Journal of Advances in Shell Programming

Volume 1, Issue 1
www.stmjournals.com

An Implementation of Extracting Data and Mining Extracted Data from Web Pages

Dhanashri Sandbhor, Archana Jadhav, Govind Kumar Dubey, Sonali Pawade, Rupesh Dhole*

Information Technology Department, Rajarshi Shahu College of Engineering, Tathawade, Pune-033, Pune University, Maharashtra, India

Abstract

Web contains various information of particular object, which could be relevant as well as non-relevant are called as data records. It is necessary to extract relevant information from web pages. Web data extraction is the system which is used for extracting data from various web pages. Data present on web pages are in un-structured format. In the process of data extraction, we convert un-structured data into structured format. This paper contains web data extraction system, stages of making a mashup and the data mining concept for data clustering. Mashup is the process which provides functionality such as data retrieval, data source modeling, data cleaning/filtering, data integration, and data visualization. We use "Xtractorz" system for data extraction and mashup for data records. By using data mining, we analyze data from different sources and using text mining we cluster all the information. And we also use those data for providing value added services

Keywords: *Web, data, mashup, data clustering*

****Author for Correspondence*** E-mail: shree.sandbhor@gmail.com

INTRODUCTION

Nowadays, we can easily search the information on internet, but whatever information we get from web pages is not completely relevant .Therefore this process of searching information will become more time-consuming, so we need various technologies to extract data. Normally web pages are in HTML format but these HTML pages are not suitable for database applications and cannot be easily represented in memory. So in this paper we are converting the unstructured HTML pages into structured XML or XHTML documents because XML pages are suitable for database applications to separate data structure from layout and represent easily [1]. When we fire the query to get relevant data from HTML pages, it requires huge amount of time and cost.

To reduce the time and cost for searching, we need a system for web data extraction and making a mashup. Mashup is a web application in which we combine the extracted data from various web pages. It provides various activities, i.e., data retrieval, data source modeling, data cleaning/filtering, data integration and data visualization. The need of mashup is to convert HTML document into XML or XHTML document. There are several problems while making mashup such as understanding the HTML content which is unstructured and difficult to sort out selected data from web pages because web data extraction process should not be done automatically. To overcome this problem, the authors designed the processes such as data extraction and making mashup to integrate data called Xtractorz by using programming such as JAVA [2].

LITERATURE REVIEW

A) Xtractorz

a) Deep Web Data Extraction

There are various technologies for data extraction but it has various limitations. The technologies automatically extract structured data from web page by identifying data region and segment into database. Data values and some records are put into same column.

b) Mashup

Mashup is the application which is used to combine data from different web pages. We can combine different robots in different mashups and can show combined result in a single mashup.

c) Robomaker

Robomaker is a technique used to implement programming robots, which perform the activity as input provided by user. It analyzes the requirement of user and on the basis of document. DOM tree represents and describes how all elements such as input field image paragraph are organized in HTML page.

e) SAX Parser

SAX parser is an event-based sequential access parser API. It is used to read data from an XML document that is alternative DOM (document object module) tree [3]. SAX parser operates on each part of XML document sequentially.

Before parsing, SAX parser does not load XML document into memory nor does it create any type of object from XML document. SAX parser has benefit over DOM-style parser.

f) Data Mining

Cluster	**Information**
1. Political	News related to politics.
2. Sports	News related to all types of sports.
3. Bollywood	News related to upcoming movies or about actors.
4. City	News from local area.

Table 1: *Example of Clustering.*

Association rule is the type of data mining which is related to relational database. This is useful for analyzing and predicting the customer behavior. There are two parts of association rules: (i) antecedent (if) and (ii) consequent (then).

Antecedent is the data which is found in extracted information. A consequent is the data related to antecedent information.

For example, if a customer buys a notebook and a pen then approximate 70% customer must buy a scale. By analyzing this

analysis decides which type of programming robot is to be created [4–6].

d) DOM Tree

DOM tree is language-independent and cross-platform and is used for representing and interacting object in HTML, XHTML or XML document. Every document contains the nodes that are organized in a tree structure called as document object in DOM tree. It is the way for accessing and manipulating an HTML

Data mining is the process of discovering actionable information from database. This paper contains mainly three ways of data mining technique:

1. Clustering
2. Association rule
3. Decision making

Clustering is the process to collect the similar kind of data and form a group of similar kind of data and form a group of similar information. It generates various clusters. Each cluster contains the similar kind of data but different from others. For example, when we collect the information regarding news, then information could be clustered like political news, sports, Bollywood, etc.

information we can predict the behavior of customers and take the decision to make discount on scales if customer buys a notebook and a pen [7].

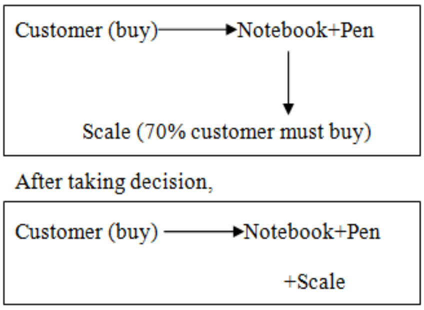


Fig. 1: *Example of Association Rule.*

Decision making is the concept, which uses the history or previous records present in database to take decisions. In this concept, the application analyzes the previous records and predicts the proper decision which is applied by user.

If a user wants to take the decision to improve his/her service then he/she fires the query related to that field. Application searches the history related to the query and after analysis of previous details it predicts the priority-based decisions. Those decisions are to be sent as output to the user.

For example, when we enter any key on the Google search engine, then it automatically displays links related to that key by analyzing the previous history on Google. This concept is called as search engine optimization.

Dissertation Work

The authors are implementing a JAVA application to search relevant information whatever user wants. The different steps of that application are:

1. To Design a Robot

When the user wants to browse information about a particular topic from web pages, it is a time-consuming process as well as requires more human effort. For that, we can implement a robot which is helpful for user to get the necessary data regularly. It could save time to search the data as well as less human effort to search.

a. Robomaker: Robomaker is the platform which provides special programming features to generate or design the robot. It provides some special features to generate the robot so that implementation of designing of robot becomes very easy.

b. Generate Robot: Whichever web page the user wants to visit again and again, firstly we have to get the URL of that web page and set the parameter of that web page where we want to search the required data. And download all the information available on that page and set it in the form of a tree. When user enters his/her query to application it gets the pattern of the node where that content is present. That pattern is saved as robot in the database [8].

c. Load Robot: As we can generate the robots as per the user requirement, so that robot can be loaded to extract information from the web page. When we need information about a particular topic, we can simply load that particular robot to extract the information; no need to visit the web page again and again.

d. Test Robot: Test robot is the function to test the functionality of the previous robot which is loaded. We can manually set the parameter available in URL to extract the data and then we can test that robot, whether it is extracted proper data or not.

e. Current Robot

This is the function that shows the extracted content as well as current path and the URL on which we are working.

2. Mashup

We can easily generate different mashup objects for different robots and then show the combined result of different robots. We can use the mashup of that robot and show the combined results of those robots.

3. Serialization

This is the method of JAVA to store data in the memory of the laptop/computer, wherever the system is running. There is no need to use the database separately. It reduces the cost of the system. It also reduces the complexity of the project as there is no need to write the connectivity code in program.

4. Excel Database

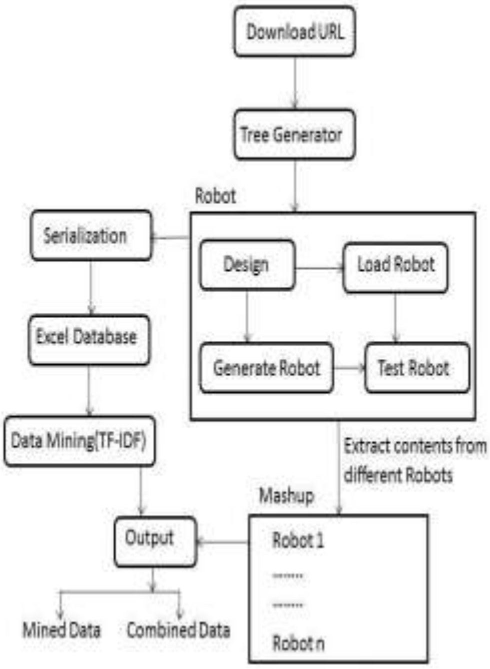
If we want to store the information which is extracted from the web, then we can store it using MS-Excel with the extension ".csv". ".csv" is used to save those files which contain different types of information in the same file. And that file is not stored into structured format.

5. Data Mining

After all the process, we can apply the data mining concept to search the data. For that, the authors are using TF-IDF algorithm.

a. TF-IDF: This is the algorithm which is used to apply text mining on the database. TF-IDF stands for term frequency and inverse document frequency. In this algorithm, when the user wants to search a query in the database, then the query is partitioned in the word and frequency is counted for each and every word in the database, then the common word is selected from the query and is ignored. After that, frequency of all words is added to get priority to the searched results. The results having maximum frequency get a higher priority than others. And the perilously searched content gets the highest priority. It is also used to show the results which get the highest priority.

Processing Block Diagram



RESULTS

Whenever the user wants to get data required by him/her, then the robot is created and that robot is executed to get the prior result.
Result of this system is data shown by the mashup system, and the data shown after firing the query.

It also stores all the data extracted from different robots in the Excel file. That file is the main database of the system [9].

FUTURE SCOPE

The future scope of this system is:

1. Extracting video and audio data from the web page.
2. Comparing data extracted from different robots.
3. Showing information into charts and tables.
4. Extract the charts and the tables available on the web page.

CONCLUSIONS

This paper proposes efficient techniques to handle data present on web pages. Web data extraction techniques and mashup those data make appropriate information available to the user. So that which project we are implementing that will be easy to understand to the user, less time consuming and very useful for the user. It provides limited content as per user requirement. Web data extraction is done on unstructured or semi-structured form of web data. The idea behind this survey is to classify existing techniques along with research areas, where web data extraction and mining on those data can provide efficient results.

REFERENCES

1. Knoblock CA, Lerman K, Minton S, et al. Accurately and reliably extracting data from the web: A machine learning approach. *Intelligent Exploration of the Web*. Springer-Verlag, Berkeley, CA; 2003.
2. Chamberlin D, et al. (Eds). XQuery: A query language for XML. http://www.w3.org, 2001.
3. Huynh D, Mazzocchi S, Karger D. Piggy bank: Experience the semantic web inside your web browser. In: *Proc. of ISWC*. 2005.
4. Google Map Facility, http://maps.google.c om, last accessed 12 October 2009.
5. Wong Jeffrey, Hong Jason I. *Making Mashups with Marmite: Towards End-User Programming for the Web*. Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, last downloaded 12 October 2009.
6. Kapow Technologies. *Kapow Mashup Server 6.3 Robomaker User Guide*. http://www.kapowtech.com, last accessed 12 October 2009.
7. Lerman K, Plangrasopchok A, Knoblock CA. Semantic labeling of online information sources. In: Pavel Shaiko (Ed.). *IJSWIS, Special Issue on Ontology Matching.* 2007.
8. Lee Y, Sayyadian M, Doan A, et al. eTuner: Tuning schema matching software using synthetic scenarios. *VLDB Journal*, Special Issue 2006.
9. Lixto Technologies. *Lixto Visual Developer*. http://www.lixto.com, last accessed 12 October 2009.