Knowledge-Based Systems 150 (2018) 85–94

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Dissimilarity-constrained node attribute coverage diversification for novelty-enhanced top-*k* search in large attributed networks

Zaiqiao Meng[a], Hong Shen[a,b,*]

[a] School of Information Science and Technology, Sun Yat-Sen University, China
[b] School of Computer Science, University of Adelaide, Australia

ARTICLE INFO

Article history:
Received 19 March 2017
Revised 28 February 2018
Accepted 2 March 2018
Available online 17 March 2018

Keywords:
Information retrieval
Query diversification
Top-k search
Submodular maximization

ABSTRACT

Query diversification as an effective way to reduce query ambiguity and enhance result novelty has received much attention in top-*k* search applications on large networks. A major drawback of the existing diversification models is that they do not consider redundancy elimination during the course of search, resulting in unassured novelty in the search result. In this paper, to improve the novelty of the search result, we propose a new method of diversified top-*k* similarity search by combining diversification of node attribute coverage with a dissimilarity constraint. Due to the non-monotonicity implied by the dissimilarity constraint, existing techniques based on monotonicity assumptions cannot be applied. Our model requires solving a new problem of *Dissimilarity Constrained Non-monotone Submodular Maximization* (DCNSM). Based on constructing a dissimilarity-based graph, we solve this problem by a greedy algorithm achieving an approximation ratio of $1/\Delta$, where Δ is the maximum node degree of the dissimilarity-based graph, in time linear to the number of edges of the graph. We show that DCNSM cannot be approximated in ratio $|V|^{1-\epsilon}$, indicating that our solution achieves an optimal ratio. We conduct extensive experiments on both synthetic and real-world attributed network datasets. The results show that our diversification model significantly outperforms the baseline methods, and confirm that combining dissimilarity constraint in diversification can significantly improve the novelty of search result.

1. Introduction

Searching for top-*k* nodes similar to a given query request in a network has numerous applications including graph clustering [15], graph query [27], object retrieval and recommendation [34]. There has been substantial research on ranking nodes and their similarity (proximity) estimation, such as the Personalized PageRank [14] and SimRank [18]. These basic methods and their variations such as P-Rank [38], TopSim [23] and Panther [37] have been successfully applied in a wide range of applications.

Nowadays with rich information available from online social networks, real social entities and their relationships can be built in a network in which nodes are associated with a set of attributes describing their properties and edges representing relationships among the nodes. In such circumstances, the problem of searching for similar nodes to a given node becomes more sophisticated and challenging.

Firstly, the top-*k* nodes resulted from the traditional similarity search methods are often highly related. It is hard to get desired similar results with such few and highly related nodes. Search result diversification has been widely studied as a way of tackling query ambiguity and enhancing result novelty in information retrieval [6,12]. Most of these diversified models tried to trade off relevance and diversity according to some parameter. In the literature there are many studies on modeling the search result diversification for network datasets. Expansion ratio [24] and expanded relevance [20] are two representative diversification models proposed recently based on node's ego features (e.g. neighborhood) diversification and addressed by the solution of the classic Monotone Submodular Maximization problem. However, a major drawback of these models is that they give no explicit mechanism for redundancy elimination, resulting in lack of novelty in their search results.

Moreover, a network with attributes has more complex characters. More precisely, node attributes with the links among them provide rich and complementary sources of information that should be used for revealing, understanding and exploiting the latent diversified structure in attributed network data. For example, in social networks users have profile information, in

* Corresponding author.
E-mail addresses: zqmeng@aliyun.com (Z. Meng), hongsh01@gmail.com (H. Shen).

https://doi.org/10.1016/j.knosys.2018.03.008
0950-7051/

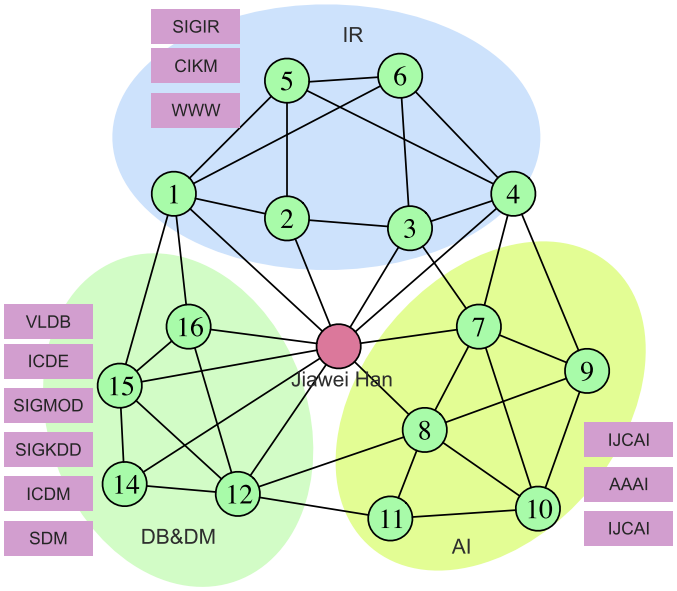

Fig. 1. A portion of Jiawei Han' s academic ego social network. 1

document networks each node also contains the text of the document that it represents. Fig. 1 depicts a portion of Jiawei Han's academic ego social network extracted from DBLP.[1] Three research communities, IR, DB&DM and AI, are distinguished by three cluster based on their common attributes (the conferences they have published papers), which are the non-topological features. Therefore, diversification should also be in presence for these non-topological features. However, the existing diversification models focus purely on network topologies only, and totally ignore that attributes also expose diversity that should be fully exploited to meet the ambiguous query intend.

In this paper, to model the diversification search problem in attributed networks, we first formulate a problem that considers only the *attribute coverage diversification* (ACD). We show that the optimization objective of the ACD problem is a non-decreasing submodular function, and give a marginal gain based greedy algorithm that yields a $(1-1/e)$-approximation near-optimal solution. To further improve novelty of the search result, we extend the ACD problem with the r-dissimilar constraint that captures the dissimilarity between result nodes based on the topological structure of the graph. We show that the new problem becomes more complicated, because adding a new node to the result set following the monotonicity may break the dissimilarity constraint and hence the existing techniques cannot be applied any more. Our model requires to solve a new problem called *Dissimilarity Constrained Non-monotone Submodular Maximization* (DCNSM). Based on constructing a dissimilarity-based graph, we propose a greedy algorithm achieving an approximation ratio of $1/\Delta$, where Δ is the maximum degree of its dissimilarity-based graph, and runs in $O(k(\bar{a}|V|+|E|))$ time.

The main contributions of this paper are: (1) We formulate the problem of node*attribute coverage diversification* (ACD) for top-k similarity search in attributed networks as that of maximizing a monotone submodular function, and give a $(1-1/e)$-approximation near-optimal solution. (2) We formulate the dissimilarity-constrained node attribute coverage diversification problem, an extended ACD problem, as that of maximizing a dissimilarity-constrained non-monotone submodular function. We prove that no $|V|^{1-\epsilon}$-ratio approximation scheme exists for this problem for any $\epsilon > 0$ and present a linear-time (to $|E|$) algorithm achieving the optimal approximation ratio $1/\Delta$, where Δ is the maximum degree of its dissimilarity-based graph. (3) We conduct extensive experiments on both synthetic network datasets and real-world attributed network datasets, and the results show that our proposed dissimilarity-constrained node attribute coverage diversification method significantly outperforms other methods.

[1] http://dblp.uni-trier.de/xml/.

2. Related work

The work presented in this paper is closely related to *similarity search on networks, submodular function maximization*, and *search result diversification on networks*.

2.1. Similarity search on networks

There have been various measures to estimate similarity between nodes on networks. Personalized PageRank (PPR) [14] is a random walk based measure evolved from the classic PageRank algorithm [29]. Similar to PPR, SimRank is defined recursively with respect to the "random surfer-pairs model", and it evaluates the similarity between two nodes as the first-meeting probability of two random surfers. Existing algorithms like P-Rank [38], TopSim [23] are extensions of SimRank. Some other examples include discounted/truncated hitting time [31], penalized hitting probability [36], and nearest neighbor [2,35] are also referred to as the random walk based method. Recently, a random path sampling based method—Panther [37] was proposed that can provably estimate the similarity between nodes efficiently and accurately.

2.2. Submodular function maximization

Submodularity is a property of set functions with deep theoretical consequences and far-reaching applications. Submodular set functions have been widely applied to many fields, including document summarization [25], image segmentation [17], sensor placement [19], diversifying search [3,24], and algorithmic game theory [8]. Submodular function maximization captures classic NP-hard problems in the combinational optimization such as *max cut, maximum facility location* and *max k-cover* problems [4,5,11,33]. Some studies dealt with submodular function maximization subject to various combinatorial constraints, such as the bases of a matroid [33], multiple knapsack constraints [22] and submodular knapsack [16]. To the best of our knowledge, there is no published work for the problem of maximizing submodular functions subject to a dissimilarity (distance) constraint.

2.3. Search result diversification on networks

There are several studies on search results diversification in network data. DivRank [26] employs a time-variant random walk process to facilitates the rich-gets-richer mechanism in node ranking. Tong, et al. [32] propose a scalable diversified ranking algorithm by optimizing a predefined diversified goodness measure. Recently a neighbor expansion based diversified ranking method was proposed, with the assumption that nodes with large expansion would be dissimilar to each other [24]. Küçüktunç [20] proposed a measure called expanded relevance which combines both relevance and diversity into a single function in order to measure the coverage of the relevant part of the graph. These methods are designed to work with the simplified structural network without considering node attributes. Our diversification model focuses on the attributed networks, and combines with a dissimilarity constraint as an explicit measure to eliminate redundancy.