International Journal of Information Technology and Knowledge Management
January-June 2011, Volume 4, No. 1, pp. 211-215

WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY

V. Bharanipriya[1] & V. Kamakshi Prasad[2]

Web mining adopts data mining techniques to automatically discover and retrieve information from web documents and services. In this Paper we have discussed the concepts of Web mining. We have mainly focused on one of the categories of Web mining namely Web Content Mining and its various tasks. We propose a six step Web content mining process in our work. Various tools for Web content mining are also discussed and their relative merits and demerits are presented.
Keywords: Web Mining, Web Content Mining, Web, Tools, Moulding

1. Introduction

Analysis and discovery of useful information from World Wide Web poses a phenomenal challenge to the researchers in this area. Such a phenomena of retrieving valuable information by adopting data mining techniques is called Web mining. Web mining is classified into following five sub tasks: 1) Resource finding, 2) Information selection and pre-processing, 3) Generalization, 4) Analysis and 5) Visualization [1]. Web mining is divided into three categories: Web content mining (WCM), web usage mining (WUM) and web structural mining (WSM).

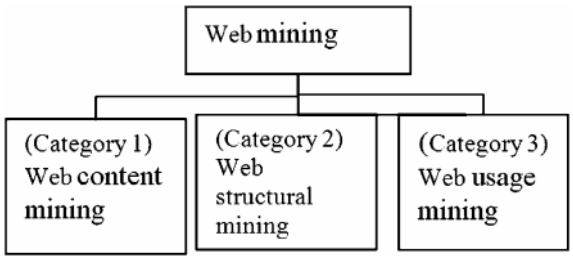

Fig. 1: Web Mining Categories

Web content mining is the process of identifying user specific data from text, image, audio or video data already available on the web. This process is alternatively called as web text mining, since text content is the most widely researched subject on the World Wide Web. The technologies that are generally used in web content mining are Information retrieval and Natural language processing.

Web structure mining is another process of using graph theory to analyze the node and connection structure of a web site. Depending upon the type of web structural data, web structure mining has been divided into two folds. The first one is extracting patterns from hyperlinks in the web. The other one is mining the document structure. This involves using the tree-like structure to analyze and describe the HTML or XML tags within the web page.

Web usage mining is to identify user access patterns from Web usage logs.

Here we discuss more on Web content mining.

[1]Dept of M.C.A, NMREC, Ghatkesar, India.
[2]School of IT, JNTUH, Hyderabad
Email: [1]g_bharanipriya@yahoo.co.in

2. Web Content Mining (WCM)

Web content mining identifies the useful information from the Web Contents/data/documents. However, such a data in its broader form has to be further narrowed down to useful information. In this section we begin with two main approaches of Web Content mining and define how it differs from Data Mining.

The web content data consist of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents. The two main approaches in WCM are (1) Unstructured text mining approach and (2) Semi-Structured and Structured mining approach.

2.1. Unstructured Text Data Mining (Text Mining)

Web content data is much of unstructured text data. The research around applying data mining techniques to unstructured text is termed knowledge discovery in texts (KDT), or text data mining, or text mining. Hence one could consider text mining as an instance of Web content mining.

To provide effectively exploitable results, pre-processing steps for any structured data is done by means of information extraction, text categorization, or applying NLP techniques.

2.2. Semi-Structured and Structured Data Mining

Structured data on the Web are often very important as they represent their host pages, due to this reason it is important and popular. Structured data is also easier to extract

compared to unstructured texts. Semi-structured data is a point of convergence for the Web and database communities: the former deals with documents, the latter with data. The form of that data is evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real-world objects like books, papers, movies, etc., without sending the application writer into contortions. Emergent representations for semi-structured data (such as XML) are variations on the Object Exchange Model (OEM). In OEM, data is in the form of atomic or compound objects: atomic objects may be integers or strings; compound objects refer to other objects through labeled edges. HTML is a special case of such 'intra-document' structure.

2.3. WCM Tasks

Apart from the five tasks enumerated under Web mining, another task viz. 'cleaning' be applied in web content mining with the objective of removing redundancy. The detailed explanation about other five tasks is given below.

i) Resource Finding: Resource finding means the process of retrieving the required data from on-line or offline. We apply data mining techniques, classification, clustering etc., to extract information.

ii) Information Selection and Pre-processing: In the pre-processing phase we deal with web pages representation. The brief explanation of different representations is given below[2].

- Binary: This is called as "Set of words". The relevance or weight of feature is a binary value {0, 1} depending on whether the feature appears in the document or not.
- Term Frequency (TF): Each term is assumed to have an importance proportional to the number of times it occurs in the text. The weight of a term t in a document d is given by: W (d; t) = TF (d; t) is the term frequency of the term t in d.
- Inverse Document Frequency (IDF): The importance of each term is assumed to be inversely proportional to the number of documents that contain the term. The IDF factor of a term t is given by: IDF(t) = logNxdf(t); where N is the number of documents in the collection and df(t) is the number of documents that contain the term t.
- TF-IDF: Salton (1998) proposed to combine TF and IDF to weight terms. Then, the TF-IDF weight of the term t in a document d is given by: W(d; t) = TF(d; t) × IDF(t).
- WIDF: It is an extension of IDF to incorporate the term frequency over the collection of documents. The WIDF weight is given by: W(d, t) = TF(d, t)/ΣdiTF(i,t).
- Other pre-processing includes Latent Semantic Indexing (LSI) which determines clusters of co-occurring keywords so that the query which uses a particular keyword can then retrieve documents perhaps not containing this keyword, but containing other keywords from the same cluster.

iii) Generalization: Pattern evaluation phase is also called as Generalization. Here we use machine learning or data mining processes to identify general patterns in individual web pages or across multiple sites.

iv) Analysis: In the analysis phase accuracy of the retrieved pattern is evaluated using accuracy measures.

v) Presentation or Visualization: To decide in which order the discovered knowledge (web pages) has to be presented.

2.4. Distinction between Web Content Mining, Data Mining and Text Mining

Web content mining though uses data mining techniques; it differs from data mining because Web data are mostly unstructured and/or semi-structured, while data mining deals mainly with structured data. It is associated to text mining because much of the Web contents are texts. Web content mining differs from text mining because of the semi-structure quality of the Web, while text mining deals with unstructured texts. Web content mining thus requires inventive applications of text mining and/or data mining techniques and also its own distinct approaches.

3. Moulding the Web Information

The information retrieved from online databases, to achieve the required knowledge the matching concept of attributes has to be achieved. We reach to the desired knowledge by comparing match attributes; and user query can be responded with the desired knowledge.

Suppose that a user is searching for books with "Web Content Mining" name and she is also looking for stores that can sell this book with the lowest price. This system, after extracting information from some query interfaces in book domain, finds matching information between them and after comparing them sends the best answer to the user query.

To attain such knowledge we have to mould the extracted information of a Web into structural form which

is in the unstructured textual format for classification.

We represent the Web data in the binary format where all of the attributes are posited in the columns and some frequent schemas posited in the rows. If an attribute is in a frequent schema, a 1 is stored in related cell and otherwise a 0 is stored in it. In the following you can see an example of this form. The attributes represented below are derived from frequent schemas. Attributes are derived in such a way that they form the keywords of the frequent schema. The attributes of frequent schemas are stated below.

QI1: Deep Web Content mining = {Deep, Web, content, mining}.

QI2: Extracting structured data from Web pages = {Extracting, structured, data, Web, pages}.

QI3: Page content Rank: An approach to the web content mining = {Page, content, rank, web, content, mining}.

QI4: Web mining: Information and pattern discovery on the World Wide Web = {Mining, information, pattern, discovery, World Wide Web}.

Table 1
A Sample Array with Input Information

	Deep	Web	Content	Mining	Extracting	Other Keywords	Category
QI1	1	1	1	1	0	0	1
QI2	0	1	0	0	1	1	2
QI3	0	1	1	1	0	1	3
QI4	0	0	0	1	0	1	1

Different types of categories listed in last column of Table 1 are represented with different unique numbers. Same domain value indicates same type of category.

4. TOOLS AVAILABLE FOR WCM

Web content mining in normal parlance is to download information available on the websites. Such a process involves tremendous stress and time-taking. To augment such a process the software related to web content mining can be used so that a computer can use this software or tools to download the essential information that one would require. It collects the appropriate and perfectly fitting information from websites that one requires. Different types of Web content mining tools are listed below:

Web Info Extractor[3]: This tool is helpful in mining web data, extracting web content, and monitoring content update. Thorny template rules are not required to be defined.

Mozenda[4]: To extract web data easily and to manage it affordably Mozenda is useful. With Mozenda, users can set up agents that regularly extract, store and circulate data to several destinations. Once information is in the Mozenda systems users can repurpose, format, and mash up the data to be used in other online/offline applications or as intelligence.

Screen-Scraper[5]: Screen-scraper allows mining the content from the web, like searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. The programming languages like Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can also be used to access screen scraper.

Web Content Extractor[6]: Most powerful and easy to use data extraction tool for web scraping, data mining or data retrieval from the internet is Web Content Extractor.

Automation Anywhere 5.5[7]: Automation Anywhere is a web data extraction tool used for retrieving web data effortlessly, screen scrape from web pages or use it for web mining.

5. DEMONSTRATING WCM TOOLS

5.1. Learning about Different Types of WCM Tools

Automation Anywhere 5.5: The Intelligent Automation Software, used for automating and scheduling business process and IT tasks in easier way.

Features:

- Intelligent automation is used for business and IT tasks.
- Unique SMART Automation Technology automates complex tasks fast! (No programming required.)
- Creating automation tasks takes few minutes-record keyboard and mouse strokes, or use easy point-and-click wizards.
- Distributes tasks to multiple computers easily, using Task to SMART Exe capability (Premier and Enterprise editions only).
- Web recorder: (Used for extracting multiple Data and to extract Table.).
- Use Automation Anywhere to automate scripts in disparate formats.
- Powerful task scheduling and auto-login - run scheduled tasks anytime, even when computer is locked.
- 385 plus actions are included: conditional, loop, prompt, file management, database, system, Internet. More great features: fast speeds, automatic email notification, task chaining, hotkey, variables, logging, etc.

Recording and running simple tasks takes just three steps:

Record → Save → Run

Screen-scraper: Screen-scraping is a tool for extracting information from web sites which can be used in other contexts. Like a database, it allows to mine the data of the World Wide Web.

Features:

Graphical interface is provided by the Screen-scraper allowing you to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data. Once these items have been created, from external languages such as .NET, Java, PHP, and Active Server Pages, Screen-scraper can be invoked. This also facilitates scraping of information at periodic intervals.

One of the most regular usages of this software and services is to mine data on products and download them to a spreadsheet. A classier example would be a meta-search engine where in a search query entered by a user is concurrently run on multiple web sites in real-time, after which the results are displayed in a single interface.

Main window Tools are as follows:

Add new proxy session, Add a new scraping session, Add new script, Save, Copy, Cut and Paste.

Web Information Extractor: For mining web data and for content retrieval it is a very powerful tool. It can retrieve unstructured or structured data from web page, reorganize into local file or save to database, place into web server. Difficult template rules are not require to be defined, browse to the web page and click what you desire to describe the retrieval task, and run it as you want, or let it run automatically.

Features:

- It is Easy to define extraction task and no need to learn boring and dense template rules.
- Retrieve unstructured data as well as tabular data to file, database.
- Monitor web pages and retrieve new content when update.
- Can deal with any kind of files like, image, text and other link file.
- Unicode support and can process web page in all languages.
- Support recursive task (child task) definition.
- Can run multi-task at the same time.

Exploring Main window Tools:

Main window Tools are as follows…..

System menu: The components of System menu are Config, Login.

Task menu: The components of Task menu are Create, Modify, Delete, Run, Pause, Stop, Import, Export, Save to File, Manual Export, Auto Save to CSV File, Auto Save To Text File, Auto Save To Website, Auto Save to Database, Clear Result.

View Menu: The components of View menu are Toolbar, Statusbar and Tasklist.

Help Menu: Components of Help menu are Online help, Website, Purchase, Activate and About.

Table2
Summary of Tools and their Respective Tasks

Name of a tool	Tasks			
	Records the data	Extract structured data	Extracts unstructured data	User friendly
Automation Anywhere	✓	✓	✓	✓
Web Info Extractor	X	✓	✓	✓
Web Content Extractor	X	✓	✓	Not for unstructured
Screen-Scraper	X	✓	✓	X
Mozenda	X	✓	✓	✓

5.2. Commonalities and Differences between the above Tools

- Commonalities:
  - All the tools automate the business task and retrieve the web data in an efficient way.
- Differences:
  - Screen-scrapper needs prior knowledge of proxy server and some knowledge of HTML and HTTP where as other tools do not require any such knowledge and it need Internet connection to run.
  - Automation-AnyWhere5.5 allows recording of actions; this facility is not provided in the other tools.

– Though we have setup file, Mozenda will not allow us to install with out Internet connection, this is not the case with other tools.

We summarize the tools and their respective tasks in Table 2.

6. CONCLUSIONS

There are many concepts regarding World Wide Web. We tried to expose Web content mining, one of the categories of Web mining. The term Web Content mining refers to a technique that encompasses broad range of issues. We provided different views to understand Web content mining, and given six tasks of WCM. We explored more on some of the tools of Web Content Mining and provided their comparisons and differences. We observed that Screen-scraper is not user friendly and some of these tools seem to be applicable for email data mining.

REFERENCES

[1] Qingyu Zhang & Richard S. Segall, "Web Mining: A Survey of Current Research", Information Technology and Decision Making, 7(4), 683-720, 2008.

[2] R Kosala, H Blockeel-ACM SIGKDD Explorations Newsletter, 2000.

[3] Web Info Extractor Manual.

[4] www.Mozenda.com

[5] www.screen-scraper.com

[6] Web Content Extractor help.

[7] Automation Anywhere 5.5 help