Theoretically and Comparatively Analysis of different Frequent Pattern Mining Algorithms

Girija Shankar Dewangan
ME (CTA Branch) Student
Computet Science and Engineering Dept.,
SSCET Bhilai, India
gsd2010@rediffmail.com,

Mrs. Samta Gajhiye
Associate Professor
Computet Science and Engineering Dept.,
SSCET Bhilai, India
samta.gajbhiye@gmail.com

Abstract

In this paper we are trying to comparing the different sequential data mining algorithms for finding the frequent data patterns. There are many algorithms have designed for finding the frequent data patterns from the transactional database. Transactional database have specified number of transactions itemsets T, which are used for knowledge discovery. For Studying and Comparatively analysis purpose, we take three methods, one is traditional method Aprori and other two methods are 1. LCM (Linear closed itemset Miner) algorithm and 2. Top-K Closed Frequent pattern mining*. Here *we show the theoretically analysis of efficiency of different algorithms with computational complexity compared

Keyword

Sequential Data mining Algorithm, Aprori Algorithm, Top-K Closed Itemset frequent patterns mining, LCM (Linear Closed Itemset Mininer) algorithm.

I. Introduction

Sequential data mining algorithm is one approach of pattern discovery in Data Mining Area which are used in many applications such as association rule mining, inductive databases, and query expansion. Here we address first the problems of traditional method Aprori algorithm and second we discuss the others efficient algorithms and their methodologies.

II. Preliminaries

I = {1,-----, n} be the set of items. A transaction database of I is a set T= {t_1, -----,t_n} such that each ti is included in I. A subset P of I called a pattern (or itemset). A transaction including P is called an occurrence of P. The denotation of P, denoted by T (P) is the set of the occurrence of P. Given constant ϴ Є N, called a minimum support. Where N is length of the pattern. F is the set of frequent patterns and C is the set of frequent closed pattern.

The difference between closed frequent patterns and frequent pattern is irredundant transaction database. Means

1. No two transactions are the same.
2. No item itself is an infrequent pattern, and
3. No item is included in all transactions.

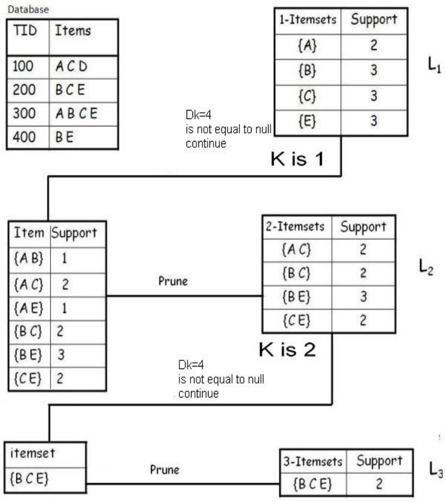
A frequent itemset P is maximal If P is included in no other frequent itemset, and closed if P is included in no other itemset included in the exactly same transactions as P. Some state of the art algorithms for closed pattern mining, such as CHARM and CLOSET [A2], use heuristic pruning method s to avoid generating unnecessary non closed pattern. Pasquier [A1] et al proposed the use of the closure operation to enumerate closed pattern s, their idea is to generate frequent pattern and check whether are closed patterns or not by closure operation. so this reduces the storage space for non closed patterns.

III. Comparatively study of the algorithms

1. Aprori Algorithm and their methodology

In the first we discuss the traditional method Aprori Algorithm, which is also called level-by-level algorithm.

for example , Let P_K be the set of frequent itemsets of size k. Aprori algorithms start with P_0, that is {ϕ}, and compute P_k from P_{k-1} in the increasing order of k=1. Any itemset in P_k is obtained from an itemset of P_{k-1} by adding an item. Aprori algorithms add every item to each itemset of P_{k-1}, and choose frequent itemsets among them. If P_k=ϕ holds for some k, then P_{k}'=ϕ holds for any k'>k. thus, Aprori algorithms stop at such k, this is the approach of Aprori algorithms.



Database

TID	Items
100	A C D
200	B C E
300	A B C E
400	B E

1-Itemsets	Support
{A}	2
{B}	3
{C}	3
{E}	3

Item	Support
{A B}	1
{A C}	2
{A E}	1
{B C}	2
{B E}	3
{C E}	2

2-Itemsets	Support
{A C}	2
{B C}	2
{B E}	3
{C E}	2

itemset
{B C E}

3-Itemsets	Support
{B C E}	2

2. LCM Algorithm and their methodology:

LCM uses prefix preserving closure extension (ppc extension in short), which is an extension from a closed itemset to another closed itemset. The extension induces a search tree on the set of frequent closed itemsets, thereby we can completely enumerate closed itemsets without duplications and generating a ppc extension needs no previously obtained closed itemset. Hence, the memory use of LCM does not depend on the number of frequent closed itemset, even if ther are many frequent closed itemset.

LCM is the backtracking algorithm Backtracking algorithm is based on recursive calls. A pruning of a backtracking algorithm inputs a frequent itemset P, and generates itemsets by adding every itemset to P, then for each itemset being frequent among them; the pruning generates recursive calls with respect to it. To avoid duplications, an iteration of backtracking algorithms adds items with indices larger than the tail of P.

2.1. Backtracking algorithm is as follows Algorithm Back Tracking (P: Current solution)

1. Output P
2. For each e Є I, e>tail (P) do
3. If (PU {e})

An execution of backtracking algorithms gives a tree structure such that the vertices of the tree are iterations, and edges connect two iterations if one of the iteration calls the other. If an iteration I recursively calls another iteration I', then we say that I is the parent of I' is a child of I. for an iteration, the itemset received from the parent called current solution.

2.2. Comparison between Aprori Algorithm and LCM:

Aprori algorithm sequential data mining approach but LCM is backtracking approach.

Sequential data mining use much memory for storing P_K in memory, while backtracking algorithm use less memory since they keep only the current solution.

Backtracking algorithms need no computational for maintaining previously obtained itemsets , so the computation time of backtracking algorithm is generally short, However , Aprori algorithms have advantage for frequency counting.

LCM algorithm are based on backtracking algorithms, and use an efficient techniques for the frequency counting , which are occurrence for the frequency counting , which are occurrence deliver and anytime database reduction desired below . Hence, LCM algorithms compute the frequency efficiently without keeping previously obtained itemset in memory.

The time complexity of LCM is theoretically bounded by a linear function in the number of frequent closed itemsets.

LCM is implemented with only arrays. Therefore, LCM is fast, and outperforms than other algorithms for some sparse datasets.

LCM does not have any routine for reducing the database, while many existing algorithms have.

Performance of LCM is not good for dense datasets with large minimum support, which involves many unnecessary items and transaction.

3. TOP-K Closed frequent pattern mining

End users of traditional frequent pattern mining applications encounter several well- known problems in practice. First, without specific Knowledge about the target data, users will have difficulties in setting the support threshold to obtain their required results. Second, the algorithms often generate an extremely large number of frequent patters, often in thousands or millions, which is even larger than the original target dataset. It is nearly impossible for the end users to comprehend or validate such large number of complex patterns, thereby limiting frequent patterns mining spread use and acceptance in many real world situations.

Top-k closed itemsets mining is combining LCM algorithm and priority queue to avoid closed checking.

Closed itemset mining is used to eliminate redundant patterns. This has to mine only the pattern having no superset with the same support. They can reduce the number of equal and less supported pattern without information loss. In general an appropriate minimum support threshold is set by the user because they need to be familiar with both mining query and task-specific data. To avoid this problem, in this method we mine only N- k itemsets the highest support for k up to certain k_{max}, where k_{max} is the upper bpund of the size of itemsets, and N is the desired number for k-itemset..This algorithm mining top-k frequent closed itemset of length no less than min l, where k is the desired number of frequent closed itemset.

3.1.Comparison between Aprori algorithm and Top-k frequent pattern mining.

If compare Top-k closed method with Aprori Algorithm we found following difference between them.

It fast finds top-K closed itemsets.

It does not need finding the final minimum support threshold before mining(the final support threshold is found when the k^{th} top –K closed itemset found),

It is an efficient pruning unpromising itemset and stopping rapidly as soon as top-K closed itemset are found,

It does not need to maintain the top=k closed itemset mined in memory (it does not need closed checking), and,

Some itemsets are skipped length by calculating their closure.

IV. Conclusion

In this paper we have presented a theoretical discussion about the sequential pattern mining and backtracking algorithms.

And here we comparatively analysis and discuss methodology of Aprori Algorithm and then LCM (Linear closed itemset Miner) algorithm and Top-K Closed Frequent pattern mining. Aprori Algorithm is basic approach to finding frequent pattern mining. But this method is not efficient because take much time to calculate frequent pattern and this method calculate in the fix length pattern. Aprori Algorithm can not find closed frequent pattern but if we use LCM algorithm then it use PPC extension to find closed frequent pattern with less consumption time, but the problem of fix support length is solve TOP-K closed frequent pattern mining by combining the LCM algorithm and priority queue to avoid closed checking.

Reference

[A1] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Discovering Frequent Closed Itemsets for Association Rules, In Proc. ICDT'99, 398-416, 1999.

[A2] M. J. Zaki, C. Hsiao, CHARM: An Efficient Algorithm for Closed Itemset Mining, In Proc. SDM'02, SIAM, 457-473, 2002.

[A3]"TFP: An Efficient algorithm for mining Top-K Frequent Closed Itemsets" Jianyoung Wang, Jiawei Han Petre TZvetkov

[A4]"Mining Closed Frequent Itemset based on FP-Tree", shenqwei Li,Institute of Data and Knowledge Engineering Henan University Kaifeng, China

[A5]"Mining Frequent Closed Itemsets from distributed Dataset", Chunhua JU and Dongjun Ni, 2008 Internaational Symposium on Computational Intelligence and Design

[A11]"Research of Top-N Frequent Closed Itemsets Mining Algorithm", Lizhi Liu, Jun Liu School of Computer Science and Enginnering, Wuhan Insititute of Technology, Wuhan Hubai, China 2008 IEEE Paper