Knowledge-Based Systems 115 (2017) 66–79

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Learning the heterogeneous bibliographic information network for literature-based discovery

Yakub Sebastian[a,*], Eu-Gene Siew[b], Sylvester Olubolu Orimaye[a]

[a] School of Information Technology, Monash University Malaysia
[b] School of Business, Monash University Malaysia

ARTICLE INFO

Article history:
Received 5 May 2016
Revised 30 September 2016
Accepted 4 October 2016
Available online 25 October 2016

Keywords:
Literature-based discovery
Heterogeneous bibliographic information network
Link prediction

ABSTRACT

This paper presents HBIN-LBD, a novel literature-based discovery (LBD) method that exploits the lexico-citation structures within the heterogeneous bibliographic information network (HBIN) graphs. Unlike other existing LBD methods, HBIN-LBD harnesses the metapath features found in HBIN graphs for discovering the latent associations between scientific papers published in otherwise disconnected research areas. Further, this paper investigates the effects of incorporating semantic and topic modeling components into the proposed models. Using time-sliced historical bibliographic data, we demonstrate the performance of our method by reconstructing two LBD hypotheses: the *Fish Oil and Raynaud's Syndrome* hypothesis and the *Migraine and Magnesium* hypothesis. The proposed method is capable of predicting the future co-citation links between research papers of these previously disconnected research areas with up to 88.86% accuracy and 0.89 F-measure.

1. Introduction

Literature-based discovery (LBD) is a systematic computational approach for making novel inferences about previously unknown connections across disparate research fields by chaining together complementary pieces of knowledge from their respective literatures [42]. Using LBD, a novel assertion such as *'dietary fish oil alleviates Raynaud's Syndrome'* can be inferred based on pre-existing assertions in the existing literatures, for example *'dietary fish oil lowers blood viscosity'* and *'high blood viscosity is observed among Raynaud's Syndrome sufferers'*. Note that these assertions have been previously published in disparate groups of research papers [53].

Basic LBD techniques search for a set of intermediate terms that frequently co-occur with a source term and a target term [42]. Following the example above, the term '*blood viscosity*' is one of the instrumental intermediate terms in associating the source term '*dietary fish oil*' with the target term '*Raynaud's Syndrome*'. More sophisticated LBD methods incorporate natural language processing (NLP) techniques with domain-specific ontologies. For instance, Hristovski et al. [20] used a third-party NLP tool to automatically extract complementary *subject-relation-object* predications from a biomedical corpus. These extracted predications could then be used for inferring novel relationships in literatures.

These existing LBD methods have several limitations. A term co-occurrence method typically suffers from the imprecise meaning of such co-occurrences [27]. On the other hand, NLP-based methods are effective only when they are applied to mining literatures in a certain domain for which the required NLP tools and ontologies are easily available [32]. Most importantly, these existing methods have not exploited the valuable bibliographic metadata that are easily available in most scientific publications.

In this paper, we extend the state-of-the-art of the current literature-based discovery research. We propose a new LBD method that harnesses the lexico-citation information found in a heterogeneous bibliographic information network (HBIN). Fig. 1 illustrates an example of HBIN graph. Unlike previous works, we view literature-based discovery as a link prediction problem with the goal of answering the following research question: '*how do we accurately predict the future co-citation links between research papers in previously disconnected research fields?*'.

A pair of research papers are said to be co-cited if they are cited together by another paper [43]. For LBD, new cross-disciplinary co-citation links that span the boundaries of previously disconnected research fields may point to the convergence of these fields [9]. For example, Swanson's seminal LBD paper formed many new co-citation links between previously disconnected fish oil and Raynaud's Syndrome research papers [48]. Consequently, the

* Corresponding author.
E-mail addresses: yakub.sebastian@monash.edu (Y. Sebastian), siew.eu-gene@monash.edu (E.-G. Siew), sylvester.orimaye@monash.edu (S.O. Orimaye).

http://dx.doi.org/10.1016/j.knosys.2016.10.015
0950-7051/

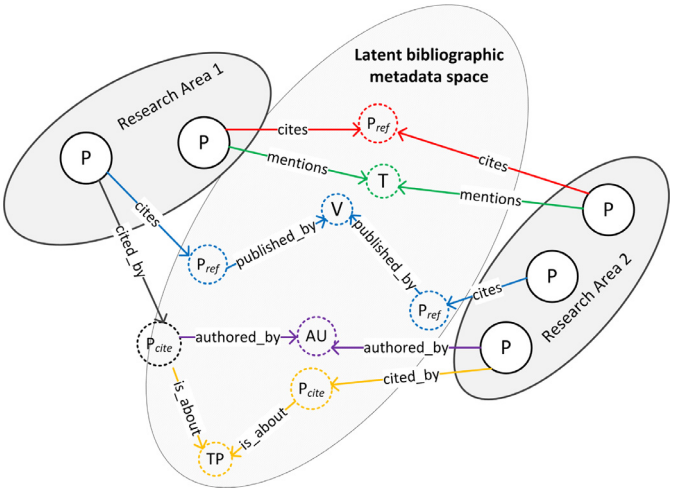

Fig. 1. Illustration of an HBIN graph. *P* nodes refer to papers published in disparate research areas. Latent metapaths between *P* nodes may be formed via various entities in the bibliographic metadata space: *term* (*T*), *author* (*AU*), *publisher* (*V*), *topic* (*TP*), *cited reference* (P_{ref}) and *citing paper* (P_{cite}).

effectiveness of an LBD method can be measured based on its ability to predict the future occurrence of these co-citation links.

Our new method, *HBIN-LBD*, addresses the research problem above by exploiting the latent interconnections between various objects in the bibliographic metadata space of a heterogeneous HBIN graph. These connections include such associations as term co-occurence, co-authorship, and shared references. In this study we also study the effects of applying word sense disambiguation on the proposed model. Finally, we explore the performance gain from incorporating topic modeling into our model.

Our contributions are two-fold. First, we propose a novel literature-based discovery method that mine the latent features in HBIN graphs. To the best of our knowledge, this is the first method that employs heterogeneous information networks for solving LBD tasks. Secondly, we demonstrate the usefulness lexico-citation features of HBIN graphs for predicting the co-citation links between papers from previously disconnected research areas. In addition, we report on the performance gain from incorporating semantic and topic modeling components into the model.

We organize the rest of this paper as follows. Section 2 presents related work. Section 3 introduces our novel technique and algorithms, including some theoretical discussions. In Section 4, we describe our evaluation methodology and present the experimental results. Section 5 further discusses our research findings and highlights the innovation in our work. Finally, Section 6 presents the conclusion and suggests some future research directions.

2. Related work

2.1. Heterogeneous Bibliographic Information Network (HBIN)

The HBIN graph is a special type of heterogeneous information network [17,46]. A collection of scientific publications can be viewed as a network of information that consists of interconnected heterogeneous bibliographic objects. Unlike homogeneous information networks, heterogeneous information networks can encode richer information and better capture different semantics between various real world objects [17]. HBIN allows various information to propagate across different types of objects and links [46]. These information can then be used to capture and model the previously unknown associations between research papers.

Most of the existing LBD methods are based on a simple discovery model known as the *ABC model* [42]. The model suggests that when term *A* co-occurs with term *B* and term *B* co-occurs with term *C*, then it may be inferred that term *A* is possibly related to term *C* [42,53]. Unfortunately, literature-based discovery cannot be solely modeled using just this simplistic model [27,42]. On the other hand, various semantics are known to propagate through different bibliographic objects in HBIN graphs [46]. These information could provide a more holistic way for understanding the previously unknown associations between disjoint research papers in LBD. Instead of performing LBD using just the lexical information (e.g. term co-occurrence), HBIN graphs provide other potentially useful non-lexical information such as citation relations. For example, [26] observed that certain intermediate terms connecting disjoint Parkinson's and Crohn's disease papers could only be found in the titles of their shared references instead of their own titles.

More specifically, mining HBIN graphs allows one to construct composite relations known as *metapaths* by adjoining different types of information links [17,46]. Through a metapath, information that propagates through lexical objects and links such as terms can be seamlessly combined with other non-lexical information propagating through non-lexical objects such as cited references, publishers or authors. As a result, metapaths provide the versatility for exploring different lexico-citation structures that could be useful for an LBD task.

A number of recent LBD methods have explored methods that utilize certain graph data structures. For example, Cameron et al. [8] introduced a method that automatically finds clusters of contextually similar paths in a *semantic predication graph*. These clusters are used to elucidate the latent associations between disjoint concepts in the literatures for reconstructing eight scientific discoveries. However, unlike the HBIN graphs, it does not use heterogeneous information networks and strongly depends on the availability of domain-specific NLP tools and ontology.

In another example, Ding et al. [10] combined the lexical and citation information from the literature in the form of an *entitymetrics graph*. The method models the latent relationships among biological entities (e.g. diseases, drugs) based on the existing citation relationships between their respective research papers. For example, assuming paper *A* cites paper *B*, the method links each biological entity mentioned in paper *A* with each biological entity mentioned in paper *B*. It then computes a clustering coefficient score from the entitymetrics graph and uses the score as a feature for predicting the interactions between genes and drugs. The prediction results are compared with in the entries in the Comparative Toxicogenomics Database (CTD)[1]. Different from HBIN graphs, the entitymetrics graph does not consider bibliographic metadata elements such as authorship or shared references.

2.2. LBD as a link prediction problem

As previously mentioned, this paper considers LBD as a link prediction problem. The goal of link prediction is to predict the occurrence of new links in the future snapshots of a network based on the existing one [14,28]. Link prediction consists of two main steps: (a) learn a number of predictive features from a network, and then (b) use the features to predict the occurrence of a link in a future snapshot of the same network [14].

Kastrin et al. [22] recently proposed formulating LBD as a problem in predicting the implicit links within a co-occurrence network of Medical Subject Headings (MeSH) terms. In contrast, we address a link prediction problem in HBIN graphs. HBIN graphs include various types of bibliographic metadata information and therefore contain richer information than just MeSH terms. Further, unlike MeSH terms which target biomedical literatures, HBIN metadata

[1] http://ctdbase.org/.