Mobile Networks and Applications (2019) 24:69–79
https://doi.org/10.1007/s11036-018-1133-9

Newly Published Scientific Papers Recommendation in Heterogeneous Information Networks

Xiao Ma[1] · Yin Zhang[1] · Jiangfeng Zeng[2]

Published online: 25 September 2018

Abstract
Millions of new research papers are published each year, making it extremely difficult for researchers to find out what they really want. Existing paper recommendation algorithms cannot effectively address the recommendation of *newly published papers* due to lack of historical information(e.g., citation information; view log), the so-called cold start problem. Furthermore, in most of these studies, papers are considered in homogeneous or bipartite networks. However, in a real bibliographic network, there are multiple types of objects (e.g., researchers, papers, venues, topics) and multiple types of links among these objects. In this paper, we study the problem of new paper recommendation in the heterogeneous bibliographic network, and a new method called HIPRec, i.e., meta-graph based recommendation model, is proposed to solve this problem. First, the top-K most interesting meta-paths are selected based on the training data. Secondly, a greedy method is proposed to select the most significant meta-graphs generated by merging the meta-paths, which can describe more sophisticated semantics between researchers and papers than simple meta-paths. In the meantime, meta-path and meta-graph based topological features are systematically extracted from the network. Lastly, a supervised model is used to learn the best weights associated with different topological features in deciding the researcher-new paper recommendations. We present experiments on a real bibliographic network, the DBLP network, which show the effectiveness of our approach compared to state-of-the-art new paper recommendation methods.

Keywords Recommender systems · Newly published research papers · Heterogeneous information networks · Meta-path · Meta-graph

1 Introduction

With the development of information science and technology, great achievements have been made in terms of electronic literatures. A huge number of research papers are published each year, making it difficult for researchers to find out what they really care about. As we all know that, newly published research papers are extremely important for researchers, which help them keep up with the latest progresses in their research fields, find new inspirations and acquire latest theories and techniques. Thus a system designed to aid researchers to quickly find relevant new publications is in high demand.

To tackle this problem, some information retrieval techniques (e.g., Google Scholar and Microsoft Academic) allow users to search publications based on keywords and properties associated with the target papers, e.g., authors, time of publication, etc. They also provide relevant papers by measuring the content similarity between papers. Although these systems make it easier for researchers to find interesting papers, keywords based systems still return thousands or millions of relevant papers. For example, Google Scholar returns more than 22,000 newly published research papers since 2018 with the query "paper recommendation". Instead of going through a large number of papers produced by matching certain query keyword, research paper recommender systems are designed to

✉ Xiao Ma
cindyma@zuel.edu.cn

Yin Zhang
yinzhang@zuel.edu.cn

Jiangfeng Zeng
jfzeng@hust.edu.cn

[1] School of Information and Safety Engineering, Zhongnan University of Economics and Law, Wuhan, China

[2] Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, China

suggest helpful papers to researchers via exploring their interests and preferences [1].

Traditional paper recommender systems are dedicated to solving the information overload problem in big academic data. Most of the existing methods focus on recommending relevant papers by analyzing the citation-ships between papers. Some typical works includes graph-based methods [2–5] and collaborative filtering based methods [6, 9, 10]. However, these methods favor the papers with higher citations. Thus newly published papers with few or no citations will seldom be discovered, which is the so-called cold-start problem in recommender systems [11].

Content-based recommendation approaches have been proved to be effective in alleviating the cold start problem [7, 30]. Some content similarity based paper recommendation methods have been proposed [9, 15, 16, 31]. Generally, these approaches measure the textual relevance between query user's publications and other research papers by computing the document similarity or the topic similarity. However, solely relying on these information are not sufficient for new paper recommendation. Although the number of newly published papers is tremendous, research topics are comparably limited. Hundreds of new papers could share a same topic, which makes textual similarity a very weak evidence in terms of new paper recommendation. In addition, many critical features closely related to new paper recommendation cannot be represented by textual similarity.

Intuitively, researchers prefer to only review a relatively small number of new publications closely related to their research topics, published in their interested journals/venues, authored by their followed researchers, etc. For example, if a new paper is published in a venue in which the query user has publications before and mentions the same research topic, the probability that the query researcher gets interested in this paper is higher than the probabilities of other papers. This phenomenon is illustrated in Fig. 1. Suppose node a_1 represents a query researcher, he/she can find the new paper p_3 along the meta-paths $a_1 - p_1 - v_1 - p_3$, $a_1 - p_1 - t_1 - p_3$ and the meta-graph $a_1 - p_1 - < t_1, v_1 > - p_3$ (The formal definitions of meta-paths and meta-graphs will be introduced in Def. 2 and Def. 3). However, researcher a_1 can find new papers p_4 only along the meta-path $a_1 - p_1 - v_1 - p_4$. Thus the probability that researcher a_1 gets interested in new paper p_3 is higher than the probability of new paper p_4. Therefore, it is intuitive to mine all the meaningful meta-path and meta-graph patterns to explain how researcher a_1 find new papers in the heterogeneous bibliographic graph.

Recently, some meta-path based approaches are proposed to solve the paper recommendation problem by employing meta-paths of various semantics in heterogeneous networks [8, 13]. However, these methods suffer from two shortcomings when tackling the new paper recommendation problem.

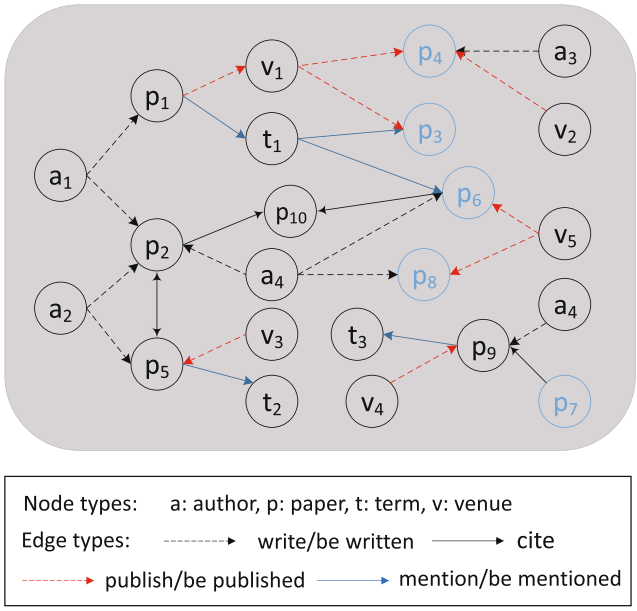

Fig. 1 A toy example of the heterogeneous bibliographic network, where the nodes in blue color denotes the set of newly published research papers which may be recommended in the future

Firstly, the meta-paths used in the existing works are simple manually created, and fail to systematically extract all the interesting topological features which describe how researchers find newly published interesting papers. Secondly, existing approaches mainly focus on the citation recommendation problem, which compute the authority of papers by analyzing the quantity of paper citations. In our case, these methods cannot boost the performance since new papers have few citations.

To tackle the afore-mentioned issues, we solve the new paper recommendation problem in the context of heterogeneous bibliographic networks aiming to simulate how researchers discover interesting new papers in the real world, as well as to help them find highly relevant new publications effectively and efficiently. The contributions of this paper are summarized as follows:

(1) We study the problem of newly published paper recommendation in the heterogeneous bibliographic networks, and a new methodology named HIPRec is developed to solve this problem; (2) We systematically extract the meta-path and meta-graph based heterogeneous topological features which are incorporated into a prediction model to measure the probability that a query researcher gets interested in the new papers. (3) Experiments on the real DBLP bibliographic network show that by considering both the meta-path and meta-graph features, the prediction accuracy can be significantly improved.

The rest of this paper is organized as follows. Section 2 gives an introduction of the related work on citation recommendation and new paper recommendation. The

background and preliminaries are introduced in Section 3. Our proposed HIPRec method is detailed in Section 4. Section 5 includes our experimental results and analysis. Section 6 concludes this study with future work.

2 Related work

2.1 Citation recommendation

A line similar to our problem is the citation recommendation task. Citation recommendation aims at recommending relevant papers to researchers for citing/referencing via the citation-ships between papers. It prefers to recommend the previously published papers with quite a considerable number of citations, but turns a blind eye to the newly published papers with no/few citations.

Huang et al. [19] propose a graph-based recommendation framework combining a content-based approach and a collaborative approach to recommend citations for query users. Strohman et al. [2] considers an unpublished manuscript as a query to a search system and use the paper's text content in the search procedure. Then they rely on the text of previous research paper and its citation graph to find relevant related papers. Yu et al. [8] introduce a citation recommendation method based on the meta-path methodology [12] and define citation probability within the scope of meta path-based feature space. Yang et al. [10] propose a joint multi-relational model to jointly recommend coauthors, citations, venues for query users. Xia et al. [14] propose a novel recommendation method, which incorporates common author relations between articles to help generate better recommendations for relevant target researchers. Zhao et al. [27] consider the knowledge gap between a researcher's background knowledge and research target in the construction of paper recommendation system. Anand et al. [29] perform random walk on the citation graph of papers to balancing the both relevance and diversity while search for research papers. Hassan [28] uses the Recurrent Neural Networks to discover continuous and latent semantic features of the papers and propose a personalized research paper recommendation based on users' feedbacks.

2.2 Newly published papers recommendation

To our best knowledge, there are few works dealing with the new paper recommendation problem because of the cold start issue. New papers have no/few historical information makes the very famous collaborative filtering method less efficient. Ha et al. [20] transform the new paper recommendation problem into a citation recommendation problem, they propose a graphical model based on the citation-ships between a new paper and its referenced papers. Wang and Blei [9] introduce a hybrid model which combines matrix factorization and topic model to recommend both old papers and new papers. However, as the cold problem of new papers, they can just rely on the topic model to recommend new papers. Sugiyama and Kan [15] recommend new papers by relying on the contents of new papers. However, it is usually nontrivial to get the contents of publications. Cai et al. [17] propose a simple graph model to recommend new papers by incorporating various valuable information. However, they consider the new papers in a bipartite graph and neglect the heterogeneous information of the new papers.

Table 1 List of notations

Notation	Description
$\mathcal{A}, \mathcal{R}$	Types of entities and relations
$\mathcal{MP}, \mathcal{MG}$	The sets of meta-paths and meta-graphs
Π, Γ	A meta-path and meta-graph
mp, mg	A meta-path and meta-graph instance
$\hat{\beta}$	Coefficient weights associated with features
P_n	The set of newly published papers
Λ	The set of training examples

3 Background and preliminaries

In this section, we briefly introduce some concepts related to heterogeneous information network and the new paper recommendation problem.

A Heterogeneous Information Network (HIN) is a directed graph, which contains multiple types of entities and links. In order to study meta-path and meta-graph based feature space and discuss new paper recommendation model, we first introduce the definitions of HIN schema, meta-paths and meta-graphs [23, 25]. The notations used in definitions as well as the rest part of the paper can be found in Table 1.

Definition 1 (HIN schema) [23] The HIN schema is a meta template of heterogeneous network $G = (V, E)$ with an object type mapping function $\phi : V \rightarrow \mathcal{A}$ and the link mapping $\varphi : E \rightarrow \mathcal{R}$, which is a directed graph defined over object types $\mathcal{A}$, with edges as relations from $\mathcal{R}$, denoted as $T_G = (\mathcal{A}, \mathcal{R})$.

The HIN schema describes all the available link types between object types. Figure 2 is an example of the HIN schema with respect to the DBLP bibliographic network.

Definition 2 (Meta-paths) [23] A meta-path Π is a path defined on an HIN schema $T_G = (\mathcal{A}, \mathcal{R})$, and is denoted

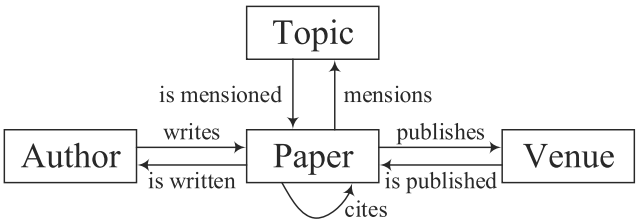

Fig. 2 DBLP network schema

in the form of $\Pi : A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_L} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \circ \ldots \circ R_l$ between type A_1 and A_{l+1}, where $\circ$ represents the composition operator on relations.

Based on the HIN schema in Fig. 2, a meta path $A - P - T - P_n$ as shown in Fig. 3 describes that the new papers (object P_n) share the same research topic with the authors' previous publications (object P), where A, P, T and P_n represent the authors, previous publications, topics and newly published papers, respectively. An instance of this meta-path in Fig. 1 is $a_1 \rightarrow p_1 \rightarrow t_1 \rightarrow p_6$.

Definition 3 (Meta-graphs) [25] A meta-graph Γ is a directed acyclic graph with a single source node n_s (i.e., with in-degree 0) and a single sink (target) node n_t (i.e., with out-degree 0), defined on an HIN schema $T_G = (\mathcal{A}, \mathcal{R})$. Formally, $\Gamma = \{N, M, n_s, n_t\}$, where N is a set of nodes and M is a set of edges. For any node $x \in N$, $x \in \mathcal{A}$; for any link $(x, y) \in M$, $(x, y) \in \mathcal{R}$.

An example meta-graph Γ is shown in Fig. 3. It can be seen that Γ is a directed acyclic graph with source node $n_s = A$ and target node $n_t = P_n$. According to Definition 3, an instance of this meta-graph in Fig. 1 is $a_1 - p_1 - < t_1, v_1 > - p_3$.

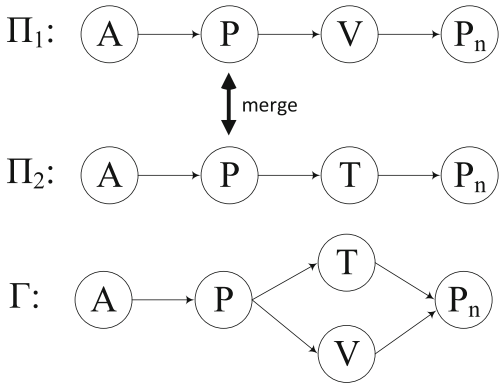

Fig. 3 Meta-path and meta-graph

4 The approach

4.1 Framework description

To summarize our approach HIPRec, we present the overall framework in Fig. 4. It consists of three main parts:

(1) The input of HIPRec is the heterogeneous bibliographic network including different types of nodes and links. Firstly, the DBLP heterogeneous information network schema is defined. Then the training and testing example pairs are extracted for discovering the significant meta-paths and meta-graphs.
(2) Given the DBLP schema, we first enumerate all the meta-paths with a length constraint L by traversing the DBLP schema, and select the top-K most interesting meta-paths $\mathcal{MP} = \{\Pi_1, \ldots, \Pi_k\}$ by employing a standard supervised learning method in Section 4.3. Then, a GreedyGraph algorithm is proposed to construct the significant meta-graphs $\mathcal{MG} = \{\Gamma_1, \Gamma_2, \ldots\}$ based on the proposed heuristic in Section 4.4. After that, all the meta-paths and meta-graphs based topological features are extracted, and we train a logistic regression model to learn the coefficients of all features in Section 4.5.
(3) Given the heterogeneous information network, the query researchers and the candidate new papers, we first extract the feature vector for each pair of query researcher and new paper, and then compute the relationship probabilities $\hat{p}$. By ranking the probabilities, our method generates a ranked list of papers as the recommendation to the query researcher.

4.2 Proximity measures

In order to calculate the relevance between two HIN objects, we introduce some proximity measures for meta-paths and meta-graphs in Sections 4.2.1 and 4.2.2 respectively.

4.2.1 Proximity measures for meta-paths

- PathCount (PC) [12] The simplest form of proximity between two nodes in a HIN is the count of the different paths which satisfy a given meta-path Π. The intuition behind this measure is that the larger the number of certain meta-path, the closer the entities are.
- Path Constrained Random Walk (PCRW) [18] For a meta-path Π, the proximity between two entities x and y is defined by the random walk starting from x and following only paths satisfying Π. PCRW indicates the probability that a walker constrained on a particular meta-path reaches the target node, and weighs the paths based on the neighborhoods of nodes along them.

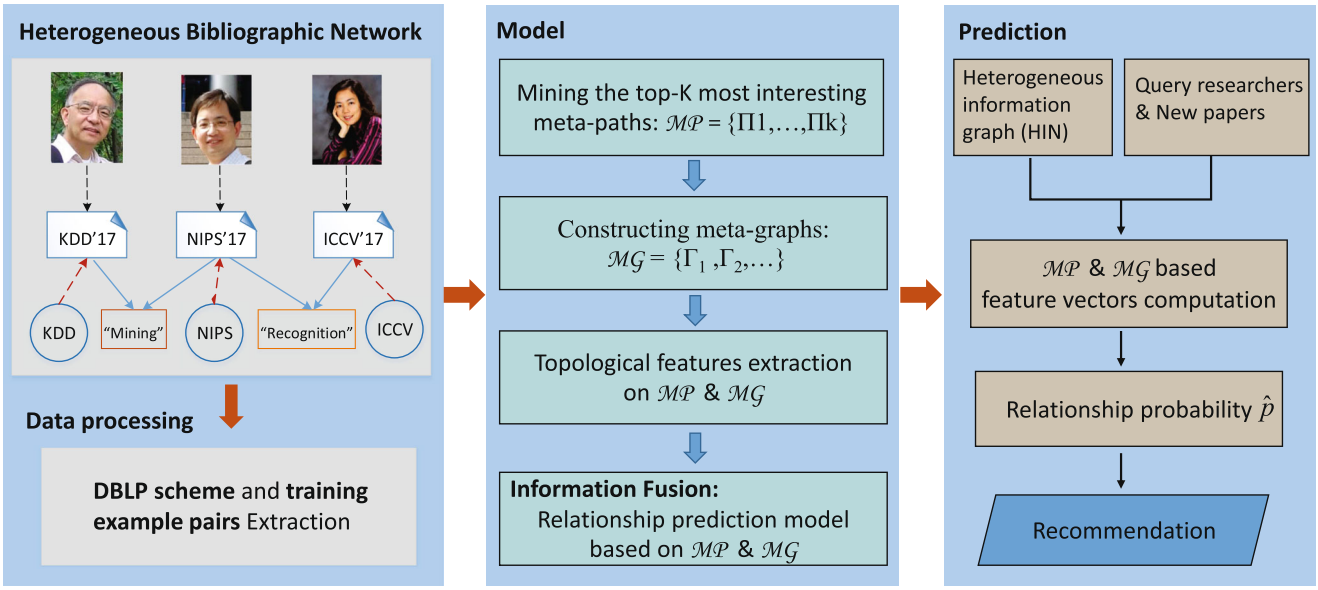

Fig. 4 Overall framework of the HIPRec approach

– Biased Path Constrained Random Walk (BPCRW) [24] BPCRW is a generalized version of PC and PCRW methods with a parameter α balancing the number of meta-paths to be counted and the contribution of neighbors.

4.2.2 Proximity measures for meta-graphs

– StructCount (SC) [25] Given a HIN schema, a meta-graph Γ, a source node x and a target node y, the proximity of x and y is defined as the number of instances of Γ.
– Structure Constrained Subgraph Expansion (SCSE) [25] Given a HIN schema, a meta-graph Γ, a source node x and a target node y, the proximity of x and y is defined as the probability that an initial subgraph of HIN (i.e., node x) would expand to an instance of Γ covering y.
– Biased Structure Constrained Subgraph Expansion (BSCSE) [25] Similarly, BSCSE is a unified version of SC and SCSE methods with a parameter α balancing the results of these two proximities.

Table 2 Proximities of different measures

	PC	PCRW	BPCRW
Π_1	1	$\frac{1}{6}$	$\frac{1}{\sqrt{6}}$
Π_2	1	$\frac{1}{6}$	$\frac{1}{\sqrt{6}}$
	SC	SCSE	BSCSE
Γ	1	$\frac{1}{2}$	$\frac{1}{\sqrt{2}}$

Take node pair (a_1, p_3) for an example. The proximities of (a_1, p_3) with respect to Π_1, Π_2 and Γ in Fig. 3 are presented in Table 2. α is set to 0.5.

4.3 Candidate meta-paths extraction

In the meta-path based citation recommendation works [8, 13], meta-paths are manually created. In this paper, we introduce a principled way to mine the top-K meaningful meta-paths from the training and test data, as well as a set of positive and negative example pairs Λ.[1]

Meta paths between two object types (e.g., A and P_n) can be obtained by traversing on the DBLP network schema as shown in Fig. 2. As the network schema is a much smaller graph compared to the original network with only 4 nodes and 7 links, it is very fast to enumerate all the meta-paths with a length constraint of L (measured by link hops). We use MP to denote the set of meta-path generated in this section. According to the results of [23], long meta-paths are not very meaningful. In the meantime, feature extractions for long meta-path patterns will be much more expensive than short ones.

Top-K interesting meta-paths With example pairs Λ, the meta-paths generated above and the proximity measures for meta-paths introduced in Section 4.2.1, the meta-path

[1] Positive example pairs denote that researchers are interested in the new papers, and vice versa. For example, (a_1, p_4) and (a_2, p_8) can be regarded as two positive example pairs, and (a_1, p_7) can be regarded as a negative example pair as shown in Fig. 1.

based feature space F_{MP} can be represented as follows: a Cartesian product of the two sets: $F_{MP} = MP \times \Phi_{MP}$, where MP is the set of possible meta-paths, Φ_{MP} is the set of meta-path based proximity measures. Then the top-K interesting meta-paths can be obtained by learning the importance of each feature in deciding the relationship building in DBLP [12]. We use $\mathcal{MP}$ to denote the selected meta-paths in the following sections.

4.4 Greedy meta-graphs construction algorithm

Huang et al. [25] introduce how to compute the proximity of two objects along a meta-graph in the heterogeneous networks, however, how to generate meaningful meta-graphs between two objects is not clear. Although subgraph mining approaches can be used to mine frequent subgraph patterns in large graphs [21, 22], it is not an efficient way to solve this problem. The reasons can be explained in two aspects.

Firstly, subgraph mining in a large graph is a NP-hard problem, that is to say, it is expensive to discover the frequent subgraphs even using some approximation strategies [22]. Furthermore, the subgraph patterns in our problem have notable features, thus there is no need to mine all the other subgraphs which violate this rule. For example, as shown in Fig. 3, all the meta-paths and meta-graphs required are those starting with the "A" type and ending with the "P_n" type. Motivated by this phenomenon, we propose a greedy meta-graph construction algorithm to generate the meta-graph Γ by simply merging meta-paths Π_1 and Π_2. Before introducing the algorithm, we first propose a heuristic on meta-graph construction.

Given the training example pairs Λ, meta-paths $\mathcal{MP}$, meta-graphs $\mathcal{MG}$, and the proximity measures Φ_{MP}, Φ_{MG}, the priorityScores S_P and S_G, which measure the capacity of meta-path and meta-graph patterns in fitting the positive training examples, are defined in Eq. 1 and 2.

$$S_P = \frac{\sum_{f \in \Phi_{MP}} \sum_{(a_i, p_{n_i})^+} f(a_i, p_{n_i}|\Pi)}{\sum_{f \in \Phi_{MP}} \sum_{(a_i, p_{n_i})} f(a_i, p_{n_i}|\Pi)} \tag{1}$$

$$S_G = \frac{\sum_{f \in \Phi_{MG}} \sum_{(a_i, p_{n_i})^+} f(a_i, p_{n_i}|\Gamma)}{\sum_{f \in \Phi_{MG}} \sum_{(a_i, p_{n_i})} f(a_i, p_{n_i}|\Gamma)} \tag{2}$$

where $(a_i, p_{n_i})^+$ denotes an positive example pair. $\Pi \in \mathcal{MP}$ represents a meta-path, and $\Gamma (\Gamma \in \mathcal{MG})$ represents a meta-graph generated by merging the meta-paths from $\mathcal{MP}$.

Definition 4 (Significant Meta-graph) Given a set of positive and negative training example pairs Λ, meta-paths $\Pi_1, \Pi_2, \ldots, \Pi_k$, and a meta-graph Γ which is a synthesis of the meta-paths. If the priorityScores $S_G(\Gamma)$ is no less than any of the priorityScore $S_P(\Pi_x)$, $x = 1, 2, \ldots, k$, the meta-graph Γ is considered as an significant meta-graph.

The intuition of Definition 4 is that, if the priorityScore of meta-graph is larger than the priorityScores of its components, i.e.,the meta-paths, we say that this meta-graph pattern is more powerful than simple meta-paths and will be used in the new paper recommendation model.

An example Suppose the positive examples are (a_1, p_3), (a_1, p_4), and the negative examples are (a_1, p_6), (a_1, p_8) as shown in Fig. 1. Meta-paths are Π_1 and Π_2, and Meta-graph is Γ as shown in Fig. 3. According to Eq. 1 and 2, $\mathcal{S}_P(\Pi_1) = 1$, $\mathcal{S}_P(\Pi_2) = 2/3$, $S_G(\Gamma) = 1$, which is no less than $\mathcal{S}_P(\Pi_1)$ and $\mathcal{S}_P(\Pi_2)$. Thus Γ is an significant meta-graph which has important semantic meanings according to Definition 4.

Meta-graphs construction Relying on the candidate meta-paths $\mathcal{MP}$ generated in Section 4.3, we define a k-path significant meta-graph which contains k different simple meta-paths. Each k-path meta-graph is generated through progressively merging k different meta-paths, and the constructed meta-graph is denoted by δ_i^k. Specifically, the simple meta-path can be represented as δ_i^1. For example, as shown in Fig. 3, meta-paths Π_1 and Π_2 can be denoted by δ_1^1 and δ_2^1. Meta-graph Γ can be represented as δ_1^2. It can be seen that we apply a merging operation to extract the 2-path meta-graph using the 1-path meta-graphs (i.e., the meta-paths). Given two meta-paths Π_1 and Π_2, we first find the common entity types of them, and then we can link them by merging the common entity types.

It is worth mentioning that the constructed meta-graphs must satisfy the Definition 3. Since there may be plenty of meta-graphs generated with respect to different merging operation, we propose a pruning strategy as shown in Definition 5 to discard the meaningless meta-graphs that will not be considered further. Thus we can recursively find the k-path meta-graphs δ_m^k by adding one meta-path a time, that is, $\delta_m^k = Merge(\delta_i^{k-1}, \delta_j^1, o_n)$. o_n represents the n-th way of the merging operation. If δ_m^k is a significant meta-graph, then we add this δ_m^k into the meta-graph set $\mathcal{MG}$.

Algorithm 1 GreedyGraph mining algorithm

Input: A HIN G; Candidate meta-paths $\mathcal{MP}$; Example pairs Λ.
Output: Significant meta-graphs $\mathcal{MG}$.

$\mathcal{MG} = \phi$
$\mathcal{MG} \leftarrow \mathcal{MP}$
Compute the S_P values for all the meta-paths $\mathcal{MP}$
$k = 2$
while $\delta_i^{k-1} \neq \phi$ **do**
  for $\delta_i^{k-1} \in \mathcal{MG}$ **do**
    for $\delta_j^1 \in \mathcal{MG}$ and $j > MaxIndex(\delta_i^{k-1})$ **do**
      Find all the common object type pairs between
      δ_i^{k-1} and δ_j^1
      Count the total number of ways to merge δ_i^{k-1}
      and δ_j^1, namely O_N
      for $o_n = 1$ to O_N **do**
        $\delta_m^k = Merge(\delta_i^{k-1}, \delta_j^1, o_n)$
        if δ_m^k satisfy Definition 3 **then**
          Compute $S_G(\delta_m^k)$ by Equation 2
          if $S_G(\delta_m^k)$ is no less than the S_P of any
            of its components **then**
            $\mathcal{MG} = \mathcal{MG} \cup \delta_m^k$
          end if
        end if
      end for
    end for
  end for
  $k = k + 1$
end while
return $\mathcal{MG}$

During the recursion procedure we must avoid searching the same $(k-1)$-path meta-graphs repeatedly. Therefore, we add the paths in the path number order and use *MaxIndex* to represent the maximum path number of the paths that are incorporated in meta-graph δ_m^k. The pseudo code of the meta-graphs construction algorithm is shown in Algorithm 1. Note that, the different proximities of the example pairs Λ on meta-paths $\mathcal{MP}$ have been calculated in Section 4.3, thus it is easy to compute the S_P of all the meta-paths. The recursion terminates when there are no significant $(k-1)$-path meta-graphs.

Definition 5 (Pruning strategy) If a k-path meta-graph δ_i^k is not a significant meta-graph according to Definition 4, then any $(k+1)$-path meta-graphs δ_i^{k+1} synthesised by δ_i^k is not significant.

4.5 The relationship prediction model

Based on the meta-paths and meta-graph features extracted in Section 4.3 and 4.4, we propose a logistic regression based prediction method to measure the probability that a query researcher gets interested in the new papers. The purpose of the model is to learn the weights associated with these features.

For each training example pair (a_i, p_{n_i}), let x_i be the $(d+1)$-dimensional vector including constant 1 and d topological features. y_i denotes the label of whether there are relationships between a_i and p_{n_i} ($y_i = 1$ if there are relationships, and otherwise 0), which follows binomial distribution with probability p_i. The probability p_i is modeled as follows:

$$p_i = \frac{e^{x_i \beta}}{e^{x_i \beta} + 1} \tag{3}$$

where β is the $d+1$ coefficient weights associated with the constant and each topological feature. Then we employ the standard MLE (Maximum Likelihood Estimation) to derive $\hat{\beta}$ which maximizes the likelihood of all the training example pairs.

5 Experiments

In this section, we conduct series of experiments to evaluate the performance of the proposed HIPRec method on DBLP-Citation-network V8[2] generated by [26].

5.1 Datasets

Note that the original DBLP dataset[3] does not contain the paper-citation relationships. Tang et al. extracted citation information from other sources and generated a DBLP citation dataset. Instead of using the entire dataset, we generate a subset which contains 32,133 papers from 20 venues[4] published during (2000, 2016), 39,530 researchers and 15,708 topics (topics are extracted from paper titles). More detailed information about the dataset used in the experiment are presented in Table 3.

[2] https://aminer.org/billboard/citation

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

[4] 20 very significant venues in the areas of Data Mining, Database, Information Retrieval and Artificial intelligence.

Table 3 Statistics of the dataset

NodeType	Author	Paper	Term	Venue
# of nodes	39,530	32,133	15,708	20
LinkType	# of link	Semantic meaning		
P-A/A-P	109,584	is published/publishes		
P-T/T-P	32,132	mentions/is mentioned		
P-V/V-P	32,133	is published/publishes		
P-P	67,435	cites/is cited		

5.2 Methodology and metrics

Example pairs Since the new papers have no historical information e.g., citations, it is impossible to evaluate the performance of the prediction model except for using the user study evaluation method. However, user study is somewhat subjective and expensive. Therefore, We consider the papers in the following condition as the new papers in the experiment.

First of all, papers are partitioned into two parts: T_0 = [2000, 2012] and T_1 = [2013, 2016]. Suppose user a published his first paper p_0 in 2008, and published 3 papers p_1, p_2, p_3 later in 2010, 2011, 2014 respectively. Then all the papers $\mathcal{P}$ that are published in T_0 but later than 2008, and are cited by $p_1/p_2/p_3$ are treated as the positive training examples. The intuition is that papers $\mathcal{P}$ are new papers at that time, and if we are able to discover how researchers find new papers in the past, it is possible to learn how researchers find interesting new papers today.

We select the papers in T_0 which are not directly cited by the researchers but within three citation hops as the negative training examples. Therefore, 1000 training example pairs (including 750 positive examples and 250 negative examples) are selected based on the methodology. Similarly, we select 500 researchers as well as one of their citing papers in T_1 to form the test example pairs. Note that, for the selected training or testing new papers, we remove all the citation links related to them in the experiment.

Meta-path instances The meta-paths we are interested in are those starting with the object type A and ending with the object type P_n, e.g., $\Pi_1 : A - P - V - P_n$. We should pay attention to the object P_n when searching the instances of this meta-path. For example, $mp_1 : a_1 - p_1(2008) - v_1 - p_2(2010)$ is an instance of the meta-path Π_1. However, $mp_2 : a_1 - p_1(2008) - v_1 - p_3(2002)$ is not. This is because p_2 is published after p_1, and it is a new paper for user a_1 if going back to that time. But p_3 is a previously published paper. Since our goal is to discover the patterns how researchers find newly published papers in the real world, mp_2 can not be treated as an instance of meta path Π_1.

Evaluation method Prediction accuracy rate is used to evaluate the performance of all the methods, which is defined in Eq. 4.

$$accuracy = \frac{N_i}{N_t} \tag{4}$$

where N_i denotes the number of interesting new papers for the query researcher, N_t represents the total number of tested new papers.

Comparative methods

- **LDA** computes the content similarity according to the topics extracted from the paper titles. For the candidate new paper p, we use the text information of paper p and the papers referenced by p to learn the topic distribution for p. As to the query researcher a, we use the titles of all his/her publications to learn the researcher's interested research topics.
- **CTR** [9] combines the traditional matrix factorization model and probabilistic topic model, which considers both the user-item ratings and textual similarity between papers.
- **UAGMT** [17] solve the new paper recommendation problem in the context of bipartite graph, which takes the readership, tag, content and citation into consideration. Since DBLP dataset doesn't have tags, we neglect the tag information in the experiment.
- **MPRec** considers the meta-path based topological features to train the model and solves the problem in the heterogeneous information networks.
- **HIPRec** systematically extracts both the meta-path and meta-graph based topological features, and fuse all the heterogeneous features into a prediction model to solve the recommendation problem.

5.3 Experiment results

In this study, we set the meta-path length L to 4 for simplicity, thus the top-5 interesting meta-paths and the significant meta-graphs are obtained as shown in Table 4. In addition, the learned coefficients of the meta-paths and meta-graphs are listed in the table. From Table 4 we can find that all the meta-paths/graphs features show positive effect on increasing the probability of establishing relationships. Note that meta-path Π_3 and meta-graph Γ_2 are more meaningful compared with the others. The results are consistent with the fact that how researchers find new papers in the reality, that is, researchers prefer new papers sharing similar research topics and published in similar venues.

Overall accuracy We first evaluate the accuracy of our method and all the comparisons. The comparative results

Table 4 The coefficients of the most significant meta-path and meta-graph features

Meta path/Graphs	Coefficients
$\Pi_1 : A \rightarrow P \rightarrow V \rightarrow Pn$	0.0627
$\Pi_2 : A \rightarrow P \rightarrow T \rightarrow Pn$	0.3018
$\Pi_3 : A \rightarrow P \rightarrow A \rightarrow Pn$	0.0315
$\Pi_4 : A \rightarrow P \rightarrow P \leftarrow Pn$	0.0446
$\Pi_5 : A \rightarrow P \rightarrow P \rightarrow T \leftarrow Pn$	0.0277
$\Gamma_1 : A \rightarrow P \rightarrow < T, A > \rightarrow Pn$	0.1164
$\Gamma_2 : A \rightarrow P \rightarrow < T, V > \rightarrow Pn$	0.4273
$\Gamma_3 : A \rightarrow P \rightarrow < A, V > \rightarrow Pn$	0.0511

are summarized in Fig. 5. We can find that LDA and CTR perform worse than the other methods. In our consideration, LDA and CTR only rely on the contents to make recommendation. However, the available contents of the papers in the experiment are the titles, which is not enough to generate accurate topic results. Thus it is difficult to find the most relevant new papers based on the content similarity. Since new papers have few historical information (e.g., citation-ships), CTR makes recommendation almost entirely based on the contents. Therefore CTR generates similar result compared to LDA. UAGMT transfers the problem into a bipartite graph, and solves it by employing readerships, contents and citations. We can observe that the accuracy of UAGMT is slightly higher than that of LDA and CTR.

It can be learned from Fig. 5 that HIPRec achieve much better accuracy than UAGMT. This is because our proposed method is designed to tackle the new paper recommendation problem in the heterogeneous information networks. Besides the links, the rich semantics underlying the links are considered, which can generate more accurate prediction results. Compared to MPRec, HIPRec exploits both the meta-path and meta-graph features, thus HIPRec performs the best among all the comparative methods. The experimental results demonstrate the effectiveness of the proposed significant meta-graphs mining algorithm. We can

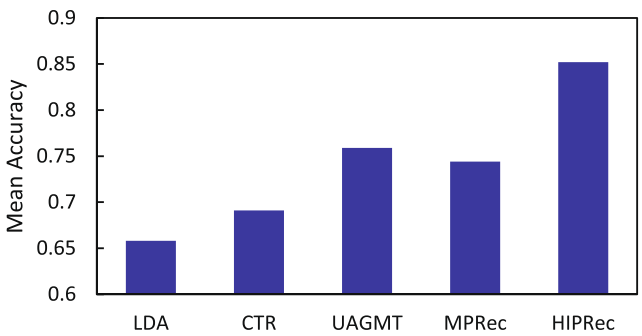

Fig. 5 Prediction accuracy of the comparative methods

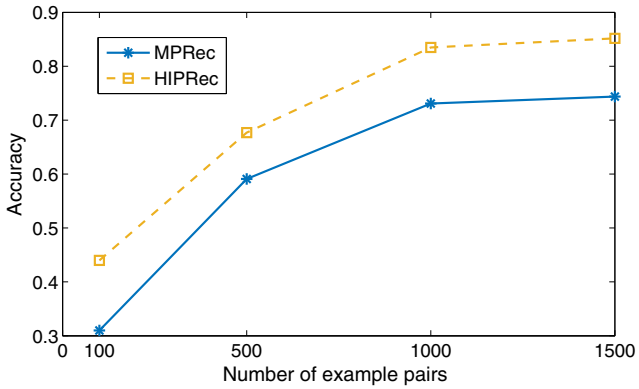

Fig. 6 Prediction accuracy for varying number of example pairs

also conclude that better prediction model can be generated by incorporating all the significant meta-paths and meta-graphs.

Second, we evaluate the accuracy of our proposed method for varying number of example pairs. The results are described in Fig. 6. It can be seen from the figure that the prediction accuracies of MPRec and HIPRec go up with the increase of the number of example pairs. This is because, with the increase of the training examples size, better meta-paths/graphs features can be learned, thus the prediction model will be more considerable. However, when the number of example pairs reaches around 1000, the set size does not greatly influence the accuracy of the model. That is to say, a relatively small portion of training examples are sufficient for HIPRec.

Efficiency Figure 7 shows the running time of MPRec and HIPRec when varying the number of example pairs given as input. It can be observed that there is a significant increase in running time when the size of example pairs increases. For MPRec, the increase is due to the computation of the instances of each meta-path for each example pair. The larger the size of example pairs is, the more time-consuming it is to train the model. For HIPRec, except the cost of searching the instances of each meta-path, it is also

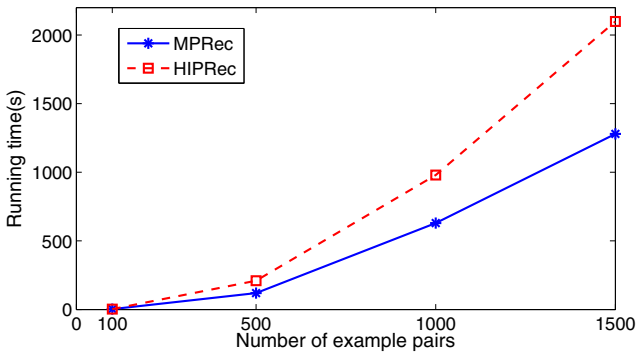

Fig. 7 Prediction accuracy for varying number of example pairs

expensive to search the instances of each meta-graphs for each example pair. However, as we can see from Fig. 6, large size of example pairs does not greatly influence the accuracy of the model but leads to worse running time. That's why we set the training example pairs to 1000 in Section 5.2 to balance the prediction accuracy and the running time.

Results analysis The experiment results give three interesting conclusions: (1) As we know that, it is usually difficult to obtain all the contents of new papers. Therefore the available contents of papers are quite limited, e.g.,only the titles can be used, which makes the content based recommendation method achieve poor results. On the other hand, the computation complexity will be very high if all the paper contents are used. (2) Traditional graph based recommendation methods are designed to tackle the new paper recommendation problem in the homogeneous or bipartite networks which ignore the semantic meanings underling the networks. However, in heterogeneous networks, different path/graph patterns between the same pair of researcher and new paper may represent different relations and denote different semantic meanings. (3) Meta-graph is a very powerful pattern which captures much richer and more complicated semantic meanings than simple meta-path. Recommendation methods that incorporate meta-graphs perform better than meta-path based recommendation methods.

6 Conclusion and future work

In this paper, we present a heterogeneous information network based newly published paper recommendation method. We introduce a principled way of discovering interesting meta-paths, and a greedy meta-graph mining algorithm named HIPRec is proposed to construct the significant meta-graphs. Then we solve the new paper recommendation problem by incorporating all the heterogeneous topological features into a logistic regression based researcher-paper prediction model. Experiments on the DBLP bibliographic network demonstrate the effectiveness of our approach, and the model using hybrid features that have combined all the significant meta-paths and meta-graphs gives the best overall performance. In the future, we plan to explore richer information to enrich the features and semantics in the network, develop more effective meta-graphs discovery methods, and apply the proposed framework to more intelligent applications [32–34].

Acknowledgements Research in this paper was partially supported by China National Natural Science Foundation (No.61702553) and MOE (Ministry of Education in China) Project of Humanities and Social Sciences (No.17YJCZH252). Appreciation also goes to anonymous reviewers for their careful work and thoughtful suggestions that have helped improve this paper substantially.

References

1. Basu C, Hirsh H, Cohen WW, Nevill-Manning CG (2001) Technical paper recommendation: a study in combining multiple information sources. J Artif Intell Res 231:14
2. Strohman T, Croft WB, Jensen D (2007) Recommending citations for academic papers. In: Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval. (ACM), pp 705–706
3. Tian G, Jing L (2013) Recommending scientific articles using bi-relational graph-based iterative rwr. In: Proceedings of the 7th ACM conference on recommender systems. (ACM), pp 399–402
4. Gupta S, Varma V (2017) Scientific article recommendation by using distributed representations of text and graph. In: Proceedings of the 26th international conference on world wide web companion. (ACM), pp 1267–1268
5. Amami M, Faiz R, Stella F, Pasi G (2017) A graph based approach to scientific paper recommendation. In: Proceedings of the international conference on web intelligence. (ACM), pp 777–782
6. McNee SM, Albert I, Cosley D, Gopalkrishnan P, Lam SK, Rashid AM, Konstan JA, Riedl J (2002) On the recommending of citations for research papers. In: Proceedings of the 2002 ACM conference on computer supported cooperative work. (ACM), pp 116–125
7. Zhang S, Yen NY, Zhu GL et al (2017) The recommendation system of Micro-blog topic based on user clustering. Mobile Networks and Applications 22(2):228–239
8. Yu X, Gu Q, Zhou M, Han J (2012) Citation prediction in heterogeneous bibliographic networks. In: Proceedings of the 2012 SIAM international conference on data mining. (SIAM), pp 1119–1130
9. Wang C, Blei DM (2011) Collaborative topic modeling for recommending scientific articles. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. (ACM), pp 448–456
10. Yang Z, Yin D, Davison BD (2014) Recommendation in academia: a joint multi-relational model. In: 2014 IEEE/ACM international conference on advances in social networks analysis and mining. (IEEE), pp 566–571
11. LL?, Medo M, Yeung CH, Zhang YC, Zhang ZK, Zhou T (2012) Recommender systems. Phys Rep 519(1):1
12. Sun Y, Barber R, Gupta M, Aggarwal CC, Han J (2011) Co-author relationship prediction in heterogeneous bibliographic networks. In: 2011 international conference on advances in social networks analysis and mining. (IEEE), pp 121–128
13. Ren X, Liu J, Yu X, Khandelwal U, Gu Q, Wang L, Han J (2014) 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), pp 821–830
14. Xia F, Liu H, Lee I, Cao L (2016) Scientific article recommendation: exploiting common author relations and historical preferences. IEEE Transactions on Big Data 2(2):101
15. Sugiyama K, Kan MY (2015) Towards higher relevance and serendipity in scholarly paper recommendation. ACM SIGWEB Newsletter, p 4. Article No. 4
16. Sugiyama K, Kan MY (2015) A comprehensive evaluation of scholarly paper recommendation using potential citation papers. Int J Digit Libr 16(2):91
17. Cai T, Cheng H, Luo J, Zhou S (2016) An efficient and simple graph model for scientific article cold start recommendation. In: Proceedings of the 35th international conference on conceptual modeling. (Springer), pp 248–259
18. Lao N, Cohen WW (2010) Relational retrieval using a combination of path-constrained random walks. Mach Learn 81(1):53

19. Huang Z, Chung W, Ong TH, Chen H (2002) A graph-based recommender system for digital library. In: Proceedings of the 2nd ACM/IEEE-CS joint conference on digital libraries. (ACM), pp 65–73
20. Ha J, Kwon SH, Kim SW, Lee D (2014) Recommendation of newly published research papers using belief propagation. In: Proceedings of the 2014 conference on research in adaptive and convergent systems. (ACM), pp 77–81
21. Zhu F, Qu Q, Lo D, Yan X, Han J, Yu PS (2011) Mining top-k large structural patterns in a massive network. In: Proceedings of the VLDB endowment, vol 4, p 807
22. Elseidy M, Abdelhamid E, Skiadopoulos S, Kalnis P (2014) Grami: frequent subgraph and pattern mining in a single large graph. In: Proceedings of the VLDB endowment, vol 7, p 517
23. Sun Y, Han J, Yan X, Yu PS, Wu T (2011) Pathsim: meta path-based top-k similarity search in heterogeneous information networks. In: Proceedings of the VLDB endowment, vol 4, p 992
24. Meng C, Cheng R, Maniu S, Senellart P, Zhang W (2015) Discovering meta-paths in large heterogeneous information networks. In: Proceedings of the 24th international conference on world wide web. (ACM), pp 754–764
25. Huang Z, Zheng Y, Cheng R, Sun Y, Mamoulis N, Li X (2016) Meta structure: computing relevance in large heterogeneous information networks. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. (ACM), pp 1595–1604
26. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z (2008) Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. (ACM), pp 990–998
27. Zhao W, Wu R, Liu H (2016) Paper recommendation based on the knowledge gap between a researcher?s background knowledge and research target. Inf Process Manag 52(5):976
28. Hassan HAM (2017) Personalized research paper recommendation using deep learning. In: Proceedings of the 25th conference on user modeling, adaptation and personalization. (ACM), pp 327–330
29. Anand A, Chakraborty T, Das A (2017) Fairscholar: balancing relevance and diversity for scientific paper recommendation. In: European conference on information retrieval. (Springer), pp 753–757
30. Pazzani MJ, Billsus D (2007) Content-based recommendation systems. In: The adaptive web. (Springer), pp 325–341
31. Kazemi B, Abhari A (2017) A comparative study on contentbased paper-to-paper recommendation approaches in scientific literature. In: Proceedings of the 20th communications and networking symposium (ACM), vol 5, pp 1–10
32. Zhang Y, Gravina R, Lu H, Villari M, Fortino G (2018) PEA: parallel electrocardiogram-based authentication for smart healthcare systems. J Netw Comput Appl 117:10–16
33. Lu H, Li Y, Chen M, Kim H, Serikawa S (2018) Brain intelligence: go beyond artificial intelligence. Mobile Networks and Applications 23:368–375
34. Pan Z, Liu S, Sangaiah AK, Muhammad K (2018) Visual attention feature: a novel strategy for visual tracking based on cloud platform in intelligent surveillance systems. J Parallel Distrib Comput 120:182–194