Data Preprocessing Algorithm for Web Structure Mining

M. Gopi chand, PG Scholar
Department of Master of Computer Applications
Vasireddy Venkatadri Institute of Technology
Nambur, Andhra Pradesh, India

B. *Laxmi Praveena ,Assistant Professor*
Department of Information &Technology
Vasireddy Venkatadri Institute of Technology
Nambur, Andhra Pradesh, India

***Abstract*—World Wide Web is an extremely large collection of information, i.e. beyond our imagination. It provides enough information according to user's need. Web is rising dreadfully as approximately 70 million pages are added daily. Knowledge Discovery on web data is referred as Web Mining. Web Structure Mining based on the analysis of patterns from hyperlink structure in the web. Like as Data Mining, Web Mining has four stages i.e. Data Collection, Preprocessing, Knowledge Discovery and Knowledge Analysis. This paper based on the first two stages Data collection and Preprocessing. Data collection is to collect the data required for analysis. Data preprocessing is considered as an important stage of Web Structure mining because of data available on web is unstructured, heterogeneous and noisy.**

***Keywords*—Data Mining; Web Mining; Web Structure Mining; Data Preprocessing.**

I. INTRODUCTION

Data mining refers to the extraction of useful information from a bulk of data or data warehouses. A large amount of data is available for users due to the growth of the web. Web structure mining can be done in several phase: Data Extraction, Data cleaning, Data fusion, Data Extraction is used to extract is used to extract the log data. Data cleaning means cleaning the irrelevant data from the extracted data. Data fusion is used to collect pages from web servers.

The web pages are the objects in the web and in-out and co-citation[4] are links. Web mining is to discover and extract the knowledge from the World Wide Web[5] [7] [8].In general web mining is classified into three categories. Web content mining , Web structure mining, Web usage mining.

Web Content Mining mainly focuses on the structure of the inner document and extracting the information from the content available on the web.

Web Structure Mining is to discover the link structure of the hyperlinks. Based on the topology of the hyperlinks, web structure mining will categorize the web pages and generate the information.

Web Usage Mining is to discover interesting patterns by analyzing the user's navigational behavior from the web data.

A. *Web Structure Mining*

Web Structure Mining[1][9] is applying of data mining techniques on large web link's structure repositories which can be used for improvement in web designing tasks. For web structure mining the data is extracted as web pages collected by crawler from all over the world wide web server . There are four steps in web structure mining.

Data Collection: The first step in any mining technique is to collect the data required for analysis. In web structure mining data collection means collect hyperlinks from web pages

Preprocessing: This is one of the techniques of data mining which is used to preprocess the data to make the data as structured and relevant to the requirements.

Knowledge Discovery: Applying various data mining techniques for processing data such as statistical explication, association, clustering, pattern analysis and such like.

Knowledge analysis: This is used to filter the irrelevant data which is available on the web and give the data which is more relevant to the user requirement.

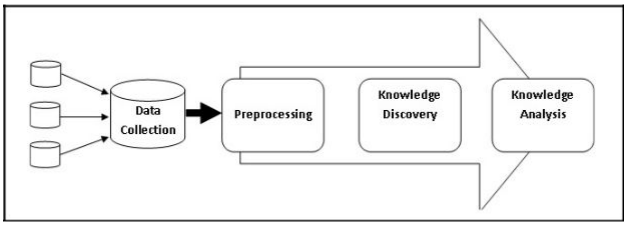

Fig. 1. *Web Structure Mining Process*

B. Web Data

Web Mining, it consists of web data which is of different types available in the web . The techniques which are used to mine the data available on the web is called web mining and its ways of execution.[7][8]

Web data is classified into the given three areas:

Content: This is about the content which is available on the web . The data on the web may be of different types like text, audio ,video and tables, lists, etc.

Structure: This is about the structure of the web pages. The content int the web pages is arranged by using the Html tags and the Cascading style sheets.

II. Data Collection

In web graphs the sites and links are considered as nodes and edges. Data collection means selection of documents from Web repository where a large amount of data is available on the web.

Hyperlinks are structural components that connect web pages from different location on the Web. The study and analysis of these hyperlinks helps to discover and valuable and rare information available in the hidden form on the web.

III. Data Preprocessing

Data Preprocessing is one of the data mining techniques used to represent the data according to the data mining techniques and their similarities.[2]

Data Preprocessing can be classified as the following :

- Data Cleaning: It means to cleaning the data i.e., the data which consists of noise and irrelevant information is cleaned to get the relevant data without any noise.
- Data Integration: In this step data collected from multiple sources that may be in different formats.
- Data Transformation: This step considers transformation of collected data into a unique format that can be used for mining technique.
- Data Reduction: This step looks to reduction of transformed data by extracting important features for mining technique.

IV. Structure Preprocessing

This is a process which is similar as data processing. Through structure processing we can process the structure of a web which consists of the hyperlinks as urls and this may have each session. [10]

A. Data Cleaning

Data Cleaning is the first step of preprocessing. Techniques. Data cleaning is the process of preparing data for analysis by removing or modifying data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted.

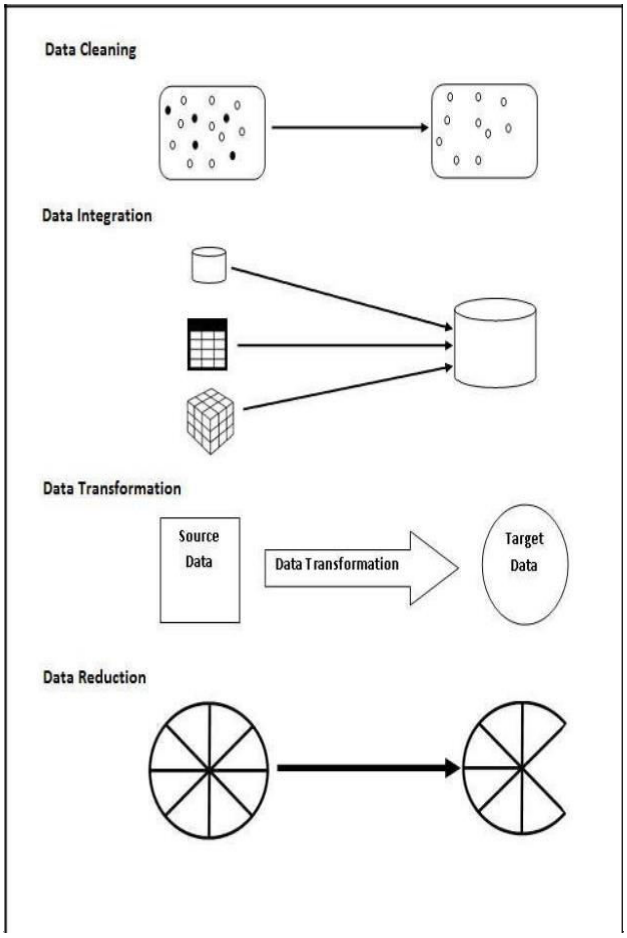

Fig. 2. *Data Preprocessing*

TABLE I. DESCRIPTION OF IRRELEVANT FILE EXTENSIONS

File Type	Description
jpeg , jpg ,gif, png, tif, bmp	Image file
mp3	Audio file
css	Html style sheet.
js	Java Script file
swf	Flash animation File
ico	Icon Image file format
cgi	Common gateway interface

TABLE II. DESCRIPTION OF ERRORS OCCUR WHILE REQUESTING A PAGE[11]

Error Code	Error Msg
400	Bad Request
403	Forbidden
404	Not Found
407	Proxy Authentication Required
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Server Unavailable
504	Gateway Timeout

V. ALGORITHM

The algorithm used for the implementation is Crawling algorithm. This algorithm is used to crawl web pages by using the hyperlinks connected in the web page. After crawling the web pages, the hyperlinks connected urls are stored for further processing of data.

These urls are retrieved from the database when the user searches for the urls. The urls then retrieve the necessary content from the web pages.

Step1

In this step we will load the data of urls which is stored in the database by using the algorithm

Step2:

The urls stored in the database must be retrieved by parsing the web page

Step3:

In this step algorithm all the hyperlinks which are available in the content of a web page.

Step4:

The urls which are extracted by the algorithm will show the errors like , NOT FOUND (404) error, Forbidden (403) error etc.

By using web crawler which is used to crawl web pages and to traverse using the relation between hyperlinks of web pages.

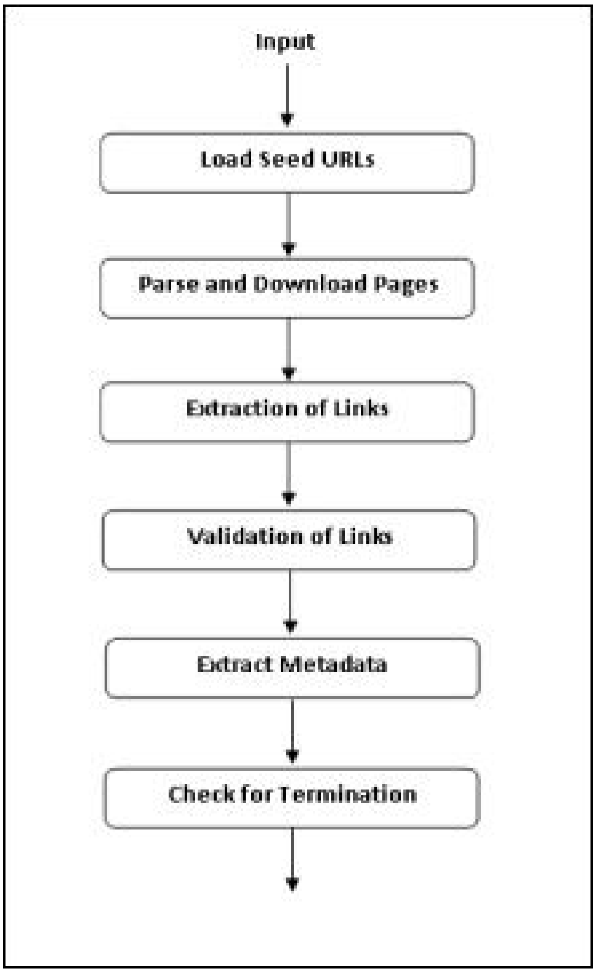

Fig 3. Preprocessing Algorithm

VI. RESULTS

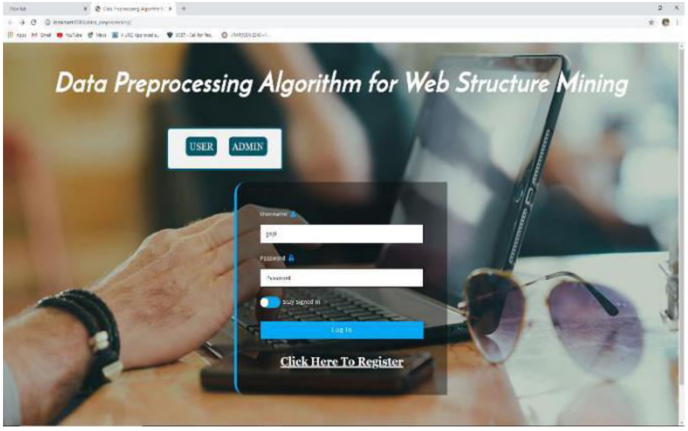

Fig 4.User login

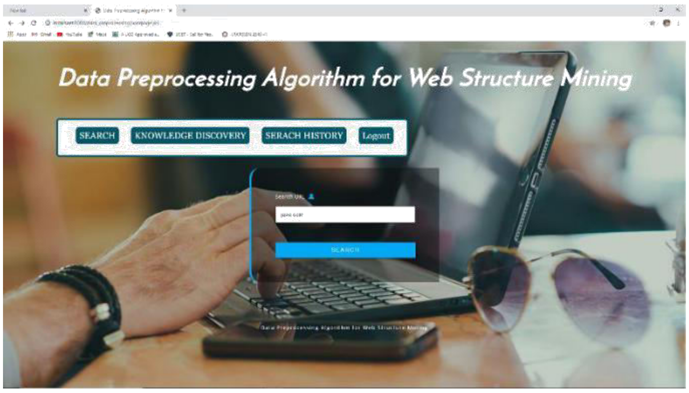

Fig 5.Searching the url from database

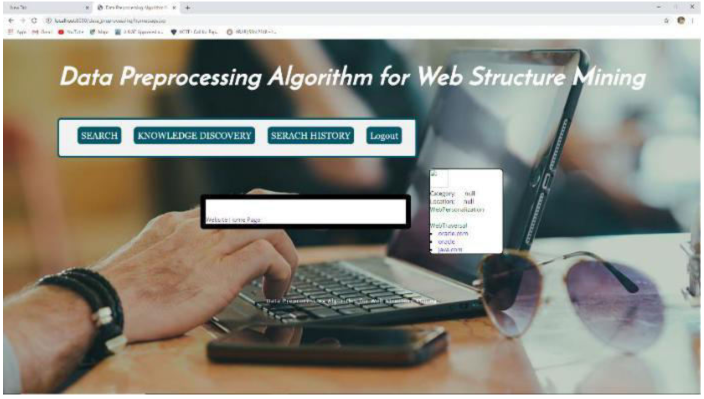

Fig 6.1 results from the database

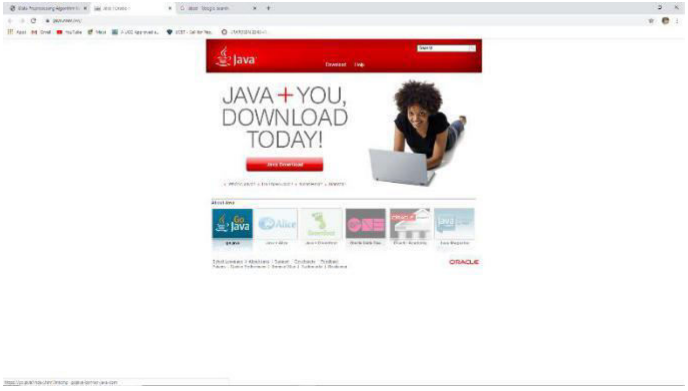

Fig 6.2 result from the web

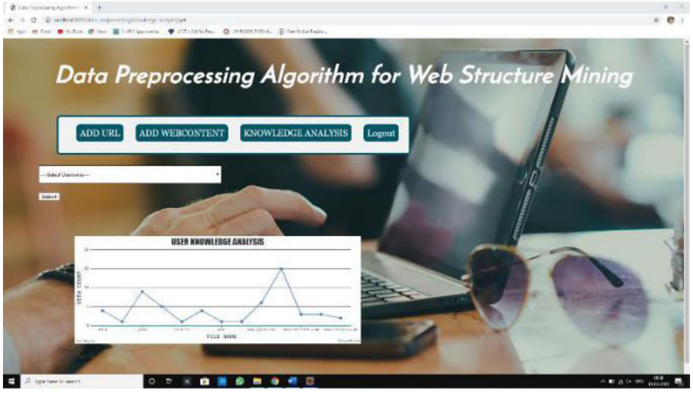

Fig 7. Knowledge analysis

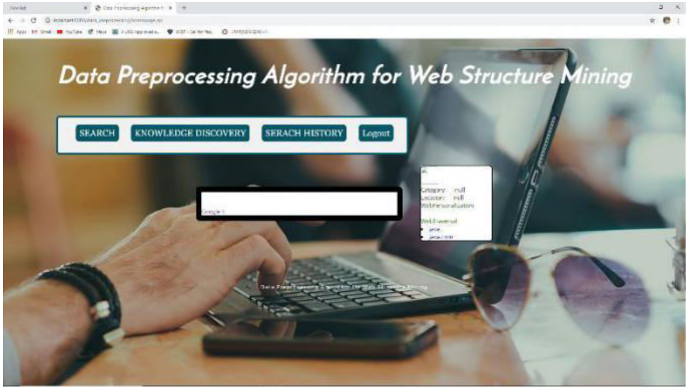

Fig 8.result for url which is not stored in database

VII. Conclusion

In this article, we are going to preprocess the data using an algorithm. This algorithm will extract the hyperlinks from all the required urls. From these hyperlinks the algorithm preprocesses the data and removes the data which is irrelevant to the user and gives the relevant information to the user which is required by the user. Finally, the outcome is to give relevant information to the user by filtering the irrelevant data.

References

[1] M. D. Costa and Z. Gong, "Web structure mining: an introduction," *2005 IEEE International Conference on Information Acquisition*, pp. 590–595.

[2] J. Han and M. Kamber, "Data Preprocessing Techniques for Data Mining," in *Data mining: concepts and techniques*, San Francisco: Morgan Kaufmann Publishers, 2001.

[3] P. Desikan, J. Srivastava, V. Kumar and P.-N. Tan, "Hyperlink Analysis – Techniques & Applications," Army High Performance Computing Center Technical Report(2002).

[4] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A.

[5] S. Tomkins, "The Web as a Graph: Measurements, Models, and Methods," Lecture Notes in Computer Science Computing and Combinatorics, pp. 1–17,

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

[7]

[8] J. E. Pitkow and J. Pitkow and K. Bharat, "WebViz: A tool for WWW access log analysis," Computer Networks and ISDN Systems, vol. 27, no. 2, p. 35-51, 1994.

[9]

[10] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining," SIGKDD Explor. Newsl. ACM SIGKDD Explorations Newsletter, vol. 1, no. 2, p. 12, Jan. 2000.

[11] S. K. Madria, S. S. Bhowmick, W.-K. Ng, and E. P. Lim, "Research

[12] Issues in Web Data Mining," DataWarehousing and Knowledge Discovery Lecture Notes in Computer Science, pp. 303–312, 1999.

[13] T. Srivastava, P. Desikan, and V. Kumar, "Web Mining – Concepts, Applications and Research Directions," Foundations and Advances in Data Mining Studies in Fuzziness and Soft Computing, pp. 275–307, 2005.

[14] M. Thelwall, "Mining the World Wide Web: An Information Search

[15] Approach20024George Chang, Marcus J. Healey, James A.M. McHugh and Jason T.L. Wang. Mining the World Wide Web: An Information Search Approach . Boston, London: Kluwer Academic

[16] Publishers 2001. 168 pp., ISBN: ISBN: 0 7923 7349 9 £79," Journal of Documentation, vol. 58, no. 2, pp. 232–234, 2002. https://en.wikipedia.org/wiki/List_of_HTTP_status_codes