A Survey Paper on the Method for Automatic Data Extraction and Alignment from Web Database.

Ms. Prital Bhure
G.S.Moze College of Engineering, Balewadi, Pune-45.
University Of Pune, Pune, India.

Prof. Ratnaraj Kumar
G.S.Moze College of Engineering, Balewadi, pune-45.
University Of Pune, Pune, India.

Abstract

The World Wide Web (www) serves a huge, widely distributed, global information service. Much information presents in the form of a web record which exists in both detail and list pages. Due to the increase of online web databases, it is required to get useful required information which is to be formatted before presenting the users which is one of the web information extraction (WIE) tasks. Using web query interfaces information is retrieved and is enwrapped in the web pages in the form of data records. There are large numbers of manually constructed, supervised, semi supervised and unsupervised WIE systems are proposed and developed. Depending on the end user query, the query results are generated by web databases. And from this query results pages, the automatic extraction of data is very important nowadays applications like data integration which need to cooperate with multiple web databases. There are many approaches presented previously for web data extraction such as wrapper induction and automatic methods; however each of them having limitations. Thus in this paper, we are presenting the new approach called CTVS which combines tag and value similarity for the efficient data extraction and alignment.

Keywords– *Data extraction, automatic wrapper generation, data record alignment.*

1. Introduction

Data mining refers to extracting hidden and valuable Knowledge and information from large data bases. It involves method s and algorithms to extract knowledge from different data repositories such as transactional databases, data warehouses text files, and www etc (as sources of data).

World Wide Web (WWW) is a vast repository of interlinked hypertext documents known as web pages. A hypertext document consists of both, the contents and the hyperlinks to related documents. Users access these hypertext documents via software known as web browser. It is used to view the web pages that may contain information in form of text, images, videos and other multimedia. The documents are navigated using hyperlinks, also known as Uniform Resource Locators (URLs).

Since its inception in 1990, World Wide Web has grown exponentially in size. As of today, it is estimated that it contains approximately 50 billion publicly accessible / index able web documents distributed all over the world on thousands of web servers. It is very difficult to search information from such a huge collection of web documents on World Wide Web as the web pages/documents are not organized as books on shelves in a library, nor are web pages completely catalogued at one central location. It is not guaranteed that users will be able to extract information even after knowing where to look for information by knowing its URLs as Web is constantly changing. Therefore, there was a need to develop information extraction tools to search the required information from WWW. Web information extraction tools are divided into three categories as follows:

- Web directories

- Meta search engines

- Search engines

Deep web consist of Online databases or web databases. The pages in the deep web are dynamically generated in response to a user query submitted through the query interface of a web database. When a web database receiving a user's query , it returns the relevant data, either structured or semi structured, encoded in HTML pages. There are some web applications, for example metaquerying, data integration and comparison shopping, need the data from multiple web databases. Automatic data extraction is necessary for such web applications to further utilize the data embedded in HTML pages. The data can be compared and aggregated, only when the data are extracted and organized in a structured manner, such as tables. Hence, accurate data extraction is important for these applications to perform correctly.

In this paper, our focus is on the problem of automatic data extraction from the query results page. In general, a query result page consist of the actual data as well as other information, such as navigational panels, advertisements, comments, information about hosting sites, and so on. The aim of web database data extraction is to remove any irrelevant information from the query result page, extract the query result records (QRRs) from the page, and align the extracted QRRs into a table such that the data values belonging to the same attribute are placed into the same table column.

2. Related Work

We can define the main problem for data extraction from web pages is that irrelevant, inappropriate, and inefficient methods of data extraction. A wrapper is a program that extracts data from a Web site or page and put them in a database. Originally a wrapper was defined as a component in an information integration system which aims at providing a single uniform query interface to access multiple information sources. In an information integration system, a wrapper is generally a program that "wraps" an information source (e.g. a database server, or a Web server) such that the information integration system can access that information source without changing its core query answering mechanism. In the case where the information source is a Web server, a wrapper must query the Web server to collect the resulting pages via HTTP protocols, perform information extraction to extract the contents in the HTML documents, and finally integrate with other data sources. Among the three procedures, information extraction has received most attentions and some use wrappers to denote extractor programs.

Wrapper induction (WI) or information extraction (IE) systems are software tools that are designed to generate wrappers. A wrapper usually performs a pattern matching procedure (e.g., a form of finite-state machines) which relies on a set of extraction rules. Tailoring a WI system to a new requirement is a task that varies in scale depending on the text type, domain, and scenario. To maximize reusability and minimize maintenance cost, designing a trainable WI system has been an important topic in the research fields of message understanding, machine learning, data mining, etc.
There are two main approaches to wrapper generation. The first approach is wrapper induction, which uses supervised learning to learn data extraction rules from a set of manually labeled positive and negative examples. Manual labeling of data is, however, labor intensive and time consuming. Additionally, for different sites or even pages in the same site, the manual labeling process needs to be repeated because they follow different templates/patterns. It mines data records in a page and extracts data from the records automatically.

The second approach is automatic extraction. In [10], a study is made to automatically identify data record boundaries. The method is based on a set of heuristic rules, e.g., highest-count tags, repeating-tags and ontology-matching. [3] Proposes a few more heuristics to perform the task without using domain ontology. However, [8] shows that these methods produce poor results. In addition, these methods do not extract data from data records. [8] proposes a method to find patterns from the HTML tag string of a page, and then use the patterns to extract data items. The method uses the Patricia tree and sequence alignment to find inexact matches. However, [8] shows that its performance is also weak. Our new method does not use tag strings for alignment but trees, which exploits nested tree structures to perform much more accurate data extraction. [5] also gives a set of heuristics to find individual product information, e.g., price and others. In [2and9], two more techniques are proposed. However, they need to use multiple pages (which are assumed to be given) that contain similar data records from the same site to find patterns or grammars from the pages to extract data records. Assuming the availability of multiple pages containing similar data records is a serious limitation. Our method works on each single page.

[7] proposes another method for data extraction. Its main idea is to utilize the detailed data in the page behind the current page to identify data records. It is common that a page with multiple data records does not contain the complete information of each data record. Instead, a link is normally used to point to the page with complete details. For example, a product record normally has a link pointing to the page that contains the detailed description of the product.

Furthermore, the method in [7] assumes that detail pages are given (in their experiments such pages are manually identified), which is not realistic. Due to a large number of links in a typical Web page, automatically identifying the correct links that point to detail pages is not a trivial task.

3. Existing System

CTVS method extracts the QRRs from a query result page p in two steps.

1. Record extraction

2. Record alignment

In first step it identifies data region and performs actual segmentation step.

In second step it aligns data values of the QRRs in p into a table.

1. QRR Extraction

Fig. 1 shows the framework for QRR extraction. Initially a query result page is given, then a tag tree is constructed for the page in <HTML> tag by the Tag Tree Construction module. Each node represents a tag in the HTML page and the tags enclosed inside it represents its children. Each internal node n of the tag tree has a tag string tsn, which includes the tags of n and all tags of n's descendants, and a tag path tpn, which includes the tags from the root to n. Next, all possible data regions are identified by the Data Region Identification module. Data regions usually contain dynamically generated data, top down starting from the root node. This identified data regions are then segmented into data records according to the tag patterns in the data regions by the Record Segmentation module. The segmented data records are given to the Data Region Merge module which merges the data regions containing similar records. Finally, the Query Result Section Identification module selects one of the merged data regions as the one that contains the QRRs.

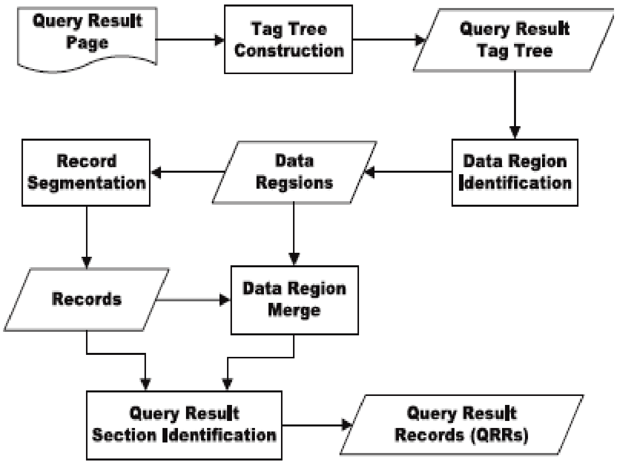

Figure 1: QRRs Extraction Framework

2. QRR Alignment

QRR alignment is performed by a novel three-step data alignment method that combines tag and value similarity.

1. Pairwise QRR alignment aligns the data values in a pair of QRRs to provide the evidence for how th e data values should be aligned among all QRRs.
2. Holistic alignment aligns the data values in all the QRRs.
3. Nested structure processing identifies the nested structures that exist in the QRRs

3.1. Limitations of Existing Methods

(1) The algorithm used for merging data regions(MDR algorithm) makes use of the HTML tag tree of the Web page to extract data records from the page. However, erroneous tags in the HTML source of some pages make it hard to build correct trees, which make it impossible to find correct data records in these pages.

(2) A single data record may be composed of multiple sub-trees. Due to noisy information, MDR may find wrong combinations of sub-trees.

(3) Another problem with most existing approaches is that they assume that the relevant information of a data record is contained in a contiguous segment of the HTML code. This is not always true.

4. Proposed Solution

We present a data extraction and alignment method called CTVS that combines both tag and value similarity. We propose some improvements in the existing CTVS method. The working of CTVS is that it automatically extracts data from query result pages by first identifying and segmenting the query result records (QRRs) in the query result pages and then aligning the segmented QRRs into a table, in which the data values from the same attribute are put into the same column. In this paper, we propose new techniques to handle the case when the QRRs are not contiguous, which may be due to the presence of auxiliary information, such as a comment, recommendation or advertisement, and for handling any nested structure that may exist in the QRRs.

One of the limitation of existing CTVS method is that, it requires at least two QRRs in the query result page. We propose solution to improve this limitation by comparing the similarity of even a single QRR with tag tree. Also we design a new algorithm for merging data regions which overcome the problem caused by existing MDR algorithm.

5. Conclusions

In this paper we survey some web data extraction methods. We studied about the method, which is combining tags and value similarities, CTVS. This method performs automatic extraction of QRRs from a query result page. CTVS does automatic extraction in two steps for. The first step is identification and segmentation of the QRRs. The improvement in our technique is that we allow QRRs in a data region to be non-contiguous. In the second step alignment of the data values takes place among the QRRs. In this paper we propose a novel alignment method which performed the alignment in three consecutive steps that are pairwise alignment, holistic alignment, and nested structure processing.

6. REFERENCES

[1]Base Paper: Weifeng Su, Jiying Wang, Frederick H. Lochovsky, Member, IEEE Computer Society, and Yi Liu, Combining Tag and Value Similarity for Data Extraction and Alignment.

[2]A. Arasu and H. Garcia-Molina, "Extracting Structured Data from Web Pages," Proc. ACM SIGMOD Int'l Conf. Management of Data, pp. 337-348, 2003.

[3]P. Bonizzoni and G.D. Vedova, "The Complexity of Multiple Sequence Alignment with SP-Score that Is a Metric," Theoretical Computer Science, vol. 259, nos. 1/2, pp. 63-79, 2001

[4]C.H. Chang and S.C. Lui, "IEPAD: Information Extraction Based on Pattern Discovery," Proc. 10th World Wide Web Conf., pp. 681- 688, 2001.

[5]D.W. Embley, D.M. Campbell, Y.S. Jiang, S.W. Liddle, D.W. Lonsdale, Y.-K. Ng, and R.D. Smith, "Conceptual-Model-Based Data Extraction from Multiple-Record Web Pages," Data and Knowledge Eng., vol. 31, no. 3, pp. 227-251, 1999.

[6]N. Kushmerick, "Wrapper Induction: Efficiency and Expressiveness," Artificial Intelligence, vol. 118, nos. 1/2, pp. 15-68, 2000.

[7]B. Liu and Y. Zhai, "NET - A System for ExtractingWeb Data from Flat and Nested Data Records," Proc. Sixth Int'l Conf. Web Information Systems Eng., pp. 487-495, 2005.

[8]L. Liu, C. Pu, and W. Han, "XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources," Proc. 16th Int'l Conf. Data Eng., pp. 611-621, 2000.

[9]V. Crescenzi, G. Mecca, and P. Merialdo, "Roadrunner: Towards Automatic Data Extraction from Large Web Sites," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 109-118, 2001.

.

[10] A.V. Goldberg and R.E. Tarjan, "A New Approach to The Maximum Flow Problem," Proc. 18th Ann. ACM Symp. Theory of Computing, pp. 136-146, 1986.