From Frequent Features to Frequent Social Links

Erick Stattner, LAMIA Laboratory, University of the French West Indies and Guiana, Pointe-à-Pitre, Guadeloupe, France
Martine Collard, LAMIA Laboratory, University of the French West Indies and Guiana, Pointe-à-Pitre, Guadeloupe, France

ABSTRACT

Standard data mining techniques have been applied and adapted for eliciting knowledge from social networks, by achieving classical tasks such as classification, search for frequent patterns or link prediction. Most works have exploited only the network topological structure, and therefore cannot be used to answer questions involving nodes features. For instance, the frequent pattern discovery task generally refers to the search for sub-networks frequently found in a single network or in a set of networks. In the same area, this paper focuses on the concept of frequent link that stands as a regularity found in a network on links between node groups that share common characteristics. The extraction of such links from a social network is a particularly challenging and computationally intensive problem, since it is much dependent on the number of links and attributes. In this study, the authors propose a solution for reducing the search space of frequent links, by filtering the nodes features on a criterion of frequency. The authors make the assumption that frequent links occur between sets of features that are themselves frequent. This property is used to reduce the search space and speed up the extraction process. The authors empirically show that it is well founded, and they discuss the efficiency of the solution in terms of computation time and number of frequent patterns found depending on several frequency thresholds.

Keywords: Filtering, Frequent Links, Frequent Patterns, Link Mining, Social Network

1. INTRODUCTION

Standard data mining algorithms have been relying on the implicit assumption that data are IID (independent and identically distributed). They consider datasets as collections of independent instances of single relations. While this restriction appears to be consistent with the classical statistical inference problem, it ignores inherent dependencies and correlations involved in numerous real-world phenomena that frequently emerge from complex interactions between entities. For instance in traditional data mining processes, it is common to search for purchase models of individuals by focusing on their own attributes (age, city, school, center of interest, etc). Nevertheless, it is obvious that social relationships between individuals, such as friendship or professional relationships,
maintain their environment and determine their behavior and decisions.

The last decade social links have become the subject of a novel active research area, the “Science of Network” (Barabasi, 2002, Watts, 2004, Borner et al., 2007), a new way of studying relationships maintained between entities with new techniques regarding older traditional approaches followed in sociology and discrete mathematics. While previous works provided most popular results among which Milgram’s experiment (Milgram, 1967) on small world phenomenon or Bott’s work (Bott, 1957) on urban families, a great emphasis has been put on network structures and gave important findings. Network structures have been analysed according to recently defined measures such as the degree, density, diameter, clustering coefficient or shortest-path, etc.

The so-called and very novel social network mining area, also called link mining, (Getoor & Diehl, 2005) has attempted to apply the concepts of data mining on networks. Standard data mining techniques (classification, prediction, clustering, frequent pattern discovery), are used to explore the topological structure of the network without considering the nodes attributes. For instance, this is the case of methods for extracting frequent patterns from social networks, which are mostly approaches that only exploit the topological structure to extract sub-networks found frequently in a set of networks or a single much larger network.

However, both sources of information (structure and nodes features) seem to be relevant to take full advantage of the knowledge hidden into the network. While the links describe the nature of the relationships between individuals into the network, the attributes may provide relevant information on their role, position or influence. In the case of the spread of diseases for example, although the social link between two individuals is the main vector of the transmission process, properties such as the age or the origins, can influence quite significantly the probability that an individual contracts the disease, and therefore can influence the whole diffusion process.

Given the increasing number of available data sets, in which, in addition to social links, some information on nodes is also available, recent methods have integrated the complementary dimension. In this field, new algorithms or adaptations of existing algorithms have been proposed in visualization (Snasel et al., 2009), classification (Kuznetsov & Ignatov, 2009) or identification of communities (Gaume et al., 2010).

In this paper, we address the problem of the search for frequent patterns in social networks and we focus particularly on a new approach that combines both the network structure and the attributes of nodes for discovering regular patterns among the links that connect nodes with common characteristics. Such patterns are called “frequent links” and provide knowledge on the groups of node the most connected in the network.

For instance let us assume a social network built on mobile phone contacts, i.e in which two individuals (nodes) are connected if we observe at least one call from one of them to the other one. Each individual is characterized by its home country, the age and the professional status. A frequent link in this kind of network would provide for instance the following knowledge: “20% of calls are between European business men and Chinese people”.

Naïve approaches to extract frequent links are impracticable considering the combinatorial amount issue of an exhaustive search. Indeed, due to the NP-complete nature of the similar problem of frequent itemset mining (Bashir et al., 2006), the efficiency of the extraction process highly depends on the efficiency of algorithm implementation.

That is why in this work we focus on the reduction of the search space and we propose to restrict the space of candidate solutions to a subset of nodes that have most frequent characteristics. This solution is based on the assumption that real human behaviors result
Related Content

An Approach Based on Hierarchical Petri Nets for the Verification of Interconnected BPEL Processes

RT-Llama: Providing Middleware Support for Real-Time SOA
[www.irma-international.org/chapter/llama-providing-middleware-support-real/66806/](www.irma-international.org/chapter/llama-providing-middleware-support-real/66806/)

IIR Filters to Sampling Rate Conversion
[www.irma-international.org/chapter/iir-filters-sampling-rate-conversion/27214/](www.irma-international.org/chapter/iir-filters-sampling-rate-conversion/27214/)

LDAP Vulnerability Detection in Web Applications
[www.irma-international.org/article/ldap-vulnerability-detection-in-web-applications/204523/](www.irma-international.org/article/ldap-vulnerability-detection-in-web-applications/204523/)

Towards Risk Based Effort Estimation: A Framework to Identify, Analyze, and Classify Risk for Early Identification at Requirement Engineering Phase
[www.irma-international.org/article/towards-risk-based-effort-estimation/220457/](www.irma-international.org/article/towards-risk-based-effort-estimation/220457/)