**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.** | Dataa Preprocessing Algorithm for Webl Structure Mining from year 2017 | Web Mining, like lData Mining, comprises four stages: data collection, lpreprocessing, knowledge discovery, and knowledge analysis. The first two stages of data collectionf and preprocessingl 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 presentedl a lpreprocessing algorithm for web structure mining; the algorithm first extracts all links| from lthe page associated with the target URL and then| constructs the Information| System using link details; finally, the Information System avoided lthe affection of redundant data and preserved the original| structure of hyperlinks and the lInformation System can be widely used to analyse| web structure and achieve high performance.  **Limitations:**  Data preparation| issues can take various forms,l but the most prevalentl are: lmissing value, manual entry, land data inconsistency.  Formats vary per region.  Unitsl of measurement.  Incorrect data types.  Manipulation of files  A|nonymizationl is lacking. |
| **2.** | Data Analytics lFor Web Structure lMining 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 lthe other webpages which are associated with it. The Graph theory will be applied to the web structure mining to analyzel the connection and node architecture of the webpage. Most popular and also well- lliked websites will get more visitors who search for than the others. lEach module will be taken from the website by severing the website's structure. Every module on lthe 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 lmining methods can examine the access pattern of webpages or websites, the features of documents, and the lbehaviour of specific consumers.  The goall of this study is to increase profit by using online services in commercial areas. They developedl algorithmic solutions for the successful execution of their suggested methodology. lTheir goal is to identify the webpages as widespread or popular by using hit counts and the quantity of adverts on that specific webpage. When lthe user inputs a certain module, the module count is increased. This will have to make comparison to the lhighest number of hits in a module to the lowest number of hits in a module. The lladvertisements lcoordinator and will then display the adl in module with the high hit rate. This will increase the effectiveness of advertising.  **Limitations:**  The page lranking algorithm is used to rank the relevant pages by treating all accessible llinks 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 usedl to determine the popularity of a webpage and the number of advertisement popups on lthat 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.** | Webl lStructure Mining:| Exploring || lHyperlinks and lAlgorithms 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 been observed .The differences are classified. These algorithms' formulas will be investigated. | **Advantages:**  The primary goal of this study is to linvestigate the| lhyperlink istructure 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 lmining methods can examine the access pattern of webpages or websites, the features of documents, and the lbehaviour of specific consumers.  This research/paper also discuss about the key methods used for hyperlink analysis, and there by exploring and comparing them.  **Limitations:**  Only few criteria have been considered in the comparison of the various lanalysis in this paper lwhich makes it not a complete study on lExploring the Hyperlinks and Algorithms . This| will be left as the Future work and Scope of this paper. |
| **4.** | Comparativel study of various Pagel lRanking Algorithms in Web lStructure Miningl from year 2012 | This study begins with a prologue to Web mining before attempting to define thorough Web Structure mining and offering link evaluation.l techniques made available by the Web. Web lMining 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 lmining methods can examine the access pattern of webpages or websites, the features of documents, and the lbehaviour of specific consumers.  The knowledge gathered from the Web may be utilised to improve lWeb information retrieval, question answering, and Web-based data warehousing performance.  **Limitatons:**  In lthis study, they discussed about the Web structure mining, and also the Link mining and block- levellink mining concerns. They also examined ltwo prominent algorithms to have a better understanding of ltheir applicability and efficacy. However, lthis 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 lbeginning point for suggesting areas for further research. |
| **5.** | lRanking WebPages Using Web Structure lMining Concepts  from year 2013 | With the Web's fast expansion, users can quickly become lost in its complex lhyperstructure. The major purpose of the lproprietors of these websites is to provide useful information to users in order to meet their demands. This method is lheavily 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 webpage's scorwweb 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 lComparison of lDimensionality lReduction Techniquesl for Web Structure Mining from year 2007. | They study the use of several dimensionalityl reduction methods (DRTs) to uncover the implicit structures buried in web hyperlink connection in this research. Wel use and compare four DRTs. Experiments on the datasets allow us to make the following claims, In terms of the stability and linterpretability of the discovered structures, MF outperforms PCA and ICA, appears to be unsuitable for this task, and they lrecommend instead using the more recent Wikipedia dataset, which is suited better. | **Advantages:**  In the paper, They have made a comparison on 4 dimensionality reduction approach. Because of its superiority over the other approaches, the results demonstrate that nonnegativel matrix factorization is a potential strategy for web structure analysis. As a result, we lintend to concentrate on this technique by examining two advanced features in greater depth. The first is the lNMF algorithm's linitialization stage, in which matrices W and H are lpresently filled with random positive entries. We feel that a ldecent seed for the NMF allgorithm would improve tlhe process of discovering web structures.  **Limitations:**  They used lthe multiplicative update methodology in this study, however lthere are many alternative approaches in the literature, such as the divergence allgorithm, the alternating lleast squares method, and the probabilistic approach. They believe that this would be useful to find the efficacy of lthese variations for web structure mining. |
| **7.** | A lReview Paper on Web Mining: lWeb 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 publicly available data.The vast volume of online data lhas 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. | **Advantages:**  This page's importance can be intentionally lincreased on the website. To determine the score for a certain lhomepage, the page ranking algorithm employs a 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 lfindings 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 ltechnology, it is now employed not just in the search engine field, Web mining ltechnology 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 linformation communicated by the links in each page. To alleviate the disadvantages highlightedl in the finding section, research may be conducted and algorithms can be updated. |