RESEARCH REVIEW JOURNALS
Volume-04
Issue-05
May-2019

ISSN: 2455-3085 (Online)
RESEARCH REVIEW International Journal of Multidisciplinary
www.rrjournals.com[UGC Listed Journal]

Algorithms and Issues of Sequential Pattern Mining

[1]Shaziya Islam and [2]Dr. Yashpal Singh

[1]*Research Scholar, Sri SatyaSai University of Technology*
[2]*Bundelkhand Institute of Engineering& Technology, Jhansi(U.P.)*

ARTICLE DETAILS

Article History
Published Online: 25 May 2019

Keywords
Sequential, pattern, mining

ABSTRACT

The sequential pattern mining on progressive databases is very new approach, in which many researchersfind the sequential patterns. Period of interest is a sliding window continuously advancing as the time goes by. Sequential Web page Access pattern mining has been a focused theme in data mining research for over a decade with wide range of applications. The aim of discovering frequent sequential access (usage) patterns in Web log data is to obtain information about the navigational behavior of the users. This can be used for advertising purposes, for creating dynamic user profiles etc. As the focus of sliding window changes, the new items are added to the dataset of interest and obsolete items are removed from it and become up to date. In general, the existing proposals do not fully explore the real world scenario, such as items associated with support in data stream applications such as market basket analysis. Thus mining important knowledge from supported frequent items becomes a non trivial research issue. This paper presents the various works done on progressive sequential pattern mining. This paper classifying sequential pattern-mining algorithms based on important key features supported by the techniques. This classification aims at understanding of sequential pattern-mining problems, current status of provided solutions.

1. Introduction

Data mining is the way toward extracting fascinating (non-paltry, verifiable, beforehand obscure and possibly helpful) data or examples from vast data vaults, for example, relational database, data warehouses, XML repository and so on. Likewise data mining is known as one of the center procedures of Knowledge Discovery in Database (KDD).Usually there are three procedures in KDD. One is called pre processing, which incorporates data cleaning, integration, selection and transformation. The principle procedure of KDD is the data mining process, in this procedure diverse calculations are connected to create shrouded knowledge. After that takesanother procedure called post processing, which assesses the mining result as per client's prerequisites and area knowledge? With respect to assessment results, the knowledge can be exhibited if the outcome is satisfactory, else we need to run a few or those procedures again until the point that we get the satisfactory outcome. Different data mining strategies are connected to the data source; distinctive knowledge turns out as the mining result. That knowledge is assessed by specific standards, for example, the space knowledge or ideas. After we get the knowledge, the last advance is to imagine the outcomes. They can be shown as crude data, tables, choice trees, rules, outlines, data cubs or 3D graphics.

In this universe of Information Technology, Every day we need to experience a few sorts of information that we require and what we do? Today, web is assuming such a fundamental job in our regular daily existence that it is exceptionally hard to make do without it. Furthermore, survival of abundant data in the network and the changing and heterogeneous nature of the web, web searching has turned into a dubious method for most of the clients. Over the most recent fifteen years, the development in number of web locales and guests to those web destinations has expanded exponentially. To mine the fascinating data from this tremendous pool, data mining strategies can be connected. Yet, the web data is unstructured or semi organized. So we cannot make a difference the data mining strategies straightforwardly. Or maybe another order is advanced called web mining which can be connected to web data. Web mining is utilized to find intrigue designs which can be connected to numerous true issues like enhancing web destinations, better understanding the guest's conduct, item proposal and so forth.

The web data is:

1. **Content:** The noticeable data in the Web pages or the information which was intended to be conferred to the clients. A noteworthy piece of it incorporates content and graphics (images).
2. **Structure:** Data which depicts the association of the website. It is isolated into two kinds. Intrapage structure information incorporates the course of action of different HTML or XML labels inside a given page. The primary sort of between page structure information is the hyper-links utilized for site navigation.
3. **Usage:** Data that portrays the usage examples of Web pages, for example, IP addresses, page references, and the date and time of gets to and different other information relying upon the log format

Data is gathered in web server when client gets to the web and may be spoken to in standard formats. The log format of the record is Common log formats, which comprises traits like IP address, get to date and time, ask for strategy (GET or POST), URL of page accessed, transfer protocol, achievement return code and so forth. So as to find get to pattern, preprocessing is essential; since raw data originating from the web server is deficient and just couple of fields is accessible for pattern discovery.

RRIJM 2015, All Rights Reserved

❖ **Types of Mining**

There are two classes of data mining descriptive and prescriptive. Descriptive mining is to condense or portray general properties of data in data storehouse, while prescriptive mining is to perform inference on current data, to make forecasts dependent on the authentic data. There are different kinds of data mining strategies, for example, association rules, classifications and clustering. In view of those procedures web mining and sequential pattern mining are additionally very much explored. Association rule mining, a standout amongst the most imperative and all around explored strategies of data mining. It expects to separate intriguing connections, frequent patterns, associations or easygoing structures among sets of things in the transaction databases or other data stores.

2. Concept Of Sequential Pattern Mining

The idea of sequence Data Mining was first presented by Rakesh Agrawal and RamakrishnanSrikant in the year 1995. The issue was first presented with regards to showcase examination. It expected to recover frequent patterns in the sequences of items obtained by clients through time requested transactions. Later on its application was stretched out to complex applications like media transmission, network recognition, DNA explore, and so forth. A few algorithms were proposed. The specific originally was Apriori algorithm, which was advanced by the originators themselves. Later progressively scalable algorithms for complex applications were produced. E.g. GSP, Spade, PrefixSpan and so forth. The zone experienced extensive headways since its presentation in a limited capacity to focus. Sequential pattern mining is a critical data mining issue, which identifies frequent sub sequences in a sequence database. The significant procedures for sequential pattern mining are

- Aprioribased Approaches
- ✓ GSP
- ✓ SPADE
- Pattern-Growth-based Approaches
- ✓ FreeSpan
- ✓ PrefixSpan

Data are changing constantly; particularly data on the web are profoundly powerful. As time goes, new datasets areembedded; old datasets are deleted while some different datasets are refreshed. It is perceptible that time stamp is a vital trait of each dataset, likewise it is essential during the time spent data mining and it can give us increasingly precise and helpful information. A database comprises of sequences of qualities or occasions that change with time is known as a time-series database, a time-series database records the legitimate time of each dataset. Time-series database is broadly used to store historical data in an assorted variety of regions, for example, financial data, medical data, and scientifically data, etc. Diverse mining methods have been intended for mining time-series data; fundamentally there are four sorts of patterns we can get from different kinds of time-series data:

A. Trend analysis: Pattern investigation is to discover the advancement patterns of characteristics after some time; they can be long-term trend developments, cyclic developments or varieties, seasonal movements and irregular/random movements.

B. Similarity search: Closeness search endeavors to discover sequences that vary just somewhat. Closeness searching is a hazy coordinating procedure that can endure a few contrasts inside a specific limit. In light of the length of sequences we are endeavoring to coordinate, sequence coordinating can be named: subsequence matching and whole sequence matching.

C. Sequential patterns: Sequential pattern mining is endeavoring to discover the relationships between events of sequential occasions, to discover if there exists an explicit request of the events. We can locate the sequential patterns of explicit individual things; additionally we can locate the sequential patterns cross diverse things. Sequential pattern mining is broadly utilized in investigating of DNA sequence.

D. Periodical patterns: Periodical patterns are those recurring patterns in the time series database; periodicity can be every day, week by week, month to month, regular or yearly. Clearly, periodical patterns mining can be seen as sequential pattern mining by accepting the periodical sequences as a set of sequences. Sequential database is an uncommon instance of time series database, thusly most researches in sequential pattern mining center around two fundamental issues. The main issue is sequential pattern mining, going for finding the frequently happened sequences to portray the data or anticipate future data. The second issue is periodical pattern mining, which can be seen as sequential pattern mining.

3. Classification of sequential pattern-mining algorithms

A Classification of existing sequential pattern-mining algorithms is given, which records the algorithms, appearing comparative analysis of their diverse critical highlights. This proposed classification is made out of three fundamental classifications of sequential pattern-mining algorithms, specifically, apriori-based, pattern-development and Early Pruning algorithms. These highlights are a consequence of watchful examination of the surveyed algorithms and speak to a superset of highlights for the most part talked about in the literature.

Frequent sequential pattern discovery can basically be thought of as association rule discovery over a worldly database. While association rule discovery covers just intra-transaction patterns (itemsets), sequential pattern mining likewise finds intra-transaction patterns (sequences), where requesting of things and itemsets is imperative, with the end goal that the nearness of a set of things is trailed by another thing in a time-requested set of sessions or transactions. The set of every single frequent sequence is a superset of the set of frequent itemsets. Because of this likeness, the prior sequential pattern-mining algorithms were gotten from association rule mining procedures. The first of such sequential pattern-mining algorithms is the Apriori. All algorithms, got from the Apriori algorithm. An algorithm can fall into at least one (crossover algorithms) of the classifications in the proposed scientific

RRIJM 2015, All Rights Reserved

categorization. Algorithms for the most part vary in two different ways:

(1) The manner by which candidate sequences are produced and stored. The fundamental objective here is to limit the quantity of candidate sequences produced in order to limit I/O cost.

(2) The manner by which support is tallied and how candidate sequences are tried for recurrence. The key system here is to wipe out any database or data structure that must be kept up all the time for support of checking purposes as it were. The data structures used to store candidate sequences have likewise been a research point and an imperative heuristic for memory use.

4. Sequential pattern-mining algorithms

4.1 Apriori-Based Techniques

a) **AprioriAll** .AprioriAll filters the database a few times to discover frequent itemsets of size k at each kth-cycle (beginning from k = 2). It additionally has the generate-and-test include by playing out the Apriori-generate join procedure to join Lk−1 with itself to generate Ck, the set of candidate sequences in the kth-emphasis, it at that point prunes

The symbol "-" implies an algorithm crashes with the parameters gave, and memory usage couldn't be estimated. Sequences in Ck which have subsequences not in Lk−1 (i.e., are not vast), makes Lk by including all sequences from Ck with support ≥ min sup until there are no more candidate sequences. The Apriori All procedure experiences expanded deferrals in mining as the quantity of sequences in the database gets bigger. It additionally experiences exponential growth of candidate sequences amid execution. A few arrangements were displayed, including hashing to diminish the span of the candidate sequences; transaction decrease; database apportioning; inspecting; and dynamic itemset tallying.

b) **GSP**- The GSP algorithm, running on database D embraces a multiple-pass candidate generate-and-test technique for finding sequential patterns. The GSP-join, similar to the apriori-generate join necessitates that two sequences in Lk join. The prune stage deletes candidate sequences that have a contiguous (k − 1)- subsequence with support not as much as min sup. Every candidate sequence has one more thing than a seed k-sequence, so all candidate sequences are a similar number of things at each dimension. Support for these candidate sequences is again found amid a pass over the data. The algorithm ends when no new sequential pattern is found in a pass, or no candidate sequence can be generated. For expanded proficiency, GSP utilizes a hash-tree to lessen the quantity of candidates in C that are checked for sequences. Note that, at each progression, GSP just keeps up in memory the officially found patterns and the k-candidates, subsequently making it not a memory-just algorithm. GSP is accounted for to be 2 to 20 times quicker than AprioriAll.

c) **PSP -**This is another apriori-based algorithm, additionally worked around GSP, yet the distinction of utilizing a prefix-tree. In PSP, client get to sequences are arranged in the web log as per the IP address. The client is permitted to give a time period _t by which get to sequences that are transiently near one another are gathered. A prefix-tree is then worked to deal with the mining procedure in a path like GSP. Following up on tree traversal, diagram traversal mining was proposed by Nanopoulos and Manolopoulos [2000], which utilizes a basic un-weighted chart to mirror the relationship between pages of websites. The algorithm is like apriori, without playing out the Apriori-generate join. The database still must be checked a few times, yet it is more productive than GSP.

d) **SPAM -**SPAM coordinates the thoughts of GSP, SPADE, and FreeSpan. The whole algorithm with its data structures fits in principle memory, and is professed to be the primary technique for mining sequential patterns to cross the lexicographical sequence tree inside and out first form. SPAM crosses the sequence tree top to bottom first search way and checks the support of each sequence-expanded or itemset-broadened youngster against min sup recursively. On the off chance that the support of a specific kid s is not as much as min sup, there is no compelling reason to rehash profundity first search on s by the Apriori property. Apriori-based pruning is additionally connected at every S-Step and I-Step of the algorithm, limiting the quantity of childrennodes and ensuring that all nodes relating to frequent sequences are visited.

Features for Apriori-Based Algorithms

- ✓ Breadth-first search
- ✓ Generate-and-test
- ✓ Multiple outputs of the database

4.2 Pattern-Growth Techniques

Not long after the apriori-based strategies for the mid-1990s, the pattern growth-technique developed in the mid 2000s, as an answer for the issue of generate-and-test. The key thought is to maintain a strategic distance from the candidate generation step by and large, and to concentrate the search on a limited segment of the underlying database. The search space apportioning highlight assumes a critical job in pattern-growth. Pretty much every pattern-growth algorithm begins by building a portrayal of the database to be mined, at that point proposes an approach to parcel the search space, and generates as few candidate sequences as conceivable by developing on the effectively mined frequent sequences, and applying the apriori property as the search space is being navigated recursively searching for frequent sequences. PrefixSpan depends on recursively developing the patterns by developing on the prefix, and at the same time, confining the search to anticipated databases. Thusly, the search space is decreased at each progression, taking into account better execution within the sight of little support limits. PrefixSpan is as yet considered a benchmark and one of the quickest sequential mining algorithms close by SPADE. Another algorithm, WAP-mine is the first of the pattern-growth algorithms to utilize a physical tree structure as a portrayal of the sequence database alongside support checks, and after that to dig this tree for frequent sequences as opposed to examining the entire sequence database in each progression.

a) **FreeSpan**FreeSpan represents Frequent Pattern-Projected Sequential Pattern Mining, and begins by making a rundown of frequent 1-sequences from the sequence database called the frequent thing list (f-show), it at that point builds a lower triangular grid of the things in this rundown. This framework contains information about the support check of each 2-sequence candidate sequence that can be generated utilizing things in the f-list, and is called S-Matrix

b) **PrefixSpan** .PrefixSpan inspects just the prefix subsequences and ventures just their corresponding postfix

RRIJM 2015, All Rights Reserved

subsequences into anticipated databases. Along these lines, sequential patterns are become in each anticipated database by investigating just local frequent sequences. The key preferred standpoint of PrefixSpan is that it doesn't generate any candidates. It just tallies the frequency of local things. It uses a gap and-vanquish system by making subsets of sequential patterns (i.e., anticipated databases) that can be additionally separated when important. PrefixSpan performs much superior to both GSP and FreeSpan. The real expense of PrefixSpan is the development of anticipated databases.

c) **WAP-mine**Pattern-Growth Miner with Tree Projection. In the meantime as FreeSpan and PrefixSpan in 2000/2001, another real commitment was made as a pattern growth and tree structure-mining procedure, that is, is the WAP-mine algorithm with its WAP-tree structure. Here the sequence database is checked just twice to fabricate the WAP-tree from frequent sequences alongside their support; a "header table" is kept up to point at the principal event for everything in a frequent itemset, which is later followed threadedly to dig the tree for frequent sequences, expanding on the postfix. The primary sweep of the database finds frequent sequences and the second output fabricates the WAP-tree with just frequent subsequences.

d) **FS-Miner**Roused by FP-tree and ISM, FS-Miner is a tree projection pattern growth algorithm that takes after WAP-mine and supports incremental and interactive mining. The hugeness of FS-Miner is that it begins mining quickly with 2-subsequences from the second (which is likewise the last sweep) of the database (at k = 2). It can do as such because of the compacted portrayal in the FS-tree, which uses a header table of edges (alluded to as connections in El-Sayed at el. [2004] as opposed to single nodes and things, contrasted with WAP-tree and PLWAP-tree. It is additionally viewed as a variety of a trie, as it stores support consider in nodes well as edges of the tree that speaks to 2-sequences and is required for the steady mining procedure.

Features of pattern growth-based algorithm

- ✓ Search space apportioning
- ✓ Tree projection
- ✓ Depth-first traversal
- ✓ Candidate sequence pruning

4.3 Hybrid Algorithms

A few algorithms consolidate a few highlights that are attributes of more than one of the three classes in the proposed scientific categorization. For instance, PLWAP joins tree projection and prefix growth highlights from pattern-growth classification with a position-coded include from the early-pruning class. These highlights are key qualities of their individual classifications, so we consider PLWAP as a pattern-growth/early-pruning hybrid algorithm. The capacity of a few algorithms to use a wide scope of proficient highlights gives them an edge over different algorithms. It is additionally critical to make reference to here that a few highlights can't be consolidated into one procedure, such as, "multiple sweeps of the database" and "tree projection;" since "tree projection" is utilized as an option in-memory database as does "support counting avoidance," it can't be joined with "multiple outputs of the database".

a) **SPADE** – Apriori-Based and Pattern-Growth Hybrid Miner SPADE is a noteworthy commitment to the writing, is as yet thought to be one of the benchmark sequential pattern-mining algorithms. It depends on the Lattice theory [Davey and Priestley 1990] to generate candidate sequences. This thought is obtained from the gradual sequential mining algorithm ISM presented before by Parthasarathy [1999]. SPADE finds sequences of subsets of things, not simply single thing sequences similar to the case with Apriori; it additionally finds sequences with discretionary time holes among things, and not simply back to back subsequences. SPADE depends on a lattice of frequent sequences generated by applying lattice theory on frequent sequences and their subsequences. SPADE works for the most part in three stages, first finding frequent 1-sequences in an apriori-like way; and second, frequent 2-sequences. The third step crosses the lattice for support checking and identification of frequent sequences. The lattice can be crossed in either expansiveness first or profundity first search.

b) **PLWAP**– Pattern-Growth and Early-Pruning Hybrid Miner PLWAP uses a binary code task algorithm to build a preordered position-coded connected WAP-tree, where every node is allocated a binary code utilized amid mining to figure out which sequences are the postfix sequences of the last occasion and to locate the following prefix for a mined addition without remaking moderate WAP-trees.

5. Issues In Sequential Pattern Mining

All in all, there are two primary research issues in sequential pattern mining.

i. Improve The Efficiency In Sequential Pattern Mining Process

As indicated by past research done in the field of sequential pattern mining, Sequential Pattern Mining Algorithms for the most part vary in two different ways: (1) the manner by which candidate sequences are generated and stored. The fundamental objective here is to limit the quantity of candidate sequences generated in order to limit I/O cost. (2) The manner by which support is checked and how candidate sequences are tried for frequency. The key system here is to dispose of any database or data structure that must be kept up all the time for support of counting purposes as it were.

ii. Extend the mining of sequential pattern to other time-related patterns.

Sequential pattern mining has been seriously considered amid late years; there exists an incredible decent variety of algorithms for sequential pattern mining. Alongside that Motivated by the potential applications for the sequential patterns, various augmentations of the underlying definition have been proposed which might be identified with different kinds of time-related patterns or to the expansion of time requirements. A few expansions of those algorithms for exceptional purposes, for example, multidimensional, shut, time interim, and requirement based sequential pattern mining are.

- Multidimensional Sequential Pattern Mining
- Discovering Constraint Based Sequential Pattern
- Discovering Time-interim Sequential Pattern
- Closed Sequential Pattern Mining

RRIJM 2015, All Rights Reserved

6. Conclusion

The paper show the classification of sequential pattern-mining algorithms, and demonstrates that present algorithms in the territory can be grouped into three principle classes, specifically, apriori-based, pattern-growth, and early-pruning with a fourth class as a hybrid of the primary three. A careful dialog of 13 characteristic features of the four classes of algorithms, with an examination of the distinctive strategies and systems, is displayed. This audit of sequential pattern-mining algorithms in demonstrates that the imperative heuristics utilized incorporate the accompanying: utilizing ideally measured data structure portrayals of the sequence database; early pruning of candidate sequences; systems to decrease support tallying; and keeping up a restricted search space. The journey for finding a solid sequential pattern-mining algorithm should think about these focuses.

The accompanying necessities ought to be considered for a reliable sequential pattern-mining algorithm. Initial, a strategy must generate a search space that is as little as would be prudent. Highlights that permit this incorporate early candidate sequence pruning and search space parceling. Testing of the database and lossy pressure (i.e., compact portrayal) can likewise be utilized to generate a littler search space. Second, it is essential to limit the search procedure inside the search space. An algorithm can have a restricted search procedure, for example, profundity first search. Third, strategies other than tree projection ought to be examined for finding solid sequential pattern-mining systems. After that this paper additionally depicts the work done on sequential pattern mining with dynamic database which is the current research region.

References

[1]. Antunes, C. And Oliveira, A. L. (2004) - Sequential pattern mining algorithms: Trade-offs between speed and memory. In *Proceedings ofthe Workshop on Mining Graphs, Trees and Sequences* (*MGTSECML/ PKDD '04*).

[2]. Chiu, D.-Y., Wu, Y.-H., and Chen, A. L. P. (2004) - An efficient algorithm for mining frequent sequences by a new strategy without support counting. In *Proceedings of the 20th International Conference on Data Engineering*. 375–386.

[3]. Ezeife, C. I. And LU, Y. (2005) - Mining web log sequential patterns with position coded pre-order linked WAP-tree. *Int. J. Data MiningKnowl. Discovery 10,* 5–38.

[4]. Ezeife, C. I., Lu, Y., And Liu, Y. (2005) - PLWAP sequential mining: Open source code. *In Proceedings of the 1st International Workshop onOpen Source Data Mining: Frequent Pattern Mining Implementation* (*SIGKDD*), ACM, New York, 26–35.

[5]. Goethals, B. (2005) - Frequent set mining. In *The Data Mining and Knowledge Discovery Handbook,* O. Maimon and L. Rokach Eds., Springer, Berlin, 377–397.

[6]. Masseglia, F., Teisseire, M., and Poncelet, P. (2005) - Sequential pattern mining: A survey on issues and approaches. In *Encyclopedia of Data Warehousing and Mining*,1–14.

[7]. Song, S., Hu, H., and Jin, S. (2005) - HVSM: A new sequential pattern mining algorithm using bitmap representation. In *Advanced DataMining and Applications.* Lecture Notes in Computer Science, vol. 3584, Springer, Berlin, 455–463

[8]. Iv´Ancsy, R. And Vajk, I. (2006) - Frequent pattern mining in web log data. *ActaPolytech. Hungarica 3*, 1,- 77–90.

[9]. JIN, X. (2006) –"*Task*-oriented modeling for the discovery of web user navigational patterns", Ph.D. dissertation, School of Computer Science. DePaul University, Chicago, IL.LIU, B. 2007. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Springer, Berlin.LU, Y. AND EZEIFE, C. I..

[10]. Navin Kumar Tyagi, A.K. Solanki & Sanjay Tyagi (2010) - "An Algorithmic approach to data preprocessing in Web usage mining", International Journal of Information Technology and Knowledge Management July-December, Volume 2, No. 2, pp. 279-283

[11]. J. Vellingiri and S. ChenthurPandian (2011) - "A Novel Technique for Web Log mining with Better Data Cleaning and Transaction Identification", Journal of Computer Science7 (5): 683-689,2011 ISSN 1549- 3636 © 2011 Science Publications

[12]. Sachinyele, Beerendra Kumar, NitinNamdev, Devilal Birla, KamleshPatidar.(2011) - "Web Usage Mining for Pattern Discovery", International Journal of Advanced Engineering & Applications, January.

RRIJM 2015, All Rights Reserved