International Journal on Emerging Technologies (Special Issue NCETST-2017) **8**(1): 1-4(2017)
(Published by Research Trend, Website: www.researchtrend.net)

ISSN No. (Print) : 0975-8364
ISSN No. (Online) : 2249-3255

Study for Social Media Mining Methods with Soft Computing

Farah Shan[1] and Dr. M.K. Sharma[2]
[1]*Research Scholar, Uttarakhand Technical University, Dehradun, (Uttarakhand), India*
[2]*Associate Professor, Amrapali Institute Haldwani, (Uttarakhand), India*

ABSTRACT: Social media sites are now very popular medium for showing your views and opinions to others with a great amount of various types of information uploaded by the social media users, a social web page can be a collection of pages, audio files, photographs, images, video files and other forms of data in structured or unstructured form. It is also huge, diverse, and dynamic, hence raises the scalability. The primary aim of web mining is to extract useful information and knowledge from web. Web mining is a part of data mining which relates to various research communities such as information retrieval, database management systems and Artificial intelligence. Soft computing methods may faster the process of web mining on social media sites, this paper attempts to analyze benefits and drawbacks of the various types of soft computing methods like ANN, ACO, GA and Fuzzy set for mining the social media data sets.

Keywords : Soft computing , social media, ANN, ACO, GA, Fuzzy set

I. INTRODUCTION

Web Mining on social media users and groups may help to collect knowledge from the group structure, hyperlink references, opinion graphs and most liked information. Web mining is a way to discover useful knowledge from the web log data obtained from some open source websites that are available on the web [7]. Web usage mining has become very critical for effective web site management, creating adaptive web sites, business and support services, personalization and network traffic flow analysis [6][8][9].

Web Usage Mining techniques can be used to anticipate the user behavior in real time by comparing the current navigation pattern with typical patterns which were extracted from past Web log. Recommendation systems could be developed to recommend interesting links to products which could be interesting to users.

Recently various soft computing methodologies have been applied to handle the different challenges posed by data mining. The main constituents of soft computing, at this juncture, include fuzzy logic, artificial neural networks, genetic algorithms, and ACO. Each of them contributes a distinct methodology for addressing problems in its domain. This is done in a cooperative, rather than a competitive, manner. The result is a more intelligent and robust system providing a human-interpretable, low cost, approximate solution, as compared to traditional techniques.

this paper attempts to analyze benefits and drawbacks of the various types of soft computing methods like ANN, ACO , GA and Fuzzy set for mining the social media data sets.

II. RESEACRH OBJECTIVES

Presently modern soft computing tools include fuzzy sets, artificial neural networks (ANNs), genetic algorithms (GAs), Ant colony Optimization (ACO). Fuzzy sets provide a natural framework for the process in dealing with uncertainty. Artificial Neural Networks (ANNs) are widely used for modeling complex functions, and provide learning and generalization capabilities. GAs and ACO algorithms are used in efficient search and optimization the search space.

The objective of this paper is to study about web mining, its various types like clustering, classifications, and to give a perspective to the research community about the potential of applying soft computing techniques to its different schemes . This paper has reviewed the existing techniques, methods, algorithms with their benefits and limitations, this paper lays emphasis on possible enhancements of these methods using soft computing framework. In this regard, the relevance of fuzzy logic (FL), ANNs, GAs, and ACO presented with some examples, along with the mention of some commercially available social media sites.

III. SOFT COMPUTING FOR WEB MINING

Soft computing is a consortium of methodologies which work with real life problems and provides in one form or another flexible information processing capabilities for handling real-life and complex situations. Its aim is to exploit the tolerance for imprecision, uncertainty,

approximate reasoning, and partial truth in order to achieve tractability, robustness, low-cost solutions, and close resemblance to human-like decision making [10]. In other words, it provides the foundation for the conception and design of high machine IQ (MIQ) systems, and, therefore, forms the basis of future generation computing systems. At this juncture, FL, ACO, ANNs, and GAs are the principal components, where FL provides algorithms for dealing with imprecision and uncertainty arising from vagueness rather than randomness, ACO for handling uncertainty and optimize results in case for real objects, ANN the machinery for learning and adaptation, and GA for optimization and searching. Relevance of soft computing to pattern recognition and image processing is extensively established in the literature [13] [11]. Recently, the application of soft computing to data mining problems has also drawn the attention of researchers. A recent review [12] is a testimony in this regard. Here, FL is used for handling issues related to incomplete/imprecise data/query, approximate solution, human interaction (linguistic information), understandability of patterns and deduction, and mixed media information (fusion). NNs are used for modeling highly nonlinear decision boundaries, generalization and learning (adaptivity), self organization, rule generation, and pattern discovery. GAs are seen to be useful for prediction and description, efficient search, and adaptive and evolutionary optimization of complex objective functions in dynamic environments.

IV. WEB USAGE MINING THROUGH ANN

The aim of Web usage mining is to discover patterns of user activities in order to better serve the needs of the users for example by dynamic link handling, by page recommendation etc. The aim of a Web site or Web portal is to supply the user the information which is useful for him. There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Web usage mining consists of four main steps:

A. Data collection
B. Preprocessing,
C. Pattern discovery
D. Pattern analysis

Different types of user data can be collected using these methods, for example (I) characteristics of the user (age, gender, etc.), (ii) preferences and interests of the user, (iii) user's behavior pattern. The aim of clustering web pages is to have groups of pages that have similar content.

Neural based approach has shown that the usage trend analysis very much depends on the performance of the clustering of the number of requests. In this test we used Self Organizing Map , which is a kind of neural network, in the process of Web Usage Mining to detect user's patterns. We are going to analyze the traditional K-Means algorithm result with comparison to SOM.

Analysis of the clusters formed by both algorithms considering above conditions with respect to percentage of occurrence of each unique url in corresponding clusters. This paper; we can see the comparison between both methods, SOM and K-Means, in gcoea.ac.in sites. For this site it has been develop a complete process involved in Web Usages Mining. This website contains information about college, student, department , staff etc. .The log file consist of one month data total 71,238 lines and 2045 unique IPs. After extracting and cleaning the information from the complicated log file, logs are divided into 30 min sessions. Total 2621 sessions, which are further represented by vector to provide input to the algorithms. The Urls will be represented by weight of url in that session. The weights will be the frequency of occurrence of url in the sessions. Analyzing the outcome of the algorithms, we can conclude that with respect K-Means we can cover more Urls but SOM works better for larger number of cases. With increase in data, learning process of SOM becomes more accurate and we can consider larger number of clusters.

V. USE OF FUZZY LOGIC IN WEB MINING

Social media Web sites personalization is the procedure of modifying the content and structure of a web site to the precise requirements of each user taking benefit of the user's directional behavior. The phases of the web personalization comprises of: 1) the collection of web data, 2) the preprocessing phase of these data, 3) the analysis of the collected data and 4) the purpose of the actions that should be performed. In the test, the log files are collected from the proxy server log. The gathered data are undergoing a preprocessing phase to remove the unwanted and noisy information. The web directories are discovered based on the user and session clustering. For grouping the user and session, the Neuro Fuzzy Clustering Approach (NFCA) is applied.

Clustering the User and Session using Neuro Fuzzy Clustering Approach (NFCA): The user is identified by the IP address of the respective user. Both the user and session information are retrieved from the log file. The user and session details are clustered using the fuzzy clustering algorithm. The clustering strategy starts by partitioning the data set into a large number of small clusters. In this approach, Advanced Apriori algorithm is utilized to analyze the sequential patterns.

Coverage is defined as the number of target web pages that are covered by the session based directories. User gain is the estimated actual gain that a user follows the interested web directories instead of accessing the

preliminary web directories to get the preferred web page. For this test we implemented Neural network and Fuzzy clustering to group the user sessions and we tested it with Advanced Apriori algorithm to analyze the frequent pattern mining. The proposed method removes the directories which are all having the threshold value less than the fixed threshold. The proposed approach provides better coverage and user gain against threshold.

VI. ANT COLONY OPTIMIZATION (ACO) IN WEB MINING

is one of the tools used for these purposes. This metaheuristic is inspired by the way ants optimize their trails for food foraging based on releasing chemical substances into the environment called pheromones. This simple idea is applied to the web user trails of visited web pages, also called sessions (Liu, 2007). Artificial ants are trained through a web session clustering method modifying an intrinsic text preference vector which represents the importance given by the users to the set of most important keywords. Furthermore, trained ants are used to predict future browsing behavior.

Deneubourg *et al.* in [7] proposed ant-based clustering and sorting. In the case of ant-based clustering and sorting, two related types of natural ant behaviors are modeled. When clustering, ants gather items to form heaps. And when sorting, ants discriminate between different kinds of items and spatially arrange them according to their properties [8]. Lumer and Faieta [9]. in proposed ant-based data clustering algorithm, which resembles the ant behavior described in [7]. The agents (ants) and data are randomly initialized on a toroidal grid. By moving agents, data is sorted according to its neighbors.

The proposed test based on an algorithm for retrieval of requested data from the users in more efficient ways. In this, the algorithm is develop on the behavior of the real ants. In this an artificial ant colony is developed and concerned over which the assumptions are formed and the implementation of the system is performed on various assumptions.

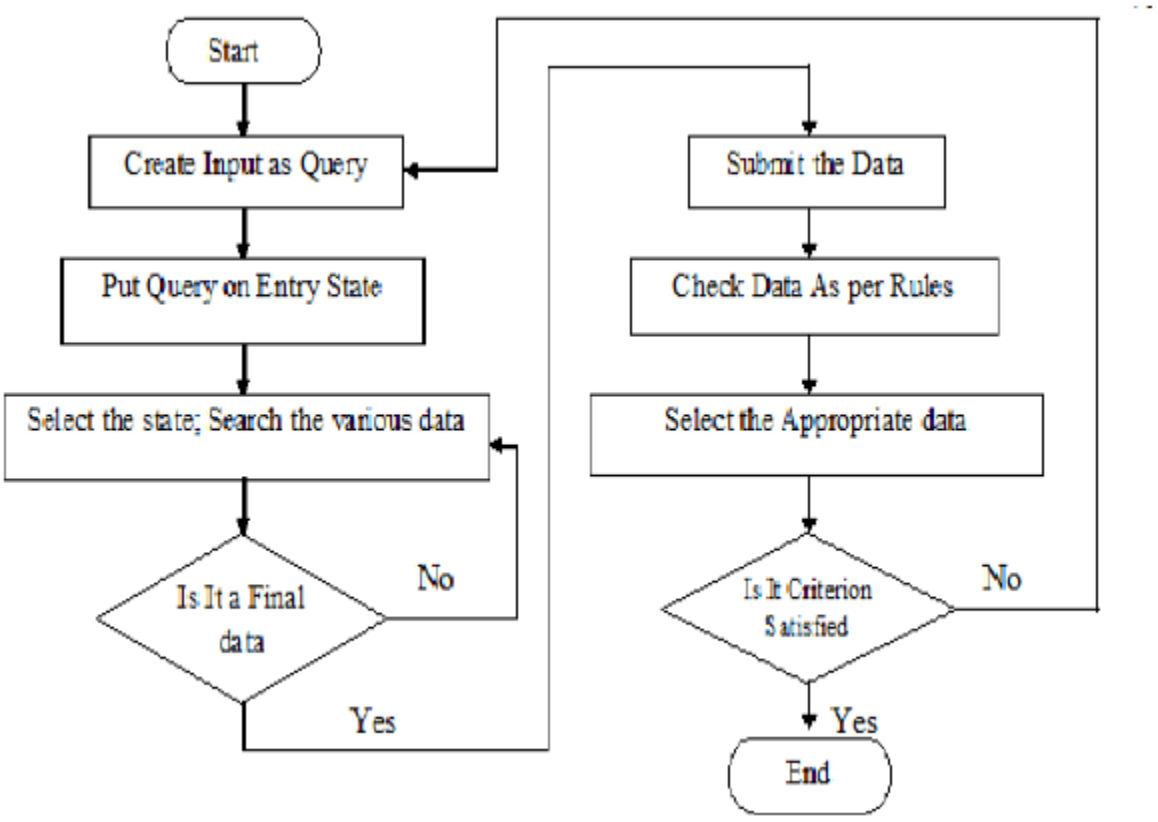

Fig. 1. Proposed ant-based clustering and sorting flowchart.

VII. CONCLUSION

Fast and hybrid social network analysis techniques are needed to mine opinion scores on social networks as a grouping on wrong opinion may create problems for a society or country . Social Network Analysis (SNA) can be used as an important tool for researchers , as the number of users and groups increasing day by day on that social sites, and a large group may influence others, but the necessary information is often distributed and hidden on social site servers, so there is

a need to design some new approaches for collection and analysis the social web data. Soft computing methodologies, involving fuzzy sets, neural networks, genetic algorithms, rough sets, and their hybridizations, have recently been used to solve data mining problems. They strive to provide approximate solutions at low cost, thereby speeding up the process. A categorization has been provided based on the different soft computing tools and their hybridizations used, the mining function implemented, and the preference criterion selected by the model. Neuro-fuzzy hybridization exploits the characteristics of both neural networks and fuzzy sets in generating natural rules, handling imprecise and mixed mode data, and modeling highly nonlinear decision boundaries. Domain knowledge, in natural form, can be encoded in the network for improved performance.

REFERENCES

[1] O.A. Mohamed Jafarand R. Sivakumar, "Ant-based Clustering Algorithms: A Brief Survey", *International Journal of Computer Theory and Engineering,* Vol. **2**, No. 5, October, 2010 1793-8201.

[2] Richa Gupta, "Web Mining using Artificial Ant Colonies: A Survey", *International Journal of Computer Trends and Technology* (IJCTT) – volume **10** number 1 – Apr 2014.

[3] J. Deneubourg -L., S. Goss, N. Franks, A. Sendova-Franks, C. Detrain, L. Chrétien, The dynamics of collective sorting: robot-like ants and ant-like robots. Proceeding of the first international conference on simulation of adaptive behavior, pp. 356–365, MIT Press, 2011.

[4] J. Handl, B. Meyer, Ant-based and Swarm-based clustering, Swarm Intelligence, **1**, pp. 95–113, 2007.

[5] E. Lumer, B. Faieta, Diversity and adaptation in populations of clustering ants. Proceeding of the third international conference on simulation of adaptive behavior, pp. 501–508, MIT Press, 2014.

[6] E.H. Chi, A. Rosien, and J. Heer, "LumberJack: Intelligent Discovery and Analysis of Web User Traffic Composition", In Proceedings of ACMSIGKDD Workshop on Web Mining for Usage Patterns and User Profiles, Canada, ACM Press, 2012.

[7] R. Cooley, "Web Usage Mining: Discovery and Application of Interesting patterns from Web Data", Ph. D. Thesis, University of Minnesota, Department of Computer Science, 2010.

[8] J. Heer, and E.H. Chi, "Identification of Web User Traffic Composition using Multi- Modal Clustering and Information Scent", In Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining, pp. 51-58, 2001.

[9] S.E. Jespersen, J. Thorhauge, and T. Bach, "A Hybrid Approach to Web Usage Mining, Data Warehousing and Knowledge Discovery", in Y. Kambayashi, W. Winiwarter, M. Arikawa , eds., LNCS 2454, pp. 73-82, 2002.

[10] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing," Commun. AGM, vol. **37**, pp. 77–84, 1994.

[11] S. K. Pal, A. Ghosh, and M. K. Kundu, Eds., Soft Computing for Image Processing. Heidelberg, Germany: Physica-Verlag, 2000.

[12] S. K. Pal and S. Mitra, Neuro-Fuzzy Pattern Recognition: Methods in Soft Computing. New York: Wiley, 1999.

[13] S. Mitra, S. K. Pal, and P. Mitra, "Data mining in soft computing framework: A survey," *IEEE Trans. Neural Networks,* vol. **13,** pp. 3–14, Jan. 2001.