Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph

Zhuoren Jiang
[1]Sun Yat-sen University
Guangzhou, China
[2]Peking University, Beijing, China
jiangzhr3@mail.sysu.edu.cn

Yue Yin
Beijing Normal University
Beijing, China
bnuyinyue@outlook.com

Liangcai Gao*
Peking University
Beijing, China
glc@pku.edu.cn

Yao Lu
Sun Yat-sen University
Guangzhou, China
luyao23@mail.sysu.edu.cn

Xiaozhong Liu*
[1]Alibaba Group
Seattle & Hangzhou, China
[2]Indiana University Bloomington
Bloomington, IN, USA
liu237@indiana.edu

ABSTRACT

While the volume of scholarly publications has increased at a frenetic pace, accessing and consuming the useful candidate papers, in very large digital libraries, is becoming an essential and challenging task for scholars. Unfortunately, because of language barrier, some scientists (especially the junior ones or graduate students who do not master other languages) cannot efficiently locate the publications hosted in a foreign language repository. In this study, we propose a novel solution, cross-language citation recommendation via Hierarchical Representation Learning on Heterogeneous Graph (HRLHG), to address this new problem. HRLHG can learn a representation function by mapping the publications, from multilingual repositories, to a low-dimensional joint embedding space from various kinds of vertexes and relations on a heterogeneous graph. By leveraging both global (task specific) plus local (task independent) information as well as a novel supervised hierarchical random walk algorithm, the proposed method can optimize the publication representations by maximizing the likelihood of locating the important cross-language neighborhoods on the graph. Experiment results show that the proposed method can not only outperform state-of-the-art baseline models, but also improve the interpretability of the representation model for cross-language citation recommendation task.

KEYWORDS

Citation Recommendation; Cross-language; Heterogeneous Graph Representation Learning

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5657-2/18/07…$15.00
https://doi.org/10.1145/3209978.3210032

ACM Reference Format:
Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018. Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210032

1 INTRODUCTION

"It takes me a lot more time to find a useful paper... and it takes me even longer to read it... " while a non-English speaking PhD student complained this in a seminar, other PhD candidates, in the similar background, agreed with her and they shared the same frustration when they are trying to find and consume the helpful English publications. Professor's (a native speaker) response came later as a relief *"well, I agree, but my problem is even bigger... I cannot read the papers in your language at all..."* This dialog initialized our thinking about this new problem - **Cross Language Publication (Citation) Recommendation**, a.k.a. how can we propose a useful method/system to assist scholars to efficiently locate the useful publications written in different languages (a typical scenario of this task is to help non-English speaking students to search for useful English papers). Increased academic globalization is forcing a scholar to break the linguistic boundaries, and English (or any other dominant language) may not always serve as the gatekeeper to scientific discourse.

Unfortunately, existing academic search engines (e.g. Google Scholar, Microsoft Academic Search, etc.) along with many sophisticated retrieval and recommendation algorithms [11, 12, 28] cannot cope with this problem efficiently. For instance, most of the existing citation recommendation algorithms work in a monolingual context, and the scholarly graph-based random walk may not work well in a multilingual environment (section 4 will prove this).

Moreover, Cross-language Citation Recommendation (CCR) can be a quite challenging problem comparing with classical scholarly recommendation due to the following reasons:

Information need shifting. Different from monolingual citation recommendation, we cannot directly calculate the relevance

between the papers written in two different languages. A straightforward solution is to utilize machine translation (MT) [1] to translate the query content (e.g., keyword, text or user profile), then, use existing matching models [10, 30] to recommend the proper papers in target language. However, MT based methods and the CCR task can be fundamentally different. The goal of MT is to find a target text given a source text based on the same semantic meaning [1] (e.g., find the papers contain exact or similar matched phrases or sentences), while the CCR task is focusing on recommending "relevant" papers in target language to the given query in the source language [29]. When research context changes, content translation may not perform well. For instance, in Chinese/Japanese research context, machine learning methods can be important for word segmentation studies, which may not be the case for the English counterpart. MT approach cannot address this kind of information need shifting problem.

Sparse inter-repositories citation relations. Besides textual information, citation relations are quite important for citation recommendation. In the prior studies, recommendation algorithms can learn the "relevance" by using citation relations on a graph [15, 22]. However, compared to the enormous monolingual citation relations, cross-language citations can be very sparse. For instance, in a computer science related bilingual (Chinese-English) context, we find the papers in ACM and Wanfang[1], on average, have about 28 times more monolingual citation relations than cross-language ones. It is difficult to effectively employ the citation relations for cross-language citation recommendation by using classical graph mining methods.

Heterogeneous information environment. Intuitively, one could integrate the cross-language content semantics, citation relations and other useful heterogeneous information (e.g., keywords and authors) to address CCR. However, most existing text or graph based ranking algorithms rely on a set of human defined rules (e.g., sequential relation path [14] and meta-path [26]) to integrate different kinds of information. On a complex cross-language scholarly graph, this kind of handcrafting features can be time-consuming, incomplete and biased.

To address these challenges, in this study, we propose a novel solution, **Hierarchical Representation Learning on Heterogeneous Graph (HRLHG)**, for cross-language citation recommendation. By constructing a novel cross-language heterogeneous graph with various types of vertexes and relations, we "semantically" enrich the basic citation structure to carry more rich information. To avoid the handcrafting feature usage, we propose an innovative algorithm to project a vertex (on the heterogeneous graph) to a low-dimensional joint embedding space. Unlike prior hyperedge or meta-path approaches, the proposed algorithm can generate a set of Relation Type Usefulness Distributions (RTUD), which enables fully automatic heterogeneous graph navigation. As Figure 1 shows, the hierarchical random walk algorithm enables a two-level random walk guided by two different sets of distributions. The global one (relation type usefulness distributions) is designed for graph schema level navigation (task-specific); while the local one (relation transition distributions) targets on graph instance level walking (task-independent).

[1]One of the biggest digital libraries in Chinese.

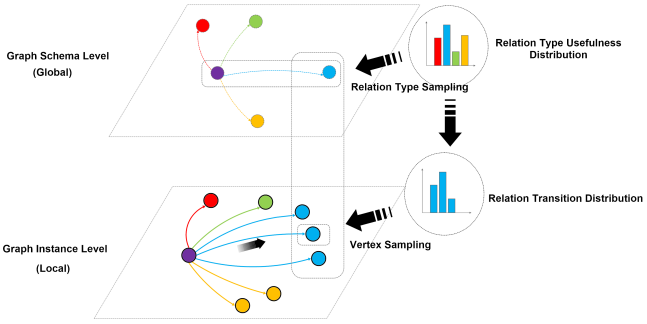

Figure 1: Hierarchical random walk illustration (different colours denote different types)

By using HRLHG, we can recommend a list of ranked cross-language citations for a given paper/query in the source language. We evaluate the proposed algorithm in Chinese and English scholarly corpora, i.e., Wanfang and ACM digital libraries. The results demonstrate that the proposed approach is superior than state-of-the-art models for cross-language citation recommendation task.

The contribution of this paper is fourfold. First, we propose a novel method (Hierarchical Representation Learning on Heterogeneous Graph) to characterize both the global and local semantic plus topological information for the publication representations. Second, we improve the interpretability of the publication representation model. By using an iterative EM (expectation-maximization) approach, the proposed algorithm can learn the implicit biases for cross-language citation recommendation, which significantly differs from classical heterogeneous graph mining algorithms. Third, we apply the proposed embedding method for a novel cross-language citation recommendation task. An experiment on real-world bilingual scientific datasets is employed to validate the proposed approach. Last but not least, although in this study we focus on cross-language citation recommendation task, the proposed method can be generalized for different tasks that are based on heterogeneous graph embedding learning.

2 PROBLEM FORMULATION

Compared to the homogeneous graph, the heterogeneous graph has been demonstrated as a more efficient way to model real world data for many applications, it represents an abstraction of the real world, focusing on the objects and the interactions between the objects [15]. Formally, following the works [6, 26], we present the definitions of a heterogeneous graph with its schema.

Definition 1. ***Heterogeneous Graph**, namely heterogeneous information network, is defined as a graph $G = (V, E, \tau, \gamma)$, where V denotes the vertex set, and $E \subseteq V \times V$ denotes the edge (relation) set. τ is the vertex type mapping function, $\tau : V \rightarrow \mathbb{N}$ and $\mathbb{N}$ denotes the set of vertex types. γ is relation type mapping function, $\gamma : E \rightarrow \mathbb{Z}$ and $\mathbb{Z}$ denotes the set of relation types. $|\mathbb{N}| + |\mathbb{Z}| > 2$.*

Definition 2. ***Graph Schema**. The graph schema is a meta template for a heterogeneous graph $G = (V, E, \tau, \gamma)$, denoted as $S_G = (\mathbb{N}, \mathbb{Z})$.*

The graph schema is used to specify type constraints on the sets of vertexes and relations of a heterogeneous graph. A graph that

follows a graph schema is then called a **Graph Instance** following the target schema [15].

Definition 3. ***Cross-language citation recommendation***. *The CCR problem can be defined as a conditional probability $Pr(p_c|p_q)$, i.e., the probability of p_c in target language given a particular query paper p_q in the source language:*

$$Pr(p_c|p_q) = \Delta\left(\phi(p_q), \phi(p_c)\right)$$

where ϕ is a representation function, which can project each paper to a low-dimensional embedding space. Δ is probability scoring function based on the learned publication embeddings.

The CCR problem can be formalized as:

- **Input**: A query paper (or partial text/keywords in the query paper) in a source language.
- **Output**: A list of ranked papers in target language that could be potentially cited or useful given the input paper.

In this study, we investigate the novel method to enhance the representation learning function ϕ for CCR. More detailed method will be introduced in Section 3.

3 HIERARCHICAL REPRESENTATION LEARNING ON HETEROGENEOUS GRAPH

In this section, we discuss the proposed method in detail. We first formulate the heterogeneous graph based representation learning framework for CCR task (3.1), then, introduce the hierarchical random walk-based strategy by leveraging the critical relation type usefulness distribution training algorithm (3.2)

3.1 Heterogeneous Graph Representation Learning Framework for CCR

Due to the aforementioned challenges of CCR task, the proposed representation model can hardly depend only on textual or citation information. In this study, we integrate various kinds of entities and relations into a heterogeneous graph (as Figure 2 shows, and the detailed node and type information can refer to Table 1). Then, the goal is to design a novel representation learning model to encapsulate both semantic and topological information into a low-dimensional joint embedding for CCR task.

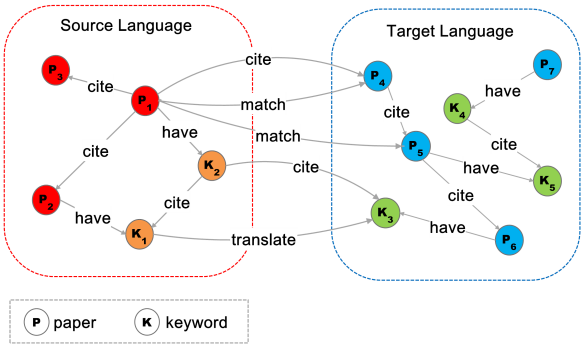

Figure 2: The constructed cross-language heterogeneous graph (different colours denote different types)

Formally, given a heterogeneous graph $G = (V, E, \tau, \gamma)$, $\tau : V \rightarrow \mathbb{N}$ is the vertex type mapping function and $\gamma : E \rightarrow \mathbb{Z}$ is the relation type mapping function. The goal of vertex representation learning is to obtain the latent vertex representations by mapping vertexes into a low-dimensional space $\mathbb{R}^d$, $d \ll |V|$. The learned representations are able to preserve the information in G. We use $f : V \rightarrow \mathbb{R}^d$ as the mapping function from multi-typed vertexes to feature representations. Here, d is a parameter specifying the number of dimensions. f is a matrix of size $|V| \times d$ parameters. The following objective function should be optimized for heterogeneous graph representation learning.

$$\max_f \sum_{v \in V} \sum_{n \in \mathbb{N}} \sum_{v_n^c \in N_n(v)} logPr(v_n^c | \overrightarrow{f(v)}) \quad (1)$$

where $N_n(v)$ denotes v's network neighborhood ("context") with the n^{th} type of vertexes. The feature learning methods are based on the Skip-gram architecture [2, 18, 20], which is originally developed for natural language processing and word embedding. Unlike the linear nature of text, the structural and semantic characteristics of graph allow the vertex's network neighborhood, $N(v)$, to be defined in various of ways, i.e., direct (one-hop) neighbors of v. It is critical to model vertex neighborhood in graph representation learning. Following the previous network embedding models [6, 9, 21], in this study, we leverage a random walk-based strategy for every vertex $v \in V$ to generate $N(v)$. For instance, in Figure 2, we can

Table 1: Vertexes and relations of cross-language heterogeneous graph

No.	Vertex	Description
1	P_s	Paper in Source Language
2	P_t	Paper in Target Language
3	K_s	Keyword in Source Language
4	K_t	Keyword in Target Language
No.	**Relation**	**Description**[§]
1	$P_s \xrightarrow{s} P_t$	A paper (in source language) is semantically related to another paper (in target language). We use machine translation* and language model (with Dirichlet smoothing) to generate this relation [30].
2	$P_s \xrightarrow{c} P_s$	A Paper (in source language) has a monolingual citation relation to another paper (in source language).
3	$P_t \xrightarrow{c} P_t$	A Paper (in target language) has a monolingual citation relation to another paper (in target language).
4	$P_s \xrightarrow{c} P_t$	A Paper (in source language) has a cross-language citation relation to another paper (in target language).
5	$P_s \xrightarrow{h} K_s$	A Paper (in source language) has a keyword (in source language).
6	$P_t \xrightarrow{h} K_t$	A Paper (in target language) has a keyword (in target language).
7	$K_s \xrightarrow{c} K_s$	A keyword (in source language) has a monolingual citation relation to another keyword (in source language)‡.
8	$K_t \xrightarrow{c} K_t$	A keyword (in target language) has a monolingual citation relation to another keyword (in target language)‡.
9	$K_s \xrightarrow{c} K_t$	A keyword (in source language) has a cross-language citation relation to another keyword (in target language)‡.
10	$K_s \xrightarrow{t} K_t$	A keyword (in source language) is translated into the corresponding keyword (in target language)*.

*As this study is not focusing on machine translation, we use Google machine translation API (https://cloud.google.com/translate) to translate the paper abstract and keywords.
‡The keyword citation relations are derived from paper citation relations.
§Because of the space limitation, the detailed relation transition probability calculation can be found at https://github.com/GraphEmbedding/HRLHG.

sample a random walk sequence $\{p_1, k_2, k_3, p_6\}$ of length $l = 4$, which results in $N(p_1) = \{k_2\}$, $N(k_2) = \{p_1, k_3\}$, $N(k_3) = \{k_2, p_6\}$ and $N(p_6) = \{k_3\}$ (window size is 1). The detailed description of this method will be introduced in section 3.2.

$Pr(v_n^c|\overrightarrow{f(v)})$ defines the conditional probability of having a context vertex $v_n^c \in N_n(v)$ given the node v's representation, which is commonly modeled as a softmax function :

$$Pr(v_n^c|\overrightarrow{f(v)}) = \frac{exp(\overrightarrow{f(v_n^c)} \cdot \overrightarrow{f(v)})}{\sum_{u \in V} exp(\overrightarrow{f(u)} \cdot \overrightarrow{f(v)})} \quad (2)$$

In this study, we use a **Heterogeneous Softmax** function for conditional probability $Pr(N(v)|\overrightarrow{f(v)})$ calculation [6]:

$$Pr(v_n^c|\overrightarrow{f(v)}) = \frac{exp(\overrightarrow{f(v_n^c)} \cdot \overrightarrow{f(v)})}{\sum_{u_n \in V_n} exp(\overrightarrow{f(u_n)} \cdot \overrightarrow{f(v)})} \quad (3)$$

where V_n is the vertex set of type n in G. Different from common Skip-gram form, the **Heterogeneous Skip-gram** with heterogeneous softmax function can specify one set of multinomial distributions for each type of neighborhood in the output layer of the Skip-gram model. Stochastic gradient ascent is used for optimizing the model parameters of $\overrightarrow{f}$. Negative sampling [20] is applied for optimization efficiency. Especially, for "heterogeneous softmax", the negative vertexes are sampled from the graph according to their type information [6].

Recall the CCR definition in section 2, given a query paper p_q in source language, the problem is to compute the recommendation probability $Pr(p_c|p_q)$ of a candidate paper p_c in target language, with representation function ϕ and probability scoring function Δ. In this study, ϕ is the optimized heterogeneous vertex representation $\overrightarrow{f}$, and we use cosine similarity with Relu function for Δ:

$$Pr(p_c|p_q) = Max(0, \frac{\overrightarrow{f(p_q)} \cdot \overrightarrow{f(p_c)}}{\left\|\overrightarrow{f(p_q)}\right\| \left\|\overrightarrow{f(p_c)}\right\|}) \quad (4)$$

3.2 Hierarchical Representation Learning on Heterogeneous Graph

In this section, we propose a novel hierarchical random walk-based strategy for vertex neighborhoods on heterogeneous graph. Before moving on, let's clarify four challenges for random walk-based graph embedding models.

(1) The existing homogeneous random walk-based embedding approaches, e.g., [9, 21], cannot be directly applied to address the heterogeneous graph problems. For instance, as Figure 2 shows, between the paper pair p_1 and p_3, there are two different types of relations: $p_1 \xrightarrow{s} p_3$ (p_1 is semantically related to p_3) and $p_1 \xrightarrow{c} p_3$ (p_1 cites p_3), but the homogeneous random walk-based approaches cannot distinguish the difference of relation types, then the neighborhood generating could be problematic for further representation learning.

(2) Recent heterogeneous graph embedding algorithms [6, 7] require a domain expert to generate random walk hypotheses, which can be inconvenient and problematic for complex heterogeneous graphs.

(3) Insufficient global information. For each step in the walk, a lot of random walk-based models are solely depending on the (local) network topology of the vertexes, but global information, i.e., graph schema information, may bring important information to navigate the walker on the graph.

(4) Most existing graph embedding methods aim to encode the topological information of the graphs, which are task independent. For instance, as described in Table 1, the vertexes and relations with transition probability are fixed after the graph is constructed. We argue that, the learned representation should be optimized for different tasks, e.g., cross-language citation recommendation task for this study. A flexible representation mechanism can be important to address recommendation problem via heterogeneous graph. For instance, on a complex graph, some kinds of relations can be more important for random walk than others given the task (conditional relation type usefulness probability given the task).

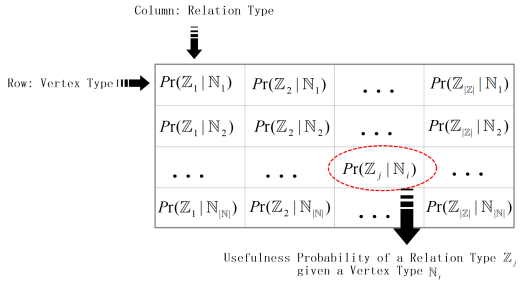

Figure 3: Relation type usefulness distributions illustration

To address these challenges, we propose a Hierarchical Representation Learning on Heterogeneous Graph (HRLHG) method. By introducing a set of Relation Type Usefulness Distributions (RTUD) on graph schema, the hierarchical (two-level) random walk algorithm can be more appropriate for heterogeneous network structure. As RTUD can be automatically learned, we don't need expert knowledge (e.g., generating meta-path) for representation learning. Meanwhile, by using RTUD, we not only bring global information for guiding the random walk, but also utilize the task specific information for optimizing the random walk generation.

Given a specific task T on a heterogeneous graph G: **Relation Type Usefulness Distributions (RTUD)** is a group of task-preferred (usefulness) probability distributions over relation types, which is defined at graph schema level (global level) of G. As Figure 3 shows, RTUD can be represented as a probability matrix β of size $|\mathbb{N}| \times |\mathbb{Z}|$, where i^{th} row of this matrix represents a relation type usefulness distribution β_i given a specific vertex type $\mathbb{N}_i$, in which $\beta_{i,j} = Pr(\mathbb{Z}_j|\mathbb{N}_i)$ denotes the usefulness probability of a relation type $\mathbb{Z}_j$ given $\mathbb{N}_i$.

Correspondingly, **Relation Transition Distributions (RTD)** is a group of task-independent probability distributions associated to relations, which is defined at graph instance level (local level) of G. Given a vertex v^* of type $\mathbb{N}_n$, $\alpha_j = Pr^j_{RTD}(V_m|v^*)$ denotes the transition distribution of a type $\mathbb{Z}_j$ relation (from vertex type $\mathbb{N}_n$ to vertex type $\mathbb{N}_m$), V_m are vertexes of $\mathbb{N}_m$ type. RTD aims to reflect the basic semantics of different types of relations in G and focuses on the local structure around v^*. For instance, Table 1 defines the RTD of cross-language heterogeneous graph.

As Figure 1 shows, with RTUD and RTD, we can simulate a hierarchical random walk of fixed length l in G. In order to avoid walking into a dead end, the directions of relations are ignored in hierarchical random walk algorithm. The hierarchical random walk process is as follows:

(1) For i_{th} vertex v_i in the walk:
  (a) Generate $\beta_n = (\beta_{n,1}, \cdots, \beta_{n,|\mathbb{Z}|})$ from RTUD based on v_i's vertex type $\mathbb{N}_n$
  (b) Probabilistically draw a relation type $\mathbb{Z}_z$ from β_n
  (c) For the generated relation type $\mathbb{Z}_z$
    (i) Generate α_z from RTD based on $\mathbb{Z}_z$
    (ii) Based on v_i, probabilistically draw one vertex from α_z as the destination vertex v_{i+1}
    (iii) Walk forward to v_{i+1}

Algorithm 1 RTUD training algorithm: a K-shortest paths ranking based EM approach

Initialize RTUD, a.k.a, the probability matrix β (Row: vertex type, column: relation type). For each row β_i, every element $\beta_{i,j}$ is set to be equal (here, $\beta_{i,j}$ denotes $Pr(\mathbb{Z}_j|\mathbb{N}_i)$, must fit for the graph schema $S_G = (\mathbb{N}, \mathbb{Z})$, the $\beta_{i,j}$ that violates S_G is set to 0)

procedure E-Step: K-shortest paths ranking
  Initialize Θ, every element is set to be zero
  for each $\{v_s, v_t\} \in Set_L$ **do**
    Calculate $w(p)$ and find K-shortest paths P^*
    for each $p^* \in P^*$ **do**
      for each relation $e \in \mathbb{Z}_j$ from p^* **do**
        Update Θ_j by $\Theta_j = F_\Theta(c)$
      end for
    end for
  end for
  Call M-Step
end procedure

procedure M-Step: Update β
  for each row $\beta_i \in \beta$ (vertex type) **do**
    for each relation $\mathbb{Z}_j \in \mathbb{Z}$ **do**
      update $\beta_{i,j}$ by $\beta_{i,j}^{n+1} = F_\beta(\beta_{i,j}^n, \Theta)$
    end for
  end for
  if P_{All}^* stabilize (a.k.a, ε% of the shortest paths ranking in the all shortest path sets are no longer changing) **then**
    Algorithm End
  else
    Call E-Step
  end if
end procedure

In this study, with a set of M labeled vertex pairs Set_L, we propose an iterative K-shortest paths ranking based EM (expectation - maximization) approach to obtain and optimize RTUD. Set_L is generated based on the task-specified relevance. For instance, for CCR task, a pair of paper vertexes $\{v_s, v_t\}$ connected via a cross-language relation could be a labeled pair for RTUD training. For a specific task, the representations of relevant pair of vertexes should be similar.

Then, in the proposed hierarchical representation learning framework, the goals are: (1) the vertex neighborhood N_v should contain the task-relevant vertexes to the greatest extent possible; (2) the distance (random walk sequence length) between two task-relevant vertexes should be as short as possible. In other words, RTUD should be trained to navigate the random walk between the related vertexes pairs, a.k.a., with the trained RTUD, there is a greater chance that one relevant vertex could random walk to another relevant one on the heterogeneous graph.

We formalize this goal as a K-shortest paths ranking problem. Let a path p from v_s to v_t in G is a sequence of vertexes and relations with the form:

$$p = \left\{ v_s \overset{relation_1}{\rightarrow} v_{s+1} \cdots \overset{relation_i}{\rightarrow} v_t \right\}$$

P denotes the set of all paths from v_s to v_t in G. Given a relation e_r of type $\mathbb{Z}_z$ from vertex v_i of type $\mathbb{N}_n$ to vertex v_j of type $\mathbb{N}_m$, the e_r's weight $w_{i,j}^z$ integrated RTUD and RTD, which can be calculated as:

$$w_{i,j}^z = \frac{1}{Pr_{RTUD}(\mathbb{Z}_z|\mathbb{N}_n) \cdot Pr_{RTD}^z(v_j|v_i)}$$

The weight function of p is $w(p) = \sum_p w_{i,j}^z$, a weight sum of all relations from p. The shortest path objective is the determination of a path $p^* \in P$ for which $w(p^*) \leq w(p)$ holds for any path $p \in P$ [8]. Then, the K-shortest paths objective is extended to determine the second, third,..., Kth shortest paths in P, that can be denoted as P^*. There are lots of efficient algorithms for this problem, we utilize the method proposed in [5]. The RTUD training utilizes an EM framework, as described in Algorithm 1.

In Algorithm 1, $\Theta = \{\Theta_1, \cdots, \Theta_{|\mathbb{Z}|})\}$ is a relation type update factor vector, Θ_i denotes the update value of ith relation type, c denotes one count for the appearance of a specific type relation in the shortest paths. We explore 3 ways for relation type update factor function F_Θ:

Raw Count (RC): $F_\Theta = c + +$. During each iteration, directly accumulate the relation type count.

Length-Normalized Count (LNC): $F_\Theta = \frac{c}{L_{p^*}} + +$. During each iteration, accumulate the relation type count that is normalized by the path length. L_{p^*} is the length of path p^*. By doing so, we try to minimize the possible bias from the long paths.

Log-Discounted Count (LDC): $F_\Theta = \frac{c}{log_2(k+1)} + +$. During each iteration, accumulate the relation type count that is discounted by path ranking. K is the rank of the path p, the shortest path's rank is 1. Using this update function, different shortest paths are given different weights.

For RTUD update function F_β, we define 2 different forms:

Direct Sum (DS): in DS, we update β by directly adding the update values, η is for normalization, $F_\beta = \frac{\beta_{i,j} + \Theta_j}{\eta}$.

Sum with a Dumping Factor (SDF): in order to avoid the extreme probability, in SDF we add a dumping factor λ for updating, $|\mathbb{Z}|^*$ is the possible relation type amount for a specific vertex type (constrained by graph schema).

$$F_\beta = (\lambda(\frac{\beta_{i,j} + \Theta_j}{\sum_{|\mathbb{Z}|}(\beta_{i,j} + \Theta_j)}) + \frac{1-\lambda}{|\mathbb{Z}|^*})/\eta$$

Note that, RTUD is constrained by graph schema. Given a vertex type $\mathbb{N}_i$, if a relation type $\mathbb{Z}_j$ violates the graph schema, the probability $Pr(\mathbb{Z}_j|\mathbb{N}_i)$ will be set to zero. For instance, for a keyword vertex, the usefulness probability of the paper citation relation (a

relation between paper vertex pairs) is zero. RTUD is task-specified, that means RTUD can dynamically change for different tasks, even though they share the same graph (e.g., we can use this graph for collaborator recommendation task, but the corresponding RTUD may change).

The pseudocode for Hierarchical Representation Learning on Heterogeneous Graph (HRLHG) is given in Algorithm 2. By applying r random walks of fixed length l starting from each vertex in G, we can minimize the implicit random walk biases. The RTUD β can be pre-trained by the K-shortest paths ranking based EM approach. The space complexity of HRLHG is $O(|E|)$, where $|E|$ is the relation number of G. The time complexity is $O(|\mathbb{Z}| + D)$ per hierarchical random walk, where $|\mathbb{Z}|$ is the relation type number, and D is the relation instance number of a specific sampled type connected to the current walking vertex. The time complexity can be further reduced, as suggested by [9], if we parallelize the hierarchical random walk simulations, and execute them asynchronously[2].

Algorithm 2 Hierarchical Representation Learning on Heterogeneous Graph (HRLHG)

RepresentationLearning (Heterogeneous Graph $G = (V, E, \tau, \gamma)$, Relation Transition Distributions (RTD) α, Dimensions d, Walks per vertex r, Random Walk Length l, Context Window size ws, Task-specified Relevance Labeled Set Set_L)
β = **K-ShortestPathEM** (G,Set_L)
Initialize *walks* to Empty
for *iter* = *1* to r **do**
  for all vertexes $v \in V$ **do**
    walk = **HierarchicalRandomWalk** (G, v, l, β, α)
    Append *walk* to *walks*
  end for
end for
f = **HeterogeneousSkipGram** (ws, d, $walks$)
return f

HierarchicalRandomWalk (Heterogeneous Graph $G = (V, E, \tau, \gamma)$, Start vertex v, Random Walk Length l, Relation Type Usefulness Distributions (RTUD) β, Relation Transition Distributions (RTD) α)
Initialize $walk$ to $\{v\}$
for $walk\_step$ = 1 to l **do**
  Generate β_n from β based on v's vertex type $\mathbb{N}_n$
  Probabilistically draw a relation type $\mathbb{Z}_z$ from β_n
  Based on v, probabilistically draw one vertex v_t from α_z
  Append v_t to $walk$
end for
return $walk$

4 EXPERIMENT

4.1 Dataset and Experiment Setting

Dataset[2]. We validated the proposed approach in a citation recommendation task between Chinese and English digital libraries. The goal was to recommend English candidate cited papers for a given Chinese publication. For this experiment, we collected 14,631 Chinese papers from the Wanfang digital library and 248,893 English papers from the Association for Computing Machinery (ACM) digital library (both in computer science). There were 750,557 English-to-English paper citation relations, 11,252 Chinese-to-Chinese paper citation relations, 27,101 Chinese-to-English paper citation relations, and 12,403 English papers had been cited by 7,900 Chinese papers. By using machine translation[3] and language modeling (with Dirichlet smoothing). We generated 158,000 cross-language semantic matching relations (from Chinese to English). There were 3,953 Chinese keywords associated to the collected Chinese papers, and the Chinese paper-keyword-associated relation number was 7,316; while there were 7,436 English keywords associated to the collected English papers and the English paper-keyword-associated relation number was 903,265. Between keywords, There were 283,268 English-to-English keyword citation relations, 2,973 Chinese-to-Chinese keyword citation relations, 9,828 Chinese-to-English keyword citation relations. 2,564 Chinese keywords could be successfully translated into the corresponding English keywords[3].

Ground Truth and Evaluation Metric. For evaluation, we generated a number of positive and negative instances to compare different algorithms for CCR task. The actual cross-language citation relation was used as ground truth (as 0 or 1 relevant scores) for evaluation. For example, if a candidate ACM paper was cited by the a testing Wanfang paper, the relevant score was 1, otherwise it was 0. We generated test and candidate collection data by using the following method: (1) randomly selected a certain proportion of papers from 7,900 Chinese papers that had cross-language citation relations to English corpus; (2) removed all cross-language citation relations from selected Chinese papers (Other relations, e.g., Chinese citation relations, were kept for model training); (3) the selected papers were used as a test collection. All the English papers cited by the Chinese papers in the test collection were used as candidate (cited paper) collection. For evaluation, the different models were compared by using the mean average precision (MAP), normalized discounted cumulative gain at rank (NDCG), precision (P) and Mean Reciprocal Rank (MRR).

Validation Set. For HRLGH, there were several hyper parameters (i.e., k for shortest path EM method) and algorithm functions (i.e., F_Θ and F_β for RTUD training) needed to be tuned. Meanwhile, for a fair comparison, we also tuned the hyper parameters of baselines (i.e., return parameter p and in-out parameter q for node2vec algorithm) for making sure the baseline algorithms could achieve the best performance. So, we constructed a validation set following the process described above (10% papers were randomly selected for validation). A comprehensive model component analysis and baseline hyper parameter tuning would be conducted via this validation set.

Baselines. We compared with three groups of representation algorithms, from text or graph viewpoints, to comprehensively evaluate the performance of the proposed method. 10-fold cross-validation was applied to avoid evaluation bias.

Textual Content Based Method.

1. Embedding Transformation [19]: We transformed the testing Chinese paper's abstract embedding into the English embedding space through a trained transformation matrix. Then, recommend

[2]The source code of HRLHG, constructed graph data (with labeled ground truth) and learned representations are available at https://github.com/GraphEmbedding/HRLHG

[3]As this study is not focusing on machine translation, we use Google translation API (https://cloud.google.com/translate) to translate the paper abstract and keywords.

the English citations based on the transformed Chinese abstract embedding, denoted as **EF**.

2. Machine Translation by Google Translation API + Language Model (with Dirichlet smoothing) [30]: We translated the testing Chinese paper's abstract into English, and then used language model to recommend English citations, denoted as **MT+LM**.

Collaborative Filtering Based Method.

3. Item-based Collaborative Filtering [23]: Recommended English citations using Item-based Collaborative Filtering based on (monolingual + cross-language) citation relations, denoted as $\mathbf{CF}_I$.

4. Popularity-based Collaborative Filtering [25]: Recommended English citations using Popularity-based Collaborative Filtering based on (monolingual + cross-language) citation relations, denoted as $\mathbf{CF}_P$.

Network Representation Learning Based Method.

5. DeepWalk [21]: We used DeepWalk to learn the graph embeddings via *uniform random walk* in the network and recommended English citations based on the learned embeddings. Because DeepWalk was originally designed for homogeneous graph, for a fair comparison, we applied DeepWalk on two graphs, (1) citation network, denoted as $\mathbf{DW}_c$; (2) all typed networks with accumulated relation weights. For this approach, we integrated all relations between two vertexes into one edge, and the weight was estimated by the sum of all integrated relations. Then, a heterogeneous graph could be simplified to a homogeneous graph, denoted as $\mathbf{DW}_{all}$.

6. LINE [27]: LINE aimed at preserving first-order and second-order proximity in concatenated embeddings. Similar as DeepWalk, we applied LINE on two graphs, with LINE 1st-order and 2nd-order representation approaches. So, there were four different baseline models, denoted as $\mathbf{LINE}_c^{1st}$, $\mathbf{LINE}_c^{2nd}$, $\mathbf{LINE}_{all}^{1st}$ and $\mathbf{LINE}_{all}^{2nd}$.

7. node2vec [9]: node2vec learned graph embeddings via *2nd order random walks* in the network. Similar as DeepWalk and LINE, we also applied node2vec on two graphs for comparison, denoted as $\mathbf{N2V}_c$ and $\mathbf{N2V}_{all}$. We tuned return parameter p and in-out parameter q with a grid search over $p, q \in \{0.25, 0.50, 1, 2, 4\}$ on the validation set, and picked up a best performed parameter setting for experiment, as suggested by [9].

8. metapath2vec++ [6]: metapath2vec++ was originally designed for heterogeneous graphs. It learned heterogeneous graph embeddings via *metapath based random walk* and *heterogeneous negative sampling* in the network. Metapath2vec++ required a human-defined metapath scheme to guide random walks. We tried 3 different metapaths for this experiment: (1) $P_s \xrightarrow{h} K_s \xrightarrow{c} K_t \xleftarrow{h} P_t$, (2) $P_s \xrightarrow{h} K_s \xrightarrow{t} K_t \xleftarrow{h} P_t$, (3) $P_s \xrightarrow{s} P_t \xrightarrow{c} P_t$. These metapaths were denoted as $\mathbf{M2V{+}{+}}_1$, $\mathbf{M2V{+}{+}}_2$ and $\mathbf{M2V{+}{+}}_3$, respectively. We also trained two learning to rank models (Coordinate Ascent [17] and ListNet [4]) to further integrate these three metapath2vec++ models (by utilizing each metapath as a ranking feature), denoted as $\mathbf{M2V{+}{+}}_{CA}$ and $\mathbf{M2V{+}{+}}_{LN}$.

For a fair comparison, for all the random walk based embedding methods, we used the same parameters as follows: (1) The number of walks per vertex r: 10; (2) the walk length l: 80; (3) the vector dimension d: 128; (4) the neighborhood size (Context Window size) ws: 10. Please note that most original baseline papers used the above parameter settings, and the proposed method also shared the same parameters. For the experiment fairness, we didn't tune those parameters on validation set. We applied the parameter sensitivity analysis in section 4.2.

4.2 Impact of Different Model Components

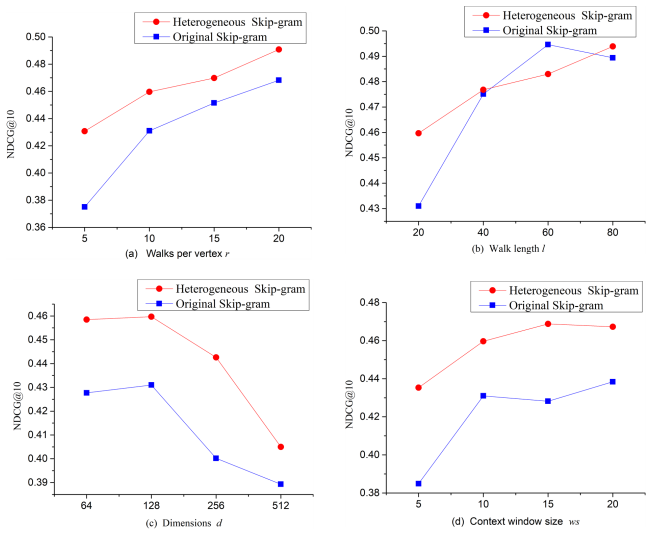

Figure 4: Sensitivity analysis for the embedding related parameters

For the proposed HRLHG, there were several important parameters and functions. To explore the effects of those model components, on the validation set, we compared the cross-language recommendation performances of proposed method under different model settings (by varying the examined model component while kept others fixed). We mainly focused on following components: (a) parameter k for K-shorest paths based EM algorithm, we compared and selected the best k from $k \in \{1, 2, 3\}$. (b) F_Θ, relation type update factor function for RTUD training: Raw Count (RC), Length-Normalized Count (LNC) and Log-Discounted Count (LDC). (c) F_β, RTUD update function for model training: Direct Sum (DS) and Sum with a Dumping Factor (SDF). (d) λ, parameter for F_β of SDF form, we compared and selected the best λ from $\lambda \in \{0.1, 0.2, ..., 0.9\}$. (e) ε, the convergence percentage of EM algorithm, we tried ε over 90% to 10%. (f) Validation of relation type usefulness distributions (RTUD) and heterogeneous skip-gram (HS).

Note that, we conducted a comprehensive comparison experiment. For each examined model component, we tried multiple combinations of other components to avoid the possible bias brought by the component setting choices. For instance, we tested the impact of different k under component combination (F_Θ = LNC, F_β= DS and ε= 80% for RTUD training, while using ordinary skip-gram for embedding) and component combination (F_Θ = RC, F_β= SDF with λ = 0.8 and ε= 20% for RTUD training, while using heterogeneous skip-gram for embedding), respectively. Because of the space limitation, we cannot report all results in this paper. The representative results on the validation set in terms of NDCG are depicted in Figure 5 (the other comparison groups showed the similar trends).

As we can see, considering more shortest paths in RTUD training brings a performance improvement. Length-normalized count (LNC) function could achieve best among the three F_Θ choices.

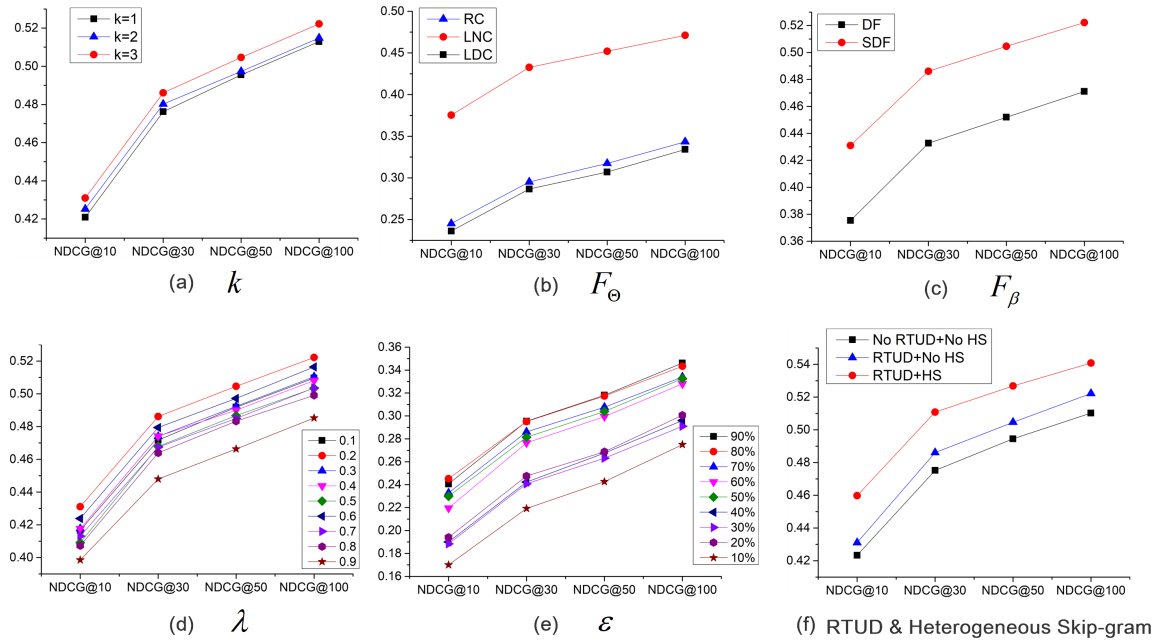

Figure 5: Hyper parameter comparison and algorithm component validation for the proposed method: (a) Comparison of k for K-shortest paths ranking based EM training; (b) Comparison of relation type update factor function F_Θ for RTUD training; (c) Comparison of RTUD update function F_β for training; (d) Comparison of the dumping factor λ for F_β of SDF form; (e) Comparison of the convergence percentage ε of K-shortest paths ranking based EM algorithm; (f) Validation of relation type usefulness distributions (RTUD) and heterogeneous skip-gram (HS). (The embedding related parameter setting is: the number of walks per vertex $r = 10$; the walk length $l = 20$; the vector dimension $d = 128$; the context window size $ws = 10$)

Table 2: Measures of different cross-language citation recommendation algorithms

Algorithm	NDCG@10	NDCG@30	NDCG@50	P@10	P@30	P@50	MAP@10	MAP@30	MAP@50	MRR
EF	0.0176	0.0301	0.0384	0.0072	0.0060	0.0054	0.0101	0.0129	0.0140	0.0300
MT+LM	0.3404	0.3811	0.3966	0.1225	0.0573	0.0387	0.2563	0.2739	0.2777	0.4343
CF_I	0.0980	0.1034	0.1059	0.0330	0.0134	0.0086	0.0772	0.0793	0.0796	0.1290
CF_P	0.0041	0.0082	0.0108	0.0026	0.0024	0.0022	0.0017	0.0023	0.0026	0.0090
DW_c	0.2713	0.3060	0.3177	0.1037	0.0502	0.0336	0.2162	0.2348	0.2381	0.3053
DW_{all}	0.3606	0.4214	0.4416	0.1463	0.0735	0.0499	0.2679	0.2979	0.3033	0.4077
$LINE_c^{1st}$	0.2258	0.2557	0.2674	0.0854	0.0421	0.0289	0.1777	0.1927	0.1958	0.2628
$LINE_c^{2nd}$	0.1499	0.1730	0.1822	0.0572	0.0295	0.0205	0.1136	0.1241	0.1263	0.1894
$LINE_{all}^{1st}$	0.3534	0.4096	0.4302	0.1386	0.0691	0.0473	0.2671	0.2936	0.2990	0.4090
$LINE_{all}^{2nd}$	0.1047	0.1385	0.1564	0.0453	0.0284	0.0221	0.0663	0.0775	0.0811	0.1544
$N2V_c$	0.2724	0.3040	0.3153	0.1025	0.0489	0.0327	0.2183	0.2353	0.2383	0.3083
$N2V_{all}$	0.4651	0.5194	0.5354	0.1730	0.0809	0.0533	0.3661	0.3951	0.3999	0.5194
$M2V++_1$	0.0195	0.0214	0.0225	0.0052	0.0023	0.0015	0.0144	0.0147	0.0148	0.0312
$M2V++_2$	0.0015	0.0031	0.0045	0.0006	0.0006	0.0006	0.0006	0.0007	0.0009	0.0037
$M2V++_3$	0.0687	0.0933	0.1058	0.0308	0.0195	0.0150	0.0409	0.0481	0.0503	0.1070
$M2V++_{LN}$	0.0198	0.0273	0.0321	0.0084	0.0054	0.0045	0.0113	0.0130	0.0135	0.0389
$M2V++_{CA}$	0.0243	0.0335	0.0380	0.0107	0.0068	0.0052	0.0136	0.0156	0.0161	0.0451
HRLHG	***0.5034***†††	***0.5522***†††	***0.5664***†††	***0.1840***†††	***0.0832***†††	***0.0543***†††	***0.4033***†††	***0.4309***†††	***0.4353***†††	***0.5598***†††

Significant test: $^{\dagger}p < 0.01$, $^{\dagger\dagger}p < 0.001$, $^{\dagger\dagger\dagger}p < 0.0001$

For F_β, sum with a dumping factor (SDF) outperforms direct sum (DS). If we utilized SDF, a small λ could be superior than a great one. A possible explanation was that a small λ would penalize the dominated relation type usefulness probability to avoid overfitting. Generally, the algorithm performed better when more P^* became stabilize in training iterations.

To validate the effectiveness of RTUD and heterogeneous skip-gram (HS), we also compared the performance of our model without them. As Figure 5 (f) showed, when RTUD was removed (treating each relation type equally when we conducted the hierarchical random walk) and HS was replaced by ordinary skip-gram, recommendation performance declined significantly. It is clear that

RTUD and HS contribute to heterogeneous graph based random walk and recommendation performance significantly. More importantly, RTUD doesn't need any human intervention or expert knowledge.

Based on the comparison and analysis, we selected a component setting (k=3, F_Θ = LNC, F_β= SDF with λ = 0.2 and ε= 80% for RTUD training, while using heterogeneous skip-gram for vertex embedding), for further experiments with Baselines.

In skip-gram-based representation learning models, there were several common parameters (see Section 4.1). We also conducted a sensitivity analysis of HRLHG to these parameters. Figure 4 showed the their impacts on recommendation performance.

4.3 Comparison with Baselines

The cross-language citation recommendation performance results of different models were displayed in Table 2. Based on the experiment results, we had the following observations: (1) The proposed method significantly outperformed ($p < 0.0001$) other baseline models for all evaluation metrics. For instance, in terms of MAP@10, HRLHG achieved at least 10% improvement, comparing with all other 17 baselines. (2) The traditional models solely relied on one kind of information, i.e., machine translation based methods (EF, MT+LM) or citation relation based collaborative filtering approaches (CF_I and CF_P) cannot work as well as other network embedding based methods. (3) Although designed for homogeneous networks, by adding more types of vertexes and relations, the performance of DeepWalk (DW_{all}), LINE ($LINE^{1st}_{all}$ and $LINE^{2nd}_{all}$) and node2vec ($N2V_{all}$) had significant improvements over the ones using only citation networks (DW_c, $LINE^{1st}_c$, $LINE^{2nd}_c$ and $N2V_c$). This observation confirmed that heterogeneous information did enhance the models' representation learning abilities. (4) metapath2vec++, designed for heterogeneous information networks, didn't work well in the experiment. Even after applying learning to rank algorithms for integrating multiple metapath2vec++ models, the recommendation results were still not good. A possible explanation was that, for CCR task, no single metapath could cover the recommendation requirement. In addition, metapath based random walk was too strict to explore potential useful neighbourhoods for vertex representation learning. This observation also indicated that metapath2vec++ was depending on domain expert knowledge. If one cannot find the optimize metapath, the embedding performances were even worse than the homogeneous network representation learning models.

For each vertex type, HRLHG trained a relation type usefulness distribution. Based on the learned RTUD (available at project website), we can obtain the task-specified knowledge and improve the interpretability of proposed graph representation model. For instance, in the experimental CCR task, when a random walker reaches an English vertex, for the next move, the probabilities of $P_E \xrightarrow{c} P_E$ (an English paper cites another English paper) and $P_E \xrightarrow{h} K_E$ (an English paper has an English keyword) are higher than other relation types. This distribution navigates the walker to prefer to stay in the English repository rather than going back to the Chinese repository. By conducting the hierarchical random walk based on RTUD, the task specific knowledge can be further embedded into the representations learned by HRLHG.

In sum, for the CCR task, the proposed HRLHG method could automatically learn the relation type usefulness distributions for random walk navigation and the new method significantly outperformed the current text, homogeneous graph and heterogeneous graph embedding methods.

5 RELATED WORK

Citation recommendation aims to recommend a list of citations (references) based on the similarity between the recommended papers and user profiles or samples of in-progress text. For instance, He et al. [11] proposed a probabilistic model to compute the relevance score based on contexts of a citation and its abstract. Jiang et al. [12] generated a heterogeneous graph with various relations between topics and papers, and a supervised random walk was used for citation recommendation. From bibliographic viewpoint, Shi, Leskovec, and McFarland [24] developed citation projection graphs by investigating citations among publications cited by a given paper. Collaborative filtering algorithm can also be used for recommending citation papers [16]. However, all of the prior studies focused on monolingual citation recommendation and cannot be directly used for cross-language citation recommendation. Intuitively, translation-based models can be addressed for cross-language recommendation. Recently, word embedding is a powerful approach for content representation [18]. Mikolov et al. [19] transformed one language's vector space into the space of another by utilizing a linear projection with a transformation matrix W. This approach is effective for word translation, but the translation effect for scholarly text has not yet been demonstrated. Tang et al. [29] proposed bilingual embedding algorithms, which were efficient for cross-language context-aware citation recommendation task. However, they ignored the important citation relations in their work.

Network embedding algorithms, namely graph representation learning models, which aim to learn the low-dimensional feature representations of nodes in networks, are attracting increasing attention recently. Based on the techniques utilized in the model, we can briefly classify these algorithms into the following categories: the graph factorization based models, e.g., GraRep [3]; the shallow neural network based models, e.g., LINE [27]; the deep neural network based models, e.g., GCN [13]; and the random walk based method, e.g., DeepWalk [21], node2vec [9] and metapathvec++ [6]. Technically the random walk based models are also using a shallow neural network. The main difference between random walk based models are the random walk algorithms used for generating the vertex sequences from the graph. A potential problem for GraRep and GCN is the space complexity ($O(N^2)$), and the computational costs of these models can be too expensive to embed the large complex networks in the real world. For instance, in this CCR experiment (a 200,000 vertexes level graph), the memory requirement of GraRep/GCN is over 600G.

In this study, we address the CCR problems and propose a novel method HRLHG to learn a mapping of publication to a low-dimensional joint embedding space for heterogeneous graph. HRLHG belongs to the random walk based network embedding models. A hierarchical random walk is proposed to cope the task-specified problem on heterogeneous graph. To the best of our knowledge, few existing studies have investigated the graph embedding approach for cross-language citation recommendation problem.

6 CONCLUSION

In this paper, we propose a new problem: cross-language citation recommendation (CCR). Unlike existing scholarly recommendation problem, CCR enables cross language and cross repository recommendation. The proposed Hierarchical Representation Learning on Heterogeneous Graph (HRLHG) model can project a publication into a joint embedding space, which encapsulate both semantic and topological information. By training a set of relation type usefulness distributions (RTUD) on a heterogeneous graph, we propose a hierarchical two-level random walk: the global level is for graph schema navigation (task-specific); while the local level is for graph instance (task-independent) walking.

Unlike most prior heterogeneous graph mining methods, which employed expert-generated or rule-based ranking hypotheses to address recommendation problems, in a complex CCR graph, it can be difficult to exhaustively examine all of the potentially useful path types to generate metapaths. Furthermore, if a large number of random walk-based ranking functions are used, the computational cost can be prohibitive. Extensive experiments prove our hypothesis that the latent heterogeneous graph feature representations learned by HRLHG are able to improve cross-language citation recommendation performance (when comparing with 17 state-of-the-art baselines). In addition, the learned RTUD is able to reveal the latent task-specified knowledge, which is important to the interpretability of the proposed representation model.

In the future, we will validate the proposed method on other heterogeneous graph embedding based tasks, e.g., music recommendation or movie recommendation. Meanwhile, we will investigate more sophisticated method to generate RTUD. For instance, add personalization component to the algorithm, and enable personalized heterogeneous graph navigation for random walk optimization.

ACKNOWLEDGMENTS

The work is supported by the National Science Foundation of China (11401601, 61573028, 61472014), Guangdong Province Frontier and Key Technology Innovative Grant (2015B010110003, 2016B030307003), Health & Medical Collaborative Innovation Project of Guangzhou City, China (201604020003) and the Opening Project of State Key Laboratory of Digital Publishing Technology.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[2] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.

[3] Shaosheng Cao, Wei Lu, and Qiongkai Xu. 2015. Grarep: Learning graph representations with global structural information. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 891–900.

[4] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. ACM, 129–136.

[5] José Augusto de Azevedo, Joaquim João ER Silvestre Madeira, Ernesto Q Vieira Martins, and Filipe Manuel A Pires. 1990. A shortest paths ranking algorithm. In *Proceedings of the Annual Conference of Associazione Italiana di Ricerca Operativa: Models and Methods for Decision Support (AIRO'90)*. 1–8.

[6] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 135–144.

[7] Tao-yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. HIN2Vec: Explore Meta-paths in Heterogeneous Information Networks for Representation Learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1797–1806.

[8] Giorgio Gallo and Stefano Pallottino. 1986. Shortest path methods: A unifying approach. *Netflow at Pisa* (1986), 38–64.

[9] Aditya Grover and Jure Leskovec. 2016. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 855–864.

[10] Jiafeng Guo, Yixing Fan, Qingyao Ai, and W Bruce Croft. 2016. A deep relevance matching model for ad-hoc retrieval. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 55–64.

[11] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proceedings of the 19th international conference on World wide web*. ACM, 421–430.

[12] Zhuoren Jiang, Xiaozhong Liu, and Liangcai Gao. 2015. Chronological Citation Recommendation with Information-Need Shifting. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1291–1300.

[13] Thomas N Kipf and Max Welling. 2016. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv preprint arXiv:1609.02907* (2016).

[14] Ni Lao and William W Cohen. 2010. Relational retrieval using a combination of path-constrained random walks. *Machine learning* 81, 1 (2010), 53–67.

[15] Xiaozhong Liu, Yingying Yu, Chun Guo, and Yizhou Sun. 2014. Meta-Path-Based Ranking with Pseudo Relevance Feedback on Heterogeneous Graph for Citation Recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. ACM, 121–130.

[16] Sean M McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K Lam, Al Mamunur Rashid, Joseph A Konstan, and John Riedl. 2002. On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. ACM, 116–125.

[17] Donald Metzler and W Bruce Croft. 2007. Linear feature-based models for information retrieval. *Information Retrieval* 10, 3 (2007), 257–274.

[18] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[19] Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168* (2013).

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[21] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 701–710.

[22] Xiang Ren, Jialu Liu, Xiao Yu, Urvashi Khandelwal, Quanquan Gu, Lidan Wang, and Jiawei Han. 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, 821–830.

[23] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*. ACM, 285–295.

[24] Xiaolin Shi, Jure Leskovec, and Daniel A McFarland. 2010. Citing for high impact. In *Proceedings of the 10th annual joint conference on Digital libraries*. ACM, 49–58.

[25] Xiaoyuan Su and Taghi M Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in artificial intelligence* 2009 (2009), 4.

[26] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment* 4, 11 (2011), 992–1003.

[27] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. 2015. Line: Large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1067–1077.

[28] Jie Tang and Jing Zhang. 2009. A discriminative approach to topic-based citation recommendation. *Advances in Knowledge Discovery and Data Mining* (2009), 572–579.

[29] Xuewei Tang, Xiaojun Wan, and Xun Zhang. 2014. Cross-language context-aware citation recommendation in scientific articles. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 817–826.

[30] Chengxiang Zhai and John Lafferty. 2001. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 334–342.