Hierarchial Approach for Frequent Closed Itemset Generation in Distributed Environment

Aditi Site
School of Computer Science and Engineering
VIT University Chennai Campus
Chennai, Tamil Nadu, India

Muralidhar A
School of Computer Science and Engineering
VIT University Chennai Campus
Chennai, Tamil Nadu, India

ABSTRACT
This paper addresses the issue of generating the frequent closed itemset in distributed environment. Some algorithms have been proposed earlier there but they are suffering from the drawbacks like: Increasing communication load or frequent communication between the nodes for transferring information. So some algorithm need to be proposed which could solve these two drawbacks simultaneously and this paper have propose one such algorithm so that the mining of the datasets present in the distributed environment could be done effectively and with less theoretical complexity.

Keywords
Frequent closed item-sets; global and local frequent closed itemsets; hierarchial system; distributed environment; subset examination techniques

1. INTRODUCTION
Data is increasing at a much faster rate and to derive the meaningful information from the data is important. This requires the mining of large datasets and many algorithms have been developed for this from past years. One such method includes mining of frequent closed itemsets. Frequent closed itemsets are basically subsets of the frequent itemsets which are frequently occurring data items in certain supersets of a single itemsets or the support of that itemsets is greater than a threshold value. From the algorithms developed in the past, only few were there which which were concerned with the parallel or distributed environment.

The algorithms which presented earlier were either concerned about the frequent closed itemsets mining in local environment or the frequent itemsets mining in distributed environment. Some newly proposed algorithms are there like DMCI (distributed mine closed itemsets) and DMCI-ML (distributed mine closed itemsets-Meta learning) which suggested two approaches on how to transfer the datasets of different sites to another sites and then perform the frequent closed mining over the datasets but these approaches also contributes to the increasing transfer costs and load over network and the nodes. Moreover it also increases the computation for mining. This paper proposes a new methodology or a new approach which is hierarchial approach, in order to efficiently perform the work of frequent closed itemsets mining for the distributed environment.

For frequent closed itemsets mining, the itemsets mining is first performed at the individual level and then their results are transferred to the upper level. At each level, the computations will be done parallel and the concurrency problem will also be not there.

2. BASIC TERMINOLOGY

- **Frequent closed itemsets**
   Support values are the values which indicate frequently that particular item appears in the dataset.
   Frequent itemsets are the itemsets whose support values are greater than the threshold value.
   Frequent closed itemsets are the itemsets for which the value of the support for its supersets is lower than the support value of the itemsets. Let us assume that the X and Y are the two itemsets which are frequent, if X is a subset of Y and sup(X)>sup(Y) then it is a frequent closed itemsets.

- **Global and local frequent closed itemsets**
  Given a closed itemsets X, if X.sup >= minsup*D then X is said to be a globally frequent closed itemsets.
  X.sup= support value of X itemsets
  Min sup=minimum support value or threshold value
  D= distributed or combined dataset
  Dsi=single site dataset
  Suppose there are n no. of sites and a complete dataset D, this dataset is divided among sites and the support value at each site is X.sui and support value at overall dataset is X.sup.
  If X.sui >= minsup*Di, then X is locally frequent closed itemsets on site Si.

- **Hierarchial system**
  Hierarchical system or hierarchical network design model breaks the complex problem of a network design into smaller problems. Each level have nodes for doing some kind of computations and the devices at each level performs similar and well defined functions.
  At lower levels, the frequent closed itemsets computation will be performed and then the results from the nodes will be sending to the parent node which will merge the results.

- **Subset Examination techniques**
  For all the candidate itemsets compare two itemsets X and Y and if X is a subset of Y and satisfies the condition of that X.sup = Y.sup then X will be pruned then repeat the steps with the other candidate itemsets.

3. EXISTING METHODOLOGY

Earlier two methodologies were proposed for the frequent closed itemsets mining in the distributed environment:

One approach is DMCI in which each and every node sends the itemsets information to the other node where dataset is scanned for the frequent closed itemsets and results are
propagated and combined at each successive node. But such type of system requires more no. of connections thus increasing load and traffic over the network.

Another approach is the DMCI-ML in which initially all the information or datasets from nodes is send and collected at a particular site and from there the information about the frequent and the non-frequent itemsets is taken and the results so obtained after the processing is send to all other sites. This type of strategy suffer from the problem that there is a load only at a particular site and if that site called as meta learning site get fails then whole system will get fail.

4. **HIERARCHIAL MODEL FOR DISTRIBUTED DATASET MINNING (HMDDM)**

Hierarchical system is a type of system in which a set of devices is arranged in the form of a hierarchical tree. The structure is:

According to this structure let:

- **Ni** from i=1 to m be the sites present at the leaf.
- **Pi** from i=1 to m be the sites which are immediate parents of the leaf.
- **Pii** from i=1 to m be the sites which are immediate parents of the **Pi**.
- **Ri** is the root node or the main site.

Their information consists of the locally mined frequent closed itemsets; the parent node at immediate higher level will therefore be processing and combining the results from the lower level nodes. This process is repeated until the information from all the datasets of the sites is combined and the knowledge about the frequent closed itemsets are obtained.

**Approach:** This approach also makes the use of the existing technology but with only limited number of nodes initially certain group of nodes will be sending their information to their parent node or a node present at a higher level.

Their information consists of the locally mined frequent closed itemsets; the parent node at immediate higher level will therefore be processing and combining the results from the lower level nodes. This process is repeated until the information from all the datasets of the sites is combined and the knowledge about the frequent closed itemsets are obtained.

**Algorithm:** Hierarchical model for distributed dataset mining (HMDDM)

- **Step1:** Perform local frequent closed itemsets mining for the leaf nodes or sites.
- **Step2:** Apply frequent closed itemsets mining algorithm to each site.
- **Step2:** Send the results to the immediate upper level nodes.

- **Step3:** Combining of the results based on rule that: If X is globally frequent closed itemsets, then X must be locally frequent in at least one of the site.
- **Step4:** Repeat step2 and step3 until the main site or the meta-learning site contains the result.
- **Step5:** The result so obtained is the combined closed itemsets.

5. **THEORATICAL COMPLEXITY AND PERFORMANCE EVALUATION AND COMPARISION**

5.1 **Communication Frequency**

**DMCI:** Here since each node has to send their information to every other node so here for n no. of nodes the complexity will be O (n^2).

**DMCI-ML:** Each node will be sending their information to a single site that is the meta-learning site and so if n no. of nodes is their communication frequency or complexity will be in O (n).
**HMDDM**: For this algorithm, the complexity of such type of structures is approximately $O(\log n)$. For $n$ no. of nodes here they will be sending the data to approximately $(n/2)$ higher level nodes so overall communication will be less than $O(n)$.

**5.2 Based on the Support Thresholds**

On increasing the support thresholds, the no. of patterns also increases.

**DMCI**: For DMCI, on increasing the support thresholds, no. of patterns for each site also increases and thus the traffic over the network increases as lots of data needed to be transferred to each and every node each time.

**DMCI-ML**: Here more patterns will be generated and have to be managed at the main site which in turn increases the load over the main node.

**HDMMD**: Always a manageable no. of patterns will be sent over the network. Moreover it is not affected by the increase in the number of patterns.

**6. FUTURE WORK**

Frequent Closed Itemsets mining can be done in the static data or in the continuous stream of data. The algorithm proposed currently is for the static data which can be extended for the continuous data streams with memory optimization as continuous or dynamic data streams leads to memory consumption.

**7. CONCLUSION**

This paper thus has proposed an efficient frequent closed itemsets mining approach in the distributed environment. This approach performs the closed itemsets mining over datasets from different sources as well as reduces the network load or communication load to a greater extent in future this work could be extended for the stream data as well.

**8. REFERENCES**


