ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

Annotating Search Results from Web Databases

Prof. P. S. Badgujar[1], Shagufta Pathan[2], Aayesha Sayyed[3], Salina Ansari[4]

Department of Computer Engineering, JESITMR, Nashik, India[1, 2, 3, 4]

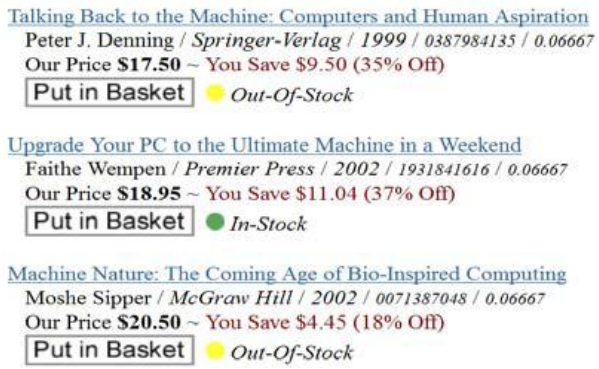
Abstract: In the world increasing number of information. This total information has become web based use of HTML files form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. For the encoded data units to be machine process able, which is essential for many applications such as deep web data collection and Internet comparison shopping, they need to be extracted out and assigned meaningful labels we present an automatic annotation approach which contains the data units on the web result page into a different groups such that same groups have the same semantic labels. Then the six annotations are combined and predict the final annotation label. An annotation wrapper for the search site is automatically constructed and can be used to annotate new result pages from the same web database. Our experiments indicate that the proposed approach is highly effective.

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

I. INTRODUCTION

Large portion of the deep web is database based, i.e., for many search engines, data encoded in the returned result pages come from the underlying structured databases. Such type of search engines is often referred as Web databases (WDB). A typical result page returned from a WDB has multiple search result records (SRRs). Each SRR contains multiple data units each of which describes one aspect of a real-world entity. Fig. 1 shows three SRRs on a result page from a book WDB. Each SRR represents one book with several data units, e.g., the first book record in Fig. 1 has data units "Talking Back to the Machine: Computers and Human Aspiration," "Peter J. Denning," etc. In this paper, a data unit is a piece of text that semantically represents one concept of an entity.

It corresponds to the value of a record under an attribute. It is different from a text node which refers to a sequence of text surrounded by a pair of HTML tags.



a) Original HTML page

```
<FORM><A>Talking Back to the Machine: Computers and
Human Aspiration</A><BR> Peter J. Denning / <FONT>
<I>Springer-Verlag / 1999 / 0387984135/0.06667</I>
</FONT> <BR>Our Price <B>$17.50</B> ~ <FONT>You
Save $9.50 (35% Off)</FONT><BR> <I>Out-Of-
Stock</I></FORM>
```

b) Simplified HTML source for first SRR

Fig 1.Example search results from Bookpool.com

II. RELATED ARTICLE

W. Liu, X. Meng, and W. Meng et al. [1] developed a paper For extracting structured data from deep Web pages is a Challenging problem due to the underlying intricate Structures of such pages. A large number of techniques have been proposed to address this problem, but all of them have inherent limitations because they are Web-page-programming-language-dependent. This approach primarily utilizes the visual features on the deep Web pages to implement deep Web data extraction, including data record extraction and data item extraction. It is also proposed as new evaluation measure revision to capture the amount of human effort needed to produce perfect extraction.

J. Madhavan, D. Ko, L. Lot, V. Ganapathy et al[2] developed a paper for content hidden behind HTML forms, has long been acknowledged as a significant gap in search engine coverage. The paper describes a system for surfacing. Deep - Web content, i.e., pre-computing submissions for each HTML form and adding the resulting HTML pages into a search engine index. The results of our surfacing have been incorporated into the Google search engine and today drive more than a thousand queries per second to Deep-Web content.

S. Mukherjee, I.V. Ramakrishnan, and A. Singh et al[3] developed a paper for identifying and annotating the semantic concepts implicit in such documents makes them directly amenable for Semantic Web processing. The paper describes a highly automated technique for annotating HTML documents, especially template-based content-rich documents, containing many different semantic concepts per document. Starting with a (small) seed of hand-labelled instances of semantic concepts in a set of HTML documents we bootstrap an annotation process that automatically identifies unlabeled concept instances present in other documents. The bootstrapping technique exploits the observation that semantically related items in content-rich documents exhibit consistency in presentation style and spatial locality to learn a statistical model for accurately identifying different

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

semantic concepts in HTML documents drawn from a variety of Web sources.

Y. Zhai and B. Liu et al [4] developed a paper for the problem of extracting data from a Web page that contains several structured data records. The first class of methods is based on machine learning, which requires human labelling of many examples from each Web site that one is interested in extracting data from. The process is time consuming due to the large number of sites and pages on the Web. The second class of algorithms is based on automatic pattern discovery. These methods are either inaccurate or make many assumptions.

III.PROPOSED WORK

A. Analysis of Search Result Records (SRR)

One-to-One Relationship (denoted as T= U).In this type, each text node holds precisely one data unit that is the text of this node encloses the value of a single attribute. Each text node is enclosed by the pair of tags <A> and </A> which refers to is a cost of the Title attribute. This can be referred to those types of text nodes known as atomic text nodes which are equal to the data units.

One-to-Many Relationship (denoted as T U). In this type of relationship, compound data units are instructed in one text node. It contains four semantic data units namely Date, ISBN, Publisher Relevance Score and Publication. As the text of those kinds of nodes can be regard as a composition of the texts of several data units, and can be called it as composite text node. Generally this examination is suitable for the reason that SRRs are produced by template programs. Finally each complex text node is divided to get real data units and annotate them.

Many-to-One Relationship (denoted as T U). In this type of relationship, multiple nodes of text jointly form a data unit. Author attribute value is composed with multiple nodes of text with each embedded contained by a distinct pair of (<A>, </A>) HTML tags. In general the webpage designers employ particular HTML tags to decorate definite information. This kind of tags is called as decorative tags since they are utilized primarily for varying the appearance of part of the text nodes.

One-To-Nothing Relationship (denoted as T U). In this type of relationship, the text nodes based on to this group are not included of any data unit inside SRRs. In addition, its examinations point out that these text nodes are frequently exhibited in a definite pattern across every SRRs. Hence, this is called as template text nodes. This identifies template text nodes by utilizing frequency-based annotator.

The following diagram in figure 2 shows the architecture of proposed system is as follows:

B. Data Alignment Algorithms

Data alignment algorithm is based on the hypothesis that attributes emerge in the similar order across every SRR on the similar result page, even though the SRRs might hold dissimilar sets of attributes.

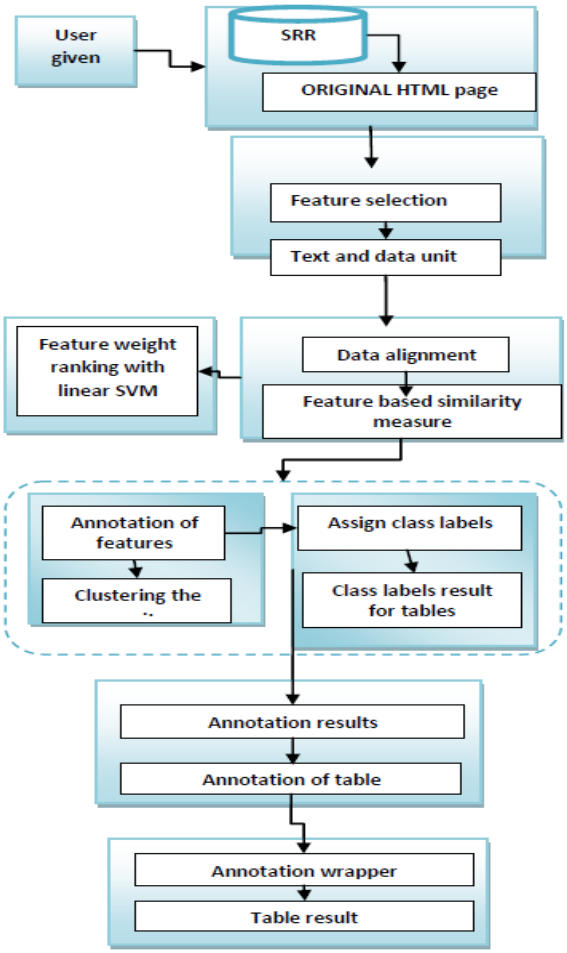


Fig2. Proposed System Architecture

C. Types of Annotators

- ✓ Table Annotator (TA)
- ✓ Query-Based Annotator (QA)
- ✓ Schema Value Annotator (SA)
- ✓ Frequency-Based Annotator (FA)
- ✓ In-Text Prefix/Suffix Annotator (IA)
- ✓ Common Knowledge Annotator (CA)

Table Annotator:

The first annotator is table based annotator, many databases to organized data units are in the table format. The table contain header which is presented in the top of the table. The table stored the data in the column and the row manner. The header is used to represent the meaning of each column and row is used to represent search result request data.

Query based annotator:

The second annotator is a query based annotator, in which user enter a query related to that data the result will be display and then user query and the data units are compared for the required column. The result is display related to that query.

Schema Value-based Annotator:

The third is a schema value based annotator, many attributes in a schema value on the search interface is predefined values. For example book related query and the related attribute authors it may have a set of predefined

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 10, October 2015

```
ALIGN(SRRs)
1.  j ← 1;
2.  while true
      //create alignment groups
3.    for i ← 1 to number of SRRs
4.      G_j ← SRR[i][j];    //j^th element in SRR[i]
5.    if G_j 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=1to|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 3.Alingment Algorithm

values i.e authors in that list.If the group having several data units, the Schema Value-based Annotator is used to find out the best synchronized attribute to the group from the IIS. The schema that firstly discovers that the uppermost matching score among the entire attribute and then it annotate the group.

Text Frequency-Based Annotator:
The fourth is the text frequency based annotator, SRR contain the records in the result page. The grouping of data is depend on the present content of that data. Same data are in one group and similarly the same for next group of data and so on. Text Frequency-Based Annotator found the general preceding units shared by all the data units of the group. The data units with the superior frequency are plausible attribute name. And the data units with the low frequency are most likely appear from databases as values.

Prefix/Suffix Annotator:
The fifth basic annotator is the prefix and suffix annotator, each search result contain a multiple search result record in the web result page. This result page also contain some prefix and the some suffix with them. For $ contain related to price so it come before price value. The prefix and suffix annotator is to check all the data units have same prefix or suffix, if it is match then it used to annotate the data units inside the next group.

Common Knowledge Annotator:
The six and the last annotator is the common knowledge, when we searching online product and the multiple results are showing to us. Then it show some related information that product, the product buy or not because it shows "in the stock" for availability and "Out of the stock" means the book not available right now, this identifications for the human because it is common thing to remember. We use alignment algorithm for align our data and extraction algorithm for extracting HTML pages.

IV. CONCLUSION

This paper proposes 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.

ACKNOWLEDGMENT

We have immense pleasure in presenting the paper "Annotating search result from web database" under the guidance of **Prof. P. S. Badgujar**. We would also like to thank Jawahar Education Societies' Institute of Technology Management and Research for providing all the required facilities.

REFERENCES

[1] Yiyao Lu,Hai He, Hongkun Zhao, WeiyiMeng , Member, and Clement Yu, Senior Member, , "Annotating Search Results from Web Databases." IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 25, NO. 3, MARCH 2013
[2] Y. Pauline Jeba, Mrs. P. Rebecca Sandra, "A Survey On Annotating Search Results From Web Databases", International Journal Of Research In Computer Applications And Robotics, Vol -1, Issue-9, 2013.
[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] Y. Lu, H. He, H. Zhao, W. Meng and C. Yu, "Annotating Structured Data of the Deep Web," Proc. IEEE 23rd Int"l Conf. Data Eng. (ICDE), 2007
[5] J. Madhavan, D. Ko, L. Lot, V. Ganapathy, A. Rasmussen, and A.Y. Halevy, "Google's Deep Web Crawl," Proc. VLDB Endowment,vol. 1, no. 2, pp. 1241-1252, 2008.
[6] 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.
[7] H. Zhao, W. Meng, and C. Yu, "Mining Templates form Search Result Records of Search Engines," Proc. ACM SIGKDD Int'l Conf.Knowledge Discovery and Data Mining, 2007.
[8] J. Zhu, Z. Nie, J. Wen, B. Zhang and W.-Y. Ma, "Simultaneous Record Detection and Attribute Labeling in Web Data Extraction, Proc. ACM SIGKDD Int' l Conf. Knowledge Discovery and Data Mining, 2006.
[9] Y. Zhai and B. Liu, "Web Data Extraction Based on Partial Tree Alignment," Proc. 14th Int'l Conf. World Wide Web (WWW '05), 2005.
[10] S. Mukherjee, I.V. Rama krishnan, and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents," Proc. IEEE Int'l Conf. Data Eng. (ICDE), 2005.
[11] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.
[12] 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 and Databases (WebDB), 2003