RESEARCH ARTICLE OPEN ACCESS

Web Noise Data Handling Using NFD and Data Mining Algorithm

SekharBabu.Boddu, Prof.RakajasekharaRao.Kurra
Asst.professor ,KLUniversity ,
Director,Usharama Engg College
AP - India

ABSTRACT

The huge development of the web in the last few years, it can adopt the society as part of it, throughout the world. The web contains different types of the data and it is represented in different forms, web contains some of the characteristics. Web data handling is a challenging task to find the knowledge from the web as per the user requirements.

Keywords:- Web Content Mining , KNN,NFD

I. INTRODUCTION

Now a day's availability of information linked with web. User can find the data from the web and the presentation of the data linked to the web is still unstoppable. Web is not confined with space limitations and it is spread worldwide. Anyone can accesses the data from anywhere from the web. The data is represented as in different formats like text, tagged, images, email, tables, audio, videos and multimedia data types. Web is support and presenting the data in to various types like structured data, unstructured data, semi structured data. Web contains mixed types of data and these are presented on web pages. The possibility of duplication of the data is available on the web, because many web pages are represented the same data in the different formats. This is the challenging research problem to get solution. The availability of the web data is also linked structure format. There is hyperlinks concept

It can be travel from on page to other page of the same website or any other. The link analysis will do the process to get the information from the multiple pages. In some situations data can be available from the different web pages at that scenario hyper link analysis will do the job. Many of the users believe that hyperlink analysis will execute the task efficiently and information fetching can be done properly. Web is data repository and it contains all types of the data, there is still question about the reliability, accuracy and quality of the data, the only one issue which gives the answer, that is the web information without any noisy. Actually the noise comes from the different issues like main content of the data , navigation panels ,copy right forms ,email ids ,number formatted data, advertisements, privacy policies etc. majority of the application need unstructured data like text data remaining they consider as noisy. Noisy data should be handled otherwise user will not get the actual information on the web .web mining contains different techniques are useful to handle the noisy data and improve the user satisfaction levels on the web. The web is not only a data repository and it is also about business and commerce. Majority of the online business processing can be performed by the users through commerce websites only. For this type of applications and its services, the web has to be ready for providing necessary resources.

The web is not a static and it is dynamic, information updated on the web is continuously. In this point of view, we need to update our requirements as per the user's request. Many applications are continuously updated so that we need to handle our data without having any noisy. For Web mining, data collection can be a substantial task, especially for Web structure and content mining, which involves crawling a large number of target Web pages. Web mining are divide into three types: Web structure mining, Web content mining and Web usage mining. Web mining tasks can be categorized into three types: Web structure mining, Web content mining and Web usage mining.

Web structure mining: Web structure mining discovers useful knowledge from hyperlinks (or links for short), which represent the structure of the Web. For example, from the links, we can discover important Web pages, which is a key technology used in search engines. We can

also discover communities of users who share common interests. Traditional data mining does not perform such tasks because there is usually no link structure in a relational table.

Web content mining: Web content mining extracts or mines useful information or knowledge from Web page contents. For example, we can automatically classify and cluster Web pages according to their topics. These tasks are similar to those in traditional data mining. However, we can also discover patterns in Web pages to extract useful data such as descriptions of products, postings of forums, etc., for many purposes. Furthermore, we can mine customer reviews and forum postings to discover consumer opinions. These are not traditional data mining tasks.

Web usage mining: Web usage mining refers to the discovery of user access patterns from Web usage logs, which record every click made by each user. Web usage mining applies many data mining algorithms. One of the key issues in Web usage mining is the pre-processing of click stream data in usage logs in order to produce the right data for mining.

Web Crawling

Web crawlers, also known as spiders or robots, are programs that automatically download Web pages. Since information on the Web is scattered among billions of pages served by millions of servers around the globe, users who browse the Web can follow hyperlinks to access information, virtually moving from one page to the next. A crawler can visit many sites to collect information that can be analyzed and mined in a central location, either online (as it is downloaded) or off-line (after it is stored). Were the Web a static collection of pages, we would have little long term use for crawling. Once all the pages are fetched and saved in a repository, we are done. However, the Web is a dynamic entity evolving at rapid rates. Hence there is a continuous need for crawlers to help applications stay current as pages and links are added, deleted, moved or modified. There are many applications for Web crawlers. One is business intelligence, whereby organizations collect information about their competitors and potential collaborators. Another use is to monitor Web sites and pages of interest, so that a user or community can be notified when new information appears in certain places. There are also malicious applications of crawlers, for example, that harvest email addresses to be used by spammers or collect personal information to be used in phishing and other identity theft attacks. The most widespread use of crawlers is, however, in support of search engines. In fact, crawlers are the main consumers of Internet bandwidth. They collect pages for search engines to build their indexes. Well known search engines such as Google, Yahoo! and MSN run very efficient universal crawlers designed to gather all pages irrespective of their content. Other crawlers, sometimes called preferential crawlers, are more targeted. They attempt to download only pages of certain types or topics.

II. WEB CONTENT MINING

Due to the heterogeneity and the lack of structure of web data, automated discovery of targeted or unexpected knowledge information still present in many challenging research problems. Moreover, the semi-structured and unstructured nature of web data creates the need for web content mining. Web content mining is the process of identifying user specific data from Text, Image, and Audio or Video data already available on the web. This process is alternatively called as web text mining, since text content is the most widely researched subjects on the World Wide Web. The technologies that are generally used in web content mining are Information retrieval and Natural language processing. Web Structure mining is another process of using graph theory to analyze the node and connection structure of a web site. Depending upon the type of web structural data, web structure mining has been divided into two fields. The first one is extracting patterns from hyperlinks on the web. The other one is mining the document structure. This involves using the tree-like structure to analyze and describe the HTML or XML tags within the web page. Web usage mining is to identify user access patterns from Web usage logs. Here we discuss more on web content mining. Web content mining identifies the useful information from the web contents/data/documents. However, such data in its broader form has to be further narrowed down to useful information. Before narrow down information it is obvious to know that what web content mining is and how web content mining differs from data mining. The web content data consists of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents. The two main approaches in WCM are (1) Unstructured Text Mining Approach and (2) Semi-Structured and Structured mining approach.

Unstructured Data Mining (Text Mining)

Web content data is much of unstructured text data. The research on unstructured text in data mining includes Knowledge Discovery in Texts (KDT), or Text Data Mining, which also called as Text Mining. Hence one could consider a text mining as an instance of web content mining. To provide effective exploitable results, pre-processing steps for any structured data is done by means of information extraction, text categorization, or applying NLP techniques.

III. SEMI-STRUCTURED AND STRUCTURED DATA MINING

Structured data on the web are often very important as they represent their host pages, due to this reason it is important and popular. Structured data is also easier to extract when compared to unstructured texts. Semi-structured data is a point of convergence for the web and database communities: the former deals with documents, the latter with data. The form of that data is evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real-world objects like books, papers, movies, etc., without sending the application writer into contortions. Emergent representations for semi-structured data (such as XML) are variations on the Object Exchange Model (OEM). In OEM, data are in the form of atomic or compound objects: atomic objects may be integers or Strings; compound objects refer to other objects through labelled edges. HTML is a special case of such 'intra-document' structure.

IV. WEB DATA EXTRACTION

Effectively identifying and filtering these noises in web pages is one crucial technique to improve the search results. Noise filtering can lead to increased search speed, simplified tag structure and reduced size of intermediary files, and hence reduced temporal and spatial complexity of search.

As input of web pages in HTML format are taken. We impose these Requirements on an output of proposed system.

1. Code removal: An output should not contain any code (a script, a style sheet . . .).

2. Boilerplate removal: Any extraneous item as for example navigation, menu, list of links, footers, copyright notices, advertisements, etc., should be removed too.

3. Structural annotation: The output text should preserve some structural information of original document. Each list item should start with <l>, each heading with <h> and each paragraph with <p>. Additional tag <doc> can be used for separating various cleaned documents within one file.

Hyper text mark up language (HTML) defines tags to describe the web page layouts [2]. One popular method to represent web page structure is to build the tag tree of the HTML document. There are different tools to build the tag tree, and DOM tree. A tag tree can be built according to the nested relations in the web pages. For the purpose of noise filtering, we classify the tags in the HTML specifications into following two function categories:

1) Tag for layout

A web page consists of several content blocks, which are formatted by container tags, such as <table>, <tr>, <td>, <p> and <div>, etc.

2) Tag for display features

Besides the tags for layout, some other tags are defined to describe the web page content in the HTML standard. For example, <b> denotes fonts display in boldface, <img> denotes a picture, <a> denotes a hyper link.

4.0 Proposed Method:

Proposed methodology:

$\boldsymbol{\emptyset_{p:}}$:- set of web pages

$\boldsymbol{\emptyset_{e:}}$:-extracted web pages

$\boldsymbol{\psi_F}$:- term frequency from $\boldsymbol{\emptyset_e}$

$\boldsymbol{\psi_a}$ **,** $\boldsymbol{\psi_x}$:-html tags and text data extracted

$\boldsymbol{\xi_D (\psi_x , \psi_2)}$:- it is a procedure to find out the keyword density

$\boldsymbol{\xi_{prom}}$ $(\boldsymbol{\psi_x , \psi_3})$:- it is procedure to find out the keyword prominence

$\boldsymbol{NL_p (\psi_y , \psi_x , \delta_t , \delta_m, \psi_u)}$**:-** it is a procedure to find out $\boldsymbol{\delta_t , \delta_m, \psi_u}$, stopping words ,

stemming words and noise data extracted.

$\boldsymbol{\emptyset}$:- set of indexed web pages.

$\boldsymbol{\beta_q}$:- Boolean Query

$\boldsymbol{\emptyset_F}$**:-** Term frequency

$\boldsymbol{\delta_v}$**:** it is a procedure to find out search volume

Tr: it is a procedure to find out position of the term.

$\boldsymbol{\lambda}$**:** term position

Algorithm: NFD

Input: choosing set of web documents that includes various sub web pages.

Output: generates noise less data from the web documents

Step: 1 Extract the webpage. The extracted web page will be the subset of web documents.

$$\emptyset_p \subset \emptyset_e$$

$$\emptyset_p \leftarrow \text{set of web pages}$$

$$\emptyset_p = \{\emptyset_1, \emptyset_2, \emptyset_3, \ldots\ldots \emptyset_n\}$$

Step: 2 Extracted Structured and Unstructured data with term frequencies

$$\emptyset_e \longrightarrow \psi_a , \psi_x$$

$$\psi_F(\psi_x , \psi_1)$$

Step: 3 Finding the keyword density and keyword prominence.

$$\xi_D (\psi_x , \psi_2)$$

$$\xi_{prom} (\psi_x , \psi_3)$$

Step: 4 find out noise data using text analyzer

$$NL_p (\psi_y , \psi_x , \delta_t , \delta_m, \psi_u)$$

Step: 5 extract the search engine indexed documents

$$\emptyset \subseteq \emptyset_p$$

$$\beta_q$$

$$\emptyset_F (\psi)$$

Step: 6 find out search volume of the each term

$$\delta_v (\emptyset_F , \psi)$$

Step: 7 find out the position of the term

$$Tr (\lambda , \kappa)$$

$$Tr (\lambda , \kappa) = \psi \equiv \kappa \in \delta_v (\emptyset_F , \psi).$$

V. RESULTS

Word Counts	Value
Total Words incl. Short Words and Stop Words:	1225
Total Unique Words incl. Short and Stop Words:	555
Total Short Words:	206
Total Unique Short Words:	54
Total Stop Words:	145
Total Unique Stop Words:	22
Total Unique Words excl. Short & Stop Words:	479
In Title:	2
In Domain Name:	1
In Path or File Name	0
In Meta Description:	1
In Meta Keywords:	1
In H1:	0/0
In H2:	0/0
In H3-H6:	66/21
In First Sentence:	0
In Bold or Strong:	103/33
In Italics:	0/0
In Image Alt or Title:	23/34
Total No. of Five Word Phrases:	10
Total No. of Four Word Phrases:	14
Total No. of Three Word Phrases:	26
Total No. of Two Word Phrases:	63
Total No. of Single Words:	134

Table 5.1

Keyword density: Is a measure, in percentage or exact number, of times the keyword appears on a page compared to the total number of words (Wikipedia Keyword Density, 2009). The recommended density is between 3-7% (bytestart.co.uk) or 2-8% (Michael Wong, 2009). Overdoing the density might get the search engine to penalize the web page because of "keyword spamming". If the keywords are relevant compared to the web pages content, a higher frequency might be allowed.

Keyword prominence: Measures how relevant the keyword is by relating it to the position within the HTML tag. That is; the earlier the keyword is used, the more prominent it is, and hence the more relevant the keyword seems with respect to the content of the web page (Michael Wong, 2009). We use the following definition of keyword prominence (ranks.nl): Prominence =(words in string –Sum of positions of keywords -1 /(repetitions of key words in a String))(100/words in String) . This definition puts more emphasis, measured in percentage, on the keywords in the beginning of a sentence, and less as the word is relegated down through the sentence. This definition is chosen because it will always reward a keyword even though they are in the end of a sentence, whereas some of the other definitions removed this reward quite early setting the prominence to zero. **Keyword proximity:** Is concerned with how far apart keywords are located.

The closer they are, the more relevant they seem for a search engine. Search engines use “stop words” to measure the proximity between keywords (Szulencki, 2008). Stop words are words that are filtered from text on a web site and are disregarded by Google’s crawler, theses are mostly adverbs. The more stop words used between keywords on a page the higher proximity. Keyword proximity is used to calculate an index, ranking the importance of the keywords; if there are stop words between the keywords, this reduces the importance of them. While there are some debate regarding the usefulness of the proximity measure (Henriksen, 2009), we do not include it in our analysis but are aware of its existence as we did not find any consistent way of calculating the proximity.

Table 5.2

Table 5.1 give details about the html tagged data extracted from the input as web page. It shows each and every tagged word count and its value.
Table 5.2 give details about the extracted useful information from the noisy data and retrieved the keywords, position and search volume.

KNN Classification:

Implement KNN (K-nearest neighbour classification algorithm on web data)

TP Rate	FP Rate	Precision	Recall	F-Measure
0.548	1	0.708	0.725	**0.798**
0.133	1	0.235	0.359	0.56
0	0	0	0	0.038
0.107	0.75	0.188	0.266	0.58
0.082	0.667	0.145	0.202	0.567
0	0	0	0	0.429
0	0	0	0	0.567
0.991	0.332	0.498	0.204	0.575
0	0	0	0	0.208
0	0	0	0	0.565
0	0	0	0	0.041
0	0	0	0	0.298
0	0	0	0	0.289
0.088	0.857	0.16	0.241	0.558
0	0	0	0	0.045
0.4	1	0.571	0.63	0.622
0	0	0	0	0.385
0	0	0	0	0.044
0	0	0	0	0.043
Weighted Avg	0.391	0.549	0.391	0.292

Table 5.3

- The Table 5.3 give details about KNN classification algorithm on web data. The data categorized on the basis of keywords retrieved by the **TP Rate**: rate of true positives (instances correctly classified as a given class)
- **FP Rate**: rate of false positives (instances falsely classified as a given class)
- **Precision**: proportion of instances that are truly of a class divided by the total instances classified as that class
- **Recall**: proportion of instances classified as a given class divided by the actual total in that class (equivalent to TP rate)
- **F-Measure**: A combined measure for precision and recall calculated as 2 * Precision * Recall / (Precision + Recall)

S.no	Keyword	Position	Search Volume
1	K1	10	40500
2	K2	70	6600
3	K3	12	2900
4	K4	57	1900
5	K5	68	1600
6	K6	0	1300
7	K7	86	1300
8	K8	0	1300
9	K9	1	1000
10	K10	80	1000
11	K11	0	1000
12	K12	72	1000
13	K13	90	880
14	K14	70	880
15	K15	50	880
16	K16	66	720
17	K17	40	720
18	K18	0	720
19	K19	0	720
20	K20	26	720

The highlighted F-measure values are high impact on the keywords and its categories.

VI. CONCLUSION

In this paper NFD algorithm implemented on web data which is extracted from the web extraction frame work. The output of the NFD is to handle the noisy data from the data. NFD is identify the useful data from the noisy data handled by the previous exusting system. NFD is do the process and filter the noisy data and get the useful from that. KNN classification algorithm will help you to identify the nearest neighbour the keywords according to their category.

REFERENCES

[1]. Hassan F. Eldirdiery, A. H. Ahmed," Detecting and Removing Noisy Data on Web document using Text Density Approach ",International Journal of Computer Applications (0975 – 8887) Volume 112 – No. 5, February 2015.

[2]. Ms. Shalaka B. Patil, Prof. Rushali A. Deshmukh," , Enhancing Content Extraction from Multiple Web Pages by Noise Reduction", International Journal of Scientific & Engineering Research, Volume 6, Issue 7, July-2015 ISSN 2229-5518

[3]. Rajni Sharma, Max Bhatia," Eliminating the Noise from Web Pages using Page Replacement Algorithm, International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3066-3068.

[4]. Hui Xiong, Member, IEEE, Gaurav Pandey, Michael Steinbach, Member, IEEE,and Vipin Kumar, Fellow, IEEE," Enhancing Data Analysis with Noise Removal", IEEE Transactions On Knowledge And Data Engineering.

[5]. Rekha Garhwal," Improving Privacy in Web Mining by eliminating Noisy data & Sessionization", International Journal of Latest Trends in Engineering and Technology (IJLTET).

[6].Erdinc,uzun,et.al," A hybrid approach for extracting informative content from web pages", Information Processing and Management: an International Journal archive Volume 49 Issue 4, July, 2013,Pages 928-944.

[7]. Shine N. Das,et.al." Eliminating Noisy Information in Web Pages using featured DOM tree, International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA Volume 2– No.2, May 2012.

[8]. Xin Qi □ and JianPeng Sun, Eliminating Noisy Information in Webpage through Heuristic Rules, 2011 International Conference on Computer Science and Information Technology (ICCSIT 2011*)*

[9]. Thanda Htwe," Cleaning Various Noise Patterns in Web Pages for Web Data Extraction", Iinternational Journal of Network and Mobile Technologies Issn 1832-6758 Electronic Version Vol 1 / Issue 2 / November 2010.

[10]. Byeong Ho Kang and Yang Sok Kim," Noise Elimination from the Web Documents by Using URL paths and Information Redundancy**".**

[11]. Thanda Htwe, Nan Saing Moon Kham," Extracting Data Region in Web Page by Removing Noise using DOM and Neural Network", 2011 3rd International Conference on Information and Financial Engineering.

[12] .Tieli Sun, Zhiying Li, Yanji Liu, Zhenghong Liu, "Algorithm Research for the Noise of Information Extraction Based Vision and DOM Tree", International Symposium on Intelligent Ubiquitous Computing and Education, pp 81-84, May 2009.

[13].Jinbeom Kang, Joongmin Choi, "Block classification of a web page by using a combination of multiple classifiers", Fourth International Conference on Networked Computing and Advanced Information Management, pp 290 -295, September 2008.

[14]. Lan Yi," Eliminating Noisy Information in Web Pages for Data Mining".2003.

[15]. Bassma, S., Alsulami, Maysoon, F., Abulkhair and Fathy E. Eassa, "Near Duplicate Document Detection Survey", International Journal of Computer Science & Communication Networks, Vol. 2, No. 2, pp. 147-151, 2011.

[16].Huda Yasin and Mohsin Mohammad Yasin, "Automated Multiple Related Documents Summarization via Jaccard's Coefficient", International Journal of Computer Applications, Vol. 13, No. 3, pp. 12-15, 2011.

[17]. Syed Mudhasir, Y., Deepika, J., Sendhilkumar, S., Mahalakshmi, G. S, "Near- Duplicates Detection and Elimination Based on Web Provenance for Effective Web Search", (IJIDCS) International Journal on Internet and Distributed Computing Systems, Vol. 1, No. 1-5, 2011.

[18]. Kanhaiya Lal & N.C.Mahanti ,"A Novel Data Mining Algorithm for Semantic Web Based Data

Cloud" International Journal of Computer Science and Security (IJCSS), Volume (4): Issue (2),Pg.160-175,2010.

[19].Ranjna Gupta, Neelam Duhan, Sharma, A. K. and Neha Aggarwal, "Query Based Duplicate Data Detection on WWW", International Journal on Computer Science and Engineering Vol. 02, No. 04, pp. 1395-1400, 2010.

[20] .Alpuente, M. and Romero, D. "A Tool for Computing the Visual imilarity of Web pages", Applications and the Internet (SAINT), pp. 45-51, 2010.

[21].PrasannaKumar, J. and Govindarajulu, P. "Duplicate and Near Duplicate Documents Detection: A Review", European Journal of Scientific Research, Vol. 32 No. 4, pp. 514-527, 2009.

[22]. Poonkuzhali, G., Thiagarajan, G. and Sarukesi, K. "Elimination of Redundant Links in Web Pages - Mathematical Approach", World Academy of Science, Engineering and Technology, No. 52, p. 562,2009.