Information Sciences 265 (2014) 211–224

Contents lists available at ScienceDirect

Information Sciences

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

Development of multidimensional academic information networks with a novel data cube based modeling method

Mehmet Kaya[a], Reda Alhajj[b,c,*]

[a] Department of Computer Engineering, Firat University, 23119 Elazig, Turkey
[b] Department of Computer Science, University of Calgary, Calgary, AB, Canada
[c] Department of Computer Science, Global University, Beirut, Lebanon

ARTICLE INFO

Article history:
Received 2 October 2012
Received in revised form 8 September 2013
Accepted 8 November 2013
Available online 19 November 2013

Keywords:
Academic network
OLAP
Data cube
Information network
Multi-agent system

ABSTRACT

A common task in many applications is to find people who are knowledgeable about a given topic, topics which are suitable for a given author or venue, and venues which are attractive for a given author or topic. This problem has many real-world applications and has recently attracted considerable attention. However, the existing methods are not very efficient in providing flexibility for multi-dimensional and multi-level view from different perspectives. In this paper, we first propose and develop three different academic networks with a novel data cube based modeling method, and then we perform automated decision processes on these networks. As the first step of the study, we integrate DBLP and CiteSeerX by employing a simple technique called canopy clustering. After the integration of the databases, the modeling stage of the academic networks is performed. In this study, each node as apart from the studies described in the literature is represented by a corresponding data cube with respect to the kind of the network being considered. In order to appropriately analyze the data cube, the OLAP technology is utilized. As the next step of the study, our aim is to automatically find relevant persons, topics and venues from each network. However, it is not an easy task to extract knowledge with low running time and high accuracy from such very huge information networks. In order to overcome this problem, a multi-agent based algorithm is proposed. We evaluate our method with the author network using a benchmark dataset of how well the expertise of the proposed experts matches a given query topic. Our experiments covering other networks show that the proposed strategies are all effective to improve the retrieval accuracy.

1. Introduction

A social network describes a group of social entities and the patterns of inter-relationship among them. The semantics and interpretation of a relationship varies from social nature such as kinship or friendship among people, to transactional nature such as trading relationship between countries, biological nature such as the interaction between molecules within the cell. Despite the variability in semantics, social networks share a common structure in which social entities, generally termed actors, are inter-linked through units of relationship between a pair of actors known as tie, link or pair. By representing actors as nodes and ties as edges, a social network can be represented as a graph which could be directed or undirected.

A constructed social network can be analyzed for many useful insights [11,16,22,27]. For instance, the important actors in the network, those with the most connections, or the greatest influence can be found. Alternatively, it may be the connection

[*] Corresponding author. Tel.: +1 403 210 9453.
E-mail addresses: kaya@firat.edu.tr (M. Kaya), alhajj@cpsc.ucalgary.ca (R. Alhajj).

http://dx.doi.org/10.1016/j.ins.2013.11.012

paths between actors that are of interest. Analysts may look for the shortest paths [24], or the most novel types of connections [27]. Sometimes, the focus may even be on finding subgroups, which are subsets of the network that are especially cohesive or interesting [2,14].

It is known that social networks are usually represented as graphs or networks. Each edge in the graph or network represents a relationship; and the strength of a relationship is depicted by the weight associated with the corresponding edge. Moreover, each edge can be directed or undirected, hence representing an asymmetric or symmetric relationship. How the construction of social networks relations varies from application to application. However, data mining and machine learning techniques can provide significant insight on the modeling on social networks based on data properties and semantics [13].

An academic information network forms a specific network different from other social networks. It may consist of scholars and researchers from various research domains [8,23]. Additionally, a lot of academic activities, such as academic conferences, workshops, or even forums are held regularly that enable researchers to capture new research trends and exchange research ideas. Currently, vast amount of these activities including scientific publications are stored in bibliographic databases, such as DBLP and CiteSeerX. Inspired by social networks analysis, works on bibliographic databases have proposed different alternatives for modeling bibliographic information using graphs. These can be divided into two main categories. For methods in the first category, n-partite graphs are created (which contain for instance authors, conferences or topics as nodes) that connect nodes of different types and represent relations (e.g., an author has published a paper in a conference) [19–21,25]. For methods in the second category, graphs with a single node type and edges that may vary in meaning depending on the application are constructed [7,10].

The aim of our work described in this paper is to develop three different academic information networks, *Authors*, *Topic* and *Venue*, with a novel data cube based modeling method. We integrate DBLP and CiteSeerX in a way that allows us to automatically perform decision processes on these networks. DBLP is an on-line resource providing bibliographic information on major computer science conference proceedings and journals. It includes publications, authors, venues and years. Unlike DBLP, CiteSeerX allows for retrieving citations, co-authorships, addresses, and affiliations of authors and publications.

The first step of the work described in this paper is to integrate the two databases. For this purpose, we employ a simple technique called canopy clustering. After the integration of the databases, we moved to the modeling stage of the academic information networks. In these networks, each node may represent an author, a topic or a venue with respect to the kind of network. Next, each node will be represented by a data cube. This cube contains some features related to the node. In order to appropriately analyze the data cube, the OLAP technology is utilized. In this paper, we realize some important advantages of using data cube based modeling method. First, the proposed method may be used easily in online social networks because the most attractive characteristic of OLAP is that it keeps summary of the kept data and facilitates for "compute once, use many". Second the dimensions of the data cube can be generalized/specialized dynamically with respect to a given queries and it performs a multi-dimensional analysis by viewing the constructed networks from different aspects.

In the second step, our aim is to automatically find relevant persons, topics and venues for each network. These persons may be reviewers, panelists or program committee members for the *Author* network. However, it is not an easy task to extract the knowledge with low running time and high accuracy from very huge information networks. In order to overcome this problem, in this study, a multi-agent based algorithm is proposed. Unlike single agent based mining on a social network, in multi-agent systems, each agent cooperates and the system finds the results with higher accuracy which will not be obtained by individual agents. Due to parallel working features of multi-agent systems, the time of extracting valuable knowledge is predictable and acceptable in even large networks. We evaluate our method with the author network using a benchmark dataset based on human relevance judgments of how well the expertise of the proposed experts matches a query topic. Our experiments conducted using other networks show the proposed strategies are all effective to improve the retrieval accuracy.

The rest of the paper is organized as follows: Section 2 presents some preliminary concepts regarding the construction of networks from bibliographical databases, OLAP technologies and ranking approaches, and discusses related work. Section 3 introduces the proposed three academic information networks. Section 4 presents our multi-agent based approach for mining potential authors, topics and venues from these networks. In Section 5, we define the experimental setup for our methods. Section 6 reports our findings obtained from the application of the proposed approach to the bibliographical data. Finally, summary, conclusions and future work are included in Section 7.

2. Preliminaries and related work

2.1. Constructing networks from bibliographical databases

From the methods that create n-partite graphs, Zaiane et al., proposed a bipartite model that connects conferences to authors [25]. Tripartite graph models for authors-conferences-topics have also been introduced in the past [21,25]. In these cases, the topics information is extracted from the paper titles and the resulting tripartite models expand the authors-topics model presented in [19]. Finally, in [20] the authors perform domain specific author and conference ranking by analyzing a bipartite author-conference graph using clustering and ranking heuristics.

In another category, the graph nodes are usually the authors, with the edges representing either citation or co-authorship relations between the connected nodes. The first one is a directed citation graph [10], which is usually employed for ranking authors, whereas the second is undirected co-authorship graph, which is mainly used for finding author communities [7].