Data Preprocessing Algorithm for Web Structure Mining

Suvarn Sharma, Research Scholar
Department of Mathematics and Computer Applications
Maulana Azad National Institute of Technology
Bhopal, India
suvarnasharma.mtech@gmail.com

Amit Bhagat,Assistant Professor
Department of Mathematics and Computer Applications
Maulana Azad National Institute of Technology
Bhopal, India
am.bhagat@gmail.com

***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 is the process of analyzing data from different perspectives and summarizing it into useful information. With the growth of Web, a large amount of data is now available for users on web. Web .Preprocessed data improves efficiency and scalability of later stages of Web structure mining. This can be done in several phases: Data fusion, Data Extraction, Data cleaning, links and metadata extraction, Path completion etc. Data fusion includes collecting of pages collected from various Web servers. Data extraction is used to extract log data according to time duration of analysis. Data cleaning refers to the cleaning of irrelevant links which is not useful for the purpose of structure analysis i.e. multimedia files html style sheet etc.

The objects in the web are web pages, and links are in-, out- and co-citation [4]. The structure of Web is a graph structured by documents and links; results of mining may be Web contents or Web structures. Web mining is the use of the data mining techniques to automatically discover and extarctthe knowledge from World Wide Web[5][7][8]. In general Web Mining categories into three area: Web content mining, Web structure mining and Web usage mining.

Web Content Mining focuses on the extracting useful information from the content or data or services available on the web document.

Web Structure Mining deals with the extraction of patterns from hyperlinks within the web itself. Hyperlink is a structural component that connect web pages from different loacation.

Web Usage Mining tries to identifing browsing patterns by analyzing the users navigational behavior.

A. Web Structure Mining

Web Structure Mining[1][9] also referred as web link's structure analysis is the application of data mining techniques on large web link's structure repositories to generate structural summary about web sites and web pages by analyzing the hyperlinks, that can be used for improvement in web designing tasks. The origin of data source for web structure mining consists of textual web pages collected by crawler from all over the world diverse 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 associated with seed Urls from various servers.

Preprocessing: Implements a sequence of process of web links file performing data cleaning, links validation, links identification, links uniqueness and links completion.

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

Knowledge analysis: After final discovery of information from web links, filter unrelavant information and to estimate and delibrate the interesting pattern to users. Knowledge about link's structure also allowsfiltering un-useful knowledge.

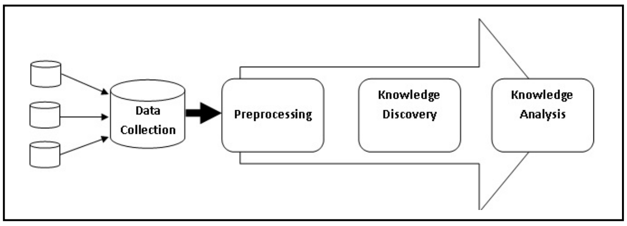

Fig. 1. *Web Structure Mining Process*

978-1-5090-2541-1/16/$31.00 ©2016 IEEE

Working of all steps is depicted through the Fig.1:

B. Web Data

Web Mining, web data can be of different type, where each form sample datadiffers not only in rapport of the location of the data source, but also the variants of data present, the fraction of population from where the sample data was collected, and its ways of execution.[7][8]

Web datais classified into the given three areas:

Content: The real data exist on the Web pages,may consists of text, images, audio, video or structured records such as tables and lists. This is not only limited to text and graphics.

Structure: The organization of the content of web pages arranged with the HTML or XML tags and represented as a tree structure. Knowledge about this organization is reffered as web structure. Data used to exaplain this structure known as web structure data.The majormode of inter-page structure information is hyper-links 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 the Web services. This includes registration data and customer profile information.

II. Data Collection

Web graph is a graph that contains sites and links as nodes and edges. Now a day's web is a typical hypertext system: a widely distributed collection of documents, connected through links present in the body of each document. Data collection means selection of documents from Web repository.

Hyperlinksare structural components that connect web pages from different loacationon the Web. By studying and analysing links and finding structure between them ease to identify relationships and the patterns of occurrences. The study and analysis of these hyperlinks helps to discover and rediscovervaluable and rare information available in the hidden form on the web.

III. Data Preprocessing

Data Preprocessing is a process to represent the data in format as per mining techniques. There are different ways for representing the data such as chart, graph, etc. Many features are also used for weighing the documents and their similarities.[2]

Data Preprocessing is classified into four categories:

- Data Cleaning: Data cleaning is the first step and it plays an important role in preprocessing to fetch cleaned data for afore processing.
- Data Integration: In this step data collected from multifarious sources that may be in different formats.
- Data Transformation: This step considers transformation of collected data into a unique format that appropriate for mining technique.
- Data Reduction: This step looks to reduction of transformed data by extracting important features for mining technique.

IV. Structure Preprocessing

The structure of a web page is designed by the hypertext links between page view codes. The structure can be possessed and pre- processed in the similar way as the content of a web page. A different web page structure may have to be raised for each server session. [10]

A. Data Cleaning

Data Cleaning is the first step of preprocessing. Techniques to clean link's list to evictunvaluedlinks are of importance for any type of link analysis. By checking the type of the *"href"* html tag eviction of unvalued links can be reasonably accomplished.For instance, Table 1 shows the all links with type like, gif, jpeg, GIF, JPEG, jpg, JPG, and map can be evicted.So in structure mining data cleaning means selection of links related to web pages and active.

Fig.2 described all these clearly:

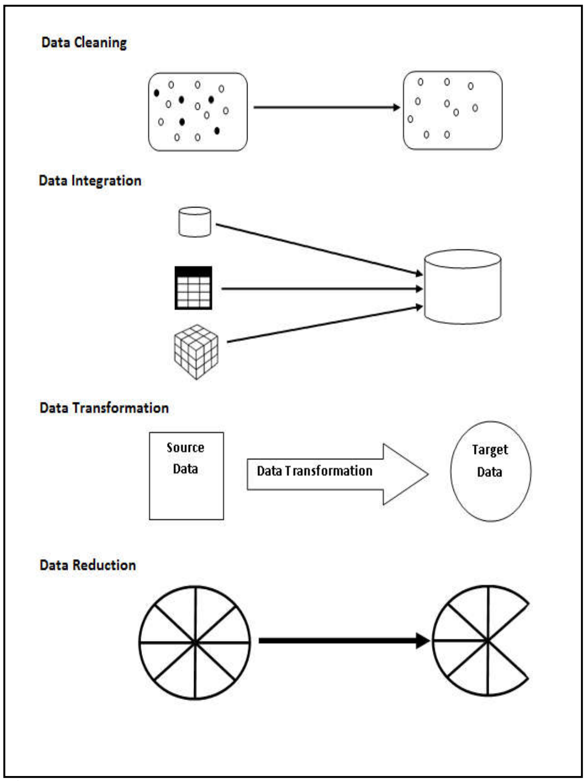

Fig. 2. *Data Preprocessing*

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

B. *Data Integration*

Extracted links, from different server and their relevant information required for process stored together.

C. *Data Transformation*

Transform the collected data in a unique format for process ahead.

D. *Data Reduction*

Data reduction means selects only that information required for algorithm among lots of information available on web page.

V. PRILIMINARIES AND PROBLEM DEFINITION

Definition 1: Web Data

Now a day Web contains large number of documents redacted by millions of users. Web is one of the giant and most pervasively known information sources.

Definition 2: Web Documents

The information, exist on the web, to available to the users as web documents. This usually consists of text, images, audio, video, or structured records such as lists and tables.

Definition 3: Hyperlinks

The hyperlinks are used to link and connect web pages with each other.

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*.[3]

Definition 4: Crawling

Web crawler is a tool used to traverses the Web in an automated manner by downloading and following hyper-link structure between pages.

Definition 5: Seed Urls

Crawler keeps aunvisited URLS list called frontier, which is initialized with seed URLs.

The crawler begins with a *seed set*that is collection of URLs.

Definition 6: Web Page's Parsing and Extraction

Parser choose a URL from the seed urlset, and thenrevolves the web page's link at that current URL. The fetched pages are then parsed, to extract both the text and the links from the page.

Definition 7: Web Page's Meta Data

Meta data means other information relevant to pages like date of creation, date of last update, keywords etc.

VI. ALGORITHM

This crawling algorithm consist six steps in different steps. A URL server stepfetches URLs out of a file and forwards them to crawler again. Crawler procedure runs on a machine, is single-threaded, and to fetch data from servers used asynchronous I/O. The crawler transmits fetched pages to a Store Server process, which after compressionstore pages to disk. By an indexer process from diskpages are then load back, which reads new URL's links and saves them to a different disk file. A URL resolver procedure reads the link file, verifies the URLs implanted therein, and stores the absolute URLs to the disk file that is fetch by the URL server. Valid URLs are used by URL server to extract other information related to links that will use for further process.

Step1: **Load Seed URLs**

Load set of *seed set* from database to main memory for next step.

Step2: **Parse and Download Pages**

Parse and download pages that associated with requested Url.

Step3: **Extraction of Links**

In this step algorithm read and understand an HTML file and find HTML tags such as <a href= "- - - "> and other relavant tags. Extract new links from targeted pages that associated with Urls.

Step4: **Validation of Links**

Extracted Urls passes to certain URL test. Its check different types of error at Url's browsing time like, NOT FOUND (404) error, Forbidden (403) error etc.

Step5: **Extract Metadata**

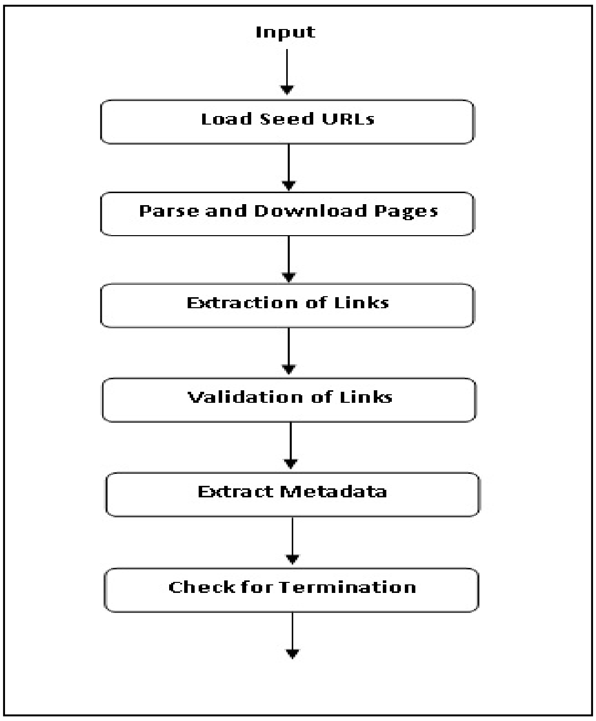

Fig. 3. *Preprocessing Algorithm*

Extract other information as attributes for algorithm required from downloaded pages. Based on these information data designed as per algorithm requirement.

Step6: **Check for Termination**

Check all the reachable links are visited or not if so then process terminate. Otherwise process will continue.

VII. Results

Depth-0

TABLE I. URL DETAILS AT DEPTH-0

Url	Media Files	Import Files	Page's Links	Intra Pages	Inactive Pages	Valid Pages
mit.edu	11	8	68	60	0	8
jiwaji.edu	113	2	95	1	0	73
javatpoint.com	81	3	171	69	0	78
manit.ac.in	28	20	208	39	0	161
dauniv.ac.in	19	5	126	63	0	49

Depth-1

TABLE II. URL DETAILS AT DEPTH-1

Url	Media Files	Import Files	Page's Links	Intra Pages	Inactive Pages	Valid Pages
mit.edu	290	131	1041	159	0	172
jiwaji.edu	3865	116	5137	56	9	73
javatpoint.com	1976	581	11785	4435	0	82
manit.ac.in	2694	1942	22917	2913	36	161
dauniv.ac.in	603	188	5061	1727	5	117

Depth-2

TABLE III. URL DETAILS AT DEPTH-2

Url	Media Files	Import Files	Page's Links	Intra Pages	Inactive Pages	Valid Pages
mit.edu	290	131	1041	159	0	172
jiwaji.edu	3865	116	5137	56	9	73
javatpoint.com	1976	581	11785	4435	0	82
manit.ac.in	2694	1942	22917	2913	36	161
dauniv.ac.in	603	188	5061	1727	5	117

VIII. Web Structure Mining Applications

Hyperlinks structure's analysis provides interesting substructures, associations among alteration of substructures, different clusters of substructures based on derived patterns. According to the way these are used in different applications. Web structure mining based on the topology of hyperlink web structure mining will categorizes the web page and generate the information such as similarity and relationship between different web sites. The mainintention to identify documents, which are linked to or linked by many relevant Web pages. Some of web structure mining application's areas is:

- ➢ User Behavior- Understanding how users behave.
- ➢ Trail Information –It is used for keeping track of browsing behavior of users.
- ➢ Filtering Behavior – Separating human and non human Web behavior
- ➢ Ranking metrics –for page selection with a better score.

IX. Conclusion

In this article, we present a preprocessing algorithm for web structure mining, the algorithm firstly extract all links from the page associated with target URL and then construct the Information System use links details, finally, achieved the Information system avoid the affection of redundant data and reserved the original structure of hyperlinks, the Information System can be widely used to the web structure analysis and achieve high performance.

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. S. Tomkins, "The Web as a Graph: Measurements, Models, and Methods," *Lecture Notes in Computer Science Computing and Combinatorics*, pp. 1–17, 1999.

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

[6] 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.

[7] 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.

[8] S. K. Madria, S. S. Bhowmick, W.-K. Ng, and E. P. Lim, “Research Issues in Web Data Mining,” *DataWarehousing and Knowledge Discovery Lecture Notes in Computer Science*, pp. 303–312, 1999.

[9] 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.

[10] M. Thelwall, “Mining the World Wide Web: An Information Search 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 Publishers 2001. 168 pp., ISBN: ISBN: 0 7923 7349 9 £79,” *Journal of Documentation*, vol. 58, no. 2, pp. 232–234, 2002.

[11] https://en.wikipedia.org/wiki/List_of_HTTP_status_codes