History Generalized Pattern Taxonomy Model for Frequent Itemset Mining

Jibin Philip
Second Year M.E. (Computer Science and Engineering)
Maharaja Prithvi Engineering College, Avinashi – 641654

Abstract--- Frequent itemset mining is a widely exploratory technique that focuses on discovering recurrent correlations among data. The steadfast evolution of markets and business environments prompts the need of data mining algorithms to discover significant correlation changes in order to reactively suit product and service provision to customer needs. Change mining, in the context of frequent itemsets, focuses on detecting and reporting significant changes in the set of mined itemsets from one time period to another. The discovery of frequent generalized itemsets, i.e., itemsets that 1) frequently occur in the source data, and 2) provide a high-level abstraction of the mined knowledge, issues new challenges in the analysis of itemsets that become rare, and thus are no longer extracted, from a certain point. This paper proposes a novel kind of dynamic pattern, namely the HIstory GENeralized Pattern (HIGEN), that represents the evolution of an itemset in consecutive time periods, by reporting the information about its frequent generalizations characterized by minimal redundancy (i.e., minimum level of abstraction) in case it becomes infrequent in a certain time period. To address HIGEN mining, it proposes HIGEN MINER, an algorithm that focuses on avoiding itemset mining followed by postprocessing by exploiting a support-driven itemset generalization approach.

I. Introduction

Frequent itemset mining is a widely exploratory technique that focuses on discovering recurrent correlations among data. The steadfast evolution of markets and business environments prompts the need of data mining algorithms to discover significant correlation changes in order to reactively suit product and service provision to customer needs. Change mining, in the context of frequent itemsets, focuses on detecting and reporting significant changes in the set of mined itemsets from one time period to another. The discovery of frequent generalized itemsets, i.e., itemsets that 1) frequently occur in the source data, and 2) provide a high-level abstraction of the mined knowledge, issues new challenges in the analysis of itemsets that become rare, and thus are no longer extracted, from a certain point. This paper proposes a novel kind of dynamic pattern, namely the HIstory GENeralized Pattern (HIGEN), that represents the evolution of an itemset in consecutive time periods, by reporting the information about its frequent generalizations characterized by minimal redundancy (i.e., minimum level of abstraction) in case it becomes infrequent in a certain time period. To address HIGEN mining, it proposes HIGEN MINER, an algorithm that focuses on avoiding itemset mining followed by post processing by exploiting a support-driven itemset generalization approach. To focus the attention on the minimally redundant frequent generalizations and thus reduce the amount of the generated patterns, the discovery of a smart subset of HIGENs, namely the NONREDUNDANT HIGENs, is addressed as well. Experiments performed on both real and synthetic datasets show the efficiency and the effectiveness of the proposed approach as well as its usefulness in a real application context.

II. Existing System

HIGEN mining may be addressed by means of a postprocessing step after performing the traditional generalized itemset mining step , constrained by the minimum support threshold and driven by the input taxonomy, from each timestamped dataset. However, this approach may become computationally expensive, especially at lower support thresholds, as it requires 1) generating all the possible item combinations by exhaustively evaluating the taxonomy, 2) performing multiple taxonomy evaluations over the same pattern mined several times from different time periods, and 3) selecting HIGENs by means of a, possibly time-consuming, postprocessing step. To address the above issues, I propose a more efficient algorithm, called HIGEN MINER. It introduces the following expedients: 1) to avoid generating all the possible combinations, it adopts, similarly to , an Apriori-based supportdriven generalized itemset mining approach, in which the generalization procedure is triggered on infrequent itemsets only.

A. *Disadvantages Of Existing System:*

- More Resource Consumption.
- More Processing Time.

III. Proposed System

Frequent weighted itemset represent correlations frequently holding in data in which items may weight differently. However, in some contexts, e.g., when the need is to minimize a certain cost function, discovering rare data correlations is more interesting than mining frequent ones. This paper tackles the issue of discovering rare and weighted itemsets, i.e., the Infrequent Weighted Itemset (IWI) mining problem. Two novel quality

A. *Advantages of Proposed System*

- Less Resource Consumption.
- Less Processing Time.
- Fast Access
- Easy Interaction to System

IV. Implementation

This paper tackles the issue of discovering rare and weighted item sets, i.e., the Infrequent Weighted Itemset (IWI) mining problem. Two novel quality measures are proposed to drive the IWI mining process. Furthermore, two

algorithms that perform IWI and Minimal IWI mining efficiently, driven by the proposed measures, are presented. Experimental results show efficiency and effectiveness of the proposed approach.

The lists of modules used .

- Data Acquisition
- HIGEN
- FP-Growth
- Result
- Comparison

A. Data Acquisition:

This module is where data required for testing the project is undertaken. There are two kinds of dataset available for processing the data mining applications. One is synthetic and other is real time dataset. The process of acquiring the dataset is carried on this module. Once the data set is acquired. It has to be converted into suitable structure for further processing by the algorithm. Java collections are used to represent the data from the dataset.

B. Higen Algoritm:

Algorithm 1 reports the pseudocode of the HIGEN MINER. The HIGEN MINER algorithm iteratively extracts frequent generalized itemsets of increasing length from each timestamped dataset by following an Apriori-based level-wise approach and directly includes them into the HIGE

C. Fp-Growth:

Given a weighted transactional dataset and a maximum IWI-support (IWI-support- min or IWI-supportmax) threshold ξ, the Infrequent Weighted Itemset Miner (IWI Miner) algorithm extracts all IWIs whose IWIsupport satisfies ξ (Cf. task (A)). Since the IWI Miner mining steps are the same by enforcing either IWIsupport- min or IWI-support-max thresholds, we will not distinguish between the two IWI-support measure types in the rest of this section.

D. Result:

The result module displays the output of the clustering, The output are shown as tabular data.

E. Comparison:

In Comparison module the algorithm is compared based on different techniques. The basic techniques included time and space complexity.

1) *Time Complexity:* The total time required for the algorithm to run successfully, and produce
an output.

T(A) = End Time – Start Time.

*2) Space Complexity :*The space complexity is denoted by the amount of space occupied by the variables or data structures while running the algorithm.

S(A) = End { Space(Variables) + Space(Data Structures) } – Start { Space(Variables) + Space(Data Structures)}

V. CONCLUSION

This paper proposes a novel kind of dynamic pattern, namely the HIstory GENeralized Pattern (HIGEN), that represents the evolution of an itemset in consecutive time periods, by reporting the information about its frequent generalizations characterized by minimal redundancy (i.e., minimum level of abstraction) in case it becomes infrequent in a certain time period. To address HIGEN mining, it proposes HIGEN MINER, an algorithm that focuses on avoiding itemset mining followed by postprocessing by exploiting a support-driven itemset generalization approach. To focus the attention on the minimally redundant frequent generalizations and thus reduce the amount of the generated patterns, the discovery of a smart subset of HIGENs, namely the NONREDUNDANT HIGENs, is addressed as well. Experiments performed on both real and synthetic datasets show the efficiency and the effectiveness of the proposed approach as well as its usefulness in a real application context.

There are different types of facilities included in the future enhancement model. The FP Growth and its advanced algorithms providing both the frequent and infrequent item set mining in fast and easy way.With the less amount of time the mining can be possible and can be provide fast access from database.The main advantages in the future enhancement are fast mining with less amount of itemset.This also provide easy interaction to the system.

ACKNOWLEDGMENTS

I express my sincere and heartfelt thanks to our chairman Thiru. K.PARAMASIVAM B.Sc., and our Correspondent Thiru. P.SATHIYAMOORTHY B.E., MBA., MS., for giving this opportunity and providing all the facilities to carry out this project work.

I express my sincere and deep heartfelt special thanks to our Respected Principal Dr. A.S.RAMKUMAR, M.E., Ph.D., MIE., for provided me this opportunity to carry out this project work.

I wish to express my sincere thanks to Mrs.A.BRINDA M.E., Assistant Professor, and Head of the Department of Computer Science and Engineering for all the blessings and help rendered during the tenure of my project work.

I am indebted to my project guide, Mr.K.Moorthy M.E., Assistant Professor for his constant help and creative ideas over the period of project work.

I express my sincere words of thankfulness to members of my family, friends and all staff members of the Department of Computer Science and Engineering for their support and blessings to complete this project successfully.

REFERENCES

[1] Luca Cagliero "Discovering Temporal Change Patterns in the Presence of Taxonomies" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINSEERING, VOL 25 NO 3, MARCH 2013

[2] Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, Andrew Joydeep Ghosh "Taxonomy with Bregman Divergences" A ACM Computing Surveys, Vol. 31,No.3 September 1999.D

[3] avid M. Blei, Andrew Y. Ng,"Frequent Itemset Allocation", J. Machine Learning Research, vol. 6, pp. 1705-1749, 2005.

[4] Inderjit S. Dhillon, Subramanyam Mallela ,Rahul Kumar "Divisive Information- Theoretic Feature Algorithm for Text Classification" J. Machine Learning

Research, vol. 3, pp. 993-1022, 2003.
[5] Jia-Ling Koh and Yuan-Bin Do " Approximately Mining Recently Representative"
[6] J. Machine Learning Research, vol. 3, pp. 1265-1287, 2003.
[7] Rakesh Agrawal, Tomasz Imielinski , Arun Swami "Mining Association Rules between Sets of Items in Large Database" Proc. ACM SIGMOD-SIGACTSIGART Symp. Principles of Database Systems (PODS), 2007
[8] T. Blaschke "TOWARDS A FRAMEWORK FOR CHANGE DETECTION BASED IMAGE OBJECTS", "Latent Dirichlet Allocation", J. Machine Learning Research, vol. 3, pp. 993-1022, 2003.

