Impact Factor (SJIF): 3.632

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 3, Issue 5, May-2016

A Survey on Web Mining with Noise Removal Technique

Patel Niral

Information Technology, Parul Institute of Engineering and Technology

Abstract *— Today, world wild web is comprised different variety of web page which is source of large amount information. Information present in web page is may be informative and non-informative (Noise information). Here non-informative contain such as copyrights, navigation bars, advertise banner etc. decreased the performance of web mining task. In web mining web page noise elimination is important research area for elimination noise information form web page and help to improve the performance of web mining task. The main aim of this paper is study about web mining area and discussed varies method are used for this purposed.*

Keywords- Web Mining, Web Content Mining, Noise.

I. INTRODUCTION

The WWW is sort form of World Wide Web and also known as Web. Web is a world largest information source which users can access this information through internet. Information available on web in form of web page. Web page is web document writing using markup language like HTML. The content of web page is image, text data, audio, video and etc. Last year survey 4.5 billion web page is available on web. But dynamic nature and unstructured, semi-structure information available in web page is make challenging task for get useful Information from web page. Web mining help to discover such information from web.

II. WEB MINING

With the growth of the data, data mining are become important and popular. Web mining is application of data mining help to extracting such information from web. It is also useful for web page segmentation, clustering, information retrieval, etc. Web mining is create new knowledge from relevant data, summarizing information, understanding user behavior and etc. The objects of Web mining include: sever logs, Web pages, Web link structures, on-line market data, and other data [9]

- Web logs: once folks browse net server, sever will manufacture 3 types of log documents: sever logs, error logs, and cookie logs. Through analyzing these log documents we will mine accessing data.
- On-line market data: used as storing e-commerce information in e-commerce sites.
- Web pages: Most of existing web mining ways are used in web content of according with hypertext mark-up language commonplace.
- Web link structures: the net pages are all connected by hyperlinks, in which there's vital mining data. So net hyperlinks are terribly authoritative resources.
- Other information: main are composed of user registrations that will facilitate mine higher.

III. WEB MINING CATEGORY

Web mining mainly categorized in to three category [1] is show in Figure 1. Web mining task can be categorized based on data used for mining purposed. Web mining task is following:

3.1 Web Structure Mining

The process of discovering structures information from the web documents are called as web structure mining. This mining can be performed either document level or hyperlink level. The hyperlinks provide clear navigation and point to the pages. This is used to retrieve the useful information in the form of structure. Hyperlink analysis can be done based on knowledge models, scope and properties of analysis and

types of algorithms. The methods that are done in the web usage mining are Data cleaning, Transaction identification, Data integration, Transformation, Pattern Discovery, Pattern Analysis.

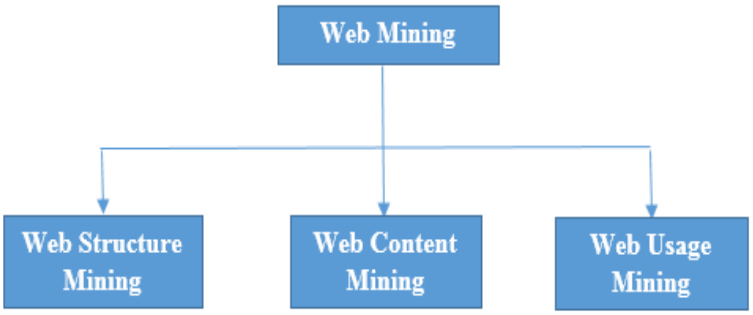

Figure 1: Web Mining Category

3.2 Web Content Mining

Web content mining is used for content or information extraction from the web page. Web page content such as image, text data and structure records like lists or tables etc. Most of the web content mining application measures similarity between two or more document. Web content mining contains two parts: searching result mining, and html Web page mining [7]

3.2.1 Searching result mining

Automatic classification in documents using searching engine Search engine can index a mass of disordered data on Web. For example, firstly, Web crawlers down load Web pages from Web sites. Secondly, searching engine extracts describable index information from these Web pages to store them with URL into searching engine base. Thirdly using data mining methods we automatically class them into usable Web page classification system organized by hyperlink structure.

3.2.2 Web Text Mining

Text mining is a comprehensive technique. It relates to data mining, computer language, information searching. Text mining uses data mining techniques in text sets to find out connotative knowledge. Its object type is not only structural data but also semi-structural data or non-structural data. The mining results are not only gene situation of one text document but also classification and clustering of text sets. The basic architecture of web text mining is show in figure 2.

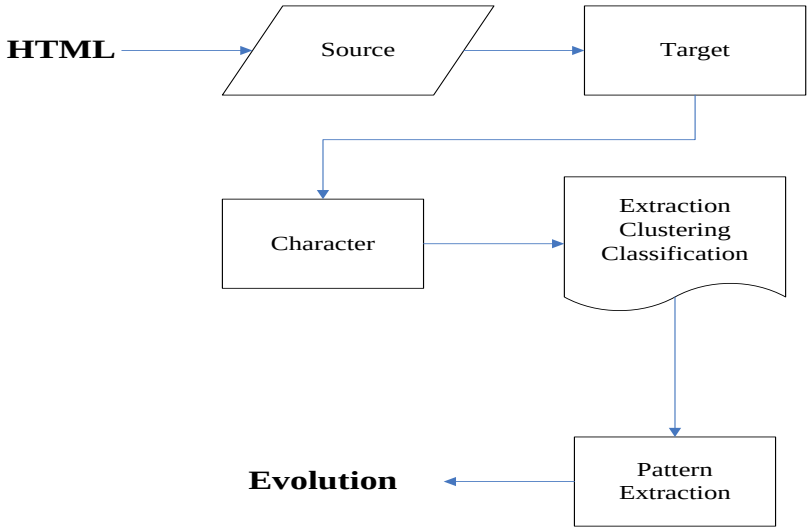

Figure 2: Architecture of Web Text Mining

3.3 Web Usage Mining

Main aim of web usage mining is help understanding user behaviors interacting with web. Web usage mining is used for finding useful usage patterns from web data. User profiles, user session, cookies, mouse click, bookmark, user queries and scrolls etc. data are used for finding useful patterns from web data which make more interesting for web based application.

4. NOISE

In web page large amount of non-information is available together with useful information known as noise. Noise available in web page like copyrights, advertisement, navigation bar, web page decoration etc. This noise information is degrade the performance of the web mining task. This noise information is hamper to web mining task like web page classification, clustering and information extraction or retrieve form web page. Here different approaches are used for removing such noise from web page.

Figure 3: Example of Web Page with Noise

Figure 3 gives a sample page from PC-Magazine [8]. This page contains an evaluation report of Samsung ML-1430 printer. The main content (segment 3 in Figure 1) only occupies 1/3 of the original Web page, and the rest of the page contains many advertisements, navigation links (e.g., segment 1 in Figure 1), magazine subscription forms, privacy statements, etc. This content is not related to main content of the web page or subject of the web page, consider as noise. This noise categorized in mainly two types [4] [6].

4.1 Local Noise

This also known as intra-page noise and it is together with main information within single page. Example of local noise is advertisement and banner, presentation style foe web page, etc.

4.2 Global Noise

This noise also known as inter-page noise and not within the single page. Example of global noise like old web site page, mirror web site, etc.

5. APPROACH FOR NOISE ELIMINATION

5.1 In this approach noise elimination is done based on Structural Analysis and Regular Expressions [2]. This method is eliminated noise using analysis of layout and content of the web page. In first stage web page filtering is done based on regular expression and after that structural analysis is done for elimination of remaining noise. In filtering phase HTML content in web page is categorized in two types: 1) Positive tag and 2) Negative tag, here positive tag comprise useful information and negative tag comprise noise information. In structural analysis phase first extracting body level tags from filter page and after comparing this page. In comparison phase similar style and content are consider as noise and different content and style consider as useful information. This approach is eliminated local noise from web page.

5.2 This approach, a noise elimination has been proposed based on regular expression using with SST [3]. In this approach first negative tag are removed using regular expression, here negative tag is consider as noise information HTML tag. But filtering stage not removed all the noise from page, so remaining noise are removed using Site Style Tree (SST) method. SST is created by using DOM tree of the particular web page. After created SST, entropy based calculation is used for identify the noise from SST. Limitation of this approach is not remove all the global noise from web page.

5.3 In this proposed method, first extracting HTML page code from web page and after that it convert into the DOM tree. Next phase Vision based page segmentation technique are used for page segmentation and forming page into content, presentation and datasets. After that calculate similarity between multiple page using FP-growth algorithm to determine informative and non-informative information. Here importance of the tag is greater than predefined threshold then is consider as useful information other then consider as noise information[4].

5.4 In this method, noise elimination is done using featured DOM tree [5]. This method is categorized into three stage. First featuring phase web page processing method are used like html tag remove etc. and web page filtering is done. After second phase is known modeling phase, here web page is converted into feature DOM tree and final phase is pruning phase, here identifying noise from feature DOM tree and Minimum Weight Overlapping (MWO) are used for similarity verification and mark noise content using some predefined threshold value. Finally noise block are remove from DOM tree.

6. CONCLUSION

Today, growth of the information on web page is difficult to extracting useful information from web page because of noise is present in web page and this noise misguided to the user. So noise elimination is important for the improving performance of web mining task. In this paper different approaches are discussed for elimination from the web page. This different approaches is help to remove local noise from the web page and improved the performance of web mining task.

7. REFERENCE

[1] Govind Murari Upadhyay, Kanika Dhingra, "Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering (IJARCSSE), 2013.

[2] Amit Dutta, Sudipta Paria, Tanmoy Golui and Dipak K. Kole, "Structural Analysis and Regular Expressions based Noise Elimination from Web Pages for Web Content Mining", IEEE, 2014.

[3] Tanmoy Golui, Sudipta Paria, Amit Dutta1, and Dipak Kumar Kole, "Noise Elimination from Web Page Based on Regular Expressions for Web Content Mining", Springer, 2014.

[4] Ms. Shalaka B. Patil, Prof. Rushali A. Deshmukh, "Enhancing Content Extraction from Multiple Web Pages by Noise Reduction", International Journal of Scientific & Engineering Research, 2015.

[5] Shine N. Das, Pramod K. Vijayaraghavan, Midhun Mathew, "Eliminating Noisy Information in Web Page using featured DOM tree", International Journal of Applied Information Systems (IJAIS), Volume 2– No.2, May 2012.

[6] Surabhi Lingwal, "Noise Reduction and Content Retrieval from Web pages", International Journal of Computer Applications", Volume 73-No.4, July 2013.

[7] Lizhen Liu, Junjie Chen, Hantao Song, "The Research of Web Mining, IEEE-2001, pg.2333-2237.

[8] Lan yi, Bing Liu, Xiaoli Li, "Eliminating Noisy Information in Web Pages for Data Mining", ACM-2003.

[9] Bing Liu, "Web Data Mining- Exploring hyperlink, Content and usage data", Springer- Verlag Berlin Heidelberg, 2007.