ISSN No. 0976-5697

Volume 4, No. 8, May-June 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Design of Information Extraction System

J.Pavithra [1] R.Monisa [2] G.Ramya [3],Assistant Professor,
School of Information Technology and Engineering,
G.Priya[4],Assistant Professor,
School of computer science and Engineering,
VIT University, Vellore – 632014.
pavithra.j15@gmail.com, monisa90@gmail.com, engr_ramyaa@yahoo.com,gpriya@vit.ac.in

Abstract- It deals with the problem of extracting specific information from a collection of documents. Information extraction has become an essential task due to the vast growth of online information. It is defined as the process of particularly structuring and combining data that are stated in one or more documents and database. It describes about the basic system architecture which is involved in designing an information extraction system. It involves the basic steps of web information extraction such as organizing web page, generating rule and the result to be displayed. XML technology can be considered as a suitable approach for information extraction because; it reduces the difficulties in extracting information from huge amount of data. Various techniques and concepts are available in XML that can be used to extract information from a document and web page. For extraction purpose we need to design an XML template to capture the information needed; the extracted information has to be placed in the template. This paper focuses on extracting information from semi-structured data.

Key Terms: Information extraction, XML, DOM, XSLT, XPath.

I. INTRODUCTION

DOM is a programming API for HTML and XML documents. DOM provides a logical structure of documents. DOM can be used to manage data. It is used by programmers to create and build documents. Also it allows the user to perform modifications such as add, delete elements to the contents. It can be used to transfer a web page to a tree structure with HTML and XML tags. It allows the user to navigate through the structure. DOM is an object model. It has been used in the conventional object oriented design logic: objects are used to model the documents; the model comprises not only the structure but also the activities of the documents and the object.

DOM can be considered as an interface that allows updating the formation, subject of a document by making use of programs and scripts. A special property of DOM model is its structural isomorphism. It means that suppose if two DOM implementations are involved in creating an illustration of similar document it will result in creating exactly the similar formation with similar objects and relationship. Dom consists of two parts DOMCORE and DOMHTML. DOMCORE provides the functionalities for XML and it also acts as base for DOMHTML. DOMHTML provides the additional functionalities that are used in HTML.

Dom is used

a. To identify the interfaces and objects.
b. To identify the relationship between the objects and interfaces.
c. To identify the semantics of the interfaces and objects.

```
<TABLE>
<ROWS>
<TR>
<TD>Benz</TD>
<TD>ambassador</TD>
</TR>
<TR>
<TD>pulsar</TD>
<TD>scooty</TD>
</TR>
</ROWS>
</TABLE>
```

The document object model of the above table will be represented like this:

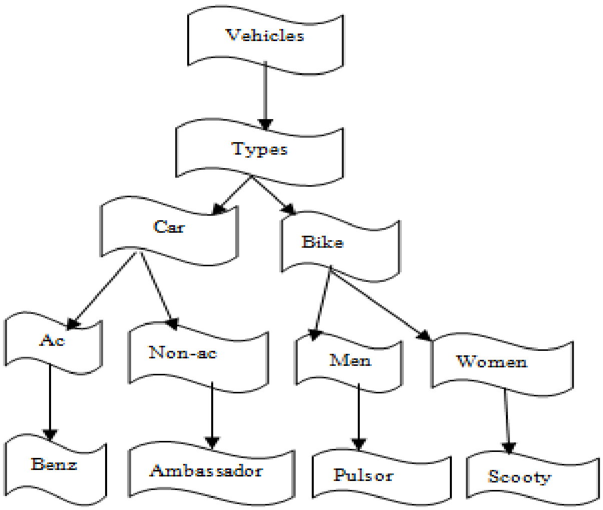

Figure1. DOM tree structure

II. XML AND XSLT

The Extensible Markup Language (XML) is a set of rules. These are used to define a tag that segregates document or web page into many partitions and sub partitions. It is used extensively with web applications. XML provides more freedom to the user for creating customized tags . It allows the user to create document and way of using is easier. SGML acts as base for XML's markup rules. XML is derived from SGML; it is a flexible text format. XML enables the report, communication and analysis of data between applications and also between organizations. XML is a meta language that allows the user to create and design documents of their own. It can be considered as a technique for placing structured information in a text file. Contents in XML should comprise three components:

a. XML document
b. Document Type Declaration(DTD)
c. Extensible Style sheet Language(XSL)

In the above components it is not mandatory that DTD and XSL should be present in all the situations. XML can be considered as essential because it solves two problems in web developments:

a. Dependence over a single , nonflexible document type
b. The complexity in SGML where the syntax provides trickier program options.

XSLT (Extensible Style sheet Language Transformations) is a language for converting XML documents into other XML document for WebPages using HTML. In XSLT two documents are used. The content of the original document will not be changed instead a new document is created related to the details in the existing original document. XSLT processor takes one or more modules from both XML source and XSLT and processes them with XSLT template processing engine to generate the output document.

III. LITERATURE SURVEY

Feihong zhuang et.al [1] proposed system describes information extraction is done automatically by extracting information from the huge collection of large scale domain independent data in a scalable manner. The objective of this paper to make use of the search engine for processing large amount of data we require in an error free manner. The paper focuses on major components like extractor, search engine interface and the database for producing large scale information.

Tao Xie et.al [2] proposed system focuses on extracting web information from complex structures and also in reducing the burden of user. This paper has produced a comprehensive model and framework which is used to combine data analysis and extraction according to user needs. Automatic web page analysis has been used in this paper.

Yan HU et.al [3] proposed system extraction has been carried by using standard XML technology. In extraction process two technologies are proposed first one is web data conversion technology based on XML, second one is generation of XPATH by DOM. It provides brief explanation about general web information extraction.

Mahmoud Shaker et.al [4] proposed system provides a framework for extracting, classifying and browsing semi structured web data sources to extract relevant information. The collected information is classified based on some standards of nokia products. Here they are used as a sample for information extraction by typing the product name in the query interface. It makes use of RIA to collect relevant data. It reduces the users complexity in searching.

Jong-seoak Jeong et.al [5] Proposed system defines information extraction with a web based information search system IHWA. It is used to collect information from semi-structured data sources. It also explains about the gatherer which is used to collect unstructured data web documents such as DTD unknown XML documents. The implemented extraction system provides java programming interface so that it can be integrated with other applications.

Chia–Hui chang et.al [6] proposed system focus on providing a survey of available web information extraction techniques. It also makes a comparative study of all the approaches in three dimensions: task domain, automatic degree and used techniques. First criteria focus on handling of web sites with a particular structure at the time of failure. Second criteria are based on the automation degree and the last one focus on the techniques in extraction.

Weicheng Xie et.al [7] proposed system uses extraction of data from the web by using XML and implemented by framework concept. In this paper a quality of data extraction from web is introduced using the concept of data mining. Data mining is used for extraction of information from the web in a repeated process and merges them to build a separate system

Chen Hong-Ping et.al [8] LBDRF algorithm which is used for solving the extraction of data records in deep from the web pages. In this paper DOM tree is used for capturing the web page and data region finder is used to locate the data in the recorded form which are used to extract data from that region

IV. INFORMATION EXTRACTION

Information extraction is the task of extracting structured data from semi-structured information. It focuses on retrieving from a particular set of data based on the domain and the user interest. The process of extracting information from semi-structured data is a tedious task because the data has to be converted to a standard structured format and from that extraction has to be carried out. IE is considered as an interested area nowadays as it evaluates and makes a comparative study of the available natural language processing technologies. IE makes use of two techniques to extract information. They are

a. Automatic information extraction
b. Wrapper induction

In this paper we are going to implement the concept of wrapper induction. It makes use of machine learning technology to extract the information. It develops rules based on machine learning. User can search information in web source by making use of these rules. In wrapper induction technique changes should be made according to the content and structure.

V. INFORMATION EXTRACTION SYSTEM ARCHITECTURE

In wrapper induction technique text is given as input and the output received will be in a proper format. It gets stored in a database. It involves five stages,

a. ***Gathering Data:*** In this process a sample HTML page is downloaded from internet. After classification it has to be placed in the domain information base.

b. ***Optimizing Pages:*** It involves optimizing the HTML pages. It involves two steps.

a) Page cleaning
b) Page analysis

Page cleaning involves removing the illegal characters, correcting errors. Semi structured HTML documents are converted to proper structured XHTML documents based on the XML standards. DOM tree is used for performing the conversion. Page analysis makes use of XML parser to parse the XHTML documents which is obtained into XML DOM tree structure.

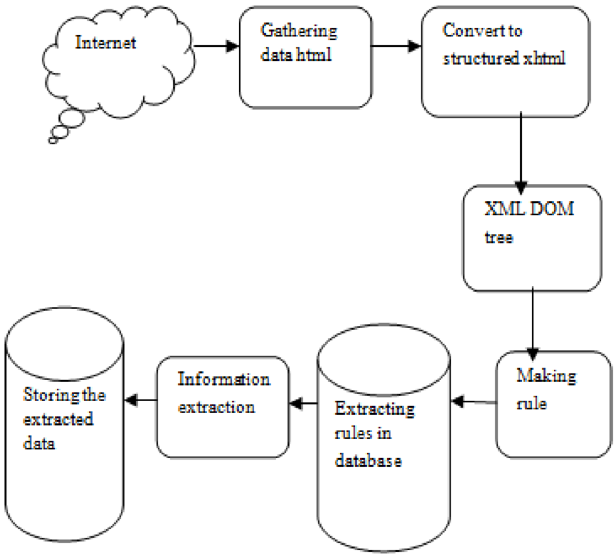
Internet
Gathering data html
Convert to structured xhtml
XML DOM tree
Making rule
Extracting rules in database
Information extraction
Storing the extracted data

Figure2. Architecture of Information extraction.

c. ***Making Rules:*** The pages obtained after optimization are considered as a sample and they are provided to the user to find their interested area of information. By making use of the user information path expression of the nodes to be extracted are known and they are combined with XSLT to frame the extraction rules. XPath is the major component of XSLT. It is a path expression and it is used to identify the content to be extracted and placed in XML document. User interested area is obtained by DOM based XPath expression and it is joined with XSLT to create the rules. The rules are represented in XML format.

d. ***Information Extraction:*** By making use of XSLT processor and extraction rules user can extract the information they require from HTML documents. If proper rule is not available for a page then new rules have to be created by verifying the HTML page again. It allows modifying the rules according to the web page structure. User can get necessary information from web pages by performing information extraction.

e. ***Storing Data:*** The extracted results are represented in XML format. Then the information is stored in a relational database. In this process of information extraction web pages of various size and structure are used, to support that we make use of relational database.

The above five stages explains the work flow of information extraction.

VI. CONCLUSION

In today's world with huge demand of gathering information from internet it is necessary to develop a system that supports information extraction. This paper supports the idea and makes use of wrapper induction technique to develop an information extraction system. For developing the system XML is used along with DOM and XSLT. In this paper XML technology is used because it is easy and efficient to develop web applications. It provides efficient extraction of data and supports modification of wrappers according to the content. This system provides usability and flexibility to make use of various wrappers.

VII. REFERENCES

[1]. Guntis Arnicans and Girts Karnitis, "Intelligent Integration of Information from Semi-Structured Web Data Sources on the Base of Ontology and Meta-Models", Riga, Lativia, 2006.

[2]. Chia-Hui Chang, Mohammed Kayed, A Survey of Web Information Extraction Systems, IEEE transactions on knowledge and data Engineering, TKDE-0475-1104.R3 1

[3]. Line Eikvil, "Information Extraction from World Wide Web A Survey", Technical Report 945, Norweigan Computing Center, 1999.

[4]. J. Hammer, H. Garcia-Molina, J. Cho, R. Aranha, and A. Crespo. Extracting Semi-structured data from the web. Proceedings of Workshop on Management Of Semi-structured Data, pages 18-25, 1997.

[5]. Chia-Hui Chang, Shao-Chen Lui, "IEPAD: information extraction based on pattern discovery," Proceedings of the 10th international conference on World Wide Web, pp.681-688, May 01-05, 2001, Hong Kong.

[6]. [MYL,01] Myllymaki, Jussi.Effective Web Data Extraction with Standard XML Technologies. International Journal of Computer and Telecommunication Networking In: 10th intl. World Wide Web Conf. Hong Kong May. 2001.

[7]. Web information extraction and its application Peng, Yan; Zhang, Chenyue. Cloud Computing and Intelligence Systems (CCIS), 2011 IEEE International Conference on Digital Object Identifier: 10.1109/CCIS.2011.6045107 Publication Year: 2011, Page(s): 448 - 451

[8]. Research on Web Information Extraction Based on XML Hu, Yan; Xuan, Yanyan Genetic and Evolutionary Computing, 2008. WGEC '08. Second International Conference on Digital Object Identifier: 10.1109/WGEC.2008.16 Publication Year: 2008, Page(s): 201 - 204

[9]. XML-based Web information extraction system design and implementation Jun, Ma; Tihong, Li Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on Volume: 8 Digital Object Identifier: 10.1109/ICCSIT.2010.5564746 Publication Year: 2010, Page(s): 551 - 554

[10]. Extracting information from semi-structured Internet sources Jeong, Jong-Seok; Oh, Dong-Ik Industrial Electronics, 2001. Proceedings. ISIE 2001. IEEE International Symposium on Volume: 2 Digital Object Identifier: 10.1109/ISIE.2001.931683 Publication Year: 2001, Page(s): 1378 - 1381 vol.2 .