International Journal of Electrical, Electronics and Computer Engineering **5**(1): 75-81(2016)

ISSN No. (Online): 2277-2626

An Efficient K-Means Clustering based Annotation Search from Web Databases

Anshul Tiwari* and Kiran Agrawal**

**Department of Computer Science and Engineering,*
***Department of Information Technology,*

(Corresponding author: Anshul Tiwari)
(Received 04 May, 2016 Accepted 02 June, 2016)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT*: In our work an efficient methodology is implemented to improve the accuracy of search results from the web databases on various keywords such as movies, CD, books etc. The improvement of Alignment algorithm using K-means clustering is proposed for the searching of annotated results from the web databases. The technique implemented is for the proficient retrieval of text nodes and data units using K-means clustering which improves precision and recall as compared to the existing approach. The methodology implemented using K-means clustering and labeling of search records is compared with existing methodology implemented for the search records.

Keywords: K-means clustering, Data Alignment, Precision, Recall

I. INTRODUCTION

Increasingly, many data sources appear as databases which are open for all and hence various queries can be used for the hidden of data after which deep web is possible [1]. First of all unique database can give data records. Second, due to the dependencies present between the hidden databases, certain databases must be queried before others. Third, some database may not be available at certain times because of some major problems, and therefore, the query planning should be capable of dealing with unavailable databases and generating alternative plans when the optimal one is not feasible [2]. One of the key principles in modern software engineering is the separation of concerns. This concept described in [3], states that by separating the basic algorithm from special purpose concerns makes each of the parts easier to write, maintain and test. For example, the separation of data and layout of a webpage into a style-sheet [4] and the html content page has become an engineering practice: It not only allows changing the layout and content independently, but it also allows different people with different skill-sets and training to work on the webpage independently, e.g. a developer and a designer. To show the effectiveness of our methods, it is necessary to demonstrate their application to some real world scenarios or conduct a study using a range of schema or ontology matching tasks. To different matching systems correspond to different evaluations in the literature, with diverse methodologies, and data sets.

It makes difficult to compare their effectiveness with respect to the state of the art. In this methods are evaluated by adopting the quality measures proposed in [7], which have been used to assess the effectiveness of several relevant schema matchers.

For example, once a book comparison shopping system collects multiple result records from different book sites, it needs to determine whether any two SRRs refer [8] to the same book.

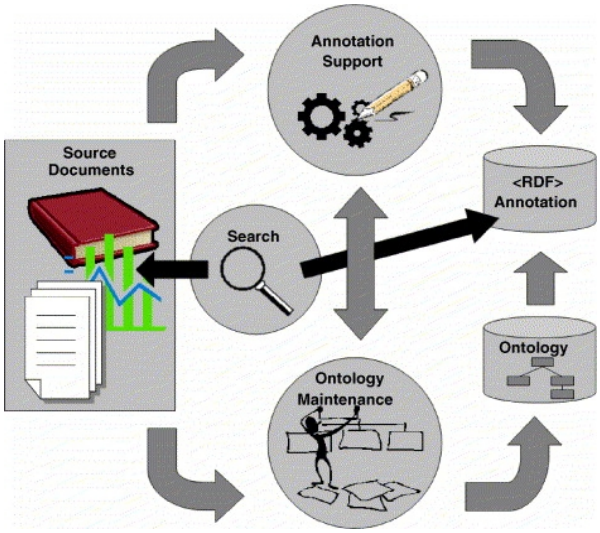

Fig. 1. Architecture of annotation based web search.

Web Databases. In Web Databases system stores information and can be accessed via an internet. Here in this paper web databases from online e-commerse website is taken with a number of categories such as movies and books etc. A Web Database contains a set of more than one tables with data and that can be communicated via a programming languugae over website. Here in the paper Web Databases are collected from various websites with its html code over various categories.

II. LITERATURE SURVEY

Marja-Riitta Koivunen, Eric Prud'Hommeaux proposed a new technique called Annotea: an open RDF infrastructure for shared Web annotations [9]. It is based on the RDF structure for the annotations extraction. The annotations are extracted on the basis of metadata. Efficient extraction of annotations such as from HTTP, Xlinks and Xpointer. Huge metadata needs to be maintained for better performance. Luis Gravano and Héctor García-Molina find a new way of search results from Text-Source Discovery over Internet [10]. Here in this paper text source discovery problem is removed that may occurs over internet. The technique removes the problem of text source discovery problem that may occur over internet. The technique works on query basis hence provides less performance for other sources. Khaled Khelif, Rose Dieng-Kuntz, Pascal Barbry. Proposed an Ontology-based Approach to Support Text Mining and Information Retrieval in the Bio logical Domain [11]. The technique includes ontology for all the text documents especially for biologists. Hence after using the technique ontology based semantic can be extracted from the sources. Efficient retrieval of information of biological domains. James Pustejovsky, Robert Gaizauskas, Graham Katz implemented Robust specification of event and temporal expressions in text [12]. Here in this paper the methodology uses the concept of TimeML which is used for the time based retrieval of text using natural language text sequence. Efficient extraction of text on spatial and temporal basis. The technique is not suitable for multi-lingual languages.

S. No.	Paper	Author	Technique Used	Advantages	Issues
1	Annotating Search Results from Web Databases [1].	Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng	Here in this paper an efficient technique is implemented for the searching of annotations in the web databases	Efficient retrieval of annotations and labeling of those annotations.	Provides less efficiency and is static.
2	Annotea: an open RDF infrastructure for shared Web annotations [13]	Marja-Riitta Koivunen, Eric Prud'Hommeaux,	It is based on the RDF structure for the annotations extraction. The annotations are extracted on the basis of metadata.	Efficient extraction of annotations such as from HTTP, Xlinks and Xpointer.	Huge metadata needs to be maintained for better performance.
3	An Ontology-based Approach to Support Text Mining and Information Retrieval in the Bio logical Domain [11]	Khaled Khelif , Rose Dieng-Kuntz, Pascal Barbry.	The technique includes ontology for all the text documents especially for biologists. Hence after using the technique ontology based semantic can be extracted from the sources.	Efficient retrieval of information of biological domains.	A different ontology needs to be maintained for different domains.
4	Robust specification of event and temporal expressions in text [14].	James Pustejovsky, Robert Gaizauskas, Graham Katz.	Here in this paper the methodology uses the concept of TimeML which is used for the time based retrieval of text using natural language text sequence.	Efficient extraction of text on spatial and temporal basis.	The technique is not suitable for multi-lingual languages.

S. No.	Paper	Author	Technique Used	Advantages	Issues
5.	An experiment using Conceptual Graph Structure for a Multilingual Information System [15]	Catherine Roussey, Sylvie Calabretto and Jean-Marie Pinon.	The technique is implemented for the extraction of information for multilingual languages. The technique allows indexing of documents and information retrieval for multilingual text documents.	Removes the problem of multilingual information retrieval and provides complex knowledge representation.	The methodology is applied for the collection of English articles.

III. PROPOSED METHODOLOGY

Web Pages are created and implemented using mark up languages such as HTML. Since almost all the things are available in web pages. Web sites such as mobiles contains a number of categories and sub categories which uses a huge databases to shows the pictures and prices and various other categories from the databases, but sometimes these web pages contains tags from which some meaningful information can be extracted. These HTML web pages contain tag structure where each of the tag structure contains either a tag node or can also contains text node. These HTML web pages contains a tag node is surrounded by decorative tag nodes such as '<' and '>' in the original HTML text source. The text that is written outside of the tag nodes are known as text nodes. The major difference between tag nodes and text nodes is that the tag nodes are not visible on the web pages while text nodes are visible on the web pages. These text nodes contain a number of data units which can be classified or detected using the major four types of relations.

The Proposed Methodology implemented here efficiently searches Annotations from the web databases using K-Means Clustering based Alignment Algorithm. The Existing methodology implemented for the Searching of Annotations from web databases can't work well for the Multiple Composite text nodes hence provides less accuracy of searching Annotations. Hence here in this paper the problem of composite text nodes can be solved using K-Means Clustering.

A. Data Alignment

The purpose of data alignment is to put the data units of the same concept into one group so that they can be annotated holistically.

B. Data Unit Similarity

Data content similarity (SimC). It is the Cosine similarity between the term frequency vectors of d1 and d2:

$$SimC(d1, d2) = \frac{V_{d1} * V_{d2}}{||V_{d1}|| ||V_{d2}||}$$

where ,Vd is the frequency vector of the terms inside data unit d, ||Vd|| is the length of Vd, and the numerator is the inner product of two vectors.

C. K-Means Clustering

K-means is an unsupervised learning algorithm which is used to classify data on the basis of clusters. It includes k number of centroids for each of the cluster. The main objective of the algorithm is the minimization of objective function which is given as:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where the parameter is the chosen distance between data point and the cluster center between n data points.

The algorithm consists of the following steps:

(i) Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.

(ii) Assign each object to the group that has the closest centroids.

(iii) When all objects have been assigned, recalculate the positions of the K centroids.

(iv) Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

The proposed methodology implemented here uses the following Alignment algorithm for the single or multiple text nodes and then uses K-Means based clustering to cluster the similar text nodes which provide same set of search results from the HTML tags.

The proposed algorithm that is implemented here consists of the following basis steps:

Step 1: Take an input dataset which contains a number of web pages for various categories such as Book and Movies and Pen Drives and Electronics, since these categories contains some basis labels such as title and author and price and ISBN for the category Book.

The Web pages taken here contains HTML tags and nods inside which some meaning information is stored.

Step 2: As soon as the input dataset is selected the next step is to extract the important features from the web pages. Here for the extraction of features for the web pages the following predefined classes are used which removes the decorative tags from the web pages and stores text nodes.

Step 3: After the extraction of features from each of the input web page dataset cosine similarity is measure for each of the text document web pages using the following formula=

```
VectAB=VectAB+(freq1*freq2);
VectA_Sq = VectA_Sq + freq1*freq1;
VectB_Sq = VectB_Sq + freq2*freq2;
sim_score = ((VectAB)/(Math.sqrt(VectA_Sq)*Math.sqrt(VectB_Sq )));
```

Step 4: The next step is the computation of Alignment of the text nodes and the data units that are available in the web pages. Here Alignment can be done using the clustering using K-Means. K-Means clustering is an efficient learning approach which takes 'X' as an input data values and 'Y' as the label values and mediods and centers.

Step 5: Finally on the basis of clustering of the labels of the text nodes value a weighted factor is decided and hence labels are assigned.

$$Sim\_Score(A,B) = \frac{V_{AB}}{\sqrt[2]{V_A} * \sqrt[2]{V_B}}$$

Where,

$$V_{AB} = V_{AB} + (F_A * F_B)$$

1. J ← 1
2. While true (means all the available text nodes from HTML tags provides labels)
3. For i ← 1 to number of search records
4. Gi ← SR[i][j]
5. If Gi contains empty labels
6. Exit
7. V ← Call_Cluster(G)
8. If |V| > 1
9. S=Call_Merge(SR[i][j])
10. V[c]=Call_Similar_Cluster(V,S)
11. Shifting of SR and V.

Call_Cluster(G)

1.Input G contains all the search records along with the labels for each text node in the search record.
2.Repeat till G[i][j]==null
3.For each A in G and B in G
4.Compute Similarity(A,B) in G using Sim(A,B)= precence(AUB)/precence(A)+precence(B)-precence(AnB)
J ← 1
5.While true (means all the available text nodes from HTML tags provides labels)
6.For i ← 1 to number of search records
7.Gi ← SR[i][j]
8.If Gi contains empty labels
9.Exit
10.V ← Call_Cluster(G)
11.If |V| > 1
13.S=Call_Merge(SR[i][j])
14.V[c]=Call_Similar_Cluster(V,S)
15.Shifting of SR and V.

Call_Cluster(G)

1. Input G contains all the search records along with the labels for each text node in the search record.
2. Repeat till G[i][j]==null
3. For each A in G and B in G
4. Compute Similarity(A,B) in G using Sim(A,B)= precence(AUB)/ precence(A)+precence(B)- precence(AnB)
5. Best ←Sim(A,B)
6. Remove text node from L
7. Remove text node from Right R
8. Add LUR to V
9. Return V;

Call_Merge(SR)

1. Repeat till SR ← empty
2. For i=1,j=1 ← length (G)
3. S[i][j] ← SR[i][j]
4. Return S;

Call_Similar_Cluster(V,S)

1. Repeat till V ←empty || S ←empty
2. Sim(V,S) = Sim(VUS) / Sim(V) +Sim(S) – Sim(VnS)
3. Check the minimum value for the cluster
4. V[c]=min (Sim(V,S)
5. Return V;

IV. RESULT ANALYSIS

The Result Analysis shows the performance of the proposed methodology. The proposed methodology shows higher precision and recall as well as has high Accuracy for the prediction of annotated search records from the web databases.

Table 1: Analysis of Existing Technique on WDB.

Domain	Precision	Recall	F-Score
Book	0.6	0.6	0.6
Pen Drive	0.468	0.732	0.565
Music	0.5	0.7	0.58
Movies	0.56	0.66	0.605098
Games	0.68	0.57	0.619238

The table shown below is the experimental analysis of the searching of annotations using SVM based clustering. The result is analyzed on various domains such as Book and Pen Drives and Music and Movies as well as Games. The result is analyzed on the basis of Precision and Recall and F-Score. Here Precision can be computed on the basis of correctly identified annotation to the total number of annotations fetched from the web databases. Recall is the computation of total number of annotations fetch from the web databases to the total number of annotated records present.

The figure shown below is the experimental analysis and comparison of Precision based on Existing and Proposed work. The result is analyzed on various domains such as Book and Pen Drives and Music and Movies as well as Games. Here Precision can be computed on the basis of correctly identified annotations to the total number of annotations fetched from the web databases.

Table 2: Analysis of Proposed Technique on WDB.

Domain	Precision	Recall	F-Score
Book	0.75	0.64	0.69
Pen Drive	0.925	0.648	0.76
Music	0.854	0.812	0.8324
Movies	0.923	0.843	0.881
Games	0.732	0.693	0.7119

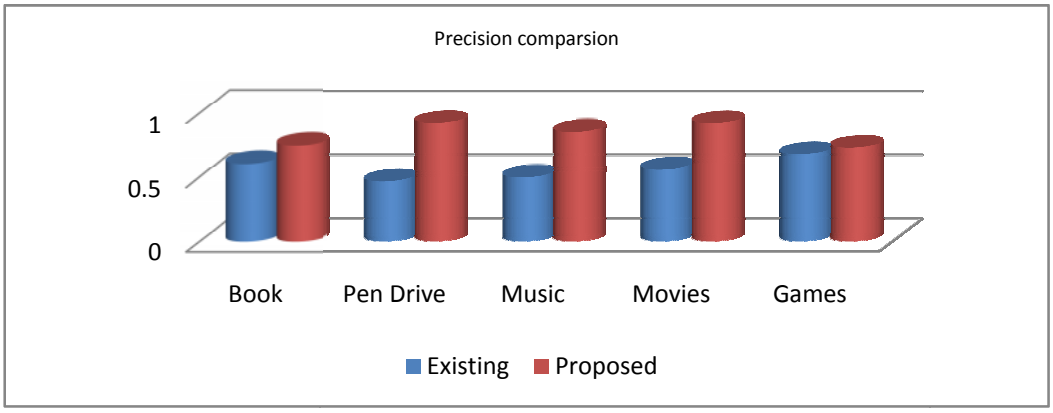

Fig. 2. Comparison Analysis of Precision on various domains.

The figure shown below is the experimental analysis and comparison of Recall based on Existing and Proposed work. The result is analyzed on various domains such as Book and Pen Drives and Music and Movies as well as Games.

Recall is the computation of total number of annotations fetch from the web databases to the total number of annotated records present.

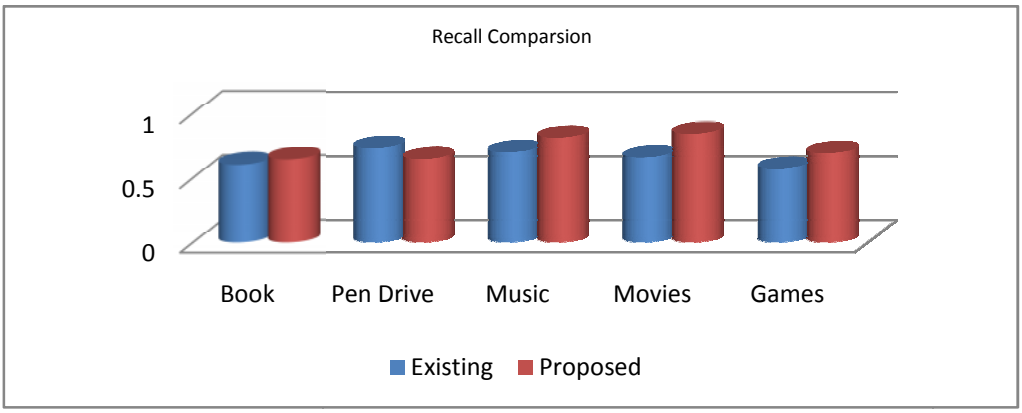

Fig. 3. Comparison Analysis of Recall on various domains.

The figure shown below is the experimental analysis and comparison of F-Score based on Existing and Proposed work. The result is analyzed on various domains such as Book and Pen Drives and Music and Movies as well as Games. It is defined as:

$$F - Score = \frac{2 * precision * recall}{precision + recall}$$

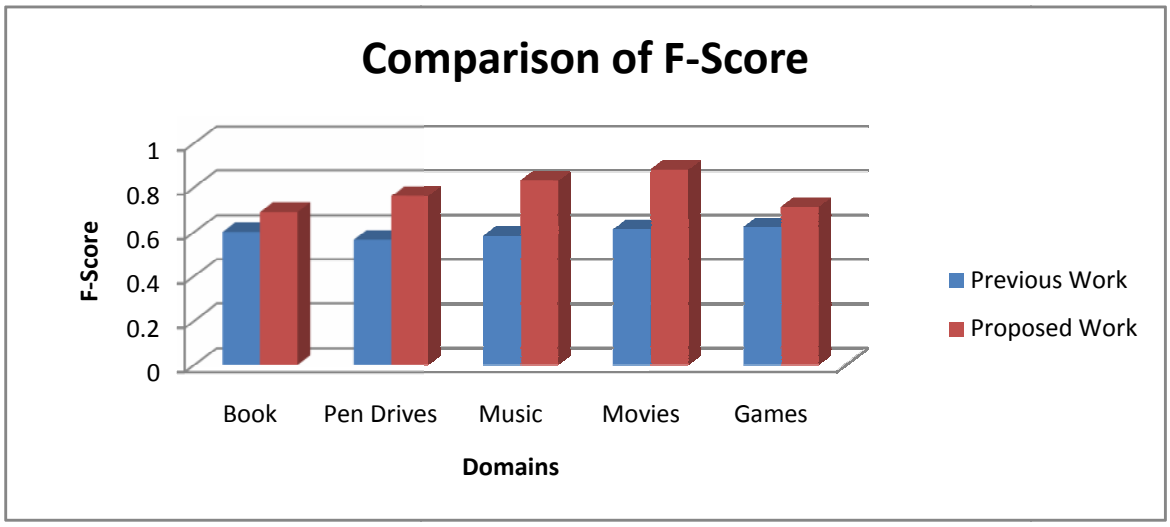

Fig. 4. Comparison Analysis of Final Score on various domains.

V. CONCLUSION

The comparison between existing and proposed techniques can be analyzed based on precision and recall and it was found that the proposed methodology not only removes the problems of the existing technique but also provides high precision and recall.

The proposed methodology implemented here for the searching of the annotations from the web databases. Here the annotations can be identified on the various categories such as Book, Movies, Electronics, Pen Drives, Auto.The proposed methodology is applied on the these categories with different web pages and hence on the basis of search web records labels are assigned to these web pages After identification of annotations in the web databases accuracy can be computed and compared to the existing technique that is implemented for the efficient search of records from the web databases and the proposed methodology provides high precision and recall as compared to the existing technique.

REFERENCES

[1]. Y. Lu, H. He, H. Zhao, W. Meng, C. Yu, "Annotating Search Result From Web databases" *In IEEE Transaction on Knowledge and Data Engineering*, Vol. **25**, No.3, 2013.

[2] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu, "Annotating Structured Data of the Deep Web," *Proc. IEEE 23rd Int'l Conf. Data Eng*. (ICDE), 2007.

[3]. W.L. Hürsch and C.V. Lopes, "Separation of Concerns". Technical Report NU-CCS-95-03, Northeastern University, 1995.

[4]. Cascading style sheets: www.w3.org/Style/CSS

[5]. A. Arsanjani et al, "S3: A Service-Oriented Reference Architecture". *IT Professional*, Vol. **9**, Issue 3, pp. 10-17, 2007.

[6]. T. Elrad, R.E. Filman and A. Bader, "Aspect-oriented programming: Introduction", *Communications of the ACM*, Vol. **44**, Issue 10, pp. 28-32, 2001.

[7]. Do, H. H., Melnik, S. and Rahm, E. Comparison of Schema MatchingEvaluations. In Chaudhri, A. B., Jeckle, M., Rahm, E., and Unland, R., editors, Web, Web-Services, and Database Systems, NODe 2002 Web and Database-Related Workshops, Erfurt, Germany, October 7-10, volume **2593** of Lecture Notes in Computer Science, pages 221–237. Springer, (2002).

[8]. Annotating Search Results from Web Databases Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng, Member, IEEE,and Clement Yu, Senior Member, *IEEE Transactions On Knowledge And Data Engineering*, VOL. **25**, NO. 3, MARCH 2013.

[9]. J. Kahan, M-R. Koivunen, Annotea: an open RDF infrastructure for shared Web annotations. Proceedings of the 10th international conference on World Wide Web, 2001.

[10]. L. Gravano, H. Garcia-Molina, A. Tomasic, "GlOSS: Text-Source Discovery over Internet", *TODS* **24**(2), 1999.

[11]. K. Khelif, R. Dieng-Kuntz, P. Barbry, An Ontology-based Approach to Support Text Mining and Information Retrieval in the Bio logical Domain, in J. UCS **13**(12), pp. 1881-1907, 2007.

[12]. A. Setzer, R. Gaizauskas, TimeM L: Robust specification of event and temporal expressions in text. In The second international conference on language resources and evaluation, 2000.

[13]. J. Kahan, M-R. Koivunen, Annotea: an open RDF infrastructure for shared Web annotations. Proceedings of the 10th international conference on World Wide Web, 2001.

[14]. A. Setzer, R. Gaizauskas, TimeM L: Robust specification of event and temporal expressions in text. In The second international conference on language resources and evaluation, 2000.

[15]. C. Roussey, S. Calabretto, An experiment using Conceptual Graph Structure for a Multilingual Information System, in the 13th International Conference on Conceptual Structures, ICCS'2005.

.