A Comparative Study of Data Mining Systems with Relevance to Social Networks

Work-in-progress paper

Mari Terblanche, Charmain Cilliers
Department of Computer Science and Information Systems
Nelson Mandela Metropolitan University, PO Box 77000, Port Elizabeth, 6031
Tel: (041) 504 2323 Fax: (041) 504 2831
Email: {Mari.Terblanche, Charmain.Cilliers}@nmmu.ac.za

Abstract –The features identified for a comparative study of data mining systems include data mining functionalities, data structures used and the ability to interface with external analysis tools. The purpose of the comparative study is to provide a foundation for the design of an extensible data mining system. The proposed system will mine email log file data in order to perform social network analysis.

Index Terms – data mining features, data mining systems, external interfaces, social network analysis.

I. INTRODUCTION

DATA mining systems perform a wide range of complex tasks, but they typically lack extensibility. The lack of extensibility means that the systems' internal data structures and algorithms are designed in such a way that it is very difficult to add new algorithms or functionalities to the system without additional major changes being made to the implementation of the system itself. Furthermore, most existing systems each exhibit customized environments and tend to function independently from external analysis tools. The data mining models resulting from data mining processes are also often not accessible to external analysis tools. No uniform internal management techniques or strategies are applied in these existing data mining systems [1].

In order to address these problems, a comparative study was conducted in order to identify the features required for a comparison of data mining systems. This comparative study will form a basis for the development of an extensible data mining system for social network analysis. This paper discusses some of the results obtained through the comparative study with special reference to the problem domain of social networks.

II. EXISTING DATA MINING SYSTEMS

In order to effectively compare different data mining systems, several features for comparison were identified, namely data mining functionalities, data structures and interfacing with external analysis tools.

A. Data Mining Functionalities

Data mining functionalities specify the type of patterns that can be discovered as a result of data mining tasks. These functionalities include:

- *characterization* – creating a general summary of the class of data under study (target class) on the basis of similar characteristics;
- *comparison* – comparing the target class with a set of other classes to show distinctions;
- *association* – discovering association rules showing frequent occurrences of attribute-value conditions appearing in the target class; and
- *classification* – creating a set of models to describe and differentiate classes of data.

Some data mining systems implement all of the previously specified functionalities [2, 3] and others, while other systems implement only those relevant to the problem domain of the system [4].

Apart from these commonly used functionalities, a data mining system called DBMiner [3] also has a meta-pattern guided miner. The miner implements a data mining technique that allows the user to specify the form of mined association rules in order to limit the volume of the discovered rules presented.

B. Data Structures

There are usually two types of data structures that are considered when implementing a data mining system [3-5], namely generalized relations and multi-dimensional data cubes. Some data mining systems use a generalized relation [5], which is a set of generalized attributes containing some sort of aggregate measure, e.g. sum or count, stored for each generalized attribute. A multi-dimensional data cube, on the other hand, can be represented by a multi-dimensional array. Multi-dimensional data cubes are the more popular choice in many data mining systems [3, 4] because they cost less to produce, use less storage space and access a specified value very quickly.

C. Interfacing with External Analysis Tools

Very few data mining systems provide interfacing with external analysis tools [3, 4, 6]. However, one system that addresses the problem of interfacing with external analysis tools is the Linear Correlation Discovery System [5]. This system forms an interface between the data mining package and the external analysis tool (a statistical package). The LCD system has two main components, namely the selection assistant and the statistics coupler. The *selection assistant* examines a schema and instances to determine the right association measurement functions, e.g. chi-square or linear regression. The *statistics coupler* applies the most suitable statistical function on the sample data set and uses the results to create statistical output.

III. Social Networks

Internet and network technologies have greatly contributed to the formation of cyber communities, where individuals with common interests can communicate with each other in a variety of ways, namely via email, file sharing and instant messaging. Virtual teams, consisting of several individuals, can form within these cyber communities. These teams are formed to serve a specific purpose, for example information sharing, collaboration for completing a project and administration [7].

A social network is the result of a computer network connecting people or organizations. In other words, it is a collection of people that have certain social relationships with one another, for example friendship, co-working or information interchange [8].

IV. Social Network Analysis

The main focus of social network analysis is on finding patterns of interaction among people [9]. The process of social network analysis involves:

- discovering important patterns in communication networks;
- tracking the movement and flow of resources (mainly information) through the network; and
- determining the impact that the relations between individuals have on the network and on the organization.

Social network analysis is only possible upon collection of sufficient relevant social network data. Traditionally, the data about social networks were gathered using interviews, questionnaires, diaries kept by individuals and through observation [8]. Nowadays, the source of social network analysis is email log files. Most modern organizations rely heavily on computer networks, therefore social network data are analyzed today using data mining techniques.

The *To* and *From* fields in email correspondence are the primary attributes mined for social network analysis. The *Subject* field and the actual content of the email are ignored, since this information covers a very broad range of subjects, and would thus be nearly impossible to classify in a way that would be useful for discovering a social network.

Ethical reasons constrain social network analysts from accessing the content of private emails. Therefore, as long as permission is granted by all the individuals involved, a data mining system is an appropriate tool for mining email data for the purpose of social network analysis.

V. Conclusion and Future Work

The data mining functionalities that apply specifically to the mining of social network analysis data are characterization and comparison. Other functionalities might also be found to apply once social network analysis algorithms have been thoroughly investigated.

Multi-dimensional data cubes are being considered as the internal data structure for the design of the framework because of the advantages identified.

An added feature that can be used to compare data mining systems is level of extensibility. Extensibility is the characteristic of a system that allows easy addition of new components without making changes to the structure of the system itself.

Future work will consist of further research into a clearer definition of extensibility in data mining systems. This definition can then be used to propose an extensible framework for the data mining of social networks.

Acknowledgment

We would like to thank the Telkom Centre of Excellence Programme and the Department of Computer Science and Information Systems at the Nelson Mandela Metropolitan University for making this research possible.

References

[1] KIMANI, S., T. CATARCI, and G. SANTUCCI. *A Visual Data Mining Environment*. in *International Workshop on Visual Data Mining*. 2002. Helsinki, Finland.

[2] AHMED, K.M., N.M. EL-MAKKY, and Y. TAHA. *Effective data mining: a data warehouse-backboned architecture*. in *1998 Conference of the Centre for Advanced Studies on Collaborative research*. 1998.

[3] HAN, J., et al. *DBMiner: A System for Data Mining in Relational Databases and Data Warehouses*. in *Conference of the Centre for Advanced Studies on Collaborative research*. 1997.

[4] ZAÏANE, O.R., et al. *MultiMediaMiner: A System Prototype for MultiMedia Data Mining*. in *ACM SIGMOD international conference on management of data*. 1998.

[5] CHUA, C., E.-P. LIM, and R.H.L. CHIANG. *An Integrated Data Mining System to Automate Discovery of Measures of Association*. in *33rd Hawaii International Conference on System Sciences - Volume 2*. 2000. Maui, Hawaii.

[6] KNOETZE, R., J. WESSON, and C. CILLIERS. *The Warehousing of Application Services Data to Facilitate Identification of Patterns of Network Usage*. in *Southern African Telecommunications and Applications Conference (SATNAC) 2004*. 2004. Stellenbosch.

[7] LIN, F.-R. and C.-H. CHEN. *Developing and Evaluating the Social Network Analysis System for Virtual Teams in Cyber Communities*. in *37th Hawaii International Conference on System Sciences (HICSS'04)*. 2004. Big Island, Hawaii.

[8] GARTON, L., C. HAYTHORNTHWAITE, and B. WELLMAN, *Studying Online Social Networks.* Journal of Computer-Mediated Communication (JCMC), 1997. **3**(1).

[9] BERKOWITZ, S.D., *An introduction to structural analysis: The network approach to social research*. 1982, Toronto: Butterworth.

Mari Terblanche received her B.Sc Hons degree in 2004 from the Nelson Mandela Metropolitan University. She is currently doing her M.Sc in Computer Science at NMMU.