Volume 4, Issue 1, January 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Map Reducing for Annotation of Search Results from web Databases

Saradha.S[1]
PG Student,
Department of CSE & Anna University,
Sri Eshwar College of Engineering, India

Aravindhan.R[2]
Assistant Professor,
Department of CSE & Anna University,
Sri Eshwar College of Engineering, India

Abstract-- The HTML form-based interfaces make a large number of database web accessible. Whenever a user submits a query to the search engine, data units are retrieved from the underlying databases. These data units have to be extracted and assigned meaningful labels for the effective use of machine processable applications like comparison web shopping and deep web data collection. This paper focuses on the multi-annotator approach which first aligns the data units in the result page into groups of the same semantic, then various annotations are performed for each group and these annotations are combined to predict the final annotation. An annotation wrapper is constructed automatically which can be used for the new queries. In addition we propose map reduce concept for efficient processing of large data.

Keywords: Web Database, Data Extraction, Data Annotation, Data Alignment, Annotation Wrapper Generation

I. INTRODUCTION

The search engine that fetches the results from the underlying structured databases and displays in the result page will be referred to as the Web Databases (WDB) in this paper. A result page returned from a WDB contains multiple search result records (SRR). Each SRR consists of multiple data units (or instances). Each data unit refers to a single concept of an entity. A text node consists of a piece of text surrounded by a pair of HTML tags. It is different from the data units referred in this paper. This paper focuses on the data unit level annotation. Annotating data units refers to assigning meaningful labels. The data units in Fig.1 are book title, author, publisher and price.

Annotation of web pages is necessary for applications such as comparison book shopping, deep web collection etc. For instance, a book comparison system collects various search records from multiple different websites and it finds whether two records points to the same book. This kind of automatic comparison can be easily done if the data units of the search results are assigned with meaningful labels. Also annotation is also essential for easier storage of data into tables and for quick retrieval or mining of data. The result page consists of many search result records. Once the search result records have been extracted from the result page, annotation of them is performed in three different phases. The first phase is the *alignment phase* which organizes the data units into groups of the same semantic. The second phase is the *annotation phase* which performs different types of annotations for the assignment of labels. The third phase is the *annotation wrapper generation phase* which constructs annotation wrapper. The wrapper is a set of rules for all aligned groups of data. This wrapper can be used for the new queries without repeating the whole process again.

Fig 2 shows the illustration of the three phases. In the figure, *i*th SRR of the *j*th concept belonging to the data unit is represented as d_i^j. Fig.2a shows the table representation of the result page with each row representing the SRR. Fig.2b shows the phase in which each column contains the data unit of the same concept. Fig.2c shows the assignment of semantic label L^j to each column. In Fig.2d rule Rj are generated for each concept.This paper focuses on the relationship between the data units and the text nodes. It also uses the cluster based shifting technique for the alignment of data units into various groups of the same semantic. To enhance the data unit level annotation integrated interface schema is used. Then six basic annotators are used for assigning labels based on some features. Once the labels are assigned, annotation wrapper is constructed for efficiently using the new queries

II. RELATED WORKS

H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu. describes an Automatic Wrapper Generation System called ViNt[16] for extraction of SRRs from web pages.

Hai He, Weiyi Meng, Hongku Zhao, Clement Yu. [9] describes the multi-annotator approach for constructing annotation wrapper. Our approach in this paper is similar to the approach referred here. It uses six different types of annotation and constructs annotation wrapper. Annotation here refers to the assignment of meaningful labels.

Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng [13] is the extension of the multi-annotator approach followed in [9]. It explains the relationship between the text nodes and the data units. It also enhances the alignment and the cluster-shifting algorithm to improve the efficiency. Arlotta et al. [2] basically annotate data units with the closest labels on result pages. However, this method has very limited applicability because many Web sites do not encode data units with

© 2014, IJARCSSE All Rights Reserved

their labels on result pages. DeLa [11] uses several heuristics to assign labels to the extracted SRRs. Also DeLa uses the local Interface Schema whereas our approach differs from it by using the Integrated Interface Schema.

H. He, W. Meng, C. Yu, and Z. Wu [7] describes the WISE Integrator approach for automatic integration of search queries from different websites.

Fig.1 Sample Search Results

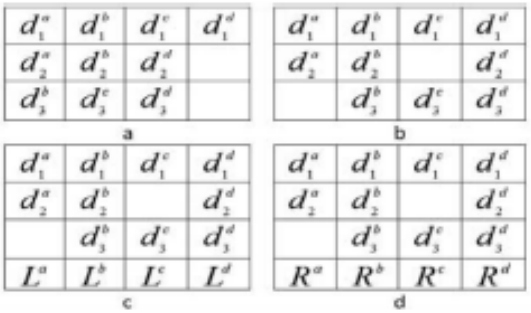

Fig.2 Illustration of three phases

III. OVERVIEW OF THE SYSTEM

This system architecture of the existing system can be broadly classified into three major phases: Alignment phase, Annotation phase, Annotation Wrapper Generation phase. The overview of our approach is shown in fig.3.

a. Data Record Extraction

For the annotation to be done, the data records or the SRRs have to be extracted from the result pages. That is the irrelevant information like advertisements, links, information about the hosting site are to be discarded from the result page. Manually writing programs to extract the records from the result page is laborious, time consuming and impractical since the search engine change the display of the result page time to time. So, a system called **ViNTs** (**V**isual **i**nformation a**N**d **T**ag **s**tructure based wrapper generator) described in [16] is employed for extracting the search records from the results page. ViNt system has its own architecture for extracting SRRs. The data extraction through ViNt is bases on the visual content features of the web page and also the HTML tag structure which is nothing but the source file of the page in HTML format.

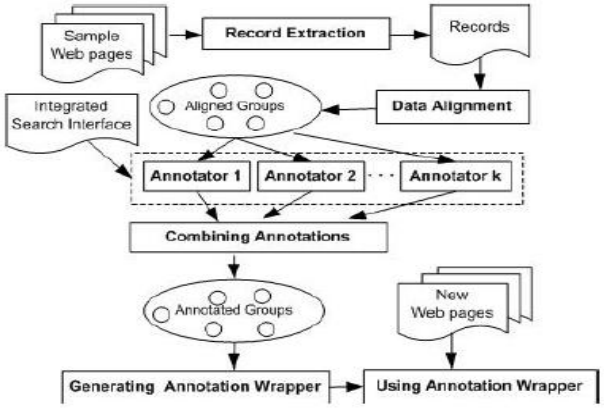

Fig.3 Overview of the approach

© 2014, IJARCSSE All Rights Reserved

b. Data Alignment

The data units are not aligned whenever the search result records are extracted from the web page. The main purpose of data alignment is to group the data units from different records into the semantically same group. This alignment of data records facilitates easier annotation of data. It is based on the assumption that the data units in different SRRs of the same semantic usually have the fixed layout and presentation. Based on this assumption a record expression (REXP)[9] is constructed for each result record. An REXP is a string comprising of sequence of symbols that represents either the presentation style of the node or the separator/ delimiter. For example, in our current implementation of REXP, \S" denotes a pure text node with bold style, \s" denotes a pure text node without bold style, \L" denotes a link node with bold style, \l" denotes a link without bold style, \^" denotes starting a new line, etc. Separators are nodes that contain only non-letter and non-digit characters appearing in HTML text. The REXP for each SRR can be constructed easily.

Example 1: The REXP of the first record in Figure 1 is \l^ss/s/s/s^sS»s^s", where \=" and \»" are the separators appearing in HTML text of the record. Note that as shown the text \Peter J. Denning" and a \=" are en-coded together, so the first \s" in the REXP represents \Peter J. Denning=".\Put in Basket" is not included because buttons, icons and images are currently ignored. The last \s" represents \Out-Of-Stock".

With this Record expression a suffix tree is constructed. Then the most common longest string (MCLS) is selected and its corresponding components are aligned to form groups. The data Alignment also concentrates on the Data Unit Similarity, Data Content Similarity, Presentation Style Similarity, Data Type Similarity, Tag Path Similarity as described in [13]. Also for improving the efficiency of data grouping and aligning, Alignment and cluster-shifting technique [13] is also used. The algorithm concentrates on four steps namely Merging text nodes, Aligning text nodes, Splitting Composite Text nodes and Align Data Units. The alignment algorithm is shown in figure 4.

c. Data Annotation

The data annotation is based on the concept that the data units corresponding to the same attribute always share some common features. These common features are the basis of our annotators. We introduce multiple basic annotators, each annotator is used for identifying a specific feature. Every basic annotator is used to produce a label for the units within their group. There are six basic annotators used for annotating the database namely Table annotator, Query- based annotator, Schema annotator, Frequency- based annotator, Prefix/Suffix annotator and Common annotator as described in [9]. Different annotations are combined to predict the final annotator.

```
ALIGN(SRRs)
1.   j ← 1;
2.   while true
       //create alignment groups
3.     for i ← 1 to number of SRRs
4.       G_i ← SRR[i][j];     //j^th element in SRR[i]
5.     if G_i is empty
6.       exit; //break the loop
7.     V ← CLUSTERING(G);
8.     if |V| > 1
         //collect all data units in groups following j
9.       S ← Ø;
10.      for x ← 1 to number of SRRs
11.        for y ← j+1 to SRR[i].length
12.          S ← SRR[x][y];
         //find cluster c least similar to following groups
13.      V[c] = min_{k=1..|V|} (sim(V[k], S));
         //shifting
14.      for k ← 1 to |V| and k ≠ c
15.        foreach SRR[x][j] in V[k]
16.          insert NIL at position j in SRR[x];
17.  j ←j+1;          //move to next group
CLUSTERING(G)
1.   V ←all data units in G;
2.   while |V| > 1
3.     best ← 0;
4.     L ←NIL; R ←NIL;
5.     foreach A in V
6.       foreach B in V
7.         if ((A != B) and (sim(A, B) > best))
8.           best ← sim(A,B);
9.           L ←A;
10.          R ←B;
11.    If best > T
12.      remove L from V;
13.      remove R from V;
14.      add L ∪ R to V;
15.    else break loop;
16.  return V;
```

Fig 4 Alignment and Cluster- Shifting Algorithm

d. Integrated Interface Schema

Integrated Interface Schema uses the Web Interfaces of Search Engine (WISE) Integrator Approach as described in [7] for automatically integrating the queries from different databases. It reduces the users difficulty of accessing numerous sites for searching a query. It provides unified access to multiple search engines of the same domain for easier comparison of products.

© 2014, IJARCSSE All Rights Reserved

e. Automatic Wrapper Generation

Once the data units have been annotated, annotation wrapper is constructed as in[13]. Annotation wrapper consists of certain rules that can be used for new queries without repeating the entire process again., V

IV. CONCLUSION AND FUTURE WORK

Annotating or analysing large data in a single website may lower the processing speed. Our future work is to implement the concept of map reducing to the existing approach to improve the processing in large database. It is a technique to reduce the list of data's from the SRR. It is used to split the large number of datasets into the small set. The map reducing technique is used to filter and sort the analysed data.

REFERENCES

[1] A. Arasu and H. Garcia-Molina. (2003) '*Extracting Structured Data from Web Pages,*' Proc. SIGMOD Int'l Conf.
Management of Data.

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

[3] P. Chan and S. Stolfo (1993) '*Experiments on Multistrategy Learning by Meta- Learning,*' Proc. Second Int'l Conf. Information and Knowledge Management (CIKM).

[4] S. Handschuh, S. Staab, and R. Volz (2003)*'On Deep Annotation*' Proc. 12th Int'l Conf. World Wide Web (WWW).

[5] S. Handschuh and S. Staab (2003) '*Authoring and Annotation of Web Pages in CREAM,*' Proc. 11th Int'l Conf. World Wide Web (WWW).

[6] B. He and K. Chang,(2003) '*Statistical Schema Matching Across Web Query Interfaces*,' Proc. SIGMOD Int'l Conf.
Management of Data.

[7] H. He, W. Meng, C. Yu, and Z. Wu (2004) '*Automatic Integration of Web Search Interfaces with WISE-Integrator,*'
VLDB J., vol. 13,no. 3.

[8] H. He, W. Meng, C. Yu, and Z. Wu (2005) '*Constructing Interface Schemas for Search Interfaces of Web Databases*' Proc. Web Information Systems Eng. (WISE)Conf.

[9] Y. Lu, H. He, H. Zhao, W. Meng, and C. Yu (2007)*'Annotating Structured Data of the Deep Web*,' Proc. IEEE 23rd
Int'l Conf. Data Eng. (ICDE)

[10] J. Wang, J. Wen, F. Lochovsky, and W. Ma (2003) '*Instance-Based Schema Matching for Web Databases by Domain-Specific Query Probing,*' Proc. Very Large Databases (VLDB) Conf.

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

[12] Z. Wu et al (2003) '*Towards Automatic Incorporation of Search Engines into a Large-Scale Metasearch Engine*,'Proc.IEEE/WIC Int'l Conf.WebIntelligence.

[13] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng(2013)' *Annotating Search Results from WeI7 b Databases'*, IEEE Transactions on knowledge and data Engineering, vol. 25, no. 3.

[14] O.Zamir and O. Etzioni (1998) '*Web Document Clustering: A Feasibility Demonstration*,' Proc. ACM 21st Int'l
SIGIR Conf. Research Information Retrieval.

[15] Y. Zhai and B. Liu(2005) '*Web Data Extraction Based on Partial Tree Alignment*,' Proc. 14th Int'l Conf. World Wide Web (WWW '05).

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

[17] H. Zhao, W. Meng, and C. Yu (2007)*'Mining Templates form Search Result Records of Search Engines,'* Proc. ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining.

[18] J. Zhu, Z. Nie, J. Wen, B. Zhang, and W.-Y. Ma (2006) '*Simultaneous Record Detection and Attribute Labeling in Web Data Extraction*,'Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining.

© 2014, IJARCSSE All Rights Reserved