Information Systems 42 (2014) 107–122

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

Assessing single-pair similarity over graphs by aggregating first-meeting probabilities

Jun He [a,d], Hongyan Liu [b,*], Jeffrey Xu Yu [c], Pei Li [a], Wei He [a], Xiaoyong Du [a]

[a] Key Labs of Data Engineering and Knowledge Engineering, Ministry of Education, China
[b] Department of Management Science and Engineering, Tsinghua University, China
[c] Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong, Hong Kong, China
[d] School of Information, Renmin University of China, China

ARTICLE INFO

Article history:
Received 30 July 2012
Received in revised form
29 July 2013
Accepted 24 December 2013
Recommended by T.G.K. Calders
Available online 7 January 2014

Keywords:
Graph mining
Similarity measure
Algorithm
Link graph
SimRank
First-meeting probabilities

ABSTRACT

Link-based similarity plays an important role in measuring similarities between nodes in a graph. As a widely used link-based similarity, SimRank scores similarity between two nodes as the first-meeting probability of two random surfers. However, due to the large scale of graphs in real-world applications and dynamic change characteristic, it is not viable to frequently update the whole similarity matrix. Also, people often only concern about the similarities of a small subset of nodes in a graph. In such a case, the existing approaches need to compute the similarities of all node-pairs simultaneously, suffering from high computation cost.

In this paper, we propose a new algorithm, *Iterative Single-Pair SimRank* (ISP), based on the random surfer-pair model to compute the SimRank similarity score for a single pair of nodes in a graph. To avoid computing similarities of all other nodes, we introduce a new data structure, position matrix, to facilitate computation of the first-meeting probabilities of two random surfers, and give two optimization techniques to further enhance their performance. In addition, we theoretically prove that the time cost of ISP is always less than the original algorithm SimRank. Comprehensive experiments conducted on both synthetic and real datasets demonstrate the effectiveness and efficiency of our approach.

1. Introduction

Measuring similarities between objects plays an important role in many real-world applications such as information retrieval and recommendation systems. Link-based similarity measures are complement to the traditional content-based similarity measures [1]. Different from the content-based similarity measures, which are based on the content and are represented by a vector space model [2], the link-based similarity measure is based on relationships between objects that are described in a graph in which objects and relationships are modeled as nodes and edges respectively. Examples of such graphs include citations between papers, social relationships in human or web hyperlink graphs. Effective and efficient computation of similarity measures between objects in a graph can greatly enhance information searching and analyzing [3–8], especially when the relationships among objects are complex.

Among the link-based similarity measures in the literature, SimRank [4] has attracted a considerable attention due to its intuition and sturdy theoretical foundation. The basic intuition behind SimRank is "*two objects are similar if they are referenced by similar objects*", which implies a mutual reinforcement naturally, by updating similarity score of (a, b) (denoted by $S(a, b)$) according to similarity scores of all in-neighbor pairs of (a, b) on the previous iteration. Based on

* Corresponding author.
E-mail address: hyliu@tsinghua.edu.cn (H. Liu).

0306-4379/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.is.2013.12.008

the random surfer model [9], SimRank owns a theoretical foundation stemming from PageRank [10] and HITS [11]. For link-based similarity measures, SimRank is considered as one of the promising ones that have a comparable impact as PageRank has for link-based ranking [1].

However, given a pair of nodes (a, b), the efficiency of computing SimRank $S(a, b)$ is an obstacle for its applicability on a large graph. For a large graph $G(V, E)$ the time complexity required for K iterations is $O(Kn^2d^2)$, where n is the number of nodes in G and d is the average incoming degree of nodes, and $O(Kn^4)$ in the worst case. Hence, new optimization techniques for SimRank computation are needed. In the literature, there exist four reported studies on SimRank optimization [1,7,12,13]. These optimization techniques have their own merits and work effectively to obtain similarity scores of every pair of nodes in a graph. However, in many cases, a user only needs to assess the similarity of some node-pairs instead of all of node-pairs. For example, given a co-author network, users may want to know who is more similar to Prof. Jiawei Han, Prof. Philip Yu or Prof. Bing Liu, in terms of research interest. In this case, users need to know the similarity of node-pairs (Jiawei Han, Philip Yu) and (Jiawei Han, Bing Liu). But if we compute the similarity of one node-pair (a, b) using the existing approaches, it becomes cumbersome for the following main reason. SimRank computes similarity $S(a, b)$ based on the similarity of all in-neighbors of a and b, which means that the SimRank of neighbors needs to be computed beforehand. In other words, every pair of nodes needs to be computed. A research issue we focus on in this paper is whether we can efficiently compute $S(a, b)$ by avoiding unnecessary computational cost on computing similarity scores of all node-pairs.

As an example, consider a simple graph G shown in Fig. 1(a). For SimRank [4] and all its current optimizations, $S(a, b)$ is updated according to the similarity of all (a, b)'s in-neighbor pairs, that is any node-pair $(x, y) \in \{c, d\} \times \{a, d\}$. Hence, the similarity of each node-pair (x, y) should be computed beforehand, analogously for the similarity of all in-neighbor pairs of (x, y). We call this kind of method *All-Pair SimRank*, in which similarities are mutually reinforced together and the output is a n-by-n similarity matrix over the whole graph. We cannot obtain a single-pair similarity $S(a, b)$ without computing similarities of other node-pairs by All-Pair SimRank.

Another problem of All-Pair SimRank is its inadaptability on time-evolving graphs. Observing that the graph structure of many real-world applications changes over time, addition/removal of edges may result in the change of many similarity scores. This effect is amplified by the nature of mutual reinforcement, which makes it hard to perform incremental computation of All-Pair SimRank with accuracy preservation. Moreover, the graph is usually very large. Frequent update of the similarity of all node pairs is time consuming and unnecessary, when user is only interested in a small number of node pairs. There have been some variants of SimRank [7] that can return similarity of single pair nodes. But they are all approximations of the real SimRank similarity scores.

In this paper, we propose a new *single-pair SimRank* approach to compute $S(a, b)$, by tactfully avoiding computation of the similarity of all node-pairs. We outline our approach below. Given a graph G, based on the viewpoint of the random surfer model, SimRank score can be modeled as the *first-meeting probability* of two random surfers on the reverse graph of G [4,6]. To be more specific, the SimRank score of the kth iteration is the sum of first-meeting probabilities within the first k steps. Hence, the key of our approach is to compute the first-meeting probability of two surfers that start from nodes a and b respectively, and meet somewhere exactly on the kth step, denoted by $M_k(a, b)$ (detail definition will be given in Section 2), as shown in Fig. 1(b). The key issue is how to compute $M_k(a, b)$. To compute $M_k(a, b)$, we first give a naive method called *Naive Single-Pair SimRank* (NSP) using matching path-trees. To enhance the performance of this approach, we propose a new data structure called position matrix and then develop an algorithm called *Iterative Single-Pair SimRank* (ISP) to compute the position matrix iteratively and efficiently. In addition, we propose two optimization techniques to accelerate this computation process and improve scalability respectively.

The main contributions of this paper are summarized below. First, we provide a deep analysis on SimRank computation. Second, we propose a new single-pair approach to compute SimRank score of a given node-pair directly. Third, we propose an algorithm together with two optimization techniques to compute the single-pair SimRank score for any given node pair. By analyzing the time complexity of the algorithms, we prove that the computational cost of ISP is always less than All-Pair SimRank, and is obviously efficient when we only need to assess similarity of one or a few node-pairs. Finally, extensive empirical studies conducted on both synthetic data and

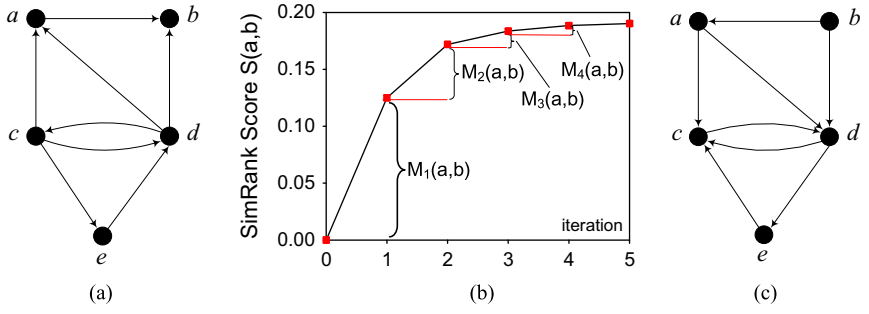

Fig. 1. (a) A tiny graph G, (b) the relationship between SimRank score $S(a, b)$ and first-meeting probability $M_k(a, b)$ on each iteration using factor $C=0.5$ and (c) the reverse graph of G.