can be defined in terms of precision and recall rates. Experiments are run five times, and average results are reported. In every experiment, an evaluation of the retrieval precision is performed so that ten images that were randomly selected from each specific category of the database are used as query images. For each query image, relevant images are considered to be those and only those which belong to the same category as the query image. Based on this concept, the retrieval precision and recall are defined as,

$$precision = \frac{N_{A(q)}}{N_{R(q)}} \quad (13)$$

$$Recall = \frac{N_{A(q)}}{N_t}$$

where N_A denotes the number of relevant images similar to the query, N_t indicates the number of images retrieved by the system in response to the query, and represents the total number of relevant images available in the database. When each precision and recall for ten images is obtained, discard the best value and the worst one and then average these values to obtain the average precision and average recall. Figure 2, observe that the

tendency of average precision and recall for the randomly selected images in each specific category is toward higher value with the proposed MOFA approach, and they can achieve higher precision values than K means clustering and Fuzzy k Means clustering.

Table 1: The Precision and Recall measurements computed for the CBIR system.

Precision vs recall	K means clustering	Fuzzy K means clustering	MOFA
0.1	0.7	0.9	0.95
0.2	0.65	0.84	0.89
0.3	0.61	0.81	0.85
0.4	0.54	0.74	0.82
0.5	0.53	0.71	0.79
0.6	0.45	0.67	0.77
0.7	0.38	0.58	0.7
0.8	0.35	0.42	0.54
0.9	0.27	0.36	0.41
1	0.15	0.23	0.27

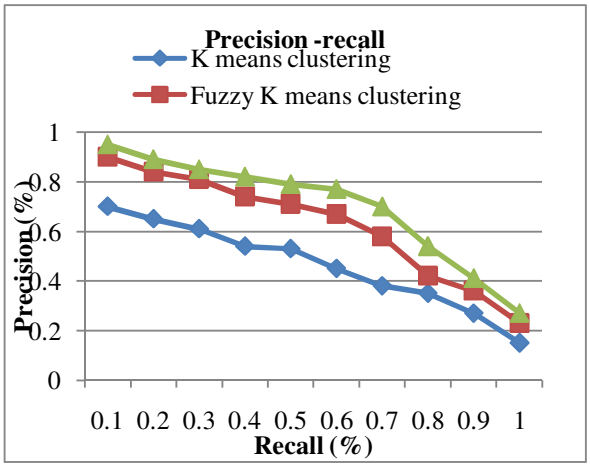


Fig. 2. The Precision and Recall measurements computed for the CBIR system.

Mean average precision (MAP). Mean average precision (MAP) is definite as the standard mean of precision values for user known query. $MAP = \frac{\sum_{q=1}^{Q} Ave}{Q}$, where Q is the number of queries. Figure 3, observe that the tendency of mean average precision and recall for the randomly selected images in each specific category is toward higher value with the proposed MOFA approach, and they can achieve higher of mean average precision values than K means clustering and Fuzzy k Means clustering.

Table 2: The Mean average Precision and Recall measurements computed for the CBIR system.

Precision vs recall	K means clustering	Fuzzy K means clustering	MOFA
0.1	0.69	0.91	0.93
0.2	0.63	0.85	0.89
0.3	0.58	0.83	0.87
0.4	0.52	0.77	0.84
0.5	0.52	0.75	0.79
0.6	0.43	0.69	0.78
0.7	0.39	0.54	0.71
0.8	0.37	0.43	0.53
0.9	0.29	0.39	0.44
1	0.23	0.29	0.33

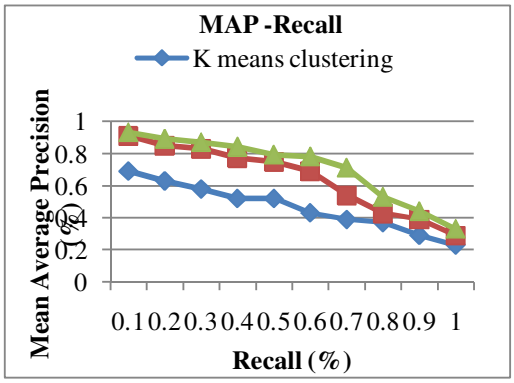


Fig. 3. The Mean average Precision and Recall measurements computed for the CBIR system.

F-measure

An f-measure identifies situations where IR results contain unnecessary information, called precision, and where the results do not contain enough information, called recall.

$$F - Measure = \frac{2 \times precision \times recall}{Precision + recall} \quad ...(14)$$

Figure 4, observe that the tendency of F measure for the randomly selected images in each specific category is toward higher F measure value for proposed MOFA approach for every features such as color ,texture and shape they can achieve higher of F measure values than K means clustering and Fuzzy k Means clustering. The values are tabulated in table 3.

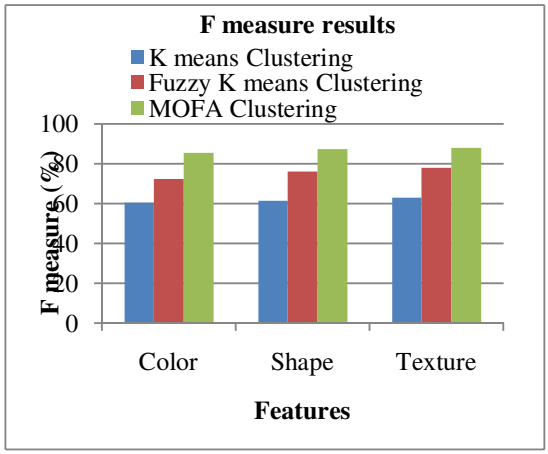


Fig. 4. The F measure measurements computed for the CBIR system.

Table 3: The F measures computed for the CBIR system.

Features	K means Clustering	Fuzzy K means Clustering	MOFA Clustering
	F measure (%)		
Color	60.52	72.35	85.68
Shape	61.25	76.2	87.52
Texture	63	78	88

V. CONCLUSION

Numerous feature extraction methods and similarity measures have been proposed in the literature but there is no generally applicable and effective framework to combine multiple features and similarity measures. In this paper, an image retrieval method using both color, texture feature based similarity measures has been proposed. Proposed Multi objective firefly optimization has used the HSV color features to improve conventional histogram features. A robust texture feature which is suitable for image retrieval has also been presented in the paper. In contrast to most conventional combined approaches which may not give better performance than individual features, our approach provides users with two alternatives, i.e., retrieval using color features only and retrieval using combined features ,best feature similarity measure estimated using MOFA then exact matching results are found. Experiments showed that the Multiobjective firefly optimization framework allows an effective way of combining feature vectors and similarity measures best matching results than K means clustering and Fuzzy K means clustering. A query image can be retrieved efficiently from a large database. In the future, plan to segment image automatically into homogenous texture regions using split and merging technique. By using regional features instead of global features, we will be able to improve retrieval performance of combined features further. We also plan to use more queries to test the retrieval performance.

REFERENCES

[1]. Ahmed S and S. Kanhere, VANETCODE: network coding to enhance cooperative downloading in vehicular ad-hoc networks, in Proceedings of International conference on Wireless Communications and Mobile Computing (IWCMC), 2006, pp. 527–532.

[2]. F. Long, H. Zhang, H. Dagan, and D. Feng, Fundamentals of content based image retrieval, in Multimedia Information Retrieval and Management: Technological Fundamentals and Applications, D. Feng, W. Siu, and H. Zhang, Eds., Berlin Heidelberg New York: Springer-Verlag, 2003, ch. 1, pp. 1-26.

[3]. Shrivastava et.al. Comparison between K-Mean and C mean Clustering for CBIR. Second International Conference on Computational Intelligence, Modelling and Simulation,2010.

[4]. T. Noreault, M. McGill, and M. B. Koll, A performance evaluation of similarity measures, document term weighting schemes and representations in a boolean environment, in 3rd ACM Conf. on Research and development in inforamtion retreval, SIGIR, 1980, pp. 57–76.

[5]. M. Kokare, B.N. Chatterji, and P.K. Biswas, Comparison of similarity metrics for texture image retrieval, in *IEEE Conf. on Convergent Technologies for Asia-Pacific Region,* 2003, vol. **2**, pp. 571–575.

[6]. D. Zhang and G. Lu, Evaluation of similarity measurement for image retrieval, in IEEE International Conf. on Neural Networks Signal, 2003, pp. 928–931.

[7]. Neumann, D., Gegenfurtner, K.R., 2006. Image retrieval and perceptual similarity. *ACM Trans. Appl. Perception* **3** (1), 31–47.

[8]. Li, B., Chang, E.Y., 2003. Discovery of a perceptual distance function for measuring image similarity. ACM Multimedia *J. Special Issue Content-Based Image Retrieval* **8** (6), 512–522.

[9]. Do, M.N., Vetterli, M., 2002. Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance. *IEEE Trans. Image Process.* **11**(2), 146–158.

[10]. Y. Liu, D. Zhang, G. Lu, andW.-Y.Ma, A survey of content-based image retrieval with high-level semantics, *Pattern Recognit.*, vol. **40**, no. 1, pp. 262–282, Jan. 2007.

[11]. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, Content-based image retrieval at the end of the early years, *IEEE Trans. Pattern Anal. Mach. Intell.,* vol. **22**, no. 12, pp. 1349–1380, Dec. 2000.

[12]. S. Antani, R. Kasturi, and R. Jain, A survey of the use of pattern recognition methods for abstraction, indexing and retrieval of images and video, *Pattern Recognit.,* vol. **35**, no. 4, pp. 945–965, Apr. 2002.

[13]. X. S. Zhou and T. S. Huang, Relevance feedback in content-based image retrieval: Some recent advances, *Inf. Sci.,* vol. **148**, no. 1–4, pp. 129–137, Dec. 2002.

[14]. H.-W. Yoo, H.-S. Park, and D.-S. Jang, Expert system for color image retrieval, *Expert Syst. Appl.,* vol. **28**, no. 2, pp. 347–357, Feb. 2005.
[15]. T.C. Lu and C.C. Chang, Color image retrieval technique based on color features and image bitmap, *Inf. Process. Manage.,* vol. **43**, no. 2, pp. 461–472, Mar. 2007.
[16]. A. Vadivel, S. Sural, and A. K. Majumdar, An integrated color and intensity co-occurrence matrix, *Pattern Recognit. Lett.,* vol. **28**, no. 8, pp. 974–983, Jun. 2007.
[17]. Chih-Chin Lai, Member, IEEE, and Ying-Chuan Chen, A User-Oriented Image Retrieval System Based on Interactive Genetic Algorithm, *IEEE transactions on instrumentation and measurement,* vol. **60**, no. 10, october 2011.
[18]. Gwénolé Quellec, Mathieu Lamard, Guy Cazuguel, Member, IEEE, Béatrice Cochener, and Christian Roux, Fellow, IEEE Adaptive Nonseparable Wavelet Transform via Lifting and its Application to Content-Based Image Retrieval IEEE transaction on Image Processing 2010.
[19]. Giacinto, G., Roli, F., 2004. Nearest-prototype relevance feedback for content based image retrieval. In: ICPR'04: Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04), vol. **2.** *IEEE Computer Society* Washington, DC, USA, pp. 989–992.
[20]. Da S. Torres, R., Falcpo, A.X., Goncalves, M.A., Zhang, B., Fan, W., Fox, E.A., Calado P., 2005. A new framework to combine descriptors for content-based image retrieval. In: Fourteenth Conference on Information and Knowledge Management, Bremen, Germany, pp. 335–336.
[21]. Haridas, K., & Thanamani, A. S. (2014). An Efficient Image Clustering and Content Based Image Retrieval Using Fuzzy K Means Clustering Algorithm.*International Review on Computers and Software (IRECOS)*, **9**(1), 147-153.
[22]. Bruno, E., Kludas, J., Marchand-Maillet, S., 2007. Combining multimodal preferences for multimedia information retrieval. In: MIR'07: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, ACM, New York, NY, USA, pp. 71–78.
[23]. VE. Chandra and K.Kanagalakshmi, Performance Evaluation of Filters in Noise Removal of Fingerprint Image, Proceedings of ICECT-2011, *3rd International Conference on Electronics and Computer Technology,(2011),* vol. **1**, pp. 117-123, ISBN: 978-1-4244-8677-9, Published by IEEE, Catalog no.: CFP1195F-PRT.
[24]. Poorani M1, Prathiba T2, Ravindran G3 Integrated Feature Extraction for Image Retrieval *IJCSMC*, Vol. **2**, Issue. 2, February 2013, pg.28 – 35
[25]. M.A. Tahir, A. Bouridane, F. Kurugollu, A. Amira, Accelerating the computation of GLCM and Haralick texture features on reconfigurable hardware, *Int. Conf. on Image Processing,* vol. **5**, pp 2857-2860, 2004.
[26]. Yang X.-S., (2010). Firefly algorithm, stochastic test functions and design optimisation, *Int. J. Bio-inspired Computation,* **2**(2), 78-84.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 121-128(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Neural Network Based Email Classification Using Tensorflow

S. Kanimozhi[1] and V. Bhuvaneswari[2]

[1]M.Phil Research Scholar, Department of Computer Applications, Bharathiar University (TN), India
[2]Assistent Professor Department of Computer Applications, Bharathiar University (TN), India

ABSTRACT: Artificial intelligence is based on human brain and imparts intelligence. Machine learning is a branch of artificial intelligence which has emerging area using massive volume of data. Deep learning is subset of machine learning which make the computation of multi-layer neural network feasible. Deep learning applications various areas are, Computer Vision for image classification, segmentation, Speech Recognition, Natural Language Processing in sentiment analysis, translation, Biology in protein structure prediction, analysis of genomic data. This paper a framework and methodology is designed for email subject classification using Tensor flow the deep learning tool. The framework consists of three phase where the first phase is data modelling and the second phase is text preprocessing to classify the subjects of email extracted as text messages and text phrases pre-processed using text preprocessing libraries using NLTK libraries. The processed text are classified using the Tensorflow neural network library. The model is evaluated and varied based on precision recall. The result are evaluated and found that 85 % accuracy.

Keyword: Machine learning, Deep learning, neural network, Tensorflow.

I. INTRODUCTION

Artificial intelligence is branch of computer science used for creating machines with intelligence as human being. Artificial intelligence is based on the architecture of human brain and imparts intelligence, to machines as human learns and design in problem solving [8]. Artificial intelligence can be created by making computing machine think like human through software as machine control robots. Artificial intelligence is applied in domains of computer vision, medical outcome analysis, and computer biology [7]. Machine learning is a branch of artificial intelligence which has emerged as important area. Arthur Samuel coined the term machine learning in 1959 at IBM [3].

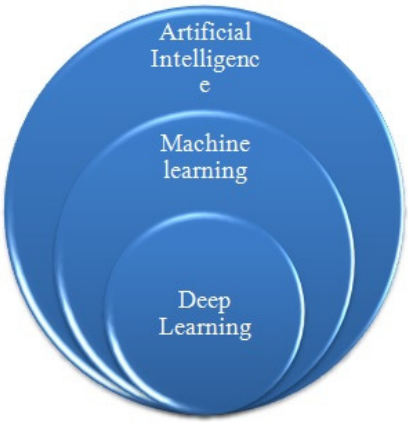


Fig. 1. Domain of artificial intelligence.

Machine learning is evolved from the study of pattern recognition, computer learning where the focus is on construction of algorithm, concerned with the design that can learn and make prediction of data [4]. Machine Learning algorithms are applied on massive volume of data to optimize performance. The processes involved in machine learning are similar to that of data mining and predictive modelling [3].

A. Machine Learning Techniques

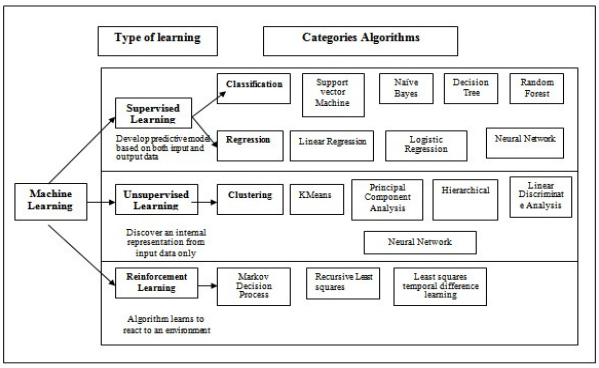


Fig. 2. Machine learning techniques.

Techniques are categorized based on the learning mechanism. Figure 2 provide with the details describing view on machine learning techniques based on machine learning patterns [5]. The section below provide with a brief overview on the various techniques used in machine learning.

B. Machine Learning Applications

Table 1: Machine learning applications.

Machine Learning algorithm and types	Uses	Analytics			
		Descriptive	Diagnostic	Predictive	Prescriptive
Supervised (inductive) learning applications Classification: Regression Applications	**Systems Biology** for gene expression microarray data	✓			
	Spam detection.		✓		
	Face detection: Character Recognitionare different handwriting styles.	✓			
	Medicine: Diagnosis a tumor is cancerous or benign. Medical diagnosis from symptoms to illnesses		✓		
	Banking: Credit/loan approval.		✓		✓
	Web Advertising is predict if a user clicks on an ad on the Internet.			✓	
	Systems Biology :based on gene expression,	✓			
	Movies ratings: Review of movie rate.			✓	
	Marketing: Segmentation of the customer based on behaviour.		✓	✓	
	Image processing: Identifying objects on an image (face detection)	✓			
	Earth-quake studies: Observed earth quake epicenters should be clustered along continent faults		✓	✓	
	Medical diagnosis: From symptoms to illnesses	✓			
	Biometrics: Recognition/authentication using physical and/or behavioral characteristics: Face, iris, signature, etc				

This paper presents a deep overview on deep learning models and tools. The main objective of this paper is to classify email subjects using neural network with tensor flow. The paper is organized as follows. Section 2 provides an brief overview on deep learning models, with their architecture and applications. Section 3 discusses the framework and methodology for classifying email subjects by creating neural network model using Tensorflow. Section 4 discusses the results and discussion followed by conclusion in section 5.

II. DEEP LEARNING - OVERVIEW

Deep Learning models contains neural network with several layers of nodes between input and output the series of layers between input and output and feature identification and sequence process [10]. It is a powerful class of machine learning model.

A. Deep Learning Model View

Huge evolution of data in BigData era requires processing different variety of data. The conventional machine learning algorithm [9], classification and clustering has challenges in arriving complex patterns of data. Artificial intelligence technique neural network is used to process complex data [11]. Neural network is an algorithm based on the concept of human brain is given in Figure 3. This section provide with brief view on traditional neural network architecture to complex deep learning models.

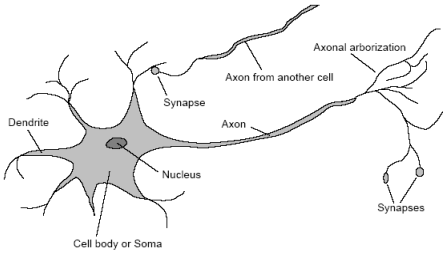


Fig. 3. Structure of neurons in brain.

Neurons the nerves of human brain consist of four parts are, dendrites, axon, synaptic and soma [12].

B. Artificial Neural Networks Architecture

A typical neural network contains three different layers, input layer, hidden layer and output layer (is given in Figure 4).

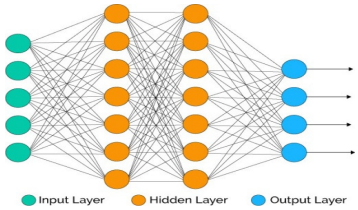


Fig. 4. Artificial Neural Network Architecture.

Input layer receives input from the input based on the trained data, the output layer contains units that respond to the information, and hidden units are between input and output layers.

A. Artificial Neural Network Model

Artificial neural networks are viewed as weighted directed graphs where neuron or nodes and directed edges with weights are connections between neuron inputs and neuron output (is given in Figure 5).

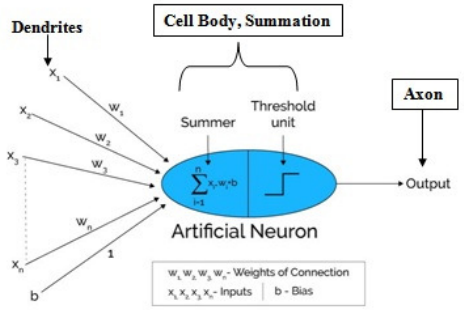


Fig. 5. Artificial Neural Network Model.

The inputs are calculated based on mathematical notation which is multiplications or addition corresponding weights [13]. The weights represent the strength of the interconnects structures in the neural network, when the weighted input sum is zero bias value as added, to make the output not zero [21]. The activation function are binary, sigmoid (linear) and tan hyperbolic sigmoid function (nonlinear).

C. Two Major Deep Learning Models

A) Deep Belief Networks: A deep belief network (DBN) is a probabilistic, generative model made up of multiple layers of hidden units [22]. Compositions of the simple learning modules are made up each layer. A deep belief network can be used to generatively pre-train a deep neural network by using the learned deep belief network weights [1].

Restricted Boltzmann Machine: A Deep Brief Network can be efficiently trained in an unsupervised, layer-by-layer mode, and the layer is normally made of restricted Boltzmann machines (RBM). Undirected, generative energy based model with a visible input layer and a hidden layer, and connections between the layers but not within layers. Boltzmann machines are then trained with the procedure is given Figure 6.

This whole process is repeated until some desired stopping condition.

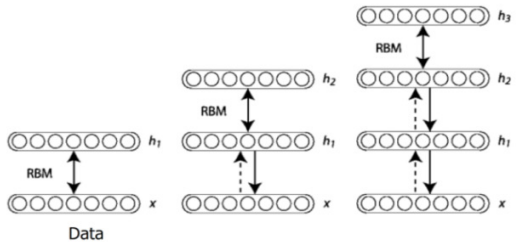


Fig. 6. Many to many connections.

B) Autoencoder: An autoencoder neural network is an unsupervised learning algorithm that applies backpropagation, set the target values to be equal to the inputs. Autoencoder purpose is dimensionality reduction in the aim is to learn a representation (encoding) for a set of data. Autoencoder concept has widely used for learning generative models of data. The simplest form of an autoencoder is a feedforward, non-recurrent neural network very similar to the multilayer perceptron (MLP) [23]. Autocoders an input layer, an output layer and one or more hidden layers connecting them. The output layer has the same number of nodes as the input layer with the purpose of reconstructing instead of predicting the target value given the inputs [24]. Autoencoder is used to the compression and decompression functions are implemented with neural networks.

D. Deep Learning for Text Mining

Structure of deep learning models on textual data requires it representation of the basic text unit, word [14]. Deep learning models for text mining used to vector representation of words. Neural network structures are Recurrent Neural Network and Recursive Neural Network [15].

***A) Recurrent Neural Networks*:**

The applications of standard Neural Networks are limited due to the only accepted a fixed-size vector as input and produce a fixed-size vector as output [2] (e.g. the number of layers in the model) is given Figure 7.

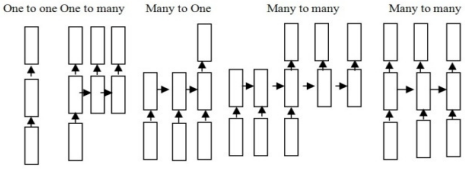


Fig. 7. Recurrent Neural Networks.

Each rectangle is a vector and arrows represent function and input vectors are in bottom to up hold the recurrent state [16]. Recurrent Neural Networks are unique as they allow us to operate over sequences of vectors.

B) Deep Learning Resources:

Table 2: Deep Learning Resources.

Name	Language	Link	Note
Pylearn2	Python	http://deeplearning.net/software/pylearn2/	A machine learning library built on Theano
Theano	Python	http://deeplearning.net/software/theano/	A python deep learning library
Caffe	C++	http://caffe.berkeleyvision.org/	A deep learning framework by Berkeley
Torch	Lua	http://torch.ch/	An open source machine learning framework
Overfeat	Lua	http://cilvr.nyu.edu/doku.php?id=code:start	A convolutional network image processor
Deep learning 4j	Java	http://deeplearning4j.org/	A commercial grade deep learning library
Word2vec	C	https://code.google.com/p/word2vec/	Word embedding framework
Doc2vec	C	https://radimrehurek.com/gensim/models/doc2vec.html	Language model for paragraphs and documents
Stanford NLP	Java	http://nlp.stanford.edu/	A deep learning-based NLP package
Tensorflow	Python	http://www.Tensorflow.org	A deep learning based python library

III. EMAIL SUBJECT CLASSIFICATION: TENSORFLOW

Classification – An Overview:

Classification is important machine learning algorithm applied in various applications to predict patterns [17]. Deep learning models are used design neural network based design architecture for classification. Neural networks are applied in various domains for supervised learning [18]. The objective of this work is to design neural network architecture. The methodology and framework is given in Figure 8.

A. Methodology and Framework

The framework and methodology for email subject classification is given in Figure 8. The framework consists of three phases. Phase I Data Modelling, Phase II Text Preprocessing and Neural Network Model Design, Phase III Prediction and Validation.

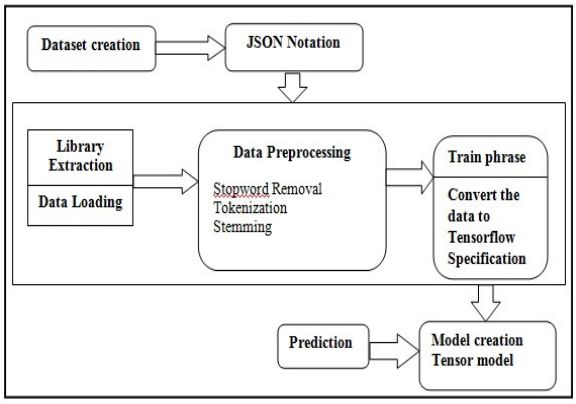


Fig. 8. Workflow of Text Classification Using Tensorflow.

Phase I: Data Modeling

Dataset : The dataset is extracted from email from users for classifying into five categories subject heading as conference, course, social media, friend, meeting, business, entertainment, tourist, news and alert related details [31]. The data collected is represented using JSON.

Data Modeling

JSON: Notation: JSON (JavaScript Object Notation) model categories are represented as the key/value pair is separated by a comma and the corresponding subject of email are represented as values.

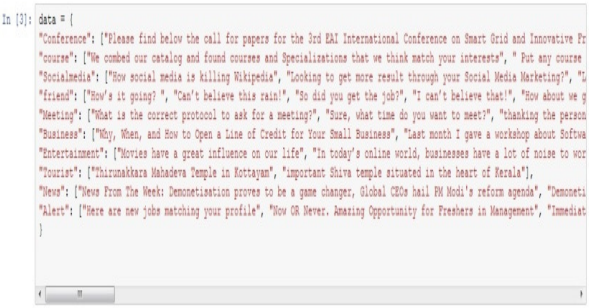


Fig. 9. JSON Dataset.

The corresponding JSON library is imported for converting text documents to JOSN format. The dataset converted into a simple JSON file that for training the neural network for email subjects classification is given in Figure 9.

Text Pre-processing. The subjects of email are extracted as text phrases represented in JSON and pre-processed using text pre-processing libraries [27]. The needed NLTK libraries are imported for text pre-processing of punctuation, stemming, tokenization, bag words of word representation.

A . Data Load and Text Preprocessing: **Stemming:** Stemming is a process used to identify the stem words which is used for classifying the sentences [28]. The Lancaster stemmer is used to do the stemming as given in #1.

#1

```
tbl = dict.fromkeys(i for i in range(sys.maxunicode)
def remove_punctuation(text):
  return text.translate(tbl)
```

Tokenization: The punctuation is removed from the sentences for tokenization. Tokenization is used to split [29] the sentence in to tokens using nltk library and the tokens are used to represent bag words to create text classification for sentence [30] as given in #2.

#2

```
# get a list of all categories to train for
categories = list(data.keys())
words = []
# a list of tuples with words in the sentence and category name
docs = []

for each_category in data.keys():
 for each_sentence in data[each_category]:
# remove any punctuation from the sentence
  each_sentence = remove_punctuation(each_sentence)
  print (each_sentence)
# extract words from each sentence and append to the word list
 w = nltk.word_tokenize(each_sentence)
 print ("tokenized words: ", w)
 words.extend(w)
```

Stemming and Removing Duplications: Duplication removal of tokens is given in #3.

#3

```
# stem and lower each word and remove duplicates
words = [stemmer.stem(w.lower()) for w in words]
words = sorted(list(set(words)))
print (words)
print (docs)
```

B. Neural Network Modelling- Email Classification

Neural network is created with twenty input nodes, eight hidden layer and five output layers is given in Figure 10.

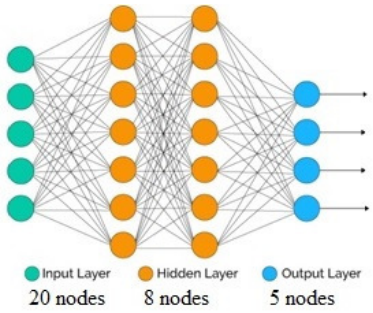


Fig. 10. Neural Network Model.

Table 3: Neural Network Layers: Email Classification.

Neural Network Model	**Number Of Nodes**	**Tensors**
Input Layer	Twenty Nodes	Conference, course, social media, friend, meeting, business, entertainment, tourist, news, alert, insurance, policy, certificate, news letter, Big data, library, book, certificate, publications, analytics.
Hidden Layer	Eight Nodes	Insurance, policy, certificate, news letter, Big data, library, book, certificate.
Output Layers	Five Nodes	Business, Friend, Conference, News, Tourist.

Neural network is created with twenty input nodes, eight hidden layer and five output nodes. The nodes in the graph are called ops (short for operations). An operation takes in zero or more tensors. Twenty five tensors are used to create the neural network model for this email classification. Creating a training dataset to train the neural network model as given in #4.

#4

```
# create our training data
training = []
output = []
output_empty = [0] * len(categories)
 training.append([bow, output_row])
random.shuffle(training)
```

Network Training Module: The neural network model is devised with varying training epochs to predict the email subject using softmax activation function. Learning module of Tensorflow in epoch 2000 as given in #5.

#5

```
#reset underlying graph data
tf.reset_default_graph()
# Build neural network
net = tflearn.input_data(shape=[None, len(train_x[0])])
net = tflearn.fully_connected(net, 8)
net = tflearn.fully_connected(net, len(train_y[0]), activation='softmax')
net = tflearn.regression(net)
# Define model and setup tensorboard
model = tflearn.DNN(net, tensorboard_dir='tflearn_logs')
model.fit(train_x, train_y, n_epoch=2000, batch_size=8, show_metric=True)
model.save('model.tflearn')
```

The results of the neural network for email classification is presented in Results and Discussion section given below.

IV. RESULTS AND DISCUSSION

A neural network is devised to classify email subjects using Tensorflow deep learning tool. The neural network is trained for various learning epoch. The performance of the network is tested with the test data. The learning performance of the architecture and prediction accuracy for the test dataset is discussed in this section.

#6 Learning Rate

```
Training Step: 3999  | total loss: 0.81059
| Adam | epoch: 2000 | loss: 0.81059 - acc: 0.8950 -- iter: 08/10
Training Step: 4000  | total loss: 0.73632 | time: 0.016s
| Adam | epoch: 2000 | loss: 0.73632 - acc: 0.9055 -- iter: 10/10
```

#6 presents the learning time for the test data with 25 instances.

```
sent_1 = "catalog"
sent_2 = "I gotta go now"
sent_3 = "Library OPAC Transaction using Hive  A Big Data Approach for book chapter"
sent_4 = "you must be a couple of years older then her!"
```

A. Prediction

The model is tested to predict the output of neural network for the test data. The prediction accuracy is validated. The snapshot of prediction using the neural network devised is given in Figure 11.

Tesorflow Using Email Subject Classification:

```
# we can start to predict the results for each of the 4 sentences
print( categories[np.argmax(model.predict([get_tf_record(sent_1)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_2)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_3)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_4)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_5)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_6)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_7)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_8)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_9)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_10)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_11)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_12)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_13)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_14)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_15)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_16)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_17)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_18)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_19)]))])
print( categories[np.argmax(model.predict([get_tf_record(sent_20)]))])
```

```
Business
friend
Conference
News
Tourist
Meeting
News
friend
friend
Business
News
Tourist
```

Fig. 11. Tesorflow using Email Subject Classification.

B. Validation Precision and Recall

The Neural network is verified using the F-Measure. F-Measure which is the harmonic mean of the precision and recall. The formula for the corresponding Precision, Recall in equation 1, equation 2, equation 3 respectively.

For any topic T and cluster X:

N1 = Number of documents judged to be of topic T in cluster X.
N2= Number of documents in cluster X.
N3=Number of document to be judged to be of Topic T in entire hierarchy.

$$\text{Precision} = \frac{N1}{N2} \quad \rightarrow \text{equation1}$$

$$\text{Recall} = \frac{N1}{N3} \quad \rightarrow \text{equation 2}$$

$$\text{F – Measure} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad \rightarrow \text{equation 3}$$

Neural Network Validation Accuracy:

Table 4: Validation Accuracy: Learning Rate 1000 epochs.

Cluster	**Precision (%)**	**Recall (%)**	**F-measure (%)**
Business	60	30	70
News	40	10	50
Social media	50	20	60
Course	35	10	20
Conference	50	10	50
Friends	80	30	80
Meeting	50	20	70
Tourist	20	0	10

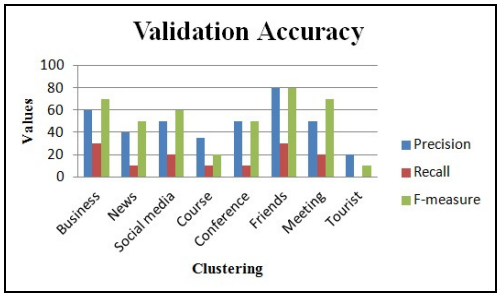


Fig. 12. Prerdiction by the values of epach 1000

Table 5: Validation Accuracy : Learning Rate 2000 epochs.

Cluster	**Precision (%)**	**Recall (%)**	**F-measure (%)**
Business	60	20	50
News	30	0	60
Social media	70	10	30
Course	30	10	20
Conference	50	20	50
Friends	80	30	80
Meeting	40	10	70
Tourist	10	20	30

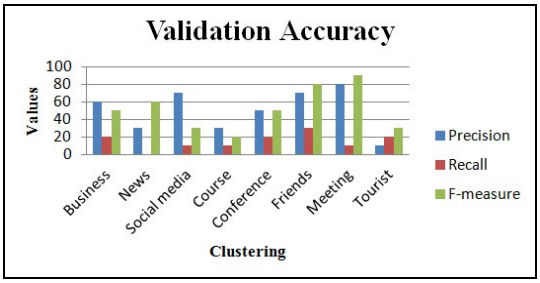


Fig. 13. Prediction by the value of epach 2000.

The Figure 12 and Figure 13 provides with a snapshot of the model is tested to predict the output of neural network for the test data [35]. The prediction of the nework was found to be varied based on the learning rate. The network is trained with 100 email subjects. The model is tested with 50 new test dataset. The overall precison and recall for the test data is found to be 85% for classifying email subject categories.

V. CONCLUSION

The exponential increase in volume of data has necessitude evolving of new algorithm and model. Machine learning algorithms are widely used for processing hug volume of data. Deep learning models are used widely to overcome the limitations of conventional neural network model. The objective of this work is to have a brief study on machine learning models and tools. This work detail study on machine learning library TensorFlow is carried out configured and tested. A neural network model with TensorFlow is design to classify subject headings of email into varies categories. The model is evaluated and varied based on precision recall. It is found from the result the model for various runs of training epochs 2000 to 500 epoch classified with % 85 as overall accuracy. In future the model will be extended with text based deep learning models for automatic classifying of email subjects in to categories which may be automatically labeled into folders.

REFERENCES

[1] Abhineet Saxena, Guru Gobind Singh, *et.al.,* "Convolutional neural networks: an illustration in TensorFlow", Volume **22**, 2016.

[2] Adam L, Berger, et.al., "A maximum entropy approach to natural language processing", March 1996, **22**: 39–71.

[3] Andrew Arnold, Ramesh Nallapati, et.al., "A comparative study of methods for transductive transfer learning", In Proceedings of the Seventh IEEE International Conference on Data Mining Workshops, 2007.

[4] Anish Talwar, Yogesh Kumar, et.al., "Machine Learning: An artificial intelligence methodology", International Journal Of Engineering And Computer Science, Volume 2 Issue 12, Dec.2013, 2319-7242

[5] Anna Rozeva, et.al., "Classification of text documents supervised by domain ontologies", Volume 8, November 2012.

[6] Aurangzeb Khan, Baharum Baharudin, et.al., "A Review of Machine Learning Algorithms for Text-Documents Classification", VOL. 1, February 2010

[7] Bo Pang, Lillian Lee, et.al., "sentiment classification using machine learning techniques", Volume 10, 2002

[8] Fabrizio, Sebastiani, et.al., "Machine Learning in Automated Text Categorization", April 2001,

[9] Fabrizio, Sebastiani et.al.,"Machine learnig Text classification", Volume 34, March 2002.

[10] Geert Litjens, Thijs Kooi, et.al., " A survey on deep learning in medical image analysis", Elsevier, 2017

[11] Illya Sutskever, James Marten, et.al., "On the importance of initialization and momentum in deep learning", 2013.

[12] Jing Bai, Ke Zhou, et.al., "Multi-task learning for learning to rank in web search", 2009.

[13] Jurgen, Schmidhuber, et.al., "Neural Networks", Published by Elsevier Ltd, 2015

[14] Kamal Nigam, Andrew Kachites Mccallum, "Text Classification from Labeled and Unlabeled Documents using EM", February 20, 1999

[15] Kotsiantis, Ikonomakis, et.al., "Text Classification Using Machine Learning Techniques", Volume 4, August 2005.

[16] Krendzelak, Jakab, et.al., "Text categorization with machine learning and hierarchical structures", IEEE , (ICETA)05 September 2016).

[17] Ladislav Rampase, Anna Goldenberg1, et.al., "Tensor Flow: Biology's Gateway to Deep Learning", Elsevier Inc., 2, January 27, 2016

[18] Monostori, Markus, H. Van Brussel, E. Westkämpfer, "Machine learning approaches to manufacturing", CIRP Annals, Manufacturing Technology, Volume 45, 1996

[19] Mart´ın Abadi, Paul Barham, et.al., "TensorFlow: A system for large-scale machine learning", (OSDI), November 6, 2016

[20] Martin Abadi, et.al., "TensorFlow: learning functions at scale", November 2016.

[21] Martin Abadi, Andy Chu, et.al., "Deep Learning with Differential Privacy", 2016

[22] MigelD. Tissera,. et.al., "Deep extremelearningmachines:supervisedautoencoding architectureforclassification", August 2015.

[23] Mohammed Abdul Wajeed, T.Adilakshmi et.al., "Text classification using machine learning", Journal of Theoretical and Applied Information Technology, 2009.

[24] Nitish Srivastava, Ruslan Salakhutdinov, et.al., "Multimodal Learning with Deep Boltzmann Machines", 2012.

[25] Ricky J. Sethi, Yolanda Gil, et.al., "Future Generation Computer Systems",elsevier, 2017,256-270

[26] Rie Kubota Ando, Tong Zhang. et.al., "A framework for learning predictive structures from multiple tasks and unlabeled data", Journal of Machine Learning Research December 2005, 1817–1853

[27] Rie Kubota Ando, Tong Zhang, et.al., "Learning Predictive Structures from Multiple Tasks", 2006

[28] Shai Ben-David, John Blitzer, et.al., "Analysis of representations for domain adaptation ", 2006.

[29] Sathya, Annamma Abraham et.al., "Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification", International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No. 2, 2013.

[30] Steffen Bickel, Michael Br¨uckner, et.al., "Discriminative learning for differing training and test distributions", (ICML) 2007.

[31] Steffen Bickel, Michael Br¨uckner, et.al.,. "Discriminative learning for differing training and test distributions", (ICML) 2007.

[32] Sumit Das, Aritra Dey, et.al., "Applications of Artificial Intelligence in Machine Learning: Review and Prospect", Volume 115 – No. 9, April 2015

[33] Vassili Kovalev, Alexander Kalinovsky, et.al., Sergey Kovalev, "Deep Learning with Theano, Torch, Caffe, Tensorflow, and Deeplearning4J: Which One is the Best in Speed and Accuracy", 2016

[34] Weike PanEmail, Erheng Zhong, et.al.,, "Transfer Learning for Text Mining", Springer, 07 January 2012.

[35] Yann LeCun, Yoshua Bengio, et.al., "Deep learning", 28 May 2015

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 129-134(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Hybrid Framework of Image Source Identification using Image Features with Conditional Probability Features

A. Jeyalakshmi[1] and Dr. D. Ramya Chitra[2]
[1]Associate Professor, Department of Computer Science,
Sri Ramakrishna College of Arts and Science, Coimbatore (Tamil Nadu), India
[2]Assistant Professor, Department of Computer Science,
Bharathiar University, Coimbatore (Tamil Nadu), India

ABSTRACT: At the advent of current era, digital images plays an important role in our daily life, even the uneducated person also has the smart phone and digital camera. From school level onwards the students are using the smart phone. They also take friendly photos with their friends. At last by some misunderstanding, one gender threatens the other person. In this stage, identifying the originality of the digital image and its acquisition device model has become more important in today's digital world. This paper studies the recent developments in the field of image source identification (ISI). Researchers have explored many methods to identify the image source of a given image using color filter array (CFA), lens distortion, demos icing artifacts, wavelet transforms. The methods and algorithm of the proposed approaches in each category is described in detail to use accurate technique in image source identification. The classification has been done by support vector machine (SVM).The simulation results produce a good classification, which proves the efficiency of this method.

Keywords: Image Source Identification, Image Features, Conditional Probability, SVM.

I. INTRODUCTION

With the availability of powerful software, digital images can be manipulated easily even by amateurs and the alterations may leave no observable traces. Blindly determining the origin and model of the image source is an important task in the image forensics, because digital images are often used as evidence during criminal investigation and intelligence in court scenarios. This paper focuses on the Image Source Identification (ISI) problem. This problem focused a lot of attention in the recent past, as witnessed by the several publications available in the literature, such as [1-5]. One of the most noteworthy contributions in this area is the method proposed by Kai san choi e*t al.* in [3].

A. Image acquisition in Digital Camera

Mostly, all digital cameras have the same architecture and general processing steps. The general structure of image formation process is illustrated in the Figure 1, after light enters into the digital camera through the lens system, a set of filters are processed. In that, anti-aliasing filter is one of the important filters, The CCD detector is used to measure the intensity of light at each pixel location on the detectors surface.

The sophisticated cameras use a separate CCD for each of the three color (RGB) channels; so, most of the manufactures use a single CCD detector at every pixel, but partition its surface with different spectral filters. Such filters are called Color Filter Arrays (CFA).
Shown in Figure 2(a) and 2(b) are CFA patterns using RGB and YMCG color space respectively for a 6x6 pixel block. Looking at the RGB values in the CFA pattern it is evident that the missing RGB values need to be interpolated for each pixel. There are a number of different interpolation algorithms which could be used and different manufacturers use different interpolation techniques.

After color decomposition is performed by CFA, a detector is used to obtain a digital representation of light intensity in each color band. Then a number of operations are done by digital camera which includes interpolation, gamma correction, color processing, white point correction, and image compression. Although the operations and stages are common to all digital cameras, the exact processing detail in each stage varies from one manufacturer to the other, and even in different camera models manufactured by the same company. Hence, it is important to know the source camera that has been used to capture the given image, in order to detect image forgeries.

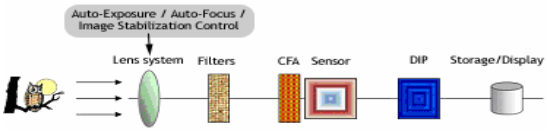


Fig. 1. Major Steps of image formation process in camera pipeline [4].

R	G	R	G
G	B	G	B
R	G	R	G
G	B	G	B

G	M	G	M
C	Y	C	Y
M	G	M	G
C	Y	C	Y

(a) RGB (b) YMCG

Fig. 2. CFA Pattern.

This paper is organized as follows: overview of recent developments in the field of image source identification(ISI) techniques in section 2, section 3 describes Proposed method of the different features extracted for image source identification(ISI), Experimental results and analysis for different feature extraction set is provided in section 4 and section 5 concludes this paper.

II. TYPICAL IMAGE SOURCE IDENTIFICATION (ISI) METHODS

The different methods and features that are used to classify image source models are given based on the differences in processing elements and the technologies. The disadvantage of these methodologies, in general, Most of the manufactures use common components, therefore processing pipeline remain the same or very similar among different models of a brand. Hence, identification of image source depends on classification of various model relevant characteristic.

A. Use of Demosaicing and Color Filter Array (CFA):

Bayram *et al* [4][6] propose a method to identify demosaicing artifacts associated with different camera-models. They have proposed two methods to describe a set of image characteristics, which are used as features in designing classifiers that distinguish between digital camera models. This work concentrated to identify, detect and classify traces of demosaicing operations. First the method estimates the differences in image formation pipeline, like processing techniques and device technologies. Next, it finds out the unique characteristics of camera model using Expectation Maximization (EM) algorithm [7]. The EM algorithm has been applied only on red channel. Using the EM algorithm, two sets of features have been obtained for classification: the weighting (interpolation) coefficients from the images and the peak location magnitudes in the frequency spectrum of the probability maps. Also, they have applied the sequential forward floating search (SFSS) algorithm to reduce the dimensionality of the feature vector by selecting the most distinguishing features.

B. Use of Sensor Imperfection

Geradts *et al* [8] proposed a method which includes pixel defects, hot points, dead pixels and cluster defects. The result shows that different model has different sensor pattern. However, it also noted that the amount of defects in the pixels for a camera varies between pictures and the content of the image. Finally it concludes that, the number of defects varied at different temperature, so, not all camera have the same problem, some of the camera has a mechanism to remove the sensor pattern noise from the captured image. Lukas *et al* [1] proposed sensor pattern noise based method. In which photo response non-uniformity (PRNU) casts a unique pattern onto every image the camera captures. This "camera fingerprint" is unique for each camera [2]. The camera fingerprint can be estimated from images known to have been taken with the camera.

$$I = I_o + I_o K + \Theta \qquad \ldots(1)$$

In this equation (1) the camera output image I is the "true scene" image that would be captured in the absence of any imperfections as I_0, and K is the PRNU factor (sensor fingerprint) , Θ includes all other noise components, such as dark current, shot noise, readout noise, and quantization noise.

The fingerprint K can be estimated from *N* images $I^{(1)}$, $I^{(2)}$, $I^{(3)}$..... $I^{(N)}$ taken by the camera. Let $W^{(1)}$, $W^{(2)}$, $W^{(3)}$… $W^{(N)}$, are their noise residuals obtained using a de noising filter F.

$$W^{(i)} = I^{(i)} - F(I^{(i)}) \qquad (2)$$

i=1….N and the PRNU factor, K has been derived as:

$$\hat{K} = \frac{\sum_{i=1}^{N} W^{(i)} I^{(i)}}{\sum_{i=1}^{N} (I^{(i)})^2} \qquad \ldots(3)$$

In both the patterns, the authors have tested 9 camera models where two of them have similar CCD and two are exactly the same model. The camera identification is accurate even for cameras of the same model. The result is also good for identifying compressed images. One problem with the conducted experiments is that the authors use the same image set to calculate both the camera reference pattern and the correlations for the images.

C. Use of wavelet transform

Using wavelet statistic features, Wang *et al* [9] proposed a method for source camera identification.

The most significant in the identification process, the frequency domain features preferred over spatial features like image color, image quality metrics and color filter array (CFA). In the wavelet domain features extraction they succeed they succeeded in distinguishing different models of the same camera brand.

D. Use of image features

Features of image somewhat applicable for the source camera identification, which involves simplifying the amount of resources required to describe a large set of data accurately, because features are common to all the images. Their accuracy differs when they were acquired. Mehdi Kharrazi *et al.* [10] proposed a method to extract 34 features from images, which are taken from different camera models and categorized into 3 groups like color features, used to train and test the classifier. In this method the result is good for uncompressed images. Also, good in the jpeg images but the accuracy dropped if the number of camera models is increased. Choi *et al* [3] have proposed a stepwise discriminate analysis method for extracting the features from digital images, which is advanced than analysis of variance (ANOVA).

III. FEATURE EXTRACTION

Digital images are customized nowadays, which can be captured by either mobile camera or digital camera. Their accuracy differed based on their captured devices.

A. Conditional Probability Features

Wahab *et al.* [12] used the conditional probability as a single feature set to classify camera models. By using this feature in intra camera model they were achieved 92.5% average accuracy. Instead of using in intra camera mode. In this paper, we apply conditional probability features in the order of non mutually exclusive events in the inter camera model classification. Figure 3 illustrates the P(A),P(B) and P(AB) in the tree diagram[11].

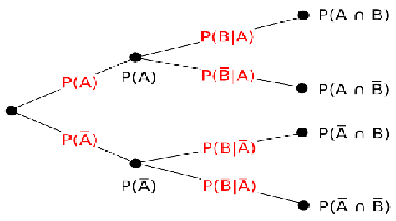


Fig. 3. Tree Diagram conditional probability P(A),P(B) and P(AB)[11].

Based on the concept of conditional probability ,from the block wise DCT coefficient select absolute values from three different locations in the order of five orientations. Combination of this with three A-events and three B-events we can get 45 conditional probability features for the identification of source camera model.

B. Color Moments

Color moments [3] are measures that characterize color distribution in an image in the same way that central moments uniquely describe a probability distribution. These are very effective for color-based image analysis. The lower-order moments generally provide enough information for image classification. The first color moment can be interpreted as the average color in the image, and it can be calculated by using the equation (4)

$$E_i = \frac{1}{N}\sum_{j=1}^{N} p_{ij} \tag{4}$$

where N is the number of pixels in the image and p_{ij} is the value of the j[th] pixel of the image at the i[th] color channel. The second color moment is the standard deviation, which is obtained by taking the square root of the variance of the color distribution.

$$\sigma_i = \sqrt{\left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^2\right)} \tag{5}$$

where E_i is the mean value, or first color moment, for the i[th] color channel of the image. The third color moment is the skewness. It measures how asymmetric the color distribution is, and thus it gives information about the shape of the color distribution. skewness can be computed with equation (6):

$$s_i = \sqrt[3]{\left(\frac{1}{N}\sum_{j=1}^{N}(p_{ij} - E_i)^3\right)} \tag{6}$$

Kurtosis is the fourth color moment, and, similar to skewness, it provides information about the shape of the color distribution. More specifically, kurtosis is a measure of how flat or tall the distribution is in comparison to normal distribution. This can be computed with equation (7):

$$K_i = \frac{\sum_{j=1}^{N}(p_{ij} - E_i)^4}{N} \tag{7}$$

C. Invariant Moments

Image moments are useful to describe objects after segmentation. An image moment is a certain particular weighted average (moment) of the image pixels' intensities. Simple properties of the image which are found *via* image moments include area (or total intensity), its centroid, and information about its orientation.

Hu's moments are invariant under translation, changes in scale, and also rotation. It describes the image despite of its location, size, and rotation.

D. Gray level co-occurrence matrix (GLCM)

Gray level co-occurrence matrix [13] has proven to be a powerful basis for use in image classification [4]. The common statistics applied to co-occurrence probabilities are discussed below.

Energy: This is called Uniformity or Angular second moment. It measures the uniformity that is pixel pair repetitions. It detects disorders in images. Energy reaches a maximum value equal to one. High energy values occur when the gray level distribution has a constant or periodic form. Energy has a normalized range. The GLCM of less homogeneous image will have large number of small entries.

Entropy: This measures the disorder or complexity of an image. The entropy is large when the image is not texturally uniform and many GLCM elements have very small values. Complex textures tend to have high entropy. Entropy is strongly, but inversely correlated to energy.

Contrast: This measures the spatial frequency of an image and is the difference moment of GLCM. It is the difference between the highest and the lowest values of a contiguous set of pixels. It measures the amount of local variations present in the image. A low contrast image presents GLCM concentration term around the principal diagonal and features low spatial frequencies.

Variance: This is a measure of heterogeneity and is strongly correlated to first order statistical variable such as standard deviation. Variance increases when the gray level values differ from their mean.

Homogeneity: This is also called as Inverse Difference Moment. It measures image homogeneity, as it assumes larger values for smaller gray tone differences in pair elements. It is more sensitive to the presence of near diagonal elements in the GLCM. It has maximum value when all elements in the image are same. GLCM contrast and homogeneity are strongly, but inversely, correlated in terms of equivalent distribution in the pixel pairs population. It means homogeneity decreases if contrast increases while energy is kept constant.

***Correlation*:** This feature is a measure of gray tone linear dependencies in the image. The rest of the textural features are secondary and derived from those listed above. They are Sum Average, Sum Entropy, Sum Variance, Difference Variance, Difference Entropy, Maximum Correlation Coefficient, and Information Measures of correlation.

In the proposed work 50 such features have been identified and extracted based on color properties, texture properties and statistical properties of the image. Image Source identification has been performed based on supervised classifier, Support Vector Machine (SVM). In the first phase, training has been performed and the data has been classified to a set of predefined classes. During the testing phase feature extraction is done on any given image and then features are used to identify the nearest class that the given image may belong to. Thus identifying best features for source camera identification has been the focus of this study and this article.

IV. EXPERIMENTAL RESULTS

Five cameras have been used in this experimentation. Table 1 lists the digital still cameras model, maximum size of image and type of the image. Each scene has been captured as an image at three different timings of the day at 9am, 12noon and 3pm on each camera. Thus 20 images of same scene have been captured at various timings by each of the camera. The camera specific parameters have not been altered.

Table 1: Digital cameras used in this experimentation.

S.No	Camera Model	Max. Image Size	Image format
C1	Canon Power Shot A495	3648x2048	JPEG
C2	SAMSUNG PL120	4320x2432	JPEG
C3	SONY-DSCW330	4320x3240	JPEG
C4	Canon-DSLR	5184x3456	JPEG
C5	NikonD90	4288x2848	JPEG

In the training phase, 300 images have been used for extracting nearly 60 features and trained using SVM classifier. During the testing phase, nearly 100 images (trained and untrained) have been used to identify the nearest matching class using SVM. Experimental results show 85.5% accuracy for single class SVM and 99.4%accuracy for a multi class SVM. The image dataset include broad range of images, from natural scenes to buildings, images with different background, light intensities etc. Using these 300 images, 60 features have been extracted by applying five techniques described in section 2.

Table 2. The confusion matrix and average classification accuracy of fifteen independent tests models of feature extraction.

Camera model	C1	C2	C3	C4	C5
C1	98%	0%	0%	1%	0%
C2	0%	99%	0%	0%	0%
C3	0%	0%	100%	1%	0%
C4	1%	0%	1%	99%	0%
C5	0%	0%	0%	0%	99%

Table 3. Simulation results-color moments of three different channels at 9.00am.

Color moments		C1	C2	C3	C4	C5
Mean	R	0.4046	0.4472	0.3712	1.8104	0.5346
	G	0.3635	0.4094	0.3466	1.7067	0.5669
	B	0.2883	0.3123	0.6557	1.6832	0.4392
Variance	R	0.22	0.2318	0.2127	1.2555	0.227
	G	0.2134	0.2333	0.2012	1.1538	0.2315
	B	0.1869	0.202	0.4795	1.1751	0.2358
STD	R	0.468	0.481	0.4595	3.3547	0.4742
	G	0.4601	0.4828	0.4574	3.1863	0.4798
	B	0.4286	0.4474	0.4226	3.1897	0.4853

Table 3 gives the results of color moments features on the same scene images captured using 5 different camera models. Each scene has been captured as an image at three different timings of the day at 9am, 12noon and 3pm on each camera. Here, The chart 1 explore the comparison of various camera models color moments accuracy of 9.00am .

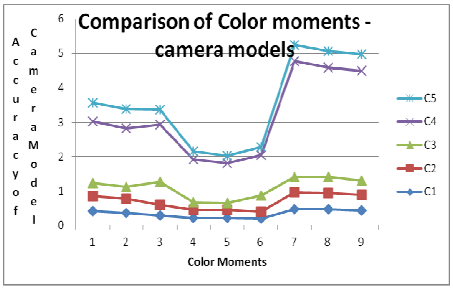


Chart 1. Comparison of color moments accuracy of various camera model.

Table 4: Simulation results-color moments of three different channels at 12.00 noon.

Color moments		C1	C2	C3	C4	C5
Mean	R	0.403832	0.443842	0.353737	0.452047	0.526137
	G	0.403958	0.444126	0.354332	0.452079	0.526221
	B	0.403895	0.444021	0.354168	0.452037	0.532028
Variance	R	0.219858	0.230963	0.218411	0.225016	0.227537
	G	0.219874	0.230926	0.218563	0.225005	0.227563
	B	0.219868	0.230968	0.218516	0.225016	0.227579
STD	R	0.467758	0.480079	0.466074	0.472763	0.475079
	G	0.467795	0.480016	0.466232	0.472768	0.475116
	B	0.467789	0.480058	0.466189	0.472784	0.475132

Table 5: Simulation results-color moments of three different channels at 3.00pm.

Color moments		C1	C2	C3	C4	C5
Mean	R	0.368758	0.493737	0.353737	0.476395	0.5613
	G	0.369221	0.494332	0.354332	0.476505	0.5618
	B	0.369221	0.494168	0.354168	0.4764	0.562
Variance	R	0.211616	0.358411	0.218411	0.233895	0.25
	G	0.211616	0.358563	0.218563	0.233889	0.25
	B	0.211726	0.358516	0.218516	0.233884	0.25
STD	R	0.458268	0.606074	0.466074	0.483132	0.5
	G	0.458247	0.606232	0.466232	0.483132	0.5
	B	0.458363	0.606189	0.466189	0.483121	0.5

Table 6: Simulation Result of Invariant Moments of various camera models at 9.00am.

	C1	C2	C3	C4	C5
M1	0.990689	1.4773	1.080768	0.817837	1.192221
M2	4.096268	5.3769	4.655453	3.278368	5.452
M3	5.677258	12.1646	5.838521	5.292605	6.7173
M4	7.031442	7.0468	7.169311	5.562432	8.017805
M5	14.54685	16.3152	14.59356	11.26311	16.47348
M6	10.09925	10.1417	10.02611	7.607258	11.64532
M7	13.83822	16.1369	14.61824	12.40639	16.07282

Table 7: Simulation Result of Invariant Moments of various camera models at 12.00 noon.

	C1	C2	C3	C4	C5
M1	0.990689	1.080768	0.817837	1.126516	1.192221
M2	4.096268	4.655453	3.278368	5.151442	5.452
M3	5.677258	5.838521	5.292605	5.997568	6.7173
M4	7.031442	7.169311	5.462432	7.245705	8.017805
M5	14.54685	14.59356	11.26311	14.52273	16.47348
M6	10.09925	10.02611	7.607258	10.72306	11.64532
M7	13.83822	14.61824	12.40639	15.01369	16.07282

Table 6-8 show the invariant moments results of different camera models. In which the moments gives better results than color moments of the digital images. The chart 2 displays the invariant moments performance of various camera models.Which shows better performance than color moments. Apart from these the combination of conditional probability features with these features produce much more better results.

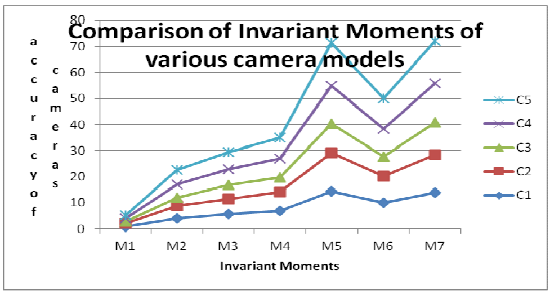


Chart 2.Comparison of Invariant Moments of various Camera Models.

Chart 3 shows the comparison performance of the conditional probability,Image features and the combination of the conditional probabilty with the invariant moments.In which the hybrid frame work of image feature with conditional probability give a very good result than orhers.

Table 8: Simulation Result of Invariant Moments of various camera models at 3.00pm.

	C1	C2	C3	C4	C5
M1	0.974305	1.4773	1.174174	0.817837	1.126516
M2	3.789932	5.3769	4.927868	3.278368	5.151442
M3	5.280074	12.1646	6.586895	5.192605	5.997568
M4	6.725768	7.0468	7.5419	5.462432	7.245705
M5	13.79208	16.3152	15.12732	11.46311	14.52273
M6	9.341337	10.1417	10.90227	7.627258	10.72306
M7	13.4623	16.1369	15.56998	12.41639	15.01369

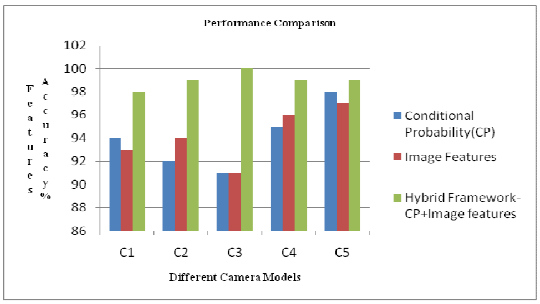


Chart 3. Performance of Hybrid framework of Conditional Probability with Image Features.

V. CONCLUSION

This work concludes that identifying best features of an image in order to perform image source identification (ISI). In which invariant moments produce better result than color moments of the image. The combined framework of these invariant moments with conditional probability features produces best result than the others. This information is essential in identifying forgeries in a given image. As an extension to this work, image forgery detection will be performed. Experimental results corroborate that these features are best to perform image analysis.

REFERENCES

[1]. J. Lukac, J. Fridrich, and M. Goljan, "Digital camera identification from sensor pattern noise," *IEEE Transactions on Information Security and Forensics*, vol. **1**, pp. 205–214, 2006.

[2]. Sintayehu Dehnie, Husrev T. Sencar and Nasir Memon, "Digital Image Forensics for Identifying Computer Generated and Digital Camera Images," *IEEE International Conference on Image Processing, Atlanta*, GA, pp. 2313-2316, 2006.

[3]. Kai San Choi, Edmund Y. Lam, and Kenneth K.Y. Wong, "Feature selection in source camera identification," *IEEE Conference on Systems, Man, and Cybernetics, Taipei*, Taiwan, pp. 3176-3180, 2006.

[4]. Sevinc Bayram, Husrev T. Sencar, Nasir Memon," Source Camera Identification Based on CFA Interpolation," *IEEE International Conference on Image Processing, Genova*, Italy, pp. 2413-2416, 2005.

[5]. M. Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. **3**, pp. 74–90, 2008.

[6]. Sevinc Bayrama, Husrev T. Sencarb, Nasir Memon, "Classification of digital camera model based on demosaicing artifacts", *Journal of Digital investigations*, vol. **5**, pp.49-59, 2008.

[7]. Todd Moon, "The Expectation Maximization Algorithm", *IEEE Transactions on Signal Processing,* November, 1996.

[8]. Z. J. Geradts, J. Bijhold, M. Kieft, K. Kurosawa, K. Kuroki, and N. Saitoh, "Methods for Identification of Images Acquired with Digital Cameras, *In. Proceeding of the SPIE, 4232*, pp. 505–512, 2001.

[9]. B. Wang, Y. Guo, X. Kong, and F. Meng, "Source Camera identification Forensics Based on Wavelet Features", *In. Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing, Los Alamitos, IEEE Computer Society*, CA, USA, pp.702-705, 2009.

[10]. Mehdi Kharrazi, Husrev T. Sencar, and Nasir Memon, "Blind source camera identification," *IEEE Int. Conference on Image Processing,* Singapore, pp. 709-712, 2004.

[11] Gouri K. Bhattacharyya and Richard A. Johnson, *Statistical Concepts and Methods*, Wiley, 1997.

[12] Ainuddin Wahid Abdul Wahab and Philip Bateman, "Conditional Probability based Camera Identification", *International Journal of Cryptology Research,* vol. **2**, no.1, pp.63-71, 2010.

[13] Anuradha. K and Dr. K. Sankaranarayanan, "Statistical Feature Extraction to classify oral cancers", *Journal of Global research in computer science*, vol. **4**, no. 2, pp.8-12, Feb'2013.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 135-142(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Study on Node Placement Strategies in Wireless Sensor Networks

R. Shanmugavalli[1] and Dr. P. Subashini[2]
[1]Research Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore (TN), India
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women Coimbatore (TN), India

ABSTRACT: In recent years, the advances in Wireless Sensor Networks (WSNs) based on wireless technology is used in variety of topologies, sensor devices and applications such as environmental monitoring, health care and military services. Wireless Sensor Networks is a collection of specialized devices and sensors, connected without any physical connections that are used to monitor more physical and environmental conditions. Sensors play a major role to observe the environmental condition results by converting the electrical signals and transferring it through base stations. The node placement of the network is organized by cooperative or allied multiple nodes to collect more environmental or physical information and pass their data through the network to base station or sink. In it, the data is transferred directly through the base station via radio signals where here the base station or sink acts like a intermediate between the WSNs and users. In this paper, we focused and reporting an the ongoing state of the research on optimized node placement in wireless sensor networks and the open problems in this node placement in WSNs with a small analysis of protocols in WSN which helps to improve the deployment of network strategy.

Keywords: Wireless Sensor Networks, Positioning, Requisitions, Node Placement, Performance Evaluation, ad hoc networks.

I. INTRODUCTION

Sensor is a device that detects physical or environmental conditions by collecting, storing, and communicating with other nodes. The nodes understandably inputs can be light, heat, motion, moisture, pressure or any one of a more environment phenomena. Normally, WSN consists of more number of distributed nodes that arrange themselves into a multi-hop wireless network and typically these nodes coordinate to perform a common task [1]. The wireless sensor networks is a large-group ad hoc, multi hop, unpartitioned network of largely homogeneous and tiny resource-constrained [2]. The sensor output is generally a different signal that is converted to human understandable display at the node location. Sensors observe the analog signals and convert the user readable languages in digits or letters (digital signal) at the sensor location. This type of sensor nodes is formatted by different types of network designs. Basically wireless sensor node placement benefits better covering, avoiding congestions, power consumption, effective coverage data collection, feasible network communication, cost effective, and network lifetime. In [4], the authors defines that the power balance problem can be considered from different aspects, including node placement and topology control and concluded that L-based topology is suitable. It also discusses about the layout optimization problem in WSNs: the layout is a different node placement focused on battery lifetime, node lifetime, energy consumption, coverage area, and etc. The authors M. Younis and K. Akkaya has discussed about careful node placement based on optimization which helps for achieving the desired design goals effectively [3]. The design is controlled by different types of topologies or formulas that are used by few researchers.

A. WSNs

Wireless Sensor Network is a collection of cooperative nodes connected without any physical connection, and communicated with end of the node via radio signals. In [4], authors have discussed about WSN that comprises of base station node and group of sensor nodes that are communicated with each other to execute a broader sensing task. Basically, the WSNs structure focuses on minimizing the number of nodes [5], in large coverage area using the equation (1)

$$\text{Coverage}(x) = 100 + \frac{\text{Covered points}}{\text{Total points}} \quad (1)$$

and $$f(x) = \frac{Cove}{Nb.o}$$

A sensing field is totally covered, if every point in the field is within the sensing range of an active sensor in a proper degree. Every sensor node comprises of memory, controller, power supply, communication device and sensing device. WSN determines the small number of energetic sensors to maintain K-coverage of a terrain as well as K-connectivity of the network that are shown in Fig.1. Global Positioning System (GPS) and local positioning algorithms can be we suited to achieve location and positioning.

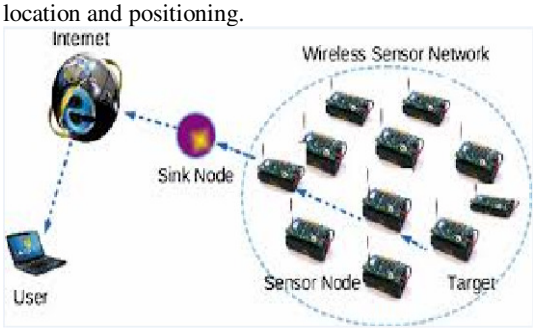


Fig. 1. Wireless Sensor Network Architecture.

The paper is organized as follows section-I gives the introduction, Section-II consists of details about WSN types, Section-III illustrates various node placements in WSN, Section-IV explains the node placement strategies and Section-V gives the performance metrics of various node placement strategies of WSN and section-V concludes the study on node placement strategy

II. WSN TYPES

There are different Wireless Sensor Networks (WSNs) available and few are explained below:

A. Terrestrial WSNs

The terrestrial wireless sensor network consists of hundreds or thousands of wireless sensor nodes and thus nodes can be deployed in an unstructured (ad hoc) or a structured (Preplanned) manner. The structured or preplanned mode considers grid placement, optimal placement, 2D placement and 3D placement models. In terrestrial WSNs, the battery power is limited and the battery is equipped with solar cells as a secondary power suppliers.

B. Underground WSNs

This network is more effective in terms of deployment, maintenance, equipment cost thoughtful and careful planning which includes the number of nodes that are hidden in the ground to monitor the underground conditions. These are effectively used to monitor the undergrounded conditions therefore their whole network is underground but to pass on the information to the base station, sink nodes are used which are present above the ground level.

C. Underwater WSNs

Underwater wireless sensor network system consists of sensor nodes or vehicles which are deployed under the water. To gather data from the sensor nodes, autonomous underwater vehicles are used. The battery of these WSNs is also limited and cannot be recharged; therefore, different techniques are being developed to solve this issue of energy usage and conservation.

D. Multimedia WSNs

Multimedia Wireless Sensor Networks have been proposed that enables tracking and monitoring of events from audio, video, and imaging for the purpose of data compression, retrieval, and correlation. These nodes are also interconnected with one another through the wireless connection. Advanced techniques like data processing and compression are used in the multimedia wireless sensor networks.

E. Mobile WSNs

These networks consist of a collection of more wireless sensor nodes that can move on their own and interact to the physical environment. It can be easily interfaced with the environment to observe the physical condition and provides better coverage, superior channel capacity and enhanced coverage and so on. The mobile sensor networks are most perfect as compared to other static sensor network systems.

III. NODE PLACEMENTS IN WSN

Kirsh Y *et al.* [12] defined about node placement as a technique to place the nodes effectively in the simulation area so as to consume the minimum energy from each node that is intended for transmission of packets or a data. In [4], authors developed the node placement algorithm which considers the number of forwarding message of each sensor and tried to balance the energy consumption of all sensor nodes. A wireless sensor network node placement includes different topologies for radio communications networks such as point-to-point, star, tree, bus, ring, mesh, and hybrid. The proper node placement [6] is necessary to ensure good sensing coverage and communication connectivity. The most favorable network topology goal is to maximize the network lifetime and increase the network throughput using optimal power-controlled topology control algorithm [7]. The topology of sensor nodes on a monitored field is a factor that is possible to reflect the overall performance of the network. In [2], author has given a study that the topology affects many network characteristics such as latency, robustness, and capacity. Energy efficient node placement algorithm is proposed in [24] by comparing energy consumed by the nodes of random placement is best with that optimized placement

of nodes. The multiple positioning of WSN has also been focused. Fig. 2 gives the different components of a sensor node.

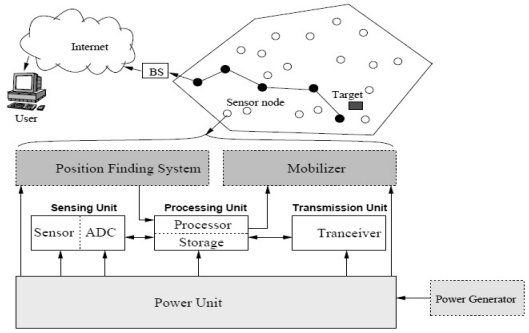


Fig. 2. The components of a sensor node.

A. Types of topologies

A sensor networks forms a single-hop network in which each sensor node being able to directly communicate with every sensor node. Sensor node placement in WSNs can be in few different topologies such as follows:

Point-to-Point Topology. The Point to Point network is also called the peer to peer network topology in which every node is directly communicated with another node without going through a centralized communications (base station). Each peer to peer topology node is able to function both as a sender and receiver to the other nodes on the network (Fig 3). In this topology, network formulation is very easy and simple to maintain communication.

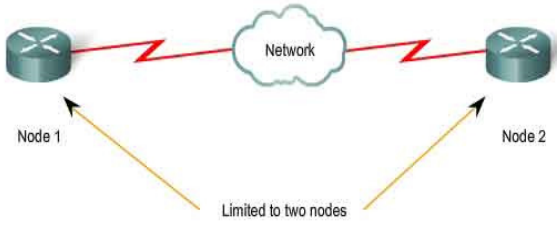


Fig. 3. A Point-to-Point topology.

Star Network. A Star network is a dissemination topology where a single base station or sink node can send and/or receive data packets to a number of other remote nodes. The base station requires message handling, routing, and decision-making then the other neighbor nodes. This type of networks considers simplicity, ability to keep the remote node's power consumption to a minimum. Jolly *et al.* [14] determined that the sensor networks are in direct communication range i.e. 30-100 meters to the central node. The base station controls all individual nodes, and a single base station node to manage the overall network. In [18], authors discussed that arranging more node placement in certain pattern decides efficiency of sensor network, so a proper arrangement is therefore required.

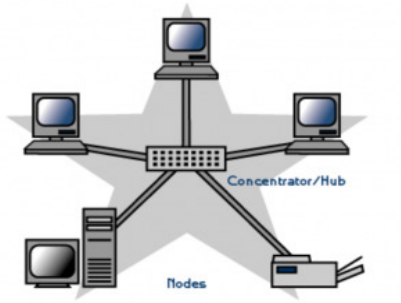


Fig. 4. A Star network topology.

Tree Network. The Tree topology is a cascaded star topology in which, each node connects to other node that is placed higher in the tree, and then to the gateway (Fig. 5). The growth of this network topology can be simply possible, and also error observation becomes easy. This networks use a base station called a root node as main communications routers. The last level of this network topology forms a star network. This sensor network path may be single hop or multi hop sensor nodes for sensing the environment and sent them to the sink and sensor forwards them to its parent after receiving data messages from child nodes [9].

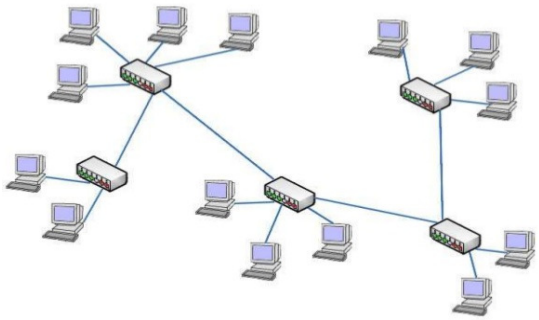


Fig. 5. A Tree network topology.

Bus Network. The Bus topology is a set of sensor nodes directly connected via a single network. In all, a sensors node receives the message but only recipient actually processes the information and rest nodes discards the message. This topology is very simple to install but congestion of traffic with single path communication and this network is best with a limited number of nodes [9]. It also creates problem of traffic congestion but it is very easy to install and uses limited number of nodes [18].

Ring Network. The Ring topology sensor node executes the same function without any base stations, generally to transfer the sensing information in single direction based on ring circle (Fig.7).

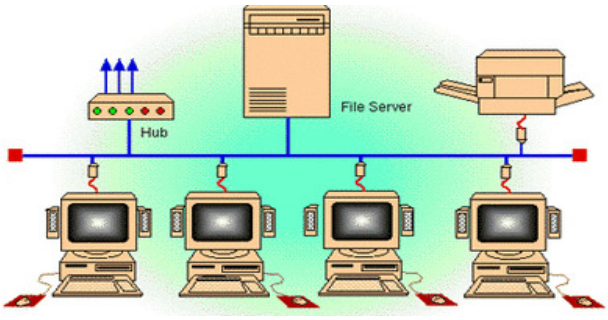


Fig. 6. A Bus network topology.

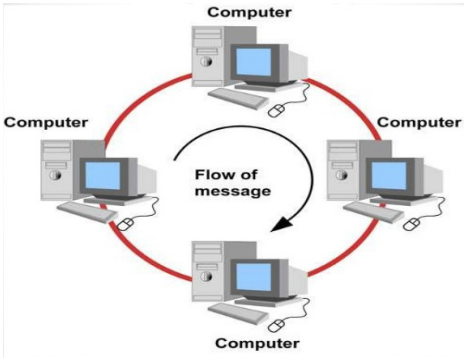


Fig. 7. A Ring network topology.

Each sensor node verifies the destination address in the message header, and processes the message addressed to it. The sensor node simply observes the messages and is not responsible for retransmitting any messages in the bus topology. Ring topology [8] is not preferred in more applications this sensor network topology does not have any leader node.

Mesh Network. A mesh network allows communicating data to one node to other node in the network that is within its radio transmission range. This topology involves messages that can be received from sender to receiver in several paths. Mesh topology is basically multi-hoping system [8] in which all sensor nodes can communicate with central node as well as with each other. Every node directly connected to other neighbors is called a full mesh. In Mesh topology nodes are to connected to each and it single network path is down or fault then the other network path is available for communication. In paper [28], authors discussed that the Mesh topology provides security, privacy, robust, and also fault diagnosis is easy.

Mesh nets can be good models for large-scale networks of wireless sensors that are distributed over a geographic region [10].

Hybrid Network. The Hybrid network combines Star and Mesh networks and provides the robust and versatile communications network with ability to keep minimum power consumption in wireless sensor nodes.

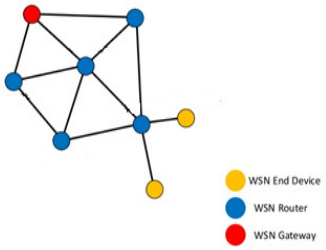


Fig. 8. A Mesh network topology.

In Hybrid network topology, sensor nodes with lowest power are not allowed with the ability to forward messages. In this topology order of the nodes will be in Star around Mesh nodes (Fig. 9), which finally turns into a mesh network.

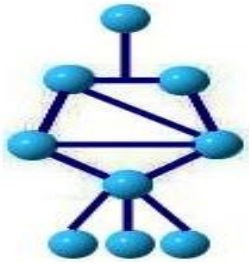


Fig. 9. A Hybrid network topology.

The hybridization of star-mesh network offers the highest degree of sensor node mobility and flexibility for fast changes in the network population and the overall low power consumption [8]. This network allows minimal power consumption to be maintained and includes multiple base stations and also communicates with mesh topology. Each base station directly communicated with all interconnected nodes and base stations. This topology provides highest degree mobility with the flexibility to rapid changes.

IV. NODE PLACEMENT STRATEGIES

A method or a plan for the wireless sensor node placement is called the placement strategies.

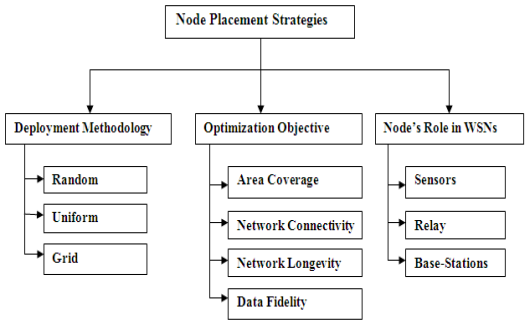


Fig. 10. Node placement strategies.

The goal of the node placement achievement is the efficient node placement with best and feasible solution to the target problems. A placement strategy comprises a deployment methodology, optimization objective and a node's roll in WSNs. Fig. 10 gives the structure of node placement strategies.

A. Deployment Methodology

The deployment methodology in wireless sensors networks is used for proper node arrangement with most effective performance results. It is used in multiple designs of node arrangements in WSNs. The best fit node placement based congestion control mechanism (Congestion Control Algorithm-CCA) [11] has the ability of increasing the bandwidth utilization over WSNs while maintaining fairness. This paper has discussed three different designs of node placements in wireless sensor networks like random placement, uniform placement and grid placement.

Random Placement. The Sensor nodes are placed randomly within the physical terrain area. In this node placement, the node are spread in fixed terrain and does not control the node density. In random distribution sensor network [12], as the name suggests the nodes are placed at un-equal distance at different positions. In [29], the author compares the two different random node placement strategies like simple random node placement and compound random node placement. In [25], authors has discussed about multiple research presenting solutions over random topologies related to WSN.

Uniform Placement. Uniform node placement is based on the number of sensor nodes in terrain divided into a number of cells, where nodes are placed randomly. The uniform topology same as random placement but major difference is keeping some nodes in uniform order. The uniform node placement [13] is based on the number of nodes in the physical terrain and divided into a number of cells, a node placed in randomly. In [12], an author discusses uniform node placement topology in a network, which mainly reduces the overall energy consumed by the network further increasing the network lifetime. This generates topology that is random, but with somewhat uniform density of nodes.

Grid Placement. The grid node placement nodes starts at (0,0) and this nodes are placed in a grid structure with each node a grid-unit away from other neighbor.

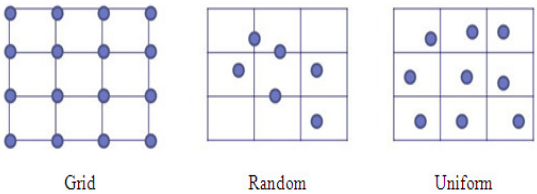


Fig. 11. Three different node placements.

Grid node placement must also be specified numerically, with the unit in meters depending on the value of coordinate system. In [12], the author discussed three node placement models and concludes that energy can be efficiently utilized in WSNs, with the grid node deployment.

B. Optimization Objectives

There are four optimization objectives available in Wireless Sensor Networks like as area coverage, network connectivity, network longevity and data fidelity.

Area coverage. The area coverage is a big challenge in wireless sensor networks and proper node placement is a best solution for area coverage. The area coverage [5] is to maximize the sensing area of the network while minimizing the number of sensors deployed with using Ant Colony Optimization (ACO) algorithm. Good area coverage is to observe the lowest place in the target area. Optimizing Coverage problem focus on placing sensors so as to get the best possible coverage while saving as many sensors as possible like SA (Simulated Annealing) algorithm, and CHC (Cross generational elitist selection, Heterogeneous recombination, and Cataclysmic mutation) [14] algorithms. In this coverage problem can be divided two types like Single coverage and multiple coverage in the wireless sensor networks. Single coverage, every target or point in the area must watch by at least single working sensor, and the multiple coverage, every target or point in the area needs to be watched by different working sensors.

Network Connectivity. A wireless sensor network does not uses any physical connection and only uses wireless connections, example radio signals. Network connectivity forms the links with other neighbor nodes so that it can communicate with each other. Various reasons of connectivity faults are node failures, security breach/denial of service, energy depletions, sparse amount of nodes, mobility of nodes and environmental changes. Network connection in between other neighbor nodes and it is broken when they communicate with each other and therefore wireless sensor networks are highly dependent on network connectivity. Laki *et al*. the wireless sensor networks may include millions of sensor nodes, each node position is not necessarily pre engineered and its manual placement in the field is unpractical [15]. WSNs nodes consume their limited battery power and without change, it can no longer be connected to the system which will disconnect some significant nodes from the access point.

Network Longevity. WSNs lifetime important goals are feasible connectivity, best area coverage and very long network life time. Extending network longevity has been the optimization objective for most of the research communication protocols for wireless sensor networks. The node positions significantly impact the network longevity.

Kenan Xu *et al.* discussed that the specific sensing task decides the number and deployment of heterogeneous devices; so that the total network cost is minimized while the constraints of lifetime, coverage and connectivity are satisfied [16].The average energy consumption per data collection round is used as a metric for measuring the sensor node lifetime. Network lifetime for a given network can be computed mathematically and use this knowledge to compute locations of deployment such that the network lifetime is maximum [17].

Data Fidelity. Wireless sensor observes information that are successfully transferred to target neighbor nodes, This is a good network communication as it follows the loyalty. Ensuring the reliability of the gathered data is an important goal of wireless sensor networks. Errol, L *et al.* discussed that the sensor network basically provides a collective assessment of the detected phenomena by merging the readings of multiple independent (and some-times heterogeneous) sensors. The data combined boosts the fidelity of the reported incidents by lowering the probability of false alarms and missing a detectable object. In [15], authors define that from a signal processing point of view, the data fusion will try to minimize the effect of the distortion by considering reports from multiple sensors so that an accurate assessment can be made reading the phenomena. Number of sensors to increasing and reporting in a particular region will surely boost the exactness of the fused data. However, redundancy in coverage would require an increased sensor node density, which can be unpleasant due to cost and other constrains the potential of detecting the sensors in a combat field.

C. Node's Role in WSNs

In Wireless sensor networks three types of nodes are used to construct a network: Sensor Node, Relay Node and Base Station Node.

Sensor. The sensor nodes are used to observe the physical or environmental conditions and to sense information directly transferred to the base station. The sensors sense the environment and report the results to the end-user [16]. Sensor includes low capacity battery and the main role of the sensors is to sense and transfer the environmental conditions. In [22], authors discussed on radio transceivers with an internal antenna or connection of an external antenna, a micro controller, an electronic circuit for interfacing with sensors and an energy source of a single sensor node. The node specific understandable input can be light, heat, motion, moisture, pressure or any one of a more environment phenomena.

The sensor output is generally a different signal that is converted to human understandable display at the node location. Sensors observes the analog signals and then convert the user readable languages like digits or letters (digital signal) at the sensor location and uses the Analogue to Digital Converter (ADC).

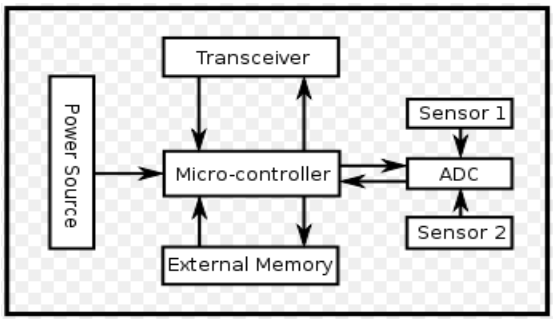


Fig. 12. Sensor Node Architecture.

Relay. The Relay node is used to transmit the information for long distance in wireless sensor networks. Long distance data transmissions maintains the relay sensor nodes and with energy efficient node. Relay nodes have a sufficient energy supply, using wall power, solar power and high capacity battery. In [20], authors defined on approach to prolong network lifetime while preserving network connectivity to deploy a small number of costly, but more powerful, relay nodes whose main task is communication with other sensor or relay nodes.

Base-Station. A Base station is also called leader node in the wireless sensor network. Sensor network normally made of a base station and with group of wireless sensor nodes. Base station to collects all the information from attached neighbor nodes. The data is forwarded, possibly via multiple hops, to a sink that can use it locally or is connected to other networks through a gateway (Sink node) [21]. This data is directly transferred through the base station via radio signals. The base station acts like a communication in between WSNs to users and internet. A single base station node controls more inter connected sensor nodes in the wireless network. A.P. Aad *et al.* compares six different SBP (Base Station Placement) schemes above and evaluated their relative performance in terms of network lifetime and amount of data delivered during the network lifetime [26].

V. PERFORMANCE ANALYSIS OF ROUTING PROTOCOLS

The Wireless Sensor Networks node placement provides different performance results in each simulation. In general, the performances of routing protocols are evaluated with packet delivery ratio, end-to-end delay, throughput and jitter. The analysis of Random Placement Strategies with different protocols is discussed in the following section. Based on the literature survey, the analysis with different number of nodes and various protocols has been compared as shown in the Fig. 13, 14, 15 and 16.

A. Packet Delivery Ratio

The Packet Delivery Ratio (PDR) is an important metric in wireless sensor networks and defines the ratio of the number of packets created by the source node. In [22] and [23], authors discussed the packet delivery as number of packets received by the destination to the number of packets generated by the source node. Packet delivery ratio is calculated by dividing the number of packets received by the base station through the number of packets generated by the overall sensor nodes. From the observation of literature in [22], [23] for the wireless sensor networks it uses DSR protocol.

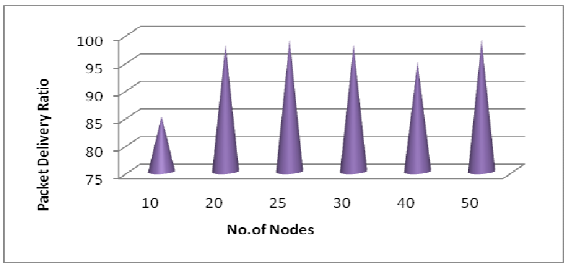


Fig. 13. Packet Delivery Ratio.

B. End-to-End Delay

The End-to-End delay is defined as the delay time during successful information transmission between sensor nodes to base station node. Total amount of time to transferred the single data in between source to destination.

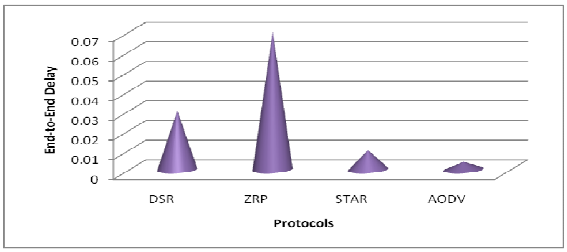


Fig. 14. Average End-toEnd Delay.

The end-to-end delay is one of the most important metrics in assessing the network performance [26]. D.Shagila et al. observed that the expected delays are more for ZRP protocols in comparison to OLSR and the end-to-end delay of OLSR is less as it has reduces routing overhead and queuing delay [17]. From the observation of literature on different end-to-end delay four different protocols like DSR, ZRP, STAR and AODV with Random Node Placement shows better results with AODV protocol with respect to node density 10 [23] and [30], as shown Fig. 14.

C. Throughput

The Throughput refers to how much information can be transferred from one location to another in a given amount of time. Throughput is a number of bits per second and must be high for a better system performance [23]. Therefore, it is good to remember that the maximum throughput of a device or network may be significantly higher than the actual throughput achieved in everyday use. Recent research works specifies different analysis results in wireless sensor networks. The Throughput of four different protocols and 20 nodes are analyzed with Random Node Placement Strategy [23],[30]. As per the results are shown Fig.15 the average throughputs of STAR Protocol high with the Random Node Placement Strategies.

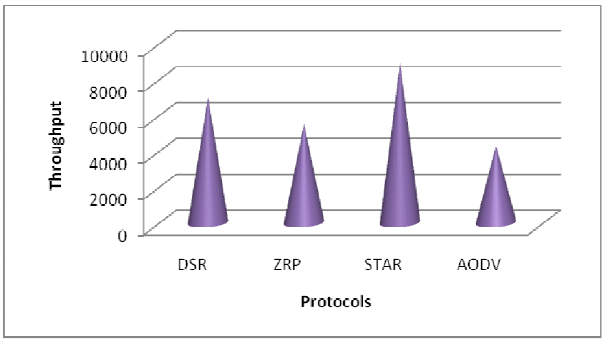


Fig. 15. Throughput.

A. Jitter

Jitter is different or variation delay time of received packets. In [23], authors discussed that average jitter is the variation of time between packets arriving, causes by network congestion, timing draft, or route changes. N. Arora mainly focused on proactive, reactive and hybrids routing protocols like AODV, DSR and ZRP and analyze the different Performance results. The sensing side transmits packets in a continuous stream and spaces them evenly apart.

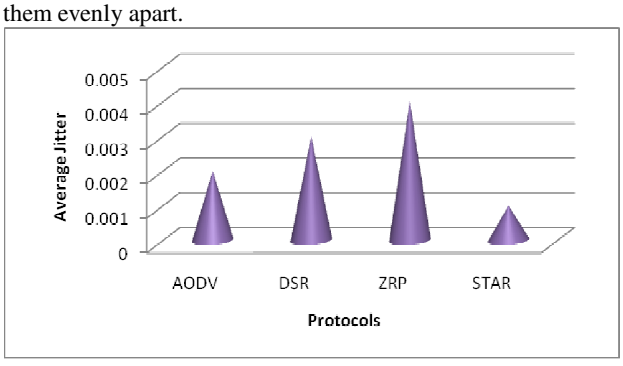


Fig. 16. Average Jitter.

Because of network congestion, improper queuing, or configuration errors, the delay between packets can vary instead of remaining constant. In [23] and [30], various routing protocols with Random Node Placement have been evaluated. As per fig.16, Star protocol provides better results in terms of average jitter when compared to ZRP and other protocols.

VI. CONCLUSION

The paper discusses the various routing protocols and placement strategies for wireless sensor networks. The performance of the random placement using different protocols was analyzed for various metrics.

REFERENCES

[1] Mo Li, Baijian Yang, "A survey on topology issues in wireless sensor networks," International onferenca on Wireless Networks, 2006.

[2] Kay Romer, Friedemann Mattern, "The design space of wireless networks," IEEE wireless communications, vol. **11**, no. 6, pp. 54-61, December 2004.

[3] Mohamed Younis, Kemal Akkaya, "Stratergies and techniques for node placement in wireless sensor networks: a survey," Ad Hoc Networks, vol. **6**, pp. 621-655, 2008.

[4] Chih-Yung Chang, Hus-Ruey Chang, " Energy-aware node placement, topology control and MAC scheduling for wireless sensor networks," Computer Networks, vol. **52**, no. 11, pp. 2189-2204, 2008.

[5] Stefka Fidanova, Pencho Marinov, "Optimal wireless sensor network coverage with ant colony optimization," International Conference on Swarm Intelligence, 2011.

[6] Charalambos Sergiou, Vasos Vassiliou, "Efficient node placement for congestion control in wireless sensor networks," INFTECH Open Science/Open Minds, Chapter 1, DIO. 10.5772/48201.

[7] Monika Bathla, Nitin Sharma, "Topology control in wireless sensor networks," *International Journal of Advances in Computer Networks and its Security*, vol. **1**, pp. 157-160, 2011.

[8] Jolly Soparia, Nirav Bhatt, "A survey on comparative study of wireless sensor network topologies," *International Journal of Computer Applications,* vol. **87**, no. 1, pp. 0975-8887, February 2014.

[9] Dhivya Sharma, Sandeep Varma, and Kanika Sharma, "Network topologies in wireless sensor networks: a review," *International Journal of Electronics & Communication Technology*, vol. **4**, no. 3, pp. 93-97, June 2013.

[10] F.L.Lewis, "Wireless sensor networks," Smart Environments: Technologies, Protocols, and Applications, 2004.

[11] Pavithra Krishna Moorthy, Karthikeyan Easwara Moorthy, "Best fit node placement based congestion control mechanism in WSNs," *7th International Conference on Engineering Technology,* July 2017.

[12] Kirshnakumar Y. Bendigeri, Jayashree D. Mallapur, "Multiple node placement strategies for efficient routing in wireless sensor networks," *Wireless Sensor Networks*, vol. **7**, pp. 101-112, 2015.

[13] Dr.Vikram Singh, Jyoti Yadav, "Impact of random, uniform node placement and grid environment on the performance of routing protocols in MANET," *International Journal on Recent Innovation Trends in Computing and Communication*, vol. **4**, no. 7, pp. 349-354, July 2016.

[14] Enrique Alba, Guillermo Molina, "Optimal wireless sensor network layout with metaheuristics: solving a large scale instance," *International Conference on Large Scale Scientific Computing,* vol. **4818**, pp. 527-535.

[15] Laki, Shri. R.N.Shukla, "Stratergies and techniques of node placement in wireless sensor networks: a survey,"

[16] Kenan Xu, Quanhong Wang, Hossam Hassam Hassanein, and Glen Takahara, "Optimal wireless sensor networks (WSNs) deployment: minimum cost with lifetime constraint," *IEEE International Conference on Wireless and Mobile Computing, Networking and Communications*, 2005.

[17] D. Shagila, N.Aparna, "Sensor deployment and scheduling for target coverage problem in wireless sensor networks," *International Journal of Mearging Technology and Advanced research in Computing*, vol. **IV**, no. 14, June 2016.

[18] Sameer Dewangan, Ashish Kumar Pandey, Nilmani Verma, and Deepak Xaxa, "Acomparative assessment of technologies and their issues in wireless sensor networks," *International Journal of Engineering Sciences & Research Technology*, pp. 388-393.

[19] Neha Singh, Kamakshi Rautela, "Literature survey on wireless sensor networks," *International Journal of Engineering and Computer Science*, vol. **5**, no. 8, pp. 17544-17548, Augest 2016.

[20] Errol L. Lloyd, Guoliang Xue, Senior Member, "Relay node placement in wireless sensor networks," *IEEE Transactions on Computers,* vol. **56**, no. 1, pp. 134-138, Jan 2007.

[21] Chiara buratti, Andrea Conti, Davide Dardari, and Roberto Verdone, "An overview on wireless sensor networks technology and evolution," *Open Access Sensors,* vol. **9**, pp. 6869-6896, August 2009.

[22] K.Prabha, R.Anbumani, "Performance evaluation of packet delivery ratio for mobile ad hoc networks," *International Journal of Computer Application Technology and Research,* vol. **6**, no. 7, pp. 306-310, 2017.

[23] Dharam Vir, Dr.S.K.Agarwal, and Dr.S.A.Imam, A simulation study on node energy constraints of routing protocols of mobile ad hoc networks use of qualnet simulator," *International Journal of Advanced Research in Eletrical, Electronics and Instrumentation Engineering*," vol. **1**, no. 5, November 2012.

[24] Kirankumar Y, Bendigeri and Jayashri D. Mallapur, "Energy aware node placement algorithm for wireless sensor network," *Advance in Electronic and Electric Electronic Engineering*, vol. **4**, no. 6, pp. 541-548.

[25] Tony Ducrocq, Michael, Nathalie Mitton, Sara Pizzi, "On the impact of network topology on wireless sensor networks performances," *Advanced Information Networking and Applications Workshops* (WAINA), 2014.

[26] A.P.Aad, A.Chockalingam, "Mobile base stations placement and energy aware routing in wireless sensor networks," IEEE: Wireless Communications and Networking Conference, 2006.

[27] Seema Pahal, Kusum Dalal, "Performance evaluation of routing protocols in WSN using qualnet 5.3," *International Journal of Recent Trends in Engineering & Research*, vol. **02**, no. 06, pp. 223-231, June 2016.

[28] Nivedita Bisht, Sapna Singh, "Analytical study of different network topologies," *International Research Journal of Engineering and Technologies,* vol. **02**, no. 01, pp. 88-90, March 2015.

[29] Mustapha Reda Senouci, Abdelhamid Mellouk, and Amar Aissani, "Random deployment of wireless sensor networks: A survey and approach," *International Journal of Ad Hoc and Ubiquitous Computing,* vol. **x**, no. x, pp, 1-14, may 2015.

[30] N.Arora, "Performance analysis of AODV, DSR and ZRP in MANETs using Qualnet simulator," *Journal of Engineering Science and Technology Review,* vol. **6**, no. 1, pp. 21-24, January 2013.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 143-146(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

A Study on the Applicability of IOT Based Technology in Mosquito Control

Dr. N. Valliammal[1] and J. Prabha[2]
[1]Assistant Professor (SS), Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (Tamil Nadu), India,
[2]PG student, Department of Computer Science
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore (Tamil Nadu), India.

ABSTRACT: In today's world everyday new technologies are arising. Among them one of the popular technologies is drones which is also known as Unmanned Aerial Vehicle (UAV). Within a short span of time drone technology has developed to a great extent in both developed as well as developing countries. UAV has skyrocketed over the past decade driving costs down and the number of potential applications up, one being mosquito control. UAV are used in numerous numbers of fields including agriculture, military, pest control and industrial plants. Since UAV critically depends on sensors, antennas and embedded software to provide two-way communications for remote control and monitoring, they play a major role in the Internet of Things. This paper deals with the study of mosquito control using IOT UAV technology. The UAV are used such that it detects the mosquitoes in a particular area and destroys them.

Keywords: UAV, Rotor system, larvicide, survillence

I. INTRODUCTION

At first Internet of Things was used by industry researcher's people. It came into trend during recent times. One group of expert claim that soon the Internet of Things will change the whole networks of computers used today at the same time some experts believe that it will not make a huge difference in the world of computers. Internet of Things represents a general concept for the ability of network devices to sense and collect data from various parts of the world and then shares it worldwide for people who have the need to utilize it for commercial or own. Some also use the term industrial Internet interchangeably with IOT. This refers primarily to commercial applications of IOT technology in the manufacturing field. The Internet of Things is not limited to a particular field or application. The UAV are of different types. They are classified depending on various parameters such as battery range, type of drone etc. The above specified parameters are most significant as they define the drone capacity, range of the battery etc. UAV are monitored and operated through the sensors attached to it and controlled by the user on the ground. If there is need for UAV to fly then wireless communication is needed instead of wired network on ground. In addition, in most cases there is a need for communication with a payload, like a camera or a sensor. The frequency spectrum is needed for the communication between the payload and the devices. All the three factors including the drone type, characteristics of flight and the payload decide the frequency spectrum [1]. Since frequency spectrum is related with both national and international borders.

Worldwide there are common rules and regulations are there for using the frequency spectrum and electronic equipment (national and international legal matters on frequency spectrum and equipment requirements) are discussed, as well as frequency spectrum and vulnerability (an insight in available frequency spectrum and associated risks in using the frequency spectrum) and surveillance and compliance (enforcement of frequency spectrum use, equipment requirements and the need for international and European cooperation). In near future very small, light weighted and cheaper UAV will come to market. The trend is for UAV to become smaller, lighter, more efficient, and cheaper [1]. Because of these advantages many sectors started using UAV in these days. Also the purpose of using them also increased. Soon UAV will replace the other technologies. This paper is described as follows. Section I depicts the types of drones. Section II focuses on the system design. Section III is about the system flow and future scope as section V then the paper is concluded.

A. Main Existing Drone Type

UAV have many technical characteristics among them the drone type is very important. Multi rotor UAV and fixed wing drone are the major common types of UAV. Most of the UAV existing today come under these two categories were described in the Table 1.

B. Fixed-Wing Systems

Aviation industry uses the term fixed-wing (Fig. 1) to define aircraft that use fixed, static wings in combination with forward airspeed to generate lift (Table 1).

Some types of UAV cannot be labelled as a fixed-wing or a multi rotor drone. Sometimes because the drone simple it is neither fixed-wing nor multi rotor, sometimes because the drone has characteristics of both types. Hybrid systems (Fig. 3) are systems that have characteristics of both multi rotor and fixed-wing systems. The hybrid quad copter is an example of such a drone. This drone uses multiple rotors to take-off and land vertically but has wings so it can fly longer distances. The neither fixed-wing nor multi rotor systems are not far less frequent. Ornithopter falls under category of this type [4]. By seeing the wings of birds and insects UAV have been developed. Most of these ornithopters are scaled to the birds or insects they represent. UAVs that are used mostly are rotary drones and fixed wing drone other small drones are in the developing stage. These small UAV are mostly still under development and are not widely used in practice. UAV using jet engines are not using both the fixed wing drone and multi rotor drone. An example for this type of UAV is phantom drone. This drone uses a turbo fan, making the drone look more like an unmanned (hydro) fixed-wing or multi rotor. The same goes for rockets and jetpacks.

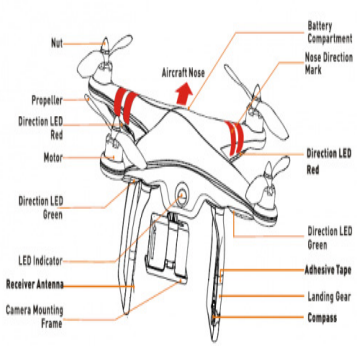


Fig. 3. Drone controller.

II. SYSTEM DESIGN

This paper is a study about UAVs used in mosquito control. Mosquito control programs are tasked with surveillance and targeted control of nuisance mosquitoes and potential vectors of pathogens that cause disease. Ideally, mosquito control personnel find and target mosquito egg-laying sites to kill aquatic immature mosquitoes before they emerge as flying blood-feeding adult females. Different mosquito species lay eggs in different locations where standing water is present. Egg-laying sites can be ephemeral or permanent. In other cases, egg-laying sites can be provided by human activity. Some mosquito control programs are using unmanned aerial vehicles to conduct surveillance and control in remote areas that are difficult to reach by land but may have mosquito production sites that impact residential or other public areas. Fixed wing or rotary wing can allow the operator to view images of potential larval production habitats, including monitoring drainage patterns, soil types and topography, in real-time [5].

The use of mosquito UAV may also be less disruptive and potentially less expensive than the use of helicopters for these remote areas. In addition to mosquito control operators, many agricultural operations are using UAV for surveillance and application of insecticides and/or for other purposes.

Fig. 4. NRI Camera.

Fig. 5. Larvicide sprayer.

Once larval habitats are identified, some UAV are capable of carrying and applying larvicide and/or adulticides to small targeted areas. Some UAV are fitted with a global positioning system (GPS) that can track flight patterns in conjunction with insecticide application. An operator can remotely pilot the drone or, in some cases, autopilot programs may be available for pre-programmed flights. UAV can be useful to target specific areas with larvicides or adulticides, as an alternative to truck-mounted applications that may require a high degree of drift of droplets in order to reach a target area in remote locations. UAV also may preclude the use of piloted aircraft for applications, thereby increasing the potential applications of smaller mosquito control programs that may not have budgets supportive of piloted aircraft [7]. States that quodapters, troopers, hex copters etc can be used with GPS. The GPS UAV is connected with satellites.

Fig. 6. Drone with NRI Camera.

It covers a large geographical used. This paper deals with the study of mosquito control using UAV and NRI camera (figure Fig. 6). Nearly 30 hectares of land can be survey by using a single drone. The camera locates the sites and they are mapped to the system. Then the larvicide dosage is calculated. For 2 hectares 400 larvicide tablets are used.

III. SYSTEM FLOW

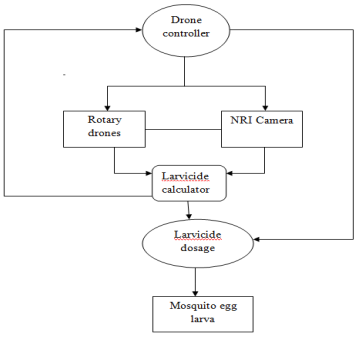


Fig. 7. Drone controller.

The flow diagram depicts the functioning of the drone controller. The controller is connected to the NRI camera. After capturing of images in the area the sprayer has to be activated. As a step process the larvicide sprayer which is attached with the camera starts its operation. After detecting the area the larvicide calculator calculates the dosage. Then the drone controller signals the sprayer to spray the larvicide on the detected area to destroy the mosquitoes. The flow diagram is depicted above (Fig. 7).

IV. FUTURE SCOPE

In today's scenario UAV have become very popular. This paper is a study about the UAVs in control mosquitoes. The review states that the technology applied will be useful to focus on a smaller land area. As a future enhancement more technically efficient UAV can be used. Instead of UV and NRI cameras GPS UAV can be used to focus and capture images in a wider geographical area. Multi rotor drones are preferable while covering huge areas. The review states that Drones works continuously for 3-4 hours under recharging. In future solar plates could be fixed for continuous recharging and the operation of UAV.

V. CONCLUSION

Internet of things provides new possibilities for a lot of problems. Many countries, now, have the best mosquito surveillance and control technology. It's true to say that mosquito control has become technically advanced in many places. This paper deals with the study about UAVs in mosquito control using NRI camera. In recent years, UAV are growing as incredible technology specifically in the domain of Internet of Things based technologies .Since the diseases caused and spread by mosquitoes are increasing day by day, new inventions and developments are becoming essential to control and destroy them at the early stage.

Table 1 Classification of UAV

S.no	Type of drone	Application	Battery	Range
1.	Single Rotor UAV	Used for business purposes including inspection,survillence, agriculture and cinematic photography.	10-20min	300-500m
2.	Multi rotor UAV	These UAV are used in aerial photography and survillence[6]	1hr	2km
3.	Tricopter	They are used in power line inspection, aerial Mapping and Pipeline.	2-6 hrs	10 km
4.	GPS UAV	These UAV are linked to satellite via GPS	24 hrs	Wide range
5.	Racing UAV	These UAV are used in competitions	2-3 hrs	2-5 km
6.	Fixed wing UAV	Mainly used in industries.	8-10 hrs	20-30 km
7.	Nano UAV	Light weight UAV used in weapons for spying	20-50 min	1km
8.	Mini UAV	They are used in diaster manadement control [7]	5-10min	2-3km
9.	Closed range UAV	These UAV are he favourite toys for kid [6].	20-45 min	5km
10.	Endurance	These UAV are popular for high end surveillance applications.	36 hrs	300km

In comparison review clearly reveals that UAV will be an efficient tool for coverage of larger geographical area and will be crucial to control mosquitoes.

REFERENCES

[1]. A. Rothstein, The UAV book: second edition, 8 may 2001, p. 25.

[2]. B. Sarah Kurney, UAV and targeted killings: third edition, 2 February 2013, p. 346.

[3]. S. Alex Elliott, Build Your Own Drone Manual: first edition, 4 January 2004, p. 568.

[4]. Duncan Still, How to build a quad copter drone: A complete guide to building a radio controlled quadcopter, third edition, 16 February 2005, p. 204.

[5]. D. Craig S Issod, Getting Started with Hobby Quad copters and UAV: Learn about, buy and fly these amazing aerial vehicles, second edition 10 June 2013, p. 309.

[6]. M. Richard Whittle, The Secret Origins of the Drone Revolution, First Edition, September 16, 2014, p. 468.

[7]. K. Jorge Dias, Lakme sanivertane, A survey of unmanned aerial vehicles: Recent development and trends, 2 November 2014,p. 235.

[8]. J. Lan cinnamon, Romi Kadri, Fitz tepper, DIY UAV for the Evil Genius: Design, Build, and Customize Your Own UAV,25 march 2014,fifth edition, p. 670.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 147-150(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Designing an Integrated Model of Organisational Commitment Among its Employees in Coimbatore using Mancova

J. Arthi
Associate Professor, Avinshilingam School of Management Technology,
Avinashilinam Institute for Home Science and Higher Education for Women, Coimbatore (TN), India.

ABSTRACT: In India over the past few years, the ITES (Information Technology Enabled Services) industry has been growing rapidly and this growth has brought a lot of HR challenges. The biggest challenge of the industry is to manage the ambitious and transient work force. Most of the research in ITES has addressed only specific problems related to challenges of Attrition, HRM systems, Retention, Compensation and Benefits, Job Stress, Job satisfaction etc. The article is a thoughtful endeavour to explore the different ways in pursuit of organisational excellence. In order to understand the organisational commitment of the employees the variable considered are quality of work life and their performance at work. In this era, when technology is taking higher strides, Human Resources Management has become a challenge with more data to be handled. The article throws light on the fourth industrial revolution called Industry 4.0 where innovation and technology working towards organisational excellence. The outcome of the article is to design a model integrating organisational commitment, Quality of Work Life and Job Performance using advanced statistical tool to ensure the future HR management becomes powerful and intellectually handled. Based on this model, a Mobile Application namely Employee Motivation Monitoring System towards enhancing employees' commitment has been developed in tune with managing future HR challenge.

Keywords: Organisational Commitment, Organisational Excellence, Quality of Work Life, Job Performance, Industry 4.0.

I. INTRODUCTION

Organisational Commitment (OC) considered being one of the foremost outcomes of the human resource strategies, Organisational commitment is seen as the key factor in achieving competitive performance. As Indian employees become more entrenched and connected to foreign organisations, it is important for both the client firm and the ITES (Information Technology Enabled Services) operation to identify applicable Human Resource Development and High Performance Management practices as given in [5].

The present article is empirical which portrays the relationship between Quality of Work Life (QWL), Organisational commitment and Job Performance (JP) could be beneficial in addressing the great issues of employee retention and attrition which are highly linked to the ITES sector. The research conducted with the employees of ITES in Coimbatore assesses the work situations that influence commitment and its effect on Job performance. Finally, an empirical model is developed to project the influence of quality of work life on Organisational Commitment and in turn its impact on performance. Based on analysis a Mobile Application in ASP.net Platform had been designed as a suggestive model to have a steady work performance which has been named as Employee Motivation Monitoring System.

II. REVIEW OF LITERATURE

The literature search indicates that OC is linked to various antecedents ranging from personal variables and organisational characteristics. The present research focuses on QWL as a factor that determines Organisational commitment and its impact on Job performance.

QWL can be traced back to the quality of working life movement that largely consisted of a number of industrial psychologists in response to a perceived disenchantment with the organisation of work in the late 1960s and early 1970s as given in [8]. QWL has been associated with organisational changes aimed at increasing the levels of job enlargement and job enrichment.

Work environment is shown in research as a dominant factor of employee performance and commitment as given in [9] .The result of employee's responses to work or organisational environment brings about work outcomes that affect their organisation's overall performance. Generally, organisational performance is indicated by the factors: profitability, market share, innovation, labour productivity, regulatory compliance, and flexibility as given in [3].

The concept Organisational commitment is described as a tri-dimensional concept, characterised by the Affective, Continuance and Normative dimensions as given in [1] Common to the three dimensions of Organisational commitment is the view that Organisational commitment is a psychological state that characterises organisational member's relationship with the organisation and has implications for the decision to continue or discontinue membership in the organisation as given in [2].

Campbell (1990) defines performance as behaviour. It is something done by the employee. This concept differentiates performance from outcomes. Outcomes are the result of an individual's performance, but they are also the result of other influences. In other words, there are more factors that determine outcomes than just an employee's behaviour and actions. It also suggested determinants of performance components. It has been proved that te psychological capital ,job performance , and work attitude are positively related[4].Individual differences on performance are a function of three main determinants: declarative knowledge (knowledge about facts, principles, objects), procedural knowledge and skill (knowledge and skill is knowing how to do it) and motivation. The literature has given greater insights for questionnaire preparation, setting hypotheses, tools for analysis and provided a concrete setting for the research [7].

III. RESEARCH METHODOLOGY

A. Research Design

The research design is descriptive.

B. Sample Size and Techniques

The employees at the operational level from these 20 organisations constitute the universe for the research. A total of 1125 employees are functioning at the operational level in these 20 organisations. The sample size was primarily designed to be 50% that is 563 employees.The procedure was Non probability sampling technique.

C. Tool for Analysis

Multivariate Analysis of Covariance (MANCOVA) is an extension of Analysis of Covariance (ANCOVA) methods to cover cases where there is more than one dependent variable and where the dependent variables cannot simply be combined. Multiple analysis of covariance (MANCOVA) is similar to multiple analyses of variance (MANOVA), but allows to control the effects of supplementary continuous independent variables – covariates. This technique is employed to find the inter effects caused between the variables and design an integrated model. Quality of Work Life (QWL) entered the MANCOVA model as fixed factors, organisational commitment as covariates and Job performance as dependent variables. The researcher decided to run Multivariate Analysis of Covariance (MANCOVA) on the variables that include Quality of Work Life, Organisational Commitment and Job Performance. The decision to run MANCOVA was made in consonant with Bray and Maxwell (1982) and the researcher -

(i) Was interested in finding the effects of Quality of Work Life on several of the Job performance variables such as Knowledge and skills, Quality and Accountability;

(ii) Was interested in the relationships among the number of criterion variables the include Knowledge and skills, Quality and Accountability;[6]

IV. RESULTS AND DISCUSSION

The MANCOVA analysis indicates that Affective commitment and all QWL factors have influence on job performance dimensions. In consolidation, it is pinpointed that QWL has significant impact on all forms of commitment and job performance and shown that QWL has an upshot on job performance through all forms of Commitment.

As specified in fig. 1. The employees in ITES sector want to belong to the organisation (Affective) and there is a possibility of an obligation to remain (Normative), carry with it a commitment to contribute towards performance. Continuance commitment has not added significantly to the prediction of Job Performance. This model is the highlight of the entire research and proves that the major objective of the research to demonstrate the integrated structure of QWL, OC and JP is completely and successfully met.

At this point of HR being managed digitally, the emotional connectedness has to be maintained so that the Human- Machine-Process - Product are well integrated to handle any kind of challenges that impact business performance. Further, based on analysis array of strategies are suggested to steadily improve the work performance. The author had made an attempt to design a mobile Application namely “Employee Motivation Monitoring System” using ASP.net Platform and compatible for android mobile set [8]. Based on MANCOVA analysis the QWL factors are considered as motivational factors and serve as input data to decide the level of motivation of employees.

This motivation index can play a greater role in determining the influence of motivational components on work performance. This application is an evidence that data based HR decision making is needed to solve organisational issues digitally. this application helps the chief administrator to understand the level of motivation of employees everyday in his/her smart phone. HRM is redefined where in future work

environment requires complex problem solving skills on the part of HR Managers. The emerging break through management strategies like Industry 4.0 will mould HR as a strategic partner.

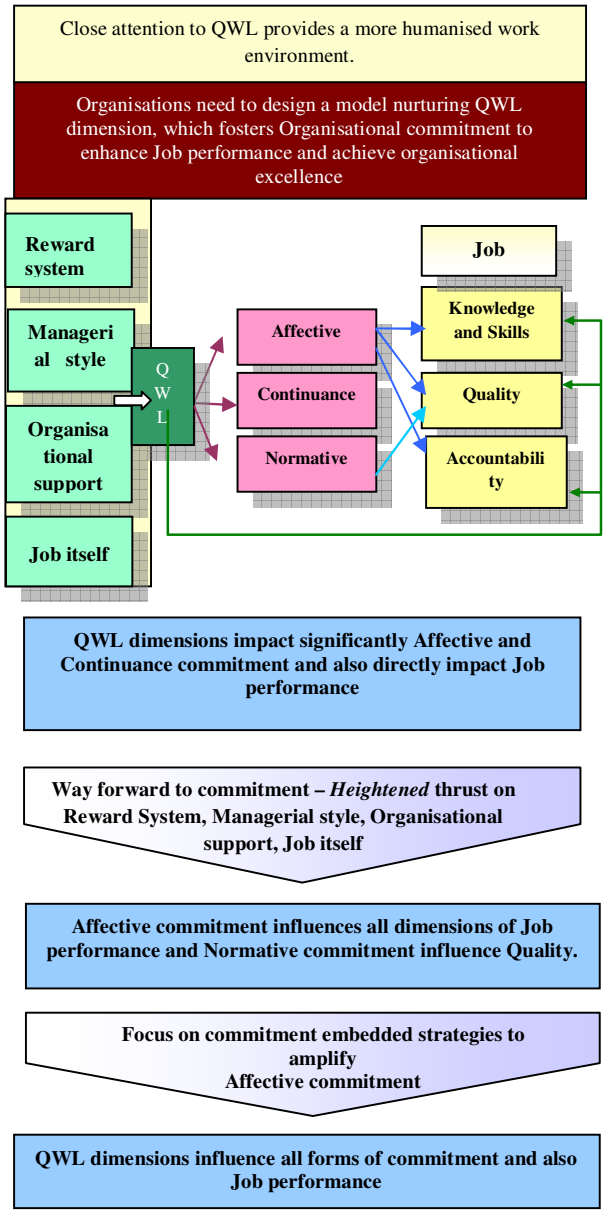


Fig. 1. Inter Effects of QWL, OC, JP Using MANCOVA.

V. CONCLUSION

In the future, teams will increasingly search for, qualify and motivate their members themselves. The human resources department must apply new technologies and concepts to enable every area to identify suitable talents and deploy them in the best possible way. The suggestions given alongside the model will help the employers in ITES sector to design creative work place strategies to activate high level of commitment and thereby augment Job performance. With globalisation and new business environment, organisations are enduring high tension to change policies and practices to stay competitive. The environment is changing towards technological management as decisions need to be taken based on data only, which ensures fair and transparent management. HR managers are profoundly expected

to participate in business transformation as the future is going to be digitalization of people management.

REFERENCES

[1]. Allen, N.J., & Meyer, J.P. The measurement and antecedents of affective, continuance and normative commitment to the organization. *J. Occup. Psychol.*, **63**,1990, pp 1-18.

[2]. Allen, M.W., & Brady, R.M. Total Quality Management, Organisational Commitment, Perceived Organisational Support, and Intraorganisational Communication, *Management Communication Quarterly*, **10**(3), 1997, pp 316-341

[3]. Bratton, J., Grint, K., & Nelson, D. (2005). *Organisational Leadership* (2nd. Ed.) Ohio, South-Western.2005.

[4]. Brett, J., Cron, W., & Slocum, J.W. Economic Dependency on Work, A moderator of the Relationship between Organizational Commitment and Performance. *Academy of Management Journal,* **38**(1), 1995, pp 261-271.

[5]. Budhwar, P., & Sparrow, P. National factors determining Indian and British HRM practices - An empirical study. *Management International Review,* **38**,1998 ,pp 105-121.

[6]. James H. Bray, Scott E. Maxwell , Multi Variate Analysis of variance, sage Publications, Issue:54 ,1985.

[7]. John P Campbell,Jeffrey J Mchenry ,Lauress L wise, Modelling Job Performance in a Population Jobs, *Personnel Psychology*, Vol **43**, 1990, pp 313-575.

[8]. Ruth Mayhew , How to monitor Employee Movitation ,Satisfaction and Performance, http://smallbusiness.chron.com/monitor-employee-motivation-satisfaction-performance-1886.html.

[9]. Walton, R.E., Quality of working life, what is it? *Sloan Management Review,* **15**(1), 1973 , pp 11-21.

International Journal of Theoretical & Applied Sciences, *Special Issue* 10(1): 151-155(2018)
International Conference on SMAC - Reshaping the Future (February 2018)
ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Hierarchical Representation with Multi-Level Fuzzy Clustering of Web Documents

Jeyasree. D and D. Kavitha
Department of Computer Science,
KGiSL Institute of Information Management, Coimbatore (TN), India

ABSTRACT: Hierarchical Representation with Multi-level Fuzzy Clustering (HRMLFC) algorithm, which consists of five-level representation of web documents and a multi-level Fuzzy Clustering algorithm is proposed based on web documents and article structure theory. Initially the proposed algorithm extracts features from the web documents using Conditional Random Field (CRF) methods and builds a fuzzy linguistic topological space based on the associations of features. To deal with the semantic similarity problem resulted from the sparse term paragraph matrix, an ontology based strategy and a Support Vector Machine (SVM) are introduced into HRMLFC algorithm. By using multi-level Fuzzy Clustering, web contents are able to be clustered into topics in the hierarchy depending on their fuzzy linguistic measures.

Keywords: Conditional random field, Support vector machine, Multi-level fuzzy clustering algorithm.

I. INTRODUCTION

A. Web Mining

One of the important technologies is developed at the intersection of data analysis and a Web technology is Web mining. Web mining consists of following three techniques are mainly used in this section:

- Web usage mining
- Web content mining
- Web structure mining

B. Web Content Mining

Web content mining is the eminent process of mining, extraction and integration of useful data, information and knowledge from web.

Web Content Mining Applications

- Identify the topics represented by a web Document.
- Categorize the web Documents. Find web Pages across different servers that are alike.
- Queries: Enhance standard Query Relevance with User, Role, and/or Task Based Relevance.
- Filters: Show/Hide the documents based upon relevance score.

II. LITERATURE SURVEY

Web document clustering is rooted in text mining techniques and shares many concepts with traditional data clustering methods. The purpose of this research is to propose a search methodology that consists of how to find relevant information from World Wide Web (WWW). Clustering method based upon fuzzy equivalence relations is being proposed for web document clustering.

Kanade and Hall present a swarm intelligence based algorithm for data clustering. The algorithm uses ant colony optimization principles to find good partitions of the data.

In the first stage of the algorithm ants move the cluster centers in feature space. The reformulated fuzzy c-means criterion is used to evaluate the cluster centers found through ants. In the second stage the best cluster centers found are used as the initial cluster centers for the Fuzzy C-Means (FCM) algorithm [1]. An original soft hierarchical Fuzzy Clustering algorithm is proposed, named Hierarchical Hyper-spherical Divisive Fuzzy C-Means (H2D-FCM) [3]. Take a different approach to solve all these issues by looking at the co-occurrence network of the keywords of the web pages. The nodes are represented by keywords and edges between keywords imply that they appear together at least once in a web page of a co-occurrence graph network [5].

III. PROPOSED SYSTEM

In this work Hierarchical Representation with Multi-level Fuzzy Clustering (HRMLFC) algorithm is proposed which consists of five-level representation of web documents and a Fuzzy Clustering algorithm is proposed based on web documents and article structure theory. At initially the proposed algorithm extracts features from the web documents using Conditional Random Field (CRF) methods and builds a fuzzy linguistic topological space based on the associations of features. To deal with the semantic similarity problem resulted from the sparse term-paragraph matrix, an ontology based strategy and a Support Vector Machine (SVM) are introduced into HRMLFC algorithm. By using Fuzzy Clustering, web contents are able to be clustered into topics in the hierarchy depending on their fuzzy linguistic measures. Thus the clusters are more efficiently and effectively captured in HRMLFC and thus web document clusters with higher quality can be generated. HRMLFC based algorithm the concepts of fuzzy linguistic topological spaces, which can discover the latent semantics in text corpora.

IV. HIERARCHICAL REPRESENTATION WITH MULTI-LEVEL CLUSTERING

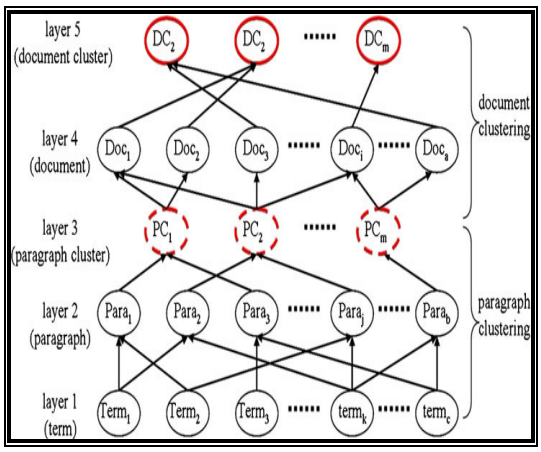


Fig. 1. Hierarchical Representations with Multi-Level Fuzzy Clustering (HRMLFC).

Propose a CRF model by using multi-level computing of document representation and clustering is illustrated in Fig 1, which contains five layers and a two-phase clustering process.

A. Support Vector Machine (SVM)

Given some training data $\mathbf{D}$, a set of n points of the form

$$D = \{x_i, y_i\} | x_i \in \mathbb{R}^p, y_i \in \{-1,1\}_{i=1}^n \quad (1)$$

Where the y_i is either 1 or −1, indicating the class to which the point x_i belongs. Each x_i is a p-dimensional real vector. Find the maximum-margin hyperplane that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of point's x satisfying the maximum-margin hyperplane and margins for an SVM trained with samples from two classes. The support vectors are the samples on the margins.

$$w.x - b = 0 \quad (2)$$

where denotes the dot product and w as not necessarily normalized normal vector to the hyperplane. The parameter $\frac{b}{||w||}$ determines the offset of the hyperplane from the origin along the normal vector w. The linear separability of the training data, only if they can select two hyperplanes in a way that they could separate the data and there are no points between them, and then try to maximize their distance. The bounded region through the hyperplanes is called "the margin".

These could be described through the equations $w.x - b = 1$ and $w.x - b = -1$.

This can be rewritten as: $y_i(w.x_i - b) \geq 1 \quad \text{for all } 1 \leq i \leq n$

Minimize $(w,b)||w||$ subject to (for any $i = 1, \ldots n$) $y_i(w.x_i - b) \geq 1$

If the kernel used is a Gaussian radial basis function, then it corresponds to a feature space is a Hilbert space of infinite dimensions., The infinite dimensions do not spoil the results as maximum margin classifiers are well regularized. Polynomial (homogeneous):

$$k(x,y) = \left(\sum_{i=1}^{n} x_i y_i + c\right)^2 = \sum_{i=1}^{n} (x_i^2)(y_i^2) + \sum_{i=2}^{n} \sum_{j=1}^{i-1} (\sqrt{2}x_i x_j)(\sqrt{2}y_i y_j) + \sum_{i=1}^{n} (\sqrt{2}c x_i)(\sqrt{2}c y_i) + c^2$$

(3)

B. Conditional Random Field (CRF)

A CRF is a simple framework for labeling and segmenting data that models a conditional distribution P(z|x) by selecting the label sequence z to label a novel observation sequence x with an associated undirected graph structure that obeys the Markov property . When conditioned on the observations that are given in a particular observation sequence, the CRF defines a single log-linear distribution over the labeled sequence. Let W = {x_1, x_2, · · · , x_2 } denote features extracted from documents, and ϕ = {z_1, z_2, · · · , z_m} be the named category set. A membership function for z_j is represented as ρ_{z_j} that the value $\rho_{z_j}(x_j)$ = P(z_j |x_i) evaluates the membership degree of xi belonging to named category z_j.

The term frequency of a feature in a document belonging to a named category is written as tf_z, and the inverse document frequency of a feature in a document belonging to a named category is written as idf_z, which are

$$tf_z(x) = tf(x) \times \rho_z(x) \quad (4)$$

and

$$idf_z(x) = idf(x) \times \rho_z(x) \quad (5)$$

The membership of a feature x of a document belonging to a named category z is defined as

$$\mu_z(x) = tf_z(x) \otimes idf_z(x) \quad (6)$$

where $tf_z(x)$ denotes the number of feature x in a document being classified into z, $idf_z(x)$ denotes inverse document frequency where document frequency is the number of documents that contain the feature x in category z, and $\otimes$ is a fuzzy operator as shown in Fig. 1 where α and β are two thresholds respectively for term frequency and inverse document frequency:

$$\mu_z(x) = \begin{cases} 1 \\ tf_z(x) \leq \alpha or\ idf_z(x) \geq \beta \\ (tf_z(x) - \alpha) \times (idf_z(x) - \beta) \end{cases} \quad (7)$$

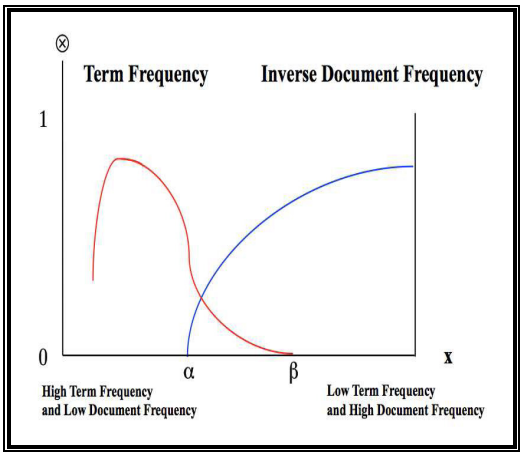


Fig. 2. Document Frequency Methods.

V. RESULTS AND DISCUSSION

A. Performance Measures

Considering the contingency table for a topic of a category (Table 1); recall, precision, and F-measure are three direct measures of the effectiveness of a NER method.

Table 1 The Contingency Table for Zi Precision and recall with respect to a topic are respectively defined as follows:

$$Precision_i = \frac{TP_i}{TP_i + FP_i} \quad (8)$$

$$Precision_i = \frac{TP_i}{TP_i + FN_i} \quad (9)$$

The F_β combines precision and recall by the following formula:

$$F_\beta = \frac{(\beta^2 + 1) \times Precision_i \times Precision_i}{(\beta^2) \times Precision_i + Precision_i} \quad (10)$$

The measure used in this work was obtained when is set to be 1, which means precision and recall are equally weighted for evaluating the performance of clustering.

B. Normalized Mutual Information (NMI)

Given the two sets of topics C and C′, let C denote the topic set defined by experts, C′ denote the topic set generated by a clustering method, and both derived from the same corpora X. Let N(X) denotes the total number of documents, N(z, X) denotes the number of documents in a topic z, and N(z, z′, X) denotes the number of documents both in topic z and topic z′, for any topics in C. The Normalized Mutual Information (NMI) metric MI(C,C′) is defined as follows

$$MI(C, C') = \sum_{z \in C, z' \in C'} P(z, z') \log_2 \left(\frac{P(z, z')}{P(z)P(z')} \right) \quad (11)$$

where P(z) = N(z, X)/N(X), P(z') = N(z'X)/N(X), and P(z, z') = N(z, z'X)/N(X). The normalized mutual information metric MI(C,C′) will return a value between zero and max(H(C),H(C')) where H(C) and H(C') define the entropies of C and C', respectively.

The higher MI(C, C') value means that two topics are almost identical, otherwise more independent. The normalized mutual information metric $\widehat{MI}$(C, C' is therefore transferred to be

$$\widehat{MI}(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))} \quad (12)$$

The Jaccard coefficient measures similarity between finite sample sets. The size of the intersection divided by the size of the union of the sample sets are defined as:

$$J(C, C') = \frac{J(C \cap C')}{J(C \cup C')} \quad (13)$$

C. Dataset Information

Table 1. Statistics of reuters-21578 corpora.

Statics	Number of topics	Number of documents	randomly selected documents
Origin	135	21578	0~3945
Single topic	65	8649	1~3945
Single topic (>=5 document)	51	9494	5~3945

Table 2: Performance comparison results of reuter S-21578.

Method	NMI	Jaccard Index	accuracy	Error	Time
Fuzzy cluster	00. 61484	0.60000	0.7619	0.2381	7mins
HRMLFC	0.62084	1.00000	0.9048	0.0952	3mins

In the following Figs. 3 to Fig. 7 shows the performance comparison results of five different metrics such as NMI, Jaccard Index comparison, Accuracy, Error Rate and the time comparison between the proposed HRMLFC and existing fuzzy clustering methods. However the proposed HRMLFC and existing fuzzy clustering produces NMI results of 0.61484 and 0.62084 respectively is illustrated in Fig. 1.

Proposed HRMLFC and existing fuzzy clustering methods produces Jaccard index results of 0.6000 and 1.000 respectively is illustrated in Fig. 2. Proposed HRMLFC and existing fuzzy clustering methods produces accuracy results of 0.7619 and 0.2381 respectively is illustrated in Fig. 3. It concludes that the proposed work performs better for all three metrics are illustrated in Figs when compared to existing methods. The proposed HRMLFC algorithm produces lesser error rate results of 0.0952, whereas the existing fuzzy clustering method produces error results of 0.2381. It concludes that the proposed work performs better when compared to existing fuzzy clustering method.

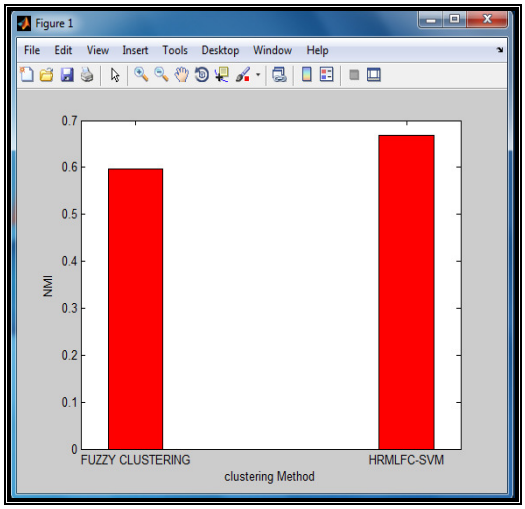


Fig. 3. NMI vs. clustering methods.

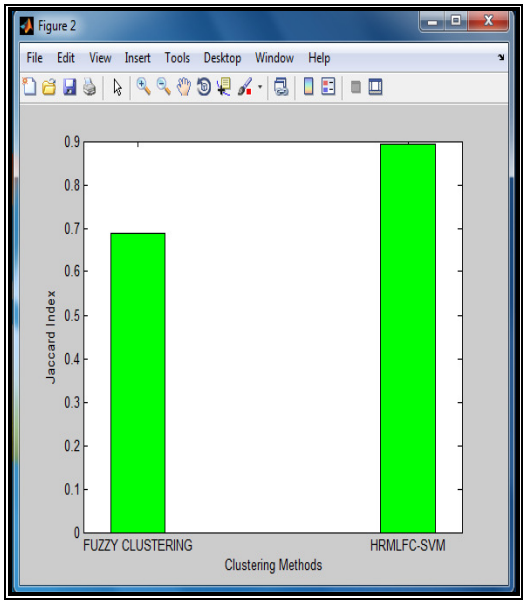


Fig. 4. Jaccard Index vs. clustering Methods.

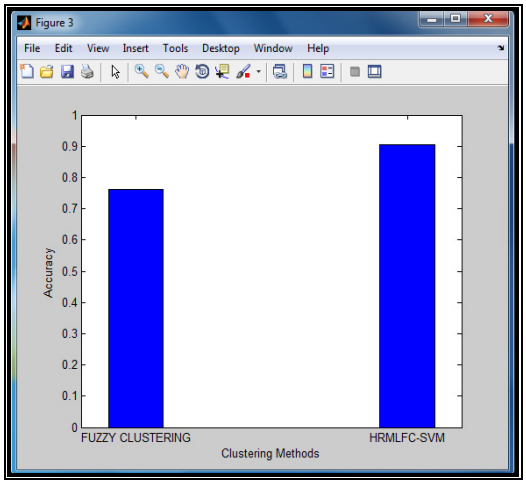


Fig. 5. Accuracy vs. clustering methods.

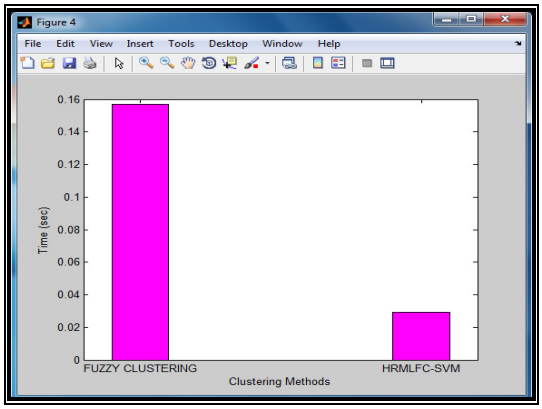


Fig. 6. Execution time vs. clustering Methods.

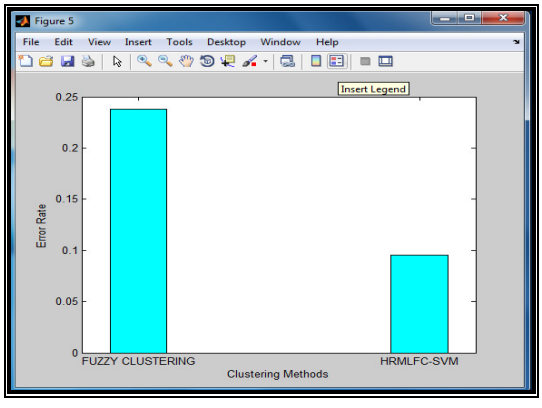


Fig. 7. Error rate vs. Clustering methods.

The proposed HRMLFC algorithm produces takes lesser execution time of 0.0014 seconds, whereas the existing fuzzy clustering method takes higher execution time of 0.0043 seconds. It concludes that the proposed

work performs better when compared to existing fuzzy clustering method.

VI. CONCLUSION

Hierarchical Representation with Multi-level Fuzzy Clustering (HRMLFC) algorithm, which consists of five-level representation of web documents and a Fuzzy Clustering algorithm is proposed based on web documents and article structure theory. At initially the proposed algorithm extracts features from the web documents using Conditional Random Field (CRF) methods and builds a fuzzy linguistic topological space based on the associations of features. To deal with the semantic similarity problem resulted from the sparse term-paragraph matrix, an ontology based strategy and a Support Vector Machine (SVM) are introduced into HRMLFC algorithm. All features in documents compose a topologically probabilistic space, more specifically a simplical complex associated with probabilistic measures to denote the underlying structure. Effectively discover maximal fuzzy simplexes and use them to cluster the collection of web documents. Based on the web site and experiments, we find that HRMLFC is a very good way to organize the unstructured and semi structured data into several semantic topics. It also illustrates that geometric complexes are an effective model for automatic web documents clustering.

REFERENCES

[1]. P. M. Kanade and L. O. Hall, "Fuzzy ant clustering by centroid positioning," in Proc. of 2004 IEEE International Conference on Fuzzy Systems, Budapest, Hungary, 2004, pp. 371–376.

[2]. Ronen Feldman and James Sanger. The text mining handbook: advanced approaches in analyzing unstructured data. Cambridge University Press, 2007

[3].. G. Bordogna and G. Pasi, "Hierarchical-hyperspherical divisive fuzzy c-means (h2d-fcm) clustering for information retrieval," in Proc. Of 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Workshops, Milano, Italy, 2009, pp. 614– 621.

[4]. Richard O. Duda, Peter E. Hart, and David G. Stork. Pattern classification. Wiley, 2001.

[5]. F. Zaidi and G. Melancon, "Organization of information for the web using hierarchical fuzzy clustering algorithm based on co-occurrence networks," in Proc. of 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Toronto, Canda, 2010, pp. 421–424.

[6]. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, 3:993–1022, 2003.

[7]. C.M. Sperberg-McQueen T. Bray, J. Paoli and E. Maler, editors. Extensible Markup Language (XML) 1.0 (Second Edition). World Wide Web Consortium., October 2000.

International Journal of Theoretical & Applied Sciences, Special Issue 10(1): 156-160(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Search Optimization in Selective Search Engines - A Survey

S. Amudha[1] and Dr. I. Elizabeth Shanthi[2]
[1]Ph.D. Research Scholar, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India.
[2]Professor, Department of Computer Science,
Avinashilingam Institute for Home Science and Higher Education for Women, Coimbatore, India

ABSTRACT: In today's technology, the role of internet is rapidly increasing and will continue to do so in the upcoming days. In this rapid increase, the exact data access and integration has become a challenge. Since 93% of internet traffic is managed by search engines, discovering the capability of search engines is crucial. Due to the extensive use of search engines like Google, search results are gaining more importance for websites to compete with other competitors. The most crucial part of managing different opponents is optimization of search engine. This paper is mainly focuses on the basics of search engine optimization techniques in yahoo, Google and Bing.

Keywords: Search Engine, Search Engine Optimization, Yahoo, Google, Bing

I. INTRODUCTION

The search engine is a program which is used to browse the content by the user in World Wide Web and retrieves the expected information by the user in short period of time [3]. The user enters the input for specific keyword or phrases and retrieves a collection of content related to those keywords or phrases. This information may not provide the exact content; but information can be retrieved using keyword or search terms. This search result is called as search engine result page [SERPs] [1]. The search result information is a combination of web pages, image, video and other type of files [6]. Most of the search engine can be retrieved only by available information in database or other directories [5]. The web directories are maintained by human editor, and also maintain real time information using algorithm on a web crawler. Search engine helps millions of users per day looking for their solution to their problem. Search engine optimization is a way to improve the concerned that so it will move to top position in the search result pages of the respective search engine.The most popular search engines used English language documents like Yahoo's web search is currently powered by Bing and Google [10]. For example while searching on Google for returning the results and complex algorithm used considers the number of factors into account to decide which web page should be placed in order like web page1, web page 2 etc. The procedure for search engines like Google, yahoo, and Bing are unique that they tap the targeted traffic and provide a roadway for what the user intends to get [9]. Every search queries takes either single word or multiple words. The performances of search engines are improving day by day [13].

The URL's that result from search query related to website appear in the search engine ranking page [SERP]. The SERP list provides the ranking position in the particular web site of the search engine.

II. HOW DOSE A SEARCH ENGINE WORK?

A search engine performs numerous activities in order to provide search results [4]. The figure 1 shows the working of search engine.

- **Crawling** – The crawler is software that performs fetching of all the web pages linked to a website. This crawler is also known as "***Spider"***. The Google search engine uses the crawler named as "***Googlebot"*** [2].
- **Indexing** –The second process is creating an index for all the web pages and maintain into a huge database. Indexing is used for identifying the words and expressions in the web pages quickly. This index is the value assigned to the web page for given keyword [7].
- **Processing** – Here the searching request for the user contain a keyword or phrase entered into the search engine is processed. Then searching

keyword is compared with indexed web pages in the database.

- **Calculating Relevancy** – The matching output of searching keyword or phrase may occur in more than one page. Now to device accurate matching the search engine is calculating the relevancy of every page with index to the search keyword or phrase.
- **Retrieving Results** – The last process of search engine is retrieving the similar results based on the search keyword or phrase. Finally it is displayed in the web browser.

The search engine Yahoo, Google, and Bing [YGB] update the relevancy algorithm and Ranking algorithm in many times based on version up gradation. Figure 2 shows the customer access of different search engines as 64% in Google, 22% in Bing and 14% in Yahoo.

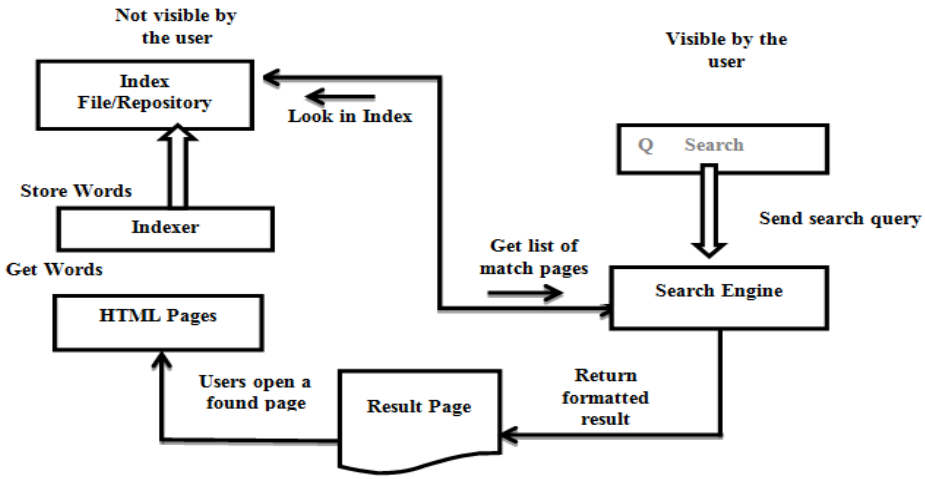


Fig 1: Working of Search Engine

III. SEARCH ENGINE OPTIMIZATION [SEO]

Search engine optimization is a set of rules that are followed to optimize the search engine websites and to improve the ranking. SEO further helps to increase the quality of the web site, provides quick access and navigate through web pages easily.

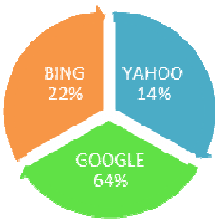


Fig. 2. Average use of Search Engine.

Search engine optimization framework considers group of rules, number of stages and controls. Search engine optimization is categorized into two stages as on-site SEO and off-site SEO.

The on-site SEO is done internally on a website that includes quality content, good sets of keywords, assign keywords on correct places, link and website structure, and more. The second category Off-page SEO - is done externally by doing the link building, increasing link popularity and reputation. Figure 3 represents the overall process of SEO.

IV. SEARCH ENGINE OPTIMIZATION TECHNIQUES

SEO techniques are classified into two wide categories [14]:

- **White Hat SEO** - This technique is the one that search engines recommend as part of a good design.

An SEO strategy is considered as White Hat if it has the following features:

- It conforms to the search engine's suggestions.
- It does no longer involve in any deception.
- It guarantees that the content Search engine indexes, and finally, ranks are the identical content material a user will see.

- It guarantees that an internet web page content material should have been created for the users and now not only for the search engines.
- It guarantees precise first-class of the web pages. It guarantees the availability of beneficial content material on the web pages.

Fig. 3. Search Engine Optimization Process.

Black Hat or Spamdexing

Black Hat SEO is also known as spamdexing. This technique used to increase a website or page rank in search Engine. A search engine optimization approach is taken into consideration as black hat or spamdexing if it has the forthcoming features:

- Attempting ranking upgrades which can be disapproved through the search engines like Google and/or involve deception.
- Redirecting users from a web page this is built for search engines like Google, Bing and yahoo and it has greater user-friendly.
- Serving one version of a page to search engine spiders/bots and every other model to human visitors. This is known as cloaking search engine optimization tactic.
- The use of hidden or invisible textual content or with the web page background coloration, the usage of a tiny font size or hiding them inside the HTML code such as "no frame" sections.
- Repeating key phrases within the metatags, and the use of key phrases that are unrelated to the internet site content. That is called metatag stuffing.
- Calculating placement of keywords within a web page to elevate the key-word count, variety, and density of the web page. That is referred to as keyword stuffing.
- Developing low-first-rate web pages that incorporate little or no content but are as a substitute full of very comparable key phrases and phrases. These pages are called doorway or gateway pages.
- Mirror websites with the aid of hosting more than one websites - all with conceptually comparable content but the use of extraordinary URLs.
- A rogue copy of a popular website which indicates contents similar to the original to a web crawler, but redirects internet surfers to unrelated or malicious websites. It is called as page hijacking.

V. COMPARITIVE ANALYSIS

The following are the observations made from Google and Bing SEO characteristics. The major differences between Bing and Google are page ranking. Bing SEO ranking factors are found to be most important by a recent search metrics analysis [8] [11] [12].

- Top brands tend to rank higher on Google.
- Google seems to give ranking preference to brand.
- Bing also gives ranking preferences to brand.
- Google considers more back link with the name of the brand in the link text alone known as “brand link”
- Bing considers more back link with the name of the brand in the link text alone.
- The Bing search engine has some difficulty distinguishing brands from related competitors.
- Social signals are relate closely with higher rankings.
- When a user searches on Bing, they can immediately view.
- Backlink numbers are closely linked to higher rankings.
- Relevant and quality content are important for search rankings.
- Relevant and quality content correlate strongly with good ranking on Bing,
- Google relies much more on text based content.
- Bing seems to be more likely to reward pictures, videos, audio and more.
- On page technical factors also available in Bing.

Common features comparison of YGB is listed in Table 1.

The step of ranking algorithms of YGB is listed in Table 2.

Table 1: Feature Comparison in YGB.

Search Engine	Boolean	Default	Proximity	Truncation	Fields	Limits	Stop	Sorting
Yahoo!	AND, OR, NOT,( ), -	and	Phrase	No word in phrase	intitle, inurl, link, site, more	Language, file type, date, domain	No	Relevance, site
Google	-, OR	and	Phrase	No Auto stem word in phrase	intitle, inurl, link, site, more	Language, filetype, date, domain	Varies	Relevance, site
Bing	AND, OR, NOT, ( ), -, +	and	Phrase	No Auto stem	intitle, inurl, link, site, more	Language, filetype, date, domain	No	Relevance, site

Table 2: Comparison of Ranking Algorithm in YGB.

Yahoo Ranking Algorithm	Google Ranking Algorithm	Bing Ranking Algorithm
1. The title of website must contain major keywords. The title is the biggest ranking factor of Yahoo's raking algorithm. 2. The description to give when submitting to yahoo web directory should include major keywords, however don't try to repeat them too much. Very important step. 3. Click Popularity is part of Yahoo's Algorithm, to determine one website ranking. The more visitors click on website from Yahoo SERP's, the more you'll get close to the Top Ranking. 4. The category are listed in Yahoo Web Directory should contain some keywords. This plays a little with the ranking. 5. Site-wide linking. It's a choice have to make, you can get one link per domain in order to rank well on Google search engine but you can also get a lot of site-wide linking from other site, Yahoo loves it.	1. The first variable is the text inside website, Google have a text machine system that check which theme of website is about and how "good" content is, this variable helps website to have relevant advertisements if using Google AdSense. 2. It also helps users on internet to have relevant SERP's regarding what keywords they used when researching on Google Search Engine. 3. The other variable of Google Algorithm is to calculate the link popularity of websites using the software PageRank. It gives website a score from 0 to 10. But PageRank is only used for Google algorithm 4. A website with a PageRank of 7/10 will have better rankings than a website with 1/10. 5. A Listing inside DMOZ Web Directory is part of the algorithm; a listing in DMOZ is a good thing so should be focus on DMOZ before focusing on Google. 6. Google Ranking Algorithm and its Text Machine System. 7. Google Robot (Googlebot) check website swiftly by looking at special keywords spots, to calculate a results with their special text machine, such as: - The title of website - The Heading Tags (H1 > H2) When using html tags - The text of website must be structured and if possible with no (grammar, writing) typos. PR(A) = (1-d) + d (PR(T1)/C(T1) + … + PR(Tn)/C(Tn)) Or PageRank of site=∑(page rank of inbound link)/(number of link on the page)	1. **Backlinks are of less importance.** If you compare the first 10 results in Bing and Google, it is noticeable that all equal, the winners in Bing have less backlinks than the winners in Google. It is unclear if no follow matters with Bing. 2. **Inbound anchor text matters more.** The quantity of quality inbound links might be of less importance for Bing but the anchor text certainly matters more. Actually, since anchor text is one of the measurements of the quality of inbound links, it isn't much different. 3. **Link spamming won't do much for you on Bing.** Since the quantity of backinks seems to be of less importance to Bing, link spamming will be even less effective than with Google. 4. **Onpage factors matter more than with Google.** This is one of the most controversial points. Many SEO experts disagree but many also think that onpage factors matter more with Bing than with Google. Still, it has nothing to do with the 90s, when onpage factors were definitive. 5. **Bing pays more attention to the authority of the site.** If this is true, this is bad news for bloggers and small sites because it means that search results are distorted in favor of older sites and/or sites of authoritative organizations. Age of domain is also very important with Bing – even more than with Google. 6. **PR matters less.** When you perform a search for a competitive keyword and you see a couple of PR2 or even PR1 sites among the top 10 results, this might make you wonder. On Google this is hardly possible but on Bing it looks quite normal. 7. **Fresh content matters less.** Bing looks a bit conservative – or maybe it just can't index sites that quickly – but it seems that fresh content is not so vital as with Google. This is related to the age of domain specifics and as a result you will see ancient pages rank high. 8. **Bing is more Flash-friendly.** Optimizing a Flash site for Google is a bit of a SEO nightmare. It is too early to say but it looks like Bing is more Flash-friendly, which is good news to all sites where Flash is heavily employed.

VI. CONCLUSION

Search engines are used in day to day activities improve for information retrieval on the web pages. Technique such as indexing, processing, calculating relevancy and relevancy of result are used in web page search optimization. This paper provides an insight on the characteristic of yahoo, Google and Bing search engine. From the analysis made it is inferred that Google is more user friendly which is widely used by internet users. On the hand Bing provides better compatibility with social media, where as Yahoo performs moderately internet transfer. Nowadays the web content is growing fast and duplicate content gets increased. Hence our future work is reducing similarity content in the multiple web pages from different servers with more accuracy and less time.

REFERENCES

[1]. Ankita Malve, Prof. P. M. Chawan," A Comparative Study of Keyword and Semantic based Search Engine", International Journal of Innovative Research in Science, Engineering and Technology ,*(An ISO 3297: 2007 Certified Organization)* Vol. **4**, Issue 11, November 2015.

[2]. Ayar Pranav, Sandip Chauhanm," Efficient Focused Web Crawling Approach for Search Engine" International Journal of Computer Science and Mobile Computing, *IJCSMC, Vol.* **4**, Issue. 5, May 2015, pg.545 – 551.

[3]. *B.T. Sampath Kumar, J.N. Prakash,"* Precision and Relative Recall of Search Engines: A Comparative Study of Google and Yahoo", *Singapore Journal of Library & Information Management* Volume **38**, 2009.

[4]. Joseph Edosomwan, Taiwo O. Edosomwan," Comparative analysis of some search engines", *South African Journal of Science.*

[5]. Kailash Kumar, Vinod Bhadu, "A Comparative Study Of Byg Search Engines ", *American Journal of Engineering Research (AJER)* e-ISSN: 2320-0847 p-ISSN: 2320-0936 Volume-**2**, Issue-4, pp-39-43.

[6]. Kamlesh Kumar Pandey, Rajat Kumar Yadu, Pradeep Kumar Shukla," Internet Search Engine: Precision and Relative Recall Analysis of Google, Yahoo and WebCrawler Search Engine Based On Their Searching Capability", *International journal of advance Research in Science and Engineering,* Vol. **5**, Issue No.01, January 2016.

[7]. Krishan Kant Lavania, Sapna Jain, Madhur Kumar Gupta, Nicy Sharma," Google: A Case Study (Web Searching and Crawling)" *International Journal of Computer Theory and Engineering*, Vol. **5**, No. 2, April 2013.

[8]. Manika Dutta, Dr. K. L. Bansal," A Review Paper on Various Search Engines (Google, Yahoo, Altavista, Ask and Bing)", *International Journal on Recent and Innovation Trends in Computing and Communication* ISSN: 2321-8169 Volume: **4** Issue: 8 190 – 195.

[9]. Ouphachay Thongsamouth," The Comparative Study of Search Engines Google versus Bing", ICT 13 January 2016.

[10]. Rakesh Agrawal, Behzad Golshan, Evangelos Papalexakis," study of Distinctiveness in Web Results of Two Search Engines", *The International World Wide Web Conference Committee (IW3C2)* ACM 978-1-4503-3473-0/15/05.

[11]. Sanjib Kumar Sahu, D.P. Mahapatra, R.C. Balabantaray," Comparative Study of Search Engines In Context Of Features And Semantics", *Journal Of Theoretical And Applied Information Technology 20th June 2016*. Vol. **88**. No.2, ISSN: 1992-8645, E-ISSN: 1817-3195.

[12]. Sridhar Neralla, Renuka Devi J, Nirmala A, Swarna M," Study and Comparison of Various Search Engines' Browsing Capabilities", *International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014),* Vol. **2**, Issue 3 (July - Sept. 2014).

[13]. Tauqeer Ahmad Usmani, Durgesh Pant, Ashutosh Kumar Bhatt," A Comparative Study of Google and Bing Search Engines in Context of Precision and Relative Recall Parameter", *International Journal on Computer Science and Engineering (IJCSE),* ISSN: 0975-3397 Vol. **4** No. 01 January 2012

[14]. Vidya Suryakant Patil, Ovhal Prajakta, "Study of Search Engine Optimization Techniques", *International Journal of Computer Engineering and Applications*, Volume **XI**, Issue XI, Nov. 17, www.ijcea.com ISSN 2321-3469.

International Journal of Theoretical & Applied Sciences, *Special Issue* 10(1): 161-165(2018)
International Conference on SMAC - Reshaping the Future (February 2018)

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

Salient Methods of Image Processing: A Fundamental Survey

Mrs. Umamaheswari. D[1] and Dr. E. Karthikeyan[2]
[1]Asst. Professor, U.G. Department of Computer Applications, N.G.M College, Pollachi
[2]Head, Dept of Computer Science, Govt. Arts College, Udumalpet

ABSTRACT: In the field of Information Technology due to the advancement of techniques, the amount of data being gathered is increasing day by day. The technique to analyse these data is known as data mining. Finding out the missing data in the images is really an important problem. Inorder to find out the missing data, the image taken as input should be denoised and enhanced. So that the processed image should be segmented to smaller pieces to find out the meaningful insights about the objects and finally the patterns present in the image are recognised and checked with the matching data sets in order to find out the missing data.

Keywords: Image denoising, Image enhancement, Image segmentation, Pattern matching.

I. INTRODUCTION

Generally the digital images produced through cameras and scanners are definitely affected by noise blur, contrast and low colour balance. The quality of image is examined thoroughly so that the expected result will be obtained. Denoising an image will make the image noise free and clean. After that image enhancement is applied to minimize the degraded effects. The enhanced image will be of good quality and better contrast. The image enhancement process brings out the hidden details and the output image of this process will be more applicable for the expected purposes. Then the image should be segmented and Image segmentation is used to divide the processed image into several pieces which provides essential information. Data analysis tool is used to discover previously unknown, valid patterns. The frequent pattern mining algorithms determine the frequent patterns from an image. So that the patterns are compared and matched in order to find the missing data present in an image.

II. LITERATURE SURVEY

Nowadays remote sensing and medical sensing are the most important image based practical applications. The images play a vital role in medical field, so the concentration is laid on image processing. Mostly the images captured by modern cameras are corrupted by noise. This paper introduces an excellent method based on Generalised Cauchy (GC) distribution. Particle Swam Optimization algorithm selects the filter parameters according to the noise level in order to remove the noise. The proposed algorithm achieves the maximum Peak Signal to Noise Ratio value and it preserves the edge and image details [5]. The main advantage of this method is that it can be implemented easily.

The quality of image and accurate results are the most important effects regarding image processing in the field of biometric identification and authentication systems. In day-to-day life biometric systems are playing an important role in the security. Noise free images provide good quality of the finger prints. Because a fetus fingerprints are fully developed at the age of seven months and it will not change throughout the lifetime. Sometimes the injury, disease or decomposition after death may cause alteration. However the pattern will grow back after the injury heals [11]. Different filtering techniques such as Average filtering, Median filtering and Adaptive Wiener filtering were applied to clear the noise and the performance was compared using a statistical approach called the correlation value.

The image quality is examined thoroughly during the lung disease processes like diagnostic, prognostic and follow-up. Due to heavy death rate, the lung cancer should be identified at an early stage. Each and every image contains noise. Hence removal of noise should be the primary objective in medical image processing. Many minor details are hidden by the noise and outliers present in the medical images. That's why removal of noise in an image plays a significant role in the image analysis. Filters are tools that are used for removing noise from an image. This paper provides the different types of noise that are present in the images and also the methods for obtaining clear images and noise removal methodologies are discussed thoroughly [13].

Definitely noise can affect quality of digital images. Sometimes filtration compromises the level of details.

In order to quantify noise level Signal-to-Noise Ratio (SNR) is checked. The higher the SNR value, better the quality of the filtered image. Noise in medical images manifests itself as single pixels brighter or much darker than the neighbourhood. So that the erroneous pixels

can be considered as local extremes of image intensity. Particular attention is paid to random noise in the author's paper. The various analysis of the received results conclude that the selected number of iterations improve signal-to-noise ratio. The proposed method in the author's paper can be very much useful in the case of noisy medical images presenting different details [2]. It can be applied successfully in all image processing and image enhancement process.

Ophthalmologists have used several techniques for early detection of disease. Optical Coherence Tomography (OCT) is one such technique that provides high resolution images. But OCT images contain speckle noise. In daily life, OCT technique is widely used by ophthalmologists to diagnose the various diseases like Glaucoma, Macular Edema etc. Speckle noise is also a special type of noise which carries information about the image, acting as a major degrading factor of an image. It is proved that the noise will be an obstacle in biomedical image for diagnosis of different diseases. This paper shows that wavelet denoising filter provides good results on all OCT images. Bilateral filter is considered to be worst and wavelet filter is considered the best for removal of speckle noise [4].

Inorder to obtain accurate structure for random vibration signals, removing the interference factors and noise components from the collected vibration signals are expected. Sometimes in research, the original or true vibration signal is not applied, instead tested signal is applied. But it tends to produce errors and wrong conclusions. By comparing the wavelet, Infinite Impulse Response numeric filter and singular entropy, the SNR value of singular entropy is the highest. The determination of order to denoise by singular spectrum is more reasonable [10]. It is very important to extract the accurate signal feature and reliable analysis.

Image processing is, taking an image as input and providing the processed image as output. The digital images produced through cameras and scanners are definitely affected by noise, blur, contrast and improper colour balance. Automatically the quality of these images will be low. Inorder to minimize the degraded effects, image enhancement is applied. The high frequency content images are produced using Discrete Wavelet Transform techniques. The performance enhancement for very dark images can be retrieved using adaptive DWT based Dynamic Stochastic Resonance (DSR) technique. The result of this technique provides better enhancement for very dark images [18]. Internoise is used to improve the performance and provides less computational complexity.

The main aim of image enhancement is used to provide good image quality for better visualization. Improvement of interpretability or perception of information in images is done by image enhancement. It is used to remove noise, enhance dark image and highlight the edges in an image. So the enhanced image is more suitable for various applications. The image contrast algorithms include gray scale manipulation, filtering and Histogram Equalization (HE) [17]. The image enhancement methods bring out hidden details in an input image and increase the contrast in a low contrast image. The major working of human brain is involved in processing and analysing images.

Image enhancement provides various methods to process the input image and supplies the output image. The output image provided by the image enhancement methods is more applicable than the original image. Image enhancement is applied in various fields where analysed images are used. This paper proposed a genetic algorithm related enhancement method. Contrast enhancement plays an important role in image processing. Histogram Equalization is an important method for image contrast enhancement. But sometimes it produces an un-natural looking images. Inorder to overcome this problem, the author's paper proposes a solution with contrast enhancement method based on genetic algorithm which provides natural looking images [8].

The processing of more useful images and improving the quality of images are done by image enhancement. Based on the original input image, Histogram Equalization helps to display the enhanced output images. HE is mainly used in the field of contrast enhancement. This paper presents an algorithm which focuses on utilization of Histogram Equalization. It also proposes a new binary preserved Histogram Equalization. This proposed algorithm reduces the complexity [20]. It also proves that HE enhances the contrast and preserves the image with proper brightness.

Image enhancement delivers clear image for edge detection, segmentation and other image processing steps. Image fusion provides better results for multi resolution images. Image fusion provides fused images with most of the information from the input images by combining relevant information from the same input images. This paper discusses the implementation of various fusion algorithms by using metric measures such as Average Difference, Normalized Mean Square Error and Peak Signal to Noise Ratio [19]. The image fusion methods provide better results than the other general image enhancement methods. Image segmentation being one of the significant steps that leads to the study of processed image data, this paper presents a study of problems being encountered and the issues regarding segmentation. The aim of image segmentation is domain independent partitioning of an image into a set of disjoint region that are visually not same. The authors investigate and then discuss the different popular segmentation techniques [12]. With this analysis, the author's paper concludes that, which segmentation suits for which application domain. But

no single algorithm serves all types of images and provides good performance.

Segmenting an entire image into several pieces which is more meaningful and rejoined will cover the entire image. This paper delivers an outline on most common segmentation techniques and exposes thresholding as the simplest technique for segmentation. The threshold value is retrieved from the edge detected image. So the edge detections are accurate, then automatically, the threshold too. Whenever the gradient is high, the gray level points are then added to threshold surface for segmentation. Hence because of this drawback, this technique is not applied to complex images. The authors summarize the various segmentation techniques and comparing to other methods, found that thresholding is the simplest one and computationally fast [21].

Image segmentation is an important step which provides essential information such as relative size, shape and orientation of blood vessels. This paper projects the increasing number of attempts at venipuncture due to the difficulty in vein localization. The authors proposed an algorithm to emphasize the features of contextually related regions including vessel size and shape. The number of local extrema is decreased and single global minimum for each vessel is obtained by the Conditional Rule Generation (CRG) method [23]. Also this CRG algorithm will be a potential method to extract vessel information for automated venipuncture systems.

By classifying image patches at different resolutions and pooling multi-channel feature information at segments, hierarchical cascade of information propagation is generated. This paper proposes a fully automated bottom-up approach for pancreas segmentation in abdominal Computed Tomography (CT) scans [1]. By using efficient supervised edge learning techniques, the strength of semantic object level boundary curves may be utilized artificially and low image boundary contrast issue in super pixel generation present in medical imaging could be solved.

This paper presents polar dynamic programming to outline complex shapes [6]. The size of the object for correct boundary delineation need not be constrained with the introduction of polar variance image. The already available implementation of polar dynamic programming cannot accomplish this task. The algorithm presented by the authors' segment high curvature objects along with low-gradient objects.

The deep Convolutional Neural Networks (CNN) based depth estimation methods are used in this paper and expose the performance of object detection and semantic segmentation can be improved by adding an explicit depth estimation process. The authors combined the task of depth estimation with object detection and semantic segmentation and propose two ways of exploiting depth information [22]. The depths from RGB image are separately estimated and adding them as a cue for detection improves the performation of the segmentation. It is proved that the performance can be improved by exploiting related data which does not share the same set of labels.

Information retrieval is the process of extracting required data from the database based on the input from the user. The algorithm which is used for this purpose is known as pattern matching algorithm. The main aim of this paper is information retrieval from desktop using string matching algorithm. Various pattern inputs such as single word, multiple word or file are used to check the accuracy of this algorithm. The authors conclude that the enhanced Knuth-Morris-Pratt (KMP) algorithm gives better accuracy than the already existing algorithms based on performance measures [9].

The main rule of data mining is to retrieve meaningful pattern from huge volume of data. In order to achieve this, finding out frequent patterns from a database is very essential. The main aim of this paper is to compare the performance of many such algorithms and evaluate them. The authors have improved the already existing Apriori algorithm with effective hash-based algorithm for the candidate item set generation which is more effective than Apriori algorithm [16].

Upon using various inputs such as noisy/denoised samples, the pattern matching algorithm performs much better than the training based in a significant manner. A comparative study of three different categories of recognition algorithms, the bootstrap aggregating tree classifiers and median filtering for high intensity noise gives the high performance. The main difference between holes filling and missing data is that the depth values are available for hole whereas they are not available for missing data [14]. Missing data generally occurs with 3D imaging system because of the reasons like self-occlusion which appears after post correction, large depth variation or imaging device inaccuracy.

Various problems arising due to more complicated patterns like trees, regular expressions, graphs, arrays and point sets use the algorithm of combinatorial pattern matching. A pattern algorithm plays an important role in finding the appropriate content in minimum time. The best algorithm can be found by applying various algorithms in various applications. This paper concludes that the KMP algorithm has less time complexity and Boyer Moore (BM) algorithm and Boyer Moore Horspool (BMH) algorithm has less processing time complexity [15]. A comparative study of the string matching algorithm has been done based on the execution time of the algorithms. In this paper, the authors researched about the efficiency of different matching algorithms. This paper concludes that, the fastest and easy to implement algorithm is BMH algorithm whereas Rabin Karp (RK) algorithm is the slowest when increasing pattern length

and pattern placement [7]. Whenever pattern is placed at the end of the target, KMP algorithm can be applied. One of the most important applications in Bioinformatics is Deoxyribo Nucleic Acid (DNA) sequence detection. The mechanism which helps to find out the exact location of a specified pattern is pattern matching. To get the expected result at the cost of sufficient time, well established pattern matching algorithm is needed.

Maximum usage of sequence information is applied in Bioinformatics analysis. The algorithm specified in this paper looks for the specified pattern in a DNA sequence and produces the expected result [3].

III. COMPARISON OF VARIOUS IMAGE PROCESSING TECHNIQUES

Table 1: Comparison of Various Image Processing Techniques.

Image Processing Methods	Author and Year	Technique / Algorithm	Applications / Advantages
Image Denoising	Azam Karami (2017)	Denoising algorithm	Achieves the maximum PSNR and preserves the edges.
	K.Kanagalakshmi (2011)	Comparative evaluation of Adaptive Filtering, Median Filtering and Adaptive Wiener Filtering	Median filtering is the best with less computational time.
	Madhura.J (2017)	Noise removal methodologies	Median filter is used for reducing impulse noise.
	Anna Fabijanska (2007)	Noise reduction algorithm	Useful in case of noisy images presenting different details
	Asim Altaf shah (2016)	PSNR and MSE	Wavelet denoising filter performed well on OCT images
	Jianwei Zhang (2010)	Singular entropy technique	Essential to extract the accurate signal feature.
Image Enhancement	Rajlaxmi Chouhan (2012)	Adaptive DWT technique	Provide better enhancement for very dark images.
	Sukhjinder Singh (2012)	Image contrast algorithms	Brings out the hidden details in an input image.
	Hashem.S (2010)	Contrast enhancement based on Genetic Algorithm	Provides natural looking images.
	Yeong-Taeg Kim (1997)	Histogram Equalization	Enhances the contrast and preserves the image with proper brightness.
	Swati Khidse (2014)	Image fusion methods and DWT Technique	DWT based fusion techniques provide good quality fused images.
Image Segmentation	Karthick.S (2014)	Region merging algorithm	This algorithm is very constant with respect to noise.
	Yogamangalam.R (2013)	Markov Random Field	MRF is the strongest method of noise cancellation and thresholding is the simplest technique for segmentation.
	Yuhe Li (2017)	Vessel segmentation and Vein localization	The CRG model decreases the number of local extrema and obtains a single global minimum for each vessel.
	Amal Farag (2016)	Pancreas segmentation in abdominal computed tomography	Enhances the strength of semantic object level boundary in 2D or surfaces in 3D.
	Christos G.Bampis (2016)	Polar dynamic programming	Segments high curvature objects.
	Yuanzhouhan Cao (2016)	Object detection and semantic segmentation	Improves the performance by exploiting the related data.
Pattern Recognition and Matching	Janani.R (2016)	Enhanced KMP algorithm	Gives better accuracy.
	Paresh Tanna (2013)	Pattern mining algorithms	Efficient for two phase transaction database pruning.
	Mhryar Emambaksh (2015)	Recognition algorithms	Significantly outperform the training-based methods.
	Nimisha Singla (2012)	String matching algorithms	KMP algorithm has less time complexity.
	DU Vidanagama (2015)	String matching algorithms	KMP algorithm serves faster.
	Ashish Prosad Gope (2014)	Novel pattern matching algorithm	DNA sequence detection.

IV. CONCLUSION

In the survey of the various methods applied and compared with input images, wavelet filter serves its purpose and clears the noise. Also the methods applied in image enhancement, Histogram Equalization helps the image in adapting medical purposes. CRG algorithm used in image segmentation helps the images to localize the search. In pattern matching, KMP algorithm fine tunes the process and provides the best result. After applying all these methods in images, the missing data in images will be found accurately.

REFERENCES

[1]. Amal Farag, LE Lu, Holger R.Roth, JIamin Liu, Ecrim Turkbey, Ronald M.Summers, "A Bottom-Up Approach for Pancreas Segmentation using Cascaded Superpixels and (Deep) Image Patch Labeling", IEEE, Vol. **26**, No.1, Jan 2017.

[2]. Anna Fabijanska, Dominik Sankowski, "Image Noise Removal – The New Approach", CADSM 2007.

[3]. Ashish Prosad Gope, Rabi Narayan Behera, "A Novel Pattern Matching Algorithm in Genome Sequence Analysis", *International Journal of Computer Science and Information Technologies*, Vol. **5**, No.4, 2014.

[4]. Asim Altaf Shah, M. Mohasin Malik, M.Usman Akram, Shafaat A. Bazaz, "Comparison of Noise Removal Algorithms on Optical Coherence Tomography (OCT) Image", *IEEE, Published in Imaging Systems and Techniques, 2016 IEEE International Conference on* Nov 2016.

[5]. Azam Karami, Laleh Tafakori, "Image Denoising Using Generalised Cauchy Filter", *IET Journals,* Vol. **11**, No.9, Sep 2017.

[6]. Christos G.Bampis, Petros Maragos, Alan C.Bovik, "Graph-Driven Fiffusion and Random Walk Schemes for Image Segmentation", *IEEE*, Vol. **26**, No.1, Oct 2016.

[7]. DU Vidanagama, "A Comparitive Analysis of Various String Matching Algorithms", Proceedings of 8th International Research Conference, Nov 2015.

[8]. S. Hashemi, S. Kiani, N. Nooroozi and M.E. Moghaddam, "An Image Contrast enhancement method based on genetic algorithm", *Pattern Recognition Letters,* Vol. **31**, No.13, 2010.

[9]. R. Janani and S.Vijayarani, "An Efficient Text Pattern Matching Algorithm for Retrieving Information from Desktop", *Indian Journal of Science and Technology*, Vol. **9**, No.43, Nov 2016.

[10]. Jianwei Zhang, Shangwei Liu, "Study on the Comparison of Several Denoising Theories and Effects About Vibration Signal", *IEEE, Published in Computational Intelligence and Design (ISCID), 2010 International Symposium on Oct 2010.*

[11]. K. Kanagalakshmi, E. Chandra, "Performance Evaluation Of Filters In Noise Removal Of Fingerprint Image", *IEEE, Published in Electronics Computer Technology (ICECT), 2011 3rd International Conference on 2011.*

[12]. S. Karthik, K. Sathiyasekar, "A Survey Based on Region Based Segmentation", *IJETT*, Vol. **7**, No.3, January 2014.

[13]. J. Madhura, D.R. Ramesh Babu, "A Survey on Noise Reduction Techniques for Lung Cancer Detection", *IEEE, Published in Innovative Mechanisms for Industry Applications (ICIMIA), 2017 International Conference on Feb 2017.*

[14]. Mehryar Emambakhsh, Jiangning Gao, Adrian Evans, "Noise Modelling for Denoising and Three-Dimensional Face Recognition Algorithms Performance Evaluation", *IEEE,* Vol. **9**, No.5, Sep 2015.

[15]. Nimisha Singla, Deepak Garg, "String Matching Algorithms and Their Applicability in Various Applications", *International Journal of Soft Computing and Engineering*, Vol. **1**, No.6, Jan 2012.

[16]. Paresh Tanna, Yogesh Ghodasara, "Foundation for Frequent Pattern Mining Algorithms' Implementation", International Journal of Computer Trends and Technology, Vol. **4**, No.7, July 2013.

[17]. Rajesh Garg et al., "Histogram Equalization Techniques for Image Enhancement", *International Journal of Electronics & Communication Technology*, Vol. **2**, No.1, March 2011.

[18]. Rajlaxmi Chouhan, C.Pradeep Kumar, Rawnak Kumar and Rajib Kumar Jha, " Contrast Enhancement of Dark Images using Stochastic Resonance in Wavalet Domain", *International Journal of Machine Learning and Computing*", Vol. **2**, No.5, Oct 2012.

[19]. Swati Khidse, Meghana Nagori, "Implementation and comparison of Image Enhancement techniques", *International Journal of Computer Applications,* Vol. **96**, No.4, June 2014.

[20]. Yeong-Taeg Kim, "Contrast Enhancement using Brightness Preservation Bi-histogram Equalization", *IEEE Transaction on Communication, Networking and Broadcasting*, Pages:1-8, 1997.

[21]. R.Yogamangalam, B.Karthikeyan, "Segmentation Techniques Comparison in Image Processing", *International Journal of Engineering and Technology,* Vol. **5**, No.1, March 2013.

[22]. Yuanzhouhan Cao, Chunhua Shen, Heng Tao Shen, "Exploiting Depth from Single Monocular Images for Object Detection and Semantic Segmentation", *IEEE,* Vol. **26**, No.2, Oct 2016.

[23]. Yuhe Li, Zhendong Qiao, Shaoqin Zhang, Zhenhuan Wu, Xueqin Mao, Jiahua Kou, and Hong Qi, "A Novel Method for Low-contrast and High – Noise Vessel Segmentation and Location in Venipuncture", *IEEE*, Vol. **36**, No.11, July2017.

International Journal of Theoretical & Applied Sciences,

ISSN No. (Print): 0975-1718
ISSN No. (Online): 2249-3247

INSTRUCTIONS TO AUTHORS

The journal invites full and mini-review research articles, full length research articles and brief communication giving exciting current information of all Applied Sciences, Pharmaceutical Sciences.

Title and authors names and address must be given in the front page. This must be followed by the abstract on separate page. Thereafter, material and method, results and discussion, acknowledgements, references. Tables, photographs and all figures including drawings, graphics, and diagrams must be attached after the references. Five type of manuscript may be submitted like, **Original Papers, Rapid communication, Technical note, Review, Advancements in instrumentation.** All manuscripts should be typed double-spaced and in 10 pt Time New Roman including, references, tables, figure). Manuscripts preferred for publication contain original work, focused on the core aims and scope of journal, clearly and correctly written and should be written clearly and grammatically. The manuscript must be accompanied by a cover letter. It should have 50-100 words that contains Significance of work, Novelty of the work, Contribution to field/community.

Paper elements

Title page with: Names of authors with address and Personal e-mail addresses and mobile numbers

Abstract, Keywords, Introduction, Material and Methods, Results and Discussion, **Acknowledgments, Reference lists**,

Tables, Figure captions Figures should be adjust in text where they needed.

Title of article – Title should be concise and informative describing the contents of pages.

Author name and affiliation – Author(s) name should be consists of first and middle name initial followed by full last name. In case with more than one author clearly indicate who is willing to handle correspondence concerning the paper i.e. correspondence author and use asterisk to indicate the corresponding author. Use the symbol *, **,***, **** etc as superscripts to relate the authors to corresponding addresses. Provide the e-mail address, full postal address, mobile numbers of each author.

Abstract – The abstract should present a brief summary of the paper including questions being addressed and the key findings of the study. It should not serve as an introduction nor contain references. It consist of one paragraph not exceed 200 words.

Keywords – It provide immediately after the abstract and between 3-10 keywords. Keywords assist readers for indexing purposes.

Introduction – It include the scientific importance, historical background relevance to other area and objectives of the paper.

Material and Methods – It should be written in sufficient detail to enable others to repeat the author(s) work.

Results and Discussion - It may be combined or kept separate and may be further divided into subsections. This section should not contain technical details.

Acknowledgments – Acknowledgments of people, grants, funds etc. should be placed in a separate section before the reference list. The names of funding organizations should be written in full.

Tables and Figure – Tables and figures should not be embedded in the text, but should be included as a separate sheets or files. A short descriptive title should appear above each table with a clear legend and any footnotes suitably identified below. Avoid vertical rules. They are numbered consecutively in accordance with their appearance in the text. Number the tables as Table 1, Table 2 etc, to in the text. Figures should be completely labeled, taking into account necessary size reduction.

References - Within the text, references should appear as consecutive numbers in brackets (e.g., [1], [1,2,3]). The list of references should be given in the order of the first appearance of references in the text. The list of references should be formatted as follows:

1. For Articles in Journals: Indicate the initials and surnames of the authors, the title of the journal in italic, the volume number, the number of the first page, the year of the reference (in parentheses), for example,

M. Smith, G. Gaur, and I. Mehta, *Laser Phys.,* **1**, 123 (1991).

2. For Books: The initials and surnames of the authors, the full title of the book in italic, and (in parentheses) the publisher, the city, and the year, for example,

S.S. Mishra, S. Morgan, and P. Charlton, *Quantum Mechanics* (Allen, New York, 1990).

3. Conference Proceedings: Please add all available data such as title, date, and place of the conference as well as publisher, place, and year of publication or, alternatively, for example,

J. Ansell, I. Harrison, and C. T. Foxon: Proceedings of the 4th International Conference on Chemistry, Colorado, USA, 2001, Part A (Wiley-VCH, Berlin, 2002), pp. 279-282.

4. References to Online Material: Should include a brief description and or title:

http://www.kzoo.edu/ajp/docs/information.html.

5. Reference to a Thesis: Author initial, surname, DSc/PhD/MSc/BSc thesis, university, town, country and year of publication in bracket.

A.J. Agutter, Ph.D thesis, Edinburgh University (Edinburgh, UK, 1995).

Equations - Each equation should appear on a separate line with proper punctuation placed before and after it. All equations must be numbered sequentially. The number of the equation, in parentheses, should be placed near the right-hand margin. Avoid bars either above or below letters. Avoid subscripts on subscripts, etc. Adequate space must be allowed for marking of inferior and superior letters or numbers. Crowded equations lead to errors in composition. Use the following format to refer to equations in the text: Equation (5) follows from substituting Eqs. (2) and (3) into Eq. (4).

Chemical Reaction Data - *(Relevant for only Chemical and Biochemical Fields*

For heterogeneous catalysis, presentation should include reaction rates normalized by catalyst surface area, surface area of the active phase, or number of active surface atoms or catalytic sites, as appropriate. Typical rate units are mol s^{-1} m^{-2} or, in the case of surface atom normalization to produce turnover frequencies, s^{-1}. For homogeneous catalysis, rates should typically be reported as turnover frequencies. Comparisons of selectivities should be made at similar conversions. Catalytic measurements need to be carried out under kinetically limited conditions. Confirming tests need to be carried out and reported, especially for all reactions occurring in the liquid phase.

Symbols and Units- Greek symbols and special characters often undergo formatting changes and get corrupted or lost during preparation of a manuscript for publication.

To ensure that all special characters used are embedded in the text, these special characters should be inserted as a symbol but should not be a result of any format styling (*Symbol* font face) otherwise they will be lost during conversion to PDF/XML^2 . Authors are encouraged to use SI units, but use of SI units is not mandatory if other units are more appropriate.

Nomenclature - Nomenclature should confirm to current American usage. Insofar as possible, authors should use systematic names similar to those used by Chemical Abstracts Service or IUPAC.

External review - The external review process is initiated when the editor sends the manuscript out for review. When the reports are returned, the Editor makes a decision based on the recommendations received and the number of previous revisions and informs the submitter in a decision letter.

Decisions on initial submissions

a. If the reviews are negative or insufficiently strong to support continued editorial consideration, the manuscript will be rejected.
b. In other cases, including cases where the reviews are mixed, the manuscript will be returned for revision with suggestions and directions for resubmission.
c. With the resubmission, authors must include a cover letter that summarizes the revisions and provides responses to the issues and questions raised by the editor and/or the reviewers. Upon resubmission the manuscript will usually be sent back to the previous reviewers and occasionally to new reviewers for re-review.

Revised manuscript - A revised manuscript should be returned within 6 days for minor changes and 10 days for major revisions. If the manuscript is not returned within this time frame, it will be considered withdrawn by the author and any revised version submitted subsequently will be considered a new contribution.

After acceptance

Copyright transfer – Authors will be asked to transfer copyright of the article to the publisher. This will ensure the widest protection and dissemination of information under copyright laws.

Offprint – Additional offprint can be ordered by the corresponding authors.

Proof reading – The corresponding author will receive a proof and should be return to publisher with in three days of receipt. Correction should be restricted to typesetting error; any other correction may be checked and corrected since the inclusion of late correction cannot be accepted.