International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 02 Issue: 05 | Aug-2015 www.irjet.net p-ISSN: 2395-0072

Feature Based Annotating Results In Content Based Retrieval System

Marina Glastin,Priya Sekhar

Mtech ,Computer Science And Engineering Lourdes Matha College Of Science And Technology,Kerala,India

Asso professor,Computer Science and Engineering,Lourdes Matha College Of Science And Technology,Kerala,India

---

Abstract - ***Most of the search engines today are called Web databases. Large databases have now become web accessible through html form based interfaces. E-commerce and digital library are good examples for that. The user enters a query term and the result records are returned from the underlying databases. The result pages returned from web databases contain data units which are usually encoded into the result pages dynamically for human browsing. Each data unit in the result records corresponds to a real world entity and for many applications these data units should be assigned meaningful labels so that they become machine processable . Annotations helps in retrieving the relevant information efficiently.A number of techniques has been developed for extracting and annotating the information. Annotations make the search meaningful.All the techniques developed for annotation have certain limitations .Another important concept used here is wrapper generation.Wrapper generations so far developed were used only for data extraction which is their main purpose and not for annotation.In order to overcome those limitations an automatic wrapper generation approach is proposed here.Intially the data units will be aligned according to certain features then automatic annotation is performed and finally wrapper generation is done.Once generated these wrappers can automatically annotate the future results from the web database. Tree alignment and clustering methods are used for wrapper generation*** **.A linear regression method is used for getting different weights of tag matching.** ***The data of interest collected by an individual from various web databases can be effectively annotated using the proposed approach. The performance analysis measures shows that this method has higher success rate.***

Key Words: Annotation, Wrapper Generation, Data Organisation,Web Databases

1.INTRODUCTION

Internet provides a huge amount of information.We need search engines which are very important tools for accessing the information on the world wide web (WWW).A large portion of deep web is database based and such search engines are called Web databases.

. Each result page returned from a WDB has numerous search result records (SRRs) and each result records includes multiple data units. The data unit is a piece of text which corresponds to a real world entity.[1]. The data units has to be extracted out and annotated so that they become machine processable.

A number of approaches have been developed for annotation. Initially labeling was done manually, which is time consuming and errors occurred frequently. Next came semi automatic annotation approach,which lacks scalability.And finally automatic annotation approach came in to exist. For performing fully automatic annotation, the result pages have to be automatically obtained and the SRRs need to be automatically extracted. web data extraction can be classified into three categories: 1) Wrapper programming languages, 2) Wrapper induction, and 3) Automatic extraction.

1) Wrapper programming languages

This approach uses the special pattern specification languages which help the user to develop extraction programs.

2) Wrapper induction method

This method is useful in systems where the resource information is formatted for use by people and so it is difficult to extract their content mechanically.

3)automatic extraction methods

To overcome the problems of wrapper induction automatic wrapper generation was developed which uses unsupervised learning. An automatic annotation wrapper is proposed in the paper. Information extraction is an important application of Data Mining.

2.EXISTING SYSTEM

. web data extraction and data annotation is a prime research area in the web database Extracting structured data from deep web pages is a challenging problem [2].Some of the limitations are webpage programming dependent, Incapable of handling ever increasing complexity of HTML source code. To overcome this problem- a vision based approach [3] vision based data extractor. ViDE is used to extract structured results from deep WebPages automatically. It can only process deep web pages containing one data region while there is significant number of multi-data region deep WebPages, which is time consuming process.[4]ODE which automatically extracts the query results records from the HTML pages. Automatic data extraction is important for many applications such as meta-querying, data integration and data warehousing. In semi automatic wrapper induction has the advantage that no extraneous data are extracted as the user can label only the data in which he/she is interested. To overcome this supervising learning methods are used [5]. Labour intensive and time consuming are drawback and also it is not scalable to a large number of websites.[Technique for extracting data from HTML sites through the use of automatically generated wrappers. A key problem with the manually coded wrappers is that writing them is usually a difficult and labour intensive task and difficult to maintain [6] has a prior knowledge about the page contents. It is an daunting task for users to access numerous web sites individually to get the desired information.[7] is a tool that perform automatic integration of web interfaces of search engines. It is to identify matching attributes. [8] ViNTS is automatically producing wrappers that can be used to extract search result records dynamically. It utilizes both the visual features on the result page displayed on browser and HTML tag structure of the source file. It helps people to locate and understand information. Existing approaches use decoupled strategies [9]. A probabilistic model to perform two tasks simultaneously. HCRF can effectively integrate all useful features by learning their importance.

3.PROPOSED SYSTEM

In this system initially user enters the query term in the form based search interface. As a result of this many search result records will be returned from the underlying databases and each SRRs contain numerous data units.The data units corresponding to the same concept often share special common features. After the feature selection data alignment is done. The main aim of alignining the data units is to put the data units corresponding to same concept into one group so that they can be annotated easily .

There are six basic annotators [1] to label data units, with each of them considering a special type of patterns/features. After annotation, the annotated data units are used to construct an annotation wrapper for the WDBs. So that the new SRRs retrieved from the same WDB can be annotated using this wrapper quickly without reapplying the entire process. Finally display the results.

Alignment algorithm is performed to move the data units in the table into well aligned alignment groups and ensures that the order of the data units within each result record is maintained and also needs the similarity between two data unit groups. Here initially create the alignment groups then clustring is performed. Then select the annotation method.Six annotation methods are available for selection. First select table annotator and perform the table annotation.Table annotators first identify the column header. Then for each SRR takes a data unit. Then select the column header with maximum overlap. At last a unit is assigned and labelled. Then query based annotation is done, in this first set of query terms are there. From that find the group with largest occurrences and label is assigned. In the schema annotator attribute is identified with highest matching score. In the frequency annotator find the common preceding units then concatenated preceding units and label the group. In text prefix/suffix annotator check the data units and share the same prefix or suffix. The common knowledge annotator from the group of data units match the patterns or values and label the group. Finally wrapper is generated

Pseduocode

Collect dataset

Clustering

Choose the annotation method

If annotation=table annotator

Perform table annotator

Elseif

If annotation=query based annotator

Perform query annotator

Elseif

If annotation= schema based annotation

Perform schema annotation

Elseif

If annotation=frequency based annotation

Perform frequency annotation

Elseif

If annotation=intext prefix/suffix annotator

Perform intext prefix/suffix annotation

Elseif

If annotation =common knowledge annotator

Perform common knowledge annotation

Elseif

Wrapper

end

3.1 Wrapper Generation

1) In this system tree alignment methods are used to calculate the similarity between input web pages and then build a wrapper on that results.The input trees are merged into one union tree whose nodes record the statistical information such as the times a node has been aligned, the text length of the node.The alignment algorithm is utilized again to detect the repeating patterns on the union tree. The wrapper is generated based on the most probable content block and the repeating patterns.

(2) A similarity series was built by calculating the similarity between the input web pages and the current wrapper using the tree alignment algorithm.

(3). A log likelihood ratio test is utilized to detect the change points on the similarity series. The wrapper generation method is applied again to generate a wrapper once a change point is detected.

3.2 Automatically getting tag-matching weight

A kind of linear regression method is employed to get the weight of various tag matching.The block elements are elements that sometimes, contain other elements. They normally act as containers of some sort. The inline elements mark up the semantic meaning of something. Furthermore, the level of the different nodes are also considered. The higher-level nodes should have higher weight as they act as bigger structure block. Different weight should be assigned to different type of tag-matching.

.Intially we found collection of similar web pages belong to the same "class". It's possible to get this kind of web pages collection automatically.Then we can use this collection for getting the weighting schema which is optimal.

Let w_i be the weight of tag-matching and $w_i > w_j$ for $i < j$.

Let D_{mn} be the sum of the gains in the best alignment between the trees T_m and T_n.

$$D_{mn} = \sum_i w_i t_i^{mn}$$

(1) Where t_i^{mn} is the number of wi occur in the alignment procedure.
(2) The sum of the gains in the collection is:

$$f = \sum_{m,n} D_{mn} = \sum_{m,n} \sum_i w_i t_i^{mn} = \sum_i w_i \sum_{m,n} t_i^{mn}$$

Because the collection is the similar web pages belonging to the same "class", a set of w_i is selected which makes the maximum f.

To get $argmax_w \sum_i w_i \sum_{m,n} t_i^{mn}$, a constraint $\sum_i w_i^2 = 1$ is added.

The group of equations is rewritten as:

$$f = \sum_i w_i C_i + \lambda\left(\sum_i w_i^2 - 1\right), \quad C_i = \sum_{m,n} t_i^{mn}, \quad \sum_i w_i^2 = 1$$

The solution of the above equations can be used as the weight of tag matching(wi).

The best weighting schema is the one maximize the sum of the gains. That means to find a set of $w_i$that output the maximum f in the equations

3.3 proposed system architecture

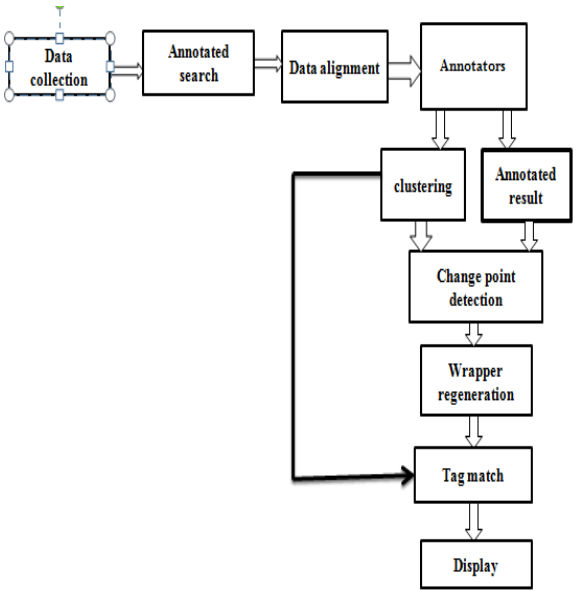


Fig 1:Wrapper generation architecture

4.PERFORMANCE ANALYSIS

We have made experiments data from various domains with respect to six annotators only. The annotators used include table annotator , schema value annotator, prefix suffix annotator, frequency annotator common knowledge annotator and query – based annotator. Both the annotators are supported by the prototype application and it is extensible so as to support more annotators in future. The performance of data alignment and annotation are presented in Table 1.

Table 1:Performance analysis table

Domain	Data Alignment performance		Annotation Performance	
	Precision	Recall	Precision	Recall
Mobile	97.2%	97.4%	93.5%	90.2%
Laptop	97.1%	97.2%	92.6%	91.3%
Camera	96.8%	97.0%	92.6%	91.3%

5.EXPERIMENTAL RESULTS

As presented in Table 1, it is evident that more than 90% precision and recall were recorded for both the performances such as data alignment and annotations. The results are presented in the following graphs. Performance of data annotation with wrappers for three domains is capable of producing annotations automatically given search results of Google. The performance of the application is encouraging and the application can be used in the real world applications
Category List

1. Query-Based Annotator (QA)
2. Schema Value Annotator (SA)
3. Frequency-Based Annotator (FA
3. In-Text Prefix/Suffix Annotator (IA)
4. Wrapper Annotation (WA)

Existing system: Manual training
Proposed system: Our Project

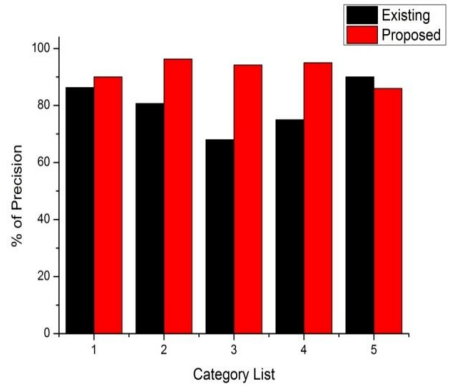
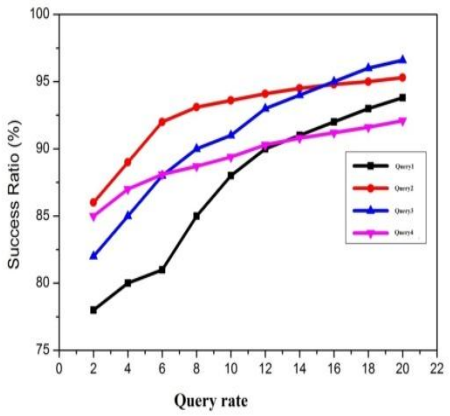


Chart 1 Precision chart



Graph 1:Success ratio of annotators

6.CONCLUSION

An automatic wrapper generation approach was proposed.Annotation wrapper generated can automatically extract the data units and annotate the search results from the WDB.The proposed method reduces the search time required for obtaining relevant information.

REFERENCES

1. Yiyao Lu, Hai He, Hongkun Zhao, WeiyiMeng "Annotating Search Results from Web Databases "IEEE Transaction on Knowledge and Data Engineering, Vol. 25, No.3, March 2013
2 S. Handschuh, S. Staab, and R. Volz, "On Deep Annotation," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.
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. W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2, article 12, June 2009.

5. N. Krushmerick, D. Weld, and R. Doorenbos, "Wrapper Induction for Information Extraction," Proc. Int'l Joint Conf. Artificial Intelligence (IJCAI), 1997.

6. L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. IEEE 16th Int'l Conf. Data Eng. (ICDE), 2001.

7.H. He, W. Meng, C. Yu, and Z. Wu, "Automatic Integration ofWeb Search Interfaces with WISE-Integrator," VLDB J., vol. 13, no. 3, pp. 256-273, Sept. 2004.

8. H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, "Fully Automatic Wrapper Generation for Search Engines," Proc. Int'l Conf. World Wide Web (WWW), 2005

9. D. Embley, D. Campbell, Y. Jiang, S. Liddle, D. Lonsdale, Y. Ng, and R. Smith, "Conceptual-Model-Based Data Extraction fromMultiple-Record Web Pages," Data and Knowledge Eng., vol. 31