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/

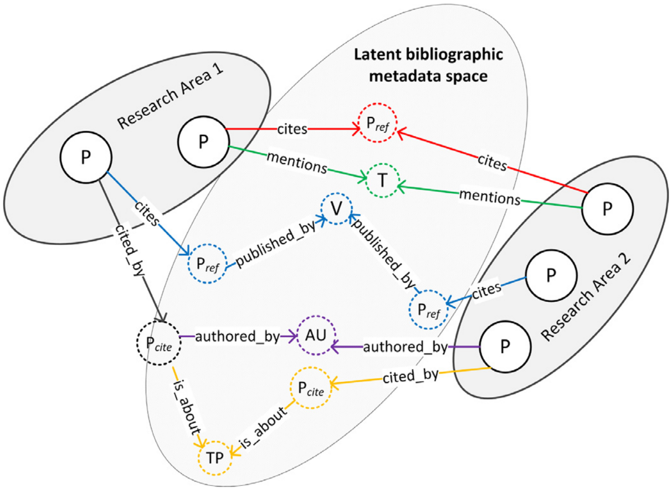


Fig. 1. Illustration of an HBIN graph. *P* nodes refer to papers published in disparate research areas. Latent metapaths between *P* nodes may be formed via various entities in the bibliographic metadata space: *term* (*T*), *author* (*AU*), *publisher* (*V*), *topic* (*TP*), *cited reference* (P_{ref})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

information is not domain-specific and can be easily obtained from literatures in various research fields.

Prior link prediction work has proposed using HBIN metadata information for predicting co-authorship links [45]. We note that their work is different from ours as their predicted co-authorship links do not necessarily span across disparate research fields. In contrast, the focus of LBD is to predict cross-disciplinary links. Besides that, compared to Sun et al.'s method, our method proposes using a different set of metadata structures.

Other works that mined information from HBIN graphs usually aim at solving citation recommendation problems [30,37]. For instance, Ren et al. [37] developed a method that learns the citation interest of a query paper from HBIN graphs in order to recommend a set of relevant citations for it. They used features such as the relevance and authority of bibliographic objects in the HBIN graphs. The algorithm outperformed other link prediction algorithms for predicting the direct citation links between papers in DBLP database[2] with 17.68% improvement on recall (0.4279 vs 0.3636). Liu et al. [30] proposed a method that extracts the citation contexts between research papers and uses this information to learn the citation topics between the papers using a supervised topic modeling algorithm. These topics are then used to compute the citation probability between research papers. Their method outperformed other HBIN-based models with nearly 60% improvement on the Mean Average Precision for predicting direct citation links between research papers in the ACM Digital Library[3].

We emphasize that the works described above are only limited to predicting direct citations between papers, not co-citations. Ren et al. [37] and Liu et al. [30] also placed no requirement that predicted relationships be spanning across two different research fields. In contrast, literature-based discovery aims at predicting the co-citation links between papers from two different research fields. LBD problem requires that these research fields are effectively disconnected from each other such that they have no research paper in common, never cite each other, and have never been co-cited before [50]. As such, our work addresses a fundamentally different problem from the ones addressed by both Ren et al. [37] and Liu et al. [30].

2.3. Semantic aspects and topic modeling

The existing LBD methods typically involve mapping raw terms in texts to corresponding standard concept names [8,20]. This allows the meaning of words to be determined more precisely. It also reduces word sense ambiguities. Owing to the availability of biomedical NLP tools and ontologies, this has been a valid approach for many biomedical LBD methods [42]. However, the effects of semantic processing on domain-independent LBD methods such as HBIN-LBD is not as widely understood [36]. In this work we explore a dictionary-based word sense disambiguation technique and study its effects on the performance of our model.

Similarly, the effects of incorporating topic modeling [6] for literature-based discovery have been rarely studied. In HBIN graphs, topical information between papers usually propagate through links between a term and a paper object. For example, a research paper on fish oil topic is more likely to form a link with the term '*fish*' compared to a paper on Raynaud's Syndrome. In this work we explore the efficacy of using topic modeling for generating new synthetic *topic* objects in the HBIN graphs. These nodes provide an alternative way for propagating topical information between papers other than through term objects. We describe these models in more detail in the following sections.

3. Method and models for learning HBIN for LBD

In this paper, we propose HBIN-LBD, a novel LBD method that learns various lexico-citation features from HBIN graphs. The goal of HBIN-LBD is to discover the latent associations between research papers through the existing interconnections of various bibliographic metadata such as author, term, publisher, cited references, and citing papers. Using a machine learning algorithm, we demonstrate the performance of these features for predicting future co-citations links between a pair of research papers from different fields. We operationalize this problem as a multiclass classification task [18,40].

We believe that HBIN-LBD is the first method that uses heterogeneous *metapath* features from an HBIN graph for performing LBD tasks. These features can be useful for inferring latent associations between two research papers. For instance, research papers that share many similar references may use a common set of background knowledge [23,26]. This information could be used to predict the possible associations between them. Fig. 2 shows an overview of the proposed HBIN-LBD method.

Three different models are proposed in this paper: the (a) *HBIN-LBD*, (b) *HBIN-LBD-Semantic*, and (c) *HBIN-LBD-Topic*. We describe each model in more detail in the following sections.

3.1. The HBIN-LBD model

This first model explores the effectiveness of various metapath features extracted from HBIN graphs for predicting future co-citation links between research papers. We define HBIN as an undirected graph $G = (\mathcal{V}, \mathcal{E})$ with a vertex type mapping function $\tau : \mathcal{V} \rightarrow \mathcal{A}$ and an edge type mapping function $\phi : \mathcal{E} \rightarrow \mathcal{R}$ [46]. We define seven types of vertices, such that $\mathcal{A} = \{P_{core}, P_{ref}, P_{cite}, AU, V, T, TP\}$, and six types of edges $\mathcal{R}$:

1. $v_1 \xrightarrow{written\_by} v_2 \ ; \ \tau(v_1) \in \{P_{core}, P_{ref}, P_{cite}\}, \ \tau(v_2) = AU, \ v_1, v_2 \in \mathcal{V}$
2. $v_1 \xrightarrow{published\_in} v_2 \ ; \ \tau(v_1) \in \{P_{core}, P_{ref}, P_{cite}\}, \ \tau(v_2) = V, \ v_1, v_2 \in \mathcal{V}$
3. $v_1 \xrightarrow{contains} v_2 \ ; \ \tau(v_1) \in \{P_{core}, P_{ref}, P_{cite}\}, \ \tau(v_2) = T, \ v_1, v_2 \in \mathcal{V}$
4. $v_1 \xrightarrow{cites} v_2 \ ; \ \tau(v_1) = P_{core}, \ \tau(v_2) = P_{ref}, \ v_1, v_2 \in \mathcal{V}$
5. $v_1 \xrightarrow{cited\_by} v_2 \ ; \ \tau(v_1) = P_{core}, \ \tau(v_2) = P_{cite}, \ v_1, v_2 \in \mathcal{V}$
6. $v_1 \xrightarrow{about} v_2 \ ; \ \tau(v_1) \in \{P_{core}, P_{ref}, P_{cite}\}, \ \tau(v_2) = TP, \ v_1, v_2 \in \mathcal{V}$

The meaning of each type of vertices is as follows. The *core paper* (P_{core}) is any paper that belongs to either one of the disparate research areas under study. Borrowing the example used at the beginning of this paper, a core paper belongs to either a fish oil or a Raynaud's Syndrome research area. The *cited reference* (P_{ref}) is a paper that is cited by a core paper, whereas the *citing paper* (P_{cite}) refers to a paper that cites a core paper. Both are considered as non-core papers and could be categorized into any research field. The *author* (AU) vertex refers to the author of a core paper or a non-core paper. The *venue* (V) is the publisher of a paper. This is usually the title of a journal or the name of a conference proceeding. The *term* (T) refers to the term appearing in the titles or abstracts of core and non-core papers, excluding general stopwords. Lastly, the *topic* (TP) vertex refers to the *most probable* topic of a paper. This will be learned using a standard topic modeling algorithm.

3.1.1. HBIN metapath features

We define different types of metapaths for our HBIN graphs. A *metapath* M is a path defined on the HBIN network schema $T_G = (\mathcal{A}, \mathcal{R})$. It joins three or more vertices using two or more edges such that $M = v_1 \xrightarrow{e_1} v_2 \xrightarrow{e_2} \ldots \xrightarrow{e_l} v_{l+1}$, where the starting and ending vertices are of the same vertex type P_{core}, $\tau(v_1, v_{l+1}) =$

[2] http://dblp.uni-trier.de/.

[3] http://dl.acm.org/.

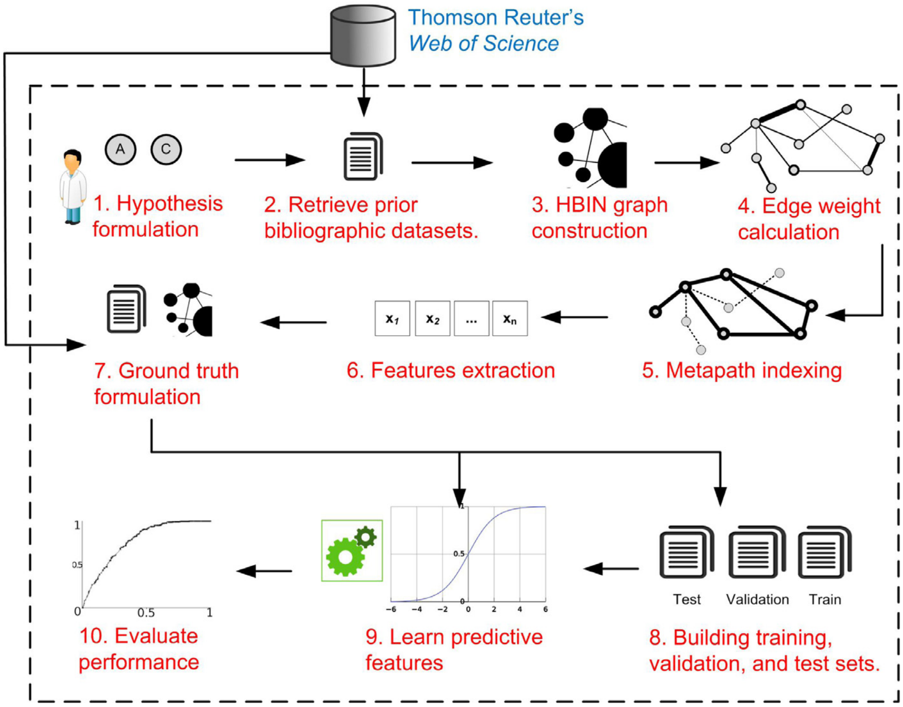

Thomson Reuter's
Web of Science
A
C
1. Hypothesis formulation
2. Retrieve prior bibliographic datasets.
3. HBIN graph construction
4. Edge weight calculation
5. Metapath indexing
x1
x2
...
xn
6. Features extraction
7. Ground truth formulation
Test
Validation
Train
8. Building training, validation, and test sets.
0.5
−6
−4
−2
0
2
4
6
9. Learn predictive features
1
0.5
0
0.5
1
10. Evaluate performance

Fig. 2. Overview of the HBIN-LBD method.

Table 1
Metapaths used in the *HBIN-LBD* model.

No.	Metapath	Description
1.	*two_term*	Core papers share a term
2.	*two_author*	Core papers share an author
3.	*two_venue*	Core papers share a publisher
4.	*two_ref*	Core papers share a reference (ref.)
5.	*three_term_ref*	Core paper shares a term with other core paper's ref
6.	*three_term_cite*	Core paper shares a term with other core paper's citer
7.	*three_author_ref*	Core paper shares an author with other core paper's ref
8.	*three_author_cite*	Core paper shares an author with other core paper's citer
9.	*three_venue_ref*	Core paper shares a publisher with other core paper's ref
10.	*three_venue_cite*	Core paper shares a publisher with other core paper's citer
11.	*four_term_ref*	Core papers' refs. share a term
12.	*four_term_cite*	Core papers' citers share a term
13.	*four_author_ref*	Core papers' refs. share an author
14.	*four_author_cite*	Core papers' citers share an author
15.	*four_venue_ref*	Core papers' refs. share a publisher
16.	*four_venue_cite*	Core papers' citers share a publisher

$P_{core}, P_{core} \in \mathcal{A}$, and $\phi(e_1, \ldots, e_l) \in \mathcal{R}$. Note that the given metapath definition is similar to a definition provided by Sun et al. [47], except for the vertex type constraint on v_1 and v_{l+1}.

Table 1 lists all metapath types defined in our first model. Since metapaths are essentially composite relations of various edge types in HBIN graph, they can capture various relationship meaning between HBIN objects. Each metapath contains rich latent information that can be used for predicting previously unknown relationships between disjoint core papers.

The *degree* of a metapath indicates its length and the distance between two core papers. An n-degree *metapath* is a sequence of n distinct edges. Unless indicated otherwise, we assume that a metapath connects two disjoint core papers p_{core_x}, p_{core_y} that belong to two disjoint sets of research papers P_x and P_y, such that $P_x \cap P_y = \emptyset$. The core papers constitute the endpoints of a metapath. This metapath structure distinguishes our method from Sun et al. [45], where both endpoints of a metapath must be the author vertices. In the following sections, we explain the three main types of metapaths.

(i) ***Two-degree metapath features.*** Two-degree metapaths such as $p_{core_x} \xrightarrow{cites} p_{ref_z} \xleftarrow{cites} p_{core_y}$ suggests a bibliographic coupling between core papers p_{core_x}, p_{core_y}. Bibliographic coupling could suggest that the scientific contributions of the two papers are built upon a common set of background knowledge. This may point to previously unknown connections between them [23].

(ii) ***Three-degree metapath features.*** Three-degree metapaths provide richer semantic interpretations. For example, the following metapath exists between a dietary fish oil core paper [31] (p_{core_x}) and a Raynaud's Syndrome core paper [4] (p_{core_y}): $p_{core_x} \xrightarrow{published\_in} v_z \xleftarrow{published\_in} p_{cite_u} \xleftarrow{cited\_by} p_{core_y}$. According to this metapath, p_{core_x} and p_{cite_u} [5] were published by the same journal *Prostaglandins* (v_z) which indicates their relevance to each other. Also, p_{core_y} is cited by p_{cite_u} which also indicate the relevance between p_{core_y} and p_{cite_u}. As such, p_{core_x} and p_{core_y} may be relevant to each other.

(iii) ***Four-degree metapath features.*** Finally, we also consider four-degree metapaths, such as:

$$p_{core_x} \xrightarrow{cites} p_{ref_u} \xrightarrow{contains} t_z \xleftarrow{contains} p_{ref_w} \xleftarrow{cites} p_{core_y}$$

This metapath suggests that, although p_{core_x} and p_{core_y} cite distinct references p_{ref_u}, p_{ref_w}, their references contain a common term t_z. This metapath configuration may point to the relevance between p_{core_x} and p_{core_y}.

For example, a dietary fish oil core paper [13] (p_{core_x}) cited [21] (p_{ref_u}) and another Raynaud's Syndrome core paper [4] (p_{core_y}) cited [11] (p_{ref_w}). On the other hand, p_{ref_u} and p_{ref_w} shared a common term *prostacyclin* in their titles. This suggests a latent association between core papers p_{core_x} and p_{core_y}. In fact, *prostacyclin* in fish oil can disrupt the platelet aggregation in human blood. It has been demonstrated that this disruption eventually alleviates the symptoms of Raynaud's Syndrome disease [8].

3.1.2. Computing metapath edge weights as features

We compute the appropriate metapath edge weights which is a key element for an effective link prediction [14]. We define the strength of a metapath is a function of the weights of its component edge and propose six scoring schemes to compute these weights. Our method is novel in that it takes into consideration the *local importance* and the *global importance* of a metapath edge. We explain these concepts later.

(i) ***Paper-to-paper edge weight.*** The weight of an edge connecting two vertices p_i and p_j is computed using either one of the following schemes:

$$w\big(p_i \longrightarrow p_j\big) = \frac{1}{\overrightarrow{count}\,\big(p_i, P^i_{ref}\big)} \cdot \frac{1}{N}\left(\sum_{k=1}^{N} \frac{\overrightarrow{count}\,\big(P_k, p_j\big)}{count\,(P_k)}\right); \quad 0 \le w \le 1 \tag{1}$$

$$w(p_i \longleftarrow p_j) = \frac{1}{\overleftarrow{count}\,(p_i, P^i_{cit})} \cdot \frac{1}{N}\left(\sum_{k=1}^{N} \frac{\overleftarrow{count}\,(P_k, p_j)}{count\,(P_k)}\right); \quad 0 \le w \le 1 \tag{2}$$

In the right-hand side of Eq. (1), the first factor of the product refers to the *local importance* of p_j which is inversely proportional to the total number of references cited by p_i which is denoted by $\overrightarrow{count}\,(p_i, P^i_{ref})$. The less references cited by p_i, the more important p_j is to p_i [30].

The second factor of the product approximates the *global importance* of p_j relative to N disjoint sets of paper $P_k ... P_N$. Function $\overrightarrow{count}\,(P_k, p_j)$ denotes the total number of papers in literature set P_k that cite p_j. This is then normalized by dividing it over the total number of papers in P_k. Hence, we assume that the more relevant p_j is to the literature set P_k, the more frequently it will be cited by the component papers in this set. By implication, if p_j is important to two sets P_k and P_{k+1}, then its total global importance score will be higher than another paper that is only relevant to one of the sets. Such reference paper could signal potential connections between two disjoint literature that are not previously known. Eq. (2) follows a similar principle for measuring the importance of paper p_j that cites p_i.

(ii) ***Paper-to-term edge weight.*** Eq. (3) measures the weight of an edge connecting p_i and t_j:

$$w\big(p_i - t_j\big) = \frac{freq\big(p_i, t_j\big)}{freq(T_{p_i})} \cdot \frac{1}{N}\left(\sum_{k=1}^{N} \frac{freq\big(P_k, t_j\big)}{freq\big(T_{P_k}\big)}\right); 0 \le w \le 1 \tag{3}$$

Factor $\frac{freq(p_i, t_j)}{freq(T_{p_i})}$ accounts for the local importance of a term t_j. It measures how frequently t_j appears in paper p_i, which is normalized by the total number of non-unique terms in p_i. We assume that term frequency is positively correlated with its importance. Besides, function $\frac{freq(P_k, t_j)}{freq(T_{P_k})}$ captures the frequency of t_j in all papers in set P_k which is normalized by the total number of non-unique terms in P_k. Altogether, Eq. (3) favors terms that frequently appear in both disjoint literature sets.

(iii) ***Paper-to-author edge weight.*** Eq. (4) measures the importance of author a_j with respect to paper p_i:

$$w\big(p_i - a_j\big) = \frac{1}{count\,(p_i, AU_{p_i})} \cdot \frac{1}{N}\left(\sum_{k=1}^{N} \frac{freq\big(a_j, AU_{P_k}\big)}{freq\big(AU_{P_k}\big)}\right); 0 \le w \le 1 \tag{4}$$

Function $count\,(p_i, AU_{p_i})$ denotes the total number of authors in p_i such that the importance of a_j to p_i is inversely proportional to the total authors of p_i (the *local importance*). Function $freq(a_j, AU_{P_k})$ denotes the total number of non-unique co-authorship pairs between author a_j and the other authors in P_k. Function $freq(AU_{P_k})$ denotes the total number of non-unique co-authorship pairs between authors in P_k. Hence, the more frequently a_j publishes papers together with other authors in P_k, the more important her contribution is to P_k (the *global importance*).

(iv) ***Paper-to-venue edge weight.*** Eq. (5) measures the edge weight connecting publisher v_j and paper p_i:

$$w\big(p_i - v_j\big) = \frac{1}{N}\left(\sum_{k=1}^{N} \frac{count\big(P_k, v_j\big)}{count\,(P_k)}\right); 0 \le w \le 1 \tag{5}$$

Function *count*(P_k, v_j) denotes the total number of papers in P_k that have been published by v_j. Function *count*(P_k) denotes the total number of papers in P_k. We assume the more papers published by v_j in P_k, the more important v_j is to P_k.

3.1.3. Edge weight features aggregation

The overall weight of a single metapath is computed by aggregating the weights of its component edges. Fig. 3 outlines the steps for generating various metapath features for our co-citation link prediction task. Referring to the figure, assume that we have two core papers p_i, p_j such that $\tau(p_i) = \mathcal{A}_1$, $\tau(p_j) = \mathcal{A}_2$, and $\mathcal{A}_1 = \mathcal{A}_2 = P_{core}$. For the sake of generality, let x represent p_i and y represent p_j. Assuming that both core papers are connected by more than one metapaths, the simplest metapath feature score can be calculated by counting the number of unique metapaths between them, i.e. the *path count*. Other features include computing the *sum* of all metapath weights as well as finding the *average*, the *minimum*, and the *maximum* edge weight score of all metapaths between x and y. Note that these aggregation techniques focus on verifying the efficacy of each of the edge weighting schemes described above.

Using these techniques, given a metapath *two_term* connecting x and y, five feature values can be computed: *two_term_pathcount, two_term_sum, two_term_avg, two_term_min*, and *two_term_max*. Since there are 16 different types of metapaths (please refer again to Table 1) and 5 possible aggregation techniques, for each pair of core papers we obtained 80 different features and used these features to train a classification algorithm. We will describe the learning process later in this paper.

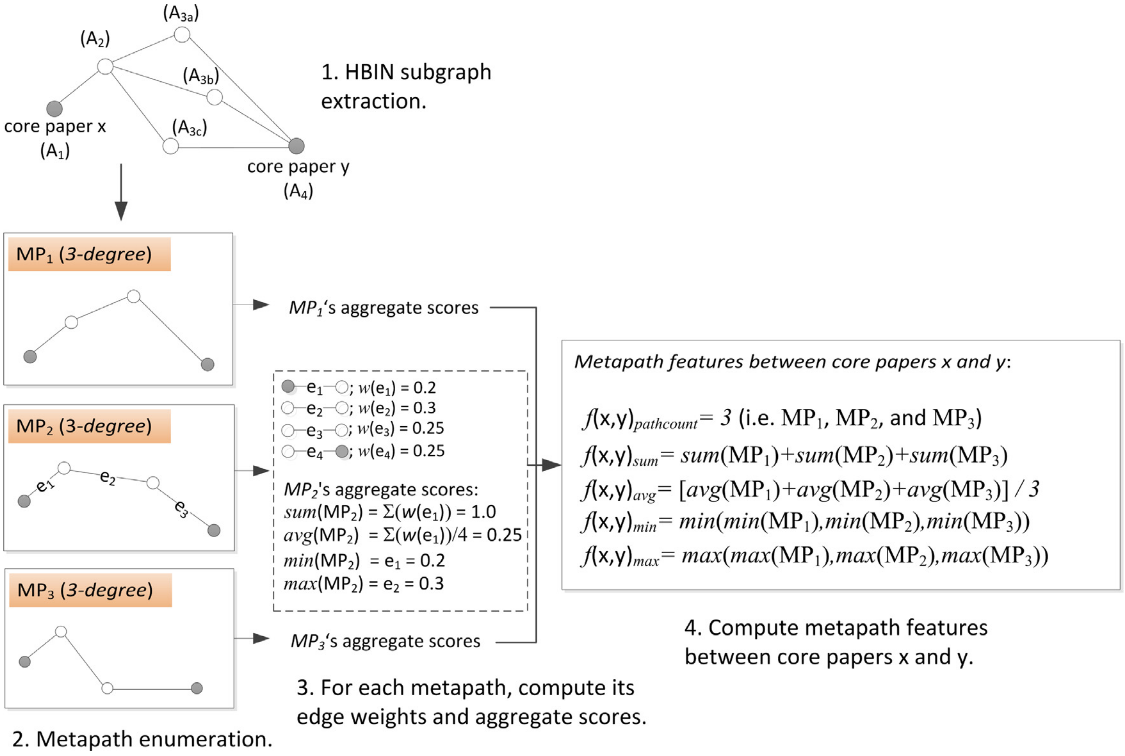

(A3a)
(A2)
(A3b)
core paper x
(A1)
(A3c)
core paper y
(A4)
1. HBIN subgraph extraction.
MP1 (3-degree)
MP1's aggregate scores
MP2 (3-degree)
e1
e2
e3
e1; w(e1) = 0.2
e2; w(e2) = 0.3
e3; w(e3) = 0.25
e4; w(e4) = 0.25
MP2's aggregate scores:
sum(MP2) = Σ(w(e1)) = 1.0
avg(MP2) = Σ(w(e1))/4 = 0.25
min(MP2) = e1 = 0.2
max(MP2) = e2 = 0.3
MP3 (3-degree)
MP3's aggregate scores
Metapath features between core papers x and y:
f(x,y)pathcount= 3 (i.e. MP1, MP2, and MP3)
f(x,y)sum= sum(MP1)+sum(MP2)+sum(MP3)
f(x,y)avg= [avg(MP1)+avg(MP2)+avg(MP3)] / 3
f(x,y)min= min(min(MP1),min(MP2),min(MP3))
f(x,y)max= max(max(MP1),max(MP2),max(MP3))
4. Compute metapath features between core papers x and y.
3. For each metapath, compute its edge weights and aggregate scores.
2. Metapath enumeration.

Fig. 3. Feature extraction for the HBIN-LBD model.

3.1.4. The symmetry of metapaths

Now we provide some theoretical considerations behind our model. It is important for an LBD method to adhere to the symmetric property of literature-based discovery. Using the *ABC model* as an example, this suggests that a hidden connection $A - C$ should be consistently found regardless of whether one starts searching for the connection from the point of view A or from the point of view of C [24,39]. The symmetry property of LBD is captured by the following theorem.

Theorem 1. *The connections A-B-C and C-B-A equally suggest the implicit connection between A and C regardless of the starting and the ending point of the search.*

Our proposed HBIN metapaths fulfil this theorem as follows.

Lemma 1. *Given core paper vertices x and y, it is true that* $f(x,y)_{pathcount} = f(y,x)_{pathcount}$, *where* $f(x,\ y)_{pathcount}$ *is a function that counts the number unique metapaths between* p_{core_x} *and* p_{core_y}.

Recall that a shortest metapath M would join at least three vertices of type $\mathcal{A}$, such that $M = (\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}_3)$. Each component edge in M can be modeled using vector $W(i,\ j)$ between adjacent vertex types A_i and A_j, where $w(a_i, a_j) = 1$ if there is an edge between a_i and a_j. Hence, for the M given above, two adjacency vectors can be defined: $W(\mathcal{A}_1, \mathcal{A}_2)$ and $W(\mathcal{A}_3, \mathcal{A}_2)$.

We can then formulate $L_{1,\ 3}$ as the scalar product of the adjacent vectors $W(\mathcal{A}_1, \mathcal{A}_2)$ and $W(\mathcal{A}_3, \mathcal{A}_2)$. $L_{1,\ 3}$ gives the total count of unique metapaths between $\mathcal{A}_1$ and $\mathcal{A}_3$. From this point, it is easy to see the symmetry of the $f(x,\ y)_{pathcount}$ can be proven as follows.

Proof. $f(x,y)_{pathcount} = L_{x,y} = L_{y,x} = f(y,x)_{pathcount}$, since $L_{x,y} = W(x,:) \cdot W(y,:) = W(y,:) \cdot W(x,:)$, where $\cdot$ means the scalar product of two vectors.

For example, in Fig. 3, we can observe that the total metapath count between $\mathcal{A}_2$ and $\mathcal{A}_4$ is 3, where $W(\mathcal{A}_2, :) = [1,1,1]$, $W(\mathcal{A}_4, :) = [1,1,1]$, and 1 refers to an edge between $\mathcal{A}_2$ or $\mathcal{A}_4$ to each vertex $\mathcal{A}_3 a, \ldots, \mathcal{A}_3 c$.

The commutativity of vectors $W(x,:) \cdot W(y,:) = W(y,:) \cdot W(x,:)$ gives HBIN metapaths their symmetry. It is easy to infer from the above that a metapath eventually consists of a set of adjacent pairwise commutative vectors, such that the symmetry property is applicable to a metapath of any length [12]. Therefore, borrowing from the previous example, it is also true that $f(\mathcal{A}_1, \mathcal{A}_4)_{pathcount} = f(\mathcal{A}_4, \mathcal{A}_1)_{pathcount}$.

Finally, since computing $f(x,\ y)_{sum}$, $f(x,\ y)_{avg}$, $f(x,\ y)_{min}$, and $f(x,\ y)_{max}$ depends on the exact metapath count, the same symmetry property applies to all of these functions. We note that a similar metapath symmetry property has been demonstrated by Sun et al. [47] for the *PathSim* algorithm.

3.1.5. The predictiveness of mixed-degree metapaths

In addition to the symmetry property of metapaths, the inclusion of metapaths of different degrees (or lengths) in the HBIN-LBD model could help increase its overall predictive accuracy. This is supported by the *latent space theory of link prediction* [19]. Hoff et al. proposed a theory which associates every vertex in a social network with a position in a *D*-dimensional latent space. The theory postulates that an edge between two vertices in a network can be predicted if their positions in the latent space are in proximity to each other. Since the vertex latent positions are unobserved, the main research problem is concerned with accurately estimating these positions [19].

Sarkar et al. [38] extended Hoff et al.'s theory to explain why certain link prediction algorithms such as the common neighbours and Adamic/Adar algorithms performed well in many prior empirical studies [28]. They showed that the performance of these algorithms can be explained by their ability to better estimate the positions of two vertices in the same latent space, as previously proposed by Hoff et al. Specifically, they found that the good performance of these algorithms is correlated with their ability to yield

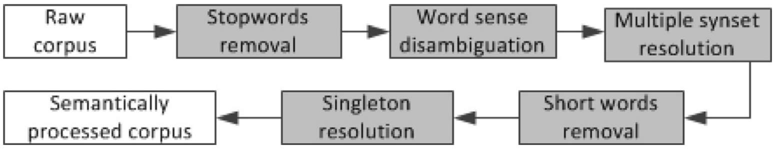

Raw corpus
Stopwords removal
Word sense disambiguation
Multiple synset resolution
Short words removal
Singleton resolution
Semantically processed corpus

Fig. 4. Semantic processing pipeline for the HBIN-LBD-Semantic model.

tighter bounds on the distance between the vertices in the latent space [38].

A further extension of the theory explains the importance of including metapaths of different degrees (2-, 3-, and 4-degree) in the HBIN-LBD model.[38] found that the lack of short paths (i.e. 2-degree paths) between two vertices makes it difficult to yield tight bounds between them in the latent space. We believe that this scenario is typical of literature-based discovery, where two groups of research papers from disparate research areas are expected to share very few common things [49,50,52]. As such, the number of short paths between them would be relatively few. On the other hand, longer paths tend to yield looser bounds between vertices, which in turn adversely affect a link prediction performance. To compensate this, a higher number of longer paths are needed to yield tighter bounds. Unfortunately, including a higher number of longer paths would lead to higher computational costs of an algorithm [3,41].

Sarkar et al. [38] suggested that this problem can be alleviated by including just a few short paths between the vertices. The existence of these short paths could dramatically increase the tightness of bounds yielded by longer paths. Consequently, the following theorem motivates our model:

Theorem 2. *The inclusion of mixed-degree metapaths could increase the overall accuracy of co-citation link prediction between disjoint groups of research papers in literature-based discovery. Having a few two-degree metapaths could increase the tightness of bounds yielded by three- and four-degree metapaths.*

We emphasise that the theorem has been derived by Sarkar et al. [38]. Due to page limit, we refer the reader to the authors' paper for proofs of the theorem.

3.1.6. Algorithm for extracting metapaths

Algorithm 1 shows our algorithm for finding instances of a four-degree metapath between two core papers s and t. Steps 4–6 exclude any bibliographic coupling or existing co-citation between the papers because the goal of four-degree metapaths is to capture distant connections between them.

We implemented the algorithm using Python-based network analysis libraries provided by the Stanford Network Analysis Project[4]. Steps 2–3 and 7–8 were achieved using the *Breadth-First Search* algorithm [41] with total time complexity $O(4(|V| + |E|))$, where $|V|$ and $|E|$ are the number of vertices and edges in G, respectively. The time complexity of steps 10 to 16 is $O(|V_s| \times |V_t| \times |V_{mt}|)$. We emphasize that the current study focuses on the effectiveness of the proposed models instead on their computational efficiency.

3.2. The HBIN-LBD-semantic model

We also introduce a second model, HBIN-LBD-Semantic, which is a variant of the HBIN-LBD model with added semantic components. This model's algorithm consists of a few stages organized as a processing pipeline as shown in Fig. 4.

Algorithm 1 Finding instances of a four-degree metapath between core papers.

```
Input: Weighted HBIN graph $G(V, E)$; a source core paper $s$ and a
    target core paper $t$, $\tau(s,t) \in P_{core}$; a metadata vertex type $mt \in$
    $\{AU, V, T, TP\}$; and a citation vertex type $ct \in \{P_{ref}, P_{cite}\}$.
Output: $M_{s,t}$, i.e. a set of metapaths connecting $s$ and $t$.
1: $M_{s,t} \leftarrow \emptyset$
2: Get vertices $v_{s_i}$ that is directly connected to $s$ in $G$, where $v_{s_i} \in$
   $G \wedge \tau(v_{s_i}) \in ct$. Call this set $V_s$.
3: Get vertices $v_{t_i}$ that is directly connected to $t$ in $G$, where $v_{t_i} \in$
   $G \wedge \tau(v_{t_i}) \in ct$. Call this set $V_t$.
4: if $V_s \cap V_t \neq \emptyset$ then
5:     Remove $V_s \cap V_t$ from $V_t$
6: end if
7: Get metadata vertices $v_{mt_{(s,i)}}$ that is directly connected to each
   vertex in $V_s$, where $v_{mt_{(s,i)}} \in G \wedge \tau(v_{mt_{(s,i)}}) \in mt$. Call this set
   $V_{mt_s}$.
8: Get metadata vertices $v_{mt_{(t,i)}}$ that is directly connected to each
   vertex in $V_t$, where $v_{mt_{(t,i)}} \in G \wedge \tau(v_{mt_{(t,i)}}) \in mt$. Call this set
   $V_{mt_t}$.
9: Build a biadjacency matrix $B = \{V_s, V_t, V_{mt}\}$, such that $B_{ij} \neq \emptyset$ if
   and only if there is a set of common metadata vertices $V_{mt} \in$
   $\{V_{mt_s}, V_{mt_t}\}$ connecting $v_{s_i}$ and $v_{t_j}$.
10: for all $B_{ij}$ in $B$ do
11:     if $B_{ij} \neq \emptyset$ then
12:         for $k \leftarrow 1, |B_{i,j}|$ do
13:             Compute the aggregate score of metapath $s - V_{s_i} -$
   $V_{mt_k} - V_{t_j} - t$ and add it into $M_{s,t}$.
14:         end for
15:     end if
16: end for
```

HBIN-LBD-Semantic applies a dictionary-based semantic processing algorithm on the title and/or abstract text of papers. This is done in order to standardize the terminological representation of different words that have a similar meaning. Basic text preprocessing such as stopwords removal and stemming are common in LBD [56], but they cannot resolve more complex linguistic problems such as word sense disambiguation [44].

HBIN-LBD-Semantic uses *WordNet*[5] as the source for dictionary terms. Unlike other LBD methods, we choose WordNet instead of the other domain-dependent ontologies such as the *Unified Medical Language Systems* (UMLS)[6] in order to ensure that our model is easily applicable to mining literature in various research domains.

3.2.1. Word sense disambiguation

The first stage of our algorithm extracts sentences from the titles and/or abstracts of papers in the raw corpus using a regex-based sentence tokenization tool *sent_tokenize*.[7] From each sentence, we identified the *n-gram* noun phrases (NP) using NLTK[8]. We then remove stopwords and lemmatize the extracted NPs and

[4] https://snap.stanford.edu/snappy/index.html.

[5] https://wordnet.princeton.edu/.

[6] https://www.nlm.nih.gov/research/umls/.

[7] http://www.nltk.org/.

[8] https://gist.github.com/alexbowe/879414#file-nltk-intro-py.

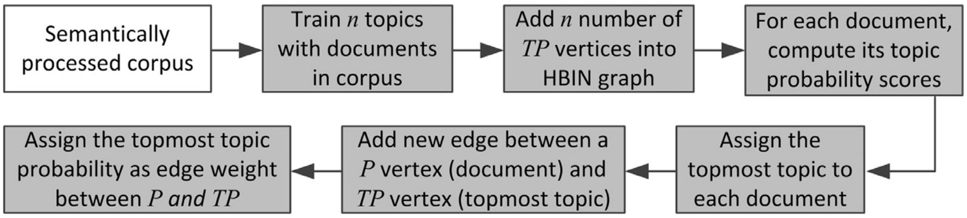

Semantically processed corpus
Train n topics with documents in corpus
Add n number of TP vertices into HBIN graph
For each document, compute its topic probability scores
Assign the topmost topic to each document
Add new edge between a P vertex (document) and TP vertex (topmost topic)
Assign the topmost topic probability as edge weight between P and TP

Fig. 5. Topic modeling components of the HBIN-LBD-Topic model.

map each NP to its standard dictionary term in WordNet, known as *synsets*. A synset represents words or lemmas that have a similar meaning. To map a single-term NP to its appropriate synset, we first search for a synset that directly maps to the phrase. If the NP maps to more than one candidate synsets, we then apply the *Adapted Lesk Algorithm*[9] to determine the most appropriate synset [1]. For instance, synset *dog.n.01* can be used to replace lemma names that have similar meaning or senses such as '*dog*', '*domestic dog*', and '*Canis familiaris*'.

3.2.2. Multiple synsets resolution

One limitation of using the Adapted Lesk Algorithm is that it is possible for an NP to be mapped to more than one synsets. As such, in order to more precisely disambiguate the sense of an NP, we select the synset that has the same morphological stem as the target phrase. If more than one synsets have the matching stem, we subsequently select the synset that has the highest frequency of occurrence in our corpus. This is akin to the *dominant synset* approach [29], except that we compute the synset's occurrence frequency in the target corpus instead of in WordNet. If more than one qualifying candidate synsets remain, we then look for the synset that has the target NP in its WordNet *gloss* or definition [1].

The final resolution step is to compute the *average path similarity* score of each candidate synset given a set of randomly selected synsets from the existing sense disambiguation results. This score is computed based on the average length of the shortest path between a candidate synset and each random synset in the WordNet semantic tree [34]. The candidate synset that has the highest average path similarity score is selected.

3.2.3. Short words removal and singleton resolution

This stage removes the single-character words such as *t* or *v* from the previous synset mapping results. Further, we also resolve *singletons*, i.e. synsets that occur only once in the entire target text corpus. Their presence may suggest possible mapping errors. Alternatively, they may point to instances of words that are too rare to be useful in our LBD task. We substitute these singletons with randomly picked hypernyms or hyponyms for retaining the singletons' information without introducing unnecessary noises into the corpus. If neither hypernym nor hyponym is found, the singleton is removed.

3.3. The HBIN-LBD-topic model

We explore a third LBD model, HBIN-LBD-Topic. This model applies the *Latent Dirichlet Allocation* (LDA) topic modeling algorithm[10] [6] in order to construct and incorporate new *topic* vertices (*TP*) in the HBIN graphs. The processing pipeline for generating the new *TP* vertices is shown in Fig. 5.

We evaluated the topic modeling quality given n number of topics; n={1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}. For each number of topics, we compute its *average topic coherence* score. Topic coherence score indicates topic modeling quality in relation to human judgment. The nearer the score to 0.00, the better quality the generated topics are [33]. In our case, we found the coherence scores consistently deteriorated for $n > 10$, hence we chose $n = 10$ as suitable number of topics.

Given n topics, we add n new *TP* vertices into an HBIN graph and link a *P* vertex to its topmost topic. The edge weight score $w(P \longrightarrow TP)$ between these nodes is assigned with the topmost topic's probability score. Using this model, we obtain five additional metapaths as shown in Table 2.

Table 2
HBIN-LBD-Topic metapaths.

No.	Metapath	Description
17.	*two_topic*	Core papers share a topic
18.	*three_topic_ref*	Core paper shares a topic with other core paper's ref.
19.	*three_topic_cite*	Core paper shares a topic with other core paper's citer
20.	*four_topic_ref*	Core papers' refs. share a topic
21.	*four_term_cite*	Core papers' citers share a topic

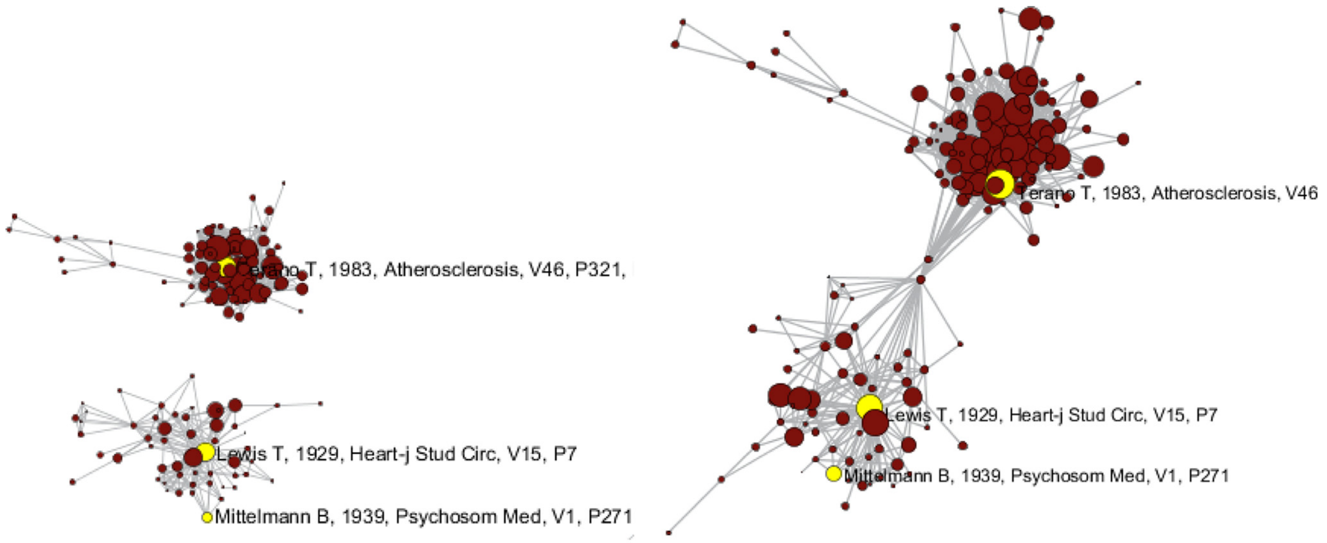
4. Experiments and results

We demonstrate the effectiveness of our models in replicating two classic literature-based discovery instances: (i) the previously unknown relationships between *Dietary Fish Oil* and *Raynaud's Syndrome* (DFORS) hypothesis [48], and (2) the neglected connections in *Migraine* and *Magnesium* (MM) hypothesis [51]. They form the evaluation ground truths for our models. For example, the following Fig. 6 shows previously disconnected clusters of fish oil and Raynaud's Syndrome papers became connected via new co-citation links following the publication of Swanson's hypothesis in 1986. Our goal is to predict the occurrence of such new links based on the available bibliographic data prior to 1986.

The DFORS and MM datasets are used in our evaluation because they have become the commonly accepted evaluation ground truths for many prior LBD methods [57]. Defining other ground truths for LBD evaluation has been challenging and often results in contentious debates among LBD researchers [25,35,42]. It is also difficult to directly use other bibliographic datasets such as DBLP [37] or ACM Digital Library [30] because no well-defined instance of a scientific discovery has been proposed for these datasets.

Using curated knowledge bases such as the CTD database [10] does not immediately address our current evaluation goal. Our goal is to test the performance of our method in predicting a future discovery based on a pre-discovery bibliographic dataset. This dataset is defined based on a specific cut-off publication date. It would take much effort to define similar publication cut-off dates for voluminous gene-disease associations in CTD database. Further, the definition of 'discovery' in relation to these associations are not well-defined [10]. To do that requires substantial effort by domain experts which is outside the scope of this paper.

[9] https://github.com/alvations/pywsd.
[10] https://radimrehurek.com/gensim/.


Terano T, 1983, Atherosclerosis, V46, P321,
Lewis T, 1929, Heart-j Stud Circ, V15, P7
Mittelmann B, 1939, Psychosom Med, V1, P271
Terano T, 1983, Atherosclerosis, V46
Lewis T, 1929, Heart-j Stud Circ, V15, P7
Mittelmann B, 1939, Psychosom Med, V1, P271

(a) Before Swanson's discovery (1900-1986).

(b) After Swanson's discovery in 1986.

Fig. 6. The evolution of DFORS clusters in time-sliced co-citation networks before and after the publication of Swanson's hypothesis. The upper cluster represents a group of fish oil papers, while the lower cluster is a group of Raynaud's Syndrome papers. Both networks are visualized using the Sci^{2}Tool.[11], based on a bibliographic dataset retrieved from the Thomson Reuter's Web of Science.

4.1. Datasets

We retrieved the bibliographic records of DFORS papers from the *Web of Science* (WoS)[12] using the same search keywords originally used by Swanson [48]. Each record in this dataset includes a paper' title, author(s), and publisher as well as the lists of its cited references and citing papers. We restricted our retrieval to papers published between year 1900 and 1985 prior to the publication of Swanson's hypothesis in 1986. Further, we limited the query results to 485 records with abstracts, consisting of 352 fish oil and 133 Raynaud's Syndrome core papers.

For the second hypothesis, we also used the original search keywords [51] to retrieve the MM bibliographic records and restricted the query to papers published prior to 1988. We obtained 43,075 migraine and magnesium records. These records include 20,043 abstracts retrieved from Pubmed using *Biopython*[13] and the rest are simply the paper titles. We then built HBIN graphs from the previously retrieved datasets. The HBIN graph for DFORS dataset has 25,176 vertices and 108,544 edges. The MM graph consists of 1,095,566 vertices and 9,010,218 edges.

4.1.1. Learning sets preparation

Each instance in our learning sets represents a unique pair of core papers. We defined three classes and labeled each instance based on the presence or absence of co-citation link between the core paper pair *following* the publication of the target hypotheses. Class '+1' refers to a pair of previously disconnected core papers from disparate clusters (*inter*-cluster) which became co-cited following the publication of the hypothesis. Class '-1' refers to a pair of previously disconnected core papers from within the same cluster (*intra*-cluster). These papers also subsequently became co-cited. Class '0' refers to a pair of core papers (either *inter* or *intra*-cluster) which never became co-cited. We obtained 2055 instances from the DFORS graph, with 685 instances in each class. From the MM graph we obtained 6366 instances, with 2122 instances in each class.

4.2. Experimental settings

4.2.1. Learning sets partitioning

We learned our HBIN models using the Weka's implementation of the *Support Vector Machine* (SVM) algorithm [15]. The hyperparameters of the learner were optimized using *Auto-Weka* [55] based on cross-validation (CV) sets as well as various training, validation and testing % splits. Note that these splits were hermetically sealed from each other and were obtained with independent random sampling.

4.3. Performance benchmarking

The performance of the existing LBD methods are usually compared by how well they rank certain target intermediate terms connecting the disjoint sets of literatures [57]. The higher and the more precise the ranking, the better the performance. Our goal is to predict future links between the papers instead of producing a list of ranked terms. As such, it is more appropriate to evaluate their performance using a link prediction-oriented evaluation method. Specifically, we examined the performance of our models in predicting different types of future co-citation links as indicated by the class labels.

For current experiments, we compared the performance of our models against three popular link prediction algorithms: *Common Neighbours* (CN), *Adamic/Adar* (AA), and *LDA topic similarity score* (LDA) [28]. We chose CN and AA due to their good performance for many social network link prediction tasks [14]. We chose LDA to find out whether topic similarity alone could account for the emergence of co-citation links between papers. In addition, we took the *ZeroR* classifier [15] as the lowest baseline. ZeroR classifies all instances in the learning set by simply choosing the majority class label.

We evaluated eight variants of our models. Two variants are derived from the HBIN-LBD model: (**i**) **HBIN-T** and (**ii**) **HBIN-TA**. These models are similar except HBIN-T includes the terms in the

[11] Sci2 Team. (2009). Science of Science (Sci2) Tool. Indiana University and SciTech Strategies.

[12] http://thomsonreuters.com/thomson-reuters-web-of-science/.

[13] http://biopython.org/DIST/docs/api/Bio.Entrez-module.html.

Table 3
Comparisons of the overall performance of HBIN models against baselines on DFORS and MM datasets (10-fold CV) in predicting instances of all classes.

Methods	DFORS dataset				MM dataset			
	Acc.%	*F1*	*P*	*R*	*Acc.%*	*F1*	*P*	*R*
HBIN-T	87.74	0.88	0.88	0.88	78.48	0.78	0.79	0.79
HBIN-TA	87.35	0.87	0.87	0.87	78.34	0.78	0.78	0.78
HBIN-T-TOPIC	87.50	0.87	0.87	0.88	77.56	0.77	0.78	0.78
HBIN-TA-TOPIC	**88.86**	**0.89**	**0.89**	**0.89**	77.66	0.78	0.78	0.78
SHBIN-T	84.87	0.85	0.85	0.85	75.40	0.75	0.75	0.75
SHBIN-TA	87.59	0.88	0.88	0.88	77.25	0.77	0.77	0.77
SHBIN-T-TOPIC	84.96	0.85	0.85	0.85	76.74	0.77	0.77	0.77
SHBIN-TA-TOPIC	88.22	0.88	0.88	0.88	**79.99**	**0.80**	**0.80**	**0.80**
AA	*60.44*	*0.61*	*0.61*	*0.60*	*49.89*	*0.46*	*0.50*	*0.50*
CN	*61.41*	*0.61*	*0.63*	*0.61*	*48.04*	*0.45*	*0.48*	*0.48*
LDA	*62.48*	*0.53*	*0.61*	*0.63*	*45.54*	*0.35*	*0.33*	*0.46*
ZeroR	*33.09*	*0.27*	*0.22*	*0.33*	*33.30*	*0.24*	*0.22*	*0.33*

Table 4
Comparisons of the overall performance of HBIN models against baselines on DFORS and MM datasets (70:15:15% split) in predicting instances of all classes.

Methods	DFORS dataset				MM dataset			
	Acc.%	*F1*	*P*	*R*	*Acc.%*	*F1*	*P*	*R*
HBIN-T	80.91	0.81	0.81	0.81	78.01	0.78	0.78	0.78
HBIN-TA	86.04	0.86	0.86	0.86	77.28	0.77	0.78	0.77
HBIN-T-TOPIC	82.85	0.83	0.83	0.83	78.01	0.78	0.78	0.78
HBIN-TA-TOPIC	**88.03**	**0.88**	**0.88**	**0.88**	77.28	0.78	0.77	0.77
SHBIN-T	81.55	0.82	0.82	0.82	74.76	0.75	0.75	0.75
SHBIN-TA	82.20	0.82	0.82	0.82	76.02	0.76	0.76	0.76
SHBIN-T-TOPIC	82.52	0.83	0.83	0.83	77.17	0.77	0.77	0.77
SHBIN-TA-TOPIC	82.20	0.82	0.83	0.82	**78.85**	**0.79**	**0.79**	**0.79**
AA	*58.58*	*0.59*	*0.59*	*0.59*	*50.26*	*0.46*	*0.49*	*0.50*
CN	*59.22*	*0.60*	*0.61*	*0.59*	*50.26*	*0.41*	*0.36*	*0.50*
LDA	*61.81*	*0.54*	*0.58*	*0.62*	*43.56*	*0.33*	*0.42*	*0.44*
ZeroR	*33.01*	*0.16*	*0.12*	*0.33*	*31.09*	*0.15*	*0.09*	*0.31*

Table 5
Feature ranking for DFORS and MM datasets.

Rank	HBIN-TA-TOPIC		SHBIN-TA-TOPIC	
	Features	*Avg. merit*	*Features*	*Avg. merit*
1	three_term_cite_max	0.611	four_venue_ref_max	0.410
2	four_topic_ref_sum	0.571	four_term_ref_pathcount	0.347
3	four_topic_ref_pathcount	0.555	four_topic_ref_max	0.368
4	four_topic_ref_max	0.448	four_term_ref_sum	0.341
5	four_term_ref_sum	0.443	four_term_ref_max	0.335
...	...	...	...	...
101	two_venue_pathcount	0.003	two_topic_max	0.004
102	two_venue_avg	0.001	two_topic_min	0.003
103	two_venue_sum	0.001	two_topic_pathcount	0.002
104	two_venue_min	0.001	two_topic_sum	0.001
105	two_venue_max	0.001	two_topic_avg	0.001

core and non-core papers' titles whereas HBIN-TA incorporates additional terms from the core papers' abstracts. Two other variants are derived from the HBIN-LBD-Semantic model: (**iii**) the **SHBIN-T** semantically process the terms in HBIN-T, whereas (**iv**) **SHBIN-TA** semantically processes the terms in HBIN-TA. Finally, four variants are derived based on the HBIN-LBD-Topic model. The (**v**) **HBIN-T-TOPIC** and (**vi**) **HBIN-TA-TOPIC** models train and incorporate topic nodes based on the terms in the HBIN-T and HBIN-TA models, respectively. Finally, the (**vii**) **SHBIN-T-TOPIC** and (**viii**) **SHBIN-TA-TOPIC** models train and incorporate topic nodes based on semantically-processed terms in SHBIN-T and SHBIN-TA models, respectively.

4.4. Results

We present our experimental results in this section.

4.4.1. Overall performance

The overall performance of our models on both DFORS and MM datasets are shown in Table 3.

The HBIN-TA-TOPIC model performed the best with 88.86% accuracy (F1: 0.89) with 10-fold CV on the DFORS dataset. The SHBIN-TA-TOPIC model performed the best on the MM dataset with 79.99% accuracy (F1: 0.80). Both models outperformed all baselines. This suggests the efficacies of our method. Table 4 shows similar performance of the models based on the 70:15:15 percentage split validation.

These overall performance results suggest that our models are useful for predicting previously unknown relationships between papers from disparate research areas. The good overall accuracies suggest that the models can effectively predict both the absence or presence of future co-citation links between research papers. They are able to discriminate between pairs of core papers that

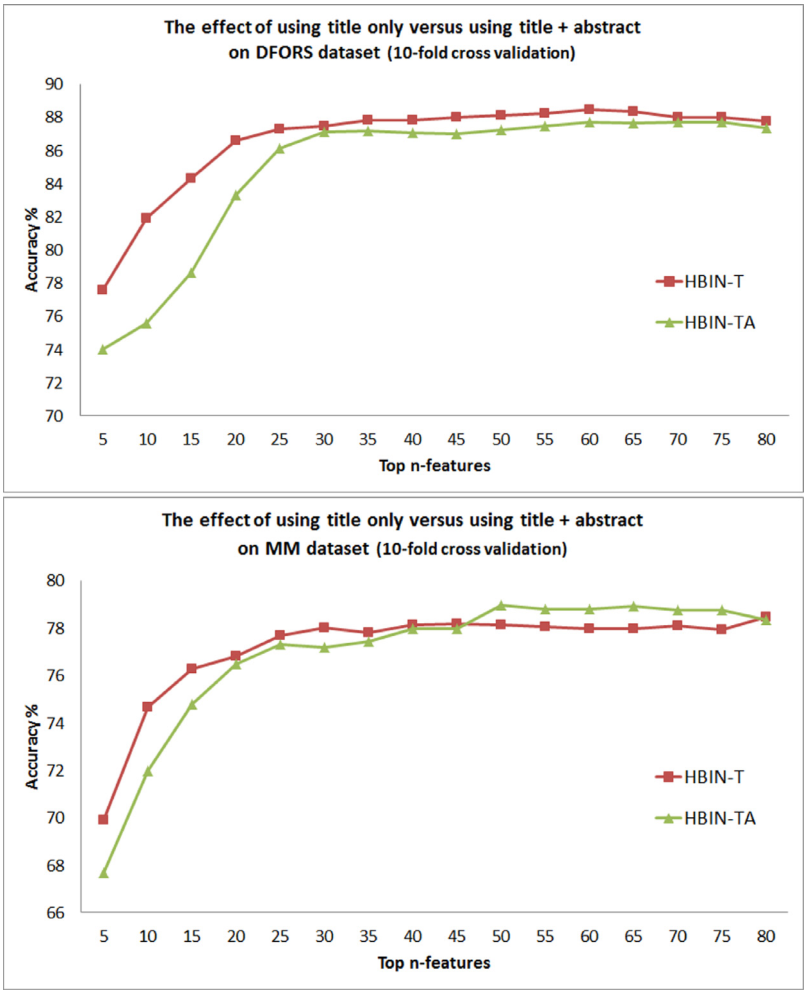


Fig. 7. Performance difference of using title text only vs. using title *and* abstract text of the HBIN models built for the DFORS and MM datasets.

are likely to form future inter-cluster co-citation links (instances of class '+1') and those that will only form intra-cluster links (instances of class '-1'). This could help surmise the hidden connections between papers in two disjoint research fields.

4.4.2. Feature rank

Table 5 summarizes the performance of different metapath features in the HBIN-TA-TOPIC model on the DFORS dataset. It also shows the performance of the features in the SHBIN-TA-TOPIC model on MM dataset. The performance are measured using *InfoGain* merit scores.

The result suggests *three_term_cite_max* as the best performing feature of the HBIN-TA-TOPIC model on DFORS dataset. In contrast, features that involve the sharing of publishers between the cited references of core papers' ($four\_venue\_ref_{max}$) gave the most performance contribution for the SHBIN-TA-TOPIC model. Both results suggest the good performance of features that combine both lexical and non-lexical information in the HBIN graphs.

4.4.3. The influence of semantic processing and topic model

Fig. 7 shows the performance of models that use terms from titles and abstracts of research papers compared to those that use terms from the titles only and without the abstracts. The result suggests incorporating more texts from the abstracts did not lead to more superior models. We note that both HBIN-T and HBIN-TA do not involve the proposed semantic and topic modeling components. The purpose of this comparison is to solely observe the performance effects from using more texts.

Next, Fig. 8 compares the performance of different models to determine the extent to which incorporating the semantic and topic modeling components influences the performance of our method. The results suggest that, on both DFORS and MM datasets, better performance can be achieved by the HBIN-LBD models without requiring the semantic components of the HBIN-LBD-Semantic models. Our finding stands in contrast with most of the NLP-based LBD methods which benefit from the incorporating semantic components into their models [20].

Fig. 9 shows that models that incorporated topic nodes outperformed the other models with up to 6% performance gain compared to models which do not incorporate topic modeling components. This result suggests that the latent topic modeling is a promising approach to enhancing the accuracy of an LBD method. Observe that the better performance is generally achieved when the semantic components and the topic modeling were used together.

5. Discussions

The results from our experiments suggest the efficacy of the HBIN-LBD method for performing LBD tasks. This is demonstrated by its good performance in predicting future co-citation links between previously disjoint papers in two real historical scientific discoveries. The good results underline the usefulness bibli-

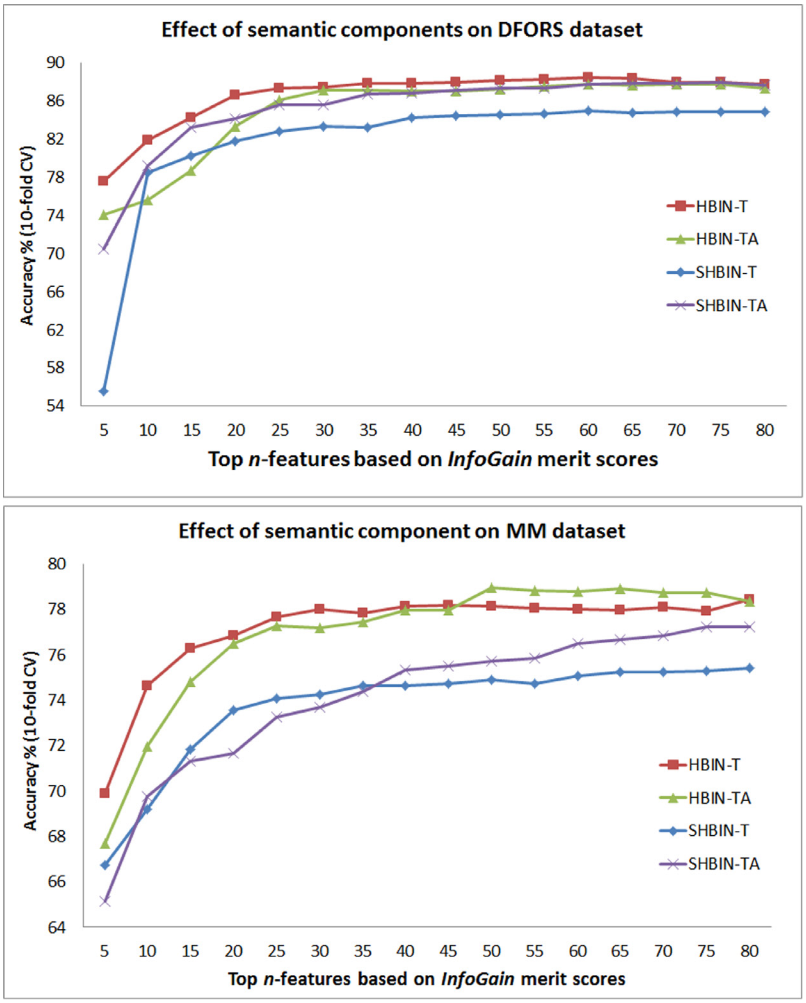


Fig. 8. Effects of incorporating the semantic components of HBIN-LBD-Semantic models on the DFORS and MM datasets.

ographic metadata for literature-based discovery. The good performance of such metapath features as *three_term_cite_max* and *four_venue_ref_max* suggests that combinations of lexical and citation features in the HBIN graphs form good link predictors between disjoint research papers.

To our knowledge, our method is the first that demonstrates an effective way of using bibliographic metadata for performing literature-based discovery. As mentioned previously, this is the first main contribution of this paper. HBIN-LBD uses simple statistics that are easy to compute from HBIN graphs without the need for sophisticated and domain-specific NLP tools and ontologies. As a result, our method can be easily extended to mining literatures in various research domains. This overcomes the limitations of many existing LBD methods whose applications are largely limited to mining biomedical literatures.

As the second research contribution, we have demonstrated the effectiveness of a hybrid approach that employs lexical and non-lexical information objects through metapaths. The result is supported by prior research. For example, hybrid information retrieval methods have been shown to mitigate the performance trade-off normally suffered by the exclusive use of either text-based or citation-based features [7,16]. Similarly, Bassecoulard and Zitt [2] found that a combination of lexical and non-lexical features could overcome the imbalance between precision and recall in many information retrieval systems. This could explain why our metapath features have yielded good accuracies.

Our methodology may complement the existing methods in technology roadmapping (TRM) [54,58]. TRM predicts the future changes in technology topics by analyzing the existing research literature. It aims at generating useful insights to help in strategic R&D planning. Our innovation comes in the form of using the interlinking of various bibliographic metadata objects in order to predict the future convergence between previously independent research fields. As such, in addition to the existing term-oriented TRM methods [58] and citation-oriented TRM methods [54], our method provides an alternative vantage point that may help better understand the future technological topic evolutions.

There are several limitations of our work. We have not applied author name disambiguation techniques during the construction of HBIN graphs, resulting in the possible duplication of author names. It will be useful to explore some existing name disambiguation techniques for addressing this limitation. Since the current focus of our work is on the effectiveness of the proposed models, we have not fully addressed their algorithmic efficiency and scalability to very large graphs. This is also another limitation of this paper.

6. Conclusion and future work

In this paper, we have presented a novel literature-based discovery method that exploits the latent information retrieved from heterogeneous bibliographic information networks or HBIN. Our results show that, with the help of word sense disambiguation

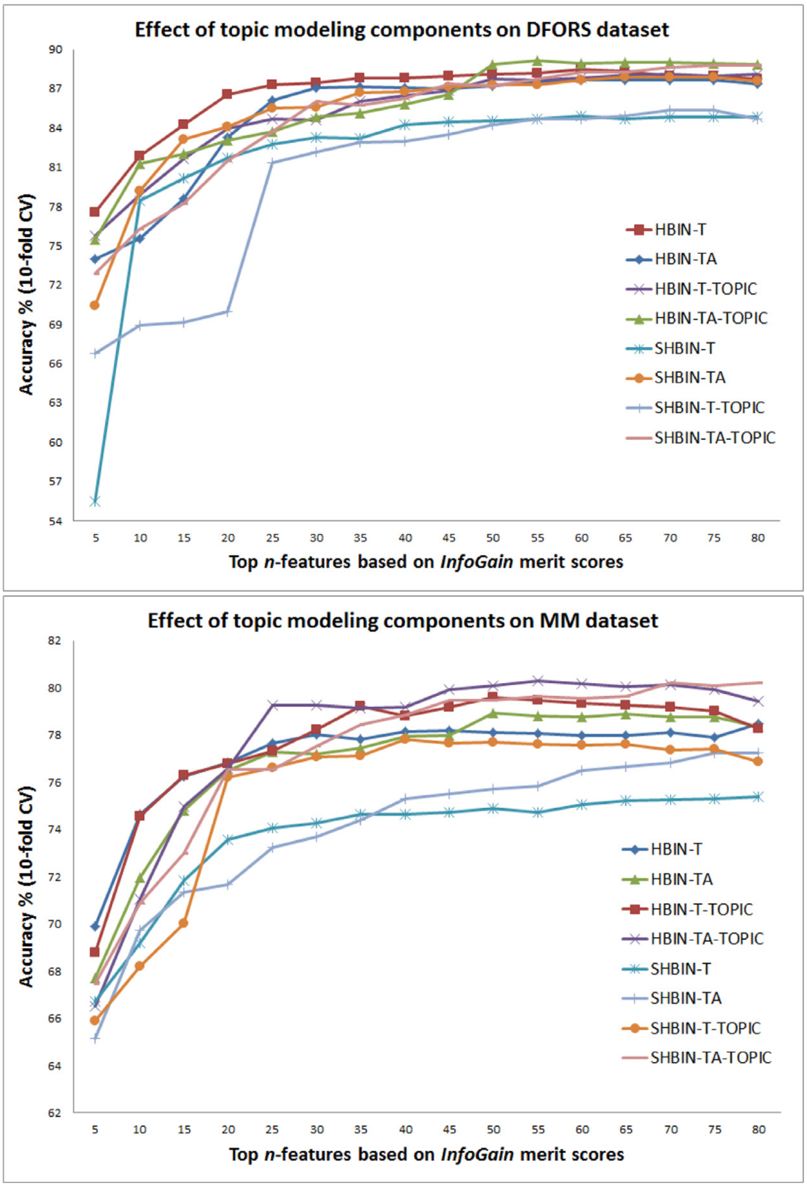


Fig. 9. Effects of incorporating the topic modeling components of HBIN-LBD-Topic models on the DFORS and MM datasets.

and topic modeling components, the combined use of lexical and non-lexical information between various bibliographic metadata in HBIN graphs yields good performance in predicting novel associations between previously disconnected research papers. Experiments showed our models outperforming other baseline link prediction algorithms in predicting the future co-citation links between research papers in the Fish Oil and Raynaud's Syndrome literatures as well as in the Migraine and Magnesium literatures.

For future work, we plan to explore an efficient graph algorithm that scales well against large HBIN graphs. We are interested in applying our models to mining the literature in other domains such as climatology [32]. Further studies of the contribution of different metapath features in our models may also shed further light on the process behind the formation of new co-citation link between disjoint research fields. Finally, future work can be directed at exploring new metapath structures that would harness other types of lexico-citation information in HBIN graphs.

References

[1] S. Banerjee, T. Pedersen, An adapted lesk algorithm for word sense disambiguation using wordnet, in: Computational linguistics and intelligent text processing, Springer, 2002, pp. 136–145.

[2] E. Bassecoulard, M. Zitt, Patents and Publications, Springer Netherlands, 2005, pp. 665–694.

[3] S. Beamer, K. Asanović, D. Patterson, Direction-optimizing breadth-first search, Sci. Program. 21 (3–4) (2013) 137–148.

[4] J. Belch, J. Drury, H. Capell, C. Forbes, P. Newman, F. Mckenzie, P. Leiberman, C. Prentice, Intermittent epoprostenol (prostacyclin) infusion in patients with raynaud's syndrome: a double-blind controlled trial, Lancet 321 (8320) (1983) 313–315.

[5] J. Belch, I. Greer, M. McLaren, A. Saniabadi, S. Miller, R. Sturrock, C. Forbes, The effects of itnravenous zk36-374, a stsable prostacyclin analogue, on normal volunteers, Prostaglandins 28 (1) (1984) 67–77.

[6] D. Blei, A. Ng, M. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.

[7] S. Brin, L. Page, The anatomy of a large-scale hypertextual web search engine, Comput. Netw. ISDN Syst. 30 (1–7) (1998) 107–117.

[8] D. Cameron, R. Kavuluru, T. Rindflesch, A. Sheth, K. Thirunarayan, O. Bodenreider, Context-driven automatic subgraph creation for literature-based discovery, J. Biomed. Inform. 54 (2015) 141–157.

[9] C. Chen, Predictive effects of structural variation on citation counts, J. Am. Soc. Inf. Sci. Technol. 63 (3) (2012) 431–449.

[10] Y. Ding, M. Song, J. Han, Q. Yu, E. Yan, L. Lin, T. Chambers, Entitymetrics: measuring the impact of entities, PLoS ONE 8 (8) (2013) e71416.

[11] P. Dowd, M. Martin, E. Cooke, S. Bowcock, R. Jones, P. Dieppe, J. Kirby, Treatment of raynaud's phenomenon by intravenous infusion of prostacyclin (pgi2), Br. J. Dermatol. 106 (1) (1982) 81–89.

[12] M. Drazin, Some generalizations of matrix commutativity, Proc. London Math. Soc. 3 (1) (1951) 222–231.

[13] J. Dyerberg, H. Bang, E. Stoffersen, S. Moncada, J. Vane, Eicosapentaenoic acid and prevention of thrombosis and atherosclerosis? Lancet 312 (8081) (1978) 117–119.

[14] R. Guns, Link Prediction, Springer International Publishing, 2014, pp. 35–55.

[15] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. Witten, The weka data mining software: an update, ACM SIGKDD Explorations Newsletter 11 (1) (2009) 10–18.

[16] M. Hamedani, S.-W. Kim, D.-J. Kim, Simcc: a novel method to consider both content and citations for computing similarity of scientific papers, Inf. Sci. 334 (2016) 273–292.

[17] J. Han, Mining heterogeneous information networks by exploring the power of links, in: International Conference on Discovery Science, Springer, 2009, pp. 13–30.

[18] M.A. Hasan, V. Chaoji, S. Salem, M. Zaki, Link prediction using supervised learning, SDM06: Workshop on Link Analysis, Counter-terrorism and Security, 2006.

[19] P.D. Hoff, A.E. Raftery, M.S. Handcock, Latent space approaches to social network analysis, J. Am. Stat. Assoc. 97 (460) (2002) 1090–1098.

[20] D. Hristovski, C. Friedman, T. Rindflesch, B. Peterlin, Exploiting semantic relations for literature-based discovery, in: AMIA Annual Symposium, 2006, American Medical Informatics Association, 2006, pp. 349–353.

[21] R. Johnson, D. Morton, J. Kinner, R. Gorman, J. McGuire, F. Sun, N. Whittaker, S. Bunting, J. Salmon, S. Moncada, The chemical structure of prostaglandin x (prostacyclin), Prostaglandins 12 (6) (1976) 915–928.

[22] A. Kastrin, T.C. Rindflesch, D. Hristovski, Link prediction on a network of co-occurring mesh terms: towards literature-based discovery, Methods Inf. Med. 55 (4) (2016) 340–346.

[23] M. Kessler, Bibliographic coupling between scientific papers, Am. Doc. 14 (1) (1963) 10–25.

[24] A. Koike, Biomedical Application of Knowledge Discovery, in: P. Bruza, M. Weeber (Eds.), Literature-based Discovery, Springer Berlin Heidelberg, 2008, pp. 173–192.

[25] R. Kostoff, Letter to the editor: validating discovery in literature-based discovery, J. Biomed. Inform. 40 (4) (2007) 448–450.

[26] R. Kostoff, Literature-related discovery: common factors for parkinsons disease and crohns disease, Scientometrics 100 (3) (2014) 623–657.

[27] R.N. Kostoff, J. Block, J. Solka, M. Briggs, R. Rushenberg, J. Stump, D. Johnson, T. Lyons, J. Wyatt, Literature-related discovery, Ann. Rev. Inf. Sci. Technol. 43 (1) (2009) 1–71.

[28] D. Liben-Nowell, J. Kleinberg, The link-prediction problem for social networks, J. Am. Soc. Inf. Sci. Technol. 58 (7) (2007) 1019–1031.

[29] S. Liu, F. Liu, C. Yu, W. Meng, An effective approach to document retrieval via utilizing wordnet and recognizing phrases, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2004, pp. 266–272.

[30] X. Liu, Y. Yu, C. Guo, Y. Sun, L. Gao, Full-text based context-rich heteregenous network mining approach for citation recommendation, in: Proceedings of the Digital Libraries 2014, ACM/IEEE, 2014.

[31] W. Lockette, R. Webb, B. Culp, B. Pitt, Vascular reactivity and high dietary eicopentaenoic acid, Prostaglandins 24 (5) (1982) 631–639.

[32] E. Marsi, P. Ozturk, E. Aamot, G. Sizov, M. Ardelan, Towards text mining in climate science: Extraction of quantitative variables and their relations, in: Proceedings of the Ninth International Conference on Language Resources and Evaluation, European Language Resources Association, 2014.

[33] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2011, pp. 262–272.

[34] T. Pedersen, S. Patwardhan, J. Michelizzi, Wordnet: Similarity: measuring the relatedness of concepts, in: Demonstration papers at HLT-NAACL 2004, Association for Computational Linguistics, 2004, pp. 38–41.

[35] W. Pratt, M. Yetisgen-Yildiz, Reply: response to validating discovery in literature-based discovery, J. Biomed. Inf. 40 (4) (2007) 450–452.

[36] J. Preiss, M. Stevenson, The effect of word sense disambiguation accuracy on literature based discovery, in: Proceedings of the ACM Ninth International Workshop on Data and Text Mining in Biomedical Informatics, ACM, 2015, p. 1.

[37] X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, J. Han, Cluscite: effective citation recommendation by information network-based clustering, in: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, 2014, pp. 821–830.

[38] P. Sarkar, D. Chakrabarti, A.W. Moore, Theoretical justification of popular link prediction heuristics., in: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, 22, 2011, p. 2722.

[39] A.K. Sehgal, X.Y. Qiu, P. Srinivasan, Analyzing Lbd Methods Using a General Framework, in: P. Bruza, M. Weeber (Eds.), Literature-based Discovery, Springer Berlin Heidelberg, 2008, pp. 75–100.

[40] N. Shibata, Y. Kajikawa, I. Sakata, Link prediction in citation networks, Journal of the American Society for Information Science and Technology 63 (1) (2012) 78–85.

[41] S.S. Skiena, The Algorithm Design Manual: Text, 1, Springer Science & Business Media, 1998.

[42] N. Smalheiser, Literature-based discovery: beyond the abcs, J. Am. Soc. Inf. Sci. Technol. 63 (2) (2012) 218–224.

[43] H. Small, Maps of science as interdisciplinary discourse: co-citation contexts and the role of analogy, Scientometrics 83 (3) (2010) 835–849.

[44] M. Stevenson, Y. Wilks, Word sense disambiguation, Oxford Handbook Comp. Ling. (2003) 249–265.

[45] Y. Sun, R. Barber, M. Gupta, C.C. Aggarwal, J. Han, Co-author relationship prediction in heterogeneous bibliographic networks, in: 2011 International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2011, pp. 121–128.

[46] Y. Sun, J. Han, Mining Heterogeneous Information Networks: Principles and Methodologies, Synthesis Lectures on Data Mining and Knowledge Discovery, Morgan and Claypool, 2012.

[47] Y. Sun, J. Han, X. Yan, P.S. Yu, T. Wu, Pathsim: meta path-based top-k similarity search in heterogeneous information networks, Proc. VLDB Endowment 4 (11) (2011) 992–1003.

[48] D. Swanson, Fish oil, raynaud's syndrome, and undiscovered public knowledge, Perspect. Biol. Med. 30 (1) (1986) 7–18.

[49] D. Swanson, Undiscovered public knowledge, Libr. Q. 56 (2) (1986) 103–118.

[50] D. Swanson, Two medical literatures that are logically but not bibliographically connected, J. Am. Soc. Inf. Sci. 38 (4) (1987) 228–233.

[51] D. Swanson, Migraine and magnesium: eleven neglected connections, Perspect. Biol. Med. 31 (4) (1988) 526–557.

[52] D. Swanson, Asist award of merit acceptance speech: on the fragmentation of knowledge, the connection explosion, and assembling other people's ideas, Bull. Am. Soc. Inf. Sci. Technol. 27 (3) (2001) 12–14.

[53] D. Swanson, Literature-based discovery? The very idea, Springer Series in Information Science and Knowledge Management, Springer-Verlag Berlin Heidelberg, Berlin, pp. 3–11.

[54] Y. Takano, C. Mejia, Y. Kajikawa, Unconnected component inclusion technique for patent network analysis: case study of internet of things-related technologies, J. Informetr. 10 (4) (2016) 967–980.

[55] C. Thornton, F. Hutter, H. Hoos, K. Leyton-Brown, Auto-weka: combined selection and hyperparameter optimization of classification algorithms, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 847–855.

[56] V.I. Torvik, N. Smalheiser, A quantitative model for linking two disparate sets of articles in medline, Bioinformatics 23 (13) (2007) 1658–1665.

[57] M. Yetisgen-Yildiz, W. Pratt, A new evaluation methodology for literature-based discovery systems, J. Biomed. Inform. 42 (4) (2009) 633–643.

[58] Y. Zhang, G. Zhang, H. Chen, A.L. Porter, D. Zhu, J. Lu, Topic analysis and forecasting for science, technology and innovation: methodology with a case study focusing on big data research, Technol. Forecast Soc. Change 105 (2016) 179–191.