International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
©2015INPRESSCO®, All Rights Reserved
Available at http://inpressco.com/category/ijcet

Research Article

Change Detection through Automatic Inference and Multiple Taxonomies

Geeta V. Poul[†*] and P.D. Lambhate[†]

[†]Department of Computer Engineering, Sacitribai Phule Pune University, Pune, Maharashtra, India

Accepted 27 July 2015, Available online 29 July 2015, **Vol.5, No.4 (Aug 2015)**

Abstract

Data mining is used to find interesting information from the raw data. The frequent itemset mining used to find number of itemsets occurred more number of times in particular time duration. It may happen that a particular item occurs for a very specific time but its frequency is more. Such itemsets are considered as non-redundant itemsets. Thus the study of temporal data mining (change mining) is important. Number of data mining algorithms introduced to find frequent itemsets in the data. The work is based on HIGEN miner algorithm to find redundant as well as non-redundant itemsets. The proposed work finds HIStory GENeralized pattern (HIGEN) using automatic inference taxonomy in a very less time. The experiment performed on both synthetic and real time datasets to find value satisfying minimum support at higher level of taxonomy.

Keywords: *Data mining; Minimum support; Association rule; change mining.*

1. Introduction

The data mining refers to the knowledge discovery in databases. The goal of data mining is to extract information from the dataset and make it into an understandable form. Data mining software is an analytical tool for analyzing data from different perspective, categorizing it, and summarizing relationship. The temporal data mining discovers how the interesting pattern varies over time. Thus the study of change mining has been evolved and become important.

A. Motivation

The necessity of data from different users changes in different times. For example we can find more reservations to the airlines during summer seasons than any other season. Or there are more peoples found online during night time than any other time as people are free in this duration. Even in mall sections we can found there is rush of customer in particular section according to the season. All these needs motivate to study the change mining or temporal data mining.

B. Background Needs

Frequent item set mining algorithm is constrained by a minimum support threshold to discover patterns whose observed support in the source data is greater than or equal to a given threshold. Generalized itemsets, which have been first introduced in the context of market basket analysis, are itemsets that provide a high level abstraction of the mined knowledge.

Table 1 Network traffic in Oct 2014

Date	Time	Location	Social website
02/10/2014	4 pm	Latur	LinkedIn
02/10/2014	7 pm	Pune	facebook
02/10/2014	7 am	Bidder	facebook
15/10/2014	10 am	Bangalore	LinkedIn
15/10/2014	11 pm	Maharashtra	facebook
15/10/2014	6 pm	Bidder	LinkedIn
27/10/2014	2 pm	Pune	LinkedIn
27/10/2014	5 pm	Karnataka	LinkedIn
27/10/2014	7 am	Latur	facebook

Table 2 Network traffic in Nov.2014

Date	Time	Location	Social website
02/11/2014	1 pm	Latur	LinkedIn
02/11/2014	10 pm	Pune	facebook
15/11/2014	3 pm	Bangalore	LinkedIn
15/11/2014	7 pm	Maharashtra	facebook
27/11/2014	9 am	Pune	LinkedIn
27/11/2014	11 am	Karnataka	LinkedIn
27/11/2014	08 am	Bidder	facebook

*Corresponding author: **Geeta V. Poul**

The table 1 and 2 shows datasets collected during October 2014 and November 2014. The two datasets shows network traffic on the two social networking sites namely LinkedIn and Facebook in two different states of country India. The analysis performed in two cities of each state in different time periods. The taxonomy structure is as shown in figure.

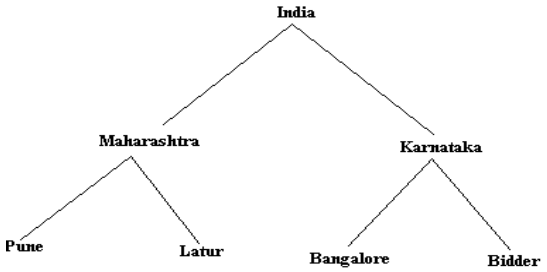

Fig.1 States: Taxonomy of States

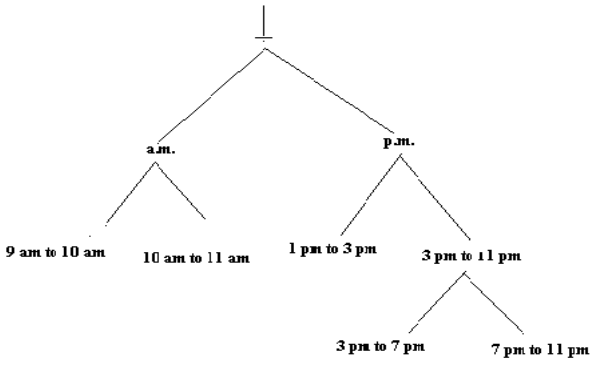

Fig.2 Time: Taxonomy of Time

Table 3 Generalised ITEMSEt

Generalized Itemset	Sup D1	Sup D2
{Facebook}	3	3
{LinkedIn}	4	5
{Facebook, pm}	2	2
{LinkedIn, pm}	2	4

Table 4 Nongeneralised Itemset

Not generalized Itemset	Sup D1	Sup D2
{facebook, am}	1(inf)	2
{LinkedIn, am}	2	1 (inf)

HIGEN's Extracted from D1 to D2

Itemset	Support D1	Support D2
{facebook}	3	3
{LinkedIn}	4	5
{facebook, pm}	2	2
{LinkedIn, pm	2	5

From the analysis we calculated support of each item and categorizes into generalized itemsets and non-generalized itemsets. The generalized itemsets are those that satisfy minimum support value in both time duration (i.e. November and December 2014). Whereas non-generalized itemsets are those whose minimum support value is not satisfying in one of both month. Given that minimum support value is 2. Sometimes it is found that although itemsets are non-generalized; their support is great for particular time. Thus these itemsets should also be considered.The taxonomy structure of different time of a day during which these social networking sites are crowded. If the minimum support value is not satisfied at lower taxonomy level then it is checked for an upper level.

2. Literature Survey

Different data mining algorithms used to calculate frequent item sets which is useful for analysis purpose.

Rakesh Agrawal and Arun Swami worked on large databases. The work focuses on analysis of past transaction to discover new rules which in turn shows association between items. The work derives qualitative rules but it does not focus on classification of items and not having pre-described classes. The algorithm gets success in buffer management. It also prunes unwanted data and reduces number of passes.

G.V. Poul et al (2015): Focus on calculating confidence of item satisfying minimum support value. During mining process many interesting as well as uninteresting patterns are generated. The algorithm uses taxonomy to prune uninteresting or redundant pattern. The paper also compare three algorithms namely Basic, Estmerge, Cumulate by varying minimum support value, number of transaction, number of items per level and number of roots.

Luca Cagliero et al (2013): The data mining paradigm that combines recent work in data mining with rich literature. The discovered rule from continuously mined data is added to common rule-base. This method includes query and history pattern of discovered rule.

(R.Agrawal et al, 193): The paper includes association rule that that changes according to time period. The DAR-C consists of rule along with comment which specifies when to apply the rule.

(R.Agrawal) et al, 1995Many algorithms used pruning step to remove non-redundant or uninteresting itemsets but it is found that such itemsets are found interesting for particular duration. The paper considers time duration to find relationship between items.

3. Problem Statement

For a given taxonomy, datasets i.e. collected at different durations of time and a minimum support value, the HIGEN MINER algorithm extracts HIGEN

and NONREDUNDANT HIGEN provided that minimum support and taxonomy is selected by user.
We find generalized as well as non-generalized itemset whose support value equal to or exceed minimum support value in frequent item set mining.

The association rule can be expressed in two step procedure:

1. Exaction of frequent and generalized itemset
2. Rule generation for given frequent itemset.

4. System Architecture

At present time business intelligence is vital tool to transform raw data into meaningful as well as useful information for business analysis. It is process of data analysis intended to boost business performance by allowing corporate executives and other end users make more informed decisions. The unstructured data coming from different real time application context varies from time to time. It shows how customer interest gets changed over time. Thus there is need to study and take perfect decisions on these data so that service providers understand what actual customer interest is in particular duration of time to satisfy them in a better way.

1.Use of multiple taxonomies for real application data and automatic inference.

2. Reduce time and space complexity of existing HIGEN miner algorithm.

Through data mining we discover knowledge from databases. As data mining consists of three stages like pattern analysis, preprocessing and pattern discovery, we are dividing our system in these stages.

Pattern Analysis

At this stage we are analyzing our database to find taxonomy hierarchy in the datasets i.e. we find which terms are related to one another according to levels.

Preprocessing

In this stage, as dataset consists of number of records, we are arranging the dataset such that the items come under taxonomy term are placed one after another, so that levels are placed in an order. For example, a state, town/city, village are arranged one after another.

Pattern Discovery

At this stage we are finding support for each item at any one level and checking whether it is less than or greater than specified minimum support value. If calculated support of an item is greater than minimum support value then we are considering as HIGEN otherwise traverse is made to upper level and same procedure is performed.

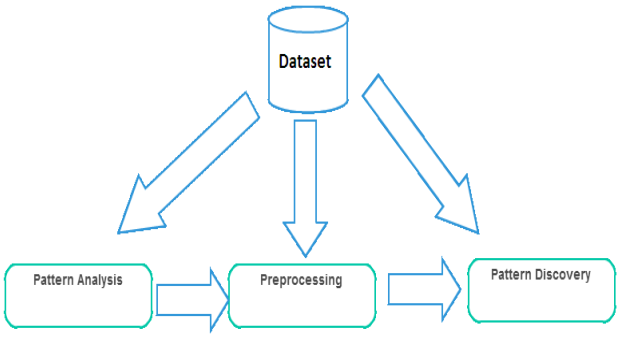

Fig.3 System Architecture

Pattern Analysis

Pattern analysis techniques are used to highlight overall patterns in data and to filter out uninteresting patterns. The techniques like Knowledge Query Mechanism, visualization, OLAP are used for pattern analysis. Here in this technique of data mining HIGEN miner algorithm is used to find generalized pattern.

This type of analysis is done by calculating support i.e. number of occurrences of each item. A value of minimum threshold is considered. If calculated support value is greater than minimum threshold then this value is considered to be important for further evaluation and if calculated support undersize minimum support then that value is neglected.The research work in this paper use HIGEN miner algorithm as a basic concept. In this project time-stamped, structured dataset is considered. The taxonomy is build over each item and If calculated support value exceeds minimum threshold then the itemsets are considered as HIGEN itemsets else the upper level in the taxonomy is studied and same experiment is done till we get HIGEN itemset. The itemset weighted frequently represents correlations frequently appear in data in which items may weights differently.

HIGEN's are classified into monotonous HIGEN's and oscillatory HIGEN's. The monotonous HIGEN's are those which appear only at once in given time periods. Whereas in oscillatory HIGEN's oscillate between two consecutive time periods.

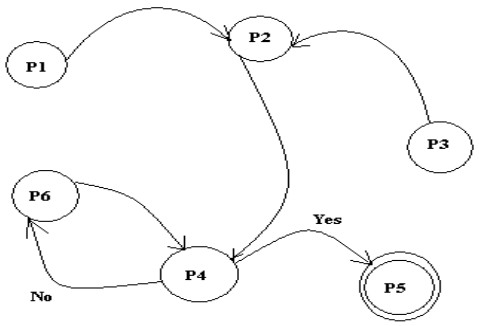

Fig.4 State Transition Diagram

5. Mathematical Modeling

Let system S= {I, O, P,
S,F}
Where, I is set of inputs {Dataset, minimum support value, Taxonomy}
O is output generated {o1, o2}
o1=Preprocessed dataset
o2 = Frequent itemsets at multiple level of taxonomy
P is number of processes
P= {p1, p2, p3, p4, p5, p6}
P1= pre-processing
P2= candidate generation
P3=support calculation
P4=Checking support
P5=HIGEN updation
P6=traverse to upper level of taxonomy
S be the success case
S= HIGEN miner for multiple level of taxonomies
F is of failure cases.

6. Algorithm

Input:
Structured time stamped dataset D, minimum support value (min_sup), taxonomy
Output:

1. Initialize HIGEN with null value (HG = 0)
2. Initialize candidate length to 1
3. Ck = set of item sets in dataset D
4. For each item (c) in k distinct item set
5. Scan dataset D and calculate support (c) in given dataset D
6. End for
7. If support(c)>/min_sup for some Di
8. Then HG= update HIGEN
9. Initialize candidate generalization level to 1
10. Initialize generalized item set container to Phi
11. For all c in ck at level 1
12. If sup(c) in dataset Di<min_sup
13. Then gen(c) go to new generalization level (l+1)
14. Gen(c)= evaluation of taxonomy
15. Gen =Gen U gen(c)
16. End if
17. Ck= Ck U Gen
18. l=l+1
19. until Gen = phi
20. k=k+1
21. until ck=phi
22. return HG

7. Result and Analysis

The results of the algorithm are captured by comparing the algorithm with the existing algorithm based on accuracy and total time required to execute the algorithm.
It is found that the proposed algorithm is more accurate than existing one.

Accuracy: Accuracy refers to the closeness of a measured value to a standard or known value. It can be also define as The ability of a measurement to match the actual value of the quantity being measured

Total Time Required: The total time required to run an algorithm is measuring difference between time at which algorithm start and end.
Total time required = End time - Start Time

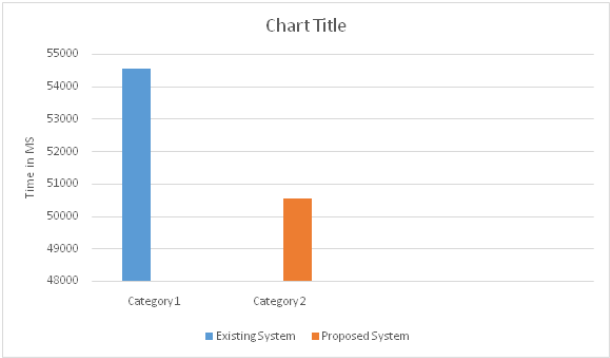

Fig.5 Time Comparison diagram

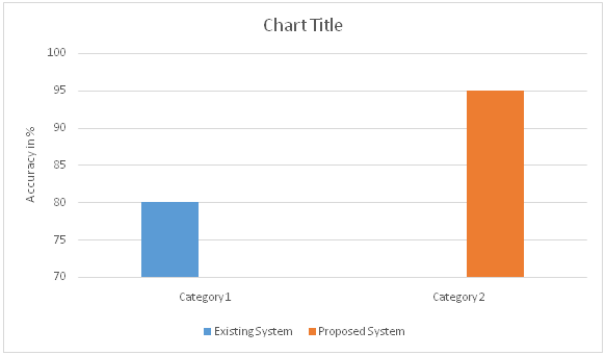

Fig. 6 Accuracy Comparison diagram

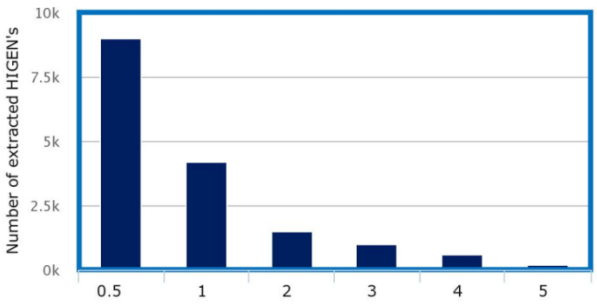

Fig.7 Impact of varying Minimum support value

Conclusion

The work addresses the problem of change mining in the context of frequent itemsets. To represent the

evolution of itemsets in different time periods, it proposes to extract generalized itemsets characterized by minimum redundancy (i.e. minimum abstraction level) in case one itemset becomes infrequent in certain time duration. To this aim, two novel dynamic patterns, namely the HIGENs and the NONREDUNDANT HIGENs for multiple levels of taxonomies have been introduced. The usefulness of the work is to support user and service profiling in a mobile environment has been validated by a domain expert.

References

G.V.Poul, P.D.Lambhate (December 2015), Ascertaining Chronological Changes Patterns in the Presence of Multiple Taxonomies vol.3, issue 1

G.V. Poul,P.D.Lambhate (March 2015), Framework for Change Detection Using Multiple Taxonomies vol.3

Luca Cagliero (March 2013), Discovering Temporal Change Patterns in the Presence of Taxonomies IEEE vol.25, No. 3

R.Agrawal, T.Imielinski, and A. Swami (1993), Mining Association Rules between Sets of Items in Large Databases,ACM SIGMOD Record, vol. 22, pp. 207-216.

R.Agrawal and G.Psaila (1995), Active Data Mining Proc. First int'1 Conf. Knowledge Discovery and Data Mining, pp. 3-8

R.Agrawal and R.Srikant, Mining Generalized AssociationRules, Proc.21th Int'l Conf. Very Large Data Bases (VLDB '95

K. Verma and O.P. Vyas (Sept.2005), Efficient Calendar Based Temporal Association Rule, ACM SIGMOD Record, vol. 34, pp. 63-70.

J. Han and Y. Fu (Sept.19), Mining Multiple-Level Association Rules inLarge Databases IEEE Trans. Knowledge and Data Eng., vol. 11,no. 7, pp. 798-805

TPC-H (2009.), The TPC Benchmark H. Transaction Processing Performance Council, http:// www.tpc.org/ tpch/default.asp

B. Liu, Y. Ma, and R. Lee (2001), Analyzing the Interestingness of Association Rules from the Temporal Dimension, Proc. IEEE Int'l Conf.Data Mining(ICDM),pp.377-384.

P. Giannikopoulos, I. Varlamis, and M. Eirinaki (2010), Mining Frequent Generalized Patterns for Web Personalization in the Presence of Taxonomies, Int'l J. Data Warehousing and Mining, vol. 6, no. 1, pp.58-76.