Group Pattern Discovery Systems for Multiple Data Sources

Shichao Zhang
University of Technology Sydney, Australia

Chengqi Zhang
University of Technology Sydney, Australia

INTRODUCTION

Multiple data source mining is the process of identifying potentially useful patterns from different data sources, or datasets (Zhang et al., 2003). Group pattern discovery systems for mining different data sources are based on local pattern-analysis strategy, mainly including logical systems for information enhancing, a pattern discovery system, and a post-pattern-analysis system.

BACKGROUND

Many large organizations have multiple data sources, such as different branches of a multinational company. Also, as the Web has emerged as a large distributed data repository, it is easy nowadays to access a multitude of data sources. Therefore, individuals and organizations have taken into account the Internet's low-cost information and knowledge when making decisions (Lesser et al., 2000). Although the data collected from the Internet (called external data) brings us opportunities in improving the quality of decisions, it generates a significant challenge: efficiently identifying quality knowledge from different data sources (Kargupta et al., 2000; Liu et al., 2001; Prodromidis et al., 1998; Zhong et al., 1999). Potentially, almost every company must confront the multiple data source (MDS) problem (Hurson et al., 1994). This problem is difficult to solve, due to the facts that multiple data source mining is a procedure of searching for useful patterns in multidimensional spaces; and putting all data together from different sources might amass a huge database for centralized processing and cause problems, such as data privacy breaches, data inconsistency, and data conflict.

Recently, the authors have developed *local pattern analysis*, a new multi-database mining strategy for discovering some kinds of potentially useful patterns that cannot be mined in traditional multi-database mining techniques (Zhang et al., 2003). Local pattern analysis delivers high-performance pattern discovery from MDSs. This effort provides a good insight into knowledge discovery from multiple data sources.

However, there are two fundamental problems that prevent local pattern analysis from widespread application. First, when the data collected from the Internet is of poor quality, the poor-quality data can disguise useful patterns. For example, a stock investor might need to collect information from outside data sources when making an investment decision, such as news. If fraudulent information collected is applied directly to investment decisions, the investor might lose money. In particular, much work has been built on consistent data. In the input to the distributed data mining algorithms, it is assumed that the data sources do not conflict with each other. However, reality is much more inconsistent than the ideal; the inconsistency must be resolved before any mining algorithms can be applied. These generate a crucial requirement: ontology-based data enhancement.

The second fundamental challenge is the efficiency of mining algorithms for identifying potentially useful patterns in MDSs. Over the years, there has been a great deal of work in multiple source data mining (Aounallah et al., 2004; Krishnaswamy et al., 2000; Li et al., 2001; Yin et al., 2004). However, traditional multiple data source mining still utilizes mono-database mining techniques. That is, all the data from relevant data sources is pooled to amass a huge dataset for discovery. These algorithms cannot discover some useful patterns; for example, the pattern that 70% of the branches within a company agrees that a married customer usually has at least two cars if his or her age is between 45 and 65. On the other hand, using our local pattern analysis, there can be huge amounts of the local patterns. These generate a strong requirement: the development of efficient algorithms for identifying useful patterns in MDSs.

This article introduces a group of pattern discovery systems for dealing with the MDS problem, mainly (1) a logical system for enhancing data quality, a logical system for resolving conflicts, a data cleaning system, and a database clustering system, for solving the first problem;

and (2) a pattern discovery system and a post-mining system for solving the second problem.

MAIN THRUST

Group pattern discovery systems are able to (i) effectively enhance data quality for mining MDSs and (ii) automatically identify potentially useful patterns from the multi-dimension data in MDSs.

Data Enhancement

Data enhancement includes the following:

1. The data cleaning system mainly includes these functions: recovering incomplete data (filling the values missed or expelling ambiguity); purifying data (consistency of data name, consistency of data format, correcting errors, or removing outliers); and resolving data conflicts (using domain knowledge or expert decision to settle discrepancy).
2. The logical system for enhancing data quality focuses on the following epistemic properties: veridicality, introspection, and consistency.
3. The logical system for resolving conflict has the property of obeying the weighted majority principle in case of conflicts.
4. The fuzzy database clustering system generates good database clusters.

Identifying Interesting Patterns

A local pattern may be a frequent itemset, an association rule, causal rule, dependency, or some other expression. Local pattern analysis is an in-place strategy specifically designed for mining MDSs, providing a feasible way to generate globally interesting models from data in multi-dimensional spaces.

Based on our local pattern analysis, three key systems can be developed for automatically searching for potentially useful patterns from local patterns: (a) identifying high-vote patterns; (b) finding exceptional patterns; and (c) synthesizing patterns by weighting majority.

(a) **Identifying High-Vote Patterns:** Within an MDS environment, each data source, large or small, can have an equal power to vote for their patterns for the decision-making of a company. Some patterns can receive votes from most of the data sources. These patterns are referred to as *high-vote patterns*. High-vote patterns represent the commonness of the branches. Therefore, these patterns may be far more important in terms of decision-making within the company. The key problem is how to efficiently search for high-vote patterns of interest in multi-dimensional spaces. It can be attacked by mining the distribution of all patterns.

(b) **Finding Exceptional Patterns:** Like high-vote patterns, exceptional patterns also are regarded as novel patterns in multiple data sources, which reflect the individuality of data sources. While high-vote patterns are useful when a company is reaching common decisions, headquarters also are interested in viewing exceptional patterns used when special decisions are made at only a few of the branches, perhaps for predicting the sales of a new product. Exceptional patterns can capture the individuality of branches. Therefore, although an exceptional pattern receives votes from only a few branches, it is extremely valuable information in MDSs. The key problem is how to construct efficient methods for measuring the interestingness of exceptional patterns.

(c) **Searching for Synthesizing Patterns by Weighting Majority:** Although each data source can have an equal power to vote for patterns for making decisions, data sources may be different in importance to a company. For example, in a company, if the sale of branch A is four times that of branch B, branch A is certainly more important than branch B in the company. (Here, each branch in a company is viewed as a data source in an MDS environment.) The decisions of the company are reasonably partial to high-sale branches. Also, local patterns may have different supports. For example, let the supports of patterns X_1 and X_2 be 0.9 and 0.4 in a branch. Pattern X_1 is far more believable than pattern X_2. These two examples present the importance of branches and patterns for decision making within a company. Therefore, synthesizing patterns is very useful.

Post Pattern Analysis

In an MDS environment, a pattern (e.g., a high-vote association rule) is attached to certain factors, including name, vote, vsupp, and vconf. For a very large set of data sources, a high-vote association rule may be supported by a number of data sources. So, the sets of its support and confidence in these data sources are too large to be browsed by users, and thus, it is rather difficult to apply the rule to decision making for users. Therefore, post-pattern analysis is very important in MDS mining. The key problem is how to construct effective partition for classifying the patterns mined.