METHODS FOR EXTRACTION AND ALIGNMENT OF DATA ON DISPLAYED PRODUCTS IN WEB PAGES

Neha Chungde[1], H. K. Chavan[2]
#Information Technology, Terna Engineering College, Mumbai University
Sector-22, Nerul, Navi Mumbai, Maharashtra, India
[1]nehavchungde@gmail.com
[2]chavan.hari@gmail.com

Abstract — **Data extraction which is important for many applications extracts the record from HTML files. The existing Machine learning method requires human labeling of web sites for extracting data from web pages. So, this process is very time consuming. Automatic pattern discovery method enables inaccurate alignment of multiple data records in web pages. Taken into consideration the limitations of data extraction methods. Many applications necessitate the automatic extraction of data from the query result pages. The proposed method of data extraction using Web extraction tool automatically extracts data from query result pages. The data extracted using web extraction tool is aligned in a structured format using Cosine-Similarity.**

Keywords—Data Extraction, automatic wrapper generation, information integration, Web Crawler.

I. INTRODUCTION

Web pages are accessed by a unique URL. These web pages are dynamically generated in response to a query submitted through the query interface of a web database. Upon receiving a query, a web database returns the relevant data, either structured or semi structured, encoded in HTML pages [1]. Many web applications, such as Meta querying, data integration need the data from multiple web databases. For these applications to further utilize the data embedded in HTML pages, automatic data extraction is necessary [2]. Only when the data are extracted and organized in a structured manner, such as tables, they can be aggregated. Hence, accurate data extraction is vital for these applications to perform correctly. A large amount of information on the Web is contained in regularly structured objects, which we call data records [6]. Such data records are important because they often present the essential information of their host pages, e.g., lists of products or services.

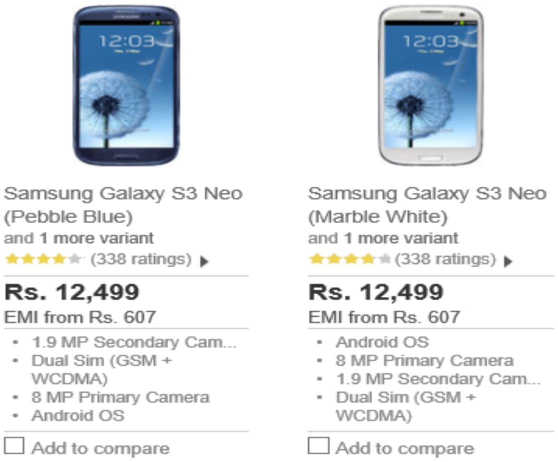
In general, a query result page contains not only the actual data, but also other information, such as navigational panels, advertisements, comments, information about hosting sites, and so on. The goal of data extraction method is to remove any irrelevant information from the query result page, extract the query result records (referred to as QRRs) and align the extracted QRRs into a table.

Usually, such software programs simulate human exploration of the World Wide Web by either implementing low-level Hypertext Transfer Protocol (HTTP), or embedding a fully-fledged web browser, such as Internet Explorer or Mozilla Firefox.

Web scraping is closely related to web indexing, which indexes information on the web using a bot or web crawler and is a universal technique adopted by most search engines. In contrast, web scraping focuses more on the transformation of unstructured data on the web, typically in HTML format, into structured data that can be stored and analyzed in a central local database or spreadsheet. Web scraping is also related to web automation, which simulates human browsing using computer software. Uses of web scraping include online price comparison, contact scraping, weather data monitoring, website change detection, research, web mashup and web data integration.

Traditionally, many web sites consist of large sets of pages generated using common templates. On the other hand, web databases consist of various result records which are encoded in HTML files. Existing methods have some limitations for data extraction and alignment. Machine learning method requires human labeling of web sites for extracting data from web pages. So, this process is very time consuming. Automatic pattern discovery method enables inaccurate alignment of multiple data records in web pages.

The existing method consists of identifying individual data records in a page and extracting data items from these identified data records. The goal of data extraction method is to remove any irrelevant information from the query result page, extract the query result records (referred to as QRRs) and align the extracted QRRs into a table. The data values belonging to the same attribute are placed into the same table column.



Samsung Galaxy S3 Neo (Pebble Blue)
and 1 more variant
★★★★☆ (338 ratings)
Rs. 12,499
EMI from Rs. 607
- 1.9 MP Secondary Cam...
- Dual Sim (GSM + WCDMA)
- 8 MP Primary Camera
- Android OS

□ Add to compare

Samsung Galaxy S3 Neo (Marble White)
and 1 more variant
★★★★☆ (338 ratings)
Rs. 12,499
EMI from Rs. 607
- Android OS
- 8 MP Primary Camera
- 1.9 MP Secondary Cam...
- Dual Sim (GSM + WCDMA)

□ Add to compare

Fig 1. An Example Data Extraction

The Rest of paper is organized as follows: Section II consists of the related work. In section III we present record extraction methods. Section IV consists of Record Alignment method and conclude the paper in section V.

II. RELATED WORK

Data Extraction Methods automatically extracts data from query result pages by identifying and segmenting the query result records (QRRs) [1]. It focuses on alignment of query result records. It also focuses on the problem of automatically extracting data records that are encoded in the query result pages generated by web databases

Arasu and H. Garcia-Molina [2] presented an algorithm that takes an input, a set of template generated pages, deduces the unknown template used to generate the pages, and extracts, as output, the values encoded in the pages.

ODE: Ontology-assisted Data Extraction [3] automatically extracts the query result records from the HTML pages. The ontology (i.e., a conceptual model instance) describes the data including relationships, lexical appearance and context keywords. ODE is extremely accurate for identifying the query result section in an HTML page, segmenting the query result section into query result records, and aligning and labeling the data values in the query result records.

Structured Data Extraction from the Web Based on Partial Tree Alignment [4] focuses on a novel partial alignment technique based on tree matching. Partial alignment means that we align only those data fields in a pair of data records that can be aligned. This approach enables very accurate alignment of multiple data records.

R. Baumgartner and et al [5] focuses on a system Lixto which presents the techniques for supervised wrapper generation and automated web information extraction. A program that makes an existing website look like a database is called a wrapper. Lixto can generate wrappers which translate relevant pieces of HTML pages into XML.

Mining Data Records in Web Pages [6] focuses on two observations about data records on the Web and a string matching algorithm. This technique is able to mine both contiguous and noncontiguous data records.

Chen and *et al* [7] proposed a novel technique for the identification of table structures in HTML documents. This identification technique is then used to automatically generate composite wrappers for applications requiring distributed resources.

A Flexible Learning System for Wrapping Tables and Lists in HTML Documents [8], it consists of wrapper learning system (WL) that can exploit several different representations of document such as two-dimensional geometric views of the page (for tabular data) and representations of the visual appearance of text.

III. RECORD EXTRACTION METHODS

Mainly web data extraction consists of HTML Entity Extraction using web harvest tool method. Fig. 2 shows the framework for QRR extraction. Given a query result page, the Tag Tree Construction module first constructs a tag tree for the page rooted in the <HTML> tag. Each node represents a tag in the HTML page and its children are tags enclosed inside it. 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, the Data Region Identification module identifies all possible data regions, which usually contain dynamically generated data, top down starting from the root node. The Record Segmentation module then segments the identified data regions into data records according to the tag patterns in the data regions. Given the segmented data records, the Data Region Merge module 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.

A. Existing Method for Data Extraction

1. Wrapper Induction Extraction method:

This method is useful in systems where the resource information is formatted for use by people and so it is difficult to extract their content mechanically. So a technique for constructing wrappers automatically from labelled examples of a resource's content is introduced called wrapper induction. In these methods there is requirement of human assistance for building a wrapper. The user labels the items present in a set of training pages or in a list of data records on page, which are to be extracted as target items. Then the system learns the wrapper rules from the labelled or marked data and further uses them to extract records from new pages. The wrapper rule is made up of two patterns. One is prefix pattern which denotes the beginning of the target item and other is suffix pattern which denotes the end of the target item.

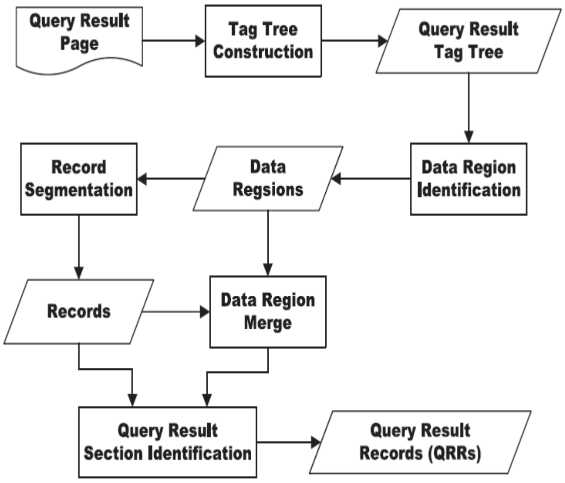


Fig 2. QRR Extraction Framework

Advantage: - As only required items of interest are labelled, no extra data are extracted.

Disadvantages:-

- Manual labelling of data is time-consuming.
- It is not scalable to a large number of web databases.
- Existing wrapper gives poor performance when the format of a query result page changes, which happens more frequently on the web.
- Continuous monitoring is needed to keep track on changes in format of pages and maintaining a wrapper when a page's format changes.

B. Proposed Method for Data Extraction

Some methods, have been proposed to automatically extract the data from the query result pages. These methods rely entirely on the tag structure in the query result pages.

This method utilizes the processes like Data Parsing, HTML Entity Extraction based on the input page structure displaying a set of products within the application.

1. Data Parsing:

Data parsing complete process is shown below in Fig. 3. Data parsing is the process of analyzing string made up of sequence of tokens to determine structure in the tree format of the set of products displayed within the application.

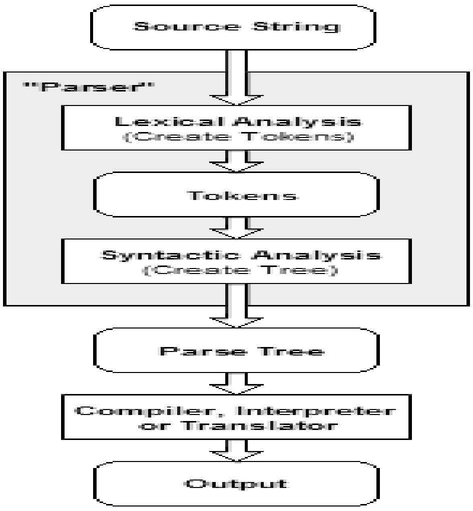
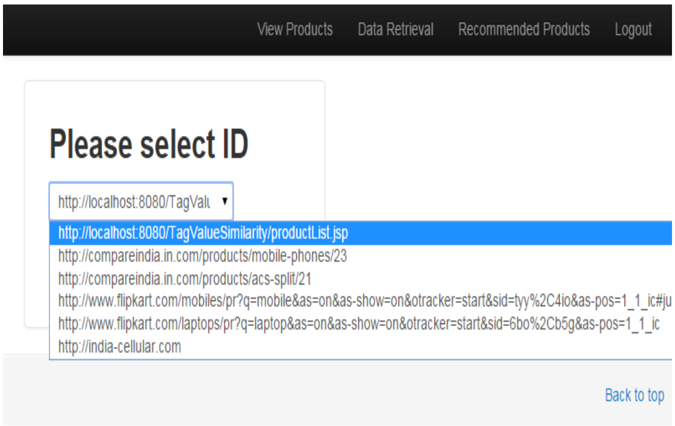

Source String
"Parser"
Lexical Analysis (Create Tokens)
Tokens
Syntactic Analysis (Create Tree)
Parse Tree
Compiler, Interpreter or Translator
Output

Fig 3. Data parsing method

2. HTML Entity Extraction using Web Harvesting tool:

Web data extraction tool leverages well proved XML/HTML and text processing technologies in order to easily extract useful data from arbitrary web pages. As per the Fig. 1, the Web harvesting tool extract data from query result page mobile provided in HTML format of various categories of mobile available depending on their distinct attributes and properties and further align this attributes in structured format depending on their similarities using cosine similarity method.



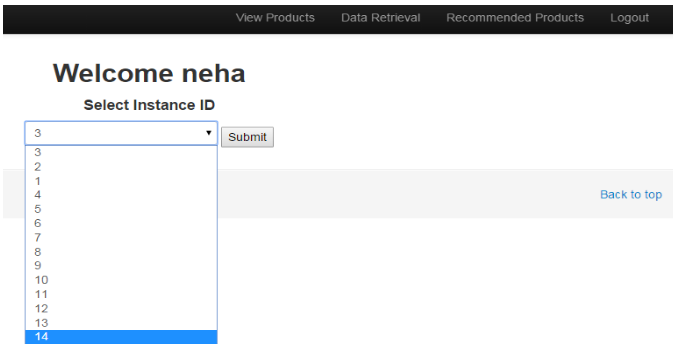
Please select ID

http://localhost:8080/TagValu

http://localhost:8080/TagValueSimilarity/productList.jsp
http://compareindia.in.com/products/mobile-phones/23
http://compareindia.in.com/products/acs-split/21
http://www.flipkart.com/mobiles/pr?q=mobile&as=on&as-show=on&otracker=start&sid=tyy%2C4io&as-pos=1_1_ic#ju
http://www.flipkart.com/laptops/pr?q=laptop&as=on&as-show=on&otracker=start&sid=6bo%2Cb5g&as-pos=1_1_ic
http://india-cellular.com

Fig 4. Data Extraction from Particular site or Stored Database

From Fig 4. It is observed that by selecting one link we can view the products depending upon the data retrieved from this website using web harvesting tool.



Welcome neha

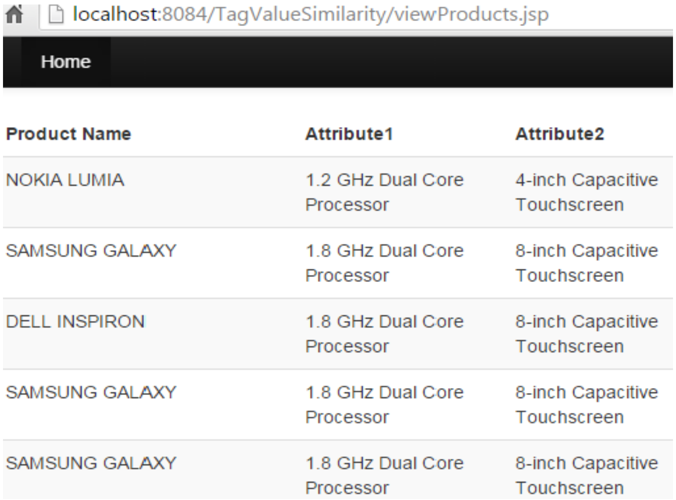
Select Instance ID

3

3
2
1
4
5
6
7
8
9
10
11
12
13
14

Fig 5. Instance Id Data Extraction

From Fig 5. after extracting the data it is stored in database and we can select instance id and depending on this instance id we can view all the product with that instance id available in the database.



Product Name	Attribute1	Attribute2
NOKIA LUMIA	1.2 GHz Dual Core Processor	4-inch Capacitive Touchscreen
SAMSUNG GALAXY	1.8 GHz Dual Core Processor	8-inch Capacitive Touchscreen
DELL INSPIRON	1.8 GHz Dual Core Processor	8-inch Capacitive Touchscreen
SAMSUNG GALAXY	1.8 GHz Dual Core Processor	8-inch Capacitive Touchscreen
SAMSUNG GALAXY	1.8 GHz Dual Core Processor	8-inch Capacitive Touchscreen

Fig 6. Products retrieved by selecting Instance Id

Advantage: As the data is extracted from particular website large amount of data and automatic data is extracted at one time.

Disadvantage: Due to large data extraction, random data is extracted so alignment is not proper.

IV. RECORD ALIGNMENT METHOD

A. Proposed Method of Alignment after Data Extraction using Cosine Similarity Method

Cosine Similarity method is used to compare the similarity between values of the two products. Once the attributes are extracted using extraction method structured alignment of this product values is done using cosine similarity method by comparing the similarities between the different values. Following are the steps to check the similarity between the different values of the product and the values depending on their attributes. The similarity s12 between two data values f1 and f2 with data type n1 and n2 is defined as

$$s_{12} = \begin{cases} 0.5 & n_1 = p(n_2)\ \&\ n_1 \neq String\ \ OR \\ & n_2 = p(n_1)\ \&\ n_2 \neq String \\ 1 & n_1 = n_2 \neq String \\ cosine\ similarity & n_1 = n_2 = String \\ 0 & otherwise, \end{cases}$$

Where p (ni) refers to the parent of ni in the data type tree.

V. CONCLUSIONS

We presented approaches to extract data from displayed products in web pages. We can summarize these web data extraction methods as follows: Existing method-Wrapper induction extraction method and proposed method-HTML entity extraction using web harvesting tool. Among the above discussed web data extraction methods, some techniques reveals flat records and some other techniques are trying to extracts nested records also.

Wrapper induction extraction method labels the items of interest, no extra data are extracted. HTML entity extraction using web harvesting tool extract large amount of data and automatic data from website at one time.

The data extracted using web extraction tool is aligned in a structured format using Cosine-Similarity.

REFERENCES

[1] 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] W. Su, J. Wang, and F.H. Lochovsky, "ODE: Ontology-Assisted Data Extraction," ACM Trans. Database Systems, vol. 34, no. 2, article 12, p. 35, 2009.

[4] Y. Zhai and B. Liu, "Structured Data Extraction from the Web Based on Partial Tree Alignment," IEEE Trans. Knowledge and Data Eng., vol. 18, no. 12, pp. 1614-1628, Dec. 2006.

[5] R. Baumgartner, S. Flesca, and G. Gottlob, "Visual Web Information Extraction with Lixto," Proc. 27th Int'l Conf. Very Large Data Bases, pp. 119-128, 2001.

[6] B. Liu, R. Grossman, and Y. Zhai, "Mining Data Records in Web Pages," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 601-606, 2003.

[7] L. Chen, H.M. Jamil, and N. Wang, "Automatic Composite Wrapper Generation for Semi-Structured Biological Data Based on Table Structure Identification," SIGMOD Record, vol. 33, no. 2, pp. 58-64, 2004.

[8] W. Cohen, M. Hurst, and L. Jensen, "A Flexible Learning System for Wrapping Tables and Lists in HTML Documents," Proc. 11thWorld Wide Web Conf., pp. 232-241, 2002.