A Review On Frequent Pattern Extraction Technique Over Static and Dynamic Data

Ms. Monika V. Deore, Prof. N. R Wankhede

Ms. Monika V. Deore, Computer Science, Sapkal Knowledge Hub Nashik, Maharashtra, India
Prof. N. R Wankhede, Computer Science, Sapkal Knowledge Hub Nashik, Maharashtra, India

ABSTRACT

Frequent itemset/pattern mining is used in variety of domains. Along with frequent itemset extraction, Utility mining is useful in retail market analysis. Lot of application in current era generates continuous streaming data. The mining strategy of static dataset differs from the continuously streaming data. This work aims to study various frequent itemset extraction technique and utility extraction techniques. Based on the study of existing work a new technique is proposed to overcome the limitations of existing system. Existing Systems implements frequent itemset extraction and utility itemset extraction separately. The proposed system aims to find utility frequent itemset over streaming data. Along with the frequent itemset extraction it also takes in to account the utility of product. The system extracts top k utility frequent itemset at the current movement.

Keyword: - *frequent itemset , utility itemset, data streams, top k itemset, CP-graph*

1. INTRODUCTION

In frequent itemset mining, set items are extracted those are frequently occur in dataset with minimum support value. The support value is the user defined threshold value. Support value adds the minimum occurrence constraint on frequent itemset mining strategy. Frequent itemset/ pattern mining is applicable in variety of domains such as market basket analysis, retail market analysis, intrusion detection, web click mining, network monitoring, bioinformatics, etc.

In real life scenarios, the data is generated in streaming format. A single or multiple streams are generated in variety of application. Hence frequent itemset mining over data stream is important task. Frequent itemset mining play vital role in variety of applications such as:

1. In social networks like facebook twitter, etc generates bulk streaming in every day. The relationship among multiple post/tweets can be extracted by matching keyword in it. More frequently occurred tweets/post can define a social media trend. Frequent itemset mining is useful in trend analysis.
2. In E-COMMERCE strategy product promotions, recommendations can be performed using frequent pattern analysis. In ecommerce the purchase history of multiple users can be traced to find sequence of items in purchase strategy.
3. Association mining: With the help of application usage statistics in smart phones, usage pattern can be extracted. Usage pattern may contain location specific application access, user profile specific application access. This helps in statistical study of various application categories.

In multiple data streaming, all streams data is merged together to generate a dataset. Along with the support evaluation closed co-occurrence patterns analysis is done in multiple streaming data mining. In closed co-occurrence pattern analysis, pattern occurrence is checked with be 2 or more streams. The frequent item should be present in at least 2 streams.

There are various algorithms are proposed for frequent pattern analysis like apriory, Eclat, Fp-Growrh, etc. These algorithms cannot be directly applied to continuously changing the streaming data. There is need to implement different strategy for stream data analysis.

A utility itemset extraction is a technique in which itemsets are extracted from a dataset that generates higher profit. The frequent itemset extraction and utility itemset evaluation are two important techniques in retail market analysis. In the following section various techniques related to frequent itemset mining and utility extraction are studied.

2. RELATED WORK

Arnaud Giacometti, Dominique H. Li, Patrick Marcel, Arnaud Soulet proposes a survey based on last 20 years work in the domain of pattern mining[2]. This survey provides the overview of 1,087 publications papers. The work includes variety of pattern extraction using association riles and itemset. The pattern mining techniques are mainly classified in following 6 categories:

1. Frequent pattern mining on static data:

A whole dataset is provided to the system at ones. The system will extract the patterns by analyzing complete data. For such analysis apriory[3] and FP growth[4] are two main techniques used in literature.

In apriory algorithm the execution is divided in two phases. In first phase candidate items are extracted and in second phase frequent items are extracted. This apriory technique uses breadth-first search to find next probable frequent itemset.

FP growth algorithm technique uses FP tree for frequent itemset evaluation. This technique reduces the database scan and it does not generate the candidate itemset. Fp growth technique is faster than apriory technique.

Mining top-k frequent closed itemsets[5] is proposed to extract top k items using apriory and fp growth technique. Differential privacy based frequent itemset[6] mining technique provides the privacy in frequent data mining. This technique adds the noise in data before analysis to provide data privacy.

2. Frequent pattern mining over a single stream.

G. S. Manku and R. Motwani[7] proposed a technique of mining frequent itemset over streaming data. FP-Stream[8] technique is used to find current frequent patterns and prediction of future pattern occurrence. These techniques provide approximate results. varying-size sliding window technique[9] is used to provide high accuracy in solution.

To extract exact frequent patterns from dataset DStree[10] technique is proposed. But this technique has several limitations like tree structure is not compact, storage overhead issue, etc. To overcome these problems CPS-tree[11] technique is proposed. This three has compact tree structure and hence reduces the storage overheads.

To mine frequent itemset over streaming data apriory based[12] and FP-growth[13] base techniques are also proposed. These techniques update the data structure incrementally with respect to every sliding window batch.

Mining Top k closed patterns from data stream [14] is also proposed in literature. The patterns are extracted from a single streaming window.

3. High utility itemset extraction

Frequent itemset mining technique extracts the items those are occurs frequently in transactions but it may discover the low value itemset i.e. low profit itemset and can lose the information of high valued itemsets. User is interested in finding high profit itemset in the transactional dataset. Apriori-based algorithm for mining High utility Closed itemsets[15] extracts the high utility itemset based on predefined threshold. To overcome the problem of user defined threshold value, Mining top k High utility itemset technique is proposed to extract the top k high profit elements from the dataset [16].

4. Utility based frequent itemset extraction:

The utility itemset extraction along with frequent itemset extraction is proposed in literature[17]. It introduces a new concept as utility frequent itemset. This algorithm is run in two phases initially frequent items are extracted as a candidate items and then based on the utility value the items are filtered.

5. Frequent pattern mining across multiple databases.

   Multiple static databases are considered to find frequent itemset. Multiple databases with user defined threshold values for inter-frequency and intra-frequency matching of itemset is proposed in[18].

6. Frequent pattern mining across multiple streams.

   Data with multiple streams is analyzed using seg-tree[19]. In this technique segment tree is generated as a data structure to find patterns in transaction dataset. A CoolMine algorithm is proposed to travel the seg-tree and obtain the common patterns. But this algorithm is computationally expensive.

   In[1] cp-grap based multiple streaming frequent itemset mining is proposed. This technique overcome the drawback of CoolMine algorithm[19]. This technique finds top k frequent itemset from multiple stream using closed co-occurrence patterns technique.

3. ANALYSIS AND PROBLEM FORMULATION

Multiple real life examples generate data streams. A same set of object can appear in more than one stream. The existing work includes variety of techniques such as single stream processing, multiple stream processing. The utility based frequent itemset are extracted from static dataset. There is need of such system that provides solution for utility based frequent itemset extraction technique for multiple data streams.

4. PROPOSED SYSTEM

The system works with multiple data streams. System read input streams and generates CP-Graph. Based on the cp graph, system finds set of tuples consist of closed co-occurrence patterns and its occurrence count. The utility value is evaluated for the filtered high co-occurrence count patterns and system generates high utility frequent itemset. Following figure represents the proposed system architecture.

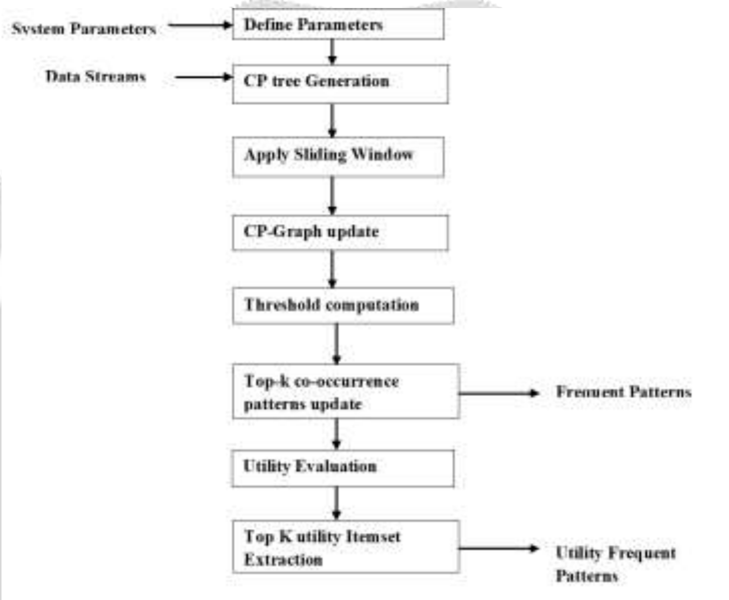

Fig -1: Proposed system architecture

To read multiple data streams system follows the sliding window protocol. After receiving transaction slot after every sliding window, system performs following actions.

1. Index update: CP-graph is updated as per the new transaction entry. The expired transactions are deleted and accordingly CP-Graph is updated.
2. Mining top-k closed co-occurrence patterns:
   A set of closed co-occurrence patterns are extracted after every sliding window slot and accordingly the threshold U is updated.
3. Mining top k Utility-frequent itemset: Based on the extracted closed co-occurrence patterns and utility values system extracts high utility itemset from frequent closed co-occurrence patterns.

4. CONCLUSIONS

Multiple real life examples generate data streams. A same set of object can appear in more than one stream. A closed co-occurrence pattern is the patterns that exist in more than the defined threshold streams. The existing work focuses on single data stream entry and utility evaluation. There is need of such system that provides solution for utility based frequent itemset extraction technique from multiple streams.

6. REFERENCES

[1] Daichi Amagata, Takahiro Hara,"Mining Top-k Co-Occurrence Patterns across Multiple Streams", in IEEE Transactions on Knowledge and Data Engineering, Vol. 29, Issue 10, pp. 2249 - 2262, Oct 2017.

[2] A. Giacometti, D. H. Li, P. Marcel, and A. Soulet, “20 years of pattern mining: a bibliometric survey,” ACM SIGKDD Explorations Newsletter, vol. 15, no. 1, pp. 41–50, 2014

[3] R. Agrawal and R. Srikant, “Fast algorithms for mining association rules,” in VLDB, 1994, pp. 487–499.

[4] [4]. J. Han, J. Pei, and Y. Yin, “Mining frequent patterns without candidate generation,” in SIGMOD, 2000, pp. 1–12.