Information Processing and Management xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

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

Search result diversification on attributed networks via nonnegative matrix factorization

Zaiqiao Meng[a,*], Hong Shen[*,a,b], Huimin Huang[a], Wei Liu[a], Jing Wang[c], Arun Kumar Sangaiah[d]

[a] School of Data and Computer Science, Sun Yat-sen University, China
[b] School of Computer Science, The University of Adelaide, Adelaide, Australia
[c] Neusoft Institute, Guangdong, China
[d] School of Computing Science and Engineering, VIT University, Vellore, India

ARTICLE INFO

Keywords:
Graph search
Diversification
Nonnegative matrix factorization
Attributed network

ABSTRACT

Search result diversification is an effective way to tackle query ambiguity and enhance result novelty. In the context of large information networks, diversifying search result is also critical for further design of applications such as link prediction and citation recommendation. In previous work, this problem has mainly been tackled in a way of implicit query intent. To further enhance the performance on attributed networks, we propose a novel search result diversification approach via nonnegative matrix factorization. Our approach encodes latent query intents as well as nodes as representation vectors by a novel nonnegative matrix factorization model, and the diversity of the results accounts for the query relevance and the novelty w.r.t. these vectors. To learn the representation vectors of nodes, we derive the multiplicative updating rules to train the nonnegative matrix factorization model. We perform a comprehensive evaluation on our approach with various baselines. The results show the effectiveness of our proposed solution, and verify that attributes do help improve diversification performance.

1. Introduction

Social network platforms such as Facebook provide a wealth of information resources that enable users to search information about their opinions, status and friends. To effectively search which users are similar to a query user in an information network is critical for further design of applications such as community detection (Cheng, Zhou, & Yu, 2011), social recommendation (Lee & Brusilovsky, 2017) and link prediction (Wu, Zhang, & Ren, 2017). As an inherently limited representation of a complex information need, queries submitted to information networks system is often ambiguous to some extent. While real social networks keep expanding in scales, the search result and recommended list that can be presented to users is only in a limited number. Hence, re-ranking the search results in order to meet the ambiguous complex information needs is challenging.

Search result diversification as an effective way to tackle query ambiguity and enhance result novelty has gained much attention in various fields, including text retrieval (Liang, Ren, & De Rijke, 2014a; Liang & de Rijke, 2015), recommender systems (Küçüktunç, Saule, Kaya, & Çatalyürek, 2015), expert finding (Liang & de Rijke, 2016) and graph based ranking (Li & Yu, 2013). A number of algorithms of search result diversification have been proposed, which can be simply classified into two categories according to whether the *query aspects*, the aspects of information needs behind a query, is implicit or explicit. For those implicit approaches,

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

https://doi.org/10.1016/j.ipm.2018.05.005
Received 20 January 2018; Received in revised form 23 March 2018; Accepted 15 May 2018

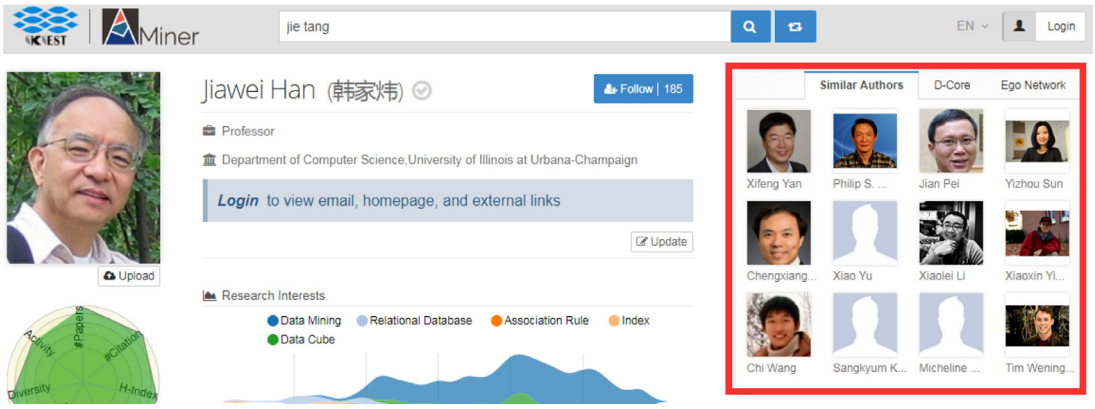

Fig. 1. An example of recommendation in AMiner academic social network.

clearly defined intent of query might be underspecified or implicit. It is not clear which aspects behind this query is what users actually interested in. To diversify the search results under this problem setting, the idea is to make the search results covering certain diversity goodness as much as possible so that the majority users with different backgrounds could find the results to be relevant to their needs (Agrawal, Gollapudi, Halverson, & Ieong, 2009; Li & Yu, 2013; Liang, Cai, Ren, & de Rijke, 2016; Liang, Yilmaz, Shen, de Rijke, & Croft, 2017). For those explicit approaches, a set of aspects underlying the query intent might be explicitly available. Rather than aiming to satisfy as many users as possible, the problem becomes to select a set of representative candidates that collectively provide maximum relevance w.r.t. these explicit query aspects (Liang, Ren, & De Rijke, 2014b). For the explicit query aspect diversification problem, quite a lot of past work has focused on explicitly modeling the possible information needs underlying a query. It has shown that explicit approaches are usually somewhat superior to implicit diversification techniques in text retrieval (Santos, Macdonald, & Ounis, 2010).

Diversifying the searching and ranking results on large information networks has also become an increasingly crucial problem, as it has a large number of potential motivating applications such as link prediction (Zhang, Yu, & Zhou, 2014), citation recommendation (Küçüktunç et al., 2015), as well as personalized services for social networks (Li & Yu, 2013).

For instance, in Fig. 1, we show a non-diversified solution of author recommendation for Professor Jiawei Han in AMiner.[1] AMiner provides a similar authors list to let users more easily find their interested authors when they visit one's personal page. However, the recommended list of this non-diversified strategy lacks of novelty. Since Professor Han's research interests are with many areas, and if only considering the similarity, the recommended result will only present one of his main research areas. As a result, users may lose their interest when they go the first few recommendations over. In this situation, making the limited sized recommended list diverse should be considered.

Existing diversification approaches in searching of networks are typically built on implicit representation of the query intent, based upon some global diversified goodness of nodes, such as rich-gets-richer mechanism (Mei, Guo, & Radev, 2010), the direction-aware goodness (Tong, He, Wen, Konuru, & Lin, 2011) and the neighbor expansion (Li & Yu, 2013). However, it is questionable that nodes with such global diversified goodness are what query users are actually interested in. These works try to identify the 'correct' interpretation of diversified intent behind any query, without taking into account the intent of the query node itself. Consequently, important aspects of the query node may be overlooked simply because they are not well represented among the whole network; conversely, marginally important aspects may be overemphasized due to the heterogeneity of networks. Moreover, nodes in real-world information networks, such as online social networks and scientific collaboration networks, are often associated with a rich set of features or attributes. Existing diversified search algorithms focus on pure networks only and ignore that the attributes also present diverse features.

Therefore, in this paper, we study the problem of *search result diversification on attributed networks*. In particular, given an information network **G** with node attributes matrix **A**, a query node q and result size k, our problem is searching for a result set of k nodes that are similar to q and diverse to each other w.r.t. the query aspects. In this problem settings, the query aspects can be represented by the local subgraphs that contain the query node, e.g. research interest groups in AMiner and social circles in Facebook.

To address this problem, we propose a new search result diversification approach called **NMFDIV**, in which query aspects and nodes are encoded as representation vectors and these representation vectors then can be directly leveraged for diversifying result nodes. More precisely, our NMFDIV model first learns low-dimensional feature representation vectors of nodes through a nonnegative matrix factorization model, then select representative nodes according to a combined diversification optimization of aspect relevance and result novelty.

Our contributions can be summarized as follows:

- We tackle the challenges of search result diversification on attributed networks in a novel way that represents the possible

[1] http://www.aminer.org/. An academic social network platform.