International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

Feature Clustering and Annotating Search Results from Web Databases

Lekshmi S.S. [1], Suryapriya S.[2]

[1]PG Scholar, Sarabhai Institute of Science & Technology, Vellanadu, Kerala-695543, India

[2]Assistant professor, Sarabhai Institute of Science & Technology, Vellanadu, Kerala-69554, India

Abstract: ***Web contain huge amount of information on Web sites the user can retrieve this with help of the search input query to Web databases & fetch the relevant information. In online shopping websites essential data will be obtained if the items are selected from the given list. But if one wants to find an item by giving a query in the search interface there is no guarantee that the retrieved data will be relevant. The information retrieval should be done automatically & arrange in a systematic way for further processing. Various methodologies like wrapper induction is been induced. The labeling is done to the extracted information as per the concept. Various types of annotators are used on the basis of the data to be annotated. In this paper survey the automatic annotation approach on the basis of different feature of text node and data units.***

General Terms: Data extraction, Web data annotation and Wrapper induction

Keywords: Data Extraction, Data annotation, Annotators, Text nodes, Data Units and Wrapper

1. Introduction

Now a day's web technology is getting an emergence importance in day to day life! Everyone is familiar with online shopping sites. There are different technologies & researches are focusing on the extraction of relevant information from large web data storage. But still there is requirement of availability of automatic annotation of this extracted information into a systematic way so to be processed later for various purposes Web information extraction and annotation has been active research area in web mining. A huge amount of the data is available on the web. The user enter the search input query in the search engine, and

it will return the dynamically search output records on browser. Many online shopping sites are available to users. There is a need for technique which should help us to provide retrieved relevant data as per user requirements. The last decade focus on multiple methodologies in firing queries, information fetching & optimization. The idea of wrapper is introduced. The wrapper is a software concept which wraps the contents of a web page using its source code via HTTP protocols [8] but it does not change the original query mechanism of that web page.

2. Literature Survey

The World Wide Web is having vital data in numerous formats the users have to deal with this data by using a search based form. The user will retrieve the information by firing the query. In traditional approach the search base form is design to fire the queries & required data is fetched. HTML form is containing the plain text. Querying, Integration etc. are used. These techniques are not effective to produce accurate search result record from web databases, because of human involvement and poor quality of the data extraction output. Two main problem arises during extracting the relevant information First: to categorized the unstructured view of data such as search engine. Second: categorized structure and semi-structure view of data. The web sites are also having heterogeneous nature due to language independent. The e commerce website or the information portals are updating their content on a regular basic. The web data is now machined process able so, we require the relevant information extraction with the semantic grouping. The semantic grouping means the data with similar meaning can form group with same concept. XML/RDF has been widely used for representing semantic web that required annotation for recognition of semantic web. These techniques provide manual mapping of unlabeled document segment to ontological concepts. In bootstrapping semantic labeling is addressed in semantic web annotation. The presentation style & spatial locality in the HTML tag is focused [3].The sites like educational, news portal and e-commerce are dynamically update contents on a regular basis so called as content-rich web sites contents management software that creates HTML pages by populating templates from databases. The structural analysis technique use to group together related elements in a HTML pages into unlabeled tree. The algorithm can use the hand-labeled concept instances from HTLM pages for identification of unlabeled concept instances in HTML pages and assigns semantic labels to them. The algorithm does not used hand-crafted ontology. For determining the consistency in presentation style we can use the feature extraction. So the data belong to same concept or set of concepts lie under similar group.

3. Types of Annotators

The returned result page contains many SRRs. The data units corresponding to the similar concept (attribute) often share special common features in certain patterns. Based on this, in this paper we used the six basic annotators have been defined to label data units, where each of them considers a special type of patterns/features. Each annotator are play unique role in labeling the name to the data units are

extracted by the wrapper. Four of these annotators (i.e., table annotator, query-based annotator, in text prefix/suffix annotator, and common knowledge annotator) are similar to the annotation heuristics used by DeLa but there different implementations for three of them (i.e., table annotator, query-based annotator, and common knowledge annotator) [1] [6] [2].

3.1 Table Annotator

The resulted page fetch from multiple website consist of various SRR. These information can be stored in the form of table. A table consist of different column header &rows. The cell of this table indicates the data unit. We can store the multiple data units. The table annotator used in Dela [2] Approach mainly focus on the <TD> tag elements. The information stored in <TD>elements is stored in the annotator table. But few websites contain the <TD> tag elements. So the table annotator is modified .The row is considered as SRR & the column is considered as attribute. The data unit having same features can be aligned under header & the column header. By considering the special feature we can annotate the SRR. Firstly we have to identify all the values of column then as per SRR we have to fill the data. In such way the limitation of Dela [2] is improved.

3.2 Query-Based Annotator

The SRR is always returned from WDB on the basis of fired query. When the user submits the data in the text box or select field from the list box on the search form, the query is fired on the WDB. Then the SRR is identified & the data is stored under the column header. The no of occurrences of matching the column header will decide the group & we can label it. The Dela uses only the local labels in the query. However, DeLa uses only local schema element names, not element names in the IIS [2].so, the new approach is use to utilize the global schema.

3.3 Schema Value Annotator

Many attributes on a search interface have predefined values on the interface. More attributes in the IIS tend to have predefined values and these attributes are likely to have more such values than those in LIS. When values from different LIS are integrated then we have to modify the schema values to perform annotation.

3.4 Frequency Based Annotator

The adjacent units have different occurrence frequencies. The data units are always associated with the higher frequency & lower frequency. The attribute names are the higher frequencies, while the data units with the lower frequency most probably come from databases as embedded values. Suppose there is a group of lower frequency then we can easily find its preceding values shared by all data units in the group .We can analysis the data unit until it is different & map its preceding. Then we can combine the preceding to form the label.

3.5 In-Text Prefix/Suffix Annotator

In some cases, the data unit is aligned with its label. The data unit consists of the comma separated vales & the labels associated with it. These lie in a particular sequence separated from each other in all multiple SRR. After alignment it will form a group. The in text prefix/suffix will check for data unit. If the same prefix is there &not a deliminator then it is removed from all data units but if the number of data nodes match with the same suffix to the data node within next group then the suffix is used for the annotation. Any group whose data unit texts are completely identical is not considered by this annotator.

3.6 Common Knowledge Annotator

Some data units on the result page are self-explanatory because of the common knowledge shared by human beings For example, “in stock” and “out of stock” occurs in many SRRs from e-commerce sites. Human users understand that it is about the availability of the product because this is common knowledge. Each common concept contains a label and a set of patterns or values.

4. Phases of Annotator

From the SRR, first identify all data units and then organize them into different groups with each group corresponding to a different concept. The data unit with same concept can fall under the same column header like table annotator.

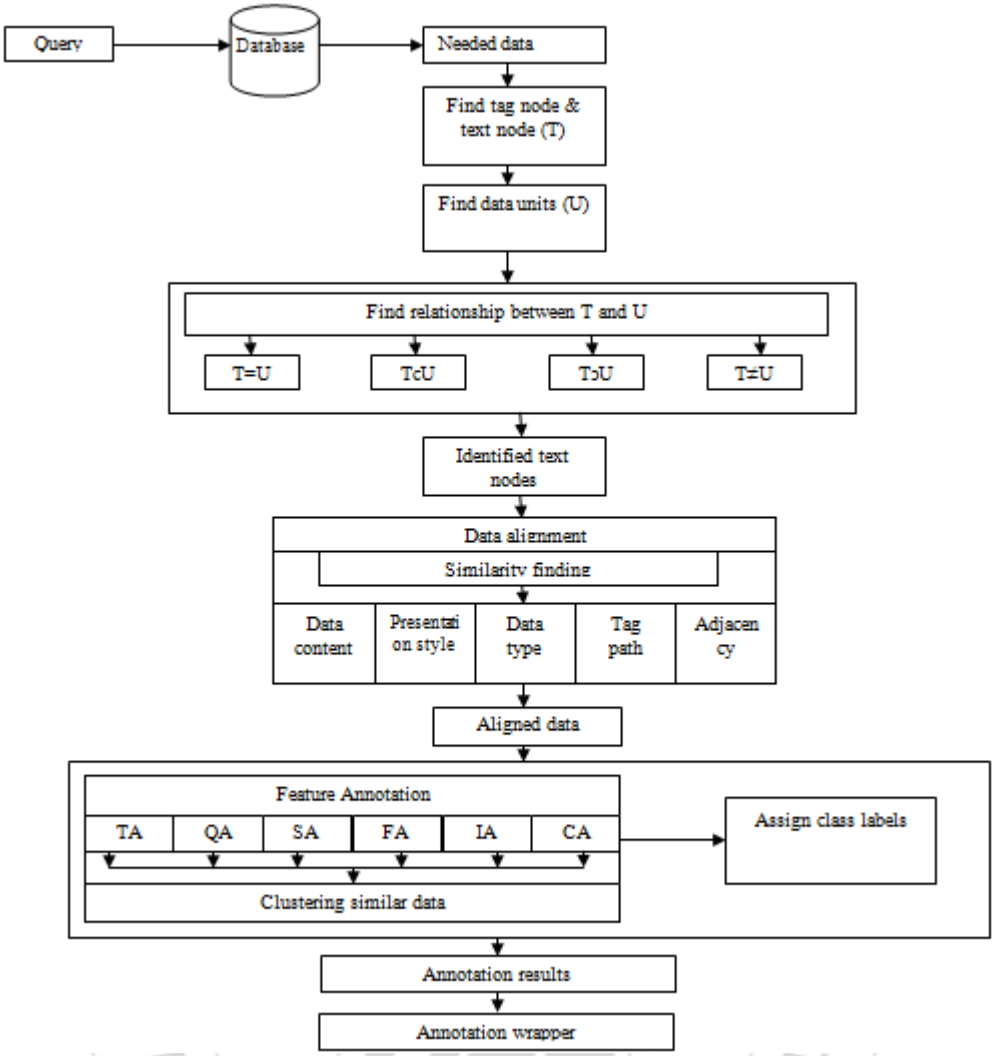

Figure 1: Feature extraction and Annotation

4.1 Alignment Phase

The alignment phases identify all data units in the SRRs and then organize them into different groups with each group corresponding to a different concept

4.2 Annotation Phase

In this phase, single or combined multiple annotators are used as per the requirement for annotation. This work on the probability based.

4.3 Wrapper Generation Phase

The wrapper set the rules for extracting the information from same WDB. The annotator wrapper can be used for further analysis. We can write the wrapper after combining the multiple annotators. For mapping the information between text node &data node we have to first find the relationship between them. The relationship between the data unit and text node are as bellow:

- **One-to-One**
  In some cases the text nodes are equivalent to data nodes so can be used for annotation in a easy way. For example the <a>… in HTML itself indicate the data value & attribute .But this is not the general case always to be considered in fig 4. Show that *title* attribute each search result considers as a one-to-one relationship [2][1].
- **Many-to-One**
  In this case, multiple text nodes together form a data unit. For example the vendor name can be embedded inside the <a>.. tag .Another example can be considered that the price can be entitled within … tag [1].
- **One-to-Nothing**
  In this case the text node is not part of any data unit. For Example vender name does not contain data unit but instead describe the meaning data unit. It is also known as *Template text node* [1].

5. Data & Text Node Alignment

Data alignment algorithm is based on the assumption that attributes appear in the same order across all SRRs on the same result page, even if the SRRs may contain different sets of attributes (due to missing values) [1]. SRRs from the same WDB are generated by the same schema. Thus, we can consider the SRRs on a result page in a table format where each row represents one SRR and each cell holds a data unit (or empty if the data unit is not available). The goal of alignment is to move the data units in the table so that every alignment group (column) contain similar data unit,

preserving the order within every SRR is preserved. The alignment algorithm is based on following steps:

- **Merge Text Nodes**
  This mainly focuses on removing the decorative or presentation style tags so that all text nodes can be merged.
- **Align Text Nodes**
  This will align the nodes with the same concept or set of concepts under one group for atomic node as well as for composite nodes.
- **Split (Composite) Text Node**
  The split node again have to be focused on the annotation work .we have to split the "values" in composite text nodes into separate data units. This step is done based on the text nodes in the same group.
- **Align Data Units**
  This step is to separate each composite group into multiple aligned groups with each containing the data units of the same concept.

6. Conclusion

In this paper we reviewed that various data extraction techniques as well as automatic annotation approach using multiple annotators from different Web data bases. We also surveyed that how the data extraction from the various web pages but the traditional approach is having many drawbacks like human interference, the inaccuracy in result and poor scalability. Some approach are used the different feature extraction techniques such as sequence based Tree edit distance, DOM tree, pattern matching and HTML tag structure. In visual data extraction approach is the language independent. This approach mainly focus on the presentation style of and extract the visually information from the template. But still there is need to identify the best technique for data annotation problems.

7. Acknowledgment

I would like to thank the University Authorities to provide basic facilities for carrying out the research work. I would like to thank my guide Mrs.Suryapriya S. and my parents for most support and encouragement, valuable advices on grammar and theme of the paper.

References

[1] Y. Lu, H. He, H. Zhao, W. Meng, C.Yu "Annotating Search Results from Web Databases", IEEE Knowledge and Data Engg"., vol. 25, March-2013.
[2] J. Wang and F.H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc. 12th Int'l Conf. World Wide Web (WWW), 2003.
[3] S. Mukherjee, I . V. Ramakrishnan and A. Singh, "Bootstrapping Semantic Annotation for Content-Rich HTML Documents", Proc. IEEE Int'l Conf. Data Eng. (ICDE)", 2005.
[4] Davi de Casto Reis, Paulo B. Golgher and Altigran S. da Silva, "Automatic Web News Extraction Using Tree Edit Distance", Proc. ACM World Wide Web (WWW), 2004.
[5] 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.
[6] 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.
[7] W. Liu, X Meng and W. Meng, "ViDE: A Vision-Based Approach for Deep Web Data Extraction," *IEEE Trans. Knowledge and Data Engg.*, vol. 22, no. 3, pp. 447-460, March 2010.
[8] H. He, W. Meng, C. Yu and Z. Wu, "Automatic Integration of Web Interface with WISE-Intigrator," VLDB J., vol. 13, no. 3 pp.256-273, Sept 2004.
[9] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis and Khaled Shaalan "A Survey of Web Information Extraction Systems" IEEE, TKDE-0475-1104.R3.
[10] 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.
[11] V. Crescenzi, G. Mecca, and P. Merialdo, "RoadRunner: Towards Automatic Data Extraction from Large Web Sites," *Proc. Int'l Conf. Very Large Data Bases(VLDB)*,pp.109-118,2001.