Integrating Entropy and Closed Frequent Pattern Mining for Social Network Modelling and Analysis

Muhammenul Adnan, Reda Alhajj, and Jon Rokne

Abstract The recent increase in the explicitly available social networks has attracted the attention of the research community to investigate how it would be possible to benefit from such a powerful model in producing effective solutions for problems in other domains where the social network is implicit; we argue that social networks do exist around us but the key issue is how to realize and analyze them. This chapter presents a novel approach for constructing a social network model by an integrated framework that first preparing the data to be analyzed and then applies entropy and frequent closed patterns mining for network construction. For a given problem, we first prepare the data by identifying items and transactions, which are the basic ingredients for frequent closed patterns mining. Items are main objects in the problem and a transaction is a set of items that could exist together at one time (e.g., items purchased in one visit to the supermarket). Transactions could be analyzed to discover frequent closed patterns using any of the well-known techniques. Frequent closed patterns have the advantage that they successfully grab the inherent information content of the dataset and is applicable to a broader set of domains. Entropies of the frequent closed patterns are used to keep the dimensionality of the feature vectors to a reasonable size; it is a kind of feature reduction process. Finally, we analyze the dynamic behavior of the constructed social network. Experiments were conducted on a synthetic dataset and on the Enron corpus email dataset. The results presented in the chapter show that social networks extracted from

---

Muhammenul Adnan
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

Reda Alhajj
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada

Department of Computer Science, Global University, Beirut, Lebanon
Department of Information Technology, Hellenic American University, Athens, Greece

Jon Rokne
Department of Computer Science, University of Calgary, Calgary, Alberta, Canada
a feature set as frequent closed patterns successfully carry the community structure information. Moreover, for the Enron email dataset, we present an analysis to dynamically indicate the deviations from each user's individual and community profile. These indications of deviations can be very useful to identify unusual events.

1 Introduction

With the advancement in information technology such as the development of storage media, the improvement of the computing power of the commodity CPUs, and the availability of internet and broadband telecommunication; business organizations are now storing more and more data in electronic format in an attempt to add value to the online services provided to the customers. But keeping data in the raw format provide no or little decision making power to the managing body of the organizations. Hence, several knowledge discovery and lately data mining techniques (such as classification, clustering, association rule mining, and time series prediction) have been developed to facilitate the decision making of the organizations. These data mining techniques help the organizations to extract the information or knowledge buried in the data which is the key to the success of today’s organizations. Most of the data mining techniques concentrate on finding information from data, or predicting future outcomes, but they provide little or no knowledge on how users of these data are connected to each other. We argue that social network is a powerful model for filling this gap. In other words, connecting people by considering their related information, connecting different information sources, and connecting people to information to which other people are connected are some of the important cases that may provide useful decision making power to the organizations, yet received little attention. In other words, an essential key factor to achieve success in a competitive environment is to avoid handling the information and people in isolation. In a social network settings information could be linked to other information and even to people. Actually, it is important to realize that people and the information are tightly coupled forming a social network which could be analyzed for interesting discoveries.

A social network represents relationships or ties among actors which depending on the problem domain could be individuals, pieces of information, genes, etc; actors are they key entities/items in the problem to be investigated. Formally, a social network can be described as a graph \( G = (V, E) \), where \( V = \{v_1, v_2, v_3, \ldots, v_n\} \) is the set of vertices representing individuals or actors and \( E \) is the set of edges representing relationships between vertices or individuals. In early days, the social networking concept was limited to analyze links between people in small communities like employees in an