Data Preprocessing Framework for Web Structure Mining

Kavita Kanathey[1], R. S. Thakur [2], Shailesh Jaloree [3]
[1] *Research Scholar, Barkatullah University, Bhopal, INDIA*
[2] *Professors, Department of Mathematics & Computer Application, MANIT, Bhopal*
[3]*Professors, Department of Applied Maths and Computer Science, SATI, Vidisha*

***Abstract*—**World Wide Web is growing exponentially as millions of pages are added on a daily basis. It provides adequate information to user to cater their need. Web mining is the application of data mining techniques and methodologies to discover knowledge or hidden information from the web. According to analysis targets, Web mining can be divided into three different types, which are web usage mining, web content mining and web structure mining. Web Structure Mining is the process of extracting knowledge from hyperlink structure of the web. Web Structure Mining has four phases namely Data Collection, Data Preprocessing, Knowledge Discovery and Knowledge Analysis. This paper mainly focuses on data collection and data preprocessing phase. In this paper, we have proposed a framework for data collection and preprocessing for web structure mining in order to determine the valid web pages connected to the given URL and relationship among web pages. During Data collection phase we scan the target web page and find out all the hyperlinks connected to the given target web page for analysis. In data preprocessing phase we eliminate all irrelevant links, validate links, identify link uniqueness and eliminates all graphics and other files such as jpg, jpeg, pdf, .css etc.. The proposed methodology is implemented using MATLAB and experiments are conducted on various academic sites.

Keywords *- Web Mining; Web Structure Mining; Data Collection; Data Preprocessing; Knowledge discovery; Knowledge analysis.*

I. INTRODUCTION

The explosive growth of both web based technologies like E-commerce, OLTP (Online Transaction Processing) and the volume of users to the internet have caused increasing information overload problem where finding relevant information to exactly meet their needs efficiently has become a challenge. Web mining is the more appropriate solution to deal with this problem. Web mining is the application of data mining techniques to discover and extract knowledge from Web [1][7]. Web mining is categorized into three research areas based on extracting knowledge. They are Web Content Mining, Web Structure Mining and Web Usage Mining. Web content mining is the process of extracting useful information from the content of web documents and services [2][9]. It may consist of text, images, audio, video, or structured records such as lists and tables.

Web structure mining is the process inferring structure information from the web. It focuses on hyperlink structure of the web [3].

Web usage mining is the process of discovering interesting user's navigation patterns in web access logs and predicting user's behavior [4].

A. Web Structure Mining

Web structure mining refers to extraction of useful knowledge from hyperlink which represents the structure of web. The hyperlink structure of web helps to discover information, site modification, and user's behavior analysis and recommendation. The web structure mining process can be decomposed into four stages as shown in Figure1 [1].

- **Data Collection** - Data collection is the process of retrieving intended web pages from web. In the context of web structure mining, data collection refers to the process of extraction of hyperlinks from web pages associated with the target URL.
- **Data Preprocessing –** It is a sequence of actions performed on web link file. It includes data cleaning, elimination of irrelevant links, link validation, link uniqueness, elimination of graphics and other files.
- **Knowledge Discovery –** In order to discover knowledge from hyperlinks, various machine learning and data mining techniques are applied on processed data such as Association rule mining, Clustering, Genetic algorithms, Fuzzy set etc.
- **Knowledge Analysis –** It is interpretation and/or validation of mined pattern.

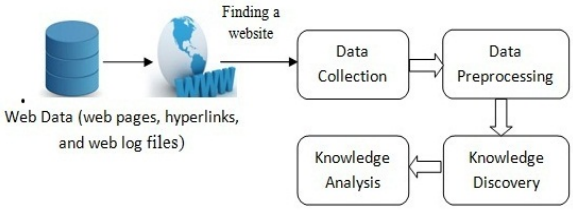

Figure 1. Web Structure Mining Process

B. Web Data

Data available on web is referred as web data.web data can be text, multimedia, hyperlinks or usage data [5]. Web data is classified into the given areas:

- **Content:** The content data exist on Web pages may consist of text, images, audio, video or structured records such as tables and lists.
- **Structure:** The way in which the content of a web page is organized is represented by structure data. They

can be HTML or XML tags or hyperlinks connecting one page to another.

- **Usage:** Data that describes the pattern of usage of Web pages by users, such as IP addresses, page references, and the date and time of accesses.
- **User Profile:** Data that dispenses demographic information about users of a Web site, as well as information about users' interest and preferences. This includes registration data and customer profile information.

As web data is unstructured, heterogeneous and noisy, so all web mining techniques described above cannot be directly applied. Thus it has to be passed through data preprocessing phase which makes it appropriate for mining. In this paper, section II describes data collection stage. Section III gives concept of data preprocessing. Section IV describes proposed methodology. Section V discusses experimental results and Section VI gives conclusion.

II. DATA COLLECTION

In web structure mining, Data collection is the process of retrieving the required web documents from the web. It is the process of extraction of hyperlinks from web pages associated with the target URL. The structure of a typical web graph consists of web pages as nodes, and hyperlink as edges connecting between two related pages as shown in figure 2. A hyperlink is a structural unit that connects a location in a web page to a different location, either within the same web page or on a different web page. A hyperlink that connects to a different part of the same page is called an intra-document hyperlink, and a hyperlink that connects two different pages is called an inter-document hyperlink.

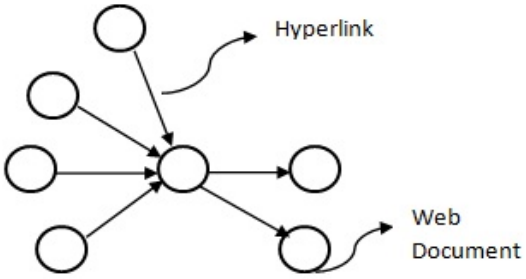

Figure 2. Web Graph

III. DATA PREPROCESSING

Data preprocessing is a set of operations that is performed upon raw data to convert it into processed data for knowledge discovery and pattern analysis [6]. It mainly includes - data cleaning, data integration, data reduction and data transformation [8].

- **Data cleaning:** data cleaning process remove noise from data and also correct inconsistencies in data.
- **Data integration:** It merges data collected from various sources into a coherent data store such as a data warehouse.
- **Data reduction:** Data reduction reduces data size by, for instance, aggregating, eliminating redundant features, or clustering.
- **Data transformations**: It may be applied, where data are scaled to fall within a smaller range.

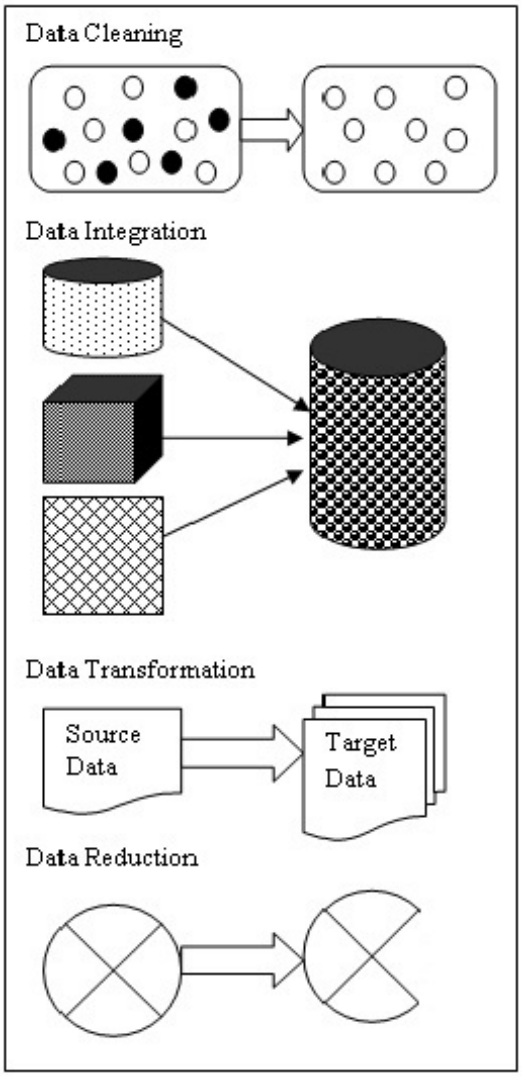

Figure 3. Data Preprocessing Techniques

Effective data preprocessing not only improves the quality of mining model but it also reduces the time and efforts needed for mining. The data preprocessing has been described in Figure 3.

Data cleaning is the first step of preprocessing. This step consists of removing and filtering irrelevant link from the list of hyperlinks. Most web pages contain graphics and multimedia files, PDF file, various style sheets and other relevant files. Table 1 shows some of the irrelevant files and Table 2 shows Unsuccessful Status code Field.

TABLE 1 DETAILS OF IRRELEVANT FILE EXTENSION

Extension Field	Description
.jpeg, .jpg, .gif, .bmp	Image file
css	Html style sheet
Js	Java script
mp3	Audio File
Swf	Flash Animation File

TABLE 2 DETAILS OF UNSUCCESSFUL STATUS CODE FIELD

Status Code	Description
401	Bad Request
403	Forbidden
404	Not Found
500	Internet Server Error
501	Not Implemented
503	Server Unavailable

IV. PROPOSED METHODOLOGY

In order to determine the valid web pages related to the given URL and to find relationship among them, we have proposed a framework for structure mining of web pages. The proposed methodology is divided into two phases: data collection and data preprocessing as shown in figure 4. During Data collection phase, the target web page is scanned and find out all the hyperlinks connected to the given target web page for analysis. Data preprocessing phase eliminates all irrelevant links, identify link validity, identify link uniqueness and remove all the graphics and other files such as jpg, JPEG, pdf, .css etc. The proposed methodology is implemented using MATLAB (version 7.0.12.635).

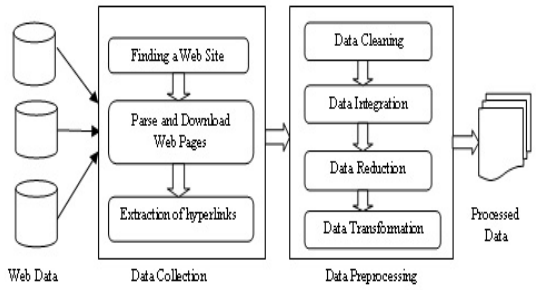

Figure 4. Proposed Framework of Data Preprocessing

The following algorithm is used to find all the valid web pages associated with the given URL and relationship among them.

Proposed Algorithm:

Input: Address of any website (URL) and NO. of link to be scan.

Output: dat file containing number of url with their address and relationship among them.

Variables:

- Uout → array containing distinct url's.
- Lout → 2D array containing connectivity among url's.
- nurl → variable that defines number of distinct url's.
- n404 → variable that defines number of broken links.
- nlink → variable that defines number of links.
- Nmeta → no of other files like jpg , pdf , css, jpeg, js

Step 1: Input URL and number of link N to be extracted from users to the system for next step.

Step 2: Create empty array Uout and Lout.

Step 3: initialize variable nurl, n404, nlink nmeta to 0.

Step 4: for each i=1 to n repeat

4.1 Read the URL and set it into variable src

4.2 Parse and download HTML page that associated with requested URL.

4.3 Scan the target web page and then find out HTML tags that contain hyperlink and other relevant tags.

4.3.1 for each (hyperlink in webpage) repeat

4.3.1.1 read the hyperlink and set it into variable Dest;
4.3.1.2 If Dest = 403 || Dest = 404 then

4.3.1.2.1 Increment n404 by 1;
4.3.1.2.2 Continue;

4.3.1.3 If Dest contains extension .gif, .pdf, .css, .js then

4.3.1.3.1 increment nmeta by 1;
4.3.1.3.2 add link into output MetaLink;
4.3.1.3.3 continue

4.3.1.4 if Dest is a new URL and if it is less then N then

4.3.1.4.1 increment nurl by 1;
4.3.1.4.2 add new url to Uout;
4.3.1.4.3 increment nlink by 1;
4.3.1.4.4 add connectivity (i,j) to output Lout;

end of step 4.3.1
end of step 4

Step 5: Create sparse matrix by using Lout;

Step 6: Save the Uout with index and non zero entries of sparse matrix in output.dat file.

V. EXPERIMENTAL RESULTS

To support our methodology, the proposed algorithm is implemented using MATLAB (version 7.0.12.635). The experiments were performed with three academic sites namely www.lnctgroup.in, www.manit.ac.in and www.w3schools.com. The following Table 3 depicts the URL of website, No. of hyperlinks to be scan, No. of images, No. of pdf, No. of css and other links.

TABLE 3 DETAILS OF FILTERED GRAPHICS,PDF CSS AND OTHER IRRELEVANT LINKS

URL	Count of Link to be scan = N	Count of images (.jpeg .jpg .gif)	Count of pdf file	Count of css file	Other link	Total no of pages
www.lnctgroup.in	50	1	0	1	20	72
	100	1	0	2	25	128
	150	4	26	2	28	160
	200	6	45	2	28	281
www.manit.ac.in	50	12	149	3	68	282
	100	22	334	3	73	532
	150	70	426	7	99	752
	200	70	426	7	101	804
www.w3schools.com	50	1	0	9	983	1043
	100	1	0	9	983	1093
	150	1	0	13	2794	2958
	200	1	0	13	2794	3008

The following Table 4 depicts the URL of website, No. of hyperlinks to be scan, connectivity and evaluation time.

TABLE 4 DETAILS OF CONNECTIVITY AND EVALUATION TIME

URL	Count of Link to be scan = N	connectivity	Evaluation time
www.lnctgroup.in	50	1280	32.10
	100	3297	104.79
	150	9176	105.76
	200	13870	131.55
www.manit.ac.in	50	2215	45.21
	100	9415	97.15
	150	14392	145.01
	200	18540	202.98
www.w3schools.com	50	1537	19.67
	100	1631	33.77
	150	5563	100.36
	200	5828	93.23

Based on the above table entries, the results are shown pictorially below. Figure 5 shows evaluation time to process all these academic sites at different no. of web pages.

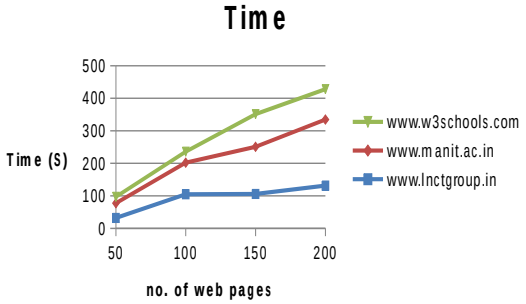

Figure 5 Evaluation Time taken by academic sites at different no. of web pages.

Figure 6 shows count of images, pdf, css and other links for "www.lnctgroup.co.in" at different no. of web pages.

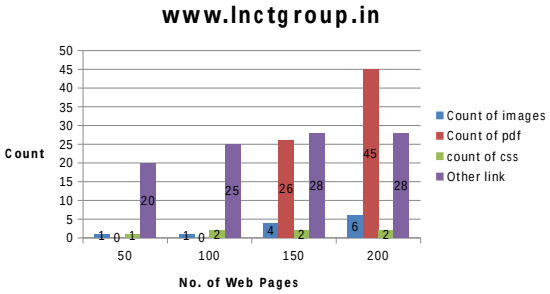

Figure 6 counts of images, pdf, css and other links at different no. of web pages.

Figure 7 shows count of images, pdf, css and other links for "www.manit.ac.in" at different no. of web pages.

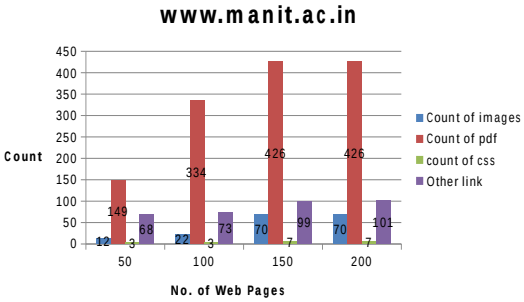

Figure 7 counts of images, pdf, css and other links at different no. of web pages.

Figure 8 shows count of images, pdf, css and other links for "www.w3schools.com" at different no. of web pages.

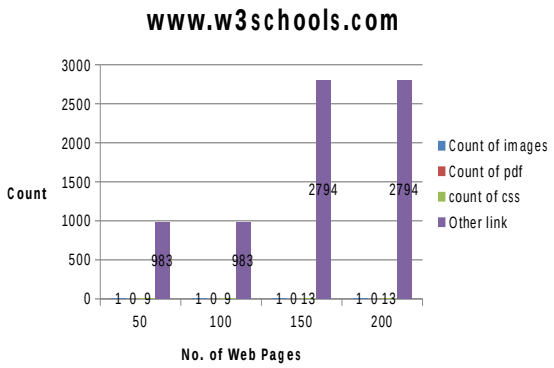

Figure 8 counts of images, pdf, css and other links at different no. of web pages.

V. CONCLUSION

In this paper, we have presented a data preprocessing framework for web structure mining. Usually web data is noisy, inconsistent and irrelevant by nature therefore this web data is not suitable for mining and analysis purpose so it needs to pass through data preprocessing phase. In this paper, data preprocessing phase removes and filter unwanted embedded objects of style, scripts and graphics to provide valid web pages and relationship among web page.

VI. REFERENCES

[1]. R. Kosala and H. Blockeel, "Web mining research: A survey" SIGKDD Explor.Newsl. ACM SIGKDD Explorations Newsletter, vol. 2, no. 1, pages 1– 15, Jan. 2000.

[2]. P. Sukumar, L. Rober, S. Yuvaraj, "Review on Modern Data Preprocessing Techniques in Web Usage Mining (WUM)", International Conference on Computational Systems and Information Systems for Sustainable Solutions, pages 64-69, 2016.

[3]. S. Jeyalatha and B. Vijaykumar , "Design and implementation of a web structure mining algorithm using breadth first search strategy for academic search application" , International Conference on Internet Technology and Secured Transactions,pages 648-654, Dec. 2011.

[4]. R.M. Suresh, R. Padmajavalli, "An Overview of Data Preprocessing in Data and Web Usage Mining", IEEE proceedings pages 194-198,2006.

[5]. Magdalini Eirinaki, "Web Mining : A Roadmap", Journal of athens University of Economics and Business", pages 1-8, 2004.

[6]. K. Sudheer Reddy, M. Kantha Reddy and V. Sitaramulu, "An effective Data Preprocessing method for Web Usage Mining", International Conference on Information Communication and Embedded Systems (ICICES), pages 1-8, 2013.

[7]. K.Amutha, Dr.M.Devapriya, "Web Mining: A Survey Paper", International Journal of Computer Trends and Technology (IJCTT), vol. 4, Issue 9, pages 3038-3048, Sep 2013

[8]. Suvarn Sharma and Amit Bhagat, "Data preprocessing algorithm for web structure mining" ,International conference on Eco-Friendly Computing and Communication System (IJECCS) , pages 94-98, 2006.

[9]. Dr. Sanjay Kumar Dwivedi and Bhupesh Rawat, "A Review Paper on Data Prerprocessing: A Critical Phase in Web Usage Mining Process", International Conference on Green Computing and Internet Of Things(ICGCIoT), Pages 506-510, 2015.