Integrating Clustering and Ranking on Hybrid Heterogeneous Information Network

Ran Wang[1], Chuan Shi[1], Philip S. Yu[2,3], and Bin Wu[1]

[1] Beijing University of Posts and Telecommunications, Beijing, China
{wangran51,shichuan,wubin}@bupt.edu.cn
[2] University of Illinois at Chicago, IL, USA
[3] King Abdulaziz University Jeddah, Saudi Arabia
psyu@cs.uic.edu

Abstract. Recently, ranking-based clustering on heterogeneous information network has emerged, which shows its advantages on the mutual promotion of clustering and ranking. However, these algorithms are restricted to information network only containing heterogeneous relations. In many applications, networked data are more complex and they can be represented as a hybrid network which simultaneously includes heterogeneous and homogeneous relations. It is more promising to promote clustering and ranking performance by combining the heterogeneous and homogeneous relations. This paper studied the ranking-based clustering on this kind of hybrid network and proposed the ComClus algorithm. ComClus applies star schema with self loop to organize the hybrid network and uses a probability model to represent the generative probability of objects. Experiments show that ComClus can achieve more accurate clustering results and do more reasonable ranking with quick and steady convergence.

Keywords: Clustering, Ranking, Heterogeneous Information Network, Probability Model.

1 Introduction

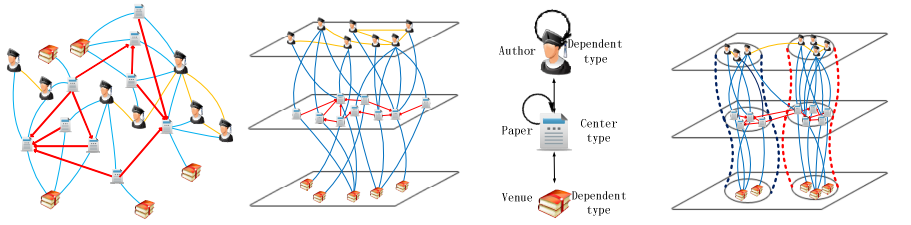
Information network analysis is an increasingly important direction in data mining in the past decade. Many analytical techniques have been developed to explore structures and properties of information networks, among which clustering and ranking are two primary tasks. The clustering task [1] partitions objects into different groups with similar objects gathered and dissimilar objects separated. Spectral method [1,4] is widely used in graph clustering. The ranking task [6,10,12] evaluates the importance of objects based on some ranking function, such as PageRank [12] or MultiRank [10]. Clustering and ranking are often regarded as two independent tasks and they are applied separately to information network analysis. However, integrating clustering and ranking makes more sense in many applications [2-3,11]. On one hand, the knowledge of important objects in a cluster helps to understand this cluster; on the other hand, knowing clusters is benefited to make more elaborate ranking. Some preliminary works have explored this issue [11].

J. Pei et al. (Eds.): PAKDD 2013, Part I, LNAI 7818, pp. 583–594, 2013.

Although it is a promising way to do clustering and ranking together, previous approaches confine it to a "pure" heterogeneous information network which does not consider the homogeneous relations among same-typed objects. For example, RankClus [2] only considers relations between two-typed objects; NetClus [3] just considers relations among center type and attribute types. However, in many applications, the networked data are more complex. They include heterogeneous relations among different-typed objects as well as homogeneous relations among same-typed objects. Taking bibliographic data as an example which is shown in Fig. 1(a), papers, venues, authors and their relations construct a heterogeneous information network. Simultaneously, the network also includes the citation relations among papers and the social network among authors. It is important to cluster on such a hybrid network which includes heterogeneous and homogeneous relations at the same time. The hybrid network can more authentically represent real networked data. Moreover, more information from heterogeneous and homogeneous relations is promising to promote the performance of clustering and ranking.

Although it is important to integrate clustering and ranking on the hybrid network, it is seldom studied due to the following challenges. 1) It is difficult to effectively organize networked data. The hybrid network is more complex than either of them. The way to organize the network not only needs to effectively represent objects and their relations but also benefits for clustering and ranking analysis. 2) It is not easy to integrate information from heterogeneous and homogeneous relations to improve clustering and ranking performances. It is obvious that more information from different sources can help to obtain better performances. However, we need to design an effective mechanism to make full use of information from these two networks.

In this paper, we study the ranking based clustering problem on a hybrid network and propose a novel ComClus algorithm to solve it. A star schema with self loop is applied to organize the hybrid network. The ComClus employs a probability model to represent the generative probability of objects and the experts model and generative method are used to effectively combine the information from heterogeneous and homogeneous relations. Moreover, through applying the probability information of objects, we propose ComRank to identify the importance of objects based on ComClus. Experiments on DBLP show that ComClus achieves better clustering and ranking accuracy compared to well-established algorithms. In addition, ComClus has better stability and quicker convergence.



(a) Bibliographic data (b)Hybrid network (c)Star schema with self loop (d)Clusters on hybrid network

Fig. 1. An example of clustering on bibliographic data