Annotation in Web Database Search Result Records with Machine Learning Technique in Frequent Pattern Clustering(FPC)

V.Sabitha[1], Dr.S.K.Srivatsa[2]

[1]Research Scholar, Sathyabama University, Chennai.

[2]Prathyusha Institute of Technology & Management, Chennai.

Abstract-

An astonishing system used for storing information which can be accessed through a website is referred to as a 'web database'. A flexible range of activities are carried out through web database. Therefore, it is important to design a proper database which involves choosing the correct data type for each field in order to reduce memory use and to increase the speed of access. Since, tiny databases do not cause any important problems, enormous web databases can grow to millions of entries and hence need to be well designed to work effectively. Thus the motive of our research is to reduce the memory and increase the s)peed of access in a web database. In this paper, we have introduced a machine learning technique based annotation to increase the speed of search result records in web database and give meaningful labels. The proposed technique is capable to efficiently reduce the recollection and add to the speed of access in a website.

Keywords: Alignment, Frequent Pattern Algorithms, Score Calculation.

I. INTRODUCTION

Internet has a major role in the day today life style of human being. Internet also has an necessary part of our lives. So the techniques in this are helpful in extract data present on the web is an motivating area of research [1, 2]. These internet techniques help to take out data from Web data, where in at any rate one of arrangement or usage data is used in the mining process [3]. Web mining is the application of data mining techniques to extract facts from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc.

By means of the volatile growth of information sources available on the World Wide Web and the quickly rising pace of support to Internet commerce, the Internet has evolved into a bullion mine that contains or animatedly generates information that is helpful to E-businesses [4]. A web site has the bulk of direct link of a company has to its current and possible customers. These companies can study the visitor's activities through web analysis, and find the patterns in the visitor's behavior [5]. The web analysis yield the rich results for a company's data warehouse, offer great opportunity for the near prospect.

Web mining is the term of applying data mining techniques to automatically discover and extract useful information from the World Wide Web documents and services [12]. Although Web mining puts down the roots deeply in data mining, it is not equivalent to data mining. The unstructured feature of Web data triggers more complexity of Web mining. Web mining research is actually a converging area from several research communities, such as Database, Information Retrieval, Artificial Intelligence [13] and also psychology and statistics as well. Business benefits of web mining affords to digital service providers include personalization, collaborative filtering, enhanced customer support, product and service strategy definition, particle marketing and fraud detection [14]. In short, the ability to understand their customers' needs and to deliver the best and most appropriate service to those individual customers at any given moment [15].

The requirement for predicting user needs in order to improve the usability and user retention of a Website can be addressed by personalizing it [16]. Web personalization is defined as any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site. The objective of a Web personalization system is to provide users with the information they want or need, without expecting from them to ask for it explicitly. Web

data are those that can be collected and used in the context of Web personalization [17, 18].

Web mining approach to detect users accessing terrorist related information by processing all ISPs traffic is suggested [19]. Automatically pages detection in a website whose location is different from where visitors expect to find them [20]. The key insight is that visitors will backtrack if they do not find the information where they expect the point from where they backtrack is the expected location for the page.

A. Objective:

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services. Many research works were developed in the area of mining from web data, documents, hyperlinks and web sites etc. One of the recently developed data annotation method which is given in [24], they initially extracts the search result records (SRRs) from the result page extracted from the WDBs and that extracted SRRs are consequently involved in the three phases namely alignment phase, annotation phase and annotation wrapper generation phase. In alignment phase, the SRRs are organized into different groups with each group corresponding to a different concept. In the second phase, a most appropriate label for each group is determined by using the probability model. Afterward, for each identified concept, an annotation rule was generated which describes how to extract the data units of this concept in the result page and what the appropriate semantic label should be. By exploiting these three phases they perform the annotation for the SRRs and achieving 98% in precision and recall. But this technique needs an improvement in the SRRs grouping process because the initial phase of the SRRs grouping is plays a vital role. The more accurate grouping of SRRs has given more accurate annotation results. Moreover in the annotation phase they exploiting probability model to compute appropriate label for each group. The probability model based label selection not acquires a more precise result in all SRRs groups from different training sites. The lack of solution for such drawbacks has motivated me to do the research work in this area.

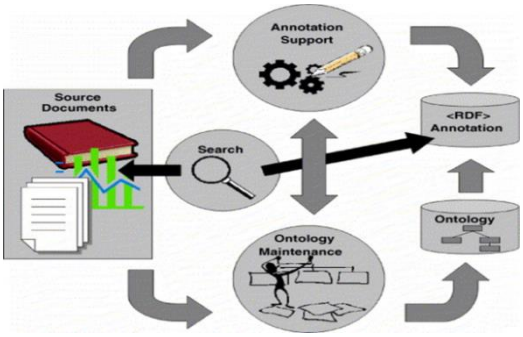

Fig 1:Block diagram of existing method

II. RECENT RELATED RESEARCHES: A REVIEW

Some of the recent so far work related to the web mining is listed as follows:-

Dusan Stevanovic *et al.* [22] have inspected the effects of applying seven well familiar data mining classification algorithm on static web server logs. Those effects were examined for the purpose of classify the user sessions as it belong to either automated web crawlers or human visitors and also identify which of the automated web crawlers 'malicious' behavior and potentially participants in a Distributed Denial of Service (DDOS) attack. The classification performance was evaluated in terms of classification accuracy, recall, precision and F1score. Seven beyond nine vector features were borrowed from earlier studies on classification of user sessions as belonging to web crawlers. Two novel web session features were introduced i.e. the successive sequential request ratio and standard divergence of page request depth. In terms of the information gain and the gain ratio metrics the efficiency of the new feature was evaluated. The experimental results of the method showed the potential of the new features to improve the accuracy of data mining classifiers in identifying malicious and well-behaved web crawler sessions.

Olatz Arbelaitz *et al.* [24] have proposed a system, which combines web usage and content mining techniques with the three principal objectives. The objectives used were creating user steering profiles used for link prediction; inspiring the profiles with semantic information to expand them to offer the Destination Marketing Organizations (DMO) with a tool to initiate links that matched the users flavor, and in addition obtaining global and language dependent user interest profiles to afford the DMO staff with important information for future

web designs, and allows them to design future marketing campaigns for specific targets. That system executed successfully, the obtained profiles vigorous in more than 60% of cases with the real user navigation sequences and in more than 90% of cases with the user interests. In addition the automatically extracted semantic structure of the website and the interest profiles were validated by the BTw DMO staff.

Yiyao Lu *et al.* [25] have presented an automatic annotation approach for the web mining application. In that approach at first aligns the data units on a result page into different groups such that the data in the same group had the same semantic. An annotation wrapper for the search site was automatically constructed and was used to annotate new result pages from the same web database. From the experimental result they have proved the high effectiveness.

V.Sabitha et al[26] have presented an automatic annotation approach give accurate best label for the web mining application. we have introduced a machine learning technique based annotation to increase the speed of search result records in web database and give meaningful labels. The proposed technique is capable to efficiently reduce the recollection and add to the speed of access in a website.

III. PROPOSED METHODOLOGY

The main aim of this research is to offer a better annotation method for web database records by solving the drawbacks that currently exist in the literary works. Hence, I have planned to propose a new annotation method with AI technique perform the annotation with different number of preparation sites. The proposed AI based annotation method includes four stages namely, alignment phase, Score Calculation, annotation phase, and annotation wrapper generation phase. In alignment phase the SRRs from the WDBs are assembled in different groups which have different concepts. To achieve this alignment phase here we will proposed an efficient frequent pattern clustering algorithm. Based on the clustering algorithm the SRRs are grouped in alignment phase. Afterward, we will calculate a score value for each SRR from different groups by using it content, domain, position and title. Thus the calculated score value is used to select the suitable label for the group. In the next phase, using the score value an ANN will be taught and selects the most suitable label for the group based on that group score value. Moreover, in the final phase an annotation wrapper generation will be performed to generate an annotation rule. The rules for all groups, collectively, form the annotation wrapper for the corresponding WDB, which can be used to directly annotate the data retrieved from the same WDB in response to new queries without the need to perform the annotation phases. Hence, the SRRs are from the different web sites will be annotated more effectively by achieving precision and recall than the existing methods

A Alignment Phase

The process carried out in alignment phase are data features extraction and data clustering which is given in the below section

B Data Features Extraction

i Data Unit Similarity

The data unit similarity is to found for the search result obtained to align the data units of same concept into a single group. Based on five features (data content, data type, tag path, adjacency and presentation style), the similarity between two data units du_1 and du_2 are found,

The similarity between two data units is given by

$$\begin{aligned} s(du_1, du_2) &= w_1 * sC(du_1, du_2) + \\ & w_2 * sP(du_1, du_2) + w_3 * sD(du_1, du_2) + \\ & w_4 * sT(du_1, du_2) + w_5 * sA(du_1, du_2) \end{aligned} \quad (1)$$

ii *Data Content Similarity (Sc)*

The data content similarity between to data units du_1 and du_2 is given by the equation

$$SC(du_1, du_2) = FV_{du_1} \bullet FV_{du_2} / \left\| FV_{du_1} \right\| * \left\| FV_{du_2} \right\| \quad (2)$$

In the above equation, FV_{du} is the frequency vector of data unit d terms, $\left\| FV_{du} \right\|$ is the length of FV_{du} , and the numerator is the inner product of two vectors.

.

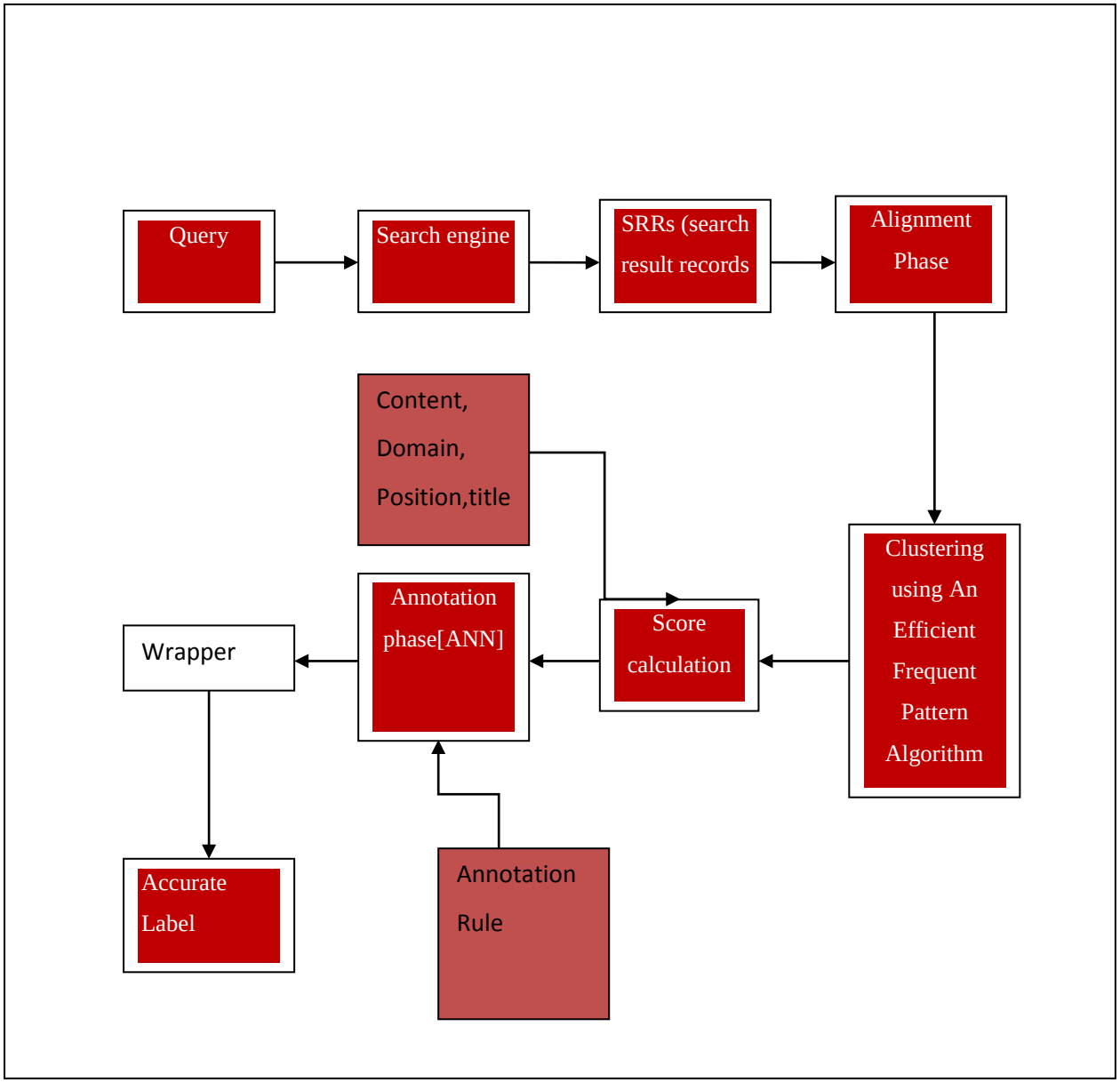

Fig.2.Block diagram of the proposed method

iii Data Type Similarity (Sd)

The data type similarity between to data units du_1 and du_2 is given by the equation

$$SD(du_1, du_2) = LCS(f_1, f_2) / MAX(Elen(f_1), Elen(f_2)) \quad (3)$$

In the above equation, LCS is the longest common sequence and f1 and f2 are the sequence of the data types of du_1 and du_2 respectively. $Elen(f)$ is the number of component types of data type f.

iv Presentation Style (Sp)

Presentation style consists of six style features. They are font weight, font size, font color, font face, text decoration and italic font. The presentation style between two data units du_1 and du_2 is given by

$$SP(du_1, du_2) = \sum_{i=1}^{6} MP_i / 6 \quad (4)$$

In the above equation, MS_i is the score of the ith style feature. Here MP_i =1 if $M_i^{du_1} = M_i^{du_2}$

Else MP_i =0. M_i^{du} Is ith style feature of data unit du

v Tag Path Similarity (St)

The tag path similarity between two data units du_1 and du_2 is given by the equation

$$ST(du_1, du_2) = 1 - EDT(t_1, t_2) / (Tle(t_1) + Tle(t_2)) \quad (5)$$

In the above equation, EDT is the edit distance between the tag paths of two data units du_1 and du_2. Here, the edit distance is referred by the number of insertions and deletions of tags needed to transform one tag path into the other. t_1 and t_2 are the tag paths of two data units du_1 and du_2 and $Tle(t)$ is the number of tags in tag paths

vi Adjacency Similarity (Sa)

The adjacency similarity between two data units du_1 and du_2 is given by the equation

$$SA(du_1, du_2) = (s^t(du_1{}^p, du_2{}^p) + s^t(du_1{}^s, du_2{}^s))/2 \quad (6)$$

In the above equation, du^p and du^s are the preceding and succeeding data units of du

C An Efficient Frequent Pattern Algorithm

1. L_1=find Frequent_1 – itemsets(D);
2. For (k=2,$L_{k-1} \neq \Phi$;k++){
3. ck=apriori_gen(Lk-1);
4. for each transaction t€ D
5. Ct=subset(ck-t);
6. for each candidates c€ Ct
7. c.count++;
8. }
9. Lk={c€ck/c.count>=min-sup}
10. }
11. return l=Uk Lk

Procedure apriori _gen(L_{k-1}:frequent(K-!)-itemsts)

1. for each itemset l1€ L_{k-1}
2. for each itemset l2€ L_{k-1}
3. if(l1[1]=l2[1]) ˆ(l1[2]=l2[2]) ˆ .ˆ(l1[k-2]=l2[k-2] ˆ(l1[k-1]=l2[k-1]) then {
4. c=l1 Union l2;
5. if has _infrequent _subset(c,L_k-1)then
6. delete c;
7. else add c to c_k;
8. }
9. Return c_k;

D Score Calculation

After clustering the data units of same concept into one group, the labels are assigned by calculating the score value of each group by using it content, domain, position and title. The content, domain, position and title based calculation is given in the below section.

i Title Based Calculation

For each link there must be a title based on which the calculation is carried out as detailed below: After separating the query words and finding the meanings for all of them, we compare them with the titles of the unique links separately to find the frequency of the words.

$$TB_s(p) = \sum_{i=1}^{n} \left(\frac{TB_s du_i - \max(TBdu_i) + 1}{\max(TBdu_i)} \times w_Q + \sum_{j=1}^{b} \frac{TB_s du_i N_j - \max(TBdu_i N_j) + 1}{\max(TBdu_i N_j)} \times w_N \right) \quad (9)$$

In the above equation, $TB_s(p)$ symbolizes the title based value of s th unique link; and $TB_s du_i$ is the number of occurrences of i th query word in the title TB of s th link, $\max(TBdu_i)$ is the maximum number of occurrence of i th query word in the title of whole unique links, $TB_s du_i N_j$ is the number of occurrence of j th meaning of i th query word in the title TE of s th link, $\max(TBdu_i N_j)$ is maximum number of occurrence of j th meaning of i th query word in the title of whole unique links, n is the total number of query word, and b, the total number of meaning of i th query word, w_Q the weight value of the query word and w_N is the weight value of the meaning word of the query word.

ii Content Based Calculation

In the content based calculation we compare the contents of each link with the separated query words and their synonyms to check the number of occurrences of separated query words and their synonyms in the contents of each link.

. The calculation based on content is shown by an equation below:

$$CB_s(p) = \sum_{i=1}^{n} \left(\frac{CB_s du_i}{\max(CBdu_i)} \times w_Q + \sum_{j=1}^{b} \frac{CB_s du_i N_j}{\max(CBdu_i N_j)} \times w_N \right) \quad (10)$$

In the above equation, $C_s(p)$ represents the calculated content based value of s th unique link; and $CB_s du_i$ is the number of occurrence of i th query word du in the content of s th unique link, $\max(CBdu_i)$ is

the maximum number of occurrence of i th query word du in the content of s th unique link, $CB\ du_i N_j$, the number of occurrence of j th synonym of i th query word du in the content of s th unique link; and $\max\left(CBdu_i N_j\right)$is the maximum number of occurrence of j th synonym of i th query word du in the content of s th unique link.

iii. Domain Calculation

Each link we have obtained from the different search engines invariably comes under a specific domain name. An example for such domain name is 'Wikipedia'. We calculate the domain value for each unique link using the domain name we found for each link in the different search engines. The equation to calculate the domain value for each unique link is given below:

$$DC_s(p) = \log_{10}\left(\frac{2N - 1 + acc_s}{2N}\right) \qquad (11)$$

In the above equation, $DC_s(p)$represents the calculated domain value of s th unique link and N , the number of search engines we used; and acc_s , the number of unique links with same domain name.

iv. Position Calculation

This calculation is based on the ranking of the link in different search engines we have used i.e. the link present in the position in each search engine which we chosen for our process. The formula to calculate the position of a link is shown below:

$$PS_s(p) = \frac{N * k - \left(\sum_{l=1}^{m} PS(p)\right)}{N * k} \qquad (12)$$

In the above equation, $PS_s(p)$ represents the position value of the link, and N , the number of search engines used, k is number of links we have taken for our process from each search engine; and $PS(p)$, the rank of a link in a particular search engine.

E Annotation Phase

Annotation phase is carried out by neural network training process which is detailed in the below section

i Neural Network Training

Once the title based calculation, content based calculation, domain calculation and position calculation are found and the labels are assigned using this score value, and the appropriate label is found out using ANN method. The given fig.2 shows the neural network of our process.

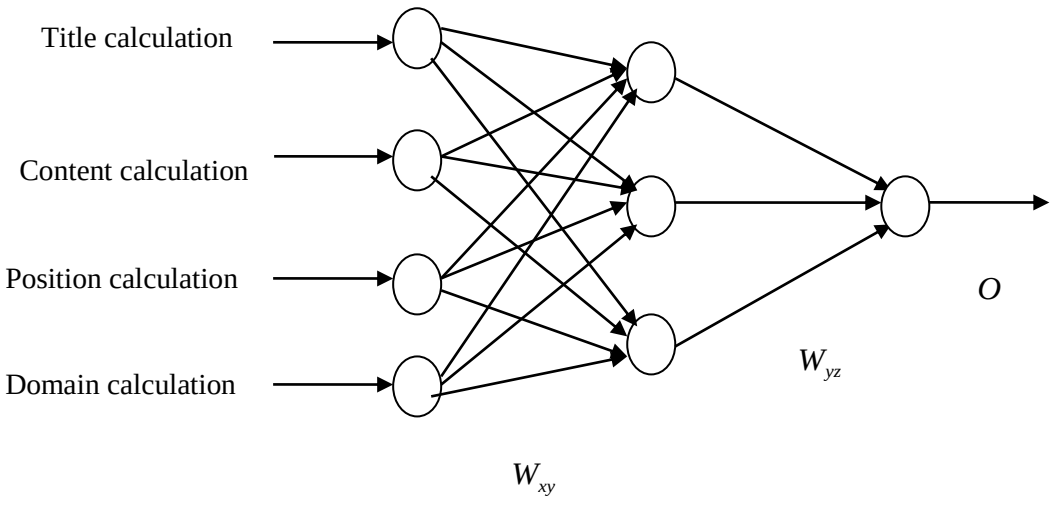

Fig.3. Neural network of our process

In Fig.3, W_{xy} represent the weight values between the input layer and the hidden layer, W_{yz} , the weight values between the hidden layer and the output layer and O, the output of the neural network. The neural network is trained based on the weight values which are adjusted as per the error we have obtained. The error is calculated by checking the difference between the target value and the output obtained using neural network. The target value is based on the user ranked list and the weight values on the back propagation algorithm. It is explained as follows: initially the weights in the neural network are random numbers and the output from the neural network for the given input is based on the weight values. Fig.3 shows a sample connection in neural network for learning back propagation algorithm

It is calculated based on the equation below:

$$er_K = O_K(1-O_K)(T-O_K) \quad (13)$$

In the above equation, er_K represents the error from the node K, O_K , the output from the node K and T , the target based on the user ranked list. Using er_K the weight values are changed as shown below:

$$W_{IK}^{+} = W_{IK} + (er_K * O_I) \quad (14)$$

$$W_{JK}^{+} = W_{JK} + (er_K * O_J) \quad (15)$$

In the above equations W_{IK}^{+} and W_{JK}^{+} symbolize newly trained weights and W_{IK} and W_{JK} , the initial weights. Thereafter, we have to calculate the errors for the hidden layer neurons. Unlike output layer we are unable to calculate it directly. So we back propagate it from the output layer. It is shown by the equations below:

$$er_I = O_I(1-O_I)(er_K * W_{IK}) \quad (16)$$

$$er_J = O_J(1-O_J)(er_K * W_{JK}) \quad (17)$$

After obtaining the error for the hidden layer, we have to find the new weight values in between input layer and hidden layer. By repeating this method we train the neural network. Subsequently, we give the query to the system which merges the unique links from the different search engines and ranks the unique links based on the trained neural network using the score generated in the neural network for each unique link.

F Annotation Wrapper

After selecting the best label for the given query in a WDB by ANN process, the annotation wrapper process is carried out. Annotation wrapper is nothing but a set of annotation rules for all the attributes on the result page with order corresponding to the ordered data unit groups.

The annotation rule is given by

$$attribute_i = < label_i, prefix_i, suffix_i, seperators_i, unitindex_i > \quad (18)$$

To use the wrapper to annotate a new result page, for each data unit in an SRR, the annotation rules are applied on it one by one based on the order in which they appear in the wrapper. If this data unit has the same prefix and suffix as specified in the rule, the rule is matched and the unit is labeled with the given label in the rule. Annotation wrapper is created so that the new search result record can be annotated by this process without reapplying the entire annotation process

IV. RESULTS AND DISCUSSIONS

The proposed method is implemented in the working platform of java. The performance of the proposed method is compared against the performance of the existing method. The performance for proposed method and existing method is evaluated for various domains (machine, job, book, animation and song) using various annotators and various calculations. From the given below results, we can analyze the performance of the proposed method.

Table 1: Existing Method Labeling Performance

DOMAINS	Existing Method							
	Precision				Recall			
	FA	QA	IA	CA	FA	QA	IA	CA
Book	82.7	88.4	89.8	89.8	55.7	80	81.9	78.9
Job	82.3	88.5	89.8	89.7	55.2	79.9	81.6	78.8
Machine	82.5	88.6	90	89.6	55.4	79.7	81.9	78.9
Animatoin	82.8	88.4	89.7	89.6	55.4	79.9	81.9	78.1
Song	83	88.9	90	89	55.9	79.9	81	78
Average	82.7	88.6	89.9	89.7	55.6	80.1	81.8	79

Table 2: Proposed Method Labeling Performance

DOMAINS	PROPOSED Method							
	Precision				Recall			
	TC	DC	PS	CC	TC	DC	PS	CC
Book	83.1	88.5	90.	90.1	55.9	80.1	82.4	80.1
Job	82.7	88.7	89.9	90.1	55.6	80.5	82	80.2
Machine	82.9	88.9	90	90.2	55.8	80.1	82.5	80.2
Animation	83.2	88.6	89.9	90.1	55.9	80.1	82.6	78.4
Song	83.2	89.2	90.3	89.9	56.4	80.2	81.3	78.4
Average	82.9	88.8	90	90	55.9	80.2	82	79.4

V. DISCUSSIONS

Table 1 and Table 2 illustrate the performance of the proposed method and the existing method. In table.1, FA, QA, IA and CA represent the frequency annotator, query annotator, In-text prefix/ suffix annotator and common knowledge annotator. In table.2, TC, DC, PS, and CC represent the title based calculation, domain based calculation, position based calculation and content based calculation.

VI. CONCLUSION

The main aim of our work is to provide a better annotation method for web database records. Although there are various annotation methods exists in the literary works, a better performance of annotation is needed in the current situation, since there are millions number of entries in the current record and also which is increasing day by day. Hence, I have intended to propose a new annotation method with AI technique that performs the annotation with different number of training sites. The results are taken for the proposed method and the existing methods and the performance is analyzed. The Search result records from different websites are taken and annotated using the proposed and the existing method. The precision and recall and Modified K-Mean Clustering Algorithm (MKMC) of the results are taken as the output for the proposed method and the existing method. From the results, the performances of both the proposed and existing methods are analyzed. As seen from the result, in most cases, the performance of the proposed method is better than the performance of the existing method for both and MKMC and precision and recall. Thus, we can conclude that the proposed method is well capable of annotating the web database records.

REFERENCES

[1] A.G. Buchner, M. Baumgarten, S.S. Anand, M.D. Mulvenna, and J.G. Hughes, "Navigation pattern discovery from Internet data," In Proceedingsof theWeb Usage Analysis and User Profiling Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (San Diego, August), pp. 25–30, 1999.

[2] S. Brin and L. Page, "The anatomy of a large-scale hyper textual Web search engine," Computer Networks and ISDN Systems, Vol. 30, pp. 107–117, 1998.

[3] J. Borges and M. Levene, "Mining Association Rules in Hypertext Databases," In Knowledge Discovery and Data Mining, pp. 149–153, 1998.

[4] Robert Cooley, Bamshad Mobasher and Jaideep Srivastava, "Web Mining: information and Pattern Discovery on the WWW," In Proceedings of Ninth IEEE International Conference on Tools with Artificial Intelligence, pp. 558-567, 1997.

[5] R. Cooley, B. Mobasher and J. Srivastava, "Data Preparation for Mining World Wide Web Browsing Patterns," Knowledge and Information Systems, Vol. 1, No.1, 1999.

[6] Monika Yadav and Mr. Pradeep Mittal, "Web Mining: An Introduction," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 3, No. 3, pp. 683-687, 2013.

[7] Brij M. Masand, Myra Spiliopoulou, JaideepSrivastava and Osmar R. Zaiane, "WEBKDD 2002 – Web Mining for Usage Patterns & Profiles," SIGKDD Explorations, Vol. 4, No. 2, pp. 125-127, 2002.

[8] M. Spiliopoulou, "Data Mining for the Web," In Proceedings of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), pp. 588-589, 1999.

[9] T. Srivastava, P. Desikan and V. Kumar, "Web Mining – Concepts, Applications and Research Directions," Foundations and Advances in Data Mining Studies in Fuzziness and Soft Computing, Vol. 180, pp. 275-307, 2005.

[10] J. Srivastava, and B. Mobasher, "Web Mining: Hype or Reality?," In Proceedings of 9th IEEE International Conference on Tools With Artificial Intelligence (ICTAI '97), Newport Beach, CA, 1997.

[11] Raymond Kosala and HendrikBlockeel, "Web mining research: a survey," ACM SIGKDD Explorations Newsletter Homepage archive, Vol. 2, No. 1, pp. 1-15, 2000.

[12] Oren Etzioni, "The World Wide Web: Quagmire or Gold Mine," Communications of the ACM, Vol. 39, No. 11, pp. 65-68, 1996.

[13] Mobasher, R. Cooley and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Communications of the ACM, Vol. 43, No. 8, 2000.

[14] Dean W. Abbott, Philip Matkovsky and John F. Elder IV, "An Evaluation of High- end Data Mining Tools for Fraud Detection," In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Vol. 3, pp. 2836-2841, 1998.

[15] Ting, I-Hsien, Wu and Hui-Ju, "Web Mining Applications in E-Commerce and E-Services," Studies in Computational Intelligence, Vol. 172, 2009.

[16] Magdalini Eirinaki and Michalis Vazirgiannis, "Web Mining for Web Personalization," ACM Transactions on Internet Technology, Vol. 3, No. 1, pp. 1-27, 2003.

[17] Mulvenna, Anand and Buchner, "Personalization on the net using web mining," Communication. ACM, Vol. 43, No. 8, pp. 123–125, 2000.

[18] J. Srivastava, Cooley, Deshpande and Tan, "Web usage mining: Discovery and applications of usage patterns from web data," SIGKDD Explorations Vol. 1, No. 2, pp. 12–23, 2000.

[19] Elovici, Kandel, M.Last, B.Shapira and O. Zaafrany, "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web," Journal of Information Warfare, 2004.

[20] RamakrishnanSrikant and Yinghui Yang, "Mining Web Logs to Improve Website Organization," Proceedings of the 10th international conference on World Wide Web, pp. 430-437, 2001.

[21] DusanStevanovic, Aijun An and Natalija Vlajic, "Feature evaluation for web crawler detection with data mining techniques," An International Journal of Expert Systems with Applications, Vol. 39, pp. 8707–8717, 2012.

[22] Juan D. Velasquez, "Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments," An International Journal of Expert Systems with Applications, Vol. 40, pp. 5228–5239, 2013.

[23] Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesus Maria Perez and Inigo Perona, "Web usage and content mining to extract knowledge for modeling the users of the Bidasoa Turismo website and to adapt it," An International Journal of Expert Systems with Applications, Vol. 40, pp. 7478–7491, 2013.

[24] Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng and Clement Yu, "Annotating Search Results from Web Databases," IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, pp. 514-527, 2013.

[25] V.Sabitha,S.K.Srivatsa,"Machine Learning Technique Based Annotation in Web Database Search Records with Aid of Modified K-Mean Clustering(MKC)," Research journal of Applied Sciences,Engineering &Technology10(8):853-862,2015.