Research Journal of Applied Sciences, Engineering and Technology 10(8): 853-862, 2015
DOI:10.19026/rjaset.10.2440
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: October 05, 2014 Accepted: January 23, 2015 Published: July 20, 2015

Research Article

Machine Learning Technique Based Annotation in Web Database Search Result Records with Aid of Modified K-Means Clustering (MKMC)

[1]V. Sabitha and [2]S.K. Srivatsa
[1]Sathyabama University, Chennai, India
[2]St. Joseph's College of Engg., Chennai, India

Abstract: To reduce the memory usage and increase the speed of access in web database, in this study, we have introduced a machine learning technique based annotation with the help of modified K-means clustering algorithm to increase the speed of search result records in web database. The proposed AI based annotation method includes four stages namely, alignment phase, Score Calculation, annotation phase and annotation wrapper generation phase. These four stages of the proposed part are spelt out in this study. The proposed technique is competent to effectively reduce the memory and increase the speed of access in a website. The proposed method is implemented in the working platform of java and the results are analyzed.

Keywords: Annotation, modified K-means clustering, score calculation, wrapper generation

INTRODUCTION

Internet takes a major role in the day today life style of human being. Internet also has an essential part of our lives. So the techniques in this are helpful in extracting data present on the web is an interesting area of research (Buchner *et al*., 1999, Brin and Page, 1998). These internet techniques help to extract information from Web data, wherein at any rate one of structure or usage data is used in the mining process (Borges and Levene, 1998). Web mining is the application of data mining techniques to extract knowledge from web data, including web documents, hyperlinks between documents, usage logs of web sites, etc.

By means of the explosive development of information sources available on the World Wide Web and the rapidly growing pace of espousal to Internet commerce, the Internet has evolved into a gold mine that contains or dynamically generates information that is beneficial to E-businesses (Cooley *et al*., 1997). A web site has the majority of direct link of a company has to its current and potential customers. These companies can study the visitor's activities through web analysis and find the patterns in the visitor's behavior (Cooley *et al*., 1999). The web analysis yields the wealthy results for a company's data warehouses, offer great opportunities for the near future.

The web mining is divided into three different categories these are Web usage mining, Web content mining and Web structure mining (Yadav and Mittal, 2013). Web usage mining is the process of extracting useful information from server logs i.e., user's history. It is the process of determine the users, whom are looking on Internet. Various users might be looking at only textual data, while some others might be interested in multimedia data. These technologies are basically concentrated upon the use of the web technologies which could help for betterment (Masand *et al*., 2002; Spiliopoulou, 1999). Web structure mining, is used to identify the relationship between Web pages linked by information or direct link connection. Spiders scanning of the Web sites are preformed to take place the completion task (Srivastava *et al*., 2005). Hyperlink hierarchy will establish the related information within the sites to the relationship of competitor links and connection through search engines and third party co-links (Srivastava and Mobasher, 1997). Web content mining is the mining extraction and integration of useful data, information and knowledge from Web page contents (Kosala and Blockeel, 2000).

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 (Etzioni, 1996). 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 (Mobasher *et al*., 2000) 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 (Abbott *et al*., 1998). In short, the ability to

Corresponding Author: V. Sabitha, Sathyabama University, Chennai, India

understand their customers' needs and to deliver the best and most appropriate service to those individual customers at any given moment (Ting and Wu, 2009).

The requirement for predicting user needs in order to improve the usability and user retention of a Website can be addressed by personalizing it (Eirinaki and Vazirgiannis, 2003). 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 (Mulvenna *et al*., 2000; Srivastava *et al*., 2000).

Web mining approach to detect users accessing terrorist related information by processing all ISPs traffic is suggested (Elovici *et al*., 2004). Automatically pages detection in a website whose location is different from where visitors expect to find them (Srikant and Yang, 2001). 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.

Objective of the study: An amazing system used for storing information which can be accessed through a website is referred to as a 'web database'. A versatile range of purposes 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 consumption and to increase the speed of access. Since, miniature databases do not cause any significant problems, gigantic 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 speed of access in a web database by developing a new annotation method.

LITERATURE REVIEW

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

Alkhattabia *et al*. (2011) have proposed an evaluation model for information quality in e-learning systems based on the quality framework. They have also proposed a framework consists of 14 quality dimensions grouped in three quality factors: intrinsic, contextual representation and accessibility in previous. They implemented a goal-question-metrics approach to develop a set of quality metrics for the identified quality attributes within the proposed framework. That proposed metrics were computed to produce a numerical rating indicating the overall information quality published in a particular e-learning system. The data collection and evaluation processes were automated using a web data extraction technique and results on a case study are discussed. That assessment model could be useful to e-learning systems designers, providers and users as it provides a comprehensive indication of the quality of information in such systems.

Stevanovic *et al*. (2012) have inspected the effects of applying seven well recognized data mining classification algorithm on static web server logs. Those effects were examined for the purpose of classify the user sessions as it belonging 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 F1 score. 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 effectiveness 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.

Velasquez (2013) has presented an integrative approach based on the distinctive attributes of web mining in order to determine which techniques and uses were harmful. The legal framework applicable to privacy affairs between private parties, the most adequate method of protecting users was considered via the contractual remedies. The contract structure was suited to the specific characteristics of the mining project. Two basic categories of web mining projects were defined for the graphical illustration, they are projects based on the mining of web logs with the intention of improving the navigation experience within a certain web site and the use of mining tools upon web data in order to make more complex inferences about an individual's attributes. The first case illustrates the publication of a clear privacy policy which details both the purposes and the pattern extraction techniques. Alternatively the second illustrates the recommendations were substantially different. At last the web miner should obtain automated decisions about individuals regarding topics with a high social impact and not to take care to not use available technology.

Arbelaitz *et al*. (2013) 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 (Bidasoa Turismo Website) BTW DMO staff.

Lu *et al*. (2013) 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.

MATERIALS AND METHODS

In our proposed method, first a set of SRRs (Search Result Records) are extracted from a result page from a WDB (Web Database). Once the SRRs are extracted, the similarity of data units (data unit in this a study is referred as a piece of text that represent a concept of an entity) are found for the whole search records based on the five features (data content, presentation style, data type, tag path, adjacency). As soon as the similarity is found for every data units, the data units are aligned in one group which is of the same concept. The alignment here is done by the modified K-means algorithm. Once the data units of same concept is arranged into one group, label is assigned to each data unit using the score value based on title calculation, content based calculation, domain calculation and position calculation and the best label is selected by the ANN (Artificial Neural Network) method for each group. Finally, an annotation wrapper phase is carried out. Annotation wrapper means simply a set of rules designed for each concept, which describes that, how to extract the data units of the same concept in the result page and what is the appropriate semantic label can be for that. Our proposed method structural diagram is presented in Fig. 1.

Given below are some of the search results of book robots.

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

Data features extraction:

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

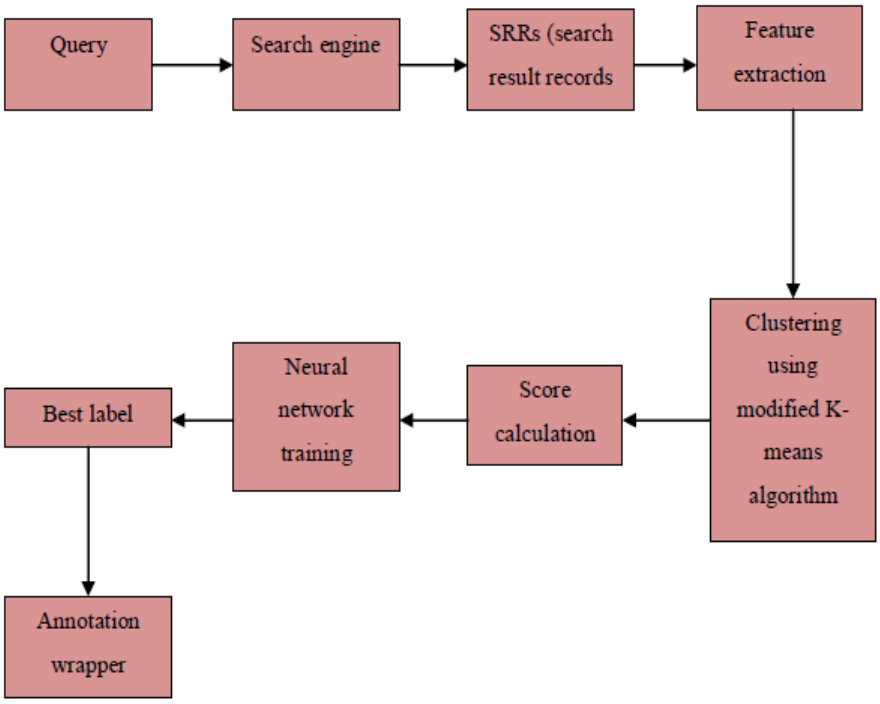

Fig. 1: Block diagram of the proposed method

features (data content, data type and 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:

$$s(du_1, du_2) = w_1 * sC(du_1, du_2) + w_2 * sP(du_1, du_2) + w_3 * sD(du_1, du_2) + w_4 * sT(du_1, du_2) + w_5 * sA(du_1, du_2) \quad (1)$$

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, $\|FV_{du}\|$ is the length of FV_{du} and the numerator is the inner product of two vectors.

Data type Similarity (Sd): The data type similarity between to data units du_1 and du_2 is given by the equation:

$$sD(du_1, du_2) = LCS(e_1, e_2) / MAX(Elen(e_1), Elen(e_2)) \quad (3)$$

In the above equation, LCS is the longest common sequence and e_1 and e_2 are the sequence of the data types of du_1 and du_2 respectively. $Elen(e)$ is the number of component types of data type e.

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:

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

In the above equation, MS_i is the score of the i^{th} style feature. Here $MS_i = 1$ if $M_i^{du_1} = M_i^{du_2}$ Else $MS_i = 0$. M_i^{du} Is ith style feature of data unit du.

Tag path Similarity (St): The tag path similarity between two data units du_1 and du_2 is given by the equation:

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

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 *Tlen* (*t*) is the number of tags in tag paths.

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.

MODIFIED K-MEANS CLUSTERING ALGORITHM

Once the data unit similarity is obtained for all data units using the data unit similarity formula, the data units of the same concept are clustered into a group by means of modified K-means clustering algorithm. The modified K-means clustering algorithm process is described below:

- Let us assume that n number of data units is given to the clustering process. The data units are represented as $N_i; i = 1,2,...n$
- Randomly choose K centroids
- Compute Euclidean distance:

$$D(N_i, K_j) = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{m} (N_i - K_j)} \quad (7)$$

where,
N_i = The data units t be clustered
K_j = The randomly chosen centroids

- Each centriods values are differenced with the all data units and compared with the user defined threshold value:

$$C_j = \begin{cases} N_i; \ if \ \ D(N_i, K_j) < T \\ otherwise \end{cases} \quad (8)$$

- The distance values which are less than the defined threshold value are stored in the cluster with the corresponding centroid value
- Now recalculate the position of k centroids
- Finally, repeats the step 3 and 4 until the centroids become fixed

At the end of this process, the data units of same concept are arranged into a group.

Table 1: Frequency of query words and meanings of query words in the title

	du_1	du_1 N1	du_2	du_2 N2	.	.	.	du_n	du_n Nb
TE1	TE1 du_1	TE1 du_1 N1	TE1 du_2	TE1 du_2 N2	.	.	.	TE1 du_n	TE1 du_n Nb
TE2	TE2 du_1	TE2 du_1 N1	TE2 du_2	TE2 du_2 N2	.	.	.	TE2 du_n	TE2 du_n Nb
TE3	TE3 du_1	TE3 du_1 N1	TE3 du_2	TE3 du_2 N2	.	.	.	TE3 du_n	TE3 du_n Nb
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
TEs	TEs du_1	TEs du_1 N1	TEs du_2	TEs du_2 N2	.	.	.	TEs du_n	TEs du_n Nb

Table 2: Frequency of query words and their synonyms in the contents of each link

	du_1	du_1 N1	du_2	du_2 N2	.	.	.	du_n	du_n Nb
CE1	CE1 du_1	CE1 du_1 N1	CE1 du_2	CE1 du_2 N2	.	.	.	CE1 du_n	CE1 du_n Nb
CE2	CE2 du_1	CE2 du_1 N1	CE2 du_2	CE2 du_2 N1	.	.	.	CE2 du_n	CE2 du_n Nb
CE3	CE3 du_1	CE3 du_1 N1	CE3 du_2	CE3 du_2 N2	.	.	.	CE3 du_n	CE3 du_n Nb
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
CEs	CEs du_1	CEs du_1 N1	CEs du_2	CEs du_2 N2	.	.	.	CEs du_n	CEs du_n Nb

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.

Title based calculation: For each document (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, which is shown in Table 1.

Table 1 explained as follows: the du_1, du_2, …., du_n represents the separated words of the query we have given and 'n' represents the number of separated words in the query and du_1 N1, the first meaning N of the respective word du_1 and du_n Nb, the b^{th} meaning N of n^{th} separated word du of the query we have given. The Ti du_1 represents the number of times the first separated word present in the title of the first unique link whereas TE1 du_1 N1 represents the number of times the first meaning of the first separated word present in the title of the first unique link. The title based calculation for each unique link is shown by an equation below:

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

In the above equation, $TE_s(p)$ symbolizes the title based value of s^{th} unique link; and $TE_s du_i$ is the number of occurrences of i^{th} query word in the title TE of s^{th} link, max ($TEdu_i$) is the maximum number of occurrence of i^{th} query word in the title of whole unique links, $TE_s du_i N_j$ is the number of occurrence of j^{th} meaning of i^{th} query word in the title TE of s^{th} link, $\max(TEdu_i N_j)$ 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.

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. Table 2 shows the number of occurrences of query words and their synonyms in the contents of each link.

Table 2 explains as follows: the CE_1, CE_2, ……., CE_s represents the contents of the unique links and Ces du_n represents the number of occurrence of n^{th} query word in the content of s^{th} unique link and CEs du_n Nb, the number of occurrence of b^{th} synonym of n^{th} query word in the content of s^{th} unique link. The calculation based on content is shown by an equation below:

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

In the above equation, $C_s(p)$ represents the calculated content based value of s^{th} unique link; and $CE_s du_i$ is the number of occurrence of i^{th} query word du in the content of s^{th} unique link, $\max(CEdu_i)$ is the maximum number of occurrence of i^{th} query word du in the content of s^{th} unique link, $CE_s du_i N_j$, the number of occurrence of j^{th} synonym of i^{th} query word du in the

content of S^{th} unique link; and $\max(CEdu_iN_j)$ is the maximum number of occurrence of j^{th} synonym of i^{th} query word du in the content of s^{th} unique link.

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:

$$DO_s(p) = \log_{10}\left(\frac{2m - 1 + acc_s}{2m}\right) \quad (11)$$

In the above equation, $DO_s(p)$ represents the calculated domain value of s^{th} unique link and m, the number of search engines we used; and acc_s, the number of unique links with same domain name. For example if we are having ten unique links out of which five are from same domain, while checking any one of the link from those five unique links which are under same domain, the acc_s value is five for that related unique link.

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:

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

In the above equation, PS_s (p) represents the position value of the link and m, 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.

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

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.

Robots, Robots Everywhere! (Little Golden Book) by Sue Fliess and Bob Staake (Aug 6, 2013)

Formats	Price	New	Used
Hardcover Order in the next **15 hours** to get it by **Tuesday, May 20**. FREE Shipping on orders over $35	~~$3.99~~ **$3.59**	$1.32	$0.01
Kindle Edition Auto-delivered wirelessly	**$2.53**		

DK Eyewitness Books: Robot by Roger Bridgman (Mar 1, 2004)

Formats	Price	New	Used
Hardcover Order in the next **15 hours** to get it by **Tuesday, May 20**. FREE Shipping on orders over $35	~~$16.99~~ **$12.23**	$9.83	$0.79

Boy and Bot by Ame Dyckman and Dan Yaccarino (Apr 10, 2012)

Formats	Price	New	Used	Collectible
Hardcover Order in the next **15 hours** to get it by **Tuesday, May 20**. FREE Shipping on orders over $35	~~$16.99~~ **$12.97**	$7.03	$6.69	$8.50
Kindle Edition Auto-delivered wirelessly	**$7.69**			

Fig. 2: Example search results from amazon.com

Figure 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. Figure 3 shows a sample connection in neural network for learning back propagation algorithm.

Figure 4, the output of node (neuron) K is formed from the neurons I and J. Let K be the output layer and I, J, hidden layers. First we calculate the error in the output from K. 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

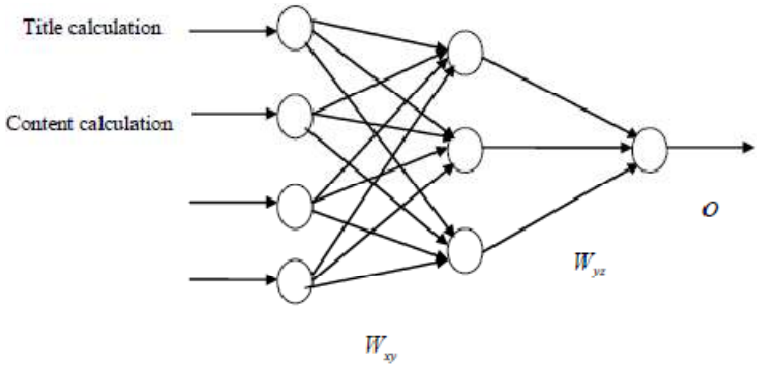

Fig. 3: Neural network of our process

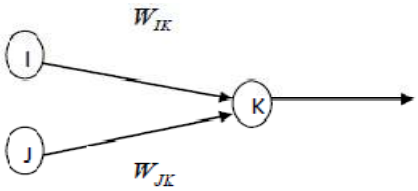

Fig. 4: Sample connection in neural network

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. Figure 4 shows the algorithm of our proposed technique.

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.

RESULTS AND DISCUSSION

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 (music, job, book, game and movie) using various annotators and various calculations. From the given below results which is given in Table 3 and 4, we can analyze the performance of the proposed method.

Discussions: Table 3 and 4 illustrate the performance of the proposed method and the existing method. Table 3, FA, QA, IA and CA represent the frequency annotator, query annotator, In-text prefix/suffix annotator and common knowledge annotator. Table 4, TC, DC, PS and CC represent the title based calculation, domain based calculation, position based calculation and content based calculation. Table 3 the average performance of 4 annotators namely frequency annotator, query annotator, In-text prefix/suffix

Table 3: Existing method labeling performance

	Existing method							
	Precision				Recall			
Domains	FA	QA	IA	CA	FA	QA	IA	CA
Book	82.7	88.3	89.9	89.8	55.6	80.0	81.9	78.9
Job	82.3	88.5	89.8	89.7	55.2	79.9	81.7	78.8
Music	82.5	88.6	90.0	89.7	55.3	79.8	81.8	78.8
Game	82.8	88.4	89.7	89.6	55.4	79.9	81.9	78.0
Movie	83.0	88.9	90.0	89.0	55.9	79.9	81.0	78.0
Average	82.6	88.5	89.9	89.6	55.4	80.0	81.7	79.0

Table 4: Proposed method labeling performance

	Proposed method							
	Precision				Recall			
Domains	TC	DC	PS	CC	TC	DC	PS	CC
Book	82.9	88.5	90.0	90.1	55.6	80.0	82.4	80.0
Job	82.5	88.7	89.9	90.1	55.4	80.3	82.0	80.1
Music	82.8	88.9	90.0	90.2	55.6	80.1	82.3	80.1
Game	83.0	88.6	89.9	90.1	55.7	80.1	82.3	78.3
Movie	83.2	89.2	90.3	89.9	56.3	80.2	81.2	78.2
Average	82.8	88.7	90.0	90.0	55.7	80.1	82.0	79.3

Fig. 5: Average performance graph of existing method

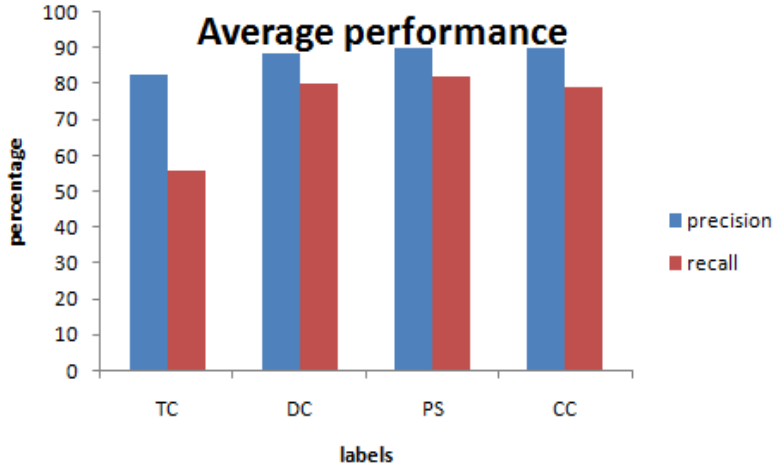

Fig. 6: Average performance graph of proposed method

annotator and common knowledge annotator is given which is compared against the performance of the proposed method namely title based calculation, domain based calculation, position based calculation and content based calculation. From the values given in the tables, two graphs are plotted for the comparison of the performance of the proposed method and the existing method by taking its precision and recall values. From the graph is given in Fig. 5 and 6, it is clear that the precision and recall of our proposed method is higher than the existing method in all the cases. Thus the performance of the proposed method performs better than the existing method. Once the precision and recall of the proposed method is higher than the existing method, it is well capable of labeling the records in search engine and thus reducing the time consumption in searching a particular file.

CONCLUSION

The main aim of our study 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 SRRs (Search Result Records) from different websites are taken and annotated using the proposed and the existing method. The precision and recall 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 precision and recall. Thus, we can conclude that the proposed method is well capable of annotating the web database records.

REFERENCES

Abbott, D.W., P. Matkovsky and J.F. Elder IV, 1998. An evaluation of high-end data mining tools for fraud detection. Proceeding of the IEEE International Conference on Systems, Man and Cybernetics, 3: 2836-2841.

Alkhattabia, M., D. Neagu and A. Cullen, 2011. Assessing information quality of e-learning systems: A web mining approach. Comput. Hum. Behav., 27: 862-873.

Arbelaitz, O., I. Gurrutxaga, A. Lojo, J. Muguerza, J.M. Perez and I. Perona, 2013. Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it. Expert Syst. Appl., 40: 7478-7491.

Borges, J. and M. Levene, 1998. Mining association rules in hypertext databases. Proceeding of the 4th International Conference on Knowledge Discovery and Data Mining, pp: 149-153.

Brin, S. and L. Page, 1998. The anatomy of a large-scale hyper textual web search engine. Comput. Netw. ISDN Syst., 30: 107-117.

Buchner, A.G., M. Baumgarten, S.S. Anand, M.D. Mulvenna and J.G. Hughes, 1999. Navigation pattern discovery from Internet data. Proceeding of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Web Usage Analysis and User Profiling Workshop (WEBKDD'99), San Diego, pp: 25-30.

Cooley, R., B. Mobasher and J. Srivastava, 1997. Web mining: Information and pattern discovery on the World Wide Web. Proceeding of 9th IEEE International Conference on Tools with Artificial Intelligence, pp: 558-567.

Cooley, R., B. Mobasher and J. Srivastava, 1999. Data preparation for mining World Wide Web browsing patterns. Knowl. Inf. Syst., 1(1).

Eirinaki, M. and M. Vazirgiannis, 2003. Web Mining for Web personalization. ACM T. Internet Techn., 3(1): 1-27.

Elovici, Y., A. Kandel, M. Last, B. Shapira and O. Zaafrany, 2004. Using data mining techniques for detecting terror-related activities on the web. J. Inform. Warfare., 3(1): 17-29.

Etzioni, O., 1996. The World-Wide Web: Quagmire or gold mine? Commun. ACM, 39(11): 65-68.

Kosala, R. and H. Blockeel, 2000. Web mining research: A survey. ACM SIGKDD Explor. Newsletter Homepage Arch., 2(1): 1-15.

Lu, Y., H. He, H. Zhao, W. Meng and C. Yu, 2013. Annotating search results from web databases. IEEE T. Knowl. Data En., 25(3): 514-527.

Masand, B.M., M. Spiliopoulou, J. Srivastava and O.R. Zaiane, 2002. WEBKDD 2002-web mining for usage patterns & profiles. SIGKDD Explor., 4(2): 125-127.

Mobasher, B., R. Cooley and J. Srivastava, 2000. Automatic personalization based on web usage mining. Commun. ACM, 43(8).

Mulvenna, M.D., S.S. Anand and A.G. Büchner, 2000. Personalization on the net using web mining: Introduction. Commun. ACM, 43(8): 123-125.

Spiliopoulou, M., 1999. Data mining for the web. Proceeding of the Symposium on Principles of Knowledge Discovery in Databases (PKDD), pp: 588-589.

Srikant, R. and Y. Yang, 2001. Mining web logs to improve website organization. Proceeding of the 10th International Conference on World Wide Web, pp: 430-437.

Srivastava, J. and B. Mobasher, 1997. Web mining: Hype or reality? Proceeding of 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '97). Newport Beach, CA.

Srivastava, T., P. Desikan and V. Kumar, 2005. Web Mining: Concepts, Applications and Research Directions. Foundations and Advances in Data Mining, Studies in Fuzziness and Soft Computing, 180: 275-307.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explor., 1(2): 12-23.

Stevanovic, D., A. An and N. Vlajic, 2012. Feature evaluation for web crawler detection with data mining techniques. Expert Syst. Appl., 39(10): 8707-8717.

Ting, I.H. and H.J. Wu, 2009. Web Mining Applications in E-commerce and E-services. Series of Studies in Computational Intelligence, Vol. 172, Springer-Verlag, Heidelberg, Germany, January 2009, ISBN: 978-3-540-88080-6.

Velasquez, J.D., 2013. Web mining and privacy concerns: Some important legal issues to be consider before applying any data and information extraction technique in web-based environments. Experty Syst. Appl., 40: 5228-5239.

Yadav, M. and P. Mittal, 2013. Web mining: An introduction. Int. J. Adv. Res. Comput. Sci. Software Eng., 3(3): 683-687.