Knowledge-Based Systems xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Analyzing future communities in growing citation networks

Sukhwan Jung *, Aviv Segev

Department of Knowledge Service Engineering, KAIST, Daejeon 305-701, South Korea

ARTICLE INFO

Article history:
Received 25 October 2013
Received in revised form 22 April 2014
Accepted 23 April 2014
Available online xxxx

Keywords:
Community
Topic detection
Link prediction
Citation network
Community detection

ABSTRACT

Citation networks contain temporal information about what researchers are interested in at a certain time. A community in such a network is built around either a renowned researcher or a common research field; either way, analyzing how the community will change in the future will give insight into the research trend in the future. The paper views the research community as a Social Web where the communication is through academic papers. The paper proposes methods to analyze how communities change over time in the citation network graph without additional external information and based on node and link prediction and community detection. Different combinations of the proposed methods are also analyzed. The identified communities are classified using key term labeling. Experiments show that the proposed methods can identify the changes in citation communities multiple years in the future with performance differing according to the analyzed time span. Furthermore, the method is shown to produce higher performance when analyzing communities to be disbanded and to be formed in the future.

1. Introduction

Citation networks represent a picture of the current situation of research information in a specific field. The network therefore represents communities centered on a specific researcher or on a shared research field. Analyzing how the community will change in the future will give insight into the research trend in the future and how a field will evolve.

Citation network analysis originated with the paper of Garfield et al. (1964) [17], which showed that the analysis indicated a high degree of coincidence between a historian's account of events and the citational relationship between these events. The present work, however, takes the opposite approach and looks to the future: it examines whether the prediction of citation networks can assist in the analysis of future events.

The paper presents a new perspective of viewing the research community as a Social Web where the communication is through academic papers. The goal of the paper is to assist a community member in making a decision as to what "products" (research topics) are good and what topics are less trendy. The idea of using a social analysis approach on the academic community is the main contribution of this work.

The paper presents several methods to analyze how communities change over time in the citation network. The methods are based on a graph representation of the citation community at given time stamps with nodes representing papers and edges representing citations. External information such as author names, institutions, and existing keyword classifications is not used. The prediction methods are composed of different combinations of proposed building block algorithms for node prediction, edge prediction, and community detection. The node prediction analyzes the change in previous years in the number of citations and gives higher probability to highly cited papers. After the node prediction, six link prediction algorithms are compared to analyze the performance. The analysis showed that the link prediction methods can be classified into two categories that contribute to the performance of the community detection. The Louvain method is used as the basic community detection method. Three topic detection methods are used to label the detected communities. The TF/IDF (Term Frequency/Inverse Document Frequency) method is a widely accepted method in IR (Information Retrieval) and is used to extract representative terms, in this case labels, from documents. The Keyword Extraction method uses a statistical algorithm and natural language process technology to analyze the text and identify terms of importance. The Concept Tagging method gives a high-level abstraction of a given text by utilizing natural language process techniques with external databases such as DBpedia or OpenCyc and returns concepts which were not directly mentioned in the text itself. The basic community analysis building blocks are organized in four different methods to provide an analysis of the order in which the methods can be used and of their individual

* Corresponding author. Tel.: +82 1042273922.
E-mail addresses: raphael@kaist.ac.kr (S. Jung), aviv@kaist.edu (A. Segev).

http://dx.doi.org/10.1016/j.knosys.2014.04.036
0950-7051/

contribution to the performance of the prediction. To analyze the models, two citation networks from the Stanford Large Network Dataset Collection from High Energy Physics Theory (18,479 papers, 136,428 citations) and High Energy Physics (30,566 papers, 347,414 citations) are used.

The rest of the paper is organized as follows. The next section reviews the related work. Section 3 describes the methods used for analyzing future communities in citation networks. Section 4 presents the experiments on citation networks, and Section 5 provides a further discussion of the results. Finally, Section 6 provides some concluding remarks.

2. Related work

2.1. Topic Detection and Prediction

Topic Detection and Prediction has been studied in many research fields, to identify newly emerging topics and to capture the possible topics of given documents respectively. Topic Detection and Tracking (TDT) [15] is a multi-site research project aiming to predict novel topics. Its goal is to find a new topic in news systems by effectively identifying the first article or report mentioning the new topic [3]. There have been many studies using Natural Language Processing (NLP) topic detection approaches. The Adaptive Auto Regression (AR) model based on the Recursive Weighted Least Square (RWLS) method is presented to capture the Internet users' psychosocial attention behavior on how 'hot' topics such as 'Olympic Games' grow on the Internet [42]. The topic-conditioned First Story Detection (FSD) method in conjunction with a supervised learning algorithm [44,45] and Document Clustering [46] are used to identify the earliest report on a certain event in news articles. Other methods are also used in topic predictions. Survey analysis has been used to predict the result of a presidential election [26].

2.2. Topic modeling

Latent Dirichlet Allocation (LDA) [5] is a generative probabilistic model using a three-level hierarchical Bayesian model to model multiple topics from collections of text corpora. Its expandable nature has enabled many researchers to build models on top of it. Hierarchical LDA [6] creates a hierarchy of topics using a random partition process called the Chinese restaurant process. LDA-dual model [38] is an extension of the LDA model introduced to simultaneously deal with two types of text to solve the author disambiguation problem. Labeled LDA (L-LDA) [34] is an extension of multinomial Naïve Bayes supervised LDA where topics are constrained to those that directly correspond to the labels of a given document. Spatial LDA (SLDA) [43] is used to graphically categorize and identify images by treating images as documents and partial sections as words.

Commonsense knowledge, or *human knowledge*, is introduced in opinion mining [9] to catch topics incomprehensible by statistical textual models, such as poetry. LDA with WordNet (LDAWN) [8] incorporated the word sense into LDA by using the WordNet lexicon. Commonsense-based Topic Modeling [33] uses human commonsense data instead of common a *bag-of-words* model. While LDA solved some of the issues such as the overfitting problem, its performance still depends on the volume of the given text corpus with which it is trained. The model proposed here does not require training and can capture the meaning of phrases such as *"getting fired"* as opposed to *bag-of-words* based models such as LDA.

2.3. Topic identification

Topic detection focuses on finding a new topic, provided by either the experimenter [19] or the NLP method. Generative models are used to generate documents by selecting a distribution over topics and then selecting each word in the document from a topic chosen according to this distribution [19]. Generative models are used to analyze research paper abstracts from Proceedings of the National Academy of Sciences (PNAS) in order to generate a number of topics which successfully resemble the data structure. Identifying communities in web pages revealed that the communities exhibit hierarchical topic generalization characteristics, showing that the communities in a general setting are shown to reveal common properties of their members such as a common viewpoint or related topics [18]. Dynamic Community Identification [4] can therefore have a large role in topic identification. The conventional definition of communities as "*unusually densely knit subsets of a social network*" is argued as misleading in dynamic social communities in [41], which proposed an optimization-based approach for modeling dynamic community structure; it is shown to accurately track the dynamic community structure of social networks.

2.4. Link prediction

Link prediction models the evolution of a network using its topological characteristics and primarily deals with the prediction of edges between existing nodes. There are a number of different approaches to link prediction [27]. The shortest path between two nodes in a graph is a simple measure of link prediction. Some methods, such as Common Neighbors [30], Jaccard's coefficient [36], Preferential Attachment [30], and Adamic/Adar [1], use the node neighborhood information. The whole path within the network can also be used in link prediction, for example Katz [21], Simrank [20], and Rooted PageRank [27]. Common to those algorithms is that they do not deal with addition of nodes and deletion of edges. Their purpose is to generate a ranked list of predictive edges between existing nodes in a given network. The Community Prediction Method in the Citation Networks section outlines the differences between these methods and the contribution of each of these methods to the prediction.

2.5. Community detection

Community detection searches structural information of a given graph to partition it into sub-graphs called communities or modules [23]. Agglomerative methods and divisive methods are commonly used in community detection. Newman's community detection algorithm [31] is a widely used agglomerative method that uses modularity as the quality function. The recently developed Louvain method [7] is an agglomerative method and is commonly used because of its low computational complexity and high performance. When merging communities, this method considers not only the modularity but also the consolidation ratio. These algorithms, however, do not consider temporal information and disregard important factors such as consistency. Community Evolution [40] and Evolutionary Clustering [10,12,13] take the temporal changes in networks into consideration. There have also been studies about utilizing communities with link predictions. Family and friendship ties can be regarded as known community structures and are shown to help in predicting links in social networks [47]. The current work proposes a set of models based on temporal graphs and community prediction techniques.

3. Community prediction method in citation networks

Citation networks are directed social networks [32] between research papers, with nodes as papers and edges as citations between them. It is a form of a network where link prediction can be applied without extra considerations about deleted edges

Please cite this article in press as: S. Jung, A. Segev, Analyzing future communities in growing citation networks, Knowl. Based Syst. (2014), http://dx.doi.org/10.1016/j.knosys.2014.04.036