On Perspective-Aware Top-k Similarity Search in Multi-relational Networks

Yinglong Zhang¹, Cuiping Li¹, Hong Chen¹, and Likun Sheng²

¹ Key Lab of Data Engineering and Knowledge Engineering of MOE, and Department of Computer Science, Renmin University of China, China
zhang_yinglong@126.com, cuiping_li@263.net
² JiangXi Agricultural University, China

Abstract. It is fundamental to compute the most “similar” k nodes w.r.t. a given query node in networks; it serves as primitive operator for tasks such as social recommendation, link prediction, and web searching. Existing approaches to this problem do not consider types of relationships (edges) between two nodes. However, in real networks there exist different kinds of relationships. These kinds of network are called multi-relational networks, in which, different relationships can be modeled by different graphs. From different perspectives, the relationships of the objects are reflected by these different graphs. Since the link-based similarity measure is determined by the structure of the corresponding graph, similarity scores among nodes of the same network are different w.r.t. different perspectives. In this paper, we propose a new type of query, perspective-aware top-k similarity query, to provide more insightful results for users. We efficiently obtain all top-k similar nodes to a given node simultaneously from all perspectives of the network. To accelerate the query processing, several optimization strategies are proposed. Our solutions are validated by performing extensive experiments.

Keywords: Random walk, Multi-relational network, Graph, Proximity.

1 Introduction

Recent years have seen an astounding growth of networks in a wide spectrum of application domains, ranging from sensor and communication networks to biological and social networks [1]. At the same time, a number of important real world applications (e.g. link prediction in social networks, collaborative filtering in recommender networks, fraud detection, and personalized graph search techniques) rely on querying the most “similar” k nodes to a given query node. The measure of “similarity” between two nodes is the proximity between two nodes
Fig. 1. A coauthor network $G$

Fig. 2. The graph $G_{DB}$ from perspective of DB

Fig. 3. The graph $G_{DM}$ from perspective of DM

Fig. 4. The graph $G_{IR}$ from perspective of IR

Based on the paths connecting them. For example, random walk with restart (RWR) [2], Personalized PageRank (PPR) [3], SimRank [4], and hitting time [5] are all such kinds of measures. These measures are computed based on the structure of graphs.

The question, computation of the most “similar” $k$ nodes to a given query node, has been studied in these researches [2,6,7,8]. Although their works are excellent, they did not consider the query under a specific viewpoint. A query, top $k$ similar authors w.r.t. Jiawei Han in the database field, is more interesting and useful than the query, that without the viewpoint, for people who are interested in the research of database.

Actually, as mentioned in [9,10,11,12,13], there may exist different kinds of relationships between any two nodes in real networks. For example, in a typical social network, there always exist various relationships between individuals, such as friendships, business relationships, and common interest relationships [10]. So different relationships can be modeled by different graphs in multi-relational networks. And these different graphs reflect relationships among objects from different perspectives. Correspondingly, the top $k$ similarity query based on these graphs will return different answers.

Here an example is given:

**Example 1.** A network $G$ of coauthor relationships is showed in Figure 1. In the figure, relationships are extracted based on the publish information from database (DB), data mining (DM), and information retrieval (IR) fields. Relationships of coauthor in different fields are denoted by different colors (DB, DM, and IR are denoted by red, green, and black edges respectively in Figure 1). The graph showed in figure 2 is modeled based on coauthor relationships of DB field. Similarly, Figure 3 and Figure 4 show the graphs from DM and IR perspective respectively. $G = G_{DB} \cup G_{DM} \cup G_{IR}$ is also considered as the graph from perspective of DB or DM or IR. Obviously, the corresponding structure from different perspective is different for $G$. The graph $G'$ in figure 5 reflects the coauthor relationship among authors without considering the specific research field, on which the traditional top-$k$ query is performed. $G'$ is the corresponding simple graph of $G$.

Given a query node $q$, if we want to know the most “similar” nodes w.r.t. $q$ from DB perspective, the result will be determined by the graph in figure 2 (rather than $G$ or $G'$). Given the query node 4, Table 1 shows its top-3 similar nodes from different perspectives. The result of the traditional query based on $G'$ (without considering a specific viewpoint) is (5, 10, 1), while the result is