An Overview of Web Content Mining Tools

Dr Eldhose T John, Bibu Skaria and P.X. Shajan

Abstract---Web is one of the most widespread platforms for information exchange today, as it is easier to publish documents. As the number of users and providers increases, the number of documents grows, searching for information becomes a difficult and time-consuming process. Web mining uses various data mining techniques to discover useful knowledge from Web hyperlinks, page content and usage log file. The mining tools are used to scan the HTML documents, images, and text, the results is provided for the search engines. It can assist search engines in providing productive results of each search in order of their relevance. In this paper, we brief introduction to the concepts related to data mining, web mining and then an overview of different Web mining tools. We conclude by presenting a comparative table of these tools based on some pertinent criteria.

Keywords---OS, PI, U, PRD, PSWD and PSUWD

I. INTRODUCTION

DATA MINING or knowledge discovery (KDD) is the process of analyzing data from different perspectives and summarizing it into useful information. Data are raw facts that can be processed by a computer[14]. Organizations have vast and growing amount of data in different formats and different databases. Data mining can be applied in various domains such as statistics, Machine learning, pattern recognition, database systems, information retrieval, World Wide Web, visualization, and many application domains, has made great progress in the last few years[1],[2]. Web mining is the application of data mining techniques to extract knowledge from web data, i.e. web content, web structure, and web usage data[14]. In this paper we are reviewing the features of a few web mining tools available.

II. WEB MINING TAXONOMY

Web mining is broadly classified into three distinct categories, according to the kinds of data to be mined.

- Web Content Mining.
- Web Structure Mining
- Web Usage Mining

A. Web Content Mining

It is a process involving extracting the useful information from the contents in a web page. Content data in a web page is the collection of facts or information that a web page is designed to contain. Content data may contain text, images, audio, video, or structured records etc. Application of text mining to web content has been the most widely researched area in the field of data mining.

B. Web Structure Mining

The complete structure of a typical web graph consists of web pages which will function as nodes and the hyperlinks as edges connecting related pages. Web structure mining is a mining process to discover information regarding the structure from the web[3]. Web structure mining can be further subdivided into two based on the type of structured information used. They are hyperlinks and document structure.

C. Web Usage Mining

Web usage mining is the application of data mining techniques to discover interesting usage patterns from web usage data, in order to understand and better serve the needs of web-based applications.

Fig. 1: Web Mining Taxonomy

III. WEB CONTENT MINING TOOLS

The content mining tools helps to download the essential information that one would require. They collect appropriate and perfectly fitting information [5], [6], [7]. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, Web Content Extractor, RapidMiner and Weka.

A. Screen-scraper[8]

Screen-scraping is a tool for extracting/mining information from web sites.

B. Automation Anywhere 6.1 (AA) [9]

AA is a Web data extraction tool used for retrieving web data[9].

Features:

- It is a (USAT) Unique SMART Automation Technology for fast automation of complex tasks.
- Record keyboard and mouse or use point and click wizards to create automated tasks quickly.
- Web record and Web data extraction.
C. Web Info Extractor (WIE)[10]

This is a tool for data mining, extracting Web content, and Web content analysis. Using this tool we can extract structured or unstructured data from Web pages, reform the results into local file or save to database, place into Web server.

Features:
- No need to learn boring and complex template rules, and it is easy to define extract tool.
- Extract tabular as well as unstructured data to file or database.
- Monitor Web pages and extract new content when update.
- Can deal with text, image and other link file.
- Can deal with Web page in all language.
- Running multi-task at the same time.
- Support recursive task definition.


This tool enables users to extract and manage Web data. Users can setup agents that routinely extract, store, and publish data to multiple destinations. The data on reaching Mozenda, users of this tool can format, recreate, and mashup with the data as per requirement so it can be used in other applications or as needed.

Features:
- Easy to use.
- Platform independency. However, Mozenda Agent Builder only runs on Windows.
- Working place independence: Tune the scraper, manage the scraping process and get scraped data from any computer connected to the Web.

E. Web Content Extractor (WCE)

Web Content Extractor[12] is a web data extraction software it is also used to increase productivity and effectiveness of the web data scraping process. Web Content Extractor is highly accurate and efficient for extracting data from websites. It offers you a user friendly, wizardry driven interface. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access etc to any ODBC data source.

Features:
- It assist in extracting the market figures, product pricing data, or real estate data, extracting the information about books, including their titles, authors, descriptions.
- ISBNs, images, and prices, from online book sellers.
- Assists users in automate extraction of auction information from auction sites.
- It supports Journalists to extract news and articles from web sites.
- Assist in seeking a job from online job websites.
- Extract the online information about vacation and holiday places, including their names, addresses, descriptions, images, and prices, from web sites.

F. RapidMiner (YALE)[13]

This RM tool developed in Java, it is advanced analytical tool through its template-based frameworks provided. An advantage of this tool is that the users hardly need to write any code. RM also has the capability to provide data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment.

Features:
- Easy data management of user data process.
- Repository structure the user data projects and serve as the major source of data to be analyzed.
- Provides learning schemes.
- Models.
- Algorithms.

G. Weka

WEKA mainly was developed for analyzing data from the agricultural domain it was a non-java platform. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. Its free under the GNU GPL. [http://www.cs.ccsu.edu/~markov/]. WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection.

Features:
- It provides many different algorithms for data mining and machine learning.
- Is open source and freely available.
- It is platform-independent.
- It is easily useable by people who are not data mining specialists.
- It is flexible for scripting experiments. And also kept up-to-date, with new algorithms being added as they appear where and there.

IV. COMPARATIVE STUDY

The diversely available Web Content Mining tools are difficult to compare, because of their variety of goals and contexts they have. We compare the existing tools based on the following six points:
- Open source- OS.
- Platform independent- PI.
- Usability (User friendly)- U.
- Possibility to Record the Data- PRD.
- Perform on Structured web Data- PSWD.
- Perform on Unstructured web Data- PSUWD.
Table 1: Comparative Study

<table>
<thead>
<tr>
<th>Tools</th>
<th>OS</th>
<th>PI</th>
<th>U</th>
<th>PRD</th>
<th>PSWD</th>
<th>PSUWD</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>AA 6.1</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>WIE</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Mozenda</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>WCE</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>RapidMiner</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Weka</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

V. CONCLUSION

The Web Data Mining tools are used to scan HTML documents, images, and text contained in Web pages. The results of these are provided to the search engines, so that it can assist search engines in providing productive results of each search in order of their relevance. In this paper, we present a list of the available Web Content Mining Tools. Through this study, we established some objective criteria for comparison (OS, PI, U, PRD, PSWD, PSUWD). Based on these criteria, we gave a comparative table of these different tools are provided. From our study, Weka is found to be one that is supporting OS, PI, U, PRD, PSWD and PSUWD. We believe that this paper in Web mining will help to understand the web content mining tools.

REFERENCE