A Comprehensive Survey on Web Content Extraction Algorithms and Techniques

Conference Paper · June 2013

DOI: 10.1109/ICISA.2013.6579445

2 authors:

Sumaia M. Al-Ghuribi
Taiz University
11 PUBLICATIONS **17** CITATIONS
SEE PROFILE

Saleh Alshomrani
King Abdulaziz University
17 PUBLICATIONS **99** CITATIONS
SEE PROFILE

A Comprehensive Survey on Web Content Extraction Algorithms and Techniques

Sumaia Mohammed AL-Ghuribi and Saleh Alshomrani
Faculty of Computing and Information Technology,
King Abdulaziz University, Jeddah, Saudi Arabia
salghoraibi0001@stu.kau.edu.sa , sshomrani@kau.edu.sa

***Abstract*—**Web Content Extraction is an important problem that has been studied through different approaches and algorithms. It is interested in extracting meaningful and useful data from the Webpage which is surrounded with many noisy data such as advertisements and navigation links. Many applications get benefits from the extracted content such as crawlers, indexers, document classification, and Information retrieval. This survey aims at providing a comprehensive overview of many approaches that constructed for extracting Webpage content. In this survey, Web Content Extraction approaches are classified into categories and for each category, some approaches are given in details with their weakness. Based on analyzing the given approaches deeply, we can draw the fundamentals factors for constructing the optimal Web content extractor.

Keywords— Web Content Extraction, useful data, noisy data, optimal extractor.

I. INTRODUCTION

The Internet has become a key communication and information medium for various organizations [30].It is a medium for accessing a great variety of information stored in different parts of the world. In recent years, the growth of the World Wide Web exceeded all expectations. Today, there are several billions of HTML documents, pictures and other multimedia files available via Internet and the number is still rising [31]. Mining the Web data is one of the most challenging tasks for data mining and data management scholars because there are huge heterogeneous, less structured data available on the Web and we can easily get overwhelmed with data [32]. Also, Analysis and discovery of useful information from World Wide Web poses a phenomenal challenge to the researchers in this area. This phenomenon of retrieving valuable information by adopting data mining techniques is called Web mining [33]. i.e., use of data mining techniques to automatically discover and extract information from World Wide Web documents and services [34]. Web content mining (WCM) is one of the Web mining categories, which is the process of identifying user specific data from text, image, audio or video data [34] already available on the Web. This process is alternatively called as Web text mining [34]. Web content is mixed together with formatting code, advertisement, or Webpage specific information such as navigation links; in that way that Webpage is human friendly, but the actual content is not machine friendly [35,36]. The mix of noisy data with the useful content (i.e. the content which is most important and interest for users) leads to the need of content extraction or content filtering. Web content extraction is kind of Web content mining that aims to extract useful and meaningful data from the Webpage. Web content extraction is an important field for many tasks as information searching, natural language processing (NLP), open source intelligence (OSInt), web documents classification, Information Retrieval (IR), machine translation and text summarization.

Also Web content extraction helps the PDA users and mobile phone users with limited features to get the relevant information quickly without taking a long time to operate the scroll bar and skip a lot of irrelevant information. In addition it is useful for blind and visually impaired users by either increasing the font of the extracted content or pass it to speech. Traditional techniques that are used in content extraction aims to remove some clutters , removing images, increasing font size to make content more readable or disabling JavaScript all of which remove the Webpage's inherent look-and-feel. In this survey most of the new techniques and algorithms for extracting Webpage content are presented with their limitations to conclude the main factors for constructing optimal Web content extractor that can be applicable for any Webpage with any style. The rest of this paper is organized as follow: section II presents five different techniques for extracting Web content; section III gives the main factor for constructing optimal extractor and finally we conclude this survey in Section IV.

II. WEB CONTENT EXTRACTION TECHNIQUES

A. *Wrappers for Content Extraction*

Wrapper in data mining is a program that extracts content of a particular information source and translates it into a form. There are two main approaches to wrapper generation: wrapper induction and automated data extraction [31]. Wrapper induction uses supervised learning to learn data extraction rules from manually labeled training examples [31]. Many researches constructed wrappers for the purpose of content extraction such as [15,19,25] but the construction of wrappers is complex task because it needs expert users to write the extraction rules for the extraction process , this leads to make construction time consuming and

978-1-4799-0604-8/13/$31.00 ©2013 IEEE

restrictive. Also, wrapper maintenance is a difficult task because whenever a page is changed, wrapper must be built again. Follow is two researches for wrappers.

Kushmerick et al. [15] introduced a method for content extraction called wrapper induction to solve the problem of regular wrappers which are hand-coded. Wrapper induction consists of three contributions. First, the wrapper construction problem is formalized as induction. Second, HLRT class, which is responsible for handling various Internet information resources which their contents are displayed in tabular layouts, is defined. To make the wrapper induction faster, HLRT is designed similar to finite state automata induction. HLRT class will enhance the extraction because it is efficiently learnable. Third, heuristic knowledge is used for composing algorithm's oracle which oracle is the key on induction. This method is too time consuming which affects its efficiency.

Tripathy et al. [25] presented a new tool for information extraction called Visual Wrapper for Extraction of Data using DOM tree (VEDD) to make the search process more efficient and reliable. VEDD filters the resulted Webpages and extracts their contents to return the most relevant Webpages according to the user's needs. This is done by using the following two processes, first is parsing the Webpages by DOM tree to understand their structures. Second is building the Wrapper by using two extraction techniques the first is BFS (Breadth First Search) Extraction using link extractor and the second is Data Filtering using text extractor which created by Hong & Fauzi [11]. Finally resulted Webpages are filtered using string matching function. This method used the text extractor [11] for extracting text which is based on heuristic techniques and observations , the main observation for extracting data is that Data Records usually consist of three HTML tags that make up their tree structure. Again Webpages are different structures and always updating, so we cannot generalize any rule for Webpage structure.

B. Template detection for extracting content

A template is a pre-prepared master HTML shell page that is used as a basis for composing new Webpages. The content of the new pages is plugged into the template shell, resulting in a collection of pages that share a common look and feel [2]. Templates can appear in primitive form, such as the default HTML code generated by HTML editors like Netscape Composer or FrontPage Express, or can be more elaborate in the case of large Websites. These templates sometimes contain extensive navigational bars that link to the central pages of the Website, advertisement banners, links to the FAQ and Help pages, and links to the Website's administrator page [2]. There are many studies and algorithms are designed to detect the templates of Webpages in order to extract Webpage's content such as [2, 4, 5, 11, 16, 26]. The weakness of these algorithms is that they extract the Webpage's content depends on Webpage's template , where Webpage's design always develops and varies from other Webpages, which makes this algorithms not suitable for most of Webpages' styles. Follow two algorithms for detecting Webpage's template is discussed.

Wang et al. [26] presented an incremental framework for detecting Webpage template where the page is processed since it has been crawled (i.e. no need for caching any Webpage). Every Webpage is passed through four steps page segmentation, text segment table expansion, template detection, and text segment table shrinkage. From the previous four steps the template of the Webpage is determined and be easily extracted its content. This framework is inaccurate for extracting because its time consuming and for each Website a model must build because of the varieties of Webpages' templates.

Hong and Fauz . [11] proposed a three-step approach (template generation , template detection and data extraction) to extract data from Webpages. In template generation step, training Webpages are used to generate a set of templates. In template detection step, a template is detected for the tested page from the set of templates generated in the previous step based on similarity. After the template of the tested page is detected, the data of the page is extracted easily which is data extraction step. This approach extracts data depends on the template of the page so it cannot perform well on Webpages with different templates or even processes two pages with different templates at the same time.

C. Content Extraction using Machine learning

Machine learning, a branch of artificial intelligence, is about the construction and study of systems that can learn from data, the core of machine learning deals with representation and generalization. Representation of data instances and functions evaluated on these instances are part of all machine learning systems. Generalization is the property that the system will perform well on unseen data instances [32]. Machine learning focuses on prediction, based on known properties learned from the training data. Both Clustering and Classification are types of Machine learning and used for content extraction. These needs to train data whenever Webpages data and structure are changed, many researches based on machine learning are done such as [13, 17,27,29,6,22] follow are two of them:

Debnath et al. [6] proposed ContentExtractor and Feature-Extractor algorithms to identify the primary informative content of Webpages of some Websites. First, the ContentExtractor algorithm used the inverse block document frequency (IBDF) to classify each block as redundant block or identified block by measuring the similarity using SIM. Second, the Feature-Extractor algorithm is used to identify blocks with some features like text, text-tag, list and style sheet. Finally using clustering algorithm, the block with the desired feature (winner) is chosen to be the primary informative content. These are efficient algorithms with single body Webpage but with multi-body are not. It assumes that primary content is one block only.

Pasternack and Roth. [22] introduced a method for extracting the texts of the Webpage's article called

Maximum Subsequence Segmentation (MSS) which combines heuristic method and supervised learning methods. For any Webpage three tasks are done, first is deciding whether the page contains an article or not. Second, if it does, a contiguous block is identified from the first word of the Webpage's article to the last word. Third is removing the rested things other than the article's block. This approach is a supervised learning which always needs training when new features are made to web structure. In most new Webpages the article text is put in either table or Div tags, this approach deletes all tables which make it not suitable for them.

D. Content Extraction using Visual Cues and/or Assumptions

Many researches are done for extracting content of Webpages based on visual cues and/or assumptions such as [18 , 3 ,12 , 28 , 37], they made assumptions on the Webpage's structure , content nodes , main content place , some tags and etc. to limit their techniques and make it easier. In another way they make generalization for the Webpage, which makes the extraction technique less efficient, follow are two of them:

Hong et al. [12] developed Visual Wrapper for Extraction of Records (ViWER) algorithm to extract data using visual cues and DOM tree. ViWER consists of two components, first is parsing the Webpage into DOM tree and second is extracting data using visual cues and DOM tree properties. A Breadth First Search (BFS) algorithm is used to detect and label the correct data regions, then four filter processes is done to obtain the output. This algorithm maybe gives effective results in some Webpages but in most Webpages they do not because it is built based on assumptions in every step of its work to limit the structure of a Webpage which is tedious, not realistic and makes it not applicable for many Webpages.

Zhang and Wang [37] proposed a method for extracting meaningful information from Chinese Webpages. The main point is to construct TCDT (Title and Content Dependency Tree) to get the meaningful content by finding the smallest dependency distance. Five steps are done for achieving the goal of the method using DOM tree and some statistical calculations. This method makes a lot of assumptions in most of its steps like extracting title and content block, these assumptions are made from some Webpages and they will be not applicable to others. The result of this method is a node which considered a meaningful node, where in fact the meaningful content contains of many nodes. Lastly, this method is designed for Chinese Webpages so it performs less efficient in English language and other languages.

E. Content Extraction Based on HTML Features and/or Statistics

The first four types of extraction techniques try to make a generalization for a Webpage, which make them less accurate and less efficient. This type of technique is the most used and studied currently and give better output than the previous techniques, Many techniques for extraction based on HTML features and/or statistics are done such as [10,20,7,23,8,1,21,24,9,14] ,follow two of them:

Gottron . [8] introduced a novel algorithm for content extraction called Content Code Blurring (CCB). Two approaches are defined to convert document into a structure in order to defining the concept of regions. As a result, the document is represented as a sequence (vector) of elements that are either code or content called content code vector. Regions, based on consisting either content or code, are calculated through calculating content code ratio for each element. The ratio will be high, if they are mainly content, but if they are mainly code, the ratio will be low. This is an efficient algorithm but it does not classify content into useful and usefulness, also this is not suitable for multi-body Webpages.

Qureshi and Memon [24] presented a hybrid model for extracting useful content from html documents which is derived from different models, one model is based on statistical feature and the other is based on formatting characteristics. To analyze the html document, Cobra is used to convert html document to DOM tree. The derivation and link densities of each unique node are calculated. Nodes that have same styles, low link densities and high derivation are considered as useful content. This model is efficient but has some weakness, because it takes styles into account, it will be resource consuming and expensive operation. Also in this model some useless nodes are considered as useful if they have the same style of useful nodes even they have high link density and low text derivation.

III. THE FUNDAMENTALS FACTORS FOR CONSTRUCTING THE OPTIMAL WEB CONTENT EXTRACTOR

Depends on analyzing the previous techniques for Web content extraction process, we can find the fundamentals factors that will improve the extraction process if they applied in constructing the extractor model. They are as follow:

- The extractor should be independent and can be applied to any language Arabic, English, Japanese, etc.
- As much as possible assumptions should not be made while constructing the extractor.
- The extractor should be applicable for both single body and multi-body Webpages.
- The useful content and noisy data should not be limited to specific tags.
- The extractor should not depend on specific template or specific structure.
- The extractor should be able to process two Webpages from different Websites (i.e. different structure) at the same time.
- The extractor should consume reasonable time and memory.

- The extractor should not depend on result of training data on extracting main content because Webpages' structures always update which leads to updating the training data (i.e. long process).

TABLE 1. WEB CONTENT EXTRACTION'S TECHNIQUES

Research No	[7]	[23]	[8]	[1]	[13]	[27]	[26]	[11]	[17]	[22]	[29]	[18]	[3]	[28]	[20]	[24]	[14]
Technique's Attributes																	
Make assumptions	√	√		√										√			
Use DOM Tree				√				√	√		√	√	√		√	√	√
Use Classification or Clustering techniques					√	√		√	√	√	√						
Depend on Template					√		√	√									
Has Statistical Operations	√		√	√	√	√			√			√		√	√	√	
Base on HTML features	√	√	√	√	√		√		√	√	√	√	√	√	√	√	√
Technique's Experiment																	
Datasets	580	188	9601	1500	26518	1778	400	4944	1781	411	5 classes	26518	140	1600	N/A	3 datasets	500
HTML document's language*	E	E	E	E	E	E/Ch	E	E	E	E	E	E	E	Ch	N/A	E	E
Recall	N/A	0.862	N/A	N/A	0.956	0.93	0.80	0.98	N/A	0.98	N/A	0.956	N/A	N/A	N/A	N/A	0.73
Precision	N/A	0.688	N/A	N/A	0.956	0.92	N/A	0.99	N/A	0.93	N/A	0.956	N/A	N/A	N/A	N/A	0.80
F1-measure	N/A	N/A	0.770	0.930	N/A	0.93	N/A	N/A	0.88	0.95	0.77	N/A	N/A	0.93	N/A	0.94	0.73
Technique's applicability																	
Single body Webpage	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√
Mutli-body Webpage				√	√	√	√	√	√	√	√					√	√

*Where E: English & Ch: Chinese

IV. CONCLUSION AND FUTURE WORK

With the enormous increasing of Webpages, Web content extraction becomes one of the important topics to improve many applications while using Web page as a source of knowledge. In this survey, we presented some techniques for extracting Web content and divide them into five groups and focus on Content Extraction Based on HTML Features and/or Statistics type as it is the most used recently. Finally we get some factors from analyzing some techniques to construct an optimal constructor. As a future work, we will try to construct an extractor model which contains the previous factors in order to enhance the extraction process.

REFERENCES

[1] Adam, G.; Bouras, C.; Poulopoulos, V., "CUTER: An Efficient Useful Text Extraction Mechanism," Advanced Information Networking and Applications Workshops, 2009. WAINA '09. International Conference on , vol., no., pp.703,708, 26-29 May 2009.

[2] Bar-Yossef Ziv and Rajagopalan Sridhar," Template detection via data mining and its applications," In Proceedings of the 11th international conference on World Wide Web, 2002, Honolulu, Hawaii, USA.ACM 1-58113-449-5/02/0005,pp. 580–591.

[3] Cai Deng ; Yu Shipeng; Wen Ji-Rong ; Ma Wei-Ying ," Extracting content structure for web pages based on visual representation," . In Proceedings of the 5th Asia-Pacific web conference on Web technologies and applications (APWeb'03), Xiaofang Zhou, Maria E. Orlowska, and Yanchun Zhang (Eds.). Springer-Verlag, Berlin, Heidelberg, 406-417, 2003.

[4] Chakrabarti Deepayan; Kumar Ravi; Punera Kunal," Page-level template detection via isotonic smoothing," In Proceedings of the 16th international conference on World Wide Web (WWW '07). ACM, New York, NY, USA, 61-70, 2007.

[5] Chen Liang ; Ye Shaozhi ; Li Xing," Template detection for large scale search engines," In SAC, pp. 1094–1098. ACM, 2006.

[6] Debnath Sandip ; Mitra Prasenjit ; Giles C. Lee,"Identifying content blocks from web documents," in: Proceedings of the 15th ISMIS 2005 Conference, 2005, pp. 285–293.

[7] Finn Aidan; Kushmerick Nicholas; Smyth Barry," Fact or fiction: Content classification for digital libraries", 2001.

[8] Gottron Thomas, "Content Code Blurring: A New Approach to Content Extraction," Database and Expert Systems Application, 2008. DEXA '08. 19th International Workshop on , vol., no., pp.29,33, 1-5 Sept. 2008.

[9] Gunasundari R. and Dr.Karthikeyan S.," STUDY OF CONTENT EXTRACTION FROM WEB PAGES BASED ON LINKS,"International Journal of Data Mining & Knowledge Management Process, 2012. Vol.2, Issue.3, ISSN 2231-007X.

[10] Gupta Suhit; Kaiser Gail; Neistadt David; Grimm Peter , "DOM-based content extraction of HTML documents," in: Proceedings of the 12th International Conference on World Wide Web, WWW '03, ACM, New York, NY, USA, 2003, pp. 207–214, http://doi.acm.org/10.1145/775152.775182.

[11] Hong Jer Lang and Fauzi Fariza, "Tree Wrap-data Extraction Using Tree Matching Algorithm," Majlesi Journal of Electrical Engineering, Iran, June 2010, pp. 43-55.
[12] Hong Jer Lang; Siew Eu-Gene; Egerton Simon, "ViWER- data extraction for search engine results pages using visual cue and DOM Tree," Information Retrieval & Knowledge Management, (CAMP), 2010 International Conference on , vol., no., pp.167,172, 17-18 March 2010 .
[13] Hung-Yu Kao; Shian-Hua Lin; Jan-Ming Ho; Ming-Syan Chen; , "Mining Web informative structures and contents based on entropy analysis," Knowledge and Data Engineering, IEEE Transactions on , vol.16, no.1, pp. 41- 55, Jan. 2004.
[14] Insa David;Silva Josep; Tamarit Salvador, "Using the words/leafs ratio in the DOM tree for content extraction," The Journal of Logic and Algebraic Programming, Available online 9 February 2013, ISSN 1567-8326.
[15] Kushmerick Nicholas; Weld Daniel S.; Doorenbos Robert," Wrapper Induction for Information Extraction," Proceedings of the International Joint Conference on Artificial Intelligence, 1997.
[16] Lei Fu; Yao Meng; YingJu Xia; Hao Yu, "Web Content Extraction based on Webpage Layout Analysis," Information Technology and Computer Science (ITCS), 2010 Second International Conference on , vol., no., pp.40,43, 24-25 July 2010.
[17] Li Zhao; Ng Wee Keong; Sun Aixin," Web data extraction based on structural similarity," . Knowl. Inf. Syst. 8, 4 (November 2005), 438-461.
[18] Lin Shian-Hua and Ho Jan-Ming,"Discovering informative content blocks from Web documents," in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '02, ACM, New York, NY, USA, 2002. pp. 588–593.
[19] Liu, L.; Pu, C.; Han, W.; , "XWRAP: an XML-enabled wrapper construction system for Web information sources," Data Engineering, 2000. Proceedings. 16th International Conference on , vol., no., pp.611-621, 2000.
[20] Mantratzis Constantine ; Orgun Mehmet; Cassidy Steve, "Separating XHTML content from navigation clutter using DOM-structure block analysis," in: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia, HYPERTEXT '05, ACM, New York, NY, USA, 2005, pp. 145–147.
[21] Mingsheng Hu; Zhijuan Jia; Xiangyu Zhang, "An approach for text extraction from web news page," Robotics and Applications (ISRA), 2012 IEEE Symposium on , vol., no., pp.562-565, 3-5 June.2012.
[22] Pasternack Jeff and Roth Dan,"Extracting article text from the web with maximum subsequence segmentation," In WWW, pages 971–980. ACM, 2009.
[23] Pinto David ; Branstein Michael; Coleman Ryan; Croft W. Bruce ; King Matthew ; Li Wei ;Wei Xing,"QuASM: A system for question answering using semi-structured data," in: Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, JCDL '02, ACM, New York, NY, USA, 2002,pp. 46–55.
[24] Qureshi Pir Abdul Rasool ;Memon Nasrullah,"Hybrid model of content extraction," Journal of Computer and System Sciences.2012. Vol. 78, Issue 4, July 2012, pp. 1248-1257, ISSN 0022-0000.
[25] Tripathy, A.K.; Joshi, N.; Thomas, S.; Shetty, S.; Thomas, N., "VEDD- a visual wrapper for extraction of data using DOM tree," Communication, Information & Computing Technology (ICCICT), 2012 International Conference on , vol., no., pp.1,6, 19-20 Oct. 2012.
[26] Wang Yu ;Fang Bingxing; Cheng Xueqi ; Guo Li; Xu Hongbo,"Incremental web page template detection", in Proc. WWW, 2008, pp.1247-1248.
[27] Weninger Tim; Hsu William; Han Jiawei, "CETR: content extraction via tag ratios, "in: Proceedings of the 19th International Conference on World Wide Web,WWW '10, ACM, New York, NY, USA, 2010, pp. 971–980.
[28] Yang Dingkui and Song Jihua , "Web Content Information Extraction Approach Based on Removing Noise and Content-Features," Web Information Systems and Mining (WISM), 2010 International Conference on, vol.1, no., pp.246-249, 23-24 Oct. 2010.
[29] Zachariasova Martina; Hudec Robert; Benco Miroslav; Kamencay Patrik,"Automatic extraction of non-textual information in web document and their classification," Telecommunications and Signal Processing (TSP), 2012 35th International Conference on , vol., no., pp.753,757, 3-4 July 2012.
[30] Q. Zhang and R. Segall," Web mining: a survey of current research,Techniques, and software", in the International Journal of Information Technology & Decision Making Vol. 7, No. 4 ,2008.
[31] V. Bharanipriy and V. Prasad," WEB CONTENT MINING TOOLS: A COMPARATIVE STUDY," International Journal of Information Technology and Knowledge Management, Volume 4, No. 1, pp. 211-215. January-June 2011.
[32] S. Ajoudanian and M. Jazi, "Deep Web Content Mining," World Academy of Science, Engineering and Technology 49, 2009.
[33] B. Liu and K. Chang,"Editorial Issue on Web Content Mining", issue2, 2004.
[34] Zubi, Z.()."Using Some Web Content Mining Techniques for Arabic Text Classification".ISSN 1790-5109.pp.73-84.
[35] S. Chakrabarti, Mining the Web: Discovering Knowledge from Hypertext Data, Morgan Kaufmann Publishers, 2003.
[36] D. Eichmann, The RBSE spider – balancing effective search against web load, in: First World Wide Web Conference, Geneva, Switzerland, April 20,1994.
[37] ZHANG Bin and 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, 2012. vol. 19, no. 2, pp. 147-151,189, October 2012,ISSN 1005-8885.