Scalable Mining and Link Analysis Across Multiple Database Relations

Xiaoxin Yin
University of Illinois at Urbana-Champaign
xyin@microsoft.com

In the last decade there have been great advances in data mining research, and many data mining methods have been applied to everyday business, such as market basket analysis, direct marketing, and fraud detection. Data mining is also providing many useful methods for other disciplines such as insurance, civil engineering, and bioinformatics.

Most existing data mining algorithms (including those for classification, clustering, association analysis, and outlier detection) work on single tables or single files. For example, a typical classification algorithm works on a table containing many tuples, each having a class label, and a value on each of the attributes. Unfortunately, most information in the world can hardly be represented by such independent tables. In a real-life data set there are usually many types of objects, which are linked together through different types of linkages. Such data is usually stored in relational databases, XML files, or other structured or semi-structured data repositories.

My thesis studies data mining in relational or other structured or semi-structured databases. A database contains multiple inter-connected relations, each of which represents a certain kind of objects or a type of relationships. Most existing data mining algorithms cannot be applied to relational data, unless the relational data is first converted into a single table. However, much valuable semantic information, especially the linkages between objects, could be lost by conversion of multi-relational data into a single table. On the other hand, a multi-relational database can often provide much richer information for data mining due to the richness of semantic data modeling in relational database design. Thus multi-relational data mining approaches can often be more effective than single-table methods. Imagine there is a database of a computer science department, which stores information about professors, students, courses, research groups, and publications. In such a database, each student is associated with various types of information, such as their courses, advisors, research groups, and publications. Moreover, the objects linked with students are also linked with each other. This rich information source provides countless opportunities of data mining. For example, we can classify students according to their academic performances, cluster students based on their research, find patterns/correlations of course registrations and publications, or detect duplicate entries among authors of publications. Because of the high popularity of relational databases, multi-relational data mining can be used in many disciplines, such as financial decision making, direct marketing, and customer relationship management.

Although very useful, multi-relational data mining faces two major challenges. First, it is more difficult to model multi-relational data. Unlike tuples in a single table which can be modelled by vectors, multi-relational data contains heterogeneous objects and relationships among them, and there has not been widely accepted model for mining such data.

Second, many data mining approaches aim at finding a model (or hypothesis) that fits the data. In a relational database, the number of possible models is much larger than that in a single table. For example, in rule-based multi-relational classification, each rule is associated with a join path. If each relation is joinable with two other relations on average, there are $2^{k+1} - 1$ join paths of length no greater than k. Thus it is more complicated or at least more time consuming to search for good models in relational databases.

The main purpose of my thesis is to study the application of data mining technology in multi-relational environments. Because most real-world relational databases have complicated schemas and contain huge amounts of data, efficiency and scalability become our major concerns as well as the accuracy and effectiveness of the algorithms. In my thesis step-by-step developments are made for multi-relational data mining. Figure 1 shows a road map of our work. There are mainly two types of multi-relational information that are widely used for data mining: Neighbor objects (objects linked to each object through certain join paths) and linkages between objects. These two types of information are complementary for each other, as neighbor objects represent the contexts of objects, and linkages indicate relationships between objects. We propose new methods for efficiently acquiring both types of information, which are *Tuple ID Propagation* [1] for finding neighbor objects in each relation, and *Path decomposition* [4] for finding all linkages between two objects efficiently.

Based on Tuple ID Propagation, we propose CrossMine [1], a highly efficient and scalable approach for multi-relational classification. Then we move to multi-relational clustering, which is a more sophisticated problem as no pre-defined classes are given. We propose CrossClus [2], which can utilize user guidance and perform efficient clustering. Based on linkage information, we propose Relom, a new approach for duplicate detection using linkages between objects. We invent LinkClus [3], a new approach for linkage-based similarity analysis and clustering. Because neighbor objects and linkages are two complementary types of information, it is very helpful to combine both of them in data mining tasks.

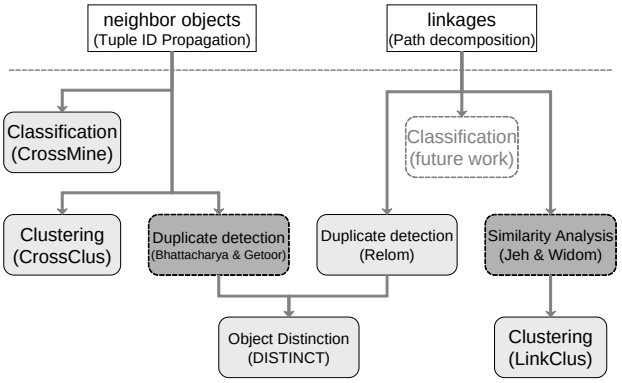

Figure 1: A road map of my thesis

We also propose DISTINCT [4], the first approach for distinguishing objects with identical names, which utilizes both types of information.

The following contributions are made in my thesis.

- **Tuple ID Propagation.** Many multi-relational data mining approaches suffer from efficiency and scalability problems because they repeatedly join different relations. We propose *tuple ID propagation*, a method for virtually joining different relations and avoiding repeated computation. It propagates the IDs of tuples in a certain relation R to other relations, so that one can easily find tuples joinable with each tuple in R. This method gives us great help in making our approaches efficient and scalable.
- **Efficient Multi-relational Classification.** With the help of tuple ID propagation and adoption of divide-and-conquer strategy, we propose CrossMine [1], a scalable and accurate approach for multi-relational classification. It uses two algorithms for classification, one being rule-based and the other decision-tree based. The experimental results show that CrossMine is tens or hundreds of times faster than existing approaches, and achieves higher accuracy.
- **Multi-relational Clustering with User's Guidance.** A relational database usually contains information of many aspects, and in most cases only a small part of information is relevant to a clustering task. Thus in multi-relational clustering it is crucial to let the user indicate her clustering goal, while it is very unlikely that the user can specify all pertinent attributes. We allow the user to provide a simple guidance, which is one or a few pertinent attributes in the relational database. We design a methodology called CrossClus [2], which selects pertinent attributes across different relations by observing whether these attributes group objects in similar ways as the user guidance.
- **Multi-relational Duplicate Detection.** The goal of multi-relational duplicate detection is to find different references to the same object in a database. Based on our linkage analysis, we find that references to the same object tend to be intensively connected to each other. We design similarity measures that reflect the strength of such connections. We also design *path decomposition*, a method that can find all linkages between two objects.
- **Object Distinction in Relational Databases.** In a databases there are often different entities with identical names (e.g., authors of papers). Distinguishing entities with identical names is a challenging problem, as there is very limited information for each tuple, and many tuples need to be grouped into clusters, so that each cluster corresponds to a real entity. We use a hybrid similarity measure [4], which combines both the context of references and linkages between them to measure their similarities.
- **Link-based Similarity Analysis.** In many applications the similarity between objects can only be inferred from their linked objects, and two objects are similar if and only if they are linked to similar objects. Inferring such similarities directly can be very expensive, as it requires storing pair-wise similarities between objects. We design hierarchical data structures that store significant similarities and compress insignificant ones, which greatly improve the efficiency and scalability of link-based similarity analysis [3].

Finally, we extended our multi-relational data mining framework to truth validation (also known as *veracity analysis*) on the web. Since multiple information providers on the web may provide *conflicting* information about the *same* entity, it is necessary to provide trustable analysis of the truthfulness of information from multiple information providers and automatically identify the correct information. We designed a general framework, called TruthFinder [5] for resolving conflictive information provided by multiple sources, based on the heuristic that *an information provider is* trustworthy *if it provides many pieces of true information, and a piece of information is likely to be true if it is provided by many trustworthy web sites*. The method links information providers, stated facts on different entities, and the corresponding entities, into a heterogeneous information network, and consolidates the trustworthiness by an iterative enhancement process with weight-propagation and consolidation across this network. In one of our experiments TruthFinder successfully finds the true set of authors and the trustable information providers based on the conflicting book author information provided by multiple sources.

Ph.D. dissertation committee

Prof. Jiawei Han, Prof. Marianne Winslett, Prof. Kevin Chen-Chuan Chang, and Prof. Philip S. Yu.

REFERENCES

[1] X. Yin, J. Han, J. Yang, and P. S. Yu. CrossMine: Efficient Classification Across Multiple Database Relations. In *ICDE'04*.

[2] X. Yin, J. Han, and P. S. Yu. Cross-Relational Clustering with User's Guidance. In *KDD'05*.

[3] X. Yin, J. Han, and P. S. Yu. LinkClus: Efficient Clustering via Heterogeneous Semantic Links. In *VLDB'06*.

[4] X. Yin, J. Han, and P. S. Yu. Object Distinction: Distinguishing Objects with Identical Names by Link Analysis. In *ICDE'07*.

[5] X. Yin, J. Han, and P. S. Yu. Truth Discovery with Multiple Conflicting Information Providers on the Web. *IEEE TKDE*, 20:796–808, 2008.