A Review for Data Processing in Web Mining Structure by Different Techniques

Ku Nalesh [1], Ghanshyam Sahu [2], Lalit Kumar P Bhaiya [2]

[1]*M. Tech Scholar (CSE), Bharti College of Engineering and Technology, Durg, Chhattisgarh, India.*

[2]*Department of Computer Science & Engineering, Bharti College of Engineering and Technology, Bharti University, Durg, Chhattisgarh, India.*

Corresponding Author: nalesh97@gmail.com

Abstract: - Now a days world wide web is beyond our imagination because it contains extreme large collections of data Which acts as repository of information. If we talk about web, it is horribly expanding the information in world, it provides the information based on the needs of the user. In this paper, we are discussing about the web structure mining which follows the method in several phases like Data extraction, Data Cleaning, Path completion etc. This review paper basically contains the knowledge and work of Knowledge discovery, Knowledge analysis and data preprocessing. Web data and documents is the goal of the emerging research field of web mining, which seeks to solve this issue. We will provide a quick review of Web mining in this work, focusing in particular on methods designed to take advantage of the Web's graph structure for better retrieval and classification accuracy.

Key Words: ***Web Mining, Web Structure Mining, Web Graph, Data Extraction, Data Processing.***

I. INTRODUCTION

In Data mining large number of data sets are stored in order to find patterns and relationships that may be used in data analysis to assist solve business challenges. Enterprises can forecast future trends and make more educated business decisions by data mining techniques and technologies. Since over 70 million pages of web data are added every day. Data collection entails gathering the information needed for analysis [1]. Because the data available on the online is unstructured, varied, and noisy, data pretreatment is regarded as a crucial component of Web structure mining. Now data mining has four stages which helps to achieve the goals i.e., data gathering, data preparation, mining the data & data analysis and interpretation.

Manuscript revised April 27, 2023; accepted April 28, 2023. Date of publication April 30, 2023.

This paper available online at www.ijprse.com

ISSN (Online): 2582-7898; SJIF: 5.59

Data mining is a technique for obtaining valuable information and impressions from numerous datasets. It may also be referred to as knowledge extraction, knowledge mining from data, knowledge discovery process, or data/pattern analysis.

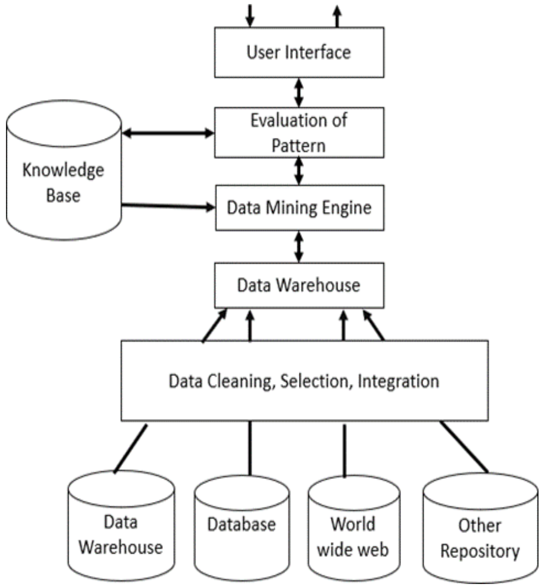

Fig.1. Architecture of Data Mining

II. Related Work

Web Mining: Web mining is used to find unobserved facts or information from net. We can also define it as to search out the helpful patterns contained in the unstructured records is referred as web mining. Web mining combines the two filed in research area: the World Wide Web and data mining. Web mining uses data mining techniques to assess the impressions contained in online material. It aids in improving web-based services. Web mining is also employed in industries like AI, business support services, and for online services, among others.

It is one of the functions of the data mining, web mining is employed for retrieving data from the internet and the internet, and which is the group of technologies to realize the capability. Web mining is a function of data mining approach that finds information from internet data by using at least one of the following: structural (hyperlink) or usage (web log) information (with or without different kind of net information). Web mining has become increasingly popular within both academic and professional circles.[2]

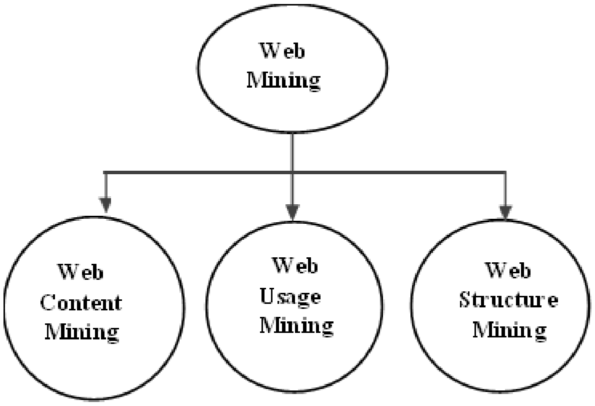

Fig.2. Types of Web Mining

Web Structure Mining: The structure of the hyperlinks within the Web itself presents a barrier for Web structure mining. A long-established field of study is link analysis. Link Mining [3], a recently emerging research field, is situated at the intersection of the work in link analysis, hypertext and web mining, relational learning and inductive logic programming, and graph mining. In contrast to traditional collections of text documents, the Web comprises a variety of items with essentially no common structure and far greater variances in authoring style and content. Web pages and links are the objects found on the Internet.

Data Collection: Gathering the necessary data for analysis is the initial stage in every mining technique. Data collection in web structure mining refers to gathering links from online sites connected to seed Urls on multiple servers.

Preprocessing: Executes a series of processes on a web links file to clean up the data and complete the links while validating, identifying, and completing the links.

Knowledge discovery: It is the process of using numerous data mining techniques, such as statistical explanation, association, clustering, pattern analysis, and similar ones, to the processing of data.

Knowledge analysis: After conclusive information discovery from online links, remove irrelevant data and calculate and provide the intriguing pattern to users. Understanding the structure of links also enables the filtering of useless knowledge.

Web Graph: It shows the link between pages of the world wide web. The documents or pages that make up the Web's nodes can be thought of as edges in a directed labelled graph, and the connections that connect them as nodes. The Web Graph is the name given to this directed graph structure. Two sets, V and E, make up a graph G, according to Horowitz et al [4]. A finite, nonempty set of vertices is the set V. The pair of vertices that make up the set E are known as edges. The sets of vertices and edges of graph G are denoted by the symbols V(G) and E(G), respectively. To depict a graph, G = (V, E) can also be used. Figure 3 shows a directed graph with three vertices and three edges.

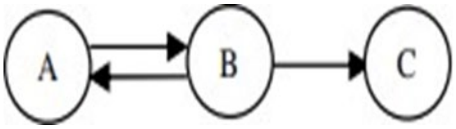

Fig.3. A directed Graph G

The vertices V of G, V(G) = {A, B, C}. The Edges E of G, E(G) ={(A, B), (B, A), (B, C)}. In a directed graph with n vertices, the maximum number of edges is n(n-1). With 3 vertices, the maximum number of edges can be 3(3-1) = 6. In the above example, there is no link from (C, B), (A, C) and (C, A). A directed graph is said to be strongly connected if for every pair of distinct vertices u and v in V(G), there is a directed path from u to v and also from v to u.

One way to visualize a Web is as a sizable graph with a few hundred million or billion nodes (or vertices) and a few billion arcs (or edges). The algorithms used in hyperlink analysis for information retrieval are explained in the section that follows:

Data Extraction: The process of gathering or extracting various forms of data from many sources, many of which may be erratically organized or entirely unstructured, is known as data extraction. Data extraction enables the consolidation, processing, and refinement of data so that it can be kept in a single location and later altered. Both ETL (extract, transform, load) and ELT (extract, load, transform) processes begin with data extraction. ETL and ELT are components of a full data integration strategy.

Data Processing: Preprocessing data converts it into a format that can be processed more quickly and effectively for the user's needs. The primary goal of data preprocessing is to choose standardized data from the original log files so that it may be used by the algorithm that finds user travel patterns Data cleaning, user identification, and session identification are all included in the data preprocessing stage. Display the data in the format required by mining techniques. The data can be represented in a variety of ways, including charts, graphs, etc. The documents are also compared using a variety of criteria to determine how similar they are. Data Preprocessing is classified into four categories:

Data Cleaning: Data cleaning is the initial stage and is critical to preprocessing in order to obtain cleansed data for later processing.

Data Integration: Data integration is the process of combining data from several sources that may use different formats.

Data Transformation: At this step, the obtained data is transformed into a special format that is suitable for mining techniques.

Data Reduction: Using crucial features for mining techniques, this stage aims to reduce altered data.

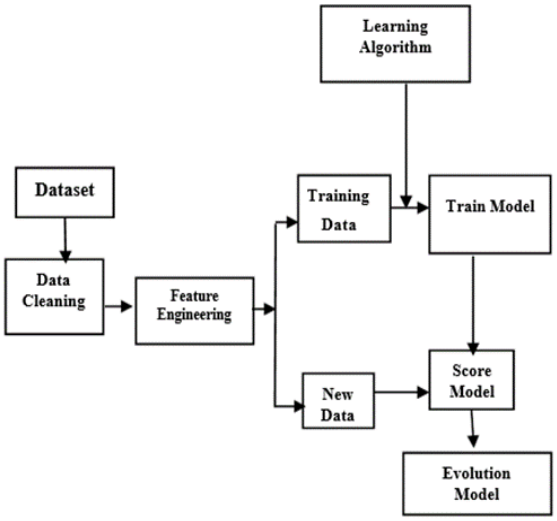

Fig.4. Data Processing

III. Literature Review

Alejandro Peña-Ayala [5] This paper, based on descriptive and predictive models, two patterns of value- instances that represent EDM techniques were found. One important discovery is that the majority of EDM approaches are based on a fundamental set that consists of three different types of educational systems, disciplines tasks, methodologies, and algorithms for each. The review's summary of the studied EDM works and analysis of the EDM come to a close. In a sense, the components of strengths, weaknesses, opportunities, and threats indicate future tasks to be completed.

Liwen Vaughan and Justin You [6] had made research and suggested a technique called keyword enhanced Web structure mining, which combines the concepts of Web structure mining with web content mining. The technique was used to gather information on the business rivalry between a number of DSLAM businesses. DSLAM was specifically incorporated into queries that looked for co-links between groups of business websites. Multidimensional scaling (MDS) was used to map the positions of the company competitors after the co-link matrix analysis. The research demonstrates that the suggested approach outperforms the prior approach of Web structure mining alone by creating a more precise map of commercial competitiveness in the DSLAM industry.

Dr. Rajesh K Shukla, Prachi Sharma, Noopur Samaiya, Monika Kherajani [7] in their research document gives a general overview of web mining techniques. Our review study focuses mostly on online usage mining and the techniques used in it. Additionally, it reviews the use of web usage mining. In web mining, the data is acquired from the server, client, proxy server, or database. Web content mining, web structure mining, and web use mining are three categories of Web mining techniques.

Vijayashri Losarwar, Dr. Madhuri Joshi [8] In their paper the significance of data preparation techniques and the numerous procedures involved in effectively obtaining the essential content are discussed in this study. It is being suggested to preprocess the web log completely in order to extract user patterns. The web log's useless entries are deleted by the data cleaning algorithm, and the log file's irrelevant characteristics are removed by the filtering algorithm. Sessions and users are identified.

Tasawar Hussain, Dr. Sohail Asghar [9] proposed the preprocessing level of web usage mining, a framework for web session clustering has been suggested. The pretreatment processes for preparing the web log data are covered by the framework, which also transforms the web log data's categorical information into numerical information. In order to cluster the web log data, an appropriate similarity and swarm optimization algorithm were performed after obtaining a session vector. According to the author, the present web session methodologies are improved by the hierarchical cluster-based approach, which provides more structured information about user sessions.

Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang [10] The enormous amounts of data that IoT generates or collects are thought to contain extremely important and helpful information. In order to make this type of technology intelligent enough to offer more convenient services and surroundings, data mining will undoubtedly play a crucial role. The IoT is discussed at the outset of this essay. The features of "data from IoT" and "data mining for IoT" are next briefly reviewed. The field's modifications, potentials, unresolved problems, and future developments are then discussed.

IV. Conclusion

In this paper, we discussed and review different research papers which gives the knowledge of web structure mining and different algorithms like data preprocessing that extracts all links from the page associated with the target URL, builds the information system using links details, and then successfully preserves the original hyperlink structure. The information system can be widely used for web structure analysis and achieve high performance. The construction of an appropriate target data set to which data mining and statistical methods can be used is a crucial effort in any application of data mining. Due to the features of click stream data and its relationship to other pertinent data acquired from numerous sources and across multiple channels.

References

[1]. Suvarn Sharma, Research Scholar, “Data Pre-processing Algorithm for Web Structure Mining”, Fifth International Conference on Eco-Friendly Computing and Communication Systems (ICECCS-2016).

[2]. Monika Yadav Mr. Pradeep Mittal, “Web Mining: An Introduction Department of Computer Science and Applications”, International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 3, March 2013 ISSN: 2277 128X.

[3]. L. Getoor, Link Mining: A New Data Mining Challenge. SIGKDD Explorations, vol. 4, issue 2, 2003.

[4]. E. Horowitz, S. Sahni and S. Rajasekaran, “Fundamentals of Computer Algorithms”, Galgotia Publications Pvt. Ltd., pp. 112- 118, 2008.

[5]. Alejandro Peña-Ayala, Educational data mining: A survey and a data mining-based analysis of recent works, Elsevier, Expert Systems with Applications 41 (2014) 1432–1462.

[6]. Liwen Vaughan and Justin You, Keyword Enhanced Web Structure Mining for Business Intelligence, Springer-Verlag Berlin Heidelberg 2009, E. Damiani et al. (Eds.): SITIS 2006, LNCS 4879, pp. 161–168, 2009.

[7]. Dr. Rajesh K Shukla, Prachi Sharma, Noopur Samaiya, Monika Kherajani, “WEB USAGE MINING-A Study of Web data pattern detecting methodologies and its applications in Data Mining”, WEB USAGE MINING-A Study of Web data pattern detecting methodologies and its applications in Data Mining, paper 7.

[8]. Vijayashri Losarwar, Dr. Madhuri Joshi, “Data Preprocessing in Web Usage Mining”, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012.

[9]. Tasawar Hussain, Dr. Sohail Asghar, “Hierarchical Sessionization at Pre-processing Level of WUM Based on Swarm Intelligence”, 6th International Conference on Emerging Technologies (ICET) IEEE, pp. 21-26, 2010.

[10]. Chun-Wei Tsai, Chin-Feng Lai, Ming-Chao Chiang, and Laurence T. Yang, Data Mining for Internet of Things: A Survey, IEEE COMMUNICATIONS SURVEYS & TUTORIALS, VOL. 16, NO. 1, FIRST QUARTER 2014.