Heterogeneous Information Network Embedding based Personalized Query-Focused Astronomy Reference Paper Recommendation

Xiaoyan Cai[1], Junwei Han[1], Shirui Pan[2], Libin Yang[1]

[1] School of Automation, Northwestern Polytechnical University,
Xi'an, Shaanxi 710072, China
[2]Center for Artificial Intelligence, University of Technology Sydney
Sydney, New South Wales 2007, Australia
E-mail: xiaoyanc@nwpu.edu.cn, jhan@nwpu.edu.cn, shirui.pan@uts.edu.au, libiny@nwpu.edu.cn

Received 18 June 2017

Accepted 5 January 2018

Abstract

Fast-growing scientific papers bring the problem of rapidly and accurately finding a list of reference papers for a given manuscript. Reference paper recommendation is an essential technology to overcome this obstacle. In this paper, we study the problem of personalized query-focused astronomy reference paper recommendation and propose a heterogeneous information network embedding based recommendation approach. In particular, we deem query researchers, query text, papers and authors of the papers as vertices and construct a heterogeneous information network based on these vertices. Then we propose a heterogeneous information network embedding (HINE) approach, which simultaneously captures intra-relationships among homogeneous vertices, inter-relationships among heterogeneous vertices and correlations between vertices and text contents, to model different types of vertices as vector formats in a unified vector space. The relevance of the query, the papers and the authors of the papers are then measured by the distributed representations. Finally, the papers which have high relevance scores are presented to the researcher as recommendation list. The effectiveness of the proposed HINE based recommendation approach is demonstrated by the recommendation evaluation conducted on the IOP astronomy journal database.

Keywords: Heterogeneous information, network embedding, personalized query-oriented reference paper recommendation, distributed representation.

1. Introduction

With the fast growth of astronomy research publication quantity, researchers might find it hard to cite appropriate and necessary astronomy reference papers. Traditional approaches usually retrieve relevant papers from search engines such as Google Scholar [1] or CiteSeer[2], based on certain keywords. Then researchers need to manually review them and decide which paper should be cited. However, it is labor-intensive and especially difficult for the beginning researchers. Thus, astronomy reference paper recommendation which recommends a list of astronomy reference papers that are relevant to the researchers' information need, is an essential technology to solve this problem.

Existing astronomy reference paper recommendation approaches fall into three categories: collaborative filtering (CF), content-based filtering (CBF) and graph-based approaches. CF makes astronomy reference paper recommendation by finding correlations among other researchers with similar research interests. CBF recommends an astronomy

[1] http://scholar.google.com
[2] http://citeseer.ist.psu.edu

reference paper based on words and/or topic features of an astronomy paper and the identity of a researcher. Graph-based approaches often consider astronomy reference paper recommendation as a link prediction problem and solve the problem using properties of random walks.

As deep learning techniques have been successfully applied in image processing fields [1-3], more and more researchers focus on applying deep learning techniques in natural language processing [4,5]. In this paper, we study heterogeneous information network embedding (HINE) problem, and propose to model and formulate astronomy reference paper recommendation based on HINE. We first construct a heterogeneous information network, in which different objects (i.e., query researcher, query text, papers and authors of papers) are act as vertices, objects are connected by edges in the network. Then we simultaneously consider linkage relationships among different heterogeneous vertices and text content associated with each vertex, to represent heterogeneous vertices in a unified vector space. Finally, the top ranked astronomy papers are recommended by measuring the relevance between the given query from an individual, astronomy papers and the authors of the papers based on the meaning representations. The main contributions of this paper are three-fold:

1) A heterogeneous information network is constructed to model different relationships among different objects (i.e., query researcher, query text, papers and authors of papers).

2) A Heterogeneous Information Network Embedding (HINE) approach is developed to map different heterogeneous objects into a unified vector space so that objects from different spaces can be directly compared.

3) A Heterogeneous Information Network Embedding (HINE) based astronomy reference paper recommendation approach is proposed and the optimization of the approach is described.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 constructs a heterogeneous information network and develops a heterogeneous information network embedding based astronomy reference paper recommendation approach. Section 4 presents experiments and evaluations. Section 5 concludes the paper.

2. Related Work

2.1. Graph-based Reference Paper Recommendation

Recent studies employed graph-based approaches to investigate the reference paper recommendation problem [6-10]. Strohman et al. [6] deemed reference paper recommendation as link prediction problem. They represented each paper as a vertex, the citation relationship as the link between vertices and a new paper as a vertex without any in-link and out-link. Zhou et al. [7] measured paper similarities by combining the author-paper graph, the paper-venue graph and the paper citation graph. Then they recommended reference papers by treating some known citations as positive labels and applying semi-supervised learning on the combined graphs. Gori et al. [8] proposed to recommend research papers by a random-walk based approach. Meng et al. [9] presented a personalized reference paper recommendation approach, which incorporated different kinds of information, such as content of papers, authorship and citation etc., into a unified graph model. Pan et al. [10] proposed an academic paper recommendation approach based on a heterogeneous graph containing various kinds of features.

2.2. Network Embedding

Network embedding is to map networks into low-dimensional spaces. Classical network embedding approaches include multi-dimensional scaling (MDS) [11], Laplacian Eigenmap [12], local linear embedding (LLE) [13] and Isometric Mapping (IsoMap) [14]. However, these approaches are not applicable to large scale networks due to computational complexity. Deep learning has been successfully applied in the field of computer vision and image processing [15-18]. Perozzi et al. [19] proposed DeepWalk by using local information from truncated random walks as input and learning latent representation of vertices in a network. Tang et al. [20] proposed a network embedding method called LINE to preserve the local and global network structures. Pan et al. [21] proposed TriDNR, a tri-party deep network representation model, to simultaneously learn network structure and vertex content. However, DeepWalk, LINE and TriDNR focus on homogeneous networks. PTE [22] extends the LINE to handle with heterogeneous network embedding problem, but PTE only leverages network structure, ignoring vertex content information. In this paper, we propose a heterogeneous information network embedding approach, which utilizes network structure and vertex content for all sources of vertices in the heterogeneous information network.

3. Heterogeneous Information Network Embedding based Personalized Query-Focused Reference Paper Recommendation Framework

3.1. Heterogeneous Information Network Construction

In this section, we first construct a heterogeneous information network containing both papers and authors as $G=<V_P,V_A,E_{PP},E_{AA},E_{PA}>$, where $V_P=\{p_i\}$ ($1\leq i\leq n$, n is the number of the papers) and $V_A=\{a_j\}$ ($1\leq j\leq m$, m is the number of the authors). $E_{PP}=\{e_{ij},p_i,p_j\in V_P\}$, $E_{AA}=\{e_{ij},a_i,a_j\in V_A\}$ and $E_{PA}=\{e_{ij},p_i\in V_P,a_j\in V_A\}$ correspond to the edges between papers, the edges between authors and the edges between papers and authors, respectively. Let c_{p_i} denote the text content associated with the paper p_i and $C_P=\{c_{p_i}\}$ ($1\leq i\leq n$), c_{a_j} denote the text content associated with the papers which are written by author a_j and $C_A=\{c_{a_j}\}$ ($1\leq j\leq m$).

3.2. Heterogeneous Information Network Embedding

The purpose of the heterogeneous information network embedding is to learn a low dimensional vector $\mathbf{v}_{p_i}(or\ \mathbf{v}_{a_i})\in \mathrm{R}^r$ (r is a small number) for each vertices p_i (or a_i) in the constructed heterogeneous information network, in order that vertices sharing common edges or with similar text content are close in the embedding space. We classify the pairwise relationships among the vertices in the constructed heterogeneous information network G into three categories: intra-relationship, inter-relationship and correlation relationship. In the following, we first describe different relationship modeling and then present heterogeneous information network embedding.

3.2.1. Different relationship modeling

(i) vertex relationship modeling

- intra-relationship modeling

(a) paper-paper relationship modeling

Motivated by DeepWalk [19], we construct a random walk generated from the heterogeneous information network G. We consider each random walk path $p_1\rightarrow p_5\rightarrow\cdots\rightarrow p_l$ as a sentence and each vertex p_i as a word in neural language models. Then we use DeepWalk algorithm to train Skip-Gram model on the generated random walk corpus S_1 of papers, and obtain a distributed vector representation for each paper vertex. Given a vertex p_i for all random walks $s\in S_1$, the objective function of paper-paper relationship modeling is defined as:

$$\begin{aligned}L_1&=\sum\nolimits_{i=1}^{n}\sum\nolimits_{s\in S_1}\log P(p_{i-b}:p_{i+b}\mid p_i)\\&=\sum\nolimits_{i=1}^{n}\sum\nolimits_{s\in S_1}\sum\nolimits_{-b\leq j\leq b,j\neq 0}\log P(p_{i+j}\mid p_i)\end{aligned}\tag{1}$$

where $p_{i-b}:p_{i+b}$ is a sequence of papers inside a contextual window of p_i with length b. The probability of observing neighboring vertices given current vertex p_i is calculated using the soft-max function as:

$$P(p_{i+j}\mid p_i)=\frac{\exp(\mathbf{v}_{p_i}^T\hat{\mathbf{v}}_{p_{i+j}})}{\sum\nolimits_{p=1}^{n}\exp(\mathbf{v}_{p_i}^T\hat{\mathbf{v}}_p)}\tag{2}$$

where $\hat{\mathbf{v}}_{p_i}$ is the output representation of the paper p_i.

(b) author-author relationship modeling

Similar to the paper-paper relationship modeling approach, the objective function of the author-author relationship modeling is formulated as:

$$\begin{aligned}L_2&=\sum\nolimits_{j=1}^{m}\sum\nolimits_{s\in S_2}\log P(a_{j-b}:a_{j+b}\mid a_j)\\&=\sum\nolimits_{j=1}^{m}\sum\nolimits_{s\in S_2}\sum\nolimits_{-b\leq k\leq b,j\neq 0}\log P(a_{j+k}\mid a_j)\end{aligned}\tag{3}$$

where S_2 is the generated random walk corpus of authors. The probability of finding neighboring authors $a_{j-b}:a_{j+b}$ given an author a_j is calculated as:

$$P(a_{j+k}\mid a_j)=\frac{\exp(\mathbf{v}_{a_j}^T\hat{\mathbf{v}}_{a_{j+k}})}{\sum\nolimits_{a=1}^{m}\exp(\mathbf{v}_{a_j}^T\hat{\mathbf{v}}_a)}\tag{4}$$

where $\hat{\mathbf{v}}_{a_j}$ is the output representation of the author a_j.

- inter-relationship modeling

(a) author-paper relationship modeling

To utilize the valuable relationship information between authors and papers, we collect all the papers written by one author. Then we use the author vector as input and simultaneously learn the input author vectors and output paper vectors. It can be formalized by the following objective function:

$$\begin{aligned} L_3 &= \sum_{i=1}^{m} \log P(p_{i-b} : p_{i-b} \mid a_i) \\ &= \sum_{i=1}^{m} \sum_{-b \le j \le b} \log P(p_j \mid a_i) \end{aligned} \tag{5}$$

The probability of observing the papers given an author a_i is defined as:

$$P(p_j | a_i) = \frac{\exp(\mathbf{v}_{a_i}^T \hat{\mathbf{v}}_{p_j})}{\sum_{p=1}^{n} \exp(\mathbf{v}_{a_i}^T \hat{\mathbf{v}}_p)} \tag{6}$$

(ii) vertex-content correlation modeling

(a) paper-content correlation modeling

We collect all the text content associated with one paper. Then the paper-content correlation relationship can be modeled as the contextual information of words within a document. The objective function is achieved by maximizing the following log-likelihood function:

$$\begin{aligned} L_4 &= \sum_{i=1}^{n} \log(w_{i-b} : w_{i+b} \mid p_i) \\ &= \sum_{i=1}^{n} \sum_{-b \le j \le b} \log P(w_j \mid p_i) \end{aligned} \tag{7}$$

The probability of observing contextual words $w_{i-b} : w_{i+b}$ given current vertex p_i is:

$$P(w_j \mid p_i) = \frac{\exp(\mathbf{v}_{p_i}^T \hat{\mathbf{v}}_{w_j})}{\sum_{w=1}^{L} \exp(\mathbf{v}_{p_i}^T \hat{\mathbf{v}}_w)} \tag{8}$$

where $\hat{\mathbf{v}}_{w_j}$ is the output representation of word w_j and L is the number of distinct words in the whole heterogeneous information network.

(b) author-content correlation modeling

We collect all the text content associated with papers which are written by one author, so the objective function is formulated as:

$$\begin{aligned} L_5 &= \sum_{i=1}^{m} \log(w_{i-b} : w_{i+b} \mid a_i) \\ &= \sum_{i=1}^{m} \sum_{-b \le j \le b} \log(w_j \mid a_i) \end{aligned} \tag{9}$$

Similarly, the probability of observing the words given an author a_i is defined as:

$$P(w_j \mid a_i) = \frac{\exp(\mathbf{v}_{a_i}^T \hat{\mathbf{v}}_{w_j})}{\sum_{w=1}^{L} \exp(\mathbf{v}_{a_i}^T \hat{\mathbf{v}}_w)} \tag{10}$$

3.2.2. Heterogeneous information network embedding

We propose a heterogeneous information network embedding (HINE) approach to jointly leverage the heterogeneous information network structure and the content information associated with each vertex in the network.

(i) Different relationship integration

We first merge intra-relationship models, inter-relationship models and correlation models into an integrated framework. The objective of the integrated framework is to maximize the following log likelihood function:

$$\begin{aligned} L = {} & (1-\alpha)\sum_{i=1}^{n} \sum_{s \in S_1} \sum_{-b \le j \le b, j \ne 0} \log P(p_{i+j} \mid p_i) \\ & + (1-\alpha)\sum_{i=1}^{m} \sum_{s \in S_2} \sum_{-b \le j \le b, j \ne 0} \log P(a_{i+j} \mid a_i) \\ & + (1-\alpha)\sum_{i=1}^{m} \sum_{-b \le j \le b} \log P(p_j \mid a_i) \\ & + \alpha \sum_{i=1}^{n} \sum_{-b \le j \le b} \log P(w_j \mid p_i) + \alpha \sum_{i=1}^{m} \sum_{-b \le j \le b} \log P(w_j \mid a_i) \end{aligned} \tag{11}$$

where α is the weight balancing network structure and text content information, b is the window size of sequence and w_j is the *j*th word in a contextual window. The first three terms in Eq.(11) indicate mutual reinforced information between papers and authors, the last two terms in Eq. (11) indicate the text information and author information will joint affect $\hat{\mathbf{v}}_{w_j}$, the output representation of word w_j, which will further propagate back to influence the input representation of a_i and p_j in the network. As a result, the vertex representation (i.e., the input vectors of vertices) will be enhanced by both network structure and text information.

(ii) model optimization

We use stochastic gradient descent (SGD) [23] to train the integrated framework in Eq.(11). However, computing the gradient in getting conditional probability in Eq. (2), Eq.(4), Eq.(6), Eq.(8) and Eq.(10) is the most time-consuming operation. To resolve this problem, we use hierarchical soft-max [24], which can reduce time complexity from $O(n+m)$ to $O((n+m)\log(n+m) + K\log(L))$, where K is the total number of words in the document content. The hierarchical soft-max builds three Huffman trees, one with author vertices as leaves, one with paper vertices as leaves and the other one with distinct words as leaves. So instead of enumerating all author vertices in Eq. (4) in each gradient step, we represent the path from the root to each leaf vertex a_i ($s_{20} \rightarrow s_{21} \rightarrow \cdots \rightarrow s_{2e}$), where s_{20} is the root of the tree and s_{2e} is the target vertex a_i, then we can compute as follows:

$$P(a_{i+j} \mid a_i) = \prod_{t=1}^{e} P(s_{2t} \mid a_i) \qquad (12)$$

$P(s_{2t} \mid a_i)$ can be further modeled by a binary classifier, which is defined as:

$$P(s_{2t} \mid a_i) = \sigma(\mathbf{v}_{a_i}^T \hat{\mathbf{v}}_{v_{s_{2t}}}) \qquad (13)$$

where $\sigma(x)$ is the sigmoid function and $\hat{\mathbf{v}}_{v_{s_{2t}}}$ is the representation of tree vertex s_{2t}'s parent. Similarly, we can use the above technique to compute conditional probability in Eq.(2), Eq.(6), Eq.(8) and Eq.(10).

3.3. Personalized Query-Focused Reference Paper Recommendation

In our work, we formulate the query as query author and query text, i.e., $q = [q_a, q_t]$, q_a represents the identity of a query researcher, q_t represents combination of the testing paper, q_a as a user and all the candidate reference papers as training papers. Given a query q, the proposed personalized query-focused reference paper recommendation aims to return top ranked training papers by measuring the relevance scores $\mathbf{r}_q = [\mathbf{r}_{qp_1}, \mathbf{r}_{qp_2}, \ldots, \mathbf{r}_{qp_l}]$ between the query q and all the training papers $p_i \in P(i = 1,2,\ldots,l)$. The input to the recommendation system is the word sequence of training and testing papers, the user of the training paper and all the authors of the testing papers. All these papers and authors are embedded into vectors based on heterogeneous information network embedding approach. Thus the relevance scores can be calculated as $\mathbf{r}_q = \mathbf{V}_{PR}\mathbf{v}_{q_t}^T + \mathbf{V}_{AR}\mathbf{v}_{q_a}^T$, where $\mathbf{V}_{PR} = [\mathbf{v}_{p_1}; \mathbf{v}_{p_2}; \ldots; \mathbf{v}_{p_n}]$ is the vector representation of training papers, $\mathbf{v}_{q_t}$ is the vector representation of the query text, $\mathbf{V}_{AR} = [\mathbf{v}_{a_1}; \mathbf{v}_{a_2}; \ldots; \mathbf{v}_{a_m}]$ is the vector representation of authors related to training papers, $\mathbf{v}_{q_a}$ is the vector representation of the user. Training papers are ranked according to the relevance scores, the top ranked ones are selected as the final reference paper recommendation list. Table 1 summarizes the whole process that determines the training papers relevance scores associated with the given query.

Table 1. Heterogeneous Information Network Embedding based Personalized Query-Focused Reference Paper Recommendation Algorithm.

```
Input: The heterogeneous information network $G = \langle V_P, V_A, E_{PP}, E_{AA}, E_{PA} \rangle$, the query text $q_t$, the user $q_a$,
the training papers and all the authors of the training papers, expected number of dimension of the vector representation $k$,
window size $b$, iteration number $T$, the walk length of Random Walk $len$ and the number of the recommended papers $Q$.
Output: Recommendation reference paper list.
1. Generate random walk corpus $S_1$ from the query text and training papers, generate random walk corpus $S_2$ from the user
   and all the authors of the training paper;
2. Generate a vocabulary binary tree $T_W$, generate a paper binary tree $T_P$ and generate an author binary tree $T_A$;
3. Get the initial input vector representation $\mathbf{v}_{a_i}$, $\mathbf{v}_{p_j}$ and output vector representation $\hat{\mathbf{v}}_{a_i}$, $\hat{\mathbf{v}}_{p_j}$ for the user and each author of
   the training paper, the query text and all the training papers, respectively;
4. Get the initial output vector representation $\hat{\mathbf{v}}_{w_j}$ for each word $w_j \in W$;
5. For iter=1,2,3,…,T do
       Fix $\hat{\mathbf{v}}_{w_j}$, solve Eq.(11) to update $\mathbf{v}_{a_i}$, $\mathbf{v}_{p_j}$ and $\hat{\mathbf{v}}_{a_i}$, $\hat{\mathbf{v}}_{p_j}$ '
       // Intra-relationship and inter-relationship
       Fix $\mathbf{v}_{a_i}$, $\mathbf{v}_{p_j}$, solve Eq. (11) to update $\mathbf{v}_{a_i}$, $\mathbf{v}_{p_j}$ and $\hat{\mathbf{v}}_{w_j}$;
       // correlation relationship
   End For
6. Calculate the relevance score $\mathbf{r}_q$ for the given query and rank the candidate reference papers according to $\mathbf{r}_q$;
7. Select top ranking $Q$ training papers as recommendation list.
```

4. Experiment and Evaluation

4.1. Experiment Setup

4.1.1 Experimental data

The experiments are conducted on the IOP astronomy journal database[3], it contains four journals, such as The Astronomical Journal, The Astrophysical Journal, The Astrophysical Journal Letters and The Astrophysical Journal Supplement Series. We have 71154 papers published from 1995 to 2016. We use 66942 papers published before 2016 as a training set, use remaining 4212 papers as a testing set and pre-process these papers by extracting its title and abstract. Then we build a heterogeneous information network with all the 71154 papers and authors of these papers (we deem an author of the testing paper as a user). Moreover, each vertex in the network associates with text content, for each paper vertex, it associates the title and abstract content corresponding to the paper, while for each author vertex, it associates with the title and abstract content corresponding to all the papers written by the author. In

[3] http://iopscience.iop.org/journals

this work, we deem a query consists of the title, abstract and an author of the query paper. Table 2 below shows the basic statistics of the dataset.

Table 2. Statistics of the IOP Astronomy Journal Database.

	Year	Papers	Authors
Training Set	1995-2015	66942	143537
Testing Set	2016	4212	27979[4]

4.1.2 Evaluation methods

Our ultimate aim is to recommend more relevant reference papers. We use three common metrics as follows:

Recall@*N*: It is defined as the percentage of original reference papers that appear in the top *N* recommended reference papers. Here we use *N*={20,40,60,80,100} to evaluate the proposed approach

Mean Average Precision (MAP): Recall@*N* ignores the exact ranking position, only considering the top *N* ranking results. MAP is a precision metric that emphasises ranking relevant papers higher, which can overcome the above disadvantage. Let T_p be the set of the testing papers. For a paper p_i in T_p, the correct reference paper set of p_i is *R*, and our proposed approach returns a reference list *B*. We consider the top 40 recommended papers in the ranking list, so |*B*|=40. The MAP is defined as:

$$MAP = \frac{1}{|T_p|}\sum_{p_i \in T_p} \frac{1}{|R|}\sum_{r_j \in R, rank(r_j)\neq 0} \frac{q(r_j)+1}{q(r_j)} \quad (14)$$

where $r_j \in R$ is a correct reference paper, $rank(r_j)$ is defined as the position of r_j in *B* if r_j is in *B*, otherwise $rank(r_j)$ is defined to be zero. $q(r_j)$ is set to the number of the correct reference papers which ranks higher than r_j.

Mean Reciprocal Rank (MRR): It measures how far from the top appears the first relevant reference papers. MRR is defined as:

$$MRR = \frac{1}{|T_p|}\sum_{p_i \in T_p}\left(\frac{1}{rank(p_{first})}\right) \quad (15)$$

where $rank(p_{first})$ is the position of the first relevant reference papers in the reference list *B*.

4.1.3 Comparison with other embedding based recommendation approaches

In order to evaluate the performance of heterogeneous information network embedding based recommendation approach, we compare it with the other four embedding based recommendation approaches:

- DeepWalk [19], which learns paper network representation by utilizing network structure information;
- Line [20], which preserves local and global network structure to learn paper network representation;
- Doc2Vec [25], which embeds variable length of text into a fixed length distributed vector using neural network models
- TriDNR[21], which simultaneously considers paper network structure and paper vertex content to learn paper network representation.

After obtaining network representation, the proposed personalized query-focused reference paper recommendation is performed based on the network representation.

4.2. Experiment Results and Evaluation

4.2.1 Performance of query-focused reference paper recommendation

In the first set of experiments, we only focus on the query text, ignoring the query author information, i.e., $q=[q_t]$. Table 3 below compares the performance of the four embedding based recommendation approaches and our proposed approach on the IOP astronomy journal database.

Table 3 shows that DeepWalk based recommendation approach and LINE based recommendation approach perform fairly poor. This can be mainly credited to the paper network structure is rather sparse and only contains limited information. As results illustrate, Doc2Vec shows better performance than LINE, because the text content contains rich information comparing to network structure. TriDNR based recommendation approach shows better performance than Doc2Vec, as it considers not only paper citation structure, but also the paper vertex content information. It is glad to see that the proposed HINE based recommendation approach shows the best performance, because it utilizes not only paper information, but also author information, besides it considers intra-relationships between papers and papers, between authors and authors, inter-relationships

[4] Among them, 17179 authors have published papers before 2016.

between papers and authors, as well as author-content correlation relationship and paper-content correlation relationship.

4.2.2 *Comparison with personalized and non-personalized query-focused reference paper recommendation*

We are also interested in studying whether personalized query-focused recommendation can provide more appropriate and individualized recommendation results to the users than non-personalized query-focused recommendation. We denote a non-personalized query by $q_1 = [q_t]$ and personalized query by $q_2 = [q_t, q_a]$, respectively. When a user who inputs the query for the personalized query-focused reference paper recommendation has not yet published any papers, the proposed approach will be reduced to non-personalized query-focused reference paper recommendation for the user, because the query information contains q_t only.

Table 3. Comparison of the Different Embedding based Approaches on the IOP Astronomy Journal Database.

	MAP	MRR	Recall@20	Recall@40	Recall@60	Recall@80	Recall@100
HINE	0.198	0.211	0.323	0.397	0.458	0.471	0.512
TriDNR	0.179	0.194	0.304	0.375	0.434	0.457	0.491
Doc2Vec	0.165	0.178	0.287	0.359	0.413	0.430	0.472
LINE	0.143	0.153	0.265	0.337	0.394	0.412	0.451
DeepWalk	0.120	0.134	0.241	0.312	0.379	0.395	0.438

Table 4.Comparison of Performance on Personalized and Non-Personalized Query-Focused Reference Paper Recommendation on the IOP Astronomy Journal Database .

	MAP	MRR	Recall@20	Recall@40	Recall@60	Recall@80	Recall@100
HINE, q_2	0.217	0.234	0.346	0.416	0.477	0.492	0.531
HINE, q_1	0.198	0.211	0.323	0.397	0.458	0.471	0.512

From Table 4, we can see that the performance of non-personalized recommendation is inferior to that of personalized recommendation. The personalized recommendation achieves a gain of about 5.9% on average. When we compare the correct recommended papers with regard to non-personalized and personalized recommendation approaches, we observe that the personalized recommendation approach can find more papers published by co-authors. We study the distinction of the top-60 recommendation results returned by HINE with q_1 and HINE with q_2. The overlap of them is about 69.38% of each. For the top-3 recommended results, the accuracy of HINE with q_2 is about 76.18% more than that of HINE with q_1.

4.2.3 *Illustration of Generated Recommendation Results*

Besides the above numerical analysis, we take an example to further illustrate the proposed recommendation approach and the limitations of existing embedding based recommendation approaches. The query paper is, Evidence for a Distant Giant Planet in the Solar System, which is one of the most read papers in The Astronomical Journal. Due to the page limit, we only list the top 5 retrieved reference papers of HINE with q_2, HINE with q_1 and TriDNR approaches in Table 5. We can see that the top 3 recommended reference papers returned by HINE with q_2 is the same with the correct reference papers, although the remaining two recommended reference papers returned by the approach is not in the top 5 correct reference papers, these two papers are in the correct reference paper list. As for the recommended reference papers returned by HINE with q_1, the top 2 results is the same with the correct reference paper results, though the other 3 recommended results are also in the correct reference paper list, the ranking positions of them are different from that of the correct reference papers. We attribute it to the HINE with q_2 approach incorporates user information in the query, while the HINE with q_1 does not do it. The top 5 recommendation results returned by TriDNR contain a paper which does not appear in the correct reference paper list, this is due to TriDNR only consider paper information, ignoring author information in heterogeneous information network construction.

5. Conclusion

In this paper, we propose a heterogeneous information network embedding (HINE) based approach to recommend personalized query-focused astronomy reference papers. Experiments demonstrate that personalized query-focused reference paper recommendation performs better than non-personalized query-focused reference paper recommendation on IOP astronomy journal database. In the future, we will extend the proposed approach to incorporate astronomy venue information.

Table 5. Comparison of the Different Embedding based Approaches on IOP Astronomy Journal Database.

Title of the Query Paper	Approaches	Top-5 System Generated Reference Papers
Evidence for a Distant Giant Planet in the Solar System	HINE with q_2	1. Retention of A Primordial Cold Classical Kuiper Belt in An Instability-Driven Model of Solar System Formation 2. Dynamical Measurements of The Interior Structure of Exoplanets 3. The Fate of Scattered Planets 4. Early Excitation of Spin-Orbit Misalignments in Close-In Planetary Systems 5. Discovery of A Candidate Inner Oort Cloud Planetoid
	HINE with q_1	1. Retention of A Primordial Cold Classical Kuiper Belt in An Instability-Driven Model of Solar System Formation 2. Dynamical Measurements of The Interior Structure of Exoplanets 3.Gas Giant Planets as Dynamical Barriers to Inward-Migrating Super-Earths 4.Jumping Neptune Can Explain the Kuiper Belt Kernel 5.The Fate of Scattered Planets
	TriDNR	1.Retention of A Primordial Cold Classical Kuiper Belt in An Instability-Driven Model of Solar System Formation 2. Dynamical Measurements of The Interior Structure of Exoplanets 3.Chaos in the Test Particle Eccentric Kozai-Lidov Mechanism 4.Tracking Neptune's Migration History Through High-Perihelion Resonant Trans-Neptunian Objects 5.Early Excitation of Spin-Orbit Misalignments in Close-In Planetary Systems

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No.61772429), the China Postdoctoral Science Foundation funded project (No. 2017M613205, 2017M623241), the Fundamental Research Funds for the Central Universities (No.3102016QD009, 3102016QD010) and the Basic Research Project Funded by Shenzhen Science and Technology Innovation Committee (No. 201703063000511).

References

1. G. Cheng, Z. Li, X. Yao, L.Guo and Z.Wei Remote sensing image scene classification using bag of convolutional features. *IEEE Geoscience and Remote Sensing Letters*,14(10): 1735-1739, 2017.
2. X. Yao, J. Han, D. Zhang and F. Nie, Revisiting Co-Saliency Detection: A Novel Approach Based on Two-Stage Multi-View Spectral Rotation Co-clustering, *IEEE Trans. on Image Processin*g, 26(7): 3196-3209, 2017.
3. J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey, *IEEE Signal Processing Magazine*, 35(1): 84-100, 2018.
4. Z.Cao, W.Li, S.Li, F.Wei and Y.Li. AttSum: Joint learning of focusing and summarization with neural attention, In *Proc. of 26th COLING Conf.* (2016).
5. L.Yang, X.Cai, S.Pan, H.Dai and D.Mu. Multi-document summarization based on sentence cluster using non-negative matrix factorization. *Journal of Intelligent and Fuzzy Systems* 33(3):1867-1879, 2017.
6. T.Strohman, W. Croft and D. Jensen, Recommending citations for academic papers, *in Proc.of 30th SIGIR Conf.*, (2007), pp.705-706.
7. D. Zhou, S. Zhu, K.Yu, X. Song, B. Tseng, H. Zha and C. Giles, Learning multiple graphs for document recommendations, In *Proc. of 17th WWW Conf.* (2008), pp.141-150.
8. M. Gori and A. Pucci, Research paper recommender systems: a random-walk based approach, in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*(2006), pp.778-781.
9. L. Pan, X. Dai, S. Huang and J. Chen, Academic paper recommendation based on heterogeneous graph, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, vol.9427 of the series Lecture Notes in Computer Science, (2015), pp. 381-392.
10. F. Meng, D. Gao, W. Li, X. Sun and Y. Hou, A unified graph model for personalized query-oriented reference paper recommendation, in *Proc. of 22nd CIKM Conf.* (2013), pp. 1509-1512.
11. T. F. Cox and M. A. Cox, *Multidimensional scaling* (CRC Press, 2000).

12. M. Belkin and P. Niyogi, Laplacian eigenmaps andspectral techniques for embedding and clustering, In *NIPS*, 14(2001), pp. 585-591.
13. S. T. Roweis and L. K. Saul, Nonlinear dimensionality nreduction by locally linear embedding, *Science*, 290(5500) (2000), pp.2323-2326.
14. J. B. Tenenbaum, V. De Silva, and J. C. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science*, 290(5500)(2000), pp.2319-2323.
15. D. Zhang, D. Meng, J. Han, Co-saliency Detection via A Self-paced Multiple-instance Learning Framework, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 39(5): 865-878, 2017.
16. J. Han, R. Quan, D. Zhang, F. Nie, Robust Object Co-Segmentation Using Background Prior, *IEEE Trans. on Image Processing*, 27(4): 1639–1651, 2018.
17. G.Cheng, P. Zhou and J. Han. Duplex metric learning for image set classification. *IEEE Transactions on Image Processing*, 27(1): 281-292, 2018.
18. C.Yao, J. Han, F. Nie, F. Xiao and Xu. Li. Local Regression and Global Information-Embedded Dimension Reduction. *IEEE Transactions on Neural Networks and Learning Systems*, doi: 10.1109/TNNLS.2017.2783384.
19. B. Perozzi, R. Al-Rfou, and S. Skiena, Deepwalk:Online learning of social representations, *in Proc. of 20th ACM SIGKDD Conf.,* (2014), pp.701-710.
20. J.Tang, W.Qu, M.Z.Wang, M.Zhang, J.Yan and Q.Z.Mei, Line: large-scale information network embedding, *in Proc. of 24th WWW Conf.* (2015), pp.1067-1077.
21. S.R.Pan, J.Wu, X.Q.Zhu, C.Q.Zhang and Y.Wang, Tri-party deep network representation, *in Proc. of 25th IJCAI Conf.* (2016), pp.1895-1901.
22. J.Tang, M.Qu, Q.Z.Mei, PTE: predictive text embedding through large-scale heterogeneous text networks, *in Proc.of 25th KDD Conf.* (2015), pp.1165-1174.
23. L. Bottou, Stochastic gradient learning in neural networks, in *Proc. of Neuro-Nımes*, 91(8), 1991.
24. F. Morin and Y. Bengio, Hierarchical probabilistic neural network language model, In Proc of the international workshop on artificial intelligence and statistics (2005), pp. 246-252.
25. Q.V. Le and T. Mikolov, Distributed representations of sentences and documents, *in Proc.of ICML Conf.* (2014), pp. 1188-1196.