A Review on Design and Development Of Sequential Patterns Algorithms In Web Usage Mining

V Aruna [a], Dr. Harsh Pratap Singh [b] and Dr. D. Sujatha [c]

[a] *Research Scholar, Dept. of Computer Science & Engineering,*
Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal Indore Road, Madhya Pradesh, India
[b] *Research Guide, Dept. of Computer Science & Engineering,*
Sri Satya Sai University of Technology & Medical Sciences, Sehore, Bhopal Indore Road, Madhya Pradesh, India
[c]*Research Co-Guide,Malla Reddy College of Engineering & Technology*

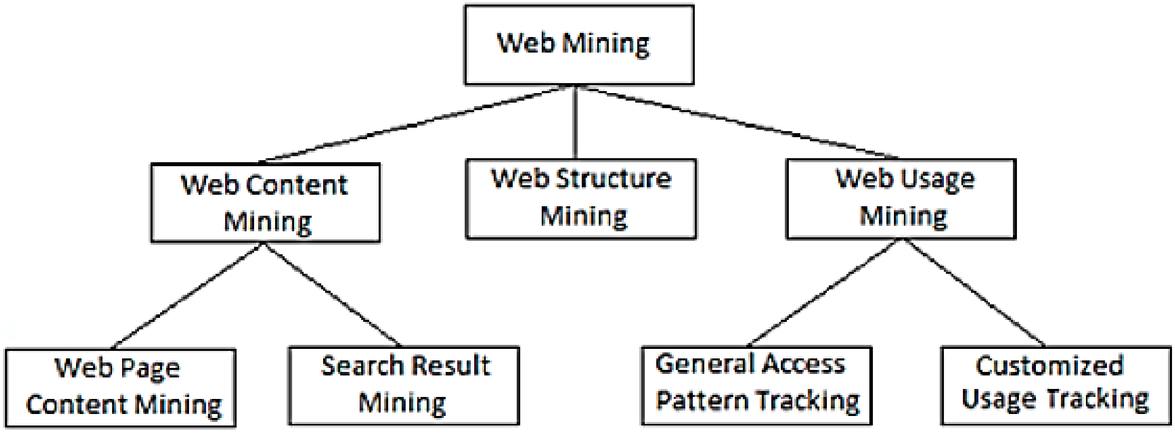
Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: In the recent years with the advancement in technology, a lot of information is available in different formats and extracting the knowledge from that data has become a very difficult task. Due to the vast amount of information available on the web, users are finding it difficult to extract relevant information or create new knowledge using information available on the web. To solve this problem Web mining techniques are used to discover the interesting patterns from the hidden data .Web Usage Mining (WUM), which is one of the subset of Web Mining helps in extracting the hidden knowledge present in the Web log files , in recognizing various interests of web users and also in discovering customer behaviours. Web Usage mining includes different phases of data mining techniques called Data Pre-processing, Pattern Discovery & Pattern Analysis. This paper presents an updated focused survey on various sequential pattern mining algorithms like apriori-based algorithm , Breadth First Search-based strategy, Depth First Search strategy, sequential closed-pattern algorithm and Incremental pattern mining algorithm which are used in Pattern Discovery Phase of WUM. At last , a comparison is done based on the important key features present in these algorithms. This study gives us better understanding of the approaches of sequential pattern mining.

KEYWORDS: Web usage mining, sequential patterns, clustering, patterns summary, Web Access Patterns (WAP), and neural networks.

Introduction

Presently, humans produce and publish more amount of data when compared to past. In fact ,there is more amount of information available on World Wide Web. Users are required to spend more amount of time to identify the hidden information. So, some techniques or methods are required to handle data very effeiciently.Web mining is one of the technique used to analyze and discover the useful information available in Web. Web mining is generally categorized into three types as seen in



Web content mining :Web content mining is the process of extracting information from Web documents. The contents of a web document can be text,image,video , sound or records like lists and tables which are used to convey information to the users. Most of the data which is available on wed is unstructured. Database approach and agent based approach are used in web content mining. This unstructured data from web documents is retrieved by making use of database approach. The agent based approach is used for searching the relevant information and organizing the collected information. The information retrieval view and database view are two different views of web content mining. The content mining from information retrieval view helps the users in filtering and finding the information from the web, where as the database view manages the web data.

Web structure mining :In web structure mining ,links between the documents can be represented as a graph with nodes and edges. Web structure mining deals with the extraction of structural information from the web. The main aim of web Structure Mining is to produce the structural summary of similar websites or similar web pages. This type of mining can be applied either at the level of document or at hyperlink level. The process of extracting the structural information can be further divided into two types that is based on . a) Hyperlinks: Hyperlinks can be used to connect to different location in same web page or to a location present on different web page. A hyperlink is generally divided into two categories i.e. intra-document hyperlink and inter-document hyperlink. Intra-document hyperlink is used to connect different parts of the same page and inter-document hyperlink is used to connect two different pages.

Web usage mining basically deals with finding out what users are looking for on the internet. Web Usage Mining (WUM) is the process of extracting useful information from web log based on users' needs .Web data preprocessing should be done on the huge data present in the web iinorder to get the user needed information . The different phases in web usage mining include data cleaning process, data preparation process, user identification, session identification, data integration, data transformation, pattern discovery and pattern analysis. The data pre-processing is most critical phase in the WUM. The data preprocessing step is applied by taking original data or on the data that is integrated from multiple sources. The purpose of web usage mining is to retrieve the rawdata from web and analyze the pattern after it is discovered. Log files provide information about the activity of user, viz., which web site he/she using, whom you send/receive e-mail etc.

Business web usage mining uses statistical methods to explore. But researches focus on developing knowledge extraction techniques that are used analyze the web usage mining data. Three main methods that are used in web usage mining are : Association rules, sequential patterns, and clustering.

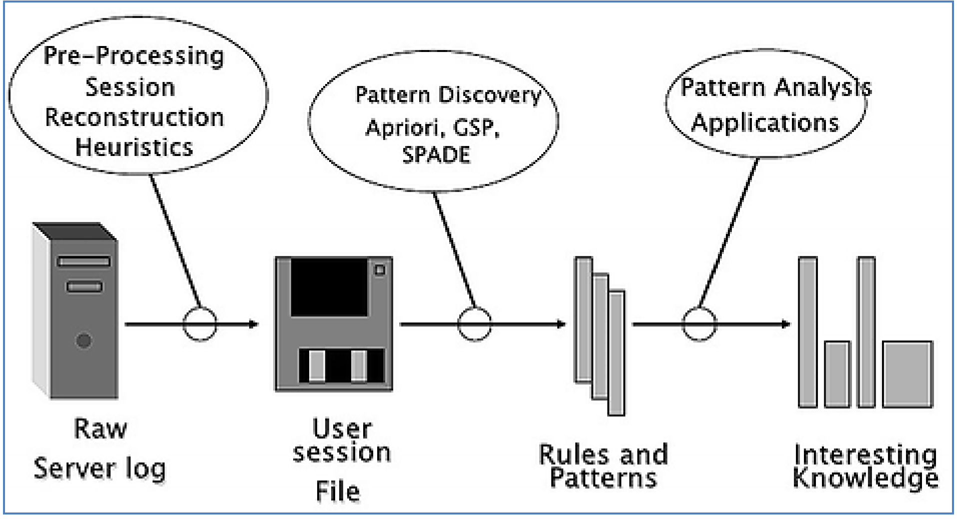


Figure 1.1 Phases of Web Usage Mining

Target data sets for data mining in the context of the web are classified into the following types:

Content data: The data meant to be conveyed to the web user. Naturally, web content mining is the process of extracting knowledge from the content of web documents.

Structure data: The Meta data that define the organization of the web information systems. Web structure mining is the process of inferring knowledge from the structure of data.

Usage data: The data collected from user interactions with the web. As mentioned before, WUM is the process of discovering and interpreting patterns based on the user access from web information system.

Data required for web usage mining is present in Web log file which is created based on the client communications on the web. The web log information is normally introduced in some standard formats, for example, Common Log Format and Extended Log Format indicated by the World Wide Web Community (W3C). The main sort of web log information is the one produced by the Web Server. Web log information mainly contains the IP address of customers, web page accessed and the log in time. The second kind of web log information is created by the Application Server. Information contained in this log will be more explicit, addressing different sorts of business occasions identified with the applications. The third sort of web log is called Application Level Log which is related to single application. The substance of the log is much more explicit than the past kind. Because of the trouble to gathering and getting to web log from different areas, most WUM strategies depend on the Web Server Log file. Web log information can be utilized for business knowledge to improve deals and promotion by providing item proposal. It can distinguish continuous access conduct to improve the general presentation of future access. To improve dormancy time, storing and pre-getting approaches can be made dependent on regular got to pages. Furthermore, web log information can likewise be utilized to improve website plan on the off chance that we know access conduct of clients. At last, personalization for a client can be gotten through WUM.

WUM, from the information mining viewpoint, is the utilization of information mining methods to find valuable information on client conduct from Web information to comprehend and encourage perusing experience for the client. Like most other information mining task, the interaction of WUM contains three principle steps to be specific pre-handling, design revelation and example examination. In the pre-handling step information are gathered, at that point cleaned to eliminate random articles, for example, realistic and mixed media passages. From that point onward, the way toward arranging meetings as per various clients is performed. A meeting is addressed by a bunch of exchanges of a client throughout some undefined time frame as he crosses a web website. The yield of pre-handling stage fills in as the contribution for the example disclosure stage whereby learning calculations are applied to dig for potential fascinating examples which might be installed in the log information. At last, in the example examination stage, tiresome guidelines or examples found from the disclosure stage are distinguished to be precluded. Example investigation strategy is a lot of ward upon the particular application that pre-owned them and one of the more mainstream design examination applications is SQL. Sessionization, the interaction to distinguish the meetings from the crude information, is a significant test, on the grounds that the worker logs don't generally contain all the information required. Additionally, the information must be changed into an appropriate configuration before they can be utilized as the contribution for the mining calculations. When the information are ready for mining, an information mining method that suits the proposed objective will be applied to the information. At last, the aftereffects of the mining calculations will be broke down and deciphered into valuable information which can be utilized to encourage dynamic.

SEQUENTIAL PATTERN

Sequential pattern is a bunch of thing sets organized in arrangement database which happens sequentially with a particular order. A grouping database is a bunch of ordered components or occasions, put away with or without a solid idea of time. Everything set contains a bunch of things which incorporate similar exchange time esteem. While affiliation rules demonstrate intra-exchange connections, sequential patterns represent the relationship between's exchanges. Sequential pattern mining (SPM) is the cycle that separates certain sequential patterns whose help surpasses a predefined negligible help limit. Moreover, sequential pattern mining assists with extricating the arrangements which mirror the most incessant practices in the succession database, which thusly can be interpreted as space information for a few purposes. To lessen the exceptionally enormous number of groupings into the most fascinating sequential patterns and to meet the diverse client necessities, it is essential to utilize a base help which prunes the sequential pattern with no interest. Obviously a higher help of a sequential pattern is preferred for all the more fascinating sequential patterns. Sequential pattern mining is utilized in a few areas. SPM is utilized in business associations to contemplate client practices. Furthermore, SPM is utilized in computational biology to investigate the amino corrosive change patterns. SPM is likewise utilized in the region of web utilization mining to mine few web logs conveyed on different servers. All these represent a test to specialists to find the web the board strategies and compelling extraction of data from the web. Sequential pattern mining algorithms can be ordered into apriori-based, pattern-development, early-pruning, and mixtures of these three methods. Breadth-first search, create and-test, and various outputs of the database, are for the most part key highlights of apriori-based techniques that posture testing issues and upset the exhibition of the algorithms. The apriori-based algorithms are discovered to be excessively slow and have a huge search space; while pattern-development algorithms have been tried widely on mining the web log and discovered to be quick, early-pruning algorithms have had accomplishment with protein arrangements put away in thick databases. The algorithm plays out a Breadth-first search, with the usage of Hash Map data structure in Java, the help checking is tried not to prompt not many outputs of database and helps in extending the database in vertical design just as places of the thing sets are coded. The algorithm by applying the Set tasks brings about database contracting. Consequently the algorithm handles Candidate arrangement pruning by applying the crossing point activity that permits them to prune competitor groupings right off the bat in the mining cycle. With the database contracting the connections among the patterns produced is profoundly expanded.

LITERATURE REVIEW

Girish et al,In the current time of technology, each association utilizing a dynamic cycle of correspondence that changing in intelligent way which is satisfied by their own website facilitated by their own web servers or business web server. This website accumulates data about client access at each time when client interfaces for correspondence for certain assets accessible to them. Furthermore, this makes most straightforward path for website head to look at the data about the client's navigational patterns investigated from the web access logs and can likewise be valuable to analyze the client access conduct about what will be the following page that ought to be given to the client following the patterns. To accomplish this reason sequential pattern procedure can be utilized. Sequential pattern procedure finds incessant succession alluded to as patterns in sequential database. In this investigation we are proposing the algorithm for finding the sequential pattern that could recommend and contrast the created patterns and the base help measure.

ThabetSlimani et al,This paper presents the existing sequential pattern mining algorithms. It presents an ordering investigation of sequential pattern-mining algorithms into five broad classes. First, based on Apriori-based algorithm, second on Breadth First Search-based system, third on Depth First Search technique, fourth on sequential closed-pattern algorithm and five based on incremental pattern mining algorithms. Towards the end, a relative comparision is done based on significant key highlights upheld by different algorithms. This examination gives an upgrade in the comprehension of the methodologies of sequential pattern mining.

DmitriyFradkin et al,while various proficient sequential pattern mining algorithms were created throughout the long term, they can at present take quite a while and produce a colossal number of patterns, a considerable lot of which are excess. These properties are particularly disappointing when the objective of pattern mining is to discover patterns for use as highlights in arrangement issues. In this paper we depict BIDE discriminative, an adjustment of BIDE that utilizations class data for direct mining of predictive sequential patterns. We at that point play out a broad assessment on 9 genuine datasets of the various manners by which the fundamental BIDE-Discriminative can be utilized in genuine multi-class order issues, including 1-versus rest and model-based search tree draws near. The aftereffects of our analyses show that 1-versus rest furnishes a proficient arrangement with great order execution.

PoojaAgrawal et al,the sequential pattern mining on reformist databases is new methodology, in which numerous researchers logically find the sequential patterns in time of interest. Time of interest is a sliding window persistently progressing as the time passes by. As the focal point of sliding window changes, the new items are added to the dataset of interest and out-dated items are taken out from it and become exceptional. When all is said in done, the current proposition don't completely investigate this present reality situation, for example, items related with help in data stream applications, for example, market basket examination. Accordingly mining significant knowledge from upheld incessant items turns into a non-trifling research issue. This paper present the different works done on reformist sequential pattern mining .This paper presents a survey of sequential pattern-mining strategies in the writing. This paper characterizing sequential pattern-mining algorithms based on significant key highlights upheld by the strategies. This arrangement targets comprehension of sequential pattern mining issues, momentum status of gave arrangements, and heading of research here. This paper additionally attempts to give a similar exhibition investigation of large numbers of the key methods.

Philippe Fournier-Viger et al,Finding sudden and valuable patterns in databases is an essential data mining task. Lately, a pattern in data mining has been to plan algorithms for finding patterns in sequential data. Perhaps the most famous data mining tasks on arrangements is sequential pattern mining. It comprises of finding fascinating aftereffect's with regards to a set of arrangements, where the intriguing quality of an aftereffect can be estimated as far as different models, for example, its event recurrence, length, and benefit. Sequential pattern mining has some genuine applications since data is encoded as groupings in numerous fields, for example, bioinformatics, e-learning, market basket examination, text investigation, and webpage click-stream investigation. This paper overviews ongoing examinations on sequential pattern mining and its applications. The objective is to give both a prologue to sequential pattern mining, and a review of late advances and research openings. The paper is partitioned into four principle parts. First, the task of sequential pattern mining is characterized and its applications are investigated. Key ideas and terminology are presented. Also, primary methodologies and procedures to take care of sequential pattern mining issues are presented. Limits of customary sequential pattern mining approaches are likewise featured, and famous varieties of the task of sequential pattern mining are presented. The paper likewise presents research openings and the relationship to other mainstream pattern mining issues. Finally, the paper additionally talks about open-source executions of sequential pattern mining algorithms.

ChintaSomeswara Rao et al, Lately, arising applications presented new requirements for data mining strategies. These imperatives are average of another kind of data. Sequential pattern mining is relevant around there, since numerous kinds of data sets are in a period related arrangement. Other than mining sequential patterns in a solitary measurement, mining multidimensional sequential patterns can give us more educational and helpful patterns. Because of the colossal expansion in data volume and furthermore very enormous search space, productive answers for discovering patterns in multidimensional grouping data are these days vital. Thus, in this paper, we present a multidimensional arrangement model; Simulation tests show great burden adjusting and adaptable and worthy speedup over various data sets and issue sizes.

Chandra Shekhar Rao et al,Sequential pattern mining is a critical data-mining technique for determining time-related conduct in arrangement databases. The data accomplished from sequential pattern mining can be utilized in marketing, clinical records, deals examination, etc. Existing strategies just spotlight on the idea of recurrence due to the supposition that successions' practices don't change after some time. A few effective algorithms for keeping up sequential patterns have been created. , old datasets are erased while some other datasets are refreshed. It is evident time stamp as a significant characteristic of each dataset, additionally it is significant during the time spent data mining and it can gives us more precise and valuable data. Despite the fact that there have been numerous new examinations on the sequential patterns in static database. In any case, the multifaceted nature of sequential pattern mining is while expanding the data in dynamically, as time passes by new data sets are embedded.

Navin Kumar Tyagi et al,Web utilization Mining is a region of web mining which manages the extraction of intriguing knowledge from logging data created by web server. Distinctive data mining methods can be applied on web utilization data to separate client access patterns and this knowledge can be utilized in assortment of uses, for example, framework improvement, web website adjustment, business insight and so on Web utilization mining requires data reflection for pattern revelation. This data reflection is accomplished through data preprocessing. In this paper we study about the data preprocessing exercises like data cleaning, data decrease and related algorithms.

Masseglia et al,in this paper we consider the issue of the incremental mining of sequential patterns when new exchanges or new clients are added to a unique database. We present another algorithm for mining continuous arrangements that utilizes data gathered during a previous mining cycle to reduce down the expense of finding new sequential patterns in the refreshed database. Our test shows that the algorithm performs fundamentally quicker than the gullible methodology of mining all in all refreshed database without any preparation. The thing that matters is articulated to such an extent that this algorithm could likewise be helpful for mining sequential patterns, since much of the time it is quicker to apply our algorithm than to mine sequential patterns utilizing a standard algorithm, by breaking down the database into a unique database in addition to an augmentation.

Shuting Yan et al,During the time spent incremental mining, when the help is changed, the capacity structure in existed incremental mining algorithms of sequential patterns confirms that the algorithms need to mine the database indeed. In this paper, we propose a novel data stockpiling structure, called regular succession tree, and give the development algorithm of incessant arrangement tree, called Con_FST. The root hub of the continuous grouping tree stores the incessant succession tree uphold edge and the way from the root hub to any leaf hub represents a sequential pattern in the database. Successive grouping tree stores all the sequential patterns with its help that meet the continuous arrangement tree uphold limit, so when the help is changed, the algorithm which uses incessant grouping tree as the capacity design can locate all the sequential patterns without mining the database by and by. A pruning methodology is proposed to improve the development algorithm. Trials show that the incremental mining algorithm of sequential patterns which utilizes the successive arrangement tree as the capacity structure outflanks Prefix Span in space cost.

CONCLUSION

This paper gives an insight into sequential pattern mining algorithms like Apriori, DFS, BFS, closed sequential pattern, and incremental pattern based algorithms. Apriori based algorithm was the first classical sequential pattern mining algorithm but this generates more number of frequent sub sequences , which is expensive both in time and space. Breadth first search allows level by-level search to be conducted to find the complete set of patterns .closed sequential pattern mining algorithm need not mine all frequent patterns but only the closed ones as it reduces the number of frequent sub sequences. In sequential pattern mining, incremental algorithm can be used for the mining incremental database updates (insertions and deletions).This paper also presents a comparative analysis of these algorithms based on some key features..

REFERENCES

1. Shuting Yan,Jiaxin Liu, JiadongRen“The Design of Frequent Sequence Tree in Incremental Mining of Sequential Patterns” (2011)
2. F. Masseglia, P. Poncelet, M. Teisseire “Incremental Mining of Sequential Patterns in Large Databases” (2014)
3. [3]Navin Kumar Tyagi, A.K. Solanki& Sanjay Tyagi“An Algorithmic Approach To Data Preprocessing In Web Usage Mining” (2010)
4. PoojaAgrawal, Suresh kashyap, Vikas Chandra Pandey, “An Analytical Study on Sequential Pattern Mining with Progressive Database” (2013)
5. DmitriyFradkin, Fabian Morchen “Mining Sequential Patterns for Classification” (2015)
6. Dr.Girish S. Katkar “Web Usage Mining for Comparing User Access Behaviour using Sequential Pattern” (2015)
7. Jiaxin Liu, “The design of storage structure for sequence in incremental sequential patterns mining,” Networked Computing and Advanced Information Management (NCM), pp. 330 - 334, 2010
8. Philippe Fournier,Viger,Roger Nkambou and Vincent Shin-Mu Tseng, “Rule Growth: Mining Sequential Rules Common to Several Sequences by Pattern-Growth,” Symposium on Applied Computing, pp . 951-960, 2011
9. Rathod K. and Valera, “A Novel Approach of Mining Frequent Sequential Pattern from Customized Web Log Preprocessing”, International Journal of Engineering Research and Applications, 2013
10. Kesavan, S., Saravana Kumar, E., Kumar, A., & Vengatesan, K. (2019). An investigation on adaptive HTTP media streaming Quality-of-Experience (QoE) and agility using cloud media servicesInternational Journal of Computers and Applications, 1–14.
11. Tyagi N., Tyagi S. and Solanki A., “An Algorithmic approach to data preprocessing in Web usage mining”, International Journal of Information Technology and Knowledge Management, 2010

12. J. X. Liu, S. T. Yan, Y. Y. Wang and J. D. Ren, “Incremental mining algorithm of sequential patterns based on sequence tree,” Proc. 3th Int. Colloquium on Computing, Communication, Control, and Management, Yangzhou IEEE Press, pp. 1–4, August 2010
13. Kumar, A., Vengatesan, K., Rajesh, R., Parthibhan, M., & Singhal, A. (2018). Review of Gene Subset Selection using Modified K-Nearest Neighbor Clustering Algorithm. In 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 570–574).