ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Extracting the Information by Ranking Techniques to Increase the Privacy of Search Engine

C.Parimala[1], Prof.B.Sakthivel[2]

M.E, Department of CSE, P.S.V College of Engineering and Technology, Krishnagiri, Tamilnadu, India[1]

Head of the Department, Department of CSE, P.S.V College of Engineering and Technology, Krishnagiri, Tamilnadu, India[2]

ABSTRACT : An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine process able, this is essential for many applications such as deep web data collection and Internet.

Comparison shopping, they need to be extracted out and assigned meaningful labels. In this paper, we present an automatic annotation approach that first aligns the data units on a result page into different groups such that the data in the same group have the same semantic. The ability to accurately judge the semantic similarity between words is critical to the performance of several applications such as Information Retrieval and Natural Language Processing.

Therefore, in this paper we propose a semantic similarity measure that uses in one hand, an online English dictionary provided by the Semantic Atlas project of the French National Center for Scientific Research (CNRS) and on the other hand, a page counts based metric returned. The ranking techniques and login security concepts for increase the efficiency and privacy of search engine.

KEYWORDS—Data alignment, data annotation, dynamic code generation, ranking prediction.

I.INTRODUCTION

A large portion of the deep web is database based, for many search engines, data encoded in the returned result pages come from the underlying structured databases. Such type of search engines is often referred as Web databases (WDB). A typical result page returned from a WDB has multiple search result records (SRRs). Each SRR contains multiple data units each of which describes one aspect of a real-world entity. Three SRRs on a result page from a book WDB. Each SRR represents one book with several data units. In this paper, a data unit is a piece of text that semantically represents one concept of an entity. It corresponds to the value of a record under an attribute. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags and describes the relationships between text nodes and data units in detail. In this paper, I perform data unit level annotation.

There is a high demand for collecting data of interest from multiple WDBs. 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 to the same book. To enable fully automatic annotation, the result pages have to be automatically obtained and the SRRs need to be automatically extracted. In a meta search context, result pages are retrieved by queries submitted by users (some reformatting may be needed when the queries are dispatched

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

to individual WDBs).In the deep web crawling context, result pages are retrieved by queries automatically generated by the Deep Web Crawler.

II.RELATED WORK

The main contribution of this paper is the design of a system to align the data's into semantic order, based on clustering technique and comparing with online virtual shopping. As we have pointed out in the introduction, to the best of our knowledge, I propose a new method for display the web search results set using annotator and also table format. And this website result datasets display orders based on previous user recommendation and ranking of the record sets. The user to view and analysis the search results easily and also user get good results from this sits, and providing a high security.

III.K-MEANS CLUSTER ALGORITHM – DATA FILTERING

k-means is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed apriori. The main idea is to define k centers, one for each cluster. These centers should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-calculate k new centroids as bary center of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new center. A loop has been generated. As a result of this loop we may notice that the k centers change their location step by step until no more changes are done or in other words centers do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function given by:

$$J(V) = \sum_{i=1}^{c} \sum_{j=1}^{c_i} \left(\left\| x_i - v_j \right\| \right)^2$$

where,

'$||x_i - v_j||$' is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

Steps for filtering a data using k-means clustering

Let $X = \{x_1,x_2,x_3,\ldots\ldots..,x_n\}$ be the set of data points and $V = \{v_1,v_2,\ldots\ldots.,v_c\}$ be the set of centers.

1) Randomly select 'c' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using:

$$v_i = (1/c_i) \sum_{j=1}^{c_i} x_i$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step 3

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Advantages

1) Fast, robust and easier to understand.
2) Relatively efficient: O (tknd), where n is # objects, k is # clusters, d is # dimension of each object, and t is # iterations. Normally, k, t, d << n.
3) Gives best result when data set are distinct or well separated from each other.

IV.APRIORI ALGORITHM GRAPH

Apriori is a classic algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association rules which highlight general trends in the database: this has applications in domains such as market basket analysis

Apriori is designed to operate on databases containing transactions (for example, collections of items bought by customers, or details of a website frequentation). Other algorithms are designed for finding association rules in data having no transactions (Winepi and Minepi), or having no timestamps (DNA sequencing).

Each transaction is seen as a set of items (an *itemset*). Given a threshold C, the Apriori algorithm identifies the item sets which are subsets of at least C transactions in the database.

The pseudo code for the algorithm is given below for a transaction database T, and a support threshold of ϵ. Usual set theoretic notation is employed; though note that T is a multiset. C_k is the candidate set for level k. Generate () algorithm is assumed to generate the candidate sets from the large item sets of the preceding level, heeding the downward closure lemma. $count[c]$ Accesses a field of the data structure that represents candidate set c, which is initially assumed to be zero. Many details are omitted below, usually the most important part of the implementation is the data structure used for storing the candidate sets, and counting their frequencies.

$$
\begin{array}{l}
\textbf{Apriori}(T, \epsilon) \\
\quad L_1 \leftarrow \{\text{large 1 - itemsets}\} \\
\quad k \leftarrow 2 \\
\quad\quad \textbf{while } L_{k-1} \neq \textit{emptyset} \\
\quad\quad\quad C_k \leftarrow \{a \cup \{b\} \mid a \in L_{k-1} \wedge b \in \bigcup L_{k-1} \wedge b \notin a\} \\
\quad\quad\quad \textbf{for transactions } t \in T \\
\quad\quad\quad\quad C_t \leftarrow \{c \mid c \in C_k \wedge c \subseteq t\} \\
\quad\quad\quad\quad \textbf{for candidates } c \in C_t \\
\quad\quad\quad\quad\quad count[c] \leftarrow count[c] + 1 \\
\quad\quad\quad L_k \leftarrow \{c \mid c \in C_k \wedge count[c] \geq \epsilon\} \\
\quad\quad\quad k \leftarrow k + 1 \\
\quad\quad \textbf{return } \bigcup_k L_k
\end{array}
$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

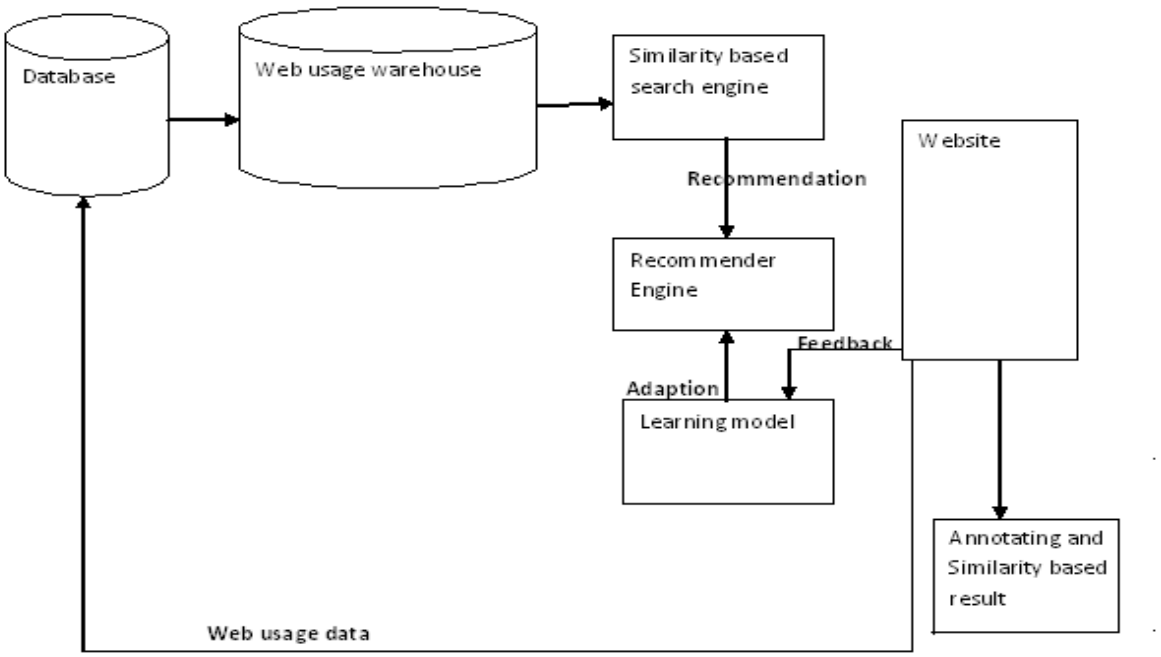
(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

V.SYSTEM ARCHITECTURE



The main components of the architecture of the system are ontology processor, ranker module, document processor. In the approach given here, I first determine the keywords of the document using syntactic analysis and making the vector space model of the documents. For this a domain specific dictionary is prepared having the words, their synonyms along with their meaning which are related to the domain.

The words present in the dictionary are assigned weights based on their relevance/relation with the domain using fuzzy set approach. The query which can be given related to the domain have been framed using the standards of the Word Net. I have thus developed a database of words, their weight age, their synonyms, ontology etc.

I have represented this database as a vector space model for the processing purpose. The mapping of each words present in the vector space models stored in the document repository is done with the domain specific dictionary weighted terms called semantic dictionary database. This is done sentence wise.

Each sentence will contain a relevance value and then the integration of all the sentences is done by using statistical approach. This integrated relevance value obtained above will form the relevance of the individual paragraphs.

Finally, the relevance value of the document is obtained by integrating the relevance score of all the paragraphs present in the document again by using statistical model. The relevance score of the document obtained above is for the query specified for the domain as the mapping of vector space model is done with the help of fuzzy weighted terms in the domain specific dictionary stored in dictionary repository. The query process: using those data structures to produce a ranked list of documents for users.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

VI. REGISTRATION FORM

In the Registration Form, the users have to make registration here. As per the registration a jar will be downloaded as per the random value. User has to install the jar in the java supporting mobile. Using the jar only we will do the login form. In the jar there will be expression calculation. Expression varies for each jar. Expression will be stored in the database.

VII.LOGIN FORM

In the login form the user will give the user name and password first. If the username and password is same means a random key will be sent to the access page. User has to install the jar and enter the random key contain in access page. As per the user expression calculation will be done and viewed in the access code text field. Please enter the value in the website if the value is correct means enter to the user's page.

VIII.SECRET LITTLE FUNCTIONS

There will be 11 jars the secret value and secret Function will vary for each jar. Calculation Part in the Secret Little Function module is as Follows.The access code values will be split into 3 parts. We split the value in 3 parts and assign to the 3 variables eg a, b, c. Then a will be added with X variable b will be subtract with x variable and c will be multiplied with x. Here x value will vary for each jar. Assign the value as a1, b1, c1. Secret Function will vary for each user. The expression calculation will be in a1 b1 c1 format only. The values will be passed to the expression and generated code will be generated.

IX.VIRTUAL PASSWORDS

In the Virtual Password module the Secret Function calculation will vary for different jar. The use gets the random value and generated value in dynamic format. Virtual means dynamic. Random numbers keep on changing so that Generated code will also keep on changing dynamically.

Posting the opinion:

In this module, we get the opinions from various people about business, e-commerce and products through online. The opinions may be of two types. Direct opinion and comparative opinion. Direct opinion is to post a comment about the components and attributes of products directly. Comparative opinion is to post a comment based on comparison of two or more products. The comments may be positive or negative.

Object identification:

In general, people can express opinions on any target entity like products, services, individuals, organizations, or events. In this project, the term object is used to denote the target entity that has been commented on. For each comment, we have to identify an object. Based on objects, we have to integrate and generate ratings for opinions. The object is represented as "O". An opinionated document contains opinion on set of objects as $\{o_1, o_2, o_3 \ldots o_r\}$

X.RATING PREDICTION

First the ratings of unrated items are estimated based on the available information (typically using known user ratings and possibly also information about item content) using some recommendation algorithm. Heuristic techniques typically calculate recommendations based directly on the previous user activities.

Each user ranks all the predicted items according to the predicted rating value ranking the candidate (highly predicted) items based on their predicted rating value, from lowest to highest (as a result choosing less popular items).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

XI.QUERY SUGGESTIONS

In order to recommend relevant queries to Web users, a valuable technique, query suggestion, has been employed by some prominent commercial search engines.

This extends the original query with new search terms to narrow down the scope of the search. But different from query expansion. Query suggestion aims to suggest full queries that have been formulated by previous users so that query integrity and coherence are preserved in the suggested queries.

XII.COLLABORATIVE FILTERING AND DISPLAY RESULTS

In this module used to filter the results based on user queries, user recommendation, rating. And display the final result sets using annotator and also table format. So user easy to view and analysis the results set.

XIII.CONCLUSION

In this paper, I have studied the data annotation problem and proposed a multi annotator approach to automatically constructing an annotation wrapper for annotating the search result records retrieved from any given web database. This approach consists of six basic annotators and a probabilistic method to combine the basic annotators. Each of these annotators exploits one type of features for annotation and our experimental results show that each of the annotators is useful and they together are capable of generating high quality annotation.

XIV.FUTURE WORK

A special feature of our method is that, when annotating the results retrieved from a web database, it utilizes both the LIS of the web database and the IIS of multiple web databases in the same domain. We also explained how the use of the IIS can help alleviate the local interface schema inadequacy problem and the inconsistent label problem. In this paper, we also studied the automatic data alignment problem. Accurate alignment is critical to achieving holistic and accurate annotation.

Our method is a clustering based shifting method utilizing richer yet automatically obtainable features. This method is capable of handling a variety of relationships between HTML text nodes and data units, including one-to-one, one-to-many, many-to-one, and one-to-nothing. Our experimental results show that the precision and recall of this method are both above 98 percent. For example, we need to enhance our method to split composite text node when there are no explicit separators. We would also like to try using different machine learning techniques and using more sample pages from each training site to obtain the feature weights so that we can identify the best technique to the data alignment problem.

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely fortunate to have got this all along the completion of my project work. Whatever I have done is only due to guidance and assistance and I would not forget to thank them. I respect and thank my guide Mr. B. SAKTHIVEL M.E., Head of the Department of Computer Science and Engineering for providing all support and guidance which made me to complete the project on time.The blessing, help and guidance given by him time to time shall carry me a long way in the journey of life on which I am about to embark .Lastly, I thank almighty, my parents for their constant encouragement without which this assignment would not be possible.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

REFERENCES

[1] Arasu.A and Garcia-Molina.H, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.
[2] Arlotta.L,Crescenzi.V, Mecca.G, and Merialdo.P, "Automatic Annotation of Data Extracted from Large Web Sites," Proc. Sixth Int'l Workshop the Web and Databases (WebDB), 2003.
[3] Bruce Croft.W, "Combining Approaches for Information Retrieval," Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic, 2000.
[4] Chan.P and Stolfo.S, "Experiments on Multistrategy Learning by Meta-Learning," Proc. Second Int'l Conf. Information and Knowledge Management (CIKM), 1993.
[5] Crescenzi.V, Mecca.G, and Merialdo.P, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf., 2001.
[6] Dill et al.S, "SemTag and Seeker: Bootstrapping the Semantic Web via Automated Semantic Annotation," Proc. 12th Int'l Conf. World Wide Web (WWW) Conf., 2003.
[7] Embley.D, Campbell.D, Jiang.Y, Liddle.S, Lonsdale.D, Ng.Y, and Smith.R, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.
[8] Elmeleegy.H, Madhavan.J, and Halevy.A, "Harvesting Relational Tables from Lists on the Web," Proc. Very Large Databases (VLDB) Conf., 2009.
[9] Freitag.D, "Multistrategy Learning for Information Extraction," Proc. 15th Int'l Conf. Machine Learning (ICML), 1998.
[10] Goldberg.D, Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, 1989.
[11] Handschuh.S and Staab.S, "Authoring and Annotation of Web Pages in CREAM," Proc. 11th Int'l Conf. World Wide Web (WWW), 2003.
[12] Handschuh.S, Staab.S, and Volz.R, "On Deep Annotation," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.
[13] He.B and Chang.K, "Statistical Schema Matching Across Web Query Interfaces," Proc.
[14] He.H, Meng.W,Yu.C, and Wu.Z, "Automatic Integration of Web Search Interfaces with WISE-Integrator," VLDB J., vol. 13, no. 3, pp. 256-273, Sept. 2004.