

Annotation on SRRS with Decorative Tag Detection & Composite Text Node Splitting

Thushara K P¹, Neethu T²

M. Tech Student, Department of Computer Science and Engineering, MIT, Kannur, India¹

Assistant Professor, Department of Computer Science and Engineering, MIT, Kannur, India²

Abstract: Data Mining is a computational process of discovering useful information from large data sets. Web Mining is recently an active research area under Data Mining, where discovery of information from web documents or web pages are done. There is a big influence of technology in our daily life, especially Internet, is more and more important for our life and surely it will be the medium of the future. So extracting information from web and organizing them for human use has a great importance. In this work, we propose an Automatic Annotation Approach, in which information from web databases are collected and grouped according to their similarity. We assign meaningful labels to these groups. A web database is an organized listing of web pages, which have a searching interface through which user can enter their query. Annotation is important since shopping comparison and multiple domain searches are necessary for our day-to-day life. In our work we use four different domains and we search through these domains and results are shown in an annotated format, these annotated results will be easier to interpret by human users and can be used in comparison of data.

Keywords: SRR, Annotation, WDB, Wrapper.

I. INTRODUCTION

Data mining is the computational process of discovering information from large data sets. The goal is to extract information from a data set and transform it into an understandable structure for further use. A web database contains multiple SRRs, each SRR contain several data units. Data units represent the single concept of a real-world entity. Here annotation is performed on the basis of data units.

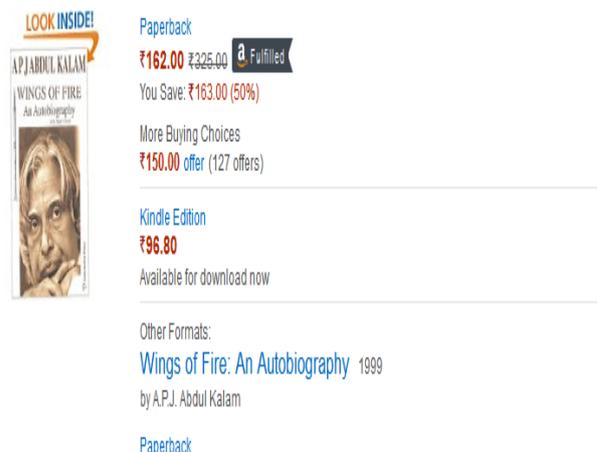
Figure present SRR on a result page of book WDB. The SRR have one book with several data units, e.g., the first book record in Fig. 1 has data units "Wings of Fire: An Autobiography," "A P J Abdul Kalam," etc. Now-a-days there is high demand for retrieving data from multiple WDBs. Earlier applications need human effort to annotate data units manually, so they have limited scalability due. In our Work, we propose an approach for automatically annotate SRRs returned from web databases.

Initially we perform extraction of data from each SRR, and then these data are grouped based on their similarity. Then assigning label is done and finally we will generate some rules for future. The work can be divided into three phases as in the previous work [1]. The first phase (alignment phase) all data units in the SRRs are identified and then aligned them into groups corresponding to same concept. In the second phase assigning labels to data units. In the next phase an annotation wrapper (annotation rule) is generated, which shows how to extract data units of same concept for the corresponding web database. This annotation wrapper makes the annotation faster in future.

Our work has following contributions:

- Our Work analyse different relationships between text nodes and data units. Perform a data unit level annotation.

- We use a tag based clustering technique, in which data units are grouped according to similar tags. We use DOM tree structure to find tags and corresponding data.
- We have included multiple web data base of multiple domains. Hence when user searches he will get results of different domains.
- For annotation we use named entity technique. By analysing clusters we generate some rules to identify entities, and these identified entity used as labels.
- We generate an annotation wrapper [1] for each WDB. This wrapper can be applied to annotate the SRRs retrieved from the WDB with new queries.
- When user searches for with a query, he will get not only result one domain but also he will get results of multiple domains and the result will be displayed in annotated form.



```
<a class="a-link-normal s-access-detail-page a-text-normal" href="http://www.amazon.in/Wings-Fire-P-7-Abdul-Kalam/dp/81737146_0ks&ie=UTF8&qid=1448275877&sr=1-1&keywords=a+p+j+abdul+kalam" title="Wings of Fire: An Autobiography">
  <h2 class="a-size-medium a-color-null s-inline s-access-title a-text-normal">
    Wings of Fire: An Autobiography</h2>
  </a>
  <span class="a-letter-space"></span>
  <span class="a-letter-space"></span>
  <span class="a-size-small a-color-secondary">1999</span>
</div class="a-row a-spacing-none">
  ::before
  <span class="a-size-small a-color-secondary">by</span>
  <span class="a-size-small a-color-secondary">A.P.J. Abdul Kalam</span>
  ::after
</div>
  ::after
</div>
<a class="a-size-small a-link-normal a-text-normal" href="http://www.amazon.in/Wings-Fire-Autobiography-Abridged-Everc_0ks&ie=UTF8&qid=1448275877&sr=1-1&keywords=a+p+j+abdul+kalam">Paperback</a>
<span class="a-size-small a-color-secondary"></span>
```

Fig. 1: Example search result from a book web

The rest of the paper is organized as follows: Section 2 describes related work. Section 3 introduces data alignment and wrapper generation process. Section 4 describes performance evaluation of our work and section 5 concludes the paper.

II. RELATED WORK

Web information extraction and annotation is an active research area. The early system Wrapper Induction for Information Extraction [5] depends on human efforts for extracting data from a particular information source. A wrapper is a procedure that is specific to a single information resource. These are applicable to tabular pages. The paper proposes an automatic wrapper construction. They use positions of particular strings and their method is known as HLRT (Head Left-Right Tail) approach.

The left - right strategy uses tags such as $\langle B \rangle \dots \langle =B \rangle$, $\langle I \rangle \dots \langle =I \rangle$ etc. $\langle P \rangle$ is used to identify head of the page. Since $\langle HR \rangle$ separates last tuple from tail it is used in identifying tail of the page.

The methods for extracting structured data from web pages [2] only extract database values from template generated web page, does not consider annotation. The template generated pages will have a general layout. During first stage it recognizes token set associated with the same type constructor in the template which is used to create the input pages. In the Analysis stage it uses the above sets to deduce the template. This template is then used to extract the values.

The effort to automatically assign labels that is Automatic Annotation of Data Extracted from Large Web Sites [3] depends on important information about semantics of data. And this work assumes the semantics of data will be available on web pages itself. This approach assumes that published data are followed by textual description to help human user. The textual description means conveying the real world meaning of entity. The main drawback of this work is limited applicability because many WDBs encode the data units without their labels.

One of the methods for split SRR is Harvesting Relational Tables from Lists on the Web [4]. This work is domain

independent and works in an unsupervised manner. Web pages contain data structured in lists these can be split into multiple columns.

Phase 1: (Splitting) it first split the individual lines into multiple fields.

Phase 2: (Alignment) construct a table by determining single number of likely column in output table. Records having too many fields are re-merged and re-split. Records with too few fields are expanded by inserting null values.

Phase 3: (Refinement) the field assignments in table are analysed to detect and fix those that are likely to be incorrect.

The DeLa [6] most similar to our work. The alignment process of this approach depends only on HTML tags. It uses a regular expression based data tree algorithm for alignment purpose. It sends query through HTML forms, automatically generate regular expression wrapper to extract data objects from result page, form crawler extracts form elements and uses as labels, restore retrieved data into an annotated table. Main disadvantage is label set are predefined so only small number of values are available.

The effort for automatic wrapper generation, Fully Automatic Wrapper Generation For Search Engines [7], focus on how to extract search result records from dynamically generated results pages returned in response to submitted queries. It uses visual content feature on the result page as displayed on a web browser and HTML tag structure of HTML source file of result page. Each SRR is stored in a tree structure.

More efficient automatic annotation, Annotating Search Results from Web Databases [1] in which a clustering-based shifting technique is used to align data units and an annotation wrapper is generated which describes how to extract data units and what semantic label should use. Their data alignment approach differs from the previous works. They handle all types of relationships between text nodes and data units such as

- One-to-One Relationship: In this, each text node contains exactly one data unit. We refer to such type of text nodes as atomic text nodes.
- One-to-Many Relationship: In this multiple data units are encoded in one text node. Such kind of nodes is known as composite text node.
- Many-to-One Relationship: In this multiple text nodes together form a data unit.
- One-To-Nothing Relationship: The text node belongs to this category will be displayed in a certain pattern across all SRRs. These are called template text nodes.

The existing approaches consider only one-to-one, one-to-many relationships. Second, they use a variety of features together. Third, they launch a new clustering-based shifting algorithm to perform alignment.

Our work is based on the paper [1]. The difference from their work is we did not use Wise integrator concept to combine search interfaces. They only deal with multiple web databases of same domain.

We are dealing with multiple domains with multiple web databases. For alignment we use tag based clustering and we use DOM tree structure for splitting data and tags. For annotation we use named entity recognition technique, and can be annotated more precisely.

III. DATA ALIGNMENT AND WRAPPER GENERATION

Each SRR has a tag structure which determines how the contents of the SRRs are displayed on a web browser. Each node in the tag structure is a tag node or a text node. A tag node is an HTML tag surrounded by "<" and ">" and a text node is the text outside the "<" and ">". Text nodes are the visible and it contain data units. From Fig. 1, text nodes are not always identical to data units.

A. System Architecture

Fig 2 shows the proposed system architecture. Each web database will give their URLs and by using these URLs we will extract information from SRRs. This information split into tag and data b using DOM tree Structure. These are stored and perform a clustering based on similarity of tags. Generate rules by using Named Entity Recognition technique and use recognized entities as labels. Generate rules for annotation for each web databases. When user searches for a specific key, information is extracted and b using annotation wrapper the data units are identified and annotated. The annotated results from multiple domains will be displayed.

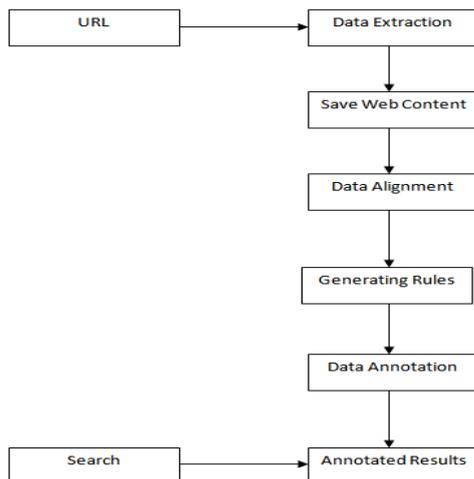


Fig.2: System Architecture

B. Alignment

After extracting data from web pages, we use DOM tree to split the data into tags and data. We will find similarity of tags to find similar data. Based on this tag similarity we will perform clustering. The output of clustering will be the aligned groups. Alignment process consists of following steps:

Step 1: Removing style information web pages. It is a data cleaning process.

Step 2: Align tags: In this step text nodes are aligned into groups so that each group contains the text nodes with the same concept.

Step 3: Align data units: This step groups the data units corresponding to their tag based clusters.

Algorithm 1: Align (HTML document doc)

- 1: Parse HTML document
- 2: Separate data and tags
- 3: Save tag and corresponding data in hash table
- 4: Repeat until hash table is empty
 - 4.1: Check the similarity of tags
 - 4.2: Group data of similar tags

C. Annotation

The data units of the same concept often share some common features, and these common features are associated with the data units. We use named entity recognition technique to assign labels. We utilize six basic annotator concepts form [1] to write rules:

1) Query-Based Annotator (QA) on the search interface a query with a set of query terms are submitted, first it finds the group that has the largest total occurrences of these query terms and then assign global name as the label to the group.

2) Frequency-Based Annotator (FA) The frequency - based annotator finds common preceding units shared by all the data units of the group. This is found easily by checking the tags.

3) In-Text Prefix/Suffix Annotator (IA) The in-text prefix/suffix annotator checks if all data units in the aligned group have the same prefix or suffix. If there is same prefix and it is not a delimiter, then it is removed from the group and is used to annotate values following it. If the same suffix is identified and if the number of data units having the same suffix and the number of data units inside the next group matches, the suffix is used to annotate the data units inside the next group.

4) Common Knowledge Annotator (CA) Human users can understand some data units on the result page because of the common knowledge. Common concept contains a label and a set of patterns or values these can be used for annotation. Example price related data units will have "Rs" this is a common knowledge.

D. Named Entity Recognition Technique

By using this technique we can write rules to identify the entities in clusters. By observing the tags and common behaviour of clusters writing rules become easy. The identified entities used as labels and these are stored in a database.

E. Annotation Wrapper

Once the data units are annotated, we generate an annotation wrapper by using these annotated data units. The new SRRs retrieved from the same WDB can be annotated using this wrapper quickly without repeating the entire annotation process. This annotation wrapper describes how to extract data units and what should be the label.

To annotate a new result page the wrapper can be used, the annotation rules are applied on each data unit in an SRR, one by one depending on the order in which they appear in the wrapper. The rule is matched if this data unit has the

same prefix and suffix as stated in the rule, and then data unit is labelled with the given label in the rule.

IV. PERFORMANCE EVALUATION

To evaluate the performance of our methods we use the precision and recall measures from information retrieval. For evaluation we selected wdb from four domains: Books, Electronics, Job and Music. Three wdb from each domain are selected. We choose two groups of keywords, one is domain specific and other is general keywords. The collected result pages are split into two groups. The first group contains 144 result pages and is used for training, and the second group has 70 result pages and is used for testing. Data set S1 is formed by obtaining sample result pages from each WDB by using two kinds of keywords. Testing data set S2 generated by collecting sample result pages from each testing site using different queries. The data alignment using tag based clustering improved precision by 1.3 points. The annotation by using the named entity recognition improved precision by 2.5 points. The precision and recall rates of Data Alignment and Data Annotation are shown in table I & II. From graphical representation of results it is obvious that as precision increases the recall decreases.

V. CONCLUSION

TABLE I. PERFORMANCE OF DATA ALIGNMENT

DOMAIN	PRECISION	RECALL
BOOKS	99.5%	96.3%
ELECTRONICS	100%	100%
JOB	100%	98.44%
MUSIC	100%	99.2%
OVERALL AVG.	99.875%	98.485%

TABLE II. PERFORMANCE OF DATA ANNOTATION

DOMAIN	PRECISION	RECALL
BOOKS	99.5%	92.59%
ELECTRONICS	100%	96.19%
JOB	99.5%	97.95%
MUSIC	99.5%	95%
OVERALL AVG.	99.6%	95%

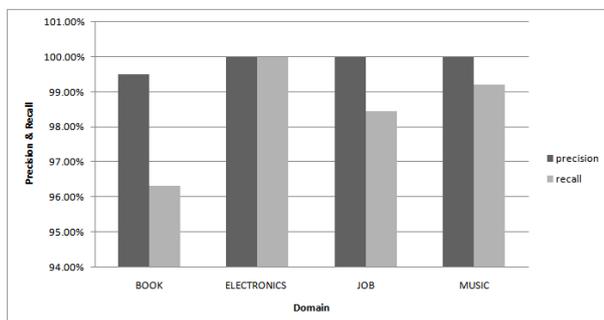


Fig.3: Evaluation of Data Alignment

In this paper, we studied the data annotation problem. To solve the automatic annotation problem, we proposed an annotation approach which automatically construct an annotation wrapper for annotating the SRRs retrieved from any given web database. Since we use only use recognized entities as the labels for annotation, there is only a limited number of labels. Also annotation repair is quite difficult in our work, so we would like to improve our wok by finding a method for annotation wrapper repair.

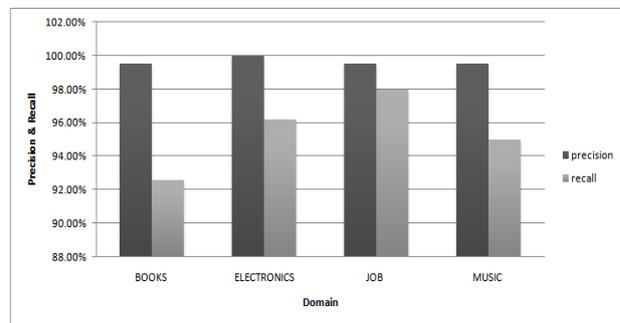


Fig.4: Evaluation of Data Annotation

ACKNOWLEDGMENT

The authors would like to express their gratitude to all who have taken particular interest on our work and supported through out. The authors would also like to express their gratitude to the anonymous reviewers for providing very constructive suggestions to improve the manuscript.

REFERENCES

- [1] Yiyao Lu, Hai He, Hongkun Zhao, Weiyi Meng and Clement Yu "Annotating Search Results from Web Databases"
- [2] A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages" Proc. SIGMOD Intl Conf. Management of Data, 2003.
- [3] L. Arlotta, V. Crescenzi, G. Mecca, and P. Merialdo, Automatic Annotation of Data Extracted from Large Web Sites, Proc. Sixth Intl Workshop the Web and Databases (WebDB), 2003.
- [4] H. Elmeleegy, J. Madhavan, and A. Halevy, Harvesting Relational Tables from Lists on the Web, Proc. Very Large Databases (VLDB) Conf., 2009.
- [5] N. Krushmerick, D. Weld, and R. Doorenbos, Wrapper Induction for Information Extraction, Proc. Intl Joint Conf. Artificial Intelligence (IJCAI), 1997.
- [6] J. Wang and F. H. Lochovsky, Data Extraction and Label Assignment for Web Databases, Proc. 12th Intl Conf. World Wide Web (WWW), 2003.
- [7] H. Zhao, W. Meng, Z. Wu, V. Raghavan, and C. Yu, Fully Automatic Wrapper Generation for Search Engines, Proc. Intl Conf. World Wide Web (WWW), 2005.
- [8] L. Liu, C. Pu, and W. Han, "XWRAP: An XML-Enabled Wrapper Construction System for Web Information Sources," Proc. IEEE 16th Intl Conf. Data Eng. (ICDE), 2001.
- [9] Thushara K P and Varsha Philip, "Mining Annotated Search Results From Web Databases" International Journal of Computer Science and Information Technologies volume 06, 2015.

BIOGRAPHY



Thushara K P currently doing M Tech in Computer Science and Engineering. She completed her B Tech in Information Technology. Her research interests include web mining, and information extraction.