A Review on Noise Removal from Web pages for Web Content Mining

Mrs Madhushree B
L. J. Institute of Engineering and Technology
Mentoring Masters in Computer Engineering
Gujarat Technological University
Ahmedabad-Gujarat-India.
Email: bmadhushree.lj@gmail.com
Telephone:(+91)7405285827

Yogita K Patel
L. J. Institute of Engineering and Technology
Student of Masters in Computer Engineering
Gujarat Technological University
Ahmedabad-Gujarat–India.
Email: yogitapatel.ks@gmail.com
Telephone: (+91)9409437004

Abstract—Today internet has made the life of human depen-dent on it. Almost everything and anything can be used for discovering useful knowledge or information from the web page. A web page typically contain large amount of information that is not part of the main contents of the pages. E.g. Banner ads, navigation bars, copy right, privacy notices, advertisements which are not related to the main content. So noisy data affect the performance of web content mining. The main objectives of this area is removing such irrelevant information in web pages. The main purpose of this paper is to review and discuss the research work that has been done in this area and identifying the issues in this area.

Index Terms—Web Mining, WWW, Web Pages, DOM Tree, Noise.

I. Introduction

Data mining is define as extracting the information from the large set of data. The massive information available on the World Wide Web has search for data. A web mining has important task to discover useful knowledge or extract information from the web. Web mining can be divided in to three categories: web structure mining, web usage mining and web content mining. Web structure mining is the process of discovering hyperlink and document structure information from the web. Web usage mining is the application of data mining techniques for finding interesting and useful usage patterns from web data which makes it more demanding for web based applications. Web content mining is the process of extracting useful information from the contents of web documents.

In the World Wide Web Noise on the web pages are not the part of main content and irrelevant information in web pages can affect web mining task.Noises present in web pages can be grouped into two categories according to their granularities. Local or Intra-page noises: These are noisy information blocks are jumbled with the main contents within a single web page. Local noises include banner and advertisements, navigational guides, some pictures for decoration, etc. Global or Inter-page noises: These are noises on the web, which are usually no smaller than single web pages. Global noises include mirror web sites, legal or illegal duplicated web pages, some old versioned web pages, etc.

There are several methods are available to segment web pages into blocks. In the DOM based segmentation approach an HTML document is represented as a DOM tree. DOM tree is generally provides a useful structure and better representa-tion for a web page. [1]

A. APPROACHES OF NOISE REMOVAL FROM WEB CONTENT MINING

Web Content mining has following approaches to mine data: unstructured mining, structured mining, and semi-structured mining. [12]

B. Unstructured Data Mining

Text document is the form of unstructured data. Most of the data that is available on web is unstructured data. The research of applying data mining techniques to unstructured data is known as knowledge discovery in texts. [12]

1) Information Extraction : To extract information from unstructured data that is present on web pattern matching is used. It traces the keywords and phrases and then finds out the connection of keywords within text. When large volume of text is there then the technique is very useful. Information extraction transforms unstructured text to more structured form. First, from extracted data the information is mined, then using different types of rules, the missed out information is found. Information extraction making incorrect predictions on data is discarded . [12]

2) Topic Tracking : This technique checks the documents viewed by the user and studies the user profile. It predicts the documents related to users interest. The topic tracking applied by yahoo, user give a keyword and if anything related to keyword pops then the user is informed about that. This technique can be applied by many fields. The two fields where it is used is medical field and education field. In medical field doctors easily come to know about the latest treatments. In education field it is used to find out the latest reference for research related work. [12]

3) Summarization : The technique is used to reduce the length of the document by maintaining the important points. It helps the user to decide whether to read the topic or not. The summarization technique uses two methods that is the extractive method and the abstractive method. The extractive method selects a subset of phrases, sentences and words to form the summary from the original text. The abstractive method builds an internal semantic representation and then uses natural language generation technique to create the sum-mary. [12]

4) Categorization : This technique identifies the main theme by placing the documents in a predefined set of group. The technique counts the number of words in the document and this decides the main topic. According to the topic the rank is given to the document. The documents with majority contents on particular topic are given first rank. This technique helps in providing customer support to the industries and business. [12]

5) Clustering : The technique is used to group similar documents. In this grouping of documents is not done on the basis of predefined topics. It is done on fly basis. Some documents may appear in different group. As a result useful documents are not omitted from search results. This technique helps user to select the topic of interest. [12]

6) Information Visualization : Visualization utilizes feature extraction and key term indexing. Documents having similarity are found out through visualization. Large textual materials are represented as visual maps or hierarchy where browsing facility is allowed. It helps in visually analyzing the content. The user can interact by scaling, zooming and creating sub maps of the graphs. [12]

C. Structured Data Mining

The techniques are used to extract structured data from web pages. Data in the form of list, tables and tree is structured data. The structured data is easy to extract as compared to unstructured data. [12]

1) Web Crawler : Crawlers are computer programs which traverse the hypertext structure in web. Web crawlers can be used by anyone to collect information from the web. Search engines use crawlers are two types of crawlers. They are internal and external web crawler. Internal web crawler crawls through internal pages of the website and the external crawler crawls through unknown websites. [12]

2) Page Content Mining : Page content mining is a tech-nique that is used to extract structured data which works on the pages that are ranked by the traditional search engines. The pages are classified by comparing the page content rank. [12]

3) Wrapper Generation : The information is provided by the wrapper generator on the capability of sources. Web pages are ranked by traditional search engines. By using the page rank value the web pages are retrieved according to the query. [12]

D. Semi-Structured Data Mining

1) Object Exchange Model : The relevant information is extracted from semi-structured and is collected in a group of useful information and is then stored in Object Exchange Model (OEM). This helps the user to accurately understand the structure of the information that is available on web. [12]

2) Top down Extraction : This technique helps in extracting complex objects from a rich web sources and decompose them into less complex objects until atomic objects have been extracted . [12]

3) Web Data Extraction Language : This technique helps in converting web data to structured data and then delivers this data to end users . [12]

II. RELATED WORK

There are many Researchers have worked in this area for extracting main content and removing noisy data from web pages. Most of have focused on detecting main content and informative blocks in web pages. Although cleaning noisy data is an important task, relatively list of the work has been done in this field such as,

In Structural analysis and Regular Expressions based Noise Elimination from Web pages for Web content Mining Amit Dutta et al. proposed the noise elimination method that uses tag based filtering followed by structural analysis of the web page. The system are uses two phases: first Filtering based on Regular Expression and second Structural analysis of the crawled web pages after filtering.This paper focus on detecting and eliminating local or intrapage noises from web pages. [1]

In Noise Elimination from Web page Based on Regu-lar Expressions for Web content mining Amit Dutta,Dipak Kole,Tanmoy Golui et al. proposed approach to detect the global noises or inter-pages from web pages. The proposed technique consists of two phases. In the first phase, filtering method based on regular expression is used on web pages to remove noisy HTML tags The filtered document then undergoes to second phase where an entropy based measured is used for removing further noise. [2]

In Mining Contents in web page using Cosine Similarity Swe Nyein et al. propose an approach to extract the main content from the web documents. The algorithm based on con-tent structure tree (CST).firstly, proposed system use HTML parser construct DOM tree from construct construct DOM tree from content structure tree which can easily separate the main content blocks from the other blocks. The proposed system introduce cosine similarity measure to which part of tree represent less important and which part of tree represent the more important of the page. [3]

In An Efficient Method of Eliminating noisy information in web pages for data mining Tripathy, Singh et al. proposed cleaning technique that is based on the analysis of both the layouts and the actual contents of the Web pages in a given Web site. Thus, in first task of proposed technique is to find a suitable data structure to represent both the presentation styles and the actual contents of the Web pages in the site. They propose a Pattern Tree (ST) to capture those frequent

presentation styles and actual contents of the Web site. The site pattern tree (SPT) provides us with rich information for analyzing both the structures and the contents of the Web pages and used an information based measure to evaluate the importance of element nodes in SPT so as to detect noises to clean a page from a site, they simply map the page to its SPT. [4]

In Noise Removing from web pages using Neural Network Thanda Htwe et al. propose the mechanisms to eliminate multiple noise patterns in Web pages to reduce irrelevant and redundancy data by applying Case-Based Reasoning technique to detect multiple noise patterns in current Web page and also present back propagation neural network algorithm for matching current noise with storing noise patterns for noise classification, and then they remove this noise pattern in current page for content extraction. [5]

In Elimination of Noisy Information from web page using DOM and Ant Colony Optimization Shaikh Sakina Banu et al. Proposed method to eliminate noisy information from web page using DOM tree approach and Ant Colony Optimization to improve the efficiency of mining and also apply neural network algorithm to detect noisy data. [6]

In Extraction of web news from web pages using a ternary tree approach Debina Laishram et al. proposed a new approach to extraction of news from multiple news web sites. The proposed method using ternary tree approach measure which expands when series of common tags are found in the web pages. [7]

In LBDA: A Novel Framework for Extracting Content from web pages C Deepa et al. proposed approach extracts the main content from the web page and remove the irrelevant information like header, footer contents, navigation bars, advertisements and other noisy images. The proposed methodology uses the following techniques: tag tree parsing to get the analysis structure, block acquiring page segmentation method to remove unwanted tags, and data extraction to retrieve the necessary contents. It can eliminate noise and extract the main content blocks from web page. [8]

In Content Extraction from web pages Based on Chinese Punctuation Number Mingqiu Song et al. proposed approach which can discover web page content according to the Chinese punctuation number. It can eliminate noise and extract main content blocks from web page. This approach is accurate and suitable for most Chinese web sites. [9]

In Automatic News Extraction System for Indian Online News Papers Dipali B, Sachin Deshmukh et al.proposed approach for the Indian online newspaper which extract contents from news web databases. The system first browse Web pages as per the input URL given by user and Next generate a DOM tree of the news Web page data. And at last, we not only identify and extract valuable news from the Indian news web pages but also remove noisy data. This paper proposed the novel approach for extract data from online Indian newspapers written in the many popular languages such as Marathi, Hindi, Tamil, Gujarati, Kannada, Oriya, Telugu, Punjabi, etc. [10]

In Using Visual Clues Concept for Extracting Main Data from Deep Web Pages Satish J.Pusdekar et al.proposed approach extracting main data from web pages. This paper vision-based approach is web page programming-language-independent approach is proposed. This approach utilizes the visual features of the web pages to extract data from deep web pages including data record extraction and data item extraction. [11]

III. Conclusion

WWW is a source of information where large amount of data is stored. The information present in the local and global noise. The purpose of noise elimination is to improve web mining. Extracting useful information from the web pages is very complex task. Accurate and effective method to find more relevant document from the web pages. Organizing and removing noise from web pages will get better on correctness of search results as well as explore results. We saw various approach which enhance the capacity of information extraction and remove the noise from the web pages. In this paper most of the approach based on the DOM tree. The DOM tree approach is always feasible as it converts the complex page into simplified form.so in the future extended work on the DOM tree approach and will get accurate results.

References

[1] 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 978-1-4799-3080-7/14,PP.1445-1451,2014.

[2] Dutta Amit, Paria Sudipta, Golui Tanmoy and Kole Dipak, Noise Elimination from Web Page based on Regular Expressions for Web Content Mining Proceedings of the Second International Conference on Advanced Computing, Networking and Informatics (ICACNI-2014), Volume 27, pp 545-554, June 2014

[3] Swe Swe Nyein,Mining Contents Web Page Using Cosine Similarity, IEEE 978-1-61284-840-2/11,pp 472-475,2011.

[4] A. K. Tripathy, A. K. Singh 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 (CIT04) 0-7695-2216-5/04 2004 IEEE

[5] Thanda Htwe, Khin Haymar Saw Hla Noise Removing from Web Pages Using Neural Network 2010 ,978-1-4244-5586-7/10/26.00 C 2010 IEEE Volume 1.

[6] Shaikh Sakina Banu, Hitesh Kumar Bhatia Elimination of Noisy In-formation from Web Page using DOM and Ant Colony Optimization International Journal of Engineering Research And Technology,Vol.3 Issue 2,PP 1227-1231,feb-2014.

[7] Debina Laishram, Merin Sebastian Extraction of web news from web pages using a ternary tree approach Second International Conference on Advances in Computing and Communication Engineering ICACCE, 978-1-4799-1734-1/15 2015 IEEE,PP 628-633, 2013.

[8] C Deepa, LBDA: A Novel Framework for extracting content from web pages International Conference on Advanced Computing and Communication Systems (ICACCS -2013), IEEE 978-1-4799-3506-2/13 2013 IEEE

[9] Mingqiu Song, Xintao Wu Content Extraction from Web Pages Based on Chinese Punctuation Number 1-4244-1312-5/07/ 2007 IEEE,PP 5573-5575

[10] Vivek D. Mohod, Dipali B. Gaikwad, Sachin N. Deshmukh , Automatic News Extraction System for Indian Online News Papers 978-1-4799-6896-1/14/ 2014 IEEE

[11] Satish J. Pusdekar, Using Visual Clues Concept for Extracting Main Data from Deep Web Pages, International Conference on Electronic Systems, Signal Processing and Computing Technologies 978-1-4799-2102-7/14 2014 IEEE DOI 10.1109/ICESC.2014.39, PP 190-193

[12] Shipra Saini, Hari Mohan Pandey, Review on Web Content Mining Tech-niques, International Journal of Computer Applications (0975 8887) Volume 118 No. 18, May 2015.