Ascertaining Chronological Change Patterns in the Presence of Multiple Taxonomies

Poul Geeta Venkatrao
Department of Computer Engineering
JSPM's JSCOE, Hadapsar,
Pune, India
gtpoul@gmail.com

Prof.P.D.Lambhate
Department of Computer Engineering
JSPM's JSCOE, Hadapsar,
Pune, India
lambhatepoonam9@gmail.com

***Abstract*—** Data mining and knowledge discovery (KDD) is the technique of converting raw data into useful information. It is predictive technique for interesting data analysis. Change mining is technique of data mining that finds and reports changes in mined item set from one time to another time. Different data mining algorithms are evolved to show correlation among data mined. The data association changes from one time to time. The project highlights the HIGEN (HIGHLY GENERLISED PATTERN) algorithm that reports minimum level of abstraction of frequently generalized pattern. Association between items shown by algorithm for data coming from real time applications at multiple level of taxonomy. The experiment performed on artificial and factual datasets to show competence and effectiveness of proposed approach as well as usefulness of real time application context.

Keywords-*Data mining; Change mining; Minimum support*

*****

I. INTRODUCTION

A. *Overview*

Association rule mining is a best way to discover interesting relations between variables in large databases. Suppose in a super market A customer who buys outerwear tends to buy shoes.This relationship among data shown through association rule mining. The association rule mining helps seller to keep related stocks together.

B. *Motivation*

A traditional DBMS stores data that is implied to be valid at the current point-in-time, it does not track the past or future states of the data. Many types of data change over time, and different users have requirements to access data at different points in time. Traffic on facebook is more from 7 pm to 12 pm. So it is required to give more support in this duration than any other time duration. Thus the study temporal changes in databases are required for service providers to give a better service to the customers for better customer satisfaction.

C. *Background Need*

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 that provides a high level abstraction of the mined knowledge. 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.

	Date	Time	Location	Product
1	2013-06-02	11.15 p.m.	Bihar	Jeans
2	2013-06-02	10.28 p.m.	Bihar	Jeans
3	2013-06-05	08.51 a.m.	Paris	Jerkin
4	2013-06-05	05.00 p.m.	Cannes	Jerkin
5	2013-06-28	05.05 p.m.	Asam	Jerkin

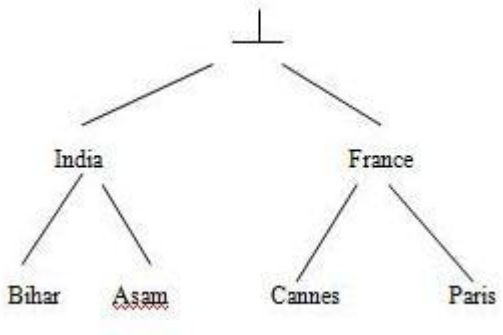
Dataset D1 - Good sales in June 2013

	Date	Time	Location	Good Description
1	2013-07-02	11.00 p.m.	Bihar	Jeans
2	2013-07-03	11.10 p.m.	Asam	Jeans
3	2013-07-05	08.40 a.m.	Paris	Jerkin
4	2013-07-30	05.05 p.m.	Paris	Jerkin
5	2013-07-31	05.00 p.m.	Cannes	Jerkin

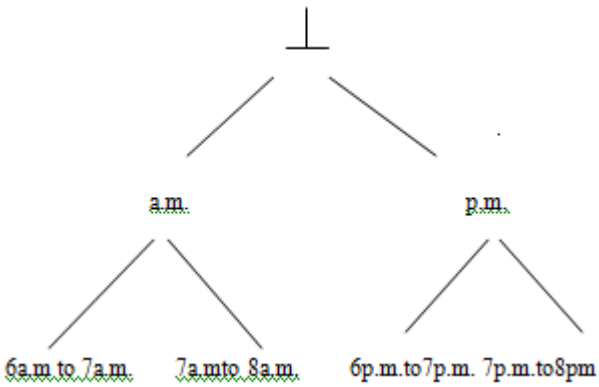
Dataset D2 - Good sales in July 2013

Tables 1 and 2 represent two examples of datasets, related with two following months of year 2013 June and July. Each record be in contacts to a product sale the good description, the time, the Date, and the location in which position are reported. Fig. shows a simple tree structure defined on the location attribute. Table 1 reports the all set of probable frequent non-generalized and generalized itemsets, and their equivalent support values mined from the example datasets D1 and D2,by exploiting the tree reported in Fig. and by implementing an absolute minimum support threshold equal to 2. In both dataset that is D1 and D2 itemset are suit the support threshold and other are frequent in one month out of two months. In the case of frequent itemsets by exploiting

generalized itemsets to represent patterns that become unusual with respect to the support threshold, and therefore they are no longer extracted, at a specific point. This type of algorithm is used to discovering temporal changes. Consider the itemsets (Jeans, Bihar) and (Jerkin, Paris). The former itemset is nonfrequent in D1 means its support count value in D1 is lower than the support threshold and frequent itemset in D2 support value increase from June to July, after that shows an opposite. The discovery of higher level relationship, in the form of generalized itemsets,. Item set generalization allowed only avoiding infrequent itemset. At the same time, the information is provided by the series of generalizations or specializations of the same pattern in successive time durations. The generalized itemset Jeans, India has a specialization Jeans, Bihar) that is infrequent in June and frequent in the July might be relevant for decision making and, reported as a temporal correlation among Recurrences about product Jeans. Same as, the information that, although (Jerkin, Paris) infrequent with respect to the minimum support threshold transferring from June to July, its generalization (Jerkin, France) remains frequent in both months can be relevant for analyst decision making. This type of information can't be mined directly or to the point represented means of any existing concept.



Good sales in June 2013



Good sales in July 2013

We are using one dynamic pattern is History Generalized Pattern that is a higen represents the minimum sequence of generalizations required to keep information provided by a non generalized itemset which is frequent, with respect to the minimum support threshold, in each time duration . If an itemset is frequent in each time duration, the respected higen reports its support variations. Otherwise, when the itemset becomes infrequent at a specific point, the higen reports the minimum number of generalizations required to make its enclosed information frequent at a higher level of abstraction. Frequent generalization at minimum abstraction level represents the information with minimal redundancy.

When an infrequent or not generalized itemset having multiple generalizations belonging to the same minimal aggregation level, number of higens associated with the same itemset.

II. LITURATURE REVIEW

There are different data mining algorithm having some weaknesses. To overcome these drawbacks HIGEN miner algorithm is used. The Association rules are already introduced in number of research papers. All the algorithms use the same scenario to calculate frequent itemset i.e. the itemset supposed to be frequent if whose support is greater than user specified minimum threshold value.

Rakesh Agrawal and Arun Swami researched on large databases [2] .They analyzed past transaction to discover new rules on database. The rule shows associvity between items and derive qualitative rule on it but it does not focus on classification on items and it doesn't have any pre-described classes.

The work is carried out on customers buying habits i.e. if customer purchases certain thing then confidence that he may buy other related thing too[4] if and only if both the item satisfy minimum support value. The same algorithm finds association between item at any taxonomy level.[4] generates many uninteresting rules along with interesting rules so pruning step is used to remove uninteresting rules.[2][4]shows if lower level of taxonomy do not satisfy minimum support value then upper level doesn't satisfy the same in the same way if upper level satisfies the minimum support value the all its subsets satisfies user specified minimum support.

Active data mining [3]method to represent and query the history pattern of the discovered association rule quality indices. It mines rules the rules, from datasets collected in different time periods, by adding rules and their related quality indices (support and confidence). Next, the analyst could specify a history pattern when trigger is fired.

The work [6] carried out which hold time stamped (temporal) association with another item. It is observed that some item are not frequent but for particular time duration they became frequent, So time duration must consider to find associationship between item.[6] uses H-mine and H-struct algorithms which are sensitive to time. H-mine algorithm proposes frequent itemset based on time and H-struct focuses on fast mining on time based datasets.

III. PROBLEM DEFINATION

For a given Taxonomy, timestampped dataset and minimum support (minsup) threshold value,the HIGEN MINER algorithm is highlighted to extract HIGEN and NONREDUNDANT HIGEN. We find generalized and non-generalized itemset whose support value is greater than or equal to minimum support value in frequent item set mining.
The generalized association rule can be expressed by two step procedure-

A. Exaction of frequent generalized itemset

B. Generation of rule for given frequent itemset

IV. SYSTEM ARCHITETURE

Business intelligence (BI)is important tools for the transformation of raw data into meaningful and useful information for business analysis. Business intelligence (BI) is a data analysis process aimed to boost business performance by helping corporate executives and other end users make more informed decisions. The data which is not structured arriving from real time application context changes from one time period to another. It shows how customer interest changes over time. Thus there is need to study and make perfect decisions on these data so that service providers come to know what actual customer interest is in a particular time duration to efficiently satisfy them. let us take an example, A customer buys jerkins or sweaters in a winter seasons whereas cotton cloths in summer. Knowing this analysis the service provider come to know which stock must be kept more in the market.

The system architecture include data mining stages like preprocessing, pattern discovery and pattern analysis.

A. Preprocessing

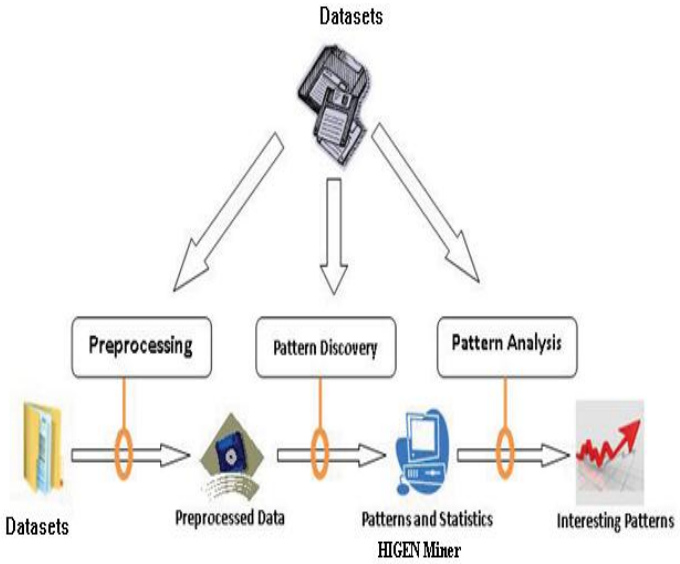
pre-processing is process of converting raw information coming from various data sources into the data abstractions which is necessary for pattern discovery The data collected at different times having different datasets is incomplete and not suitable for mining of data. Pre-processing is necessary to convert the data into suitable form for discovery of pattern. It can provide accurate, concise data for data mining. Pre-processing of data includes data cleaning, user and user sessions identification and integration of data.

B. Pattern Discovery

To discover the novel, interesting and potentially useful information, data mining algorithms are applied such as Association Rules, Path Analysis, Clustering and Classification, Sequential Patterns etc. Here in this method we employ association rule in different entities. For this we applied logical AND operation among entities.

C. 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, OLAP and visualization are used for pattern analysis. Here in this technique of web mining HIGEN miner algorithm is applied to find generalized pattern.



Architeure

This type of analysis is done by calculating support for each item. A value of minimum support (threshold) is considered. If calculated support value is greater than minimum threshold then this value is considered to be interesting for further study and if calculated support is undersized to minimum support then that value is neglected. The research work in this project is based on HIGEN miner algorithm. In this project time-stamped dataset is considered. The taxonomy is build over each item and If calculated support value exceeds minimum threshold then the itemset is considered to be HIGEN itemset else the upper level in the taxonomy is considered and same experiment is done till we get HIGEN itemset. Frequent weighted itemset represent correlations frequently appear in data in which items may weight differently.

V. PROPOSED SYSTEM

A. The automatic inference and multiple taxonomies evolution for real application data (e.g., social network data).
B. Getting complex HIGEN constraints quality into the process of mining.

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 minimal 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, have been introduced. The usefulness of the proposed approach to support user and service profiling in a mobile environment has been validated by a domain expert.

ACKNOWLEDGMENT

I am thankful to Prof. P. D Lambhate for her expert guidance and continuous encouragement throughout the work.

REFERENCES

[1] Luca and Cagliero, "Discovering Temporal Change Patterns in the Presence of Taxonomies " IEEE vol.25, NO. 3 ,March 2013

[2] R. Agrawal, T. Imielinski, and A. Swami, Mining Association Rules between Sets of Items in Large Databases, ACM SIGMOD Record, vol. 22, pp. 207-216, 1993.S. Zhang, C. Zhu, J. K. O. Sin, and P. K. T. Mok, "A novel ultrathin elevated channel low-temperature poly-Si TFT," *IEEE Electron Device Lett.*, vol. 20, pp. 569–571, Nov. 1999.

[3] R. Agrawal and G. Psaila, Active Data Mining Proc. First Intl Conf. Knowledge Discovery and Data Mining, pp. 3-8, 1995.

[4] R. Agrawal and R. Srikant, Mining Generalized Association Rules, Proc. 21th Intl Conf. Very Large Data Bases (VLDB 95), pp. 407-419, 1995

[5] W.-H. Au and K.C.C. Chan, Mining Changes in Association Rules: A Fuzzy Approach, Fuzzy Sets Systems, vol. 149, pp. 87-104, Jan. 2005.

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

[7] S. Baron, M. Spiliopoulou, and O. Gnther, Efficient Monitoring of Patterns in Data Mining Environments Advances in Databases and Information Systems, L. Kalinichenko, R. Manthey, B. Thalheim, and U. Wloka, eds., vol. 2798, pp. 253-265, Springer, 2003.

[8] M. Bo ttcher, D. Nauck, D. Ruta, and M. Spott, Towards a Framework for Change Detection in Datasets Research and Development in Intelligent Systems XXIII, M. Bramer, F. Coenen, and A. Tuson, eds., pp. 115-128,Springer, 2007.