Noise Reduction using Character Density Approach

Jincymol Joseph, J R Jeba

Abstract--- Web mining is an application of data mining to extract informative content from World Wide Web(WWW). It has become one of the most significant resources nowadays. It may contain informative as well as non-informative contents. Non-informative contents may be header, footer, advertisements, copyright information, etc. These are called noisy data. A user needs only main contents. Web mining methods are useful for removing noisy parts and extract main contents from a web page, The advantage of using web mining methods is reduced time. Also, it provides users the needed information. This paper describes various methods for eliminating non-informative content from the large volume of data present in World Wide Web.

Keywords- Noisy data, web mining, cluster, outlier

I. INTRODUCTION

In WWW, meaningful data is extracted using data mining application called web mining. Anyone can upload or download any information at any time. So, internet grows rapidly and it leads to continuous expansion of irrelevant, redundant, structured and unstructured data on the web. In current development of data, extraction is a difficult work. Many algorithms were developed for extracting core contents from web pages. Using web mining methods, the needed information can be accessed timely and efficient manner.

A web content extractor extracts the core contents by removing the noises such as header, footer, copyright information, advertisements, etc. In this algorithm, character_density of each tag is calculated and compares it with threshold value.

II. WEB MINING CATEGORIES

It has 3 classes: Web usage mining, Web structure mining and Web content mining.

A. Web content mining

It is the process of extraction of meaningful data from huge quantity of material available on the web pages. Web page consists of information in the form of text, audio, images, tables, video etc. Most of the data present on the web is structured, semi-structured or unstructured form. Data extraction from structured pages is easy when compared with semi-structured or unstructured pages. Web content mining follows two approaches.

a. Agent based approach

The objective of agent based approach is to find relevant and significant information available.

(a) Intelligent search agents- It searches for information automatically besides a particular query.
(b) Information filtering/categorizing agents- Filters the data present on web.
(c) Personalized web agents-IT discovers those documents which are any how related to the user profiles.

b. Database Approach

Database approach consists of databases which contain attributes, tables and schema. By using standard query, organize the semi-structured data present on the web pages into structured data.

B. Web structure mining

A web graph structure contains nodes and edges. Here nodes represent web pages and edges represent hyperlinks which connect two interrelated pages. Web structure mining uses graph theory for analyzing the node and link structure of a web site. Web structural mining can be classify on the basis of web structural data types:

1. Hyperlink analysis- It links the webpage from current page to some other location or page.
2. Document structure- It is the process of mining document structure which includes analysis of the tree-like structure of page structures to describe XML or HTML tag usage.

C. Web Usage Mining

A web consists of inter-related files on one or more servers. Web usage mining is used to find interesting and meaningful usage pattern from data on web to understand the requirements of web based applications. Web usage mining can be classified upon the type of usage data considered.

- Web server data: Web server data includes user logs, access time, page reference and IP address.
- Application server data: To build an e-commerce websites and applications is a feature of commercial application server. Create application server log is a key attribute or feature of application server data.
- Application level data: New type of event is defined and then logging of that event is turned on. [1]

Revised Manuscript Received on December 22, 2018.
Jincymol Joseph, Department of Computer Science, St.Pius X College Rajapuram, Kasargod,Kerala, India (E-mail-jincyjosek@gmail.com)
Dr. J R Jeba, Department of Computer Applications, Noorul Islam Centre for Higher Education, Kumaracoil, Tamil Nadu, India

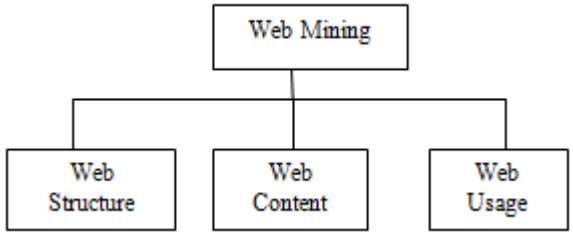

Fig 1:

This page covers enormous volume of undesirable material, named as noisy information. The header, footer, copyright, navigational panel and advertisements are known as noisy content. Noises can be classified as :

- Global noise: Global noises are redundant web pages on the internet. Global noises spread over websites such as duplicate page, Mirror sites and so on.
- Local noise: Local noises are also known as intra page noise. These are unimportant information in a web pages such as copy right, advertisements, footer, header, navigation bar and so on

Web content extraction is to identify the main content blocks by removing global and local noises from a web page [2].

It extracts valuable information by exploring the hyperlinks, audio, text, video, metadata, image and so on. It aids to know customer behavior and analyze the performance of a web site. It aids to develop companies indirectly. Web content mining estimates the hunt outcome of search engine and extracts the core contents from web pages. It takes more time manually. If analyzed information is in huge quantity, tough to discover out appropriate data [3].

Web sites may contain structured, unstructured, semi structured and multimedia data. Web content mining becomes complicated when it has to analyse these types of data. Unstructured data mining techniques are Information extraction, topic tracking, summarization, categorization, clustering and information visualization. Web crawler, wrapper generation and page content mining are structured data mining techniques. Semi- structured methods are OEM, Top down extraction and web data extraction language. Color histogram matching, multimedia miner and Shot Boundary Detection are multimedia data mining methods.

Categorization classified documents into various classes. It aids to recognize the key subject of a document pool.

Cluster is the collection of associated documents. The clustering is process of clustering documents depend on the similarity measure. Some common clustering algorithms available are hierarchical, binary relational and fuzzy. Similarity measure is the most important factor in clustering algorithms.

Summarization reduces the amount of text in a document. For this, the user defined a parameters list. [4]

III. WEB MINING METHODS

Some methods are developed for web content mining. Each and every method has advantages as well as disadvantages. Some of them are:

a. Hybrid Approach

It includes rule generation and automatic extraction methods. Mining of useful content from HTML pages called rule generation method. DOM tree is built to know the visual content of the web page along with features initially. Feature extraction technique is used between and <td> tags. To generate rules and well-formed document, machine learning methods like Decision tree classification and Naïve Bays Classification are applied Rules generated from these methods are used for extracting the core content from the web pages.

b. Outliers Detection Method

Outlier detection is a process for detecting the noises that are irrelevant to the informative content. First the documents are preprocessed for outlier detection. A list of word are generated by tokenizing the document and these are stored in the repository. These tokens are presented in the repository is used to generate a vector numerically representing the preprocessed document. Outliers are identified in the data set depends on the distance to their k nearest neighbors. It uses a distance search through the k-th nearest neighborhood, so it implements some type of locality as well. Those objects with the largest distance to their k-th nearest neighbors are considered as outlier respective to the data set. [5]

IV. PROPOSED METHOD

The objective of the proposed method is to extract informative content from a web page. So, proposed system, the non-informative contents are eliminated and display only the informative contents. A set of web pages is given as input and set of informative contents within the web page provides as output.

Depends on the contents of HTML tags, tags can be classified into two types: positive tag and negative tag. Positive tag contains useful or core content in a web page. Negative tags are also called noises and all tags except the

positive tags are negative tags. Negative tag does not contain any useful information and it reduces the performance of web pages. Removing the negative tags or noisy data will improve the performance of web content mining. Some negative tags are Anchor tag(<a>), Style tag(<style>), Link tag(<link>), Script tag(<script>), Comment tag(<!--…-->), Noscript tag(<noscript>), Horizontal ruler(<hr>) and Line Break(
).

A. Algorithm Noise Removal

Input: An HTML page

1. Convert HTML into XHTML
2. Remove tags<script>, <img>, <style>, <declaration>, <option>, <comment> and <meta>.
3. Construct a DOM Tree T for web page.
4. Compute threshold for web page
5. For every child node c in T do

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Compute character_density(c)
6. If character_density(c) > threshold then
6.1 content = c
7. else remove (c)
8. end if
9. end for

Output: Web page without noises.

Noise in the web page contains less text. Also content contains lengthy texts. In a tree structure , <body> is the root node and it contains noise as well as content. So, it has more text than noise and more hyperlinks than content. Its character_density should be an intermediate value for compare each node's density. Character_density of <body> tag is considered as threshold for our study. If character_density of a node is greater than given threshold then we can say that the node is a content node and not a noise block. If it is a noise block, discard that block.

B. DOM Tree construction

[17]Document Object Model (DOM) is a standardized and language independent interface for accessing and updating content and structure of documents. It is a logical structure of a document. A DOM tree is constructed corresponding to each HTML page. Tags are internal nodes and the information within these tags are leaf nodes.

[18] DOM tree structure permit dynamic access of programs and scripts. It is used to update the content and structure of a page. DOM defines the logical structure of documents and the way for accessing and manipulating the document. Using DOM method , structure of the web page can be constructed. Web pages include noise and relevant data. Consider the example given below.

Example 1.

```
<html>
        <head>
                <title> Welcome:DOM Tree</title>
        </head>
        <body>
           <table>
                <tr>
                        <td> Character Number</td>
                </tr>
           </table>
           <div>
                <p>Tag</p>
           </div>
        </body>
</html>
```

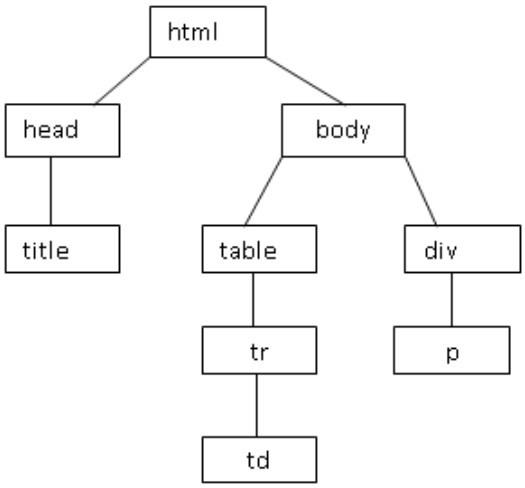

Fig 2: DOM tree construction

Algorithm character_density(c)
Input: Node c
1. for all child node i under c do
2. c.CharNum=CountChar(c)
3. c.TagNum= CountTag(c)
4. if i.TagNum=0 then
5. i.TagNum=1
6. End if
7. Calculate
character_density(c) = i.CharNum / i.TagNum
Output: character_density(c)

Here CharNum is the number of characters present in the subtree.

TagNum is the number of tags in the subtree. Character_density is the ratio of CharNum to TagNum. TagNum value is set to 1 if TagNum equal to 0.

V. RESULT

Character_density for the tags in example1 can be calculated as follows.

1. <body>: CharNum=41, TagNum=5 ,Character_Density=8.2
2. <td>: CharNum=16, TagNum=1,Character_Density=16
3. : CharNum=3,TagNum=1,Character_Density=3

Threshold value is 8.2, Character_Density of each tag is compared and the tag with below this threshold value is considered as noise.

VI. ISSUES IN WEB MINING

- Large Data Sets- Web data sets can be very large. It require huge amount of storage on the database.
- Large no. of Servers- A single server can not mine all the data.
- Hardware and Software Management- Proper organization of software and hardware is required to mine multitera byte data set which is not an easy task.

- Data Cleaning- Automated data cleaning is required on large scale to find out useful information from data.
- Relevant Information- Difficult to find out important information from large database on web.
- New knowledge Mining- Extracting new knowledge from the web by using traditional methods.[1]

VII. CONCLUSION

The technique is proposed for content extraction from web pages in this paper. In this method, the HTML tags are analysed and we divide the tags as Positive tag and Negative tag. All the negative tags are removed, since negative tags are considered as noises present in the web pages. To find out negative tags, character_density is calculated for each tags. After removing all the noises, the informative contents are extracted. It produces effective result for user query and thus the user get accurate information. This method takes less time considering the other methods.

REFERENCES

1. Kavitha,Priyanka Mahani, Dr.Neelam Ruhil, "Web Data Mining Aperspective of research and challenges", IEEE Transactions on knowledge and data Engineering, October 2016.
2. Sandeep Kaur and Abhishek Tyagi "Noise reduction and Content Extraction from web pages using DOM Based Page Segmentation", International Journal of Computer Technology & Application, Vol 5(6),2022-2027, December 2014.
3. Faustina Johnson and Santosh Kumar Gupta, "Web content mining techniques: A survey", International Journal of Computer Applications, Vol. 47, No. 11, June 2012
4. Vishal Gupta, Gurpreet S. Lehal, "A Survey of Text Mining Techniques and applications.", Journal of emerging technologies in web intelligence, Vol. 1, No. 1, August 2009.
5. Surabhi Lingwal, "Noise Reduction and content retrieval from web pages ", International Journal of Computer Applications, Vol. 73, No. 4, July 2013
6. Ms. Pranjali G. Gondse, Professor Anjali B. Raut "Main Content Extraction From Web Page Using Dom", International Journal of Advanced Research in Computer and Communication Engineering Vol. 3, Issue 3, March 2014
7. D.S.Misbha, J.R.Jeba, "Scheduling Effective Cloud Updates in Streaming Data Warehouses using RECSS Algorithm" IJAER Vol.11 No.7
8. K. Nethra1, J. Anitha2 and G. Thilagavathi, "Web Content Extraction Using Hybrid Approach", ICTACT Journal On Soft Computing, January 2014, VOLUME: 04, ISSUE: 02
9. Fei Sun,Dandan Song and Lejian Liao ,"DOM based Content extraction via Text Densiy ", International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3066-3068
10. J.R Jeba, S.P.Victor,"A novel approach for finding item sets with hybrid strategies", International Journal of Computer Applications., Vol.17,No.5,2011
11. [12] J.R.Jeba, S.P.Victor," Comparison of frequent item set Mining algorithms", International Journal of Computer Science and Information Technologies, Vol 2 (6), 2011
12. Jincymol Joseph, J.R.Jeba , "Survey on web Content Extraction" , IJAER Vol.11 No.7
13. J R Jeba, S.P.Victor, "Effective measures in Association Rule mining", International Journal of Scientific and Engineering research, Vol 3,Issue 8,2012.
14. A.F.R Rahman, H.Alam and R.Hartono, "Content extraction from HTML documents", International workshop on Web document Analysis, pp.7-10, 2001.
15. CincyW.C,J.R.Jeba, "A method of A-BAT Algorithm based Query Optimization for crowed Sourcing System, IJ Intelligent Systems and applications, March 2018.
16. R.Gunasundari, "A study of content extraction from web pages using links"International Journal of Data Mining & knowledge management process, Vol.2,No.3,May2012
17. S.S. Bhamare,Dr.B.V.Pawar, "Survey on Web Page Noise Cleaning for Web Mining", International Journal of Computer Science and Information Technologies, Vol. 4 (6) , 2013
18. Pralhad S. Gamare, G. A. Patil, "Web document clustering using hybrid approach in data mining" International Journal of Advent Technology, Vol.3, No.7, July 2015
19. CincyW.C,J.R.Jeba, Performance Analysis of Novel Hybrid A-BAT Algorithm in Crowdsourcing Environment" , IJAER, Vol.12 No.24
20. W3C document object model. Website, 2009.
21. http://www.w3.org/DOM.
22. Bhavdeep Mehta,Meera Narvekar, "DOM Tree based approach for web content extraction", IEE

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication