# A REVIEW: ALGORITHMS FOR MINING FREQUENT CLOSED ITEMSET AND CLOSED SEQUENTIAL PATTERNS



| Sonamdeep Kaur | Mrs. Sarika Chaudhary | Mrs. Neha Bishnoi |
|---|---|---|
| M.Tech (C.S.E) | Assistant Professor | Assistant Professor |
| Amity University, Haryana | Amity University, Haryana | Amity University, Haryana |




## ABSTRACT

*Sequence pattern mining finds novel and potentially useful patterns amongst a large database. In this review we have studied different mining algorithms for "closed itemsets" basically divided in two broad categories: Frequent Closed Item set Mining and Closed Sequential Pattern Mining. We have studied the advantages of different algorithms over others on parameters like memory and time consumption of these algorithms. Algorithms are compared and the results were tabulated on the basis of parameters like original author, year of publication, type, advantages and disadvantages.*


## INTRODUCTION

Data mining is defined as the analysis step of the Knowledge Discovery in Databases process, or KDD, as a branch of computer science, which is interdisciplinary and is the computational process of getting patterns in large data sets involving tasks at the intersection of synthetic intelligence, machine education, data, and database systems. The in general goal of the data mining method is to haul out in rank from a data set and convert it into an logical structure for further use. Aside from the raw study step, it involves database and data organization aspects, data pre-analysis, model and inference considerations, interestingness metrics complexity considerations, post-processing of discovered structures, visualization, and online updating.

### Frequent Itemset Mining

Frequent item set mining is an interesting branch of data mining that focuses on looking at sequences of actions or events, for example the order in which we get dressed. Frequent sets play an crucial role in many Data Mining jobs that try to find appealing patterns from databases, like association rules, correlations, sequence, episode, classifiers and clusters. The

withdrawal of connection rules is one of the main popular troubles of all these. The classification of sets of items, harvest, symptoms and characteristics, which often occur mutually in the given database, can be seen as one of the most basic jobs in Data Mining.

**Sequential Pattern Mining**

Sequential Pattern mining is a theme of data mining anxious with finding statistically related patterns between data strings where the values are given in a sequence. In most cases it is supposed that the values are distinct, and thus time chain mining is strongly correlated, but regularly measured a different activity. Sequential pattern mining is a out of the ordinary case of structured data mining. There are several key conventional computational problems associated within this field. These include construction of potential databases and indexes for sequence in rank, taking out the frequently happening patterns, colliding sequences for similarity, and improving missing sequence members. In common, sequence mining problems can be confidential as the string mining which is typically based on string dispensation algorithms and item set mining which is typically based on organization rule learning.

**Closed itemset**

Let IS be a set of simple binary-valued properties, called items. A set X? IS is called an item set. A transaction database DB is a multi-set of item sets, where each item set, called an operation, has a unique identifier, called a TID. The support of an item set X in a dataset DB, denoted sup(X), is the fraction of transactions in DB where X comes as a subset. X is said to be a frequent item set in DB if sup(X) ò minimum support, where minimum support is a user defined minimum support threshold. An (frequent) item set is called closed if it has no (frequent superset having the same support. An association rule is an expression A1 ? B1, where A1 and B1 are item sets, and A1 ï B1 = í. The support of the rule is joint probability of a transaction containing both A1 and B1, given as support (A1 ? B1) =P (A1 ? B1) = support (A ? B). The confidence of a rule is the conditional probability that a transaction contains B1, given that it contains A1, given as: confidence (A1 ? B1) = P (B1|A1) =P (A1?B1)P (A1) =support (A1?B1) support (A1). A1 rule is frequent if the item set A1? B1 is frequent. The aim of non-redundant association rule mining is to generate a rule basis, a small, non-redundant set of rules, from which all other connection rules can be derived.

## LITERATURE SURVEY

In [1] the author has formulated and explained every step of our ClaSP algorithm. Clasp has two main phases as explained further the first one generates a subset of FS (and superset of FCS) called Frequent Closed Candidates (FCC), that is kept in main memory; and the second step executes a post-pruning phase to eliminate from FCC all non-closed sequences to finally obtain exactly FCS.

It is discussed in [1] when CMAP is integrated with CLaSP then CMCLaSP is formed. It shows the suitability of the vertical database format in obtaining the frequent closed sequence set, and how, under some database configurations, a standard vertical database format algorithm can already be faster than Pattern Growth algorithms for closed sequences, by only adding a simple post-processing step.

In this [2] paper the author has formulated and explained in comparison with our earlier technique. It generate the LS set which is a superset of closed frequent sequences and then store it in the prefix sequence lattice. It also does post-pruning to eliminate non-closed sequence.

This paper [3] explains about BIDE+, which has enumerated the complete set of frequent sequences and how to check upon getting a frequent sequence if it is not closed. We had to design some search space pruning methods or other optimization techniques to accelerate the mining process.

FPClose algorithm [4] is also preceded in two phases: first, we had found that the frequent closed item sets for each node (these are known as local closed item sets)and we then had checked whether the local closed item sets are also globally closed using various inter-node pruning techniques. Here we had given the formal definitions of local and global closed item sets.

In this paper [5] it is discussed that CHARM is unique algorithm because it concurrently explores both the item set space and TID set space, unlike all prior connection mining methods which only make use of the item set space. Further, CHARM gets over enumerating all possible subsets of a closed item set in cases of enumerating the closed frequent sets, which avoid a neat and hard down to top scheme.

DCHARM [6] algorithm is based on dissimilar sets data structure .DCHARM does a search for closed frequent sets by checking out both the item set space and operation space over an IT-tree (item set-TID set tree). It uses diverse set vertical data illustration for fast support

computations. It is also noted that at a given level of sustain; the completing time linearly goes up with increasing number of connections.

DCI_Closed [7] algorithm of the method receives three attributes: a closed item sets CLOSED SET, and two sets of items, i.e. th PRE SET and POST SET .The method will output all the non-similar closed item sets that neatly contain CLOSED SET. The prime target of the method is to deeply explore each valid new originator obtained from CLOSED SET by increasing it with all the element in POST SET.

LCM [8] The existing enumeration algorithm for frequent closed item sets are based on backtrack algorithm, which traverse a tree composed of all frequent item sets in F, and skip some item sets by pruning the tree. Our algorithm traverses a tree composed only of frequent closed item sets.

Apriori closed [9] algorithm is based on two prime ways explained as Minimum support is put on to find all frequent item sets in a .These frequent item sets and the leasr confidence restiction is used to form rules. The AprioriTid algorithm [10] is a transform of the Apriori algorithm. The AprioriTid algorithm also uses the "apriori-gen" function to know the candidate Item-sets before the pass starts. The main difference from the Apriori algorithm is that the Apriori TID algorithm does not use the database for including support after the initial pass.

## CONCLUSION

In this survey we considered both types of mining algorithms. The frequent item set mining algorithms perform better in terms of time consumption and are highly recommended for problems in which the individual item set is of greater importance. Whereas sequential pattern mining algorithms are much time consuming but they can be used and in fact are used to solve problems where a particular occurrence of a particular item set is not of great importance. Both types of mining algorithm perform well in different types of problem solving scenario. The approach taken by both of them is different but still if a head to head comparison is done, The FIM algorithms will definitely be considered better because of less time consumption and solving day to day problems in business marketing. For potential working of SPM algorithm, the need of large databases also comes to fore.

## REFERENCES

[1] Antonio Gomariz, Manuel Campos , Roque Marin, and Bart Goethals. "ClaSP:

An Efficient Algorithm for Mining Frequent Closed Sequences". 2013.

[2] Xifeng Yan, Jiawei Han and Ramin Afshar. "CloSpan: Mining Closed Sequential Patterns in Large Datasets".

[3] Jianyong Wang and Jiawei Han. "BIDE: Efficient Mining of Frequent Closed Sequences".

[4] H. D. K. Moonesinghe, Samah Fodeh, Pang-Ning Tan. "Frequent Closed Itemset Mining Using Prefix Graphs with an Efficient Flow-Based Pruning Strategy".

[5] Mohammed J. Zaki and Ching-Jui Hsiao. "CHARM: An Efficient Algorithm for Closed Association Rule Mining".

[6] Shaymaa Mousa. "COMPARISON BETWEEN RISS AND DCHARM FOR MINING GENE EXPRESSION DATA". 2013. International Journal of Data Mining & Knowledge Management Process.Claudio Lucches and Salvatore Orlando. "DCI Closed: a Fast and Memory Efficient Algorithm to Mine Frequent Closed Itemsets".

[8] Takeaki Uno, Tatsuya Asai, Yuzo Uchida and Hiroki Arimura. "LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets".Mining Frequent Itemsets – Apriori Algorithm

[9] Anurag Choubey, Ravindra Patel, J.L. Rana. "A Survey of Efficient Algorithms and New Approach for Fast Discovery of Frequent Itemset for Association Rule Mining(DFIARM)". 2011. International Journal of Soft Computing and Engineering (IJSCE)

[10] Y.Elovici, A.Kandel ,M.Last ,B.Shapira and O. Zaafrany. "Using Data Mining Techniques for Detecting Terror-Related Activities on the Web".

[11] Jiawei Han and Micheline Kamber. "Data Mining: Concepts and Techniques".2000.

[12] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. "From Data Mining to Knowledge Discovery in Databases".

[13] Chris Clifton, Murat Kantarcioglu. "Tools for Privacy Preserving Distributed Data Mining".

[14] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer. "The WEKA Data Mining Software: An Update".

**Appendix-A**

| Algorithim | Type Of Mining | Advantages | Year of Publication | Author |
|---|---|---|---|---|
| Apriori | SPM | Produces association rules that indicate what combinations are used. | 2010 | Hannu Toivonen |
| AprioriTIDClosed | SPM | The AprioriTid algorithm also uses the "apriori-gen" function to determine the candidate Item-sets before the pass begins. | 2011 | AnuragChoubey, Ravindra Patel, J.L. Rana |
| BI-Directional Extension | FIM | It consumes order(s) of magnitude less memory and can be more than an order of magnitude faster. It is also linearly scalable in terms of database size. | 2011 | Jianyong Wang _and JiaweiHan |
| CHARM | SPM | It is also linearly scalable in the number of transactions and the number of closed item sets found. | 2002 | M. J. Zaki and Ching-Jui Hsiao |
| ClaSP | FIM | Solves the problem of pattern mining on the vertical databases. | 2013 | Antonio Gomariz, M. Campos, Roque Marin, and Bart Goethals |
| CloSpan | FIM | Clospan produces a significantly less number of discovered sequences than the traditional methods while preserving the same expresie power. | 2003 | Xifeng Yan, Jiawei Han, RaminAfshar |
| DCharm | SPM | performs a search for closed frequent sets by exploring both the itemset space | 2013 | Shaymaa Mousa |

| | | | | |
|---|---|---|---|---|
| DCI Closed | SPM | It detects and discards duplicate closed itemsets, without the need of keeping in the main memory the whole set of closed patterns. | 2004 | Claudio Lucchese, Salvatore Orlando |
| FpClose | SPM | It facilitates fast frequency counting of itemsets viaintersection operations. | 2007 | H. D. K. Moonesinghe , SamahFodeh , Pang-Ning Tan |
| LCM | SPM | This algorithm exactly enumerates the set of frequent closed item sets within polynomial time perclosed item set in the total input size. | 2002 | Takeaki Uno, Tatsuya Asai, Yuzo Uchida, Hiroki Arimura |