Comprehensive Study of Weighted Sequential Pattern Mining

Niti Desai, Amit Ganatra

Abstract: ***Extensive growth of data gives the motivation to find meaningful patterns among the huge data. Sequential pattern provides us interesting relationships between different items in sequential database. In the real world, there are several applications in which specific sequences are more important than other sequences. Traditional Sequential pattern approaches are suffering from two disadvantages: Firstly, all the items and sequences are treated uniformly. Second, conventional algorithms are generating large number of patterns for lower support. In addition, the unimportant patterns with low weights can be detected.***

This paper addresses problem of traditional framework and various framework of weighted sequential pattern. Paper also discuses how algorithm mines sequential pattern which reduces the search space and new pruning technique prune the unimportant pattern and pick only those patterns which leads to important and emerging pattern. Later section of paper discuses results of simulation study and how researcher can lead current research.

Keywords: Weighted Sequential Pattern Mining, Weighted Association Mining Framework, Weighted sequential pattern Mining Framework

I. INTRODUCTION

Data mining problem, discovering sequential patterns, was introduced in [1][2]. Sequential pattern mining is an important data mining task of discovering time-related behaviours in sequence databases. Sequential pattern mining algorithms GSP[11], SPAM[3], SPADE[19], Freespan[6], Prefixspan[10] etc. are worked on two threshold values like minimum support and minimum confidence.The entire data mining is processing under the two threshold value restriction. Therefore, support and confidence play an important role in the mining process. Support threshold value becomes a key factor in sequential pattern mining because on basis their value item/itemset/pattern is prune. But in real life, each item is having different significance like bread and butter is having high occurrence as compare to gold-chain and pendent but to extract gold-chain and pendent from database is more important for decision maker because its significance weight is more stronger than bread and butter. In real life each item is having different importance, which is not at all consider in well known sequential pattern mining approaches like GSP[11], SPAM[3], SPADE[19], Freespan[6], Prefixspan[10] etc. Existing sequential pattern mining suffers following problem:

- If the given minimum support is too high, then the items with low frequency of appearance can't be mined.
- Otherwise, if the given minimum support is too low, then a large number of non-meaningful patterns will be mined in the mining process.

To solve above problems, various researchers have proposed various frameworks which can be useful to solve problem of traditional techniques. To meet the user objective and business value, various weighted association rule mining methods were proposed based on the weightage to items.

II. EXISTING WORK

2.1. Association Rule Mining(ARM)

Association rule mining aims to explore large transaction databases for association rules. Classical Association Rule Mining (ARM) model worked on Support and confidence measures.

Support: The Support of an itemset expresses how often the itemset appears in a single transaction in the database i.e. the support of an item is the percentage of transaction in which that items occurs.

Formula: $\mathbf{I = P(X \cap Y) = \frac{(X\cap Y)}{N}}$

Range: [0, 1]

If I=1 then Most Interesting

If I=0 then Least Interesting

Confidence: Confidence or strength for an association rule is the ratio of the number of transaction that contain both antecedent and consequent to the number of transaction that contain only antecedent.

Formula: $\mathbf{I = P\left(\frac{Y}{X}\right) = \frac{P(X\cap Y)}{P(X)}}$

Range: [0, 1]

If I=1 then Most Interesting

If I=0 then Least Interestin

Limitation of Association Rule Mining (ARM) model:

- ARM assumes that all items have the same significance without taking their weight into account
- It also ignores the difference between the transactions and importance of each and every itemsets.

2.2. Weighted Association Rule Mining (WARM)

Weighted Association Rule Mining (WARM) does not work on databases with only binary attributes. It makes use of the importance of each itemset and transaction[15]. Weighted version of SPAM is able to extract the item within the sequence are expensive and the which are having low frequencies (supports)[16].

2.2.1. Weighted support – Confidence framework

Weighted Support:

Manuscript received on May, 2013.

Niti Desai, Deprtment of Computer Engg, Uka Tarsadia University, Bardoli, Surat, Gujarat, India.

Amit Ganatra, U and P U Patel Department of Computer Engineering, Charotar University of Science and Technology, Changa 388421, Anand, Gujarat,India.

The weighted value of the subset X ($X \subset I$) of items in transaction t_i is calculated as

$$W(x)=\frac{\prod_{k=1}^{|x|}(\forall[i_k \in X])^{t_i[i\kappa[w]]}}{\sum_{k=1}^{|x|}(\forall[i_k \in X])^{t_i[i\kappa[w]]}}$$

Where $t_i[i\kappa[w]]$ is the weighted value of the kth item $i\kappa$ in the ith transaction. The weighted support is a summary of weighted value of the transaction item set containing this item in the transaction database, that is

$$W(x)=\frac{N_x\prod_{k=1}^{|x|}(\forall[i_k \in X])^{t_i[i\kappa[w]]}}{n\sum_{k=1}^{|x|}(\forall[i_k \in X])^{t_i[i\kappa[w]]}}=\frac{N_x}{N_x}W(x)$$

Where N_x is the count of X in the database , n is the total number of database records.

Weighted Confidence:

The weighted confidence is the weighted support ratio of the weighted support of $X \cup Y$ and X in the transaction database :

$$W\text{Conf}(X \Rightarrow Y)=\frac{W\text{Sup}(X \cap Y)}{W\text{Sup}(X)}$$

$$W(x)=\frac{N_{X\cup Y}}{N_X}\cdot\frac{\prod_{k=1}^{|X\cup Y|}(\forall[i_k \in X\cup Y])^{t_i[i\kappa[w]]}}{\sum_{k=1}^{|x|}(\forall[i_k \in X])^{t_i[i\kappa[w]]}}\cdot\frac{\prod_{k=1}^{|x|}(\forall[i_k \in X])^{t_i[i\kappa[w]]}}{\sum_{k=1}^{|X\cup Y|}(\forall[i_k \in X\cup Y])^{t_i[i\kappa[w]]}}$$

Where, $N_{X\cup Y}$ is the count of $X \cup Y$ in the database.
Fuzzy association Rule Mining (FARM), is used to mine fuzzy association rules for quantitative values [9].FARM is an efficient solution for a special case that user-supplied thresholds are hard to determine. Using fuzzy set concept, the discovered rules are more understandable to human. In real world applications, transaction data are usually composed of quantitative values.
Most of the researchers have consider weight of item based on its monitory value which will be helpful to extract the items which are having high monitory value but less frequency, WAR[17],WARM[15]. Most of the work done on pre-assign weight but in Time-interval weighted sequence (TiWS) weights is calculated on generation time and time-interval[8].Few work has been done in dynamic weight assignment. Each page of web site has different importance. So, it’s not possible to pre-assign the weight to items. This difficulty been solved by Dynamic significance is assign to items: Frequent sequential traversal pattern mining with weight (FSTPMW)[14]. New measure w-support is defined to give the significance of item sets and applicable to those problems where item does not having pre-assigned weights [12].

Table 1 : Comparative Study of Weighted Association Rule/Sequential Mining

Weighted Association Rule (WAR)[17]	
Core idea	Associate weight parameter with each itemuses a post-processing approach by deriving the maximum weighted rules from frequent itemsets.
Method	breadth first traversal Two fold approach: Generate frequent item sets. Consider only frequency without considering weight. Apply conventional frequent item discovery algorithm. For each item set , Find WAR(s) that meets support , confidence and density threshold. WAR derived using an ”ordered” shrinkage approach.
Advantages	Shorter average execution time as compare to conventional method. Produce high quality result
Weighted Sequential pattern mining WSpan[16]	
Core idea	Worked on weighted item.
Method	Weight are assign to items’ importance or priority. Weight are normalized as minw <= weight <= maxw Use the prefix projected sequential pattern growth method with two measures of weight and support) to prune items
Advantages	WSpan generates fewer patterns than SPAM[3] by adjusting the weight range. WSpan is faster than SPAM The number of patterns discovered by WSpanfewer than the number of sequential patterns found by SPAM with the same minimum supports
WARM (Weighted Association Rule Mining)[15]	
Core idea	The problem of invalidation of the “downward closure property” in the weighted setting is solved by using an improved model of weighted support measurements and exploiting a “weighted downward closure property”
Method	Frequent pattern/association rule mining algorithm is applied. Use “significant – weighted support” metric framework instead of the “large – support” framework used in previous works Consider only weight values (not consider supports of patterns)
Advantages	Solve the downward closure property by developing a weighted downward closure property Algorithm is scalable
Ke Sun and Fengshan Bai [12]	
Core idea	Algorithm is applicable to those problem where item does not having pre-assigned weights, such as web site click-stream data W-support: A new measurement is introduced Hyperlink-induced Topic search (HITS) apply link-based models to association rule mining
Method	new measure w-support is defined to give the significance of item sets. Apriori-like algorithm is proposed based on w-support and w-confidence.
Advantages	w-support, we are able to discover some significant item sets that are not frequent.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Disadvantage	w-support measurement is not recommended for dense data sets.
Frequent sequential traversal pattern mining with weight (FSTPMW) [14]	
Core idea	Each page of web site has different importance. So, can't assign the weight to items. Dynamic significance is assign to items.
Method	The information gain of each item is calculated which is used to mine surprising pattern.
Advantages	Information gain metric helps to discover surprising patterns. Downward closure property is maintained.
Time-interval weighted sequence(TiWS) [8]	
Core idea	Weight is calculated based on Generation times and time-intervals
Method	Time-interval weight from the time-interval and the strength of each pair of data elements are calculated. Weight is calculated with mean value (meanTI approach), standard deviation value (sdTI approach) and coefficient of variation(cvTI approach) for time-interval for pairs of consecutive data elements.
Advantages	Find more interesting sequential patterns in a sequence database. cvTI approach perform better as compare to other two.
Fuzzy Weighted Association Rule Mining (FWARM) [13]	
Core idea	Weighted Association Rule Mining (WARM) [15] with fuzzy weighted support and confidence
Method	Two-fold pre processing approach Uses breadth first traversal of Apriori Weighted support and confidence framework for both boolean and quantitative items for classical and fuzzy WARM.
Advantages	It handles downward closure property, solved using fuzzy weighted support and confidence.
Association rule mining algorithm of dual confidence (DPNAR) [18]	
Core idea	Introduce the concept of weighted dual confidence
Method	solve a great number of redundant and wrong association rules that the association rules based on "support – dual confidence – correlation framework" , Generate association rules using frequent itemset and using the correlation between items using positive and negative association rule
Advantages	The algorithm can reduce to produce a large number of useless and wrong association rules. Mine a large number of significant negative association rules and overcome the shortcomings of low efficiency and not enough accuracy of the traditional algorithm.
Disadvantage	Algorithm assumes that each item in the database has the same importance and effect, but the actually each items are having different weights.

III. EXPERIMENTAL RESULTS

In this section we have performed a simulation study to compare the performances of the conventional Apriori [2], FARM[9] and weighted Pattern Mining Algorithms: WARM[15], FWARM [13]. Comparison is based on frequent sequence patterns on various (20 % to 60%.) support threshold. We have also perform simulation study of generation of Association rule for conventional tree based FP-Growth [7], WARM[15], FARM[9], FWARM[13].

These algorithms were implemented in Sun Java language and tested on an Intel Core Duo Processor with 2GB main memory under Windows XP operating system. Dataset description is given below. Following is the description of Dataset:

Table 2: Dataset Description

DataSet	breast.D20.N699.C2.
Number of records |N|	699
Number of columns |D|	20

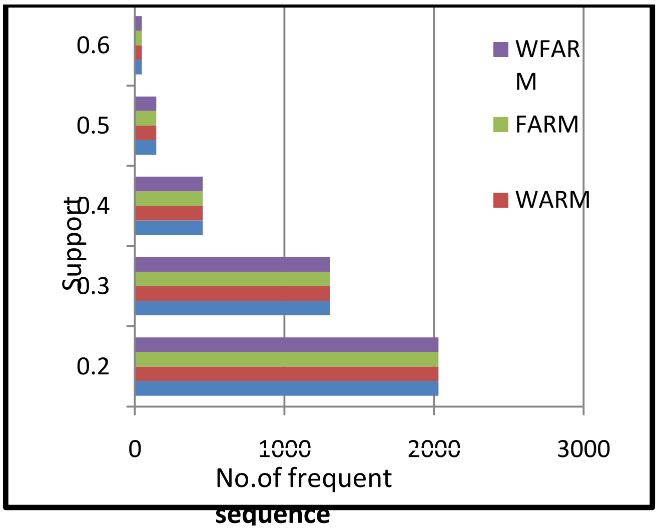

Fig 1: *No. of frequent sequence vs. Support*

On comparing the different algorithms above results have been obtained. The following points can be observed from above simulation:

- Same no. of frequent sequence are generated by weighted algorithms: WARM, FuzzyARM as well with Weighted Fuzzy ARM and non-weighted conventional algorithm Apriori.(fig.1)
- Number of sequential pattern is decreasing by 36%-66%-69%-67% with respect to increasing support threshold values (from 20% to 60%).(fig.1)
- Out of 699 records and 20 various data items 98% more associations are generated by tree based FP-Growth with respect to weighted Apriori algorithms: WARM, FARM, WFARM for lower threshold(20%).Number of association generation is decreasing by 96.5% and 83.5% in case of 30% and 40% support values. For higher values of support

(50%-60%) both weighted and Tree based non- weighted algorithm generates exactly same association. 78% less associations are generated in case of higher support threshold (60%) for WARM, FARM, WFARM. (fig.2)

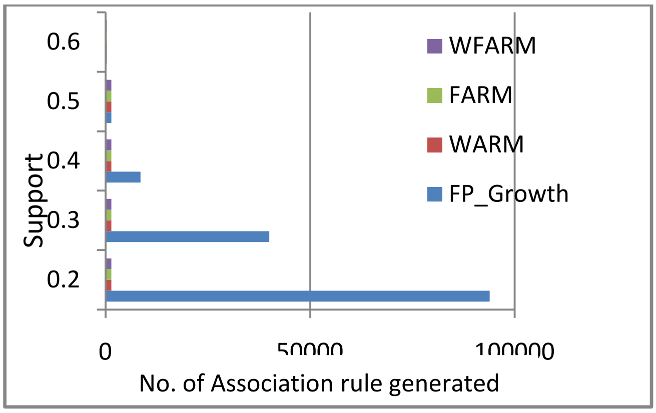

Fig 2: *No. of Association rule generated Vs. Support*

IV. CONCLUSION AND FUTURE SCOPE:

Lots of work already been done in field of sequential pattern mining. Most of the existing SPM algorithms work on objective measures Support and Confidence. All the items are having equal weight but in real life each item is having different significance. So it is important to consider weight of each item. Result of simulation shows, Sequential pattern with weighted framework generate less number of association or more interesting associations as compared to support and confidence based framework. The items or transactions are having more potential in terms of benefit can be selected.

Comparatively less work has been done in field of weighted sequential mining. Most of the researchers have consider weight of item based on its monitory value which will be helpful to extract the items which are having high monitory value but less frequency. Most of the research done in field of pre-assign weight but very few work has been done in dynamic weight assignment which is really necessary in domain of web-log analysis because every page of web site has different importance. So, it's not possible to pre-assign the weight to pages. Little work has been done on weight based on gap or time-interval. Still there are some unexplored constraint like recencey, length, aggregation etc. which has not been considered as weight measure of item, which can be leads to emerging pattern. Profitable sequential pattern in long time and short time can also be identifying through novel weight constraint.

REFERENCE

[1] R. Agrawal and R. Srikant, "Fast Algorithms for Mining Association Rules," Proc. 1994 Int'l Conf. Very Large Data Bases (VLDB '94), pp. 487-499, Sept. 1994.

[2] Agrawal R. And Srikant R. 'Mining Sequential Patterns.', In Proc. of the 11th Int'l Conference on Data Engineering, Taipei, Taiwan, March 1995

[3] AYRES, J., FLANNICK, J., GEHRKE, J., AND YIU, T., 'Sequential pattern mining using a bitmap representation', In Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining-2002.

[4] D. Chiu, Y. Wu, A.L. Chen : An efficient algorithm for mining frequent sequences by a new strategy without support counting, in: Proc. the Twentieth International Conference on Data Engineering, March/April 2004, 2004, pp. 375–386.

[5] M. Garofalakis, R. Rastogi, and K. Shim, 'SPIRIT: Sequential pattern mining with regular expression constraints', VLDB'99, 1999.

[6] Han J., Dong G., Mortazavi-Asl B., Chen Q., Dayal U., Hsu M.-C.,' Freespan: Frequent pattern-projected sequential pattern mining', Proceedings 2000 Int. Conf. Knowledge Discovery and Data Mining (KDD'00), 2000, pp. 355-359.

[7] J. Han, J. Pei, and Y. Yin, 'Mining Frequent Patterns without Candidate Generation', Proc. 2000 ACM-SIGMOD Int'l Conf. Management of Data (SIGMOD '00), pp. 1-12, May 2000.

[8] Joong Hyuk Chang a, Nam Hun Park : Comparative analysis of sequence weighting approaches for mining time-interval weighted sequential patterns Expert Systems with Applications Science Direct

[9] Keith C. C. Chan and Wai-Ho Au, Mining Fuzzy Association Rules, In Proceeding of the 6" lnfemutionul Conference on ltnformution and Knciwledge Munugemetit, Pages 209-2 15, 1997

[10] J. Pei, J. Han, B. Mortazavi-Asi, H. Pino, 'PrefixSpan: Mining Sequential Patterns Efficiently by Prefix- Projected Pattern Growth', ICDE'01, 2001

[11] Srikant R. and Agrawal R.,'Mining sequential patterns: Generalizations and performance improvements', Proceedings of the 5th International Conference Extending Database Technology, 1996, 1057, 3-17.

[12] Ke Sun and Fengshan Bai : Mining Weighted Association Rule without Preassigned Weights IEEE Transactions on Knowledge and Data engineering Vol. 20, No. 4, April 2008, pp. 489-495

[13] M. Sulaiman Khan, Maybin Muyeba, Frans Coenen : Fuzzy Weighted Association Rule Mining with Weighted Support and Confidence Framework

[14] S.P.Syed Ibrahim and K.R.Chandran: COMPACT WEIGHTED CLASS ASSOCIATION RULE MINING USING INFORMATION GAIN , International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.1, No.6, November 2011 DOI : 10.5121/ijdkp.2011.1601 1

[15] F. Tao: Weighted association rule mining using weighted support and significant framework, in: Proc. the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 2003, 2003, pp. 661–666.

[15] Unil Yun: A new framework for detecting weighted sequential patterns in large sequence databases, Science direct Knowledge-Based Systems 21 (2008) 110–122

[16] W. Wang, J. Yang, and P.S. Yu, "Efficient Mining of Weighted Association Rules (WAR)," Proc. ACM SIGKDD '00, pp. 270-274, 2000.

[17] Yihua Zhong,Yuxin Liao Mining Effective and Weighted Association Rules Based on Dual Confidence Fourth International Conference on Computational and Information Sciences Research 2012

[18] M. Zaki, 'SPADE: An Efficient Algorithm for Mining Frequent Sequences', Machine Learning, vol. 40, pp. 31-60, 2001.

.