International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 4, Issue 1, January 2015 **ISSN 2319 - 4847**

Multi-annotating approach for searching web databases

Prof.Dr.K.Rajeswari [1], Sandhya Pawar[2], Vishakha Ramteke[3], Trupti Phatan[4]

[1] Pimpri Chinchwad College of Engineering, University of Pune

[2] Pimpri Chinchwad College of Engineering, University of Pune

[3]Pimpri Chinchwad College of Engineering, University of Pune

[4]Pimpri Chinchwad College of Engineering, University of Pune

ABSTRACT

An increasing number of databases have become web accessible through HTML form-based search interfaces. The data units returned from the underlying database are usually encoded into the result pages dynamically for human browsing. Basically every search engines shows the web content and web links related to our input query in the search box. It is just a text node which refers to a sequence of text surrounded by a pair of HTML tags. There is no relationship between text nodes and data units. We perform data unit level annotation.When we search any content in a search engine, it will group the content into different categories related to what we are searching about and also provides data unit level annotation which means order or group the content which belongs to our need .We will get summarised data in structured form.

Keywords: Data Units ,Text nodes ,Annotation ,wrapper generation.

1.INTRODUCTION

A 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.

Simplified HTML code for SRR

<FORM><A>Principles of database Query Processing for Advanced Applications.</A>
Clement T.Yu.et al/ <FONT> <I>Morgan Kaufmann,published 1997,ISBN 1558604340</I></FONT>
Our Price <B> $70.50 </B>~<FONT> You Save $ 11.45 </FONT>
 <B> Availability</B> Out of Stock </FORM>

Fig 1: Original HTML page

2. PHASES OF ANNOTATION

Earlier the data is randomized data .The data from the SRR are inserted into the table in row wise manner. That means the data of same SRR are in one row.In this the data units are having ith row of SRR and jth concept.Aim is get similar concept of data into one column.

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	

Phase 1 is the alignment phase. In this phase we are going to arrange same concept of data in one column.

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

In **Phase 2** is the annotation phase.In this we are assign labels to each column by using annotators.Annotation means assigning labels

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
L^a	L^b	L^c	L^d

In **Phase 3** is the annotation wrapper generation phase,.In this rules are assigned to each column.

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
R^a	R^b	R^c	R^d

3. DATA UNITS AND TEXT NODE RELATIONSHIPS

Data unit is each identical entity in text node.. Data unit is totally different from text node where, text node is a sequence of text surrounded by pair of HTML tags. Text node is visible element on the web page and data unit located in the text nodes.

Relationships between text node and data unit features are

3.1 One-to-One Relationship: (referred as atomic text nodes). Text node containing only one data unit i.e. the text of this node contains the value of a single attribute. Each text node surrounded by the pair of HTML tags <A> and </A>.

3.2 One-to-Many Relationship: (referred as composite text nodes) A text node consists of multiple data units i.e. multiple data units are encodes into single text nodes.

3.3 Many-to-One Relationship: (referred as decorative tags) multiple text nodes are encoded into single data unit.

3.4 One-To-Nothing Relationship: (referred as template text nodes) Text nodes are not part of any data unit inside SRRs. This relationship for text nodes and data units are represents the relation in between them.

4. FEATURES SHARED BY DATA UNITS

Data units be characterized into different features.

Following are the features of the data units.

4.1 Data content: To search information quickly data unit or text node of same concepts shares certain keywords.

4.2 Presentation style: This feature describes how a data unit is displayed on the web page by using few styles are out face, font size, color, text decoration etc.

4.3 Data type: These features are predefined characteristics that have their own meaning. Basically used data types are date, time, currency, integer, decimal etc.

4.4 Tag path: Sequence of tags traversing from root to corresponding node in the tree.

4.5 Adjacency: Adjacency refers to the data units that are immediately before and after in the SRR.

5. DATA UNIT SIMILARITY

In this two data units are compared and the similarity is calculated.Data units are compared according to there features.In this d1 and d2 are two data units and w1,w2,w3,w4,w5 are the weights of respective features.Sim(d1,d2) represents similarity between two data units.

Sim(d1,d2)=w1*SimC(d1,d2)+w2*SimP(d1,d2)+w3 * SimD(d1,d2) + w4 * SimT(d1,d2) + w5*SimA(d1,d2)

Above formula is used to calculate similarity between two data units.

- **Data content similarity (SimC).** This gives the content similarity between d1 and d2.

$$SimC(d1,d2) = \frac{V_{d1} * V_{d2}}{\| V_{d1} \| \| V_{d2} \|}$$

- **Presentation style similarity (SimP)**. The presentation are the viewing part to the user.This formula shows the similarity between the presentation styles of the compared data units.

$$\text{SimP(d1,d2)} \sum_{i=1}^{6} FSi/6$$

- **Data type similarity (SimD**). Data types are the type of the data.For example whether it is having same data types.LCS is the least count search.t1,t2 are the sequences of data types.Tlen(t)is number of component types of datatype t.

$$SimD(d1,d2) = \frac{LCS(t1,t2)}{Max(TLen(t1),TLen(t2))}$$

- **Tag path similarity (SimT).** This gives the tag path similarity.EDT(p1,p2) is the edit distance between two tagpaths.p1 and p2 are the tag paths.

$$\text{SimC(d1,d2)} = 1 - \frac{EDT(p1,p2)}{PLen(p1)+PLen(p2)}$$

- **Adjacency similarity (SimA).**

This checks the preceding and succeding similarity of the data units. SimA(d1,d2)=Sim'($d1^p$,$d2^p$)+Sim'($d1^s$,$d2^s$)) When computing the similarities (Sim0) between the preceding/succeeding units, only the first four features are used. The weight for adjacency feature (w5) is proportionally distributed to other four weights.

6. ALIGNMENT ALGORITHM

This algorithm is used for aligning.that is similar kind of concept are in a group.If all the data in a group are having similar concept than we call it as a well aligned group.For example in a table each row will have a same concept that is they are well aligned.By doing this we can further give name to it by using annotators.

Alignment Algorithm: Alignment algorithm has following four steps.

Step 1: Merge text nodes: This step detects and removes decorative tags from each SRR to allow the text nodes corresponding to the same attribute merge into a single one.

Step 2: Align text nodes: After the merging aligns text nodes into different groups. So that same group has the same concepts.

Step 3: Split text nodes: In this step split the composite text nodes into separate data unit.

Step 4: Align data units: This is the last step for alignment in which separates each composite group into multiple aligned groups with each containing the data units of the same concept.

Let number of SRRs=n

number of Data units=m

ALIGN(SRRs)

```
1.  j=1;
2.  while true
                //create alignment groups
3.   for i=1 to number of SRRs
4.   Gj=SRR[i][j];  //j th  element in SRR[i]
5.   if Gj 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=NULL;
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(sim(V[k],S));
      k=1to|V|
  // shifting
14.    for k=1 to |V| and k !=c
15.   for each 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=O;
4.     L=NIL; R=NIL;
5.     for each A in V
6.     for each 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 U R to V;
15.   else break loop;
16.    return V;
```

7. BASIC ANNOTATORS

The extracted data from web database is analysed and represented using annotators.Annotation means labelling the data units.By using the 6 annotators we can label the data units.

There are 6 Types of annotators

7.1 Table-based Annotator,
7.2 User Query-based Annotator,
7.3 Prefix/Suffix-based Annotator,
7.4 Common Knowledge-based Annotator,
7.5 Schema Value-based Annotator,
7.6 Text Frequency-Based Annotator.

7.1 Table-based Annotator (TA)

In this each SRR are inserted into the table .Each row contains the one SRR.Than consider many SRR which to be inserted in the table.Each cell represents one data unit.Than one column will be compared if each unit will have same concept.According to the concept the label is given to the column.

7.2 User Query-Based Annotator (QA)

In this many queries are given and in that the similar words from the given user querries are compared and label is assigned according to the common word.

7.3 Text Frequency-Based Annotator (FA)

Some texts are occurred in all records in the result page. Data units are grouped depends upon the concept. Some group has lesser frequency. The higher frequency data units are attribute names and lower frequency data units are their values. To calculate this, compute the cosine similarity between the attribute and the data unit. 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 lesser frequency are most likely appear from databases as values

7.4 Prefix/Suffix Annotator (IA)

Prefix are the data which is before the data unit and suffix are the data after that data unit.The data units are checked and the prefix and suffix are seen .If all the prefix and suffix are similar then that will be give as label.

7.5.Common Knowledge Annotator (CA)

As the name suggest labels are assigned according to the common knowledge.If name of the countries are written in the column.Then it is commonly understood that it is in the common knowledge country.

8. DATA ALIGNMENT AND LABELLING

The existing works differs when compared with the automatic annotation approach. They are based on one or few features. In automatic annotation the alignment approach first handles the relationship between data units and text nodes and utilizes different types of data unit features. And a cluster-based shifting algorithm is used in alignment process. Label assignment is performed using IIS (Integrated Interface Schema) and LIS (Local Interface Schema).IIS contains the attributes in all the LIS and thus eliminates label inadequacy and inconsistent label problems. Few basic annotators in are introduced to annotate the aligned groups and a probability model is used to combine the results of multiple annotators. This approach is called multi-annotator approach

9. CONCLUSION

Assigning meaningful labels to the extracted data unit of each SRR is a challenging task. The automatic annotation approach considers several types of data unit and text node features and makes annotation scalable and automatic. Multiple annotators of different features are used to annotate the extracted information from the result pages. Each annotators exhibit one special type of feature and they are together used to automatically construct a high quality annotation wrapper. Uses both LIS and IIS for label assignment and alleviates local interface schema inadequacy and inconsistent label problem.

REFERENCES

[1] A. 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 and Databases (WebDB), 2003.
[3] P. Chan and S. Stolfo, "Experiments on Multistrategy Learning by Meta-Learning," Proc. Second Int'l Conf. Information and Knowledge Management (CIKM), 1993.
[4] W. Bruce Croft, "Combining Approaches for Information Retrieval," Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval, Kluwer Academic, 2000

AUTHOR

Prof. Dr. K. Rajeswari has completed her Ph.d. in July 2014 from SASTRA University, Tamil Nadu and has a teaching experience of a 17 + years .She is an expertise in Data Mining, Machine Intelligence, Artificial Intelligence, Nano Technology. She has published and presented over 50 research papers in reputed journals and conferences.

Sandhya Arjun Pawar pursuing Bachelor of Computer Engineering from Pimpri Chinhwad College Of Engineering ,Savitribai Phule Pune University.

Trupti Pandurang Phatan pursuing Bachelor of Computer Engineering from Pimpri Chinhwad College Of Engineering ,Savitribai Phule Pune University.

Vishakha Dharmapal Ramteke pursuing Bachelor of Computer Engineering from Pimpri Chinhwad College Of Engineering ,Savitribai Phule Pune University.