International Journal of Electronics Communication and Computer Engineering
Volume 4, Issue 1, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

A Survey on Frequent ItemSet Mining Over Data Stream

Rajesh Rawat
Student of Software System
S.A.T.I (Vidisha), M.P., India
neesuraj@gmail.com

Asst. Prof. Nidhi Jain
Department of Information Technology
S.A.T.I (Vidisha), M.P., India
j.nidhi9@yahoo.co.in

***Abstract* - The growing importance of data streams from a wide range of advanced applications such as fraud detection and learning trend has led to the study of Frequent Item-Set Mining over Data Stream. A data stream is an ordered sequence of instances that arrive at a rate that does not permit to permanently store data in memory. A frequent item-set is a set of items that appears at least in a pre-specified number of transactions. Frequent item-sets are typically used to generate association rules. In this paper we are discussing different type windowing techniques and the important algorithms available in this mining process.**

***Keywords* – Data Stream Mining, Frequent Itemset, Association Rule Mining, Windowing Techniques.**

I. INTRODUCTION

A data stream is an ordered sequence of instances that arrive at a rate that does not permit to permanently store data in memory. A frequent item-set is a set of items that appears at least in a pre-specified number of transactions. Frequent item-sets are typically used to generate association rules.

II. ASSOCIATION RULE MINING

Association rule mining extracts interesting correlation and relation between large volumes of transactions. This process is divided into two phases. First phase is frequent item-set generation, finding all item-sets that sufficiently exceed minimum support. Second phase is rules construction. From the frequent item-sets generated all association rules having confidence higher than minimum confidence are formed. Frequent item-set generation is the resource consuming task and is the active area of research.

III. FREQUENT ITEMSETS, CLOSED ITEMSETS

Let I = {I1 , I2 , . . . , Im } be a set of items. Let D, the task-relevant data, be a set of database transactions where each transaction T is a set of items such that T ⊆ I. [5] Each transaction is associated with an identifier, called TID. Let A be a set of items. A transaction T is said to contain A if and only if A ⊆ T. An association rule is an implication of the form A ⇒ B, where A ⊂ I , B ⊂ I , and A ∩ B = ̃The rule A ⇒ B holds in the transaction set D with support s, where s is the percentage of transactions in D that contain A ∪ B (i.e., the union of sets A and B, or say, both A and B). This is taken to be the probability, P(A ∪ B).1 The rule A ⇒ B has confidence c in the transaction set D, where c is the percentage of transactions in D containing A that also contain B. This is taken to be the conditional probability, P(B|A). [8]That is,

$$support(A \Rightarrow B) = P(A \cup B) \qquad (1.1)$$

$$confidence(A \Rightarrow B) = P(B|A) \qquad (1.2)$$

Rules that satisfy both a minimum support threshold (min sup) and a minimum confidence threshold (min conf) are called strong. By convention, we write support and confidence values so as to occur between 0% and 100%, rather than 0 to 1.0.

A set of items is referred to as an itemset. An itemset that contains k items is a k-itemset. The set {computer, antivirus software} is a 2-itemset. The occurrence frequency of an itemset is the number of transactions that contain the itemset. This is also known, simply, as the frequency, support count, or count of the itemset. Note that the itemset support defined in Equation (1.2) is sometimes referred to as relative support, whereas the occurrence frequency is called the absolute support. If the relative support of an itemset I satisfies a pre specified minimum support threshold (i.e., the absolute support of I satisfies the corresponding minimum support count threshold), then I is a frequent itemset. The set of frequent k-itemsets is commonly denoted by Lk .[7]
From Equation (1.2), we have

$$\text{Confidence}(A \Rightarrow B) = P(B \mid A) = \frac{\text{support}(A \cup B)}{\text{support}} = \frac{\text{support count}(A \cup B)}{\text{support count}(A)} \qquad (1.3)$$

Equation (1.3) shows that the confidence of rule A⇒B can be easily derived from the support counts of A and A ∪ B. That is, once the support counts of A, B, and A ∪ B are found, it is straightforward to derive the corresponding association rules A⇒B and B⇒A and check whether they are strong. Thus the problem of mining association rules can be reduced to that of mining frequent itemsets.

IV. WINDOWING TECHNIQUES

According to different stream processing models, the research of mining frequent itemsets in data streams can be divided into three categories:[1] landmark windows as shown in Figure 1.1, sliding windows as shown in Figure 1.3, and damped windows as shown in Figure 1.2. In the landmark windows model, knowledge discovery is performed based on the values between a specific time stamp called landmark and the present. In the sliding

windows model, knowledge discovery is performed over a fixed number of recently generated data elements which is the target of data mining. In the damped windows model, recent sliding windows are more important than previous ones. In other words, older transactions contribute less toward itemset frequencies.

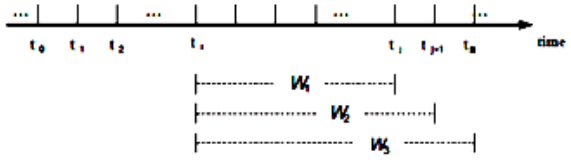

Figure 1.1 The landmark window model

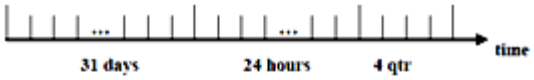

Figure 1.2 The tilted-time window model with logarithmic partition

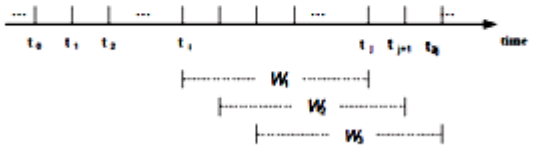

Figure 1.3 The sliding window model

A transaction data stream is a sequence of incoming transactions and an excerpt of the stream is called a window. [10] A window, W, can be either time-based or count-based, and either a landmark window or a sliding window. W is time-based if W consists of a sequence of fixed-length time units, where a variable number of transactions may arrive within each time unit. W is count-based if W is composed of a sequence of batches, where each batch consists of an equal number of transactions. W is a landmark window if W = (T1, T2, . . . , T); W is a sliding window if W = (TT−w+1, . . . , TT), where each Ti is a time unit or a batch, T1 and TT are the oldest and the current time unit or batch, and w is the number of time units or batches in the sliding window, depending on whether W is time-based or count-based. Note that a count-based window can also be captured by a time-based window by assuming that a uniform number of transactions arrive within each time unit.

The frequency of an item set, X, in W, denoted as freq(X), is the number of transactions in W support X. The support of X in W, denoted as sup(X), is defined as freq(X)/N, where N is the total number of transactions received in W. X is a Frequent Item set (FI) in W, if sup(X) ≥ σ, where σ (0 ≤ σ ≤ 1) is a user-specified

V. Conclusion

As association rule mining is one of the hottest topics in the area of data mining. The research activities on this topic is reviewed hence, the survey guides the researchers to get an idea about the recent advancements with association rule mining. We also discussed some of the issues of the windowing concept for the online stream mining to develop an effective, performance oriented algorithm.

References

[1] Pramod S., O. P. Vyas “Data Stream Mining : A Review on Windowing Approach” Vol 12, No 11-C (2012): Global Journal of Computer Science and Technology, 2012.

[2] Wen-Yang Lin, Yi-Ching Chen, He-Yi Li, “ Mining Indirect Associations Over Data Streams “ ICSSE, 2012

[3] C. Aggarwal, J. Han, J. Wang, P. S. Yu, “A Framework for Clustering Evolving Data Streams, Proc. 2003 Int. Conf. on Very Large Data Bases”, Berlin, Germany, Sept. 2003.

[4] C. Aggarwal, J. Han, J. Wang, and P. S. Yu, “A Framework for Projected Clustering of High Dimensional Data Streams”, Proc. 2004 Int. Conf. on Very Large Data Bases, Toronto, Canada, 2004.

[5] Ashraf El-sisi (2010) “ Fast Cryptographic Privacy Preserving Assocaition rules mining on Distributed Homogenous database”, The IAJIT vol 7, No.2, April 2010.

[6] AhaiLiang, TANG Xingmang, LI Lin JIANG Wen Liang, (2005)“ Temporal Association with Rule mining Based on T-Apriori algorithms and its typical application”, Proceedings of International Symposium on Spatio-temporal Modeling, Spatial Reasoning, Analysis, Data Mining and Data Fusion.August 27-29, Peking University, China -2005.

[7] Alva Erwin, Raj P. Gopalan, N. R. Achuthan.(2007) “A Bottom – Up Projection Based Algorithm for Mining High Utility Item sets“, 2nd workshop on Integrating AIDM Gold Coast, Australia. CRPIT- vol 84, kok-leong Ong, Junbin Gao, Wenyuan aLi, Ed. constraint based mining and learning at ECML/PKDD2007, CMILE, str 10-20, 2007.

[8] Alexandre Evfimievski, Ramakrishnan srikant, Rakesh Agrawal, Johannes Gehrke(2002) “ Privacy preserving mining of Association Rules “, SIGKDD ACM.

[9] Chang-Hung Lee, Ming-syan chen,Cheng-Ru Lin (2003) ,“An Efficient algorithm for mining general Temporal Association rules “,IEEE transactions on knowledge and data engineering , vol 15 , No.4, july/august- 2003

[10] Chang-HungLee, Cheng-Ru Lin, Ming-syan Chen (2005), “Sliding Window filtering: an efficient method for incremental mining on a time-variant database “, Information systems 30, 227-244, 2005.