ANNOTATING SEARCH RESULTS FROM WEB DATABASE USING IN-TEXT PREFIX/SUFFIX ANNOTATOR

K.Aparna[1], G.Murali[2]

[1]*M-Tech, Computer Science & Engineering, JNTUA College of Engineering Pulivendula, AP, India*
[2]*Assistant Professor, Computer Science & Engineering, JNTUA College of Engineering, Pulivendula, AP, India*

Abstract

Due to development of web search engines databases towards web reachable widely through HTML based search techniques it is important to extract the necessary information in the form of search result records SRR's and analysis of the data units returned from the web is also necessary. Commonly the data units which are retrieved from web reachable search engine database is normally prearranged into the end result pages vigorously for personage browsing. In this paper, considering automatic data units assignment used for SRRs pages returned from the novel search engine databases along with usage of in-text prefix/suffix annotation method. To overcome these problems we proposed an automatic semantic annotation method in the course of semantic comparison measure for text units and data unit's which results from features for the Search results records SRR's. The features of text and data units are obtained from Particle Swarm Optimization (PSO) techniques for each and every data unit semantic similarities based measurements are measures by extracting important features of the search result records SRR's. Semantic similarity between the terms in the web pages are measured by the system measures which are based on Ontology, after that aligns the data units in the efficient manner. In the present work we competently analysis the data units and most brilliant alignment of search result records. For the annotation of fresh search result from network search engines for different domains in databases we go to the usage of annotation wrapper.

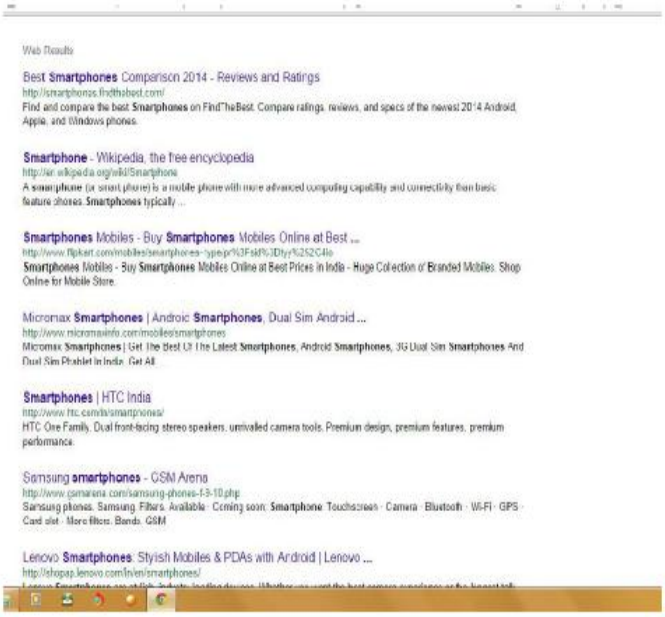
Keywords: *Data annotation, Data alignment, wrapper generation, web database.*

--***--

1. INTRODUCTION

Web is mainly based on the databases i.e., data encoded in the form of result pages for many search engines comes from the underlying databases. The databases from which the results are being extracted are known as the web databases. Because of the wide range of development in the search engines now a day's analysis of information in deep manner from database or web search engines is also important to receive exact data in search result web pages. For a result page returned from the web database consists of several search result records (SRR's). Specifically while searching the query term entered in the search attribute on the local search interface will most likely appear on the result page in form of search result records (SRR's). Here in the below figure search attribute is "Smartphone's". When we search with this key word on the local search interface the result will be as shown in the figure-1. If we observe the search result records that are present in the result page will appear most likely as the query term or the search attribute entered in the search interface. Here the searching is done based on the query term hence called as the query based search. Some times while searching with this keyword, the results that are related to the query may or may not be displayed on the result page. This is because, the web database consists of various results records with several names that are belonging to the same domain. But when the query term does not match the result record, it may or may not be displayed, hence results in loss of data even it is related to the query term entered in the local search interface. The below figure-2 is an example for the search that is based on the query. Here the query term is Software books. If we observe the database, there are three result records named 1.List of Software books, 2. Java Software books, 3. Software testing books. As the result records 1and2 matches directly with the query term Software books, they are displayed on the result page. In case of the third result record there is a word "testing "in between the query attributes, so it may or may not be displayed which leads to loss of date even it is related to the search query. By using the proposed techniques user can retrieve the data without any loss of data that is related to the query. Basically the SRR is a collection of data units, each SRR represent a book with several number of data units. The semantic labels of the data units are often not provided in the result page. We use annotation wrapper for the annotation of fresh search result from search engines designed for various domains in databases. Earlier applications are required terrific effort to manually annotate data units which results in the limitations in case of scalability.

The number of search result records (SRR) and each one data unit of the SRRs are related to separate individual concepts. Suppose if we consider reliance mart as an example in case of online web store the encoded HTML page consists of the information about the products. Generally every SRR consists of several data units such as product name, brand name, price, company, etc. Normally, n every data unit in the SRRs is not encoded in the semantic manners or meaningful manner. To overcome this problem previous research introduced an efficient algorithm which automatically interprets the data units present in the SRRs pages through Web database. But those results does not investigate search sites of the SRR's with the intention that encloses web service interfaces, since the exact semantic meaning of data unit labels are most efficiently described in WSDL.



Web Results

Best **Smartphones** Comparison 2014 - Reviews and Ratings
http://smartphones.findthebest.com/
Find and compare the best **Smartphones** on FindTheBest. Compare ratings, reviews, and specs of the newest 2014 Android, Apple, and Windows phones.

Smartphone - Wikipedia, the free encyclopedia
http://en.wikipedia.org/wiki/Smartphone
A smartphone (or smart phone) is a mobile phone with more advanced computing capability and connectivity than basic feature phones. **Smartphones** typically ...

Smartphones Mobiles - Buy **Smartphones** Mobiles Online at Best ...
http://www.flipkart.com/mobiles/smartphones~type/pr%3Fsid%3Dtyy%252C4io
Smartphones Mobiles - Buy **Smartphones** Mobiles Online at Best Prices in India - Huge Collection of Branded Mobiles. Shop Online for Mobile Store.

Micromax **Smartphones** | Android **Smartphones**, Dual Sim Android ...
http://www.micromaxinfo.com/mobiles/smartphones
Micromax **Smartphones** | Get The Best Of The Latest **Smartphones**, Android **Smartphones**, 3G Dual Sim **Smartphones** And Dual Sim Phablet In India. Get All

Smartphones | HTC India
http://www.htc.com/in/smartphones/
HTC One Family. Dual front-facing stereo speakers, unrivalled camera tools. Premium design, premium features, premium performance.

Samsung **smartphones** - GSM Arena
http://www.gsmarena.com/samsung-phones-f-9-10.php
Samsung phones. Samsung. Filters. Available · Coming soon. **Smartphone** · Touchscreen · Camera · Bluetooth · Wi-Fi · GPS · Card slot · More filters. Bands. GSM

Lenovo **Smartphones**: Stylish Mobiles & PDAs with Android | Lenovo ...
http://shopap.lenovo.com/in/en/smartphones/

Fig -1: Example of result page with SRR's

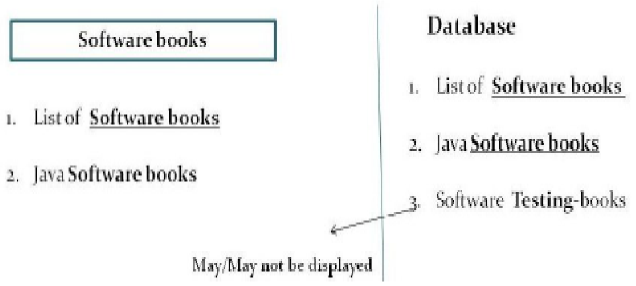


Fig 2: Illustration of query based search

2. LITERATURE SURVEY

Many Web sites contain a wide collection of "ordered" or structured web pages. These pages encode facts from an original structured resource, and are classically generated dynamically. Extracting the structured data units from the wide collection of the web pages as stated above is our goal, without providing any human inputs such as manually generating rules and training sets etc. Extracting the structured information provides us higher querying command over the data and also it is useful in information integration systems. Our approach is of two stages. In the initial stage, the template that is not known and is used to create the web pages is deduced. In the second, the above deduced template is used for extracting the values. We focus on the initial stage as it is more challenging. The complete version consists of formal meaning of high occurrence correlation and the algorithm. We go on evaluating our approach by considering the real collections of the web pages [1].

Data extraction process from web pages is performed with the help of software modules called as wrappers. Recently, few systems for generating wrappers automatically have been introduced in the previous research. These systems are based on unconfirmed inference techniques: taking small set of sample pages as input, they can generate an ordinary wrapper that extract related information. Because of the automatic nature of this approach, the information extracted by using these wrappers has unspecified names. In the present framework of the ongoing project named Roadrunner, we have developed a prototype, named Labeler which automatically annotates data units extracted by the wrappers that are generated automatically. Although Labeler is using as a companion system for our wrapper generator, its fundamental approach has a universal validity and also it can be applied together with the other wrapper generating systems [2].

The quantity of the useful semi-structured information on the web going on increasing at a stunning pace. Often the interesting web data units are not present in the database systems but in XML pages, HTML pages, or text files. Data that is present in these formats cannot be directly used by standard SQL such as query processing engines which support complicated querying along with reporting beyond keyword-based retrieval. As the result the web users or applications finds a smart way of extracting information from these web resources. One of the most popular approaches is writing wrappers around sources, either manually or with the help of software assistance, to get the web data within the reach of more complicated query tools and common mediator-based data integration systems [3].

3. FUNDAMENTALS

3.1 Data Unit and Text Node

Each SRR that is extracted from the web has a tag structure which determines how the contents of the SRRs must be displayed on a web browser. Every node in the tag structure will be either tag node or a text node.

3.2 Features of Data Unit and Text Node

3.2.1 Data Content

This indicates that the content or the information present in two data sets may be same. In this case the data nodes match's with same keywords while searching. Sometimes the developers place the keywords those are popular in front of the data set to make the search easier to the user.

3.2.2 Data Type

Every data has its own semantic style even it is just a common text string in HTML. Data type may be of textual representation, pictorial, graphs, charts, diagrammatical.

3.2.3 Representation Style

It indicates that how the data units are presented on the search result record on the webpage. It consists of style features such as *font face, font face, font color, font weight, text decoration.*

4. BASIC ANNOTATORS

- Table Annotator
- Query-Based Annotator
- Schema value Annotator
- Frequency based Annotator
- In-text prefix/suffix Annotator
- Common Knowledge Annotator

Table -1

Annotator	Applicability	Success Rate
Table Annotator	6%	1.0
Query-Based Annotator	35%	0.95
Schema value Annotator	41%	0.5
Frequency based Annotator	14%	0.86
In-text prefix/suffix Annotator	7%	0.85
Common Knowledge Annotator	25%	0.84

Table -1 explains the applicability and the success rate of various annotators. In this proposed work we go through the in-text prefix/suffix method to get efficient results.

5. EXISTING & PROPOSED SYSTEMS

Additional to all the annotation approaches we have concentrated on the search procedure, wrapper generation, and data alignment. As mentioned in the Fig-2 the existing system search is based on the query based method, in this paper along with the query based search we are using the intext prefix/suffix annotator which compares the prefix and suffix of the query term entered in the personal search interface with the terms in the web database those are matching with the query word entered. Along with this we use also propose Measurement of semantic similarity to each and every data, text unit nodes. Alignment of data units, search result records in efficient manner. We use Particle Swarm Optimization (PSO) method to obtain features of data and text units

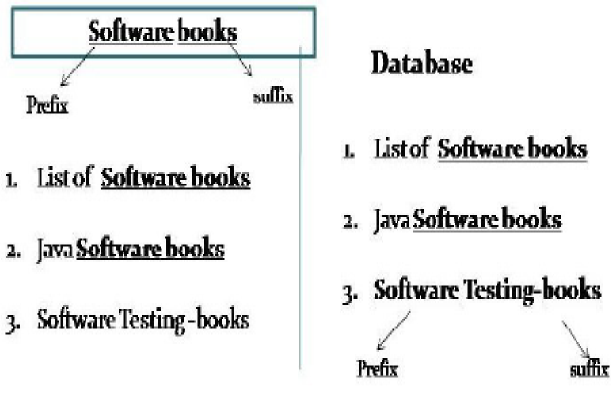
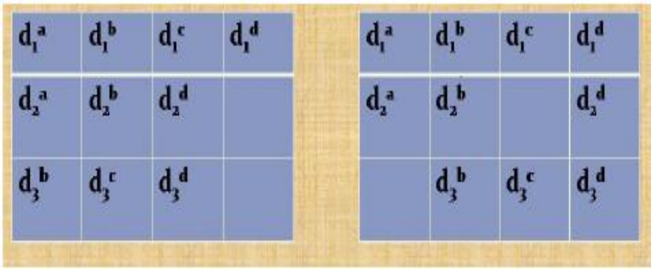

Software books
Prefix
suffix
1. List of Software books
2. Java Software books
3. Software Testing -books
Database
1. List of Software books
2. Java Software books
3. Software Testing-books
Prefix
suffix

Fig -3: Example to Explain in-text prefix/suffix method.

The above figure represents the search procedure by using the in-text prefix/suffix method. With the previous methods data that is related to the search query may be lost while searching. This is due to the additional text that is present between the search queries in the web database. By the proposed method user can get the exact results for the query attribute because it compares the prefix and the suffix of the search key word with the related data units in the web database.

5.1 Alignment Representation

The above figure-4 represents the proposed work for alignment of data units, search result records in the efficient manner.



d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b	d_2^d	
d_3^b	d_3^c	d_3^d	

d_1^a	d_1^b	d_1^c	d_1^d
d_2^a	d_2^b		d_2^d
	d_3^b	d_3^c	d_3^d

Fig -4: Alignment procedure

In the above figure the left table represents the data units those are not aligned in the specific manner, if we observe after the arrangement of the data units d_1^a, d_2^a, in the place of d_3^a, d_3^b has been placed which may leads to the difficulty which searching the data unit. In order to avoid these problems efficient alignment procedure has been proposed. Here after arrangement of d_1^a, d_2^a if the third data unit of “a” is not present we keep the cell empty without filling it with any other elements. This procedure helps us to avoid the problems facing while searching a data unit in the huge databases.

5.2 Practical Swan Optimization

Practical swan optimization is a computational technique which optimize a problem by trying iteratively to improve the candidate result with regard to given measure of quality. PSO optimizes a problem by having population of candidate results. When Compared with genetic algorithms the advantage of PSO is that PSO is very easy to implement and there is a need of few parameters that are to be adjusted . PSO has been successfully applied in various areas: function optimization, fuzzy system control, artificial neural network training, and other areas where genetic algorithms can be applied.

5.3 Data Alignment

Data alignment can be carried out by comparing the following terms,

- Data unit similarity
- Data content similarity
- Presentation style similarity
- Data type similarity

Basically the data is grouped based on the above, while searching this alignment procedure helps us to find the results for the related search attribute very accurately and efficiently.

5.4 In-Text Prefix/Suffix Annotator

In few cases, a piece of data unit is encoded with the help of its label to form a single unit without having any separator between the label and value, but it contains both the value and the label. Those nodes may occur in few SRRs. In the previous search techniques the searching is done based on query term entered in the local search interface. Here we proposed a technique for searching by the use of in-text prefix/suffix annotation method.

5. CONCLUSION

In this paper we go through feature extraction results for SRRs that is based on a practical swan optimization that extract individual data and text unit nodes individually. Here we use in-text prefix/suffix annotation method in order to extract the results in the efficient manner for the given query. Proposed method every self concerned server alternates data units are based on semantic similarity measure which results by using ontology and different data unit's relationship measures whereas hard to achieve a compromise. The returned result pages as the best data alignment and then perform the annotation approach using wrapper methods for search results returned from whichever specified web database. Experimental results says that proposed method based feature extraction with proficient semantic similarity measurement best data alignment is supportive and they simultaneously are proficient of creating high-quality explanation of several web databases in the equivalent field.

REFERENCES

[1]. Arasu and H. Garcia-Molina, “Extracting Structured Data from Web Pages,” Proc. SIGMOD Int’l Conf.Management of Data, 2003.

[2]. L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, “Automatic Annotation of Data Extracted from Large Web Sites,” Proc. Sixth Int’l Workshop the Web& Databases (WebDB), 2003.

[3]. L. Liu, C. Pu, and W. Han, “XWRAP: An XMLEnabled Wrapper Construction System for Web Information Sources,” Proc. IEEE 16th Int’l Conf. Data Eng. (ICDE), 2001.

[4]. N. Krushmerick, D. Weld, and R. Doorenbos, “Wrapper Induction for Information Extraction,”Proc. Int’l Joint Conf. Artificial Intelligence (IJCAI),1997.

[5]. W. Meng, C. Yu, and K. Liu, “Building Efficient and Effective Metasearch Engines,” ACM Computing Surveys, vol. 34, no. 1, pp. 48-89, 2002.

[6]. Z. Wu et al., “Towards Automatic Incorporation of Search Engines into a Large-Scale Meta search Engine,” Proc. IEEE/WIC Int’l Conf. Web Intelligence (WI ’03), 2003.

[7]. Cohen, W., Hurst, M., and Jensen, L.” A flexible learning system for wrapping tables and lists in HTML documents”,WWW-2002, 2002.

[8]. Kushmerick, N,” Wrapper induction: efficiency and expressiveness”, Artificial Intelligence, 118:15-68, 2000.

[9]. Lerman, K., Getoor L., Minton, S. and Knoblock, C. “Using the Structure of Web Sites for Automatic Segmentation of Tables.” SIGMOD-04, 2004.

[10]. Pinto, D., McCallum, A., Wei, X. and Bruce, W. “Table Extraction Using Conditional Random Fields”, SIGIR-03.

[11]. Buttler, D., Liu, L., Pu, C ,”A fully automated extraction system for the World Wide Web “, IEEE ICDCS-21, 2001.

[12]. Liu, B., Grossman, R. and Zhai, Y. “Mining data records from Web pages.” KDD-03, 2003.

[13]. W. Bruce Croft ,” Combining approaches for information retrieval”, In Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic,2000.

[14]. V. Crescenzi, G. Mecca, and P. Merialdo ,” Road RUNNER: Towards Automatic Data Extraction from Large Web Sites “, VLDB Conference, 2001.