A Review on Annotating Search Results from Web Databases

Richa Saxena, Sushil Chaturvedi
M. Tech. CSE Final Year, Assistant Professor
SRCEM, Banmore

Abstract- Now a day's development of internet leads to larger size of data. This data is usually stored in a database. When these data is accessed from the web and web extraction techniques were used to maintain such types of records then these databases are known as web databases. Current trends show the vital use of such data bases are growing rapidly. The data units returned from such database are usually encoded into the result pages dynamically for human browsing. These are mainly used by various search engines. In this paper we are going to present some techniques for annotating search based result from web databases.

Keywords- Search, Annotation, Web, Human Browsing.

I. INTRODUCTION

Data that are omitted from various search engines as a search result may be generated or accessed from various inbuilt databases managed by webs. These types of databases are known as Web Database abbreviated as WDB. Many search engines uses large part of the deep web is database based, it means, data encoded in the returned result pages come from the underlying structured databases. These search engines are omits result are mainly generated by web databases (WDB). These are multiple results known as search result record (SRR). Each SRR contains multiple data units each of which describes one aspect of a real-world entity. Here data unit is a piece of text that semantically represents one concept of a thing. It communicates to the value of a record under a characteristic. It is diverse from a text node that passes on to a sequence of text surrounded by a pair of HTML tags [1].

Deep Web primarily contains some dynamic pages returned by the underlying databases. Compared with Surface Web, Deep Web contains more abundant, valuable and focusing on specified field information that cannot be indexed by traditional search engines. But its characteristics of heterogeneity, large-scale, distributed and autonomy makes how to take effective advantage of Web information demanding. Structural information of database model is lost through query-processing and query-results are only for browsing, so applications cannot understand this information extracted from result pages. Semantic annotation is to find an explicit semantic vocabulary for every data unit in result pages, for making these data understandable and process-able for computers. Annotation process of deep web data uses the heuristic information summarized in above the adapter to annotate Deep Web query results [2].

The emergence of the deep-web is posing many new challenges in data integration. Standard search engines like Google are not able to crawl to these web-sites. At the same time, in several domains, manually submitting online queries to frequent query forms, keeping track of the gained results and merging them collectively is a tedious and error-prone process. A challenge associated with deep web systems, which has not received attention so far, arises because the deep web databases within a specific domain are often not independent, i.e., the output results from one database are needed for querying another database. For a given user query, multiple databases may need to be queried in an intelligent order to retrieve all the information desired by a user. Thus, there is a need for techniques that can produce query plans, accounting for dependencies among the data sources and extracting all information desired by a user [3].

Record Matching is a process to identify the duplicate records in web databases. It is an important step for data integration. In earlier systems, the record matching is addressed through the Unsupervised Online Record Matching method, UDD, i.e. for a given user query, can effectively identify duplicates from the query result records of multiple web databases. This process of record matching is done through a single domain which provides limited number of non-duplicate data results. Now-a-days, numerous number of databases that generate web pages in accordance to the user queries on the web. These web databases include redundant and unreliable data. To eliminate the duplicates or redundant data an unsupervised duplicate detection (UDD) and SVM are made used [4].

Unsupervised Duplicate Detection (UDD) for the specific record matching problem of identifying duplicates among records in query results from multiple Web databases were used. UDD focused on techniques for adjusting the weights of the record fields in calculating the similarity among two records. Two records are measured as redundant if they are similar enough on their fields. This method can remove duplicates for multiple web databases, where a single domain can include multiple web databases; similarly the user query can include multiple domains. So a system is build that helps the users to compare query result returned from multiple domain, multiple data-bases, i.e., to match the different sources records that refer to the same real-world entity, to match records that are identical [4].

Feature Selection method, the absolute values of the common vector elements are calculated for each class. Based on the consideration of indifference subspace projection, and justified by extensive testing, it is observed that elements of the common vector, which are low-in-magnitude, correspond to relatively irrelevant features compared to the features corresponding to elements which are high–in-magnitude. In other words, the common vector elements which have large magnitudes correspond to more common, hence representative, properties of respective class. Therefore, since the elements of the common vector that have small values carry relatively small information, their use in classification is redundant. In the feature selection process, labeling a feature vector index as useless according to the results obtained for one class is not appropriate. One useless feature for one class may be quite critical for the expression of the other class. Therefore, features can only be eliminated if they prove to be redundant for both of the classes. In this study, the redundancy analysis is carried out for both spam and non-spam classes, and common redundant features are eliminated. As expected, the irrelevant tagged features are intuitively justifiable. Besides, the classification performance does not deteriorate [5].

II. RELATED WORK

In recent year 2013, Yiyao Lu et al offered a technique under title Annotating Search Results from Web Databases. They 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. Proposed method divided in six basic annotators and a probabilistic method to combine the basic annotators. All annotators utilize one type of features for annotation. They also employ a probabilistic model to combine the results from different annotators into a single label. This representation is highly bendable so that the existing basic annotators may be modified and new annotators may be added easily without affecting the operation of other annotators [1].

An annotation wrapper is constructed for any WDB. The packaging can be useful to efficiently annotating the SRRs retrieved from the same WDB with new queries. A clustering-based shifting technique is offered to align data units into different groups so that the data units inside the similar group have the equal semantic. The integrated interface schema (IIS) utilize over multiple WDBs in the same domain to enhance data unit annotation. 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. Their experimental results show that the precision and recall of this method are both above 98 percent [1].

In 2013 Yong FENG and Wei LU proposed a protocol which uses heuristics-based semantic annotation method. This work summarizes some heuristic information. It uses this heuristic information to analyze the data to be annotated, which identifies a semantic vocabulary for each data unit. It performs a semantic annotation experiment on the Deep Web data of various areas in the UIUC standard dataset. Assigning semantic labels to the data units extracted from result pages is a challenging task. Experimental results show that offered method has a good annotation effect. This method is very effective and gives better performance [2].

Fan Wang et al introduced a new protocol for query planning for deep-web integration system. It uses dynamic query planner to generate an efficient query order based on the database dependencies. It selects the top K query plans and develops cost models for query planning for deep web mining. They consider a query that asks for the amino acids occurring at the corresponding position in the orthologous gene of non-human mammals with respect to a particular gene. The system is designed to support a very simple and easy to use query interface, where each query comprises a query key term and a set of query target terms that the user is interested. The query key term is a name and the query target terms capture the properties or the kind of information that is desired for this name. In the context of such a system, develop a dynamic query planner to generate an efficient query order based on the deep web database dependencies.

According to results, this algorithm gave very similar results as the optimal algorithm. The scalability of this system with respect to the number of data sources used and the number of query terms [3].

In 2012 P. Kowsiga and T. Mohanraj proposed a new technique for a Multi-domain record matching process and uses N-Staged SVM that separates the duplicate and non-duplicate records based on the classifiers. This method separates the duplicate and non-duplicate data by iterative process. In this model a single domain includes multiple web databases, a single database includes multiple hyper planes and a single hyper plane include multiple data, which are made separated as duplicate and non-duplicate using the N-Staged SVM. This process is repeated for multiple domains by constructing hyper planes for each. Hence the result produced will be efficient and more reliable results are provided for the user query [4].

Duplicate detection is an important step in data integration to extract duplicate free information and the problem of duplicate detection is done through UDD which checks for duplicates in a single domain at a stretch, which provides less query result as search is done in single domain. Although, the UDD solves the problem of Duplicate detection, it just partially accomplished the task. The resulted query results are limited as it extracted data from a single domain. To overcome this problem, N-Staged SVM algorithm is proposed which provides more results by detecting the duplicate records at multiple domains simultaneously by separating each item set in the domain through the SVM classification, similarly the process continues up to N-Stages until the duplicates are removed and the user query are extracted with non-duplicate data which are accurate result [4].

In year 2012, A.Srinivas and R. Srinivas offered Refining Redundant Query results From Multiple Web databases. In the Web database scenario, where records to match are greatly query-dependent, a pre-trained method is not applicable as the set of records in each query‘s results is a biased subset of the full data set. Duplicate detection is an important step in data integration and most state-of-the-art methods are based on offline learning techniques, which require training data. They use a unsupervised technique named Unsupervised Duplicate detection (UDD) which uses two classifiers for record matching and duplicate detection. This eliminates the user preference problem in supervised learning. By employing two classifiers that collaborate in an iterative manner, UDD identifies duplicates based on the dissimilarity among these records, field‘s weight is set and record matching is done by the first classifier. These results i.e., the matched records form the duplicate or positive set. The second classifier uses both duplicate and the non duplicate sets to identify the duplicate record pairs [6].

Supervised learning methods use only some of the fields in a record for identification. This is the reason for query results obtained using supervised learning to contain duplicate records. Unsupervised Duplicate Detection (UDd) does not suffer from these types of user reference problems. The exact matching method is applicable only for the records from the same data source. Element identification thus merges the records that are exactly the same in relevant matching fields. Ontology basically refers to the set of concepts such as things, events and relations that are specified in some way in order to create an agreed-upon vocabulary for exchanging information. Ontologies can be represented in textual or graphical formats. Usually, graphical formats are preferred for easy understand ability. Ontology matching is used for finding the matching status of the record pairs by matching the record attributes. Ontology matching helps in finding the relevant records based on the user queries by considering all the record attributes information [6].

In year 2011, Wei Zheng et al offered Search Result Diversification for Enterprise Data. They study the problem of result diversification for enterprise search with the focus on extracting high quality query subtopics from the integrated enterprise data. They suggest integrating the structured and unstructured data to discover meaningful query subtopics in search result diversification. To evaluate the proposed query subtopic identification methods use the results as query subtopics and then diversify search results using a state of the art sub topic based diversification method. The experimental result shows that the integrated subtopics can cover different information of the query and are effective in diversifying the documents. [7].

In same year 2011, Sarawuth Sonnum et al proposed Approximate Web Database Search Based on Euclidean Distance Measurement. In most of databases they lost eliminate data due to improper match of data properties. Thus a large number of possible choices are unnecessary eliminated. They develop a new method that can include as many possibly relevant objects as possible to facilitate approximate search through the Web database. This method finds objects with similar or closely properties and then identifies objects that share closest properties based on the

Euclidean distance measurement. This research is to devise an efficient similarity search based on closest Euclidean distance to extract the most similar objects to the users‟ preferences in the Web-based application. The result is a database that has some records cut out. When used in the process of selecting data using Euclidean data selection algorithm, this database will reduce processing time [8].

The implementation of Euclidean distance based data selection algorithm using Erlang as a programming language. Erlang is a functional language with a declarative style of function declaration. This allows program implementation to be done in a short time. The purpose of research is to find the most similar objects to the user's preferences in the Web-based applications. The user's queries identify some attributes of the desired objects. This algorithm then computes the Euclidian distances of the target object and the surrounding data. From the given distance threshold, the algorithm can produce the most relevant objects to the user interest. They implement the proposed method with the Erlang programming language and test the program with the telecommunication Web database. The experimental results reveal that the program can display similar objects within the short period of time. The proposed method can thus be applied to other Web applications [8].

In same year, Anuradha and A.K.Sharma offered A Novel Technique for Data Extraction from Hidden Web Databases. The traditional search engines use inverted index as a data structure to index the web data and keyword interface to retrieve the data. But, surfacing the Hidden Web is more difficult task in many respects. First, the index structures for the hidden web deal with the structured data as well as the large volume of data. Second, the search query interfaces often have more than one attribute and requires their respective values to be submitted. Therefore, there arises the need for new information services that can help users to find the information in integrated form. To minimize user effort, the problem of automatically interaction with information sources in the hidden web is explored [9].

Search Query Interface is considered as an entrance to the websites that are powered by backend databases. User can find the desired information by submitting the queries to these interfaces. These queries are constructed as SQL queries to fetch data from hidden sources and send it back to user with desired results. The proposed approach is presented in four phases. Firstly, different query interfaces are analyzed to select the attribute for submission. In the second phase, queries are submitted to interfaces. Third phase extracts the data by identifying the templates and tag structures. Fourth phase integrates the data into one repository with all duplicate records removed. It is the scheme of a new information retrieval service that can help users to find the desired information in integrated form. To minimize user effort, the problem of automatically interaction with hidden web sources is explored. Repository is formed by the data extracted from various sites. In addition to this, duplicate records are also removed from repository and this repository is prepared for user search. Later on, When user fills the global form for searching, SQL query is fired o this repository to get the desired result [9].

Sanjay Agrawal et al offered Exploiting Web Search Engines to Search Structured Databases. Novel integrated search architecture was offered to establish and exploit the relationships between web search results and the items in structured databases to identify the relevant structured data items for a much wider range of queries. This focus on the \mentions relationship", that is, a web document is related to an entity if the entity occurs the document. Current entity extraction techniques use machine learning and natural language techniques to parse documents and break it into sentences, and assign parts of speech tags for extracting entities. These techniques can be quite resource-intensive. Even if entity extraction is per- formed at document indexing time in a web search engine, the additional overhead is typically unacceptable as it adversely affects the document crawl rates. It is shown that establishing and exploiting relationships between web search results and structured entity databases significantly enhances the effectiveness of search on these structured databases [10].

In year 2012, S. Gunal offered Hybrid feature selection for text classification. n text classification studies, widely used feature selection methods are univariate filter approaches due to the mass amount of features that require significant processing time. Once the individual discriminatory powers of the features are obtained, the best N features are selected while the others are eliminated. Hence, a compact subset of features is attained, although feature dependencies are ignored. In spite of the extensive number of feature selection studies on text classification, there is no significant work investigating the efficacy of a combination of features, which are selected by a variety of selection methods, under different conditions. Therefore, a hybrid feature selection strategy, which consists of both filter and wrapper feature selection steps, is proposed. This hybrid selection process is repeated

under particular conditions, including different feature set sizes, dataset characteristics, classifiers, and success measures. This approach enables us to discover which features or feature combinations are better identifiers for text classification and whether there is a correlation among the useful features, the desired feature size, the utilized classification method, and the success measure. During the assessment, information such as the uniqueness and coverage rate of the features are utilized. The results of the experimental study reveal that a combination of the features selected by various methods is more effective than the features selected by the single selection method. The profile of the combination, however, depends on the characteristics of the dataset and the choice of the classification algorithm and success measure [11].

In year 2011, Yuanchun Zhu and Ying Tan proposed a Local-Concentration-Based Feature Extraction Approach for Spam Filtering. They propose a local concentration (LC)-based feature extraction approach for anti-spam by taking inspiration from the biological immune system (BIS). The LC approach is considered to be able to effectively extract position-correlated information from messages by transforming each area of a message to an analogous LC feature. Two implementation approaches of the LC approach are designed using a fixed-length sliding window and a variable-length sliding window. To integrate the LC scheme into the whole process of spam filtering, a generic LC model is designed and presented. BIS is an adaptive distributed system with the capability of discriminating "self cells" from "non-self cells." It protects our body from attacks of pathogens. Antibodies, produced by lymphocytes to detect pathogens, play core roles in the BIS. On the surfaces of them, there are unambiguous receptors which can combine corresponding specific pathogens. Thus, antibodies can detect and destroy pathogens by binding them. All the time, antibodies circulate in our body and kill pathogens near them without any central controlling node. In the BIS, two types of immune response may happen: a primary response and a secondary response. The primary response happens when a pathogen appears for the initial time. In this case, the antibodies with resemblance to the pathogen are produced slowly. After that, a corresponding long-lived B memory cell (a type of lymphocyte) is created. Then when the same pathogen appears again, a secondary response is triggered, and a huge amount of antibodies with high resemblance to that pathogen are proliferated [12].

Alper Kursat Uysal et al suggested The Impact of Feature Extraction and Selection on SMS Spam Filtering. They extensively analyses the effects of several feature extraction and feature selection methods together on filtering SMS spam messages in two dissimilar languages, explicitly Turkish and English. The whole feature set of the filtering scheme is composed of the features originated from the bag-of-words (BoW) model, and also an ensemble of structural features (SF) adopted for the spam problem. The distinctive features based on the bag-of-words model are determined using chi-square and Gini index based feature selection methods. The selected features are then combined with the structural features and fed into two separate pattern categorization algorithms, specifically k-nearest neighbor and support vector machine, to categorize SMS messages as either spam or legitimate. The filtering framework is evaluated on two separate SMS message datasets consisting of Turkish and English messages, respectively. The impact of various feature extraction and selection methodologies on SMS spam filtering, particularly for Turkish and English languages was systematically observed in terms of cataloging accuracy and dimension reduction rate. Alternatively effectiveness of the utilized feature selection strategies was not significantly superior to each other for both languages [13].

III. CONCLUSION

The high quantity of electronic information obtainable on the Internet increases the difficulty of dealing with it in modern years. In order to provide people a unified access to these Web databases and achieve information from them automatically is difficult task. In this paper we are briefly discussed some techniques to proficiently handles the web database.

REFERENCES

[1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng and Clement Yu "Annotating Search Results from Web Databases", IEEE Transactions on Knowledge and Data Engineering, Vol. 25, No. 3, March 2013.

[2] Yong FENG, Wei LU "Heuristics-based Semantic Annotation for Deep Web Query Results" Journal of Computational Information Systems, vol. 9, issue 14, pp. 5685 – 5692, July 15, 2013.

[3] Fan Wang, Gagan Agrawal, and Ruoming Jin "Query Planning for Searching Inter-Dependent Deep-web Databases", Scientific and Statistical Database Management,

Lecture Notes in Computer Science, Vol. 5069, pp 24-41, 2008.

[4] P. Kowsiga and T.Mohanraj "Multi-Domain Record Matching over Query Results from Multiple Web Databases", International Journal of Scientific & Engineering Research, ISSN 2229-5518, Volume 3, Issue 5, pp. 1-6, May-2012.

[5] Serkan Günal, Semih Ergin, M. Bilginer Gülmezoğlu, and Ö. Nezih Gerek "On Feature Extraction for Spam E-Mail Detection", Proceedings of the 2006 international conference on Multimedia Content Representation, Classification and Security (MRCS'06), pp. 635-642, 2006.

[6] A. Srinivas and R. Srinivas "Refining Redundant Query results From Multiple Web databases", International Journal of Computer Application, ISSN: 2250-1797, Volume 2, Issue 2, pp. 232 – 236, April 2012.

[7] Wei Zheng, Hui Fang, Conglei Yao and Min Wang "Search Result Diversification for Enterprise Data", Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 1901-1904, 2011.

[8] Sarawuth Sonnum, Somtida Thaithieng, Sittichai Ano, Krerkchai Kusolchu, and Nittaya Kerdprasop "Approximate Web Database Search Based on Euclidean Distance Measurement", proceedings of the International Multi Conference of Engineers and Computer Scientists (IMECS), Vol. 1, 2011.

[9] Anuradha and A.K.Sharma "A Novel Technique for Data Extraction from Hidden Web Databases", International Journal of Computer Applications, ISSN: 0975 – 8887, Volume 15– No.4, pp. 45 – 48, February 2011.

[10] Sanjay Agrawal, Kaushik Chakrabarti, Surajit Chaudhuri, Venkatesh Ganti Arnd Christian König, Dong Xin "Exploiting Web Search Engines to Search Structured Databases", Proceedings of the 18th international conference on World wide web, pp. 501-510, 2009.

[11] S. Gunal, "Hybrid feature selection for text classification", Turkish Journal of Electrical Engineering & Computer Sciences, vol. 20, No. sup.2, pp. 1296-1311, 2012.

[12] Yuanchun Zhu and Ying Tan "A Local-Concentration-Based Feature Extraction Approach for Spam Filtering", IEEE Transactions on Information Forensics And Security, Vol. 6, No. 2, pp. 486 – 495, June 2011.

[13] Alper Kursat Uysal, Serkan Gunal, Semih Ergin, Efnan Sora Gunal "The Impact of Feature Extraction and Selection on SMS Spam Filtering", Elektronika ir Elektrotechnika (Electronics and Electrical Engineering), 2012.