CSE-3024 Web Mining

Digital Assignment 1

Alokam Nikhitha (19BCE2555)

G2+TG2 Slot

Question

You have to download 6 to 8 recent journal papers from reputed journal (IEEE, Elsevier, Springer, MDPI, Hindawi etc.). Read out the paper completely and identified the methodology used, pros and cons and scope for future work. Try to find out a core pitfall and find the solution for it.

TOPIC : Web Structure Mining

SL .No	Paper title and Year	Method	Advantage and Limitation
1.	Data Preprocessing Algorithm for Web Structure Mining from year 2017	Web Mining, like Data Mining, comprises four stages: data collection, preprocessing, knowledge discovery, and knowledge analysis. The first two stages of data collection and preprocessing are the focus of this research. The process of collecting the information inorder to analyze is called Data Collection. Because the data available on the web is unstructured, varied, and noisy, data pretreatment is regarded as a key stage of Web Structure mining.	**Advantages:** They presented a preprocessing algorithm for web structure mining; the algorithm first extracts all links from the page associated with the target URL and then constructs the Information System using link details; finally, the Information System avoided the affection of redundant data and preserved the original structure of hyperlinks and the Information System can be widely used to analyse web structure and achieve high performance. **Limitations:** Data preparation issues can take various forms, but the most prevalent are: missing value, manual entry, and data inconsistency. Formats vary per region. Units of measurement.

			Incorrect data types. Manipulation of files A nonymization is lacking.
2.	Data Analytics For Web Structure Mining In Business Website from year 2020	In the model that they have mentioned, websites are taken as nodes and hyperlinks as edges. The hyperlinks will help in connecting a webpage to the other webpages which are associated with it. The Graph theory will be applied to the web structure mining to analyze the connection and node architecture of the webpage. Most popular and also well- liked websites will get more visitors who search for than the others. Each module will be taken from the website by severing the website's structure. Every module on the website is allocated a web address by default.	**Advantages:** The World Wide Web (WWW) has a massive quantity of information, which makes it an appealing prospect for web mining. Web mining methods can examine the access pattern of webpages or websites, the features of documents, and the behaviour of specific consumers. The goal of this study is to increase profit by using online services in commercial areas. They developed algorithmic solutions for the successful execution of their suggested methodology. Their goal is to identify the webpages as widespread or popular by using hit counts and the quantity of adverts on that specific webpage. When the user inputs a certain module, the module count is increased. This will have to make comparison to the highest number of hits in a module to the lowest number of hits in a module. The advertisements coordinator and will then display the ad in module with the high hit rate. This will increase the effectiveness of advertising.

			Limitations: The page ranking algorithm is used to rank the relevant pages by treating all accessible links equally while distributing the webpage's score. In this study, the company domain is used to identify and locate sub URLs relevant to that website, as well as the page rank. The URL linked with which the company domain will have to be extracted using java coding. The hit count for a certain web page is used to determine the popularity of a webpage and the number of advertisement popups on that webpage. But the PR scores have not calculated on the moment of search because that would be too expensive and takes a lot of time.
3.	Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval from year 2019	This paper is about the hyper link analysis which is required in link analysis, how such algorithms compare, and also the vitality/importance of hyperlink analysis in Web searches. The number of incoming links to a website and the number of outgoing connections from that page will be studied in the hyperlink analysis, as well as the dependability of the connecting. Web page authorities and hubs will be investigated. The various forms of lin-analysis algorithms, have	**Advantages:** The primary goal of this study is to investigate the hyperlink structure and grasp the Web graph in a straightforward manner. The World Wide Web (WWW) has a massive quantity of information, which makes it an appealing prospect for web mining. Web mining methods can examine the access pattern of webpages or websites, the features of documents, and the behaviour of specific consumers. This research/paper also discuss about the key methods used for

		been observed .The differences are classified. These algorithms' formulas will be investigated.	hyperlink analysis, and there by exploring and comparing them. **Limitations:** Only few criteria have been considered in the comparison of the various analysis in this paper which makes it not a complete study on Exploring the Hyperlinks and Algorithms . This will be left as the Future work and Scope of this paper.
4.	Comparative study of various Page Ranking Algorithms in Web Structure Mining from year 2012	This study begins with a prologue to Web mining before attempting to define thorough Web Structure mining and offering link evaluation. techniques made available by the Web. Web Mining explains the fundamentals of Web mining as well as the Web mining categories. Different algorithms are presented , and the performance are compared. PageRanks are generated for the PageRank and Weighted lPageRank algorithms in order to maintain consistent lhyperlink composition.	**Advantages:** The World Wide Web (WWW) has a massive quantity of information, which makes it an appealing prospect for web mining. Web mining methods can examine the access pattern of webpages or websites, the features of documents, and the behaviour of specific consumers. The knowledge gathered from the Web may be utilised to improve Web information retrieval, question answering, and Web-based data warehousing performance. **Limitatons:** In this study, they discussed about the Web structure mining, and also the Link mining and block- levellink mining concerns. They also examined two prominent algorithms to have a better understanding of their applicability and efficacy.

			However, this is a vast subject and so much more work to be done. They were unable to cover everything, but this report might be a good beginning point for suggesting areas for further research.
5.	Ranking WebPages Using Web Structure Mining Concepts from year 2013	With the Web's fast expansion, users can quickly become lost in its complex hyperstructure. The major purpose of the proprietors of these websites is to provide useful information to users in order to meet their demands. This method is heavily relies on WSM. In web structuremining,2 page ranking algorithms they are ,PageRank andHyperlink-Induced Topic Search, are extensively utilised. When awarding rank scores, both algorithms consider all connections equally.	**Advantages:** The page lranking algorithm is used to rank the relevant pages by treating all accessible llinks equally while distributing the ⎽web page's score. The data which was left over from the users' earlier behavior is used to find the behavior and suggestions for the user. Web pages that are linked by information or a direct link. It is quite crucial in this strategy. **Limitations:** In this study, they did not cover the performance analysis of PageRank and lHITS, but they are working on techniques to identify users and web pages in order to produce better lPageRank outcomes. This might be regarded a disadvantage or limitation for this article, as well as an area of future work.
6.	A Comparison of Dimensionality	They study the use of several dimensionality reduction methods (DRTs) to uncover the	**Advantages:** In the paper, They have made a comparison on 4 dimensionality

	Reduction Techniques for Web Structure Mining from year 2007.	implicit structures buried in web hyperlink connection in this research. We use and compare four DRTs. Experiments on the datasets allow us to make the following claims, In terms of the stability and interpretability of the discovered structures, MF outperforms PCA and ICA, appears to be unsuitable for this task, and they recommend instead using the more recent Wikipedia dataset, which is suited better.	reduction approach. Because of its superiority over the other approaches, the results demonstrate that nonnegative matrix factorization is a potential strategy for web structure analysis. As a result, we intend to concentrate on this technique by examining two advanced features in greater depth. The first is the NMF algorithm's initialization stage, in which matrices W and H are presently filled with random positive entries. We feel that a decent seed for the NMF allgorithm would improve tlhe process of discovering web structures. **Limitations:** They used the multiplicative update methodology in this study, however there are many alternative approaches in the literature, such as the divergence allgorithm, the alternating least squares method, and the probabilistic approach. They believe that this would be useful to find the efficacy of these variations for web structure mining.
7.	A Review Paper on Web Mining: Web Structure Mining from year 2021	The ubiquitous usage of the Internet has had a significant impact on how we socialise, do business, and make purchases. The growth of the internet has resulted in a massive amount of	**Advantages:** This page's importance can be intentionally increased on the website. To determine the score for a certain lhomepage, the page ranking algorithm employs a

		publicly available data.The vast volume of online data has a variety of useful data patterns and relationships that must be mined and recognisedl. It covered a wide range of research approaches, including statistics, informatics, knowledge discovery, and many more.	random navigation model. l Referrals can occasionally affect the rating of the first page in the page ranking algorithm. The hit algorithm assigns a value to a website depending on its hubs and permissions. However, it is difficult to determine if a given page is a hub or an authority, and it occasionally returns irrelevant connections. With the expansion of the website, the response time grows llinearly. According to the findings of this article's research, in a particular situation, all algorithms return useless links and the search result. **Limitations:** With the advent and advancement of internet mining technology, it is now employed not just in the search engine field, Web mining technology has evolved as the cornerstone for a plethora of new internet technologies, resulting in immeasurable value. Online structure mining most efficiently investigates the associations between web documents by utilising the information communicated by the links in each page. To alleviate the disadvantages highlighted in the finding section, research may be conducted and algorithms can be updated.