Journal of Theoretical and Applied Information Technology
31st December 2013. Vol. 58 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN IMPLEMENTATION OF WEB CONTENT EXTRACTION USING MINING TECHNIQUES

[1]BADR HSSINA, [2]ABDELKARIM MERBOUHA,[3]HANANE EZZIKOURI, [4]MOHAMMED ERRITALI, , [5]BELAID BOUIKHALENE

[1,2,3,4] Computer Sciences Department, Faculty of Sciences and Technologies
Sultan Moulay Slimane University. Beni-Mellal, BP: 523, Morocco
[5]Poly disciplinary faculty, University Sultan Moulay Slimane, Morocco
E-mail: [1]hssina.badr@hotmail.fr, [2]merbouhak@yahoo.fr , [3]ezzikourihanane@gmail.com,
[4]mederritali@yahoo.fr,[5]bbouikhalene@yahoo.fr

ABSTRACT

The Web has continued to grow up since its inception in volume of information, in the complexity of its topology, as well as in its diversity of content and services. This phenomenon was transformed the web in spite of his young age to an obscure media to take useful information. Today, they are billions of HTML documents, images and other media files on the Internet.

Taking into account the wide variety of the web, the extraction of interesting content has become a necessity. Web mining came as a rescue for the above problem. Web content mining is a subdivision under web mining, which is defined as "the process of extracting useful information from the text, images and other forms of content that make up the pages" by eliminating noisy information .This extraction process can employ automatic techniques and hand-crafted rules. In this paper, we propose a method for web data extraction that uses hand-crafted rules developed in Java.

Keywords: *Web Mining, Content Extraction, Web Cleaning*

1. INTRODUCTION

Since its appearance, the Web has grown up in volume of information, in the complexity of its topology as well as in the diversity of content and services. This phenomenon transformed it, in spite of his young age in an obscure media to taking useful information. Today, there are billions of HTML documents, images and other media files on the Internet.

Taking into consideration the wide variety of Web, the extraction of interesting content has become a necessity. A solution to this problem is to use data mining techniques.

The term Data Mining literally means, "drilling data ". As with any drilling, its purpose is to extract an element of knowledge.

Data Mining is the process of exploration and analysis, by automatic or semi-automatic, of large amounts of data to discover meaningful patterns and rules. Contrary to popular belief, Data Mining is not miracle cure that can solve all the problems or needs of the business [11]. However, a multitude of economic and commercial issues of intellectual order can be grouped in their formalization in one of the following tasks:

- ✓ Classification: The classification consists to examine the characteristics of a newly introduced element in order to assign it to a class of a predetermined set.
- ✓ Estimation: the result of estimation provides a continuous variable. The result of estimation allows classifications with a scale.
- ✓ Prediction: The prediction is similar to the classification and estimation but in a different time scale.
- ✓ Grouping by similarities: is to group items which form naturally sets. The most appropriate technique to group similarities is the analysis of consumer basket
- ✓ Segmentation (or clustering) consists in segmenting a heterogeneous population into homogeneous populations.

✓ Optimization: To solve many problems, it is common for each potential solution to involve an evaluation function. The goal of optimization is to maximize or minimize this function. [11]

Web mining is the use of data mining techniques for automatic discovery and knowledge extraction from documents and Web services. This new area of research was defined as an interdisciplinary field (or multidisciplinary) that uses techniques borrowed from: data mining, text mining, databases, statistics and machine learning…. [12]

Web Mining has two main objectives:

1. The improvement and development of websites: The analysis and understanding of user behavior on websites can enhance the content of sites by improving the organization and performance of sites.

2. Customization: The Data Mining techniques applied to data collected on the Web used to extract valuable information on the use of the site by users. The analyses of this information to personalize the service offered to users by taking into account their preferences and profile content.[8]

The three axes of development of Web Mining are Web Content Mining, Web Structure Mining and Web Usage Mining.

⊕ Web Content Mining: extraction of predictive models and knowledge of the contents of Web pages.

⊕ Web Structure Mining: the discovery of useful knowledge from the structure of links between web pages.

Web usage Mining: analyze the use of Web resources using the Logs files.

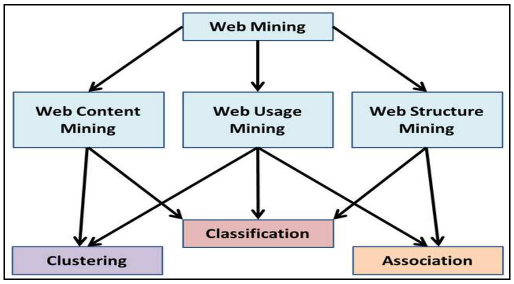

Figure 1: Axes of development of Web Mining

We focus on the extraction of Web content, which Linoff and Berry (2001) defined as "the process of extracting useful information from the text, images and other content that make up the pages."

Web Content Mining (WCM) consists into an advanced textual analysis integrating the study of hyperlinks and the semantic structure of web pages. Thus techniques of classification and analysis of strings of text mining is very useful in treating the text of the pages. The WCM is also interested in images. It allows, for example, quantifying images and text boxes for each page. In addition, through joint analysis of attendance pages, it is possible to determine whether the pages with images that are most visited pages with text.

2. RELATED WORK

Current Internet includes billions of pages consist of drowned data information layout. Whether to convert existing sites or sites semantics for intelligent use embedded data, the definition of data mining techniques is of great interest. For that reason, the extraction of data from the Internet has been and continues to be the subject of much research.

Related works can be grouped into two categories. The automatic extraction and rules handcrafted techniques. The main focus of automatic extraction techniques is inference through features extracted from HTML .Hand-crafted rules is mostly used to extract information from HTML through string manipulation functions [2].

Godoy, Schiaffino, and Amandi [13] demonstrated that the use of Web Mining can be used to extract knowledge from observed actions.

Crescenzi and al. [14], Baumgartner and al. [15], and Liu and al. [16] are based on the HTML markup generated automatically or semi-Automatically extracting useful data modules. Each extraction module is used for extracting data of pages whose information content and structure are homogeneous.

Adelberg [17] draw on the definition of a target structure for the data to be extracted. This structure is created by analyzing a sample document. According to this structure, an algorithm defines extraction rules based on delimiters (constant punctuation, text), and browsing other documents of the same type in order to extract the data in a format conforming to the target structure.

Embley and al. [18] rely on the data themselves. Prior to extraction, they define the domain of the ontology. Thus they describe the data worthy of interest and their relationships, and provide a set of constants and keywords for each element of the domain. It is on this basis they sweep documents to spot the data to be extracted. This method can deal with heterogeneous material in terms of the structure.

Chung and al. [19] Propose a mixed method (HTML markup and ontologies) to integrate homogeneous HTML documents on the informational level but heterogeneous in terms of structure and presentation. Rules to restructure documents based on structural and visual information of HTML markup are used to transform the source XML documents. To give names to represent XML elements, the user defines a first set of concepts of application domain, and examples of instances (keyword) or models of instances for these concepts. These models and keywords are compared to textual information met during the restructuring. From XML documents, a DTD file describing common structures is derived.

JIANG Chang-Bin Chen and Li [21] provide a log file preprocessing algorithm of Web data based on collaborative filtering. It can identify the user session fast and flexibly, even if the statistics are not sufficient and the historical records of visits of the user is absent.

3. PROBLEM DESCRIPTION

If the internet is the most important source of information that exists, it is not easy to use it content effectively and intelligently. In fact, HTML pages have two major drawbacks: first, they consist of a mixture of data and instructions and presentation, on the other hand, are virtually devoid of information about the structure and the meaning of the data.

In order to improve the possibilities of exploitation, it is interesting to develop methods to separate the data from their presentation, extract, interpret and find their hidden semantic structure in HTML pages.

The data structures extracted become a basis for as diverse as reengineering site applications, integration of multiple sites, migrating data to a database, statistical studies, etc.

4. PROPOSED SYSTEM

In web, most of the data are noisy and dirty in nature. The main idea of the proposed system is to extract patterns based on user interest.

Architecture of proposed system is shown in figure 2.

- ⊕ Connecting to any website and getting the data from that website.
- ⊕ Remove comments, delete Meta tags (Meta tags provide descriptions of Web documents, the user interesting patterns are not in the description) Remove Scripts (client side or server side scripting languages are used to present the document), Remove ads (pop-up ads deviate user sites spoofing the Web, delete the external or internal links.
- ⊕ Extract the content based on the user's interests (list of links, images, media ...).
- ⊕ Display content to the user.

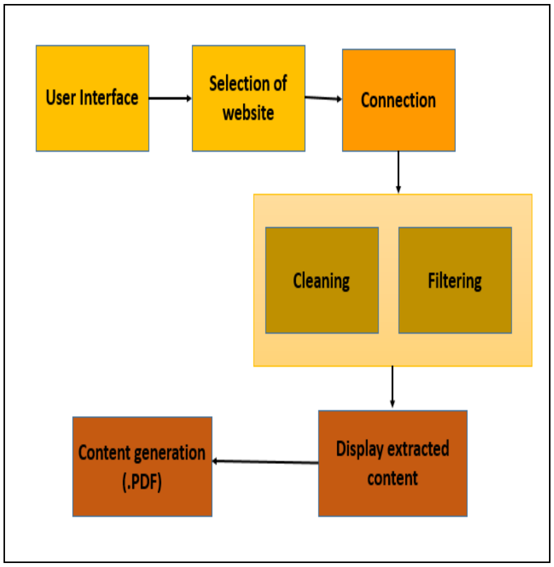

Figure 2: Architecture of proposed system

5. EXPERIMENTAL RESULTS

The experimental part of the proposed system is shown in Figure 3. Before extraction of any content, the first web document is selected. Then, application of pre-processing techniques to the data.

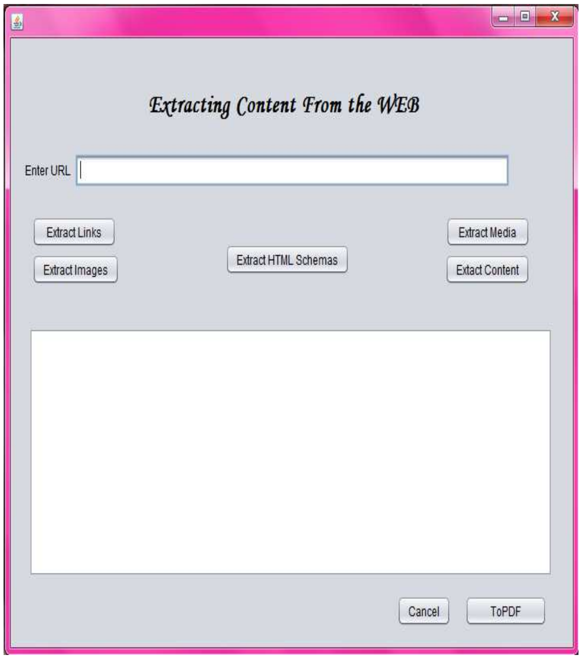

Figure 3: Experimental setup

The use of this application is very simple you just enter the URL of the website in question and choose the action (or actions) by the user.

- **Extract links**: display all links that are found throughout the web page.
- **Extract images:** display URL images on the website.
- **Extract media**: displays the list of media that is in the whole page.
- **Extract Content**: Extraction of useful content in the web page.
- **Extract Html Schemas :** provide the subset of the shaft HTML encompassing the entire contents.
- **Cancel**: clear the results above for a new use of the application.
- **ToPDF:** Generates a document included your extracted content in a PDF document.

6. CONCLUSION AND FUTURE WORK

Web page content extraction is extremely useful in search engines, web page classification and clustering process, it is the basis of many other technologies about data mining, which aims to extract the worthiest information from data-intensive web pages full of noise.

In the proposed method we extract required patterns by removing noise that is present in the web document using hand-crafted rules developed in Java. In future we plan to extend our work to the Web usage Mining.

In the introduction, we marked the considerable character figures for the use of the Internet and the number of pages available. It can be considered in parallel need, what the website owners to understand their users. The existences of these factors has increased strongly the emergence of Web Usage Mining by applying knowledge extraction algorithms on large volumes of data on one side and use the results of an other side. However, the data contained in log files results in a lack of reflection on how to proceed. The step data mining itself deserves further work to be adapted to the needs of the analysis of log files.

REFRENCES:

[1] S.Mahesha, Dr. M S Shashidhara, and Dr. M. Giri, “An Implementation of Web Content Extraction Using Mining

Techniques” IFRSA International Journal of Data Warehousing & Mining |Vol 2|issue4|November 2012.

[2] Erdinç Uzun, Hayri Volkan AgunTarık Yerlikaya,” A hybrid approach for extracting informative content from web pages”. E. Uzun and al. / Information Processing and Management 49 (2013) 928–944

[3] Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona, “Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it” O. Arbelaitz and al. / Expert Systems with Applications 40 (2013) 7478–7491

[4] ZHANG Bin, WANG Xiao-fei, “Content extraction from Chinese web page based on title and content dependency tree” The Journal of China Universities of Posts and Telecommunications. October 2012, 19(Suppl. 2): 147–151, www.sciencedirect.com/science/journal/10058885.

[5] Markus Schedl, Gerhard Widmer, Peter Knees, Tim Pohle, “A music information system automatically generated via Web content mining techniques,” M. Schedl and al. / Information Processing and Management 47 (2011) 426–439.

[6] Lihui Chen, Wai Lian Chue, “Using Web structure and summarisation techniques for Web content mining,” L. Chen, W.L. Chue /

Information Processing and Management 41 (2005) 1225–1242.

[7] Jean-Roch Meurisse," Extraction de données de sites web :Méthodologie, outils et étude de cas" Mémoire présenté en vue de l'obtention du grade de licencié en informatique, Facultés Universitaires Notre-Dame de la Paix, Namur,Institut d'Informatique. Année académique 2003-2004.

[8] Malika CHARRAD, " Techniques d'extraction de connaissances appliquées aux données du Web", National School of Computer Science, University of Manouba, Tunisia - Master in Computer Science, Option: Geniuses Documentiel and Software 2005.

[9] [Embley and al.] Embley D, Campbell D., Jiang Y., Liddle S., Kai Ng Y.-K., Quass D. and Smith R.«Conceptual-model-based data extraction from multiple-record web pages», Journal of Data and Knowledge Engineering, volume 31(3), 1999, 227-251.

[10] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", IEEE 2nd International Workshop On Education Technology and Computer Science, pp.118-121, 2010.

[11] Georges El Helou and Charbel Abou khalil," Data Mining Techniques d'extraction des connaissances", Management and NTIC project 16 February 2004, PANTHEON-ASSAS PARIS II University.

[12] CLAUDIA ELENA DINUCĂ, DUMITRU CIOBANU, "WEB CONTENT MINING", Annals of the University of Petroşani, Economics, 12(1), 2012, 85-92.

[13] Godoy, D., Schiaffino, S., & Amandi, A. (2004). "Interface agents personalizing webbased tasks".Cognitive Systems Research Journal, 207–222.

[14] Crescenzi and al., Mecca G. and Merialdo P. "Road Runner : Towards automatic data extraction from large Web sites", in Proceedings of 27th International Conference on Very Large Data Bases, Rome, 2001, pages 109-118.

[15] Baumgartner and al. ,Flesca S. and Gottlob G. "Visual Web Information Extraction with Lixto", in Proceedings of 27th International Conference on Very Large Data Bases, Rome, 2001, pages 119-128.

[16] Liu and al., Pu C. and Han W. "XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources", in Proceedings of 16th International Conference on Data Engineering, San Diego, 2000, pages 611-621.

[17] [Adelberg B. «NoDoSE—a tool for semi-automatically extracting structured and semistructured data from text documents», in Proceedings of the 1998 ACM SIGMOD international conference on Management of data, Seattle, 1998, pages 283-294.

[18] Embley and al., Campbell D., Jiang Y., Liddle S., Kai Ng Y.-K., Quass D. and Smith R.«Conceptual-model-based data extraction from multiple-record web pages», Journalof Data and Knowledge Engineering, volume 31(3), 1999, 227-251.

[19] Chung and al., Gertz M. and Sundaresan N. «Reverse Engineering for Web Data: From Visual to Semantic Structures», in Proceedings of 18th International Conference on Data Engineering, San Jose, 2002, pages 53-63.

[20] O. Zamir and O. Etzioni; "Web document clustering: a feasibility demonstration"; In Proceedings of SIGIR; 1998.

[21] JING Chang-bin and Chen Li, " Web Log Data Preprocessing Based On Collaborative Filtering ", IEEE 2nd International Workshop On Education Technology and Computer Science, pp.118-121, 2010.