International Journal of Modern Trends in Engineering and Research
www.ijmter.com
e-ISSN No.:2349-9745, Date: 2-4 July, 2015

Dynamic Annotating SRRs from Web Databases using Naïve Bayes

Ms. Nikita P. Rane[1], Prof. Dinesh D. Patil[2]
[1] Student , Department of Computer Science & Engineering, Shri Sant Gadge Baba College of, Engineering & Technology, Bhusawal, nprane.it@gmail.com
[2] Associate Professor & Head , Department of Computer Science & Engineering, Shri Sant Gadge Baba College of, Engineering & Technology, Bhusawal, dineshonly@gmail.com

Abstract— Web Databases are progressively accessible through HTML source based search engines. The knowledge or data units that are excavated from databases are generally encoded dynamically for the purpose of human browsing. Data units that are dynamically encoded are essential for multiple applications like comparison shopping, web resource rating and for deep web collection. A web user whenever submits a query, he is returned with number of SRRs which are needed to be extracted and collected at one place under their respective semantic labels. In this paper, a survey is conducted on different previous efforts taken to extract the data from multiple SRRs and to annotate them. Also the proposed system enhanced with Naïve Bayes machine learning is discussed which allows to extract the data units from SRRs , assign semantic labels to the extracted data units and to accurately align the data under their respective semantic labels.

General Terms:Information retrieval, classification.
Keywords:Web Databases, Search Result Records, Data Extraction, Data Annotation, Data Alignment, Naïve Bayes.

I. INTRODUCTION

Now- a-days a very large portion of hidden web i.e. deep web is based on database. Many search engines returns result pages which comes from structured databases having different structures. Such type of search engines are often referred to as Web Databases (WDB). These widely increasing multiple number of databases can be accessed through HTML source based search interfaces. When user submits a query, a typical result page is returned to the user. This result page consist of number of search result records (SRR) which are relevant to query. In recent years, there has become high demand or basically need for deep web collection of interest from numerous Web Databases. For example, such a collection of data can be used for comparison shopping in the domain of interest, for rating a particular web resource in comparison to other web resource, for deep web collection etc. and for all these kind of applications, system will need to collect the data from different web information providers. There is also a need to properly arrange and classify the data for the purpose of analysis. This classification of collection of data requires semantic labels under which the data will be aligned. Now here arises the purpose of the paper that how the data units will be extracted from SRRs, how the semantic labels will be generated, assigned and how properly the data will be aligned. For instance, for book shopping system, system needs to collect the books record from different books website and represent the retrieved information to user at one place. There comes the need to represent this extracted information properly at one place. User can then easily search the book by different attributes like author, title etc. With semantic labels to the value of the attributes of author title, name, ISBN, publisher etc., there is great need that properly sorted and semantic data should be aligned under the semantic label. Earlier application required tremendous human involvement and efforts for manually annotating the data units. In this paper different annotation techniques have been discussed and data unit level annotation is performed. Each of the three SRRs in Fig 1 represents the number of data units for example,

Basically, data unit and text node are distinct from each other. Data Unit is the portion of text that semantically embodies one concept of an entity. It refers to the value under an attribute of a particular record. Text node is the series of text which is surrounded by a pair of HTML tags i.e. text external to "<" and ">". Elements of text node are the visible elements on webpage and data unit is the part of text node. Naïve Bayes is used in this paper for the purpose of improving the annotations and data alignment.

II. LITERATURE SURVEY

A wrapper Induction method [2] was introduced for constructing the wrappers automatically. Basically wrappers is a process, a method used as translators where a query language is translated to a relational form. Large number of systems [2] depends on human user to mark the required information on sample web pages and label them semantically at the same time. The system then defines the series of rules that retrieve the similar set of information from the same sources. These kind of systems are usually referred to as Wrapper Induction Systems. The inductive algorithm [2] requires oracle to assign labels to examples. Secondly HLRT (Head Left Right Tail) Wrapper class [2] is used. Though [2] provides higher extraction accuracy, go through poor scalability.

A Vision based novel approach has been proposed [3] that scrutinizes the visual features of the web pages. ViDE (Vision Based Data Extractor) extracts the results which are in a structured form automatically from deep web pages. ViDE is based on extracting visual features that are visible to human users from the web pages. Four step strategy [3] is used in which first the visual representation of the web page is obtained and is converted into Visual block tree. Secondly the extraction of records from visual block tree takes place. Thirdly, the extracted data records along with data items are partitioned and the data items that are semantically closed to each other are grouped together. Fourthly, visual wrappers i.e. the set of visual features extraction rules are defined so that the data record and data item extraction is carried out for new web pages from same web database effectively and efficiently.

A wrapper generation process [4] has been investigated. The efforts are taken to generate the wrappers automatically. The wrappers [4] but just focuses on data extraction and not for annotation.

A highly automated technique [5] is described for annotating HTML documents especially which contains numerous semantic concepts per document. The bootstrapping technique is discussed which identifies the unlabeled instances in different documents. This techniques called bootstrapping exploits the examination that the items which are semantically related with each other illustrates consistency in appearance and spatial locality for identifying the semantic data items from result records accurately.

DeLa (Data Extraction and Label Assignment for Web Databases) [6] concentrates on automatically extracting the data from website and assign semantic labels to the extracted data. This technique focuses purely on HTML tags and does not considers the important features such as adjacency information, text content and data type. DeLa makes use of LIS and not of IIS. Schema-based and Frequency based annotators are not used in DeLa.

III.EXISTING SYSTEM

The databases are increasing now a days and are highly accessible through HTML source based search interfaces. Human user to perform comparison shopping, deep web collection, rating different web resources, there is a great need to extract data from different SRRs and to assign meaningful labels to different values of the attributes. Along with the semantic assignment of labels to the values, it is also necessary to align the extracted data properly under the semantic labels. The data unit level annotation is performed. Data alignment is done of the data units observed on result pages. And finally an annotation wrapper is constructed automatically to annotate and align the data units from newly retrieved result pages. This system uses Clustering based shift technique. In this, clustering based shift technique, clusters of same concept are formed after initial alignment and when new value from new result page comes it is matched with the initial cluster if it does not matches then the value is shifted one position to right.

IV. CONTEMPLATED IDEOLOGY

In proposed system, data unit level annotation is considered. The system allows the user to search the query by attributes title, author name etc. The query is then submitted to different number of websites and the information relevant to query is extracted and is displayed for the user's analysis at one place.

Now, here there is a need to properly align the extracted data under the semantic labels relevant to the values of the attributes for instance, value of the attribute title should be labeled as title or by other semantic label and the value should be aligned below this automatically generated label. This paper also contributes to also display the updated information about different attributes like price, dispatched in, discount in case of book shopping system. Also the user is redirected to the particular website in order to buy the product. The proposed system considers the book domain. In this paper, above purpose mentioned purpose is achieved along with the unwanted labels generated are extracted. In this paper, the system is implemented in three phases:

1 Data Extraction Phase
2 Annotation Phase
3 Data Alignment Phase

Fig 1: Automatic Annotating SRRs using Naïve Bayes, shows the process of annotating the search result records with above three different phases.

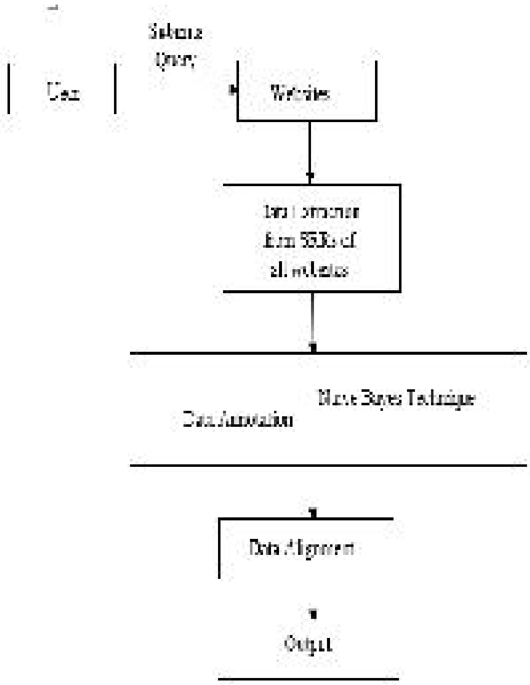

Fig 1: Automatic Annotating SRRs using Naïve Bayes

Assuming data user has inserted the query and different search result pages has been extracted for that particular query. These result pages are retrieved from different websites dynamically. Here, following four types of relationships are considered depending on number of data units, a text node contains.

1) One-to-One Relationship
2) One-to-Many Relationship
3) Many-to-One Relationship
4) One-to-Nothing Relationship

□ One-to-One relationship: For each text node there exist exactly one value for example, title attribute has only one value.

□ One-to-Many Relationship: For each text node there exist multiple values for example text node can be composed of different data units like values of title, author and publisher which are separated by a separator.

□ Many-to-One Relationship: Here, different number of text nodes form one data unit for example, there can be different values for the attribute author i.e. there can be more than one author.

□ One-to-Nothing Relationship: Here text node are not a part of any of the data units of a search result records. For example, some SRRs may not contain the data units which were there for previous result page.

Furthermore, the different data unit and text node features are considered as follows:

1) Data Content: In this type of feature, the text node and data unit of similar concept generally share some similar keywords. For example: generally price of a book in every record has leading terms as
"Price".

2) Data Type: Every data unit has its own data type. The basic data type considered are: Currency, Decimal, Symbol, Percentage and String.

3) Tag Path: A tag path is nothing but a series of tags which traversed from the root to the node.

4) Adjacency: Here, it is considered that if the successor and predecessor belongs to the same concept then the data unit also belongs to the same concept.

All the above data types helps to annotate the different data units. In this paper first the data extraction of different result pages is done from different websites. The tag nodes are removed in order to obtain the data units. Secondly in second phase the annotations are obtained from different websites and these different annotations are integrated to form a combine list of annotation. For generating more accurate annotations, Naïve Bayes machine learning technique is used. Thirdly, the data, the value from SRRs is extracted and are aligned directly under semantic labels. In second phase, for obtaining the list of annotations, different annotators are used which are as follows:

1) Table Annotator

In table annotator, for many WDBs, SRR is represented by each row where there is table header signifies the semantics of a particular column.

2) Query-Based Annotator

In query-based annotator, the keyword submitted as a query by the user is always contained in the returned SRRs. For instance," machine" named query submitted by the user in the local search engine, will be contained by SRRs which are retrieved to this query. Hence, the term will be present in the field of Title of the retrieved SRRs. Then, this Title Keyword can be used as annotator which consist of "machine" keyword

3) Schema-Based Annotator

In Schema-Based Annotator, several attributes on search interfaces consist of prerecorded values on the interfaces. IIS tends to have many more attributes with than in LISs prerecorded values as when attributes when are unified from different multiple interfaces, their values are also unified. Such integrated value set are thus utilized to perform annotation in our Schema-Based Annotator.

4) Frequency-Based Annotator

In Frequency-Based Annotator, there are such attributes which are present in maximum retrieved SRRs with their individual different values, for example "Our Price" if occurs in maximum set of retrieved records then this attribute can be used to annotate its different values in different SRRs.

5) In-Text Prefix/Suffix Annotator

In In-Text Prefix/Suffix Annotator, this annotator inspects whether all the data units in the aligned group utilizes the same prefix or suffix. If same prefix is utilized then it is excluded from all data units and is used as an annotator. For example, if "You Save" prefix is shared by different SRRs then it's obvious that value after "You Save" attribute is price. Hence, this "You Save" attribute can be used for annotation.

6) Common-Knowledge Annotator

In Common-Knowledge Annotator, every common concept consist of a label and a set of patterns or values. For instance, Country label will be there for a country concept with a set of values Canada, U.S.A etc. The patterns of all the data units extracted if matches with the pattern or value of a concept then the label of this concept can be used to annotate these extracted data units i.e. Country attribute from our example can be used as a label for its values.

Thus, the results of above mentioned annotators are combined to annotate the multiple data units. In order to improve the performance of annotations Naïve Bayes classifier from MLT is used.

V. NAÏVE BAYES MLT

Naïve Bayes is one of the probabilistic classifier which is used in supervised learning that applies Bayes theorem to train the system, to learn from data. Naïve Bayes is one of the classification technique for text with its different applications. The different areas of application for Naïve Bayes are document categorization, email spam detection, sentiment detection etc. Despite of its simplified design and assumptions, it is widely used for many complex real-world problems. Naïve Bayes evaluates the probability of selected class, is used for decision making and the decision making is precise, hence the results are accurate. The main task of Naïve Bayes classifier is to assign the sample, the pre-defined label. For example, if the spam classifier is constructed then for an email representation from a feature space, its trains the system that whether to assign the email representation to "Spam" or "Non-Spam" label. It is used to accurately assign classes to the incoming representations accurately. This classification is a supervised machine learning, where the system is trained by analyzing the training data and generates an inferred function, such that with the help of this new examples are mapped. The determine the class labels accurately and correctly. For performing this type of learning and classification prior things needed to accomplished are to determine what kind of data is utilized as a set of training. Gather such training sets along with this the representation of an input feature is determined of the learned function and the learned function should be tested for its accuracy. Naïve Bayes can be used when limited resources are available. Naïve Bayes is a very popular algorithm because of its simplicity, good performance and its reckoning efficiency. Naïve Bayes algorithm has the ability to trail the system faster.

CONCLUSION

In this paper, a survey for different system that carried out data extraction and data annotation is carried out. Proposed System works for book domain and extracts dynamic data from multiple websites. Also the data annotation problem is studied for extracted data units along with proposed system uses multiannotator approach which is applied with their combined results in order to annotate the search result records and to assign meaningful labels to the data units extracted from SRRs. Each of the six annotators are used to elicit one unique type of feature. Different relationships are considered between text nodes and data units. In this paper the proposed system is carried out in three phases: Data Extraction, Data Annotation and Data Alignment. To obtain accurate holistic data annotation and data alignment Naïve Bayes Linear classifier from machine learning is used. The proposed system provide user i.e. redirects user to website if the user wants to shop the book. This system along with Naïve Bayes approach helps to achieve the objectives.

REFERENCES

[1] Y. Lu, H. He, H. Zhao, W. Meng, C.Yu, "Annotating search results from web databases," in IEEE Transaction on Knowledge and Data Engineering, Vol. 25, No.3, 2013.

[2]] N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Arti_cial 261 Intelligence (IJCAI), 1997.

[3] W. Liu, X. Meng, and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," IEEE Trans. Knowledge and Data Eng., vol. 22, no. 3, pp. 447-460, Mar. 2010.

[4] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf.,2001.

[5] S. Mukherjee, I.V. Ramakrishnan, and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents," Proc. IEEE Int'l Conf. Data Eng. (ICDE),2005.

[6] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.