Generating of interesting patterns for text mining

V.Shankar Ganesh, G. Surendra Reddy, N.S.Jagadeesh
Department of Computer Science & Engineering, Kuppam Engineering College,
Kuppam, Andhra Pradesh, India.

Abstract

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Data Mining Approach is used which is a sophisticated statistical processing or artificial intelligence algorithms to discover useful trends and patterns from the extracted data so that it can yield important insights including prediction models and associations that can help companies understand their customer better.

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both.

Examining and analyzing the data can turn raw data into valuable information about customer's needs. It helps for finding frequent item sets is one of the most investigated fields of data mining. The Apriori algorithm is the most established algorithm for frequent item sets mining (FIM). Several implementations of the Apriori algorithm have been reported and evaluated. We revised Apriori implementation into a parallel one where input transactions are read by a parallel computer. The effect a parallel computer on this modified implementation is presented.

Keywords - Apriori, Association Rules, Data Mining, Frequent Itemsets Mining (FIM), Parallel Computing, Clustering.

Introduction

Enterprise data mining applications, such as mining public service data and telecom fraudulent activities, inevitably involve complex data sources, particularly multiple large scale, distributed, and heterogeneous data sources embedding information about business transactions, user preferences, and business impact. In these situations, business people certainly expect the discovered knowledge to present a full picture of business settings rather than one view based on a single source. Knowledge reflecting full business settings is more business friendly,

comprehensive, and informative for business decision makers to accept the results and to take operable actions accordingly.

It is challenging to mine for comprehensive and informative knowledge in such complex data suited to real-life decision needs by using the existing methods. The challenges come from many aspects, for instance, the traditional methods usually discover homogeneous features from a single source of data while it is not effective to mine for patterns combining components from multiple data sources. It is often very costly and sometimes impossible to join multiple data sources into a single data set for pattern mining.

In our project, we proposed the concepts of *combined association rules*, *combined rule pairs*, and *combined rule clusters* to mine for informative patterns in complex data by catering for the comprehensive aspects in multiple datasets. A combined association rule is composed of multiple heterogeneous itemsets from different data sets while combined rule pairs and combined rule clusters are built from combined association rules. Analysis shows that such combined rules cannot be directly produced by traditional algorithms such as the FPGrowth. This paper builds on the existing works and proposes the approach of *combined mining* as a general method for directly identifying patterns enclosing constituents from multiple sources or with heterogeneous features such as covering demographics, behavior, and business impacts. Its deliverables are *combined patterns* such as the aforementioned combined association rules. Combined patterns consist of multiple components, a pair or cluster of atomic patterns, identified in individual sources or based on individual methods.

The general ideas of combined mining are as follows.

1) By involving multiple heterogeneous features, combined patterns are generated which reflect multiple aspects of concerns and characteristics in businesses.

2) By mining multiple data sources, combined patterns are generated which reflect multiple aspects of nature across the business lines.

3) By applying multiple methods in pattern mining, combined patterns are generated which disclose a deep and comprehensive essence of data by taking advantage of different methods.

4) By applying multiple interestingness metrics in pattern mining, patterns are generated which reflect concerns and significance from multiple perspectives.

The main contributions of our paper are as follows:

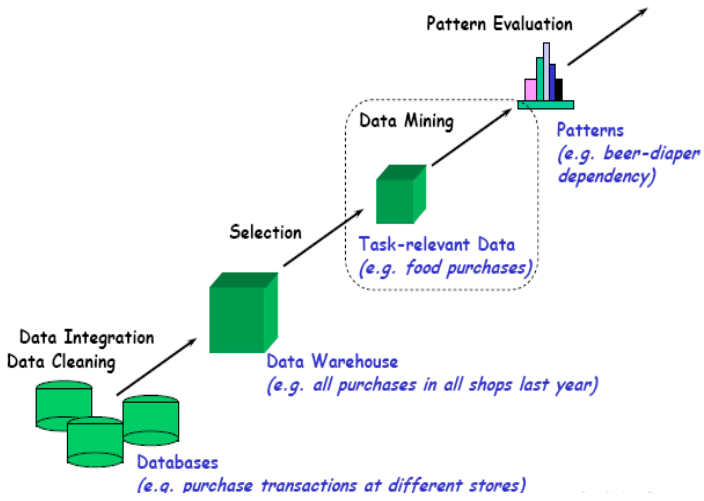
1) Building on existing works, generalizing the concept of combined mining that can be expanded and instantiated into many specific approaches and models for mining complex data toward more informative knowledge.

2) Discussing general frameworks and their paradigms and basic processes of *multi feature* and *multi method combined mining* for supporting combined mining, which contribute to *multisource combined mining*—they are flexible to be instantiated into specific needs.

3) Proposing various strategies for conducting pattern interaction when instantiating the aforementioned proposed frameworks—as a result, novel combined pattern types, such as incremental cluster patterns, can result from combined mining, which have not been investigated before.

4) Illustrating the corresponding interestingness metrics for evaluating certain types of combined patterns.

5) Demonstrating the use of combined mining in discovering combined patterns in real-world government service data for government debt prevention in an Australian Commonwealth Government agency.



Architecture Implementation

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods. In this phase, various modeling techniques are selected and applied, and their parameters are calibrated to optimal values. Typically, there are several techniques for the same data mining problem type. Some techniques have specific requirements on the form of data. Therefore, stepping back to the data preparation phase is often needed.

The wide-spread use of distributed information systems leads to the construction of large data collections in business, science and on the Web. These data collections contain a wealth of information, that however needs to be discovered. Businesses can learn from their transaction data more about the behavior of their customers and therefore can improve their business by exploiting this knowledge. Science can obtain from observational data (e.g. satellite data) new insights on research questions. Web usage information can be analyzed and exploited to optimize information access. Data mining provides methods that allow to extract from large data collections unknown relationships among the data items that are useful for decision making. Thus data mining generates novel, unsuspected interpretations of data.

Introduction to Itemsets

The existence of large amounts of scan code data collected by many businesses represents a potential wealth of information given adequate methods of transforming the data into meaningful information. One class of such data is stored in transaction databases from which all items obtained in a single transaction can be retrieved as a unit. The transactions can then be examined to determine what items typically appear together, e.g., which items customers typically buy together in a database of supermarket transactions. This in turn gives insight into questions such as how to market these products more effectively, how to group them in store layout or product packages, or which items to offer on sale to boost the sale of other items.

Recent research has focused on determining which groups of items, called itemsets, are frequently appear together in transactions. From any itemset an association rule may be derived which, given the occurrence of a subset of the items in the itemset, predicts the probability of the occurrence of the remaining items.Several algorithms have also been proposed for finding generalized itemsets from items that are classified by one or more taxonomic hierarchies Itemsets that meet a minimum support threshold are referred to as frequent itemsets. The

rationale behind the use of support is that a retail organization is only interested in those itemsets that occur frequently. However, the support of an itemset tells only the number of transactions in which the itemset was purchased. The exact number of items purchased is not analyzed and the precise impact of the purchase of an itemset cannot be measured in terms of stock, cost or profit. This shortcoming of the support measure prompted the development of a measure called itemset share, the fraction of some numerical value, such as total quantity of items sold or total profit, that is contributed by the items when they occur in an itemset

General Definitions

- **Itemset:** Set of items that occur together
- **Association Rule:** Probability that particular items are purchased together.
  - $X \Rightarrow Y$ where $X \cap Y = 0$
- **Support**, supp(X) of an itemset X is the ratio of transactions in which an itemset appears to the total number of transactions.
- **Share** of an itemset is the ratio of the count of items purchased together to the total count of items purchased in all transactions.
- **Confidence** of rule $X \Rightarrow Y$, denoted conf($X \Rightarrow Y$)
  - conf($X \Rightarrow Y$) = supp($X \cup Y$) / supp(X)
  - Confidence can also be defined in terms of the conditional probability.

conf($X \Rightarrow Y$) = P($Y \mid X$) = P($X \cup Y$) / P(X)

- **Transaction Database** stores transaction data. Transaction data may also be stored in some other form than a *m* x *n* database.

Customer	Item purchased	Item purchased
1	pizza	beer
2	salad	soda
3	pizza	soda
4	salad	tea

Fig: Data set Table

If A is “purchased pizza” and B is “purchased soda” then

Support = P(A and B) = ¼

Confidence = P(B / A) = ½

Applications of relational data mining

The use of RDM has enabled applications in areas rich with structured data and domain knowledge, which would be difficult to address with single table approaches. RDM has been

used in different areas, ranging from analysis of business data, through environmental and traffic engineering to web mining, but has been especially successful in bioinformatics (including drug design and functional genomics).Bioinformatics applications of RDM are discussed in the article by Page and Craven in this issue.

$daughter(X,Y) \leftarrow female(X), mother(Y,X).$
$daughter(X,Y) \leftarrow female(X), father(Y,X).$

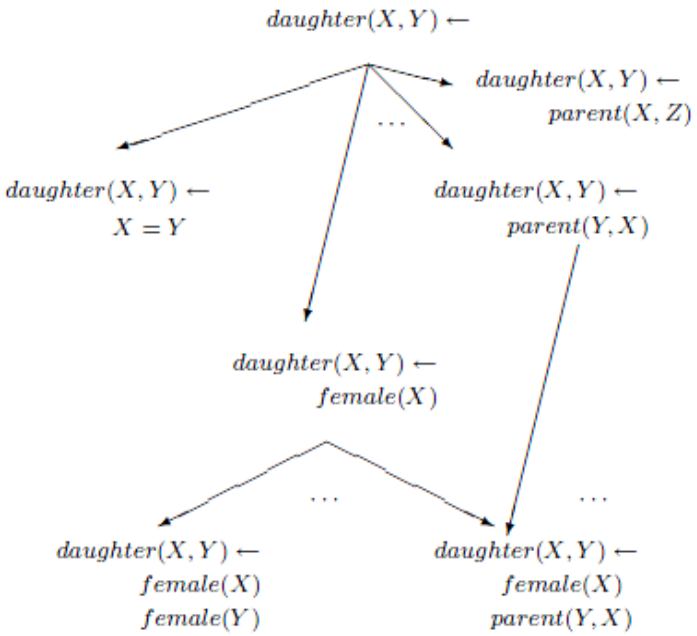


Figure 1: Part of the refinement graph for the family relations problem.

Relational Association Rules

The discovery of frequent patterns and association rules is one of the most commonly studied tasks in data mining. Here we first describe frequent relational patterns (frequent Datalog Patterns) and relational association rules (query extensions). We then look into how a well-known algorithm for finding frequent itemsets has been upgraded do discover frequent relational patterns.

Association Rules

The general aim of data mining is to find patterns in data that can become actionable information. In the case of association discovery, the involvement of domain experts is critical to the process of identification of relevant rules.

One of the reasons behind maintaining any database is to enable the user to find interesting patterns and trends in the data. For example, in a supermarket, the user can figure out which items are being sold most frequently. But this is not the only type of `trend' which one can possibly think of. The goal of database mining is to automate this process of finding interesting patterns and trends. Once this information is available, we can perhaps get rid of the original database. The output of the data-mining process should be a "summary" of the database. This goal is difficult to achieve due to the vagueness associated with the term `interesting'. The solution is to define various types of trends and to look for only those trends in the database. One such type constitutes the association rule.

1. Association rule mining

Association rule mining, a widely used data mining technique, is used to reveal the nature and frequency of relationships or associations between entities. Support and confidence, are the two major indices, which have useful applications to evaluate the rules. For instance, consider rule *X*: if *E* then *F*. Suppose that *X* has 60% confidence and 40% support. It expresses that 40% of records contain *E* and *F*. In fact, this means that in 40% of total records, rule *X* is valid. Additionally, it expresses that 60% of records that contain *E*, contain *F* as well. However, conventional association rules, as discussed above, can only provide limited knowledge for potential actions.

For example, in a supermarket, the user can figure out which items are being sold most frequently. But this is not the only type of `trend' which one can possibly think of. The goal of database mining is to automate this process of finding interesting patterns and trends. Once this information is available, we can perhaps get rid of the original database. The output of the data-mining process should be a "summary" of the database. This goal is difficult to achieve due to the vagueness associated with the term `interesting'. The solution is to define various types of trends and to look for only those trends in the database. One such type constitutes the association rule.

2. Combined association rule mining

Strictly speaking, traditional association rule mining can only generate simple rules. However, the simple rules are often not useful, understandable and interesting from a business perspective. Thus, Zhao et al. and Zhang et al, proposed combined association rules mining, which generated through further extraction of the learned rules. In other words, to present

associations in an effective way, and in order to discover actionable knowledge from resultant association rules, a novel idea of combined patterns is proposed.

Support and Confidence, are two major objective indices, which have useful applications to evaluate the association rules. Since combined association rule mining is further extracted from the simple learned rules. Support and confidence are two major metrics of combined association rules. Subjective metrics such as domain knowledge and expert knowledge are not involved in combined association rules obviously.

The proposed approach in this study consists of three steps. The first step is to employ expert knowledge and domain knowledge to ascertain the composite items generated from itemset. Then use association rules algorithms to find the rules with composite items. Finally, use combined association rule integrating the rules generated. It also discusses how to mine actionable combined patterns with composite items. Firstly, association rule, combined association rule and association rules with composite items are presented respectively. Then in order to implement actionable pattern mining in real world applications, metrics such as objective and subjective measures are discussed. Finally, this paper proposes a novel approach of mining actionable association rules with composite items integrating combined association rules and association rules with composite items.

The Algorithm

PD (transaction-set *T*)
1: *D*1 = {<*t*, 1>| *t* ? *T* }; *k*=1;
2: while (*D*k? *F*) do begin
3: forall *p* ? *D*k do *// counting*
4: forall *k*-itemset *s* ? *p.IS* do
5: Sup(s|*D*k) += *p.Occ*;
6: decide *Lk* and *~Lk* ;
//build Dk+1
7: *D*k+1= PD-rebuild(*D*k, *Lk*, *~Lk*);
8: *k*++;
9: end

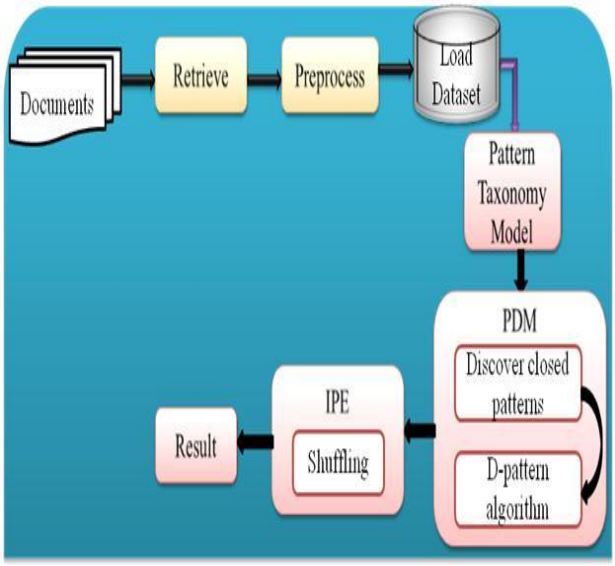
Our Implementation for parallel computing

The Apriori algorithm has been revised in several ways. One revision of the Apriori algorithm is to partition a transaction database into disjoint partitions TDB1, TDB2, TDB3, …, TDBn. Partitioning a transaction database may improve the performance of frequent itemsets mining by fitting each partition into limited main memory for quick access and allowing

incremental generation of frequent itemsets. Our implementation is a partition based Apriori algorithm that partitions a transaction database into N partitions and then distributes the N partitions to N nodes where each node computes its local candidate k-itemsets from its partition. As each node finishes its local candidate k-itemsets, it sends its local candidate k-itemsets to node 0. Node 0 then computes the sum of all candidate k-itemsets and prunes the candidate kitemsets to the frequent k-itemsets.

Frequent itemsets mining is one of the most important areas of data mining. Existing implementations of the Apriori based algorithms focus on the way candidate itemsets generated, the optimization of data structures for storing itemsets, and the implementation details. Bodon presented an implementation that solved frequent itemsets mining problem in most cases faster than other well-known implementations

Flow Diagram



Conclusion and Future Enhancement

Typical enterprise applications, such as telecom fraud detection and cross-market surveillance in stock markets, often involve multiple distributed and heterogeneous features as well as data sources with large quantities and expect to cater for user demographics, preferences, behavior, business appearance, service usage, and business impact. There is an increasing need to mine for patterns consisting of multiple aspects of the aforementioned information so as to reflect comprehensive business scenarios and present patterns that can inform decision-making

actions. This challenges existing data mining methods such as post analysis and table joining based analysis. Building on existing works, this paper has presented a comprehensive and general approach named *combined mining* for discovering informative knowledge in complex data. We focus on discussing the frameworks for handling multifeature-, multisource-, and multimethod-related issues. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents. In this research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The proposed technique uses two processes, pattern deploying and pattern evolving, to refine the discovered patterns in text documents

References:-

[1] Effective Pattern Discovery for Text Mining Ning Zhong, Yuefeng Li, and Sheng-Tang Wu, IEEE TRANSACTIONS ON KNOWLEDGE AND DATAENGINEERING, VOL. 24, NO. 1, JANUARY 2012

[2] L. Cao, Y. Zhao, and C. Zhang, "Mining impact-targeted activity patterns in imbalanced data," IEEE Trans. Knowl. Data Eng., vol. 20, no. 8, pp. 1053–1066, Aug. 2008.

[3] L. Cao, Y. Zhao, H. Zhang, D. Luo, and C.Zhang, "Flexible frameworks for actionable knowledge discovery," IEEE Trans. Knowl. Data Eng., vol. 22, no. 9, pp. 1299–1312, Sep. 2010.

[4] H. Cheng, X. Yan, J. Han, and C.-W. Hsu, "Discriminative frequent pattern analysis for effective classification," in Proc. ICDE, 2007,pp. 716–725.

[5] S. Dzeroski, "Multirelational data mining: An introduction," ACM SIGKDD Explor. Newslett., vol. 5, no. 1, pp. 1–16, Jul. 2003.

[6] G. Dong and J. Li, "Efficient mining of emerging patterns: Discovering trends and differences," in Proc. KDD, 1999, pp. 43–52.

[7] K. K. R. Hewawasam, K. Premaratne, and M.-L. Shyu, "Rule mining and classification in a situation assessment application: A belief-theoretic approach for handling data imperfections," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 37, no. 6, pp. 1446–1459, Dec. 2007.

[8] A. Jorge, "Hierarchical clustering for thematic browsing and summarization of large sets of association rules," in Proc. SDM, 2004, pp. 178–187.

[9] B. Lent, A. N. Swami, and J. Widom, "Clustering association rules," in Proc. ICDE, 1997, pp. 220–231.

[10] B. Liu, W. Hsu, and Y. Ma, "Pruning and summarizing the discovered associations," in Proc. KDD, 1999, pp. 125–134.

[11] Y. Zhao, C. Zhang, and L. Cao, Eds., Post-Mining of Association Rules: Techniques for Effective Knowledge Extraction. Hershey, PA: Inf. Sci.Ref., 2009.

AUTHORS DESCRIPTION

V.Shankar Ganesh, received B.tech (Computer Science and Engineering) from ANNA University, Chennai. Currently he is doing M.Tech Academic Project from Kuppam Engineering College, Andhra Pradesh, India. His Research interest areas are Data warehousing and Networks.

G.Surendra Reddy, currently he is working as Assistant Professor in Kuppam Engineering College, kuppam, received M.Sc (Computer Science) from Dravidian university and M.E (Computer Science and Engineering) from Satyabama university,Chennai. His Research interest areas are Data warehousing and Mining & Cloud Computing, Software Engineering.

N.S.Jagadeesh, currently he is working as Assistant Professor in Kuppam Engineering College, kuppam, received B.Tech (Information Technology) and M.Tech (Computer Science and Engineering) from JNTU,Anantapur. His Research interest areas are Data warehousing and Mining & Cloud Computing, Software Engineering.