

Heterogeneous Association Rules Mining

Badr. Al-Daihani

School of Computer Science, Cardiff University, Queen's Buildings, Newport Road, PO Box 916, Cardiff CF24 3XF,UK
badr@cs.cf.ac.uk

Abstract. Bioinformatics databases are highly heterogeneous, not only do they differ in their representation but they also offer radically different query capabilities across the diverse information held in distributed resources. The need to extract and link knowledge from these very large databases is increasing. So there is a critical need for innovative information management and knowledge discovery tools and techniques to sift through this heterogeneous data to recover appropriate data and analyse it. Data mining is a particular step of the Database Discovery Knowledge (DDK) process. It is the iterative and interactive process of discovering valid, novel, useful and understandable patterns or data sets in these databases. We will investigate the structure of a system that allows users to access such databases to extract knowledge and patterns of interest, while not needing to be aware of the representation details in the individual resources.

KEYWORDS: bioinformatics, data mining, association rules, heterogeneous.

INTRODUCTION

The ability to interact with heterogeneous data source is critical for the biologist. They nowadays spend a significant amount of time and effort in accessing/querying multiple remote or local heterogeneous data sources and integrating the results of these searches either manually, or with the aid of data integration tools. They face a number of problems. They are required to have good knowledge about the content, structure, and representation of the data in such databases, and their implementation, and query capabilities. Also when users have a complex query or information from different sources needs to be combined they need to construct a query plan on their own. A mistake in this plan may lead to poor performance or even not obtaining a result. Biological data is available in different formats [1,2,3,4,5]: Flat files such as GenBank, EMBL, DDBJ, PDB; Relational databases such as HGMD, MGMD ; Object-oriented database such as AceDB and XML databases like PIR, SwissProt, and InterPro.

The major characteristics of bioinformatics databases are: the Diversity/variety of data, the representational heterogeneity, autonomous and web-based sources and varied interface and query capabilities [1,2,3,4].

The existing systems for integrating bioinformatics databases differ from each other in the integration approach they use. Currently, there are three models of data integration: federation, warehousing and mediations [1,2,3,4,5].

In a federation, the database systems are independent and each database component communicates to another directly and the data can continue to be stored in its original locations and retrieved via a middleware component. Some examples of federation systems are K2/BioKleisli, Entrez [1,2,3,4,5].

A mediator stores no data on its own rather it provides a virtual view of the integrated sources. It interacts with data sources via wrappers and handles a user query by splitting it into sub-queries, sending the sub-queries to appropriate wrappers and integrating the results locally to provide answers to queries. Some examples of mediator systems are Discovery-Link, Transparent Access to Multiple Bioinformatics Information Source (TAMBIS), and Knowledge-based Integration of Neuroscience Data (KIND) [1, 2, 3, 4, 5].

Finally, Data warehouses often use wrappers to import data from remote sources that is then materialized locally as a global schema, and queries are evaluated against the warehoused data. While this approach simplifies the access and analysis of the data that is stored in the heterogeneous data repositories, the challenge is to keep the data in the warehouse current as changes are made to the sources, in particular when the warehouse is large. Some examples of warehousing systems are Genomics Unified Schema (GUS), Sequence Retrieval System (SRS) [1,2,3,4,5].

With massive volumes of data collected and stored in distributed and heterogeneous databases, an urgent need exists to develop tools that extract knowledge from large data sources. This information is invaluable sources for analysis and decision support. Distribution of attribute values within/across data sources or association between different structured and unstructured data components can be used to evaluate trends and discover pattern.

The focus of thesis is twofold:

First, we will investigate the structure of a system that allows users to access such databases to extract knowledge and patterns of interest, while not needing to be aware of the representation details in the individual resources.

Second, we define the problem of mining association rules over heterogeneous databases. We refer to this mining problem as the Heterogeneous Association Rules problem (HAR). Then we will propose a novel approach to efficient mining of association rules over heterogeneous databases.

Association Rules:

Data Mining also known as Knowledge Discovery in Database – KDD) has been recognized as a new research area for database research. It has been defined as “the nontrivial extraction of implicit previously unknown and potentially useful information from data”. The goal of data mining is to automate the process of finding interesting patterns and trends from a given data. Association Rule Mining is a widely used technique for large-scale data mining. Discovery of association rules is an important data mining task.

The general problem of discovering association rules was introduced in [6]. Since then, there has been considerable work on designing algorithms for mining such rules.

Given a set of transactions, where each transaction is a set of items, an association rule is an expression of the form $X \Rightarrow Y$, where X and Y are subsets of items. These rules indicate that the transactions that contain X tend to also contain Y . They identify groupings between sets of items with minimum specified confidence where confidence is defined as the percentages of transactions containing X and Y to number of transactions containing X . Only sets of items satisfying a minimum support condition are potential rule candidates where support is defined as the percentage of transactions containing X and Y to the total number of transactions. For instance, an association rule is “30% of the transaction that contain bread also contain milk; 2% of all transactions contain both items.” In this case, 30% is the confidence(c) of the rule and 2% is its support (s).

Gene mutation is one of the promising application area. Extracting interesting patterns and rules from gene mutation datasets, can be important in identifying cause of gene tumours and diseases. The discovery of interesting association relationships among huge amount of gene mutation can help in determining the cause of mutation in tumours and diseases. We will try to discover the association rules between gene mutation and diseases as well the association between gene mutation and mutagen.

MATERIAL AND METHODS

In this study, we choose the following two data sources, which are related to each other using Gene name as key:

HGMD database

The Human Gene Mutation Database (HGMD) represents an attempt to collate known (published) gene lesions responsible for human inherited disease. Thus, HGMD provides information about practical diagnoses. All HGMD entries comprise a reference to the first literature report of a mutation, the associated disease state as specified in this report, the gene name, symbol and chromosomal location. [8]

MGMD database

The Mammalian Gene Mutation Database (MGMD) is a project to collate the profiles of known (published) mutagen-induced gene mutations following analysis of the literature on mutagen-induced mutational spectra in mammalian tissues. The database stores the mutation spectra information ie. Mutagen, Species, Tissue, Cell line, Gene, Mutation Class, Mutation or first name author from the reference, or the Medline abstract number of the studies. It is a single relational table which has 39134 records [7].

From above databases, sets of items whose elements tend to be in both databases will be retrieved to discover the interesting association rules among genes, mutations, mutagens and diseases.

REFERENCES

- [1] Hernandez T. and Kambhampati S. (2004) *Integration of Biological Sources: Current Systems and Challenges Ahead*, Proc. of the ACM SIGMOD Conference.
- [2] C. Goble et al. (2001) *Transparent access to multiple bioinformatics information sources*. IBM Systems Journal, 40(2).
- [3] Barbara Eckman, Zoe Lacroix and Louiqa Raschid (2001) *Optimized Seamless Integration of Biomolecular Data*, IEEE, International Conference on Bioinformatics and Biomedical Engineering, 23-32.
- [4] Lacroix Z, Boucelma O and Essid M (2003) *The Biological Integration System*. Proc. of the 5th ACM Workshop on Web Information and data management, pp 45-49.
- [5] Aldana J., Roldán M, Navas I, Pérez A and Trelles O (2004) *Integrating Biological Data Sources and Data Analysis Tools through Mediators*, Proceedings of the 2004 ACM symposium on Applied computing.
- [6]. Agrawal, R.-Imielinski, T.-Swami, A. (1993) Mining Association Rules Between Sets of Items in Large Databases. Proc. ACM SIGMOD:207-216.
- [7] P.D. Lewis, J.S. Harvey, E.M. Waters, and J.M. Parry (2000) The Mammalian Gene Mutation Database, *Mutagenesis*, 15(5): 411- 414.
- [8] Krawczak M, Ball EV, Fenton I, Stenson PD, Abeyasinghe S, Thomas N, Cooper DN (2000): Human Gene Mutation Database - a biomedical information and research resource. *Human Mutation* 15(1):45-51.