International Association of Scientific Innovation and Research (IASIR)
(An Association Unifying the Sciences, Engineering, and Applied

ISSN (Print): 2279-0063
ISSN (Online): 2279-0071

International Journal of Software and Web Sciences (IJSWS)
www.iasir.net

A SURVEY ON DATA EXTRACTION IN WEB BASED ENVIRONMENT

Neeraj Raheja[1], Dr. V.K.Katiyar[2]
Associate Professor[1], Professor[2]
[1,2] Department of Computer Engineering, M.M.E.C, M.M.University, Mullana, Haryana, India

Abstract: *Web is a great source of information today. A lot of information is available over the internet and a lot of information is added and updated to it everyday hence web data extraction systems are necessary to use. These systems are used to find useful, hidden or related information according to user need hence also improves the efficiency of search engines. Web data extraction mainly deals with unstructured or semi-structured form of data which is main feature of web pages. In this paper we will discuss various techniques, application areas, current research areas in web data extraction.*

Keywords: *Web Data Extraction, Web mining, web Searching.*

I. INTRODUCTION

As the information is growing and changing over the internet at a very fast rate hence search engines are used to search the required information but still there are some limitations likewise to receive fast, accurate and hidden information. Hence web data extraction systems are used which provide the information according to user need and in a fast way which improves the efficiency of search engines. Web Data Extraction systems are a broad class of software applications or methods targeting at extracting information from web sources like web pages, scripts or websites in an efficient manner[1][9]. Web data extraction system works by interacting with a web page and extracts data stored in it according to user need. Suppose the source is a HTML Web page or a web script, the extracted information may consist of elements like tags of the webpage, related images or full content of the webpage. After the information is extracted it is processed and converted to most convenient structured format so that it can be easily used and stored [10] [11].Hence web data extraction system is a software for extracting the data or information automatically and repeatedly from web pages (dynamic or static) with changing contents, and then delivers extracted data to a database, email or some other application where it is to be stored.

A. Working of web data extraction:

Web Data Extraction works on the basis of three aspects:

a) Automation and Scheduling: Automating means to access the web pages or scripts as well as their localization i.e. locally storage of their elements or content [12]. It requires creating macros or programs which execute multiple instances of the same task, like filling forms, selecting menus and buttons, automatically updating of webpage etc. [13]. Scheduling means if a user wants to extract data from a web site (e.g. share market, news) which updates very frequently, hence scheduling tools are used to setup a scheduler, which launch macros or programs and execute scripts automatically and periodically.

b) Data transformation: In this step information extracted from various sources are transformed in a structured manner so that it can be processed or stored for further use easily. During this phase, processes like data cleaning [14] and resolution [15] are also performed by the users.

c) Use of extracted data: After the extraction task is complete and required data is transformed into the required format, the information is ready to be used by the user. The last step is to deliver or stored the structured form of data to a computerized system like a database, file or warehouse etc. so that this data can be used by analytical or statistical purposes for further processing and get more results [17].

B. *Challenges in Web Data Extraction*

i) Automation: It requires providing a high degree of human expertise for the process of automation. It requires high level accuracy, fast retrieval of information by web data extraction system. So the main challenge is to create an automated system for extracting data from the web sources with high performance.

ii) Processing Capability: Web Data Extraction techniques or systems should be able to process large volumes of data in very short time. Mainly it is necessary in the field of business (like share market) and artificial Intelligence because a firm needs to perform timely or fast processing according to market conditions.

iii) Privacy: Applications where privacy or security is required like in the field of social web, banking or applications dealing with human related or project related data. Therefore, potential (even if unintentional)

attempts to violate or break user privacy should be timely and adequately identified and advertised. To provide this feature in web data extraction is a great challenge.
iv) Training: Approaches relying on Machine Learning like in semantic web often require a significantly large training set of manually labeled web pages. In general, the task of labeling pages is time-expensive and error-prone and, therefore, in many cases we can not assume the existence of labeled pages.
v) Unpredictability: Web sources are continuously updating and hence structural changes happen are unpredictable. Hence in real-world situations where the need of maintaining these systems which might stop working correctly if lacking of flexibility or easiness and face structural modifications of related web sources. Hence to handle unpredictability is a great challenge.
vi) Noise: By noise we mean content other than the required content available on a webpage or script. Removing of the noise provides efficient web data extraction. It is a research challenge in web content mining and web structure mining.
vii) Semantic Web: To extract data from semantic web which is in machine understandable form and in form of ontologies is also a research challenge.
viii) Integration of web content and usage mining: By integrating both types of web mining techniques we can get useful and more efficient results.
ix) Web Page ranking: It deals with web structure mining and to remove limitations of ranking algorithms like page rank and HITS algorithms to get efficient results.

C. Techniques used for Web Data Extraction

Web data extraction techniques uses different approaches like markov chains, graph theory, neural network approaches, statistical techniques, and association mining and different data extraction tools to extract the information from the web sources. Following methods are used normally for web data extraction purpose:

1. Tree-based techniques

One of the most important features in web data extraction is the semi-structured or unstructured nature of web pages i.e. no specified structure most of the times like in databases. This type of data can be represented by labeled ordered rooted trees, where labels represent the tags of the HTML mark-up language syntax, and the tree hierarchy represents the different levels of nesting of elements constituting the web pages. The representation of a web page by using a labeled ordered rooted tree is known as DOM (Document Object Model). The general idea behind the Document Object Model is that HTML Web pages are represented by means of plain text, which contains HTML tags, i.e., particular keywords defined in the mark-up language that can be interpreted by the browser to represent the elements specific of a Web page (e.g., hyper-links, buttons, images and so forth). HTML tags may be nested one into another, forming a hierarchical structure. This hierarchy is captured in the DOM tree, whose nodes represent HTML tags. [1] This technique is easy and cheaper to implement than other techniques.

2. Web wrappers

Web wrappers are the programs or procedure, that might implement one or different classes of algorithms, which seeks and finds data required by a human user, extracting them from unstructured (or semi-structured) web sources, and transforming them into structured data, merging and unifying this information for further processing, in a semi-automatic or fully automatic way.[1] The main limitation of web wrapper is that for every website or script we have to develop a different program which extract data according to web source hence these are costly as compared to other techniques but faster than other techniques. Web wrappers are characterized by a life-cycle which constitutes wrapper generation, wrapper execution and wrapper maintenance like in life cycle of software.

3. Machine learning approaches

Machine Learning techniques fit well to the purpose of extracting domain-specific information from web sources, since they rely on training sessions during which a system acquires a domain expertise. These techniques are applied on semantic web which is based on machine learning systems. These techniques are performed by automatic systems instead of manually done. Statistical Machine Learning systems are also developed, relying on conditional models [12] or adaptive search [18] as an alternative solution to human knowledge and interaction.

4. Web Data Mining

Web data mining is an application of data mining technique which is used for searching hidden information & patterns from the WebPages. As Web has grown exponentially along with its strengths and its weakness, the strength is that one can find out information on just about anything even if the quality varies. The weakness is that there is the problem of abundance and types of information. Standard data mining techniques may be applied for mining information on the Web [2], but data mining mainly deals with structured form of data organized in well formed databases while web mining deals with unstructured form of data. So mining of web data is one of the most challenging tasks for the data mining. [2]

II. TYPES OF WEB MINING

Web mining is divided into three types [8]:

A. Web Content Mining

It deals with discovering important and useful information or knowledge from web page contents according to user needs. [3] Web consist unstructured information like text, image, audio, and video hence pattern recognitions are used according to type of information. [3] A lot of techniques and tools like statistical, neural network approaches, rapid miner, web data extractor etc. may be use for this purpose.

B. Web Structure Mining

It deals with discovering and modeling the link structure of the web and discovers important information hidden in them. Web information retrieval tools make use of only the text available on web pages but ignoring valuable information contained in web links hence main focus of web structure mining is on link information [2] [3]. It is also used for the purpose of ranking of a webpage on the web. Algorithms like Page Rank [5], Weighted Page Rank [6] and HITS (Hyper-link Induced Topic Search) [7] are available for page ranking over the web.

C. Web Usage Mining

It deals with understanding user behavior in interacting with a particular web site. Web usage mining uses web logs to record user access patterns. Log files are created by web servers and filled with information about user requests on a particular Web site. [4] Web usage mining is used to know the importance of a webpage over others hence important information may be extracted according to the importance of the webpage. According to importance of the webpage web page ranking and content quality may also be improved i.e. integration of web content mining and web usage mining may provide very important results. [24]

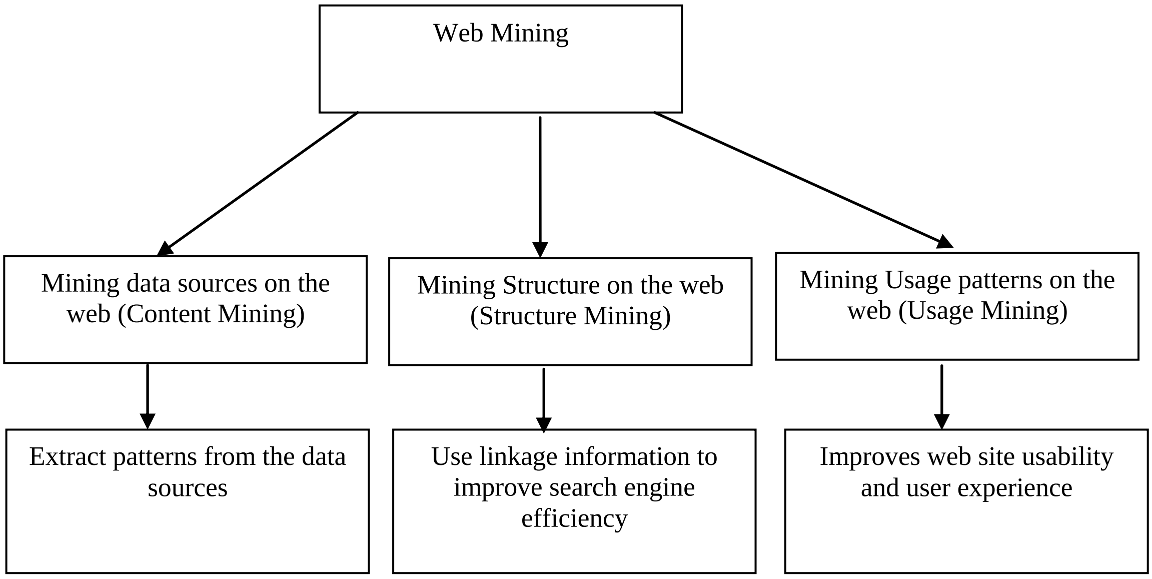
Web Mining
Mining data sources on the web (Content Mining)
Mining Structure on the web (Structure Mining)
Mining Usage patterns on the web (Usage Mining)
Extract patterns from the data sources
Use linkage information to improve search engine efficiency
Improves web site usability and user experience

Figure 1: Web mining and its types

D. Application Areas of Web Data Extraction

i) **Business**: Web Data Extraction systems are used in a wide range of applications like the analysis of text documents in a firm or company (like e-mails, support forum, technical and legal documentation, and so on).e-commerce is also an emerging application of data extraction systems.

ii) **Social Web**: It may be used in different areas like railways, social networking websites (like face book), and shopping websites for efficient information retrieval.

iii) **Real time systems:** Web data extraction systems are used where very fast processing is required like in share markets etc.

iv) **Medical:** In medical field data extraction systems can provide updated information about medicines, patients in an efficient manner.

v) **Bio-informatics**: Data extraction systems can be used in this field for efficiency, security purposes etc.

vi) **Searching Systems:** Web data extraction systems improve the efficiency of search engines by providing limited data.

III. Literature Review

Haitao Yao [26] presented an approach for noise removal from web pages by treating them as images. All of image features were used to measure similarity of noise blocks. It also provides an approach to distinguish noise blocks and information blocks after measuring similarity between them.

Thanda Htwe [25] presented an approach for cleaning web pages by using an approach called case based reasoning. The proposed approach uses Noise Eliminator that detect multiple noise patterns and remove those noise patterns from Web pages of any Web sites. It also apply back propagation neural network algorithm to classify various noise patterns, data patterns and mixture patterns in Web pages.

Leander et al. [9] presented a survey that offers a rigorous taxonomy to classify Web Data Extraction systems. They introduced a set of criteria and a qualitative analysis of various Web Data Extraction tools.

Kushmerick [19] tracked a profile of finite-state approaches to the Web Data Extraction problem. The author analyzed both wrapper induction approaches (i.e., approaches capable of automatically generating wrappers by exploiting suitable examples) and maintenance ones (i.e., the update of a wrapper each time the structure of the Web source changes).

Kushmerick [27], Web Data Extraction techniques derived from Natural Language Processing and Hidden Markov Models.

Flesca [23] surveyed approaches, techniques and tools based on the wrapper induction problem.

Baumgartner [1] is the most updated survey on the state-of-the-art in web data extraction systems. It includes various techniques, applications and current research in the field of web data extraction.

Robert Cooley, Pang-ning Tan [20] provides Web Site Information Filter System (WebSIFT) is a Web usage mining framework that uses the content and structure information from a Web site, and identifies the interesting results from mining usage data.

Haravu and Neelameghan [22] suggest two essential methods of creating these platforms. First, using natural language processing software in text mining, and second, the planning, designing, and developing of a comprehensive multiple media product that would satisfy their target audiences' needs. They trust that these text and data-mining products can only become more useful if the features of a subject classification system are incorporated into text mining techniques and products.

NStein [2009] review about TME (Text Mining Engine), a software program, which optimize the web content over the semantic web.

S. Taherizadeh and N. Moghadam [24] authors describe an approach for integration of web content and web usage mining for pattern discovery. It combines textual content of WebPages with Web server log files to discover useful information and association rules about users' behaviors.

IV. Conclusion:

Web data extraction systems provide efficient information from web sources. It provides fast access, limited content as per user need. It can work over general web as well as semantic web. Web data extraction system can process unstructured or semi-structured form of data which is available in web sources. The need for structured information urged researchers to develop and implement various strategies to accomplish the task of automatically extracting data from Web sources. Web data extraction systems provide a wide range of applications where efficient results can be retrieved. The central thread of this survey is to classify existing approaches along with research areas where web data extraction can provide efficient results.

References:

[1] Web Data Extraction, Applications and Techniques: A Survey by Emilio Ferrara, Pasquale De Meo, Giacomo Fiumara, and Robert Baumgartner published at ACM Computing Surveys, Jul 2012.

[2] Yuefeng Li and Ning Zhong: Web Mining Model and Its Applications for Information Gathering, Knowledge-Based Systems 17, pp. 207–217, 2004.

[3] Rekha Jain and Dr. G. N. Purohit,"Page Ranking Algorithms for Web Mining, International Journal of Computer Applications",ISSN: 0975 – 8887, Volume 13– No.5, pp. 22–25, January 2011.

[4] Claudia Elena DINUCA, "An Application for Data reprocessing and Models Extractions in Web Usage Mining", International Conference on "Risk in Contemporary Economy", Galati, Romania. ISSN 2067-0532, XIIth Edition, 2011.

[5] S. Brin, and L. Page, "The Anatomy of a Large Scale Hypertextual Web Search Engine", Computer Network and ISDN Systems, Vol. 30, Issue 1-7, pp. 107-117, 1998.

[6] Wenpu Xing and Ali Ghorbani, "Weighted Page Rank Algorithm", Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR '04), IEEE, 2004.

[7] J. Kleinberg, "Authoritative Sources in a Hyper-Linked Environment", Journal of the ACM 46(5), pp. 604-632, 1999.

[8] Cooley R., Mobasher B., Srivastava J. "Web mining: Information and Pattern discovery on the World Wide Web. A survey paper". In Proc. ICTAI-97, 1997.

[9] Laender, A. H. F., Ribeiro-Neto, B. A., da Silva, A. S., and Teixeira, J. S. “A brief survey of web data extraction tools”, SIGMOD Rec. 31, 2, pp. 84-93,2002.

[10] Zhao, H. “Automatic wrapper generation for the extraction of search result records from search engines”. Ph.D. thesis, State University of New York at Binghamton, Binghamton, NY, USA. Adviser-Meng, Weiyi, 2007.

[11] Irmak, U. and Suel, T. “Interactive wrapper generation with minimal user effort”. In Proc. of the International Conference on World Wide Web , ACM, Edinburgh, Scotland, pp. 553-563,2006.

[12] Phan and Horiguchi,”Automated data extraction from the web with conditional models” International journal of Business Intelligence and Data Mining. pp. 194-209, 2005.

[13] Garrett, J. J.” Ajax: A new approach to web applications”. Technical report.

[14] Rahm “Data cleaning: Problems and current approaches” IEEE Bulletin on Data Engineering, 2000.

[15] Monge “Matching algorithm within a duplicate detection system”, IEEE Techn. Bulletin Data Engineering, 2000.

[16] Rahm and Bernstein ” A survey of approaches to automatic schema matching. The VLDB Journal vol 10, pp. 334-350, 2001.

[17] Berthold, M. and Hand, D. J. “Intelligent Data Analysis: An Introduction.”, 1999. [18] Turmo, J., Ageno, A., and Catal_a, N, “Adaptive information extraction” Springer- Verlag New York, ACM Computing Survey, USA.

[19] Kushmerick, N.”Wrapper induction: effciency and expressiveness. Artif.” Intell. 118, 1-2, pp.15-68, 2000.

[20] Robert Cooley, Pang-ning Tan “WebSIFT: The Web Site Information Filter System”, In Proceedings of the Web Usage Analysis and User Profiling Workshop, 1999.

[21] Catanese, S., De Meo, P., Ferrara, E., and Fiumara, G.” Analyzing the facebook friendship graph” In Proc. of the 1st International Workshop on Mining the Future Internet, pp.14-19, 2010

[22] Haravu, L.J. and A. Neelameghan. “Text Mining and Data Mining in Knowledge Organization and Discovery: The Making of Knowledge- Based Products.” Knowledge Organization and Classification in International Information Retrieval, NY, Haworth, 2003.

[23] Flesca, S., Manco, G., Masciari, E., Rende, E., and Tagarelli, A.” Web wrapper induction: a brief survey”. AI Commun. 17, 2, pp. 57-61,2004.

[24] S. Taherizadeh and N. Moghadam “ Integrating Web Content Mining into Web Usage Mining for Finding Patterns and Predicting Users’ Behaviors” in International Journal of Information Science and Management, Vol. 7, No. 1, pp. 51-66, june 2009.

[25] Thanda Htwe,”Cleaning Various Noise Patterns in Web Pages for Web Data Extraction”, International Journal of Network and Mobile Technologies, ISSN 1832-6758,VOL 1 ,ISSUE 2 , nov 2010.

[26] Haitao Yao ,The Noise Reduction Method of Web Pages Based on Image Features, International Conference on Computational Intelligence and Software Engineering, IEEE ,pp.1-5,Dec 2009.

[27] Kushmerick, N “Finit-state approaches to web information extraction”, Proc. of 3rd Summer Convention on Information Extraction, pp. 77-91, 2002