NOISE REDUCTION INFORMATION USING BY WEB PAGES

Hilery Bhanuprasad Rathod

Researcher Scholar (Dep. Of English, Gujarat University, Ahmdabad)

ABSTRACT:

In this paper I review studies of the growth of the Internet and technologies that are useful for information search and retrieval on the Web. I present data on the Internet from several different sources, e.g., current as well as projected number of users, hosts, and Web sites. Although numerical figures vary, overall trends cited by the sources are consistent and point to exponential growth in the past and in the coming decade. Hence it is not surprising that about 85% of Internet users surveyed claim using search engines and search services to find specific information. The same surveys show, however, that users are not satisfied with the performance of the current generation of search engines; the slow retrieval speed, communication delays, and poor quality of retrieved results (e.g., noise and broken links) are commonly cited problems. I discuss the development of new techniques targeted to resolve some of the problems associated with Web-based information retrieval and speculate on future trends.

Keywords: *Information and communication engineering,Web Content Mining,Web Structure Mining,World Wide Web.*

1. INTRODUCTION

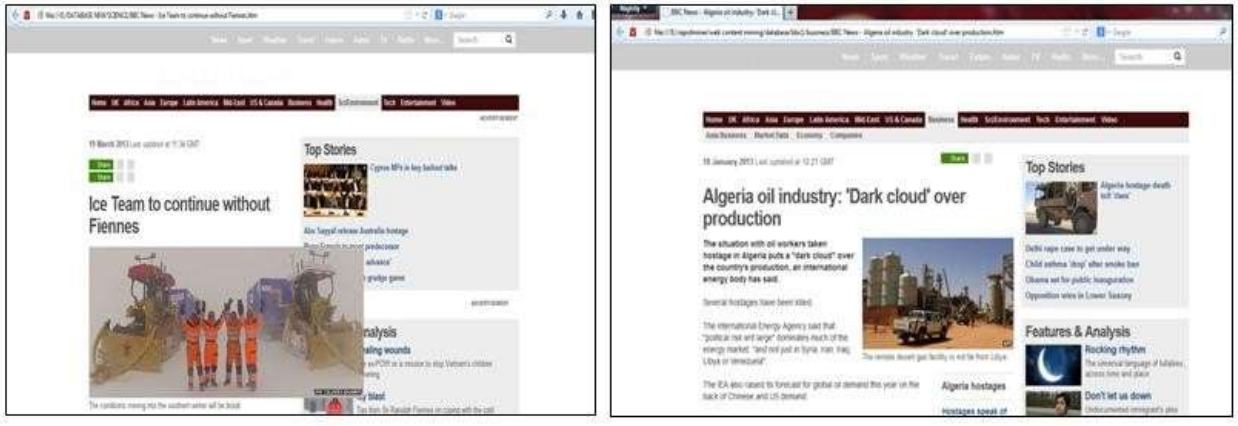
Web mining is an emerging research area due to the rapid growth of websites. Web mining is classified into Web Content Mining , Web Usage Mining and Web Structure Mining . Extraction of required information from web page content available on World Wide Web is Web Content Mining. The Web Content Mining is further classified into two categories first category is to directly mine the content on documents and second category is to mine the content using search engine. The mining method focuses on the information extraction and integration. The content of Web may be text, image, audio, video. Web pages typically contain a large amount of information that is not part of the main contents of the pages, like banner advertisements, navigation bars, copyright notices, etc. Such noises on Web pages usually lead to poor results in Web mining. This research focuses on the problem of Noise free Information retrieval on web pages, which means the pre-processing of Web pages automatically to detect and eliminate noises. This research work proposes an approach for eliminating noises from web pages for the purpose of improving the accuracy and efficiency of web content mining. The main objective of removing noise from a Web Page is to improve the performance of the search. It is very essential to differentiate important information from noisy content that may misguide users' interest. This approach mainly concentrates on removing the following noises in stages: (1) Primary noises-Navigation bars, Panels and Frames, Page Headers and Footers, Copyright and Privacy Notices, Advertisements and other Uninteresting Data such as audio, video, multiple links. (2) Duplicate Contents and (3) Noise Contents according to block importance. The removal of these noises is done by performing three operations. Firstly, using the Block Splitting operation, primary noises are removed and only the useful text contents are partitioned into blocks. Secondly, using simhash algorithm, the duplicate blocks are removed to obtain the distinct blocks..The importance of the block is then calculated using simhash algorithm. Based on a threshold value the important blocks are selected using sketching algorithm and the keywords are extracted from those important blocks. The performance of the proposed approach is evaluated with several web pages and the results ensure the effectiveness of the proposed approach in identifying the important blocks, which are relevant for knowledge extraction from web pages

2.GLOBAL NOISES

These are noises on the Web with large granularity; they are usually no smaller than individual pages. Global noises are like mirror sites, legal/illegal duplicated Webpages, old versioned Web pages to be deleted, etc.

2.1 Local(intra-page noises

Figer 1: Two different web pages from BBC news site

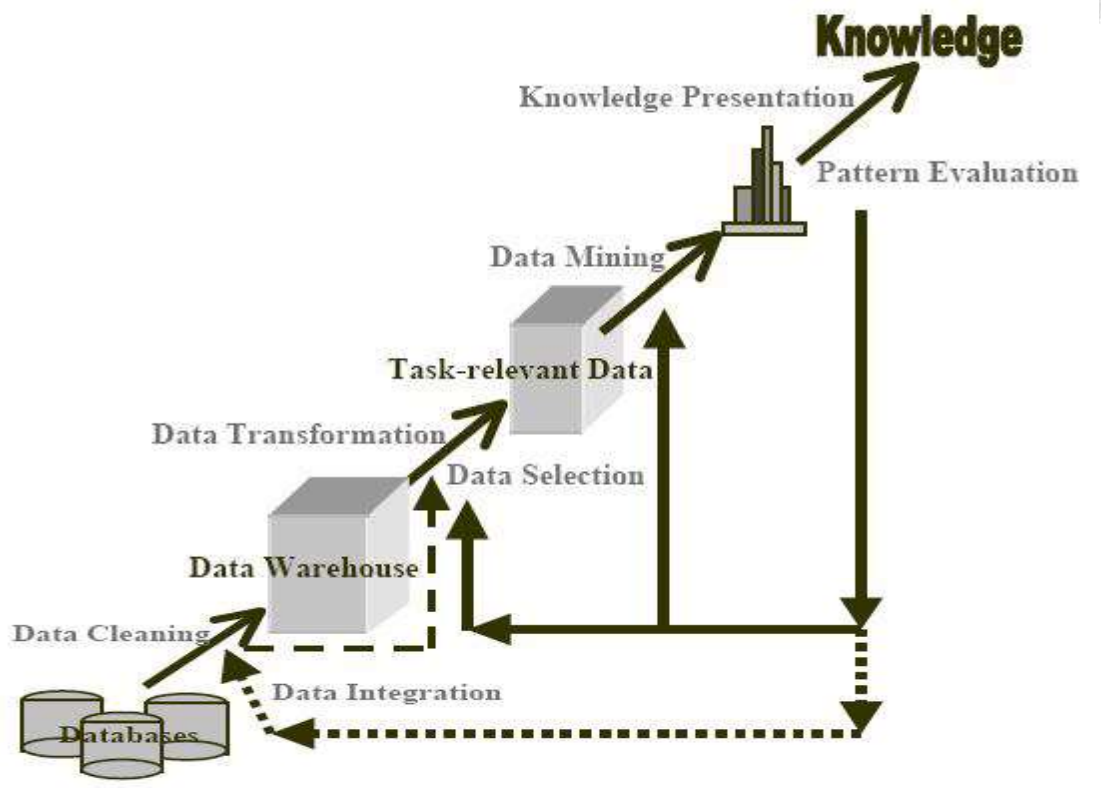


These are noisy regions within a Web page. Local noises are usually incoherent with the main contents of the Web page. Such noises include banner advertisements, navigational guides, decoration pictures, etc. This paper, first introduces a new technique to capture the actual contents of the pages in a Web site using RapidMiner. This technique is based on term frequency and inverse document frequency of words in a document and based on it; noises are detected from the web pages. Secondly, it also detects the redundant data if it exists in a document database. Web content mining intends to mine valuable information or knowledge from Web page contents. The focus of web data extraction is to extract the contents from web pages for other applications like summarization, task learning, etc.

3. DATA

Data plays an important role in today's world. It is animportant area of research, since the volume of data is available in most applications. This huge amount of data processed to extract useful information and knowledge because they are not unique. Data is the process of discovering interesting knowledge from large amount of data.

Figure-2 :The complete data process is given in



3.1 Types of web data

The World Wide Web contains different information in different formats. As indicated above, the World Wide Web is composed of three typesof data; the classification is shown in Figure 1.2. Web content data are the data that the web pages are designed for the presentation to the user. It consists of free text such as semi-structured data such as Hypertext Markup Language (HTML) pages and more structured data, like automatic generated HTML pages, Extensible Markup Language (XML) files, or data generated in the tables for Web content. Text, image, audio and video data types are all falling into this category.

4. SEARCH ENGINE

Google is one of the most popular and widely used search engines. Google provides web users with information from more than 2.5 billion web pages that it has indexed on its server. Compared to other search engines, Google has a simple and quick search facility. This property makes it the most widely used search engine. Previous engines based on the web content in order to fetch a web page as a result of submitted query. However, Google was first engine to reflect importance of web structure namely link analysis from the web. The key method of Google called PageRank, measures an importance of a page, is the underlying technology in all search products. PageRank method makes use of the information about link structure of the web graph, this method plays significant role for returning relevant results to simple query.

5. PROBLEM IDENTIFICATION

In web pages, it is very essential to differentiate important information from noisy content that may misguide users' interest. Here the noise means irrelevant data like advertisement which may be the local noise. The main goal of this approach is to remove the noises from web pages. The
removal of these noises from web pages is done by performing three operations namely, 1. Block Splitting Operation, 2.Eliminating the Duplicate Blocks, and 3.Finding the Block Importance of each and every block. Using the threshold value, the blocks with less importance are eliminated. The remaining blocks with high importance are considered as important blocks and the keywords are extracted from those important blocks.

6. LITERATURE REVIEW

the related work for this research work and existing techniques for noise reduction information retrieval of web pages is discussed. It is observed that the web page noise reduction information retrieval is related to feature selection, feature weighting, block splitting, duplicate block elimination, and important block calculation in the of web content mining field where text files or databases are preprocessed to improve subsequent mining tasks by filtering irrelevant or useless information.

Isabelle Guyon and Andrne Elisseef (2003) have proposed a feature selection technique is dealing with text categorization the high dimensionality of feature space. Some feature selection methods remove non-informative terms according to some prior criteria like information gain, document frequency, term frequency mutual information and etc.

Bekkerman et al (2003) has proposed a further dimension by constructing higher level dimensions from combining lower level dimensions. Textual documents or web document are typically modeled as a term vector space where features are personal terms. However, local noise in web pages is regularly blocked by items like images, Texts, hyperlinks etc and Instead of single individual terms. Furthermore, the vector space model cannot capture the occurring position of terms in web pages.

7. WEB PAGE NOISE

A large number of web pages contained useful information is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices, etc. These noise data can seriously harm for web miners by extracting whole document rather than the informative content and also retrieve non-relevant results. It is also important to distinguish valuable information from noisy data within a single web page. The web pages are constructed not only main contents information like product information in shopping domain, job information in a job domain but also advertisements bar, static content like navigation panels, copyright sections, etc. When web documents are processed, the main content is surrounded by noise in the retrieved data. Therefore, without removing such data, the efficiency of feature extraction and finally text classification is certainly degraded. Web noise can be classified as global noises and local noise (Yi et al 2003). Global noises include mirror sites; legal/illegal duplicated web pages, old versioned web page with advertising segments, unnecessary images, or navigation links, etc.

8. NOISE IDENTIFICATION

Information in a web page is not uniformly significant. For example, consider the web page in Figure the caption in a news website is much more attractive to users than the navigation bar. And users only just pay attention to the advertisement or the copyright when they browse a web page

Figure 3 Sample Webpage Containing Multiple Regions with Different Importance

Therefore, dissimilar information in a web page has dissimilar importance weight according to its location, occupied area, content, etc. Thus, it is to assign importance to a region in a web page, and need to segment a web page into a set of blocks.

9. TYPES OF NOISE

Noise data of web documents can be categorized into two groups such as global noise and local noise. Global noises are redundant web pages over the Internet such as mirror sites and legal or illegal duplicated web pages. Local noises, only related intra-page redundancy and exist in the web page. This research work focuses on the local

noise elimination method. There are at least four different known categories of noise pattern within Web pages of any web sites including banners with links including search panels, advertisements, navigational panel (directory list) and copy right and privacy notice in each web site. It can be seen that many web pages contain these four noise categories together but most of noise patterns are structured by using sectioning tags such as <TABLE> and <DIV> and sectioning separating tags like <FRAMESET>, and interactive tags like <SELECT>, <FIELDSET> , Input moreover, anchor tag <A> and <IMG> tag are most commonly used to link another web page or another web site. However, these four noise categories can be structured by using various noise patterns.

9.1 fixed noise

Fixed noise is usually descriptive the information on a webpage or a website. It consists of three sub-types:

1. Decorating noise like site logos and decorative graphics or text, etc.
2. Statement noise, such as copyright notices, privacy statements, license notices, terms and conditions, partners or sponsors statement and etc.
3. Page description noise like date, time and visit counters of the current page and etc.

Figure 4 :Example of Fixed Noise

Figure 4 :show some examples of fixed noise description of an actual web page. The fixed noise description is usually found either in the form or content.

9.2 web service noise

Figure 5 : Examples of Web Services Noise

Many web pages contain useful service blocks by the way to page content or to manage the server to communicate. This web service blocks are known as noise. There are three types of web service noise illustrated in the Figure 5

1. Page noise, as the management of this page and page relocation, etc. services to print the current page and e-mail, or services to jump to other parts of the current page.
2. Little information board, like the weather reports and stock board / council reporting market, etc.
3. Interactive noise service for users to put their needs. These include inputs based services such as search bars, sign forms, subscription forms, etc., and the selection based services such

10. BLOCK SPLITTING

Figure 6: Part of a Web Page Taken as an Example

The Figure 6: is taken an example of a sample web page which consists of local noises such as images, multiple links, etc. and also the main content useful for mining.

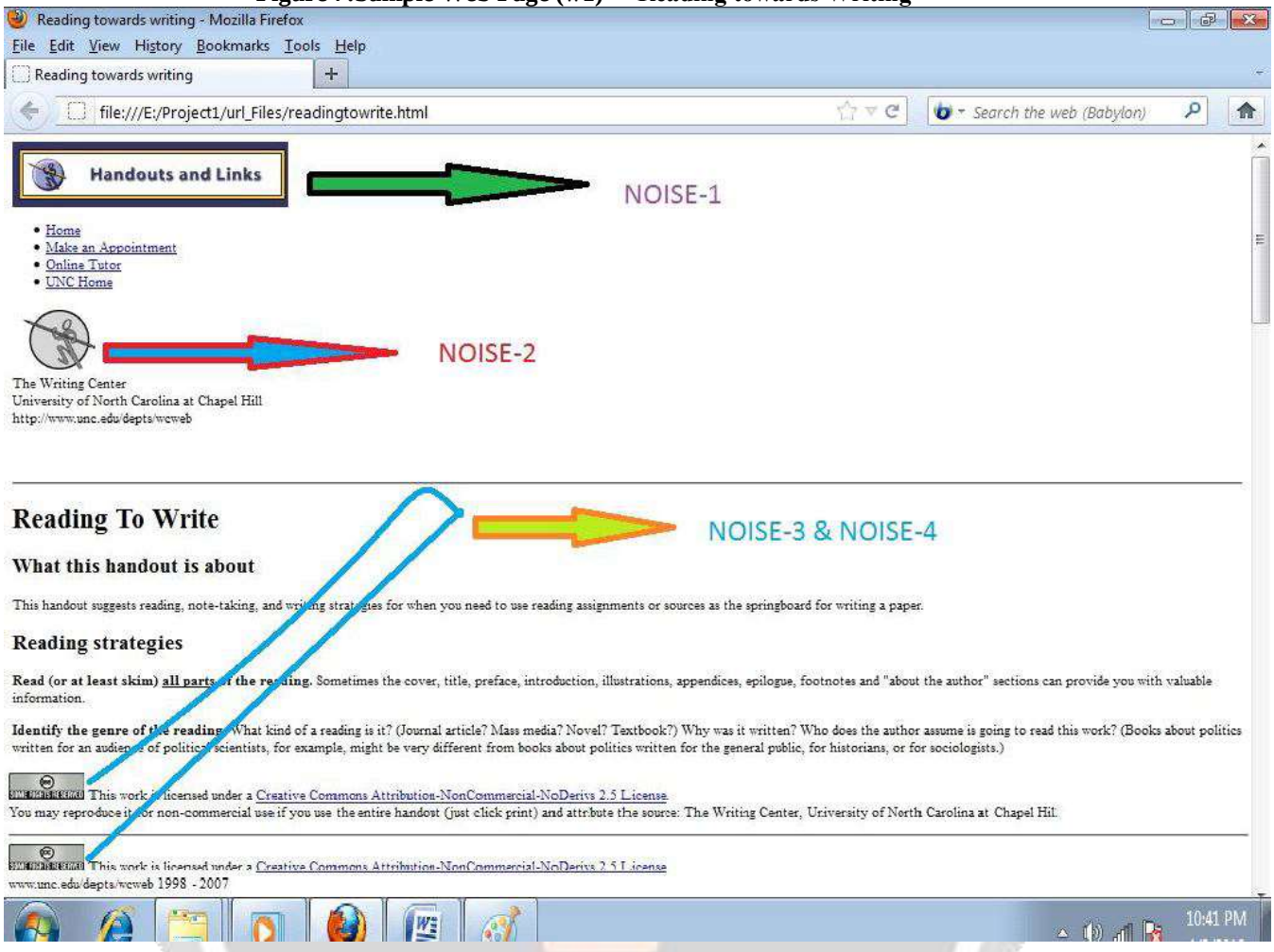
11.RESULT ANALYSIS

The results analysis focuses on experimentation and evaluation of the research work. Various real time scenarios are taken and applied to this proposed approach in the experimentation. Then the results are evaluated with existing and proposed method.

The minimum configuration required to simulate this research work is given as below:

Processor - Intel Pentium Core 2 Duo - 2.4 GHz
Mother Board - Intel
RAM - 2 GB DDR - II
HDD - 160 GB
Monitor - 17" LCD
Mouse - Optical Mouse
Keyboard - Standard Multimedia Keyboard

Figure 7:Sample Web Page (w1) - "Reading towards Writing"



the web page w1 shown in Figure 7 is taken; named "reading towards writing" and this web page contain images which are noises that need to be removed first, to separate the main contents from noises. On this web page w1, there are four images which are not required for further processing. Hence, only the content inside the HTML tag "div" is considered as the main content.

12.CONCLUSION

Unlike conventional data, Web pages typically contain a large number of information items that are not part of the main contents. Such information items like Banner Advertisement, Navigation Bars, and Copyright Notices and etc. which are irrelevant or incoherent to the main content of web pages are called Web page noise in this study. This research work first categorizes the web page noise in the World Wide Web then a web page cleaning approach is proposed, which detects and eliminates web page noise and improves web mining results. In this research work, two new approaches are proposed, the first approach is for eliminating web page local noise then block splitting using Simhash and important block values are calculated, the second approach is Sketching algorithm is used for selecting
important blocks in web pages . Every web page noise is categorized into fixed description noise, Web service noise and navigational guidance according to their functionalities and formats. Fixed description noise provides descriptive information about the host web site or page. Web service noise provides convenient and useful ways for managing web page content or to communicate between server and web users. Navigational guidance works as intermediate guidance or shortcut to other web pages in/out of the host web site. Navigation guidance includes directory guidance like a list of hyperlinks linking to crucial index/portalpages within a site and recommendation guidance like guidance suggests web.

13. REFERENCES

(1) A. K. Tripathy and A. K. Singh. 2004. An Efficient Method of Eliminating Noisy Information in Web Pages for Data Mining. In Proceedings of the Fourth International Conference on Computer and Information Technology

(2) C. Li, J. Dong, J. Chen. 2010. Extraction of informative blocks from Web pages based on VIPS. Journal of Computational Information Systems .

(3) D. Alassi, R. Alhajj. 2013. Effectiveness of template detection on noise reduction and websites summarization.

(4) D. Fernandes, E. Moura, B. Ribiero-Neto, A. Silva, M. Goncalves. 2007. Computing block importance for searching on Web sites.

(5) D. Gibson, K. Punera, A. Tomkins. 2005. The volume and evolution of Web page template, in: International World Wide Web Conference.

(6) F. Akthar, C. Hahne,. 2012. RapidMiner 5 Operator Reference. August 2012. www.rapid-i.com.

(7) G. Poonkuzhali , G.V. Uma, K. Sarukesi. 2010. Detection and Removal of redundant web content through rectangular and signed approach, International Journal of Engineering Science and Technology ,

(8) G. Poonkuzhali, K. Thiagarajan, K. Sarukesi and G.V. Uma. 2009. Signed Approach for Mining

(9) Web Content Outliers. World Academy of Science, Engineering and Technology,

(10) L. Yi, B. Liu, X. Li. 2003. Eliminating Noisy Information in Web Pages for Data Mining, SIGKDD .

(11) L. Yi and B. Liu. 2003. Web Page Cleaning for Web Mining Through Feature Weighting. In Proceedings of the 18th International Joint Conference on Artificial Intelligence.

(12) M. Agyemang, K. Barker and R. S. Alhajj. 2005. Mining Web Content Outliers using Structure Oriented Weighting Techniques and N-Grams. In Proceedings of the ACM Annual Symposium on Applied Computing, pp. 482-487, New Mexico.

(13) P. Sivakumar, R. M. S Parvathi. 2011. An Efficient Approach of Noise Removal from Web Page for Effectual Web Content Mining. European Journal of Scientific Research. http://www.eurojournals.com/ejsr.htm

(14) S. Akbar, L. Slaughter, Ø. Nytrø. 2010.Extracting main content-blocks from blog posts. In: Proceedings of the International Conference on Knowledge Discovery and Information Retrie.

(15) S. Debnath, P. Mitra, N. Pal, and C. Lee Giles. 2005. Automatic Identification of Informative Sections of Web Pages, IEEE Transactions on knowledge and data engineering, vol. 17, no. 9.

(16) S. Gupta, G.E. Kaiser. 2005. Automating Content Extraction of HTML Documents. World Wide Web: Internet and Web Information Systems.

(17) Z. Cheng-li and Y. Dong-yun. 2004. A Method of Eliminating Noises in Web Pages by Style Tree Model and Its Applications.

(18) J. Kang and J. Choi. 2007. Detecting Informative Web Page Blocks for Efficient Information Extraction Using Visual Block Segmentation. International Symposium on Information Technology Convergence