ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Annotating Multiple Web Databases Using Svm

M.Yazhmozhi[1], M. Lavanya[2], Dr. N. Rajkumar[3]

PG Scholar, Department of Software Engineering, Sri Ramakrishna Engineering College, Coimbatore, India[1, 3]

Head of the department, Department of Software Engineering, Sri Ramakrishna Engineering College, Coimbatore, India[2]

ABSTRACT: There is a far above the ground demand for deep web data search. By using search interfaces, backend database can be accessed through web. The result for the user query in search interfaces is search result record (SRR) that consists of data units. For efficient deep web data search, the SRR should be extracted out and then meaningful labels would be added to data units. This process is referred as annotation. This methodology is a time consuming process for the user and also weight values are fixed while finding the similarity. To overcome these problems, Support Vector Machine (SVM) has been used. With the help of old annotation results, the trained data has been created. This is given as input for SVM. Whenever the user gave search query, Label will be generated automatically using SVM. The SVM algorithm has been enhanced the annotation course of action.

KEYWORDS: Data alignment, data annotation, web database, wrapper generation

I. INTRODUCTION

Web search engine is intended to search the required information from the World Wide Web. The data returned from the search engines are encoded in the returned result page. These encoded data are returned from the databases called Web databases (WDB). A result page returned from WDB has multiple search result records (SRRs) and the SRR contains multiple data units. 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. The data unit concept is different from a text node. Text node is defined as a series of text that surrounded by a pair of associated HTML tags.

The International Standard Book Number (ISBNs) can be compared to achieve this. If ISBNs are not available, their attributes (titles and authors) could be compared. The system may also need to compare the semantic of each data unit. For instance it may list the prices of book offered by each site. Thus, the system needs to know the semantic of each data unit. Unfortunately, the semantic labels of data units are often not provided in result pages. The automatic annotation approach consists of three stages. If the user enters his query then the corresponding SRRs are displayed in the result page. The first stage is alignment stage and it consists of two steps. In the first step of alignment stage, the data units presented in the SRRs units is not only important for the above record linkage task, but also for storing collected SRRs into a database table for later analysis.

The older applications require dreadful human efforts to annotate data units manually, so the scalability is limited. Automatic annotation approach, consider how to automatically assign labels to the data units within the SRRs returned from WDBs.Step of alignment stage, the identified data units in the SRRs are organized into different groups. And each group corresponding to a different concept that is same semantic (e.g., all titles are grouped together). These grouping of same semantic for the data units are used to identify the common patterns and features among the data units. The second stage is the annotation stage. In annotation stage, multiple basic annotators are introduced with each exploiting one type of features. The common features for the data units from the first stage are the basis of annotators. Each and every basic annotator is used to produce a label for the data units. To precede an appropriate label for data units the probability model is used. The third stage is the annotation wrapper generation stage. In this annotation wrapper generation stage an annotation rule is created. These rules are used to explain how to extract the data units of the identified concept in the result page. And it also determines the appropriate label. If the new quire is submitted for the same WDB, these rules are used to directly annotate the data from the same WDB without need to perform the alignment and annotation stages again.

Using these annotation wrappers, annotation process was performed quickly. To make the annotation process efficient, the Support Vector Machines algorithm is used. The training data has been generated by using the previous annotation process results. With the help of this, SVM algorithm will learn about the label assignment for the WDB.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

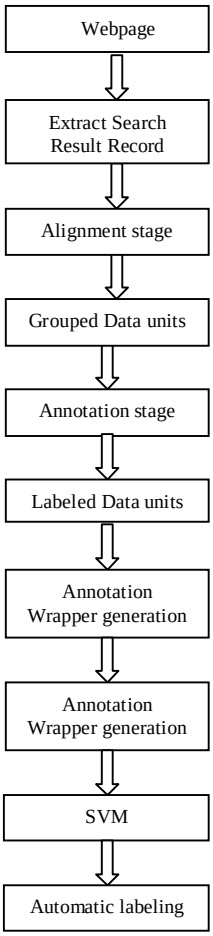


Fig. 1. Architecture diagram

II. RELATED WORKS

J. Wang and F.H. Lochovsky[2], present a scheme for annotation that simply based on HTML tags. If some web sites uses the newer version (XML), This method is not suitable for some the newer version .This approach uses one-to-one and one-to-many relationship. But many-to-one and one-to-nothing relationship are not used. DeLa approach uses only the local interface schema (LIS).

Meng.W, Yu.C, and Liu.K[1] proposed a arrangement styles and the spatial locality for data arrangement. They only use the only one relationship. This scheme mainly focused on human for labeling. But it is fully domain in need. These methods are not fully automatic.

Crescenzi.V, Mecca.G, and Merialdo.P[3] proposed a technique that mainly focused on extracting information from HTML sites. These techniques utilize the automatic generated wrappers. But these wrappers are used simply intended for data aligning method and not for annotation development.

Zhai.Y and Liu.B [5] proposed a technique, which gave key to the problem for extraction of data. This system mainly used intended for the Web page that contains more than a few ordered data records. This system fragments the data records and extracts data items. And they are ordered in to table. The main shortcoming of this system is the starting stage of class methods are based on machine learning. So they need the human labeling method.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

RoadRunner, Arlotta.L, Crescenzi.V, Mecca.G, and Merialdo.P[6] proposed a prototype, called Labeler. This prototype is used for the annotation. This labeler annotates data units with the closest labels on result pages. The main disadvantage of this system is, many WDBs fail to encode data units with their labels on the result page.

Salton.G.M and McGill[7] propose a system that uses domain(field) ontology to assign labels. These ontologies are constructed using query interfaces and result pages from WDBs in the similar domain. Later than the label is assigned, the data values with the same label are logically aligned. This method is susceptible to the superiority and wholeness of the ontologies generated.

III. RESEARCH METHODOLOHY

The major contribution was

1. The first work is to analyze the relationships between text nodes and data units. For that four types of relationships are used.
2. Then the alignment should be performed. For that the clustering-based shifting technique is proposed. Clustering-based shifting technique is used to align the data unit.
3. Then the important features are also identified. The features are
    1. Data type feature ,
    2. Data content feature ,
    3. Presentation style feature ,
    4. Tag Path feature,
    5. Adjacency feature.
4. To make the annotation process over multiple WDB the integrated interface schema (IIS) is used.
4. The six basic annotators were used to allocate labels to data units
5. A probabilistic model was used to unite the consequences from different annotators into a single label.
6. Then annotation wrappers were constructed for each WDB.

3.1 Relationship between text node and data unit.

If the user enters his query then the corresponding SRRs are displayed in the result page. Each SRR extracted from the particular website and each web site has its own tag structure. These tag structure are used to determine the structure of web browser. That is how the content of SRRs is displayed on a web browser.

The tag structure has both text node and tag node. Tag node is defined as a tag of HTML. Tag node comes in-between the open and close angle brackets. The text node comes outside of the open and close angle brackets. Text nodes are the able to be seen on the webpage and data units are positioned in the text nodes. In this work, the data unit level annotation is focused. So the data units and the text nodes should be recognized. To separate the data units from the Text node, four types of relationships are used.

The relationships between the text node and data unit are,

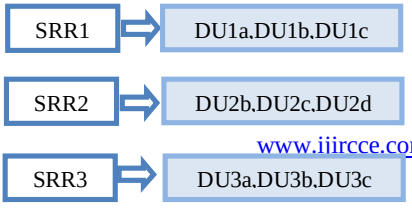
- One-to-One Relationship:
  In One-to-One Relationship, the text node refers to only one data unit.
- One-to-Many Relationship:
  In this One-to-Many Relationship, a single text node holds multiple data units.
- Many-to-One Relationship:
  In Many-to-One Relationship, a single data unit has multiple text nodes.
- One-To-Nothing Relationship: In One-To-Nothing Relationship, text nodes are not element of any data unit.

3.2 Alignment stage

The alignment stage consists of two processes.

Process1:

In the first process of alignment stage, the data units presented in the SRRs are identified.



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Fig.2. Extracting Data Units from SRR

In Fig.2, Data Units are defined as DU and the number 1,2,3 specify the number of SRRs. The concepts are denoted as alphabets (a,b,c). Each SRR will have n numbers of DU. So the DU should be identified first. Fig.2 shows the extracting of DU from the SRR.

Process2:

In the second process of alignment stage, the identified data units in the SRRs are organized into different groups.

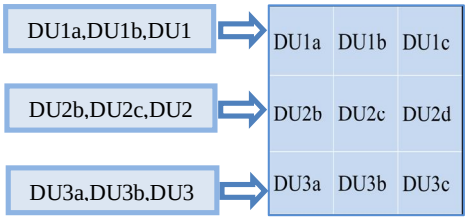


Fig.3. Grouping of Data Units

In Fig.3, DUs has been grouped and structured as table format. And every group analogous to different notion that is similar semantic (e.g., all titles are grouped together).These assemblage of identical semantic for the data units are used to identify the common patterns and features in the middle of the data units.

Fig.4 shows the organization of Data Units. This organization was based on SRR's concept. DU of Same concept has been arranged in to same column in the strutted table.

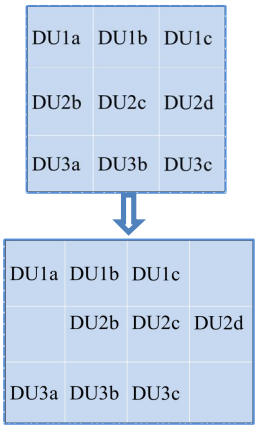


Fig.4. Organization of Data Units

In Fig.4 DU1a and DU3a were in same column because both DUs have same concept. i.e. (Concept a). If the concept of the particular DU was not presented then the particular column was left and checks for the next concept.

3.3 Annotation stage

The second stage is the annotation stage. In annotation stage, multiple basic annotators are introduced with each exploiting one type of features.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

The common features for the data units from the first stage are the basis of annotators. Each and every basic annotator is used to produce a label for the data units. To precede an appropriate label for data units the probability model is used.

3.3.1Basic annotators

Annotators are used to create a label for the data units. The common features are got from the first alignment stage. In this annotation stage we use the common features to produce a label. In this paper, six basic annotators were introduced to label data units.

The result page returned from the required website has multiple SRRs. The data units from the SRR may share some special common features. So, these common features are usually associated with the data units on the result page in certain patterns. All these annotators are used for labeling. Single annotator labeling was not perfect.

The six basic annotators are

- Table-based Annotator,
- User Query-based Annotator,
- Prefix/Suffix-based Annotator,
- Common Knowledge-based Annotator,
- Schema Value-based Annotator,
- Text Frequency-Based Annotator.

3.3.1.1 Table-based Annotator (TA)

Tables are used by many web data bases to organize the data units. Each table has its header and it always presented in the top. The header is used to represent the meaning of each column. And the row is used to represent the SRR. The arrangement of this table is get during extraction of SRR.

The Table-based Annotator has two processes.

Process1:

In the first process of Table Annotator, column header for the each data unit should be identified.

Process2:

It takes the first SRR and the first data units and checks with the header for maximum vertical overlap and finally chooses the header. It was determined by the coordinates. Then take the second data unit and do the process given above.

These processes are applied to all data units. And finally all data units are arranged to their corresponding header.

3.3.1.2 User Query-Based Annotator (QA)

The User Query-Based Annotator uses the user query for annotation. If user enters his query the required result will be displayed. Then the user query and the data units are compared for the required column. This returned SRR always related to the user query. So the data unit related to that SRR should also relate to that concept. It checks the user query and title i.e. first data unit. Because title was presented in the first data unit in many web sites. If it not matches, then moves to the next annotator.

3.3.1.3 Schema Value-based Annotator (SA)

A lot of attributes on a search interface had predefined values. For example, the attribute authors may have a set of predefined values (i.e., authors) in its selection 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 value annotator initially discovers the feature that has the uppermost matching score among all attributes and then to annotate the group. Note that multiplying the above summation by the number of nonzero similarities is to provide preference to attributes that have additional matches (i.e., having nonzero similarities) over those that have fewer matches.

3.3.1.4 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.

3.3.1.5 Prefix/Suffix Annotator (IA)

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

The result page may contain some prefix and suffix with them. For example, '$' may come before the prize value. These prefix and suffix may repeat in all SRRs. The in-text prefix/suffix annotator checks the data units in the SRR have the same prefix or suffix. If it matches then the suffix or prefix is used to annotate the data units inside the next group.

3.3.1.6 Common Knowledge Annotator (CA)

The data units on the result page for the user query may self-explanatory. The cause is that the widespread knowledge is communal by human beings. For example, some text may occur in many SRRs from e-commerce sites. This is identified by the human users because of their common knowledge.

The common knowledge annotator aims to exploit this situation by means of a number of predefined widespread concepts. For example, a state concept has a label "state" and a set of values such as "Tamil Nadu," "Kerala," and so on. Common knowledge annotator considers both patterns and certain value sets such as the set of states.

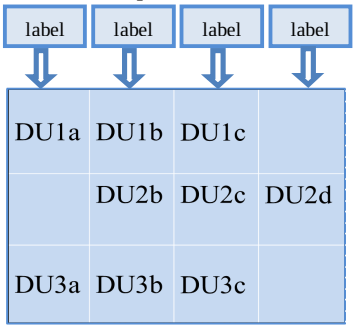


Fig.5. Assigning labels to each group

3.4 Annotation wrapper generation stage

The third stage is the annotation wrapper generation stage. In annotation wrapper generation stage, an annotation rule is created. These rules are used to explain how to extract the data units of the identified concept in the result page.

And it also determines the appropriate label. If the new query is submitted for the same WDB, these rules are used to directly annotate the data from the same WDB without need to perform the alignment and annotation stages again. Using these annotation wrappers, annotation process was performed quickly.

3.5 Automatic Label Generation using SVM

The training data has been generated by using the previous annotation process results. With the help of this, SVM algorithm will learn about the label assignment for the WDB. Whenever user gave the input query, SVM algorithm will generate the label for that WDB without going for annotation each time.

The experimental result shows that the Time consumption has been reduced due to the automatic generation of label using SVM algorithm. Also the weight values are fixed, while calculating similarities in annotation process. Where as, in SVM the weight a value has been generated automatically, since it was learned through previous results.

IV. Conclusion and Future Work

Automatic label generation i.e. SVM approach has been used to assign labels automatically without any human involvement. In this work, the annotation wrappers have been generated where the data units are extracted automatically for the identified concept. To make the annotation process over multiple WDB the integrated interface schema (IIS) is used. The clustering based shifting scheme was performed for precise alignment. To make the annotation process efficient, the Support Vector Machines algorithm is used. The integrated interface schema is used to carry out the annotation process over multiple web data base.

Performance evaluation : To evaluate the performance of the project, the precision and recall has been measured. The precision is defined as the percentage of the correctly aligned data units over all the aligned units by the system. Recall is the percentage of the data units that are correctly aligned by the system over all manually aligned data units by the expert. Fig.6 gives the performance measure for the precision values.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

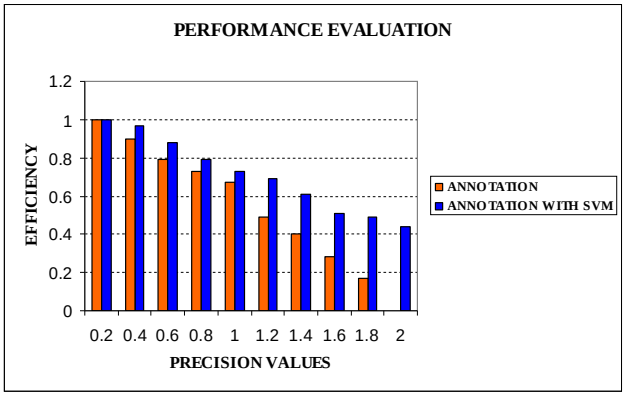


Fig.6. performance evaluation

REFERENCES

[1] 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.
[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] Crescenzi, G. Mecca, and P. Merialdo, "RoadRUNNER: Towards Automatic Data Extraction from Large Web Sites," Proc. Very Large Data Bases (VLDB) Conf., 2001.
[4] 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.
[5] 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.
[6] 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.
[7] 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.
[8] Embley.D, Campbell.D, Jiang.Y, Liddle.S, Lonsdale.D, Y. Ng.Y, and Smith.R, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.
[9] Freitag.D, "Multistrategy Learning for Information Extraction," Proc. 15th Int'l Conf. Machine Learning (ICML), 1998.
[10] Goldberg.D, Genetic Algorithms in Search, Optimization and Machine Learning. Addison Wesley, 1989.
[11] Arasu.A and Garcia-Molina.H, "Extracting Structured Data from Web Pages," Proc. SIGMOD Int'l Conf. Management of Data, 2003.
[12] Arlotta.L, Crescenzi.V, Mecca.G, and Merialdo.P, "Automatic Annotation of Data Extracted from Large Web Sites," Proc. Sixth Int'l Workshop the Web and Databases (Web DB), 2003.