Journal of Ambient Intelligence and Humanized Computing
https://doi.org/10.1007/s12652-021-03153-5

ORIGINAL RESEARCH

A content-sensitive citation representation approach for citation recommendation

Lantian Guo[1,2] · Xiaoyan Cai[3] · Haohua Qin[1] · Fei Hao[4] · Sensen Guo[3]

Received: 27 November 2020 / Accepted: 12 March 2021

Abstract

Citation recommendation systems mainly help researchers find the lists of references that related to their interests effectively and automatically. The existing approaches face the issues of data sparsity and high-dimensional in large-scale bibliographic network representation, which hinder the citation recommendation performance. To address these problems, we proposed a Content-Sensitive citation representation approach for Citation Recommendation, named CSCR. Firstly, the Doc2vec model is used to generate a paper embedding according to paper content. Then, utilizing the similarity between the paper content embeddings to select the assumed neighbours of the target paper, append the auxiliary links between target paper and its new neighbours in the bibliographic network. Thirdly, distributed network representation method is implemented on appended bibliographic network to obtain the paper node embedding, which can learn interpretable lower dimension embedding for paper nodes. Finally, the embedding vectors of these papers can be used to conduct citation recommendation. Experimental results show that the proposed approach significantly outperforms other benchmark methods in Normalized Discounted Cumulative Gain (*NDCG*) and the positive rate (*Recall*).

Keywords Citation recommendation · Citation network · Distributed network representation · Content-sensitive

1 Introduction

Along with the vigorous development of scientific research, scientific publications are increasing exponentially. However, in numerous of scientific papers, how to find out the suitable papers is an challenge for researchers. Recommendation system is an active and effective information retrieval method, which is a very effective approach to solve the problem of information overload. It has been widely used in many fields, e.g. item recommendation (Liu et al. 2020), social recommendation (Xiong et al. 2019), vacation rental recommendation (Li et al. 2020).

In order to address information overload problem in scientific research, citation recommendation system, a kind of scientific paper recommendation system that can recommend a small list of high-quality and suitable references (Dai et al. 2019) to accelerate researchers' work, has been paid to increasing attention (Chen et al. 2019) in recent years.

There are some citation recommendation works that explored the Collaborative Filtering (CF) and Content-based Filtering (CBF) technologies (Dai et al. 2018). CF-based citation recommendation usually considers that researcher's like-minded, where two researchers are considered

✉ Sensen Guo
guosensen@mail.nwpu.edu.cn

Lantian Guo
guolt0211@gmail.com

Xiaoyan Cai
xiaoyanc@nwpu.edu.cn

Haohua Qin
80649469@qq.com

Fei Hao
feehao@gmail.com

[1] School of Automation and Electronic Engineering, Qingdao University of Science and Technology, Qingdao 266044, China

[2] School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China

[3] School of Automation, Northwestern Polytechnical University, Xi'an 710072, China

[4] Department of Computer Science, University of Exeter, Exeter EX4 4QF, UK

Published online: 21 March 2021

like-minded when their citation list is similar (Yang et al. 2016). The CF algorithm has been employed in some e-commerce recommendations but cannot be effectively implemented in citation recommendation. The reason is that it is usually limited by data sparsity and scalability problem (Wang et al. 2020b). According to the item's content of researchers' citation list in the past, CBF-based citation recommendation usually provides the recommendations to researchers with similar research content (Guo et al. 2019; Wang et al. 2020a). However, due to mainly considering the content information, CBF falls into traditional information retrieval problems, such as semantic ambiguity (Dai et al. 2018). In the bibliographic dataset, there are various types of relationship information. However, both CF and CBF-based citation recommendation is difficult to deal with various types of relationship information and capture deep level features in bibliographic data.

Recently, with the rise of research on large-scale and heterogeneous information networks, graph-based recommendation approaches have been developed rapidly (Ali et al. 2020). The heterogeneous graph model can be constructed with the multiple types of links in the bibliographic data (Hu et al. 2021; Ma and Wang 2019). It can exploit various relationships among heterogeneous objects, e.g. co-author network, paper citation network, and paper's content, etc. (Guo et al. 2017; Ma and Wang 2019). Although the graph-based approaches can deal with various types of relationship information in bibliographic data, these methods faced with the sparsity and uncontrollable dim,ensions problems in lar,ge scale and heterogeneous information networks (Wang et al. 2020b).

To address these problems, network representation comes into use, e.g. Grover and Leskovec (2016), Tang et al. (2015), which encodes each node in a low-dimensional space while preserving the neighborhood relationship between node. Network representation approaches are mainly based on network structure, and model latent features of the nodes that capture neighborhood similarity and community membership (Pan et al. 2019). However, since existing citation recommendation methods based on the heterogeneous bibliographic network still rely on superposition or expansion of the adjacency matrix to represent heterogeneous networks, the dimensionality of the original information is inevitably increased. It leads to an increase in the burden of network representation learning. Neglecting the characteristic of bibliographic data will result in a long-tail problem and one-sided correlations in citation recommendation schema. One phenomenon caused by these problems is highly cited to be recommended continuously. For example, system is difficult to find some papers without citation relationships but with high content relevances.

Aiming at these problems, we proposed a content-sensitive citation representation approach for citation recommendation. In our approach, instead of having only network information to conduct node embedding, we also incorporate content information into the node embedding procedure. Firstly, we use the Doc2vec model to generate paper embedding according to paper content. Then, utilizing the similarity between the paper content embeddings to select the assumed neighbours of the target paper, append the link between the target paper and its new neighbours as auxiliary relationships in the citation network. Thirdly, a distributed network representation method is implemented on an appended citation network to generate the paper node embedding, which can jointly learn interpretable lower dimension vector for paper node. Experimental results show that the proposed approach significantly outperforms other baseline methods.

The main contribution of this paper include:

- A new content-sensitive citation representation approach is proposed, incorporating citation and content information. Unlike these approaches rely on superposition or expansion of the adjacency matrix to incorporate citation and content information, e.g. multi-layered graph (Cai et al. 2018; Guo et al. 2017), our method does not increase the dimension of the adjacency matrix. To some extent, it can solve the high-dimensional problem in citation and content information incorporating. Specifically, the content similarity between papers is seen as a basis for linking the papers without citation relationships but with high content relevances. More relevant papers with similar content are linked in the citation network, so that the data sparsity problem on the citation network would be alleviated.
- A series of extensive experiments are carried out on ACL Anthology Network (AAN) dataset to evaluate the effectiveness of our proposed method and implement parameter analysis. The experimental results illustrate that the performances of the citation recommendation method with our proposed CSCR outperform other methods relying on superposition or expansion of the adjacency matrix to incorporate citation and content information. These results further demonstrate that the CSCR method can capture more structure feature information in the citation network representation task. The rest of this paper is organized as follows. In Sect. 2, the related work on citation network representation and citation recommendation are reviewed. The Sect. 3 explain problem definition. The Sect. 4 details our citation representation method and citation recommendation framework. The Sect. 5 is the experimental results and analysis. The Sect. 6 concludes this paper.

2 Related work

2.1 Citation recommendation

Among the recommendation technologies used in citation recommendations, according to the implementation approaches, they mainly include the following three types: Collaborative Filtering (CF), Content-based Filtering (CBF), and Graph-based Approaches.

The CF-based citation recommendation method focuses on the relationship between researchers or the scoring matrix created by the citation network. Because the traditional collaborative filtering algorithm cannot effectively recommend the literature with fewer citations, and the collaborative filtering method has problems such as data sparseness and cold start (Wang et al. 2020b). Researchers make many improvements to reduce the recommendation errors caused by these problems. To solve the problem of cold start, Torres et al. (2004) integrated the collaborative filtering method and the recommendation result generated in the content-based filtering method, but the paper only used the term frequency-inverse document frequency (Term Frequency-Inverse Document Frequency, TF-IDF) to calculate content similarity. Liu et al. (2015) used citation context to mine the co-occurrence relationship between citations, and regard this co-occurrence relationship as associated information, which is added to the model as supplementary information to improve the CF-based citation recommendation accuracy.

CBF is also one of the most widely used and researched recommendation methods, which is derived from information filtering and retrieval methods. CBF method retrieves and matchs papers related to the citation recommendation target. The foundation of CBF-based citation recommendation is to analyze the paper content, e.g. paper title, abstract, the main body of paper and reader's attention or reading history, by content analysis method, e.g. bag-of-words (BOW) model (Ko 2012), the term frequency-inverse document frequency (TF-IDF) (Zhang et al. 2011), and Latent Dirichlet allocation (LDA) (Wei and Croft 2006). The content analysis method is to extract key features to form a representation of the paper content (Wang et al. 2020b), and further to recommend suitable papers with establishing descriptions. For example, Ding et al. (2014) used syntactic and semantic analysis techniques to analyzes the semantic similarity of the paper text for matching. Amami et al. (2016) uses the LDA topic model to model the paper text topic.

In addition, there are some CBF-based citation recommendations that use local or/and global contextual to rank the papers and achieve recommendation. He et al. (2010) combined language model, text similarity, topic model, feature dependence model, etc. to find suitable citation context. Livne et al. (2014) proposed a contextual citation recommendation exploiting various machine learning methods. In Livne et al. (2014), the paper similarity is integrated into the coupling of citation context to achieve the purpose of enriching the citation context of the paper content. At present, it is generally difficult to enhance the content feature extracting and the potential interest representation of users. Therefore, most of the citation recommendation systems are based on other algorithms (such as CF), and supplemented by CBF methods to solve the problem of cold start and inaccurate recommendations in the main algorithm.

Since heterogeneous objects and the relationship between them can be simply represented by a graph, the graph-based method is gradually paid more and more attention. The graph-based method can be easily applied to a bibliographic dataset containing multiple types of bibliographic network to construct a corresponding model and generate a list of recommended results (Cai et al. 2018). For example, West et al. (2016) proposed a hierarchical clustering algorithm to calculate the correlation between papers. However, since this method only uses citation relationship information, the constructed graph model has sparse and noise problems. To construct a heterogeneous network model to enhance the citation recommendation, Ren et al. (2014) proposed a cluster-based recommended citation framework named ClusCite, which uses the citation, venue, term, and author of papers. Links among nodes play a significant role in graph-based methods. Most of the graph-based methods treat citation recommendation as a citation link predication task (Yang et al. 2019). That assumption is impracticable. When the link prediction task is used to recommend potential citations, it is demand difference, that is the query content varies from person to person.

In order to solve these problems, the graph-based citation recommendation begins to regard a brief description of the query as the recommendation goal. Pan et al. (2015) used a graph-based similarity learning algorithm to achieve recommendation for query content. The graph-based model method uses two attribute information—-the citation relationship and the content of the paper. To reduce the data sparse problems and achieve better performance, Dai et al. (2018) proposed a model in a bipartite bibliographic network, which not only considers the content similarity of the paper topics, but also considers the community similarity between authors. To achieve effective citation recommendation, Guo et al. (2017) proposed a query-based personalized citation recommendation method with a fine-grained co-author relationships. A fine-grained co-author relationship is constructed by integrating co-author relationship network structure and publication content of author. On the basis of fine-grained co-author relationship, authors without citation

relationships but with high publication content relevances would be linked. However, Dai et al. (2018) and Guo et al. (2017) focus on representation optimization of co-author relationship network, and overlooked representation optimization of the citation network.

2.2 Bibliographic network representation

The adjacency matrix is the most commonly used graph representation form in graph-based information retrieval and recommendation task. It is employed to indicate whether there is a connection between two nodes. However, since most of the nodes in the dataset are not related directly, the adjacency matrix is extremely sparse as well as not conducive to calculate and store. The existing citation recommendation approaches usually rely on superposition or expansion of the adjacency matrix to incorporate citation and content information. For example, Guo et al. (2017) utilized a three multi-layered graph to achieve citation recommendation. The three multi-layered graph contains three types of the entity including author, paper, and content (keyword). Cai et al. (2018) utilized another kind of three multi-layered graph to achieve citation recommendation. The three multi-layered graph is composed of three types of the entity including author, paper, and venue. The problems of these representation methods are that the dimension of the adjacency matrix increases as the number of nodes increases. Even when incorporating new node types, the dimensionality of the adjacency matrix directly increases significantly. Liu et al. (2016) found that the network representation constructed by entity relationships is vulnerable to data sparsity. The above situations will cause the data sparsity problem to be more serious.

With the success in lots of link prediction and classification tasks, network representation has been paid tremendous attention to the researchers (Wang et al. 2020b). For example, Perozzi et al. (2014) proposed a DeepWalk model, which generates sequence data that similar to textual context with a random walk procedure beginning from a node in the network graph, and then to obtain latent representations of vertices in the network by skip-gram. Tang et al. (2015) proposed the concepts of first-order proximity and second-order proximity. With these two concepts, they optimized an objective of preserving both global and local network structures, Thus, that method is very suitable for large-scale data processing. Grover and Leskovec (2016) proposed a higher applicability model that similar to DeepWalk. It improves the strategy of random walk that reaches a balance between Depth-first Sampling(DFS) and Breadth-first Sampling(BFS). Besides, account local and macro information is also taken into account by this model. In order to simplify computational storage without manually extracting features, Zhang et al. (2018) represent nodes in a network by dense, low-dimensional, and real-valued vectors. This method can project heterogeneous information into the same low-dimensional space and facilitate large-scale network representation.

Although these approaches do not increase the dimension of the adjacency matrix, they rely on the superposition of the adjacency matrix to incorporate citation and content information in the node learning procedure. To some extent, it can solve the high-dimensional problem in citation and content information incorporating, however, those problems still exist and seriously hinder the improvement of bibliographic network representation and citation recommendation performance. Especially, some nodes in a network are not related directly, but they may have implied relevance. For example, in the citation network, some papers have no citation relationships but the content relevances are high.

3 Problem definition

In this paper, the citation network in the bibliographic data can be modeled as a paper–paper relationship graph. Let $G(V, E)$ be a directed weighted graph. We set $V = V_p$ is the paper vertex set, $V_p = \{p_i\}(1 \le i \le n, \cdot n$ is the total number of papers. $\mathrm{E} = \{E_{pp}\}$ is the edge set, $E_{pp} = \{e_{ij}, p_i, p_j \in V_p\}$. We aim to link some papers without citation relationships but with high content relevances. So, the data sparsity problem can be to alleviate as well without increasing the dimensionality of the network. Moreover, papers with high content relevances can be found more easily during the recommendation process, which can promote citation recommendation performance.

Therefore, a problem is how to find a paper with high content relevance. Our solution is: firstly, the Doc2vec model is used to generate a paper content embedding vector. Then, the similarity between the paper content embeddings is calculated by the paper content vector. After that, a more significant problem is how to utilize these similarities. Our solution is to link relevant papers as a complement for the citation network. The relevance papers are regarded as the assumed neighbours of the target paper, and appended the auxiliary links between the target paper and its new neighbours in the bibliographic network. Set the appended links among papers as the edge set A_{pp}, the appended citation network as the edge set EA_{pp}p. According to the principle of our method, the appended citation network can be denoted as $EA_{pp} = \{E_{pp}\} + \{A_{pp}\}$. Our proposed method is to find

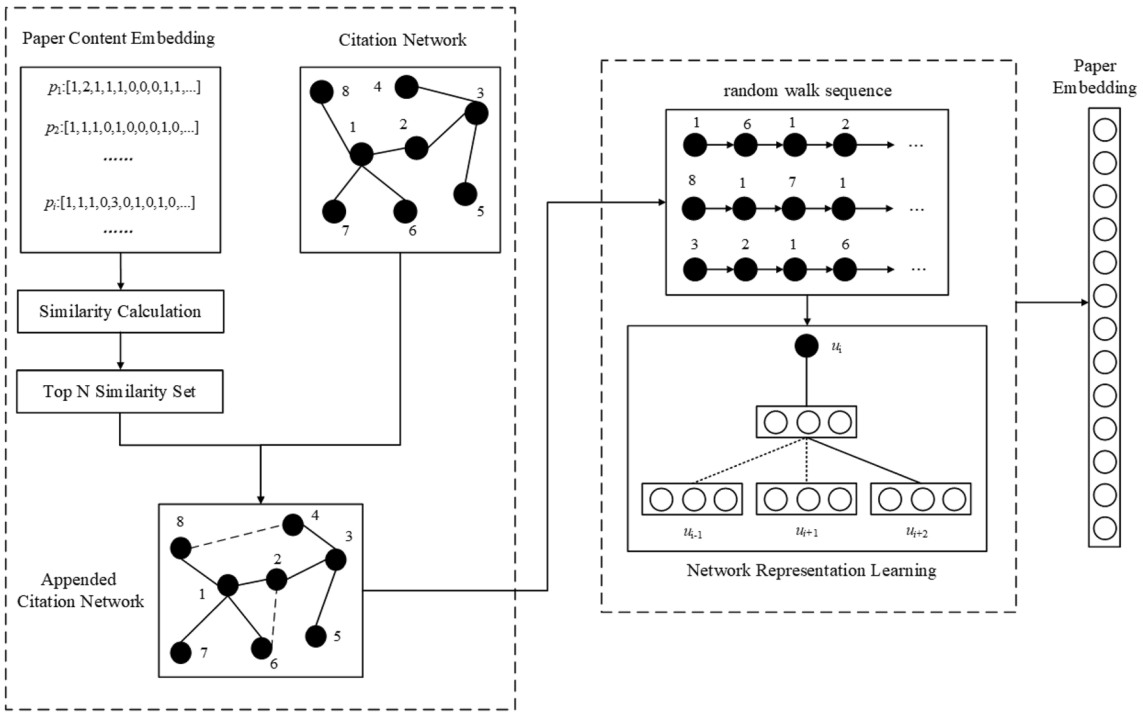

Fig. 1 Framework of the CSCR method

the A_{pp}, and further obtain the EA_{pp}. The distributed vector representation of each paper node can be obtained on EA_{pp}. We can use the vector representations of these papers to conduct citation recommendations.

4 Proposed method

4.1 Appended citation network

As shown in Fig. 1, the framework of the CSCR method in this paper includes regenerating the appended citation network part and paper embedding part. Based on these parts, we can learn the feature representation vectors of paper nodes paper embedding, which contains not only network structure information but also vertex content information. The generated paper embedding can be applied to achieve the citation recommendation.

a) **Paper content similarity calculation**: In our method, the paper content employs the paper's title and abstract. Each paper content is associated with a paper vertex in the citation graph. As the architecture of the proposed method, the Doc2vec (Lau and Baldwin 2016) model is employed to learn a paper content embedding vector. Then, utilizing the similarity between the paper content embeddings to select the assumed neighbours for each paper. The content similarity of paper content embeddings can be calculated as follows:

$$\begin{aligned} \text{content}(p_i, p_j) &= \frac{\left\| V_{PT}(p_i) \cdot V_{PT}(p_j) \right\|}{\left\| V_{PT}(p_i) \right\| \cdot \left\| V_{PT}(p_j) \right\|} \\ &= \frac{\sum_{i=1}^{\text{size}} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{\text{size}} x_i^2} \cdot \sqrt{\sum_{i=1}^{\text{size}} y_i^2}} \end{aligned} \tag{1}$$

where V_{PT} is the paper text embedding vector in candidate paper set P. x_i, y_i are the components of the paper vectors $V_{PT}(p_i)$, $V_{PT}(p_j)$. We can set the matrix S as the text content similarities among papers, which the i-th row of matrix S is the content vector representation of p_i.

b) **Appending the auxiliary link in citation network**: After paper content similarity calculation, a more significant problem is how to utilize these similarities. Our solution is to link relevant papers as a complement for the citation network. In our proposed method, the top n papers with the greatest similarity are regarded as rel-

evant papers. We set the top n papers with the greatest similarity as Top N Similarity Set (*TNSS*). The text content similarities among papers S is utilized to select the assumed neighbours of the target paper p_i, obtaining a *TNSS* set from $S(p_i)$. Regarding *TNSS* set as its new neighbours, the auxiliary links can be appended between the target paper and *TNSS* set in the bibliographic network. Appended auxiliary link set A_{pp}, which are the link between target paper and *TNSS*, can be obtained. Further, appended citation network EA_{pp} can be obtained by $EA_{pp} = E_{pp} + A_{pp}$.

4.2 Paper embedding in appended citation network

In our method, we use typical network representation methods to achieve paper embedding on citation bibliographic network, e.g. DeepWalk (Perozzi et al. 2014), LINE (Tang et al. 2015). Take Deepwalk as an example, a basic principle is to simulate the text generation process by constructing random walk paths of nodes on the network. It first samples a random walk sequence of nodes, and then use the Hierarchical Softmax and Skip-gram models to the context windows in the sequence of nodes. The node pairs are probabilistically modeled to maximize the likelihood probability of the random walk sequence, and finally use a random gradient to update parameters of the model. In the random walk, transitivity is an axiom of logic and mathematics. All in all, the above procedure can be used for the paper embedding on regenerated paper–paper citation network. Its objective function of embedding learning can be expressed as follow:

$$L_u = \frac{1}{|U|} \sum_{i=1}^{|U|} \sum_{i-t \le j \le i+t, j \ne i} \log \mathcal{P}(v_j|v_i) \tag{2}$$

where $\mathcal{P}(v_j|v_i)$ can be expressed as:

$$\mathcal{P}(v_j|v_i) = \frac{\exp(v_j^{'} \cdot v_i)}{\sum_{v_j' \in V} \exp(v' \cdot v_i)} \tag{3}$$

where U is the number of nodes in the paper-paper citation network. $\mathcal{P}(v_j|v_i)$ is the transition probabilities between two node v_j and v_i. It is defined by a softmax function to generate embedding of node. According to our above principle, the target paper and its corresponding *TNSS* would be linked, which can be further regarded as candidate paths in the random walk sequence. As a result, in the random walk sequence, two similar papers that never had a citation relationship may be discovered. According to network representation methods, the paper node embedding PE embodying the information in appended citation network EA_{pp} can be obtained. The newly generated paper embeddings PE embody not only citation network structure information but also content information.

4.3 Citation recommendation

Algorithm 1 Citation Recommendation with CSCR-DeepWalk

Input: Query: Query information from user: a keywords set of the submitted text q.
Matrix: Citation Network E_{pp}.
Parameters Setting: the dimension of the feature vector represention d, embedding size n, the width of the context sliding window c, Length of walking path l;

Output: Ranking scores of candidate paper set P, $R(q, P)$

$V_{PT} \leftarrow$ Doc2vec(P) //get the text embedding for each paper in P,
for a paper p_i in P **do**
  obtain $S(p_i)$ on Eq. 1 //calculate the text similarity with other papers in P
  obtain $TNSS$ from $S(p_i)$ //find top n set for p_i,
  obtain $A_{pp}(p_i)$ with $TNSS$ //append the link between p_i and $S(p_i)$, .
end for
$EA_{pp} \leftarrow \{E_{pp}\} + \{A_{pp}\}$. //obtain appended Citation Network
$V_{PE} \leftarrow$ DeepWalk(EA_{pp}) // obtain paper nodes embedding by DeepWalk
Initialize query information q.
$r_q \leftarrow$ sim(q,V_{PE}). //According to V_{PE} calculate the score of the relevance of each paper to the query r_q.
$R(q, P) \leftarrow Sort(r_q)$. //Sort the paper score
Return $R(q, P)$

A query-based citation recommendation framework is employed in this paper. Given a query manuscript q from a user, and a candidate paper set (P). The citation recommendation problem is to learn a recommended score list $R(q, P)$. The learned score list $R(q, P)$ indicates the matching degree of each candidate paper $p \in P$ specifically for the query q.

In our citation recommendation scenario, the query information can be formally expressed as q=[q_w], the keyword set including the description text is q_w. In the course of recommendation, after a query information q is initialized, the recommendation model can measure correlation between query information q and each candidate paper $p_i \in P, (i = 1, 2, \cdots, n)$. The recommendation result is generated by sorting the paper nodes in the training set according to the correlation values $r_q = [r_{qp_1}, r_{qp_2}, \cdots r_{qp_n}]$. The highest ranked papers is selected a set of most relevant ones and returned as a citation recommendation list.The relevance r_{qp_i} between query information q and each candidate paper p_i can be determined by the vector representation similarity between q and p_i. That is, r_{qp_i} can be calculated by the function 4.

$$\begin{aligned}\mathrm{sim}\left(q, V_{PE}(p_i)\right) &= \frac{\left\|q \cdot V_{PE}(p_i)\right\|}{\|q\| \cdot \left\|V_{PE}(p_i)\right\|} \\ &= \frac{\sum_{i=1}^{\mathrm{size}} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{\mathrm{size}} x_i^2} \cdot \sqrt{\sum_{i=1}^{\mathrm{size}} y_i^2}}\end{aligned} \tag{4}$$

where x_i, y_i are the components of the paper vectors q and $V_{PE}(p_i)$ respectively. $V_{PE}(p_i)$ is learned paper node embedding by the CSCR method.

5 Experiments

In this section, a series of experiments is conducted. We first introduce the pre-processing of the Association of Computational Linguistics (ACL) Anthology Network (AAN) dataset. Then, to evaluate the performance of the proposed approach, an experiment is conducted to compare the proposed approach with baseline methods in terms of *Recall* and *NDCG* metrics. Finally, to observe and analyze the impact of different parameter settings on the recommendation results, an experiment is also conducted to obtain the accuracy and effectiveness of the proposed approach. Moreover, detailed analysis and discussion of the experimental results are conducted on these experiments.

5.1 Dataset

The AAN 2013 Release dataset (Radev et al. 2013)[1] was used to evaluate our method. This dataset contains a collection of papers published in most ACL venues. Specifically, this release contains 21,236 papers published from the year 1965 to 2013. Moreover, it provides information such as the citation list, the content of the paper, author, year of publication, title, and venue. The preprocessing for each paper in this dataset consists of four steps: (1) extracting abstracts and titles. (2) deleting words with no more than 3 characters. (3) deleting stop words. (4) using porter stemmer to stem the remaining words.

To reduce noise, we also deleted papers with no citation relationship and appeared less than 10 times in the dataset. After preprocessing, only 12,504 of 21,236 papers remained. Among these papers, 11,129 papers are published before 2013. Our experiment takes the 11,129 papers as the candidate papers set (training set). The remaining 1,375 papers published in 2013 are used as a test set. The title and abstract of the paper is seen as its content, which is used to learn a content vector representation of the paper. In the citation recommendation framework, each query term is simulated by the title and abstract of a paper. The actual citation list of each paper in the test set was adopted as a groundtruth result for citation recommendation.

5.2 Evaluation metrics

Recall (Liu et al. 2011) and Normalized Discounted Cumulative Gain (*NDCG*) (Totti et al. 2016) have been widely used in the field of statistical classification and information retrieval, and here we use it to evaluate the quality of the recommended results and the accuracy of our method. The formula of *Recall* and *NDCG* can be expressed as Eqs. 5 and 6.

$$Recall@L = \frac{1}{M}\sum_{i=1}^{M}\frac{R(p) \cap T(p)}{T(p)} \tag{5}$$

where M is the total number of queries (test papers) and L is the length of the recommended list.

$$NDCG@L = \frac{1}{M}\sum_{i=1}^{c}\left(\left(\sum_{j}^{L}\frac{2^{r_i}-1}{\log_2(j+1)}\right)/IDCG@N\right) \tag{6}$$

where M is the total number of queries (test papers) and L is the length of the recommended list.

5.3 Performance comparison

To validate the effectiveness, we implement CSCR method in citation recommendation framework, and compared it with three baseline methods. Moreover, since distributed network representation is employed in our CSCR method, we also adopt three different distributed network representation methods to construct three types of CSCR method, and compare their performances in citation recommendation framework.

For the three types of CSCR method, there are two common parameters: the size of Top N Similarity Set *TNSS* and Embedding size d. *TNSS* indicates the number of papers with the largest similarity n to the target paper. Embedding size d is the dimensionality of distributed graph feature representation. For a fair comparison, in the three types of CSCR method—-CSCR-LINE, CSCR-GraRep, and CSCR-DeepWalk—-the embedding size setting d is 75 and Top N Similarity Set *TNSS* setting n is 2 respectively. As the experimental parameters analysis subsection discussed (Sect. 5.4), when the embedding size setting d is set to 75 and Top N Similarity Set *TNSS* in CSCR-DeepWalk is set to 2 respectively, the recommendation performance achieve a relatively high level. Thus, these setting values are employed in this experiment.

[1] http://clair.eecs.umich.edu/aan/downloads/.

Table 1 *Recall* and *NDCG* Performance comparison in terms of different methods

Top-*N*	25		50		75		100	
Metric	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*	*Recall*	*NDCG*
Link-PLSA-LDA	0.121	0.272	0.163	0.284	0.217	0.315	0.251	0.342
PWR	0.151	0.259	0.227	0.282	0.277	0.335	0.315	0.353
PWMP	0.197	0.337	0.277	0.350	0.318	0.356	0.379	0.358
CSCR+LINE	0.205	0.322	0.291	0.347	0.317	0.358	0.403	0.363
CSCR+GraRep	0.214	0.336	0.301	0.354	0.325	0.371	0.403	0.381
CSCR+DeepWalk	**0.229**	**0.360**	**0.311**	**0.379**	**0.366**	**0.390**	**0.419**	**0.392**

Bold indicates that the CSCR+DeepWalk method obtains the best performance, comparing with other two types of CSCR based method (CSCR+LINE and CSCR+GraRep)

- **Baseline Approach 1 (Link-PLSA-LDA)**. It is an unsupervised topic model that incorporates the citation link relationship based on the LDA topic model (Nallapati and Cohen 2008). It can better capture the probabilistic distribution information of the paper topic that implies the potential link relationship between the papers. So that, this model can predict the citation relationship, and achieve the recommendation purpose. In this experiment, the recommended citation list is sorted according to the relevance between their text topic distribution and the query text. The number of Link-PLSA-LDA topics is also set to 300.
- **Baseline approach 2 (PWR)**. As Literature (Guo et al. 2017; Pan et al. 2015), a two-layer graph model is established for incorporating citation relationships and paper content information. This two-layer graph contains paper–word subgraph and paper–paper subgraph. We call it PW graph. A random walk-based algorithm is applied on PW graph to learn paper ranking and achieve citation recommendation. We call it PWR method.
- **Baseline Approach 3 (PWMP)**. This method also adopts the above two-layer PW graph to model citation relationship and paper content information. Different from PWR method, a Meta-Path-based (Liu et al. 2014) algorithm is applied on PW model to learn paper ranking and achieve citation recommendation. We call it PWMP method.
- **Our Approach 1 (CSCR-LINE)**. Following the CSCR framework, CSCR-LINE first appends auxiliary link between papers with similar content and then learn paper embedding with distributed network feature representation model. In CSCR-LINE, the distributed graph feature representation unit employs the LINE model (Tang et al. 2015). LINE model uses two independent stages to learn the characteristic representation of the d dimension, and designs an optimized objective function that considers the first-order and second-order transition information between nodes. In the first stage, it learns by sampling the direct neighbors of the node by BFS (Breadth-first Search); in the second stage, it is limited to sampling nodes at a 2-hop distance from the target node to learn. In the training process, LINE adopts a link-based negative sampling optimization algorithm, and the parameters are set as follows: the number of negative examples is 5; the mini-batch size of stochastic gradient descent is 1; the initial learning rate is 0.025.
- **Our Approach 2 (CSCR-GraRep)**. In CSCR-GraRep, the distributed network feature representation unit employs GraRep model (Cao et al. 2015). GraRep is a representation model based on matrix factorization. The difference with DeepWalk and LINE is that GraRep is based on a probability transition matrix and takes into account higher-order context information. It uses the SVD matrix factorization training model to obtain network representation. In CSCR-GraRep, the maximum number of transition steps in the model is set to 6.
- **Our Approach 3 (CSCR-DeepWalk)**. In CSCR-DeepWalk, the distributed network feature representation unit employs DeepWalk model (Perozzi et al. 2014). The DeepWalk model simulates the text generation process by constructing random walk paths of nodes on the network. It first samples a random walk sequence of nodes, and then use the Hierarchical Softmax and Skip-gram models to the context windows in the sequence of nodes. The node pairs are probabilistically modeled to maximize the likelihood probability of the random walk sequence, and finally use a random gradient to update parameters of the model. Basic parameters setting follows the experiments in Grover and Leskovec (2016). Context window size for neural network model training in random walk sequence is set to 10, which controls the number of sampling neighbor nodes for the source node in the biased random walk. The maximum path length of a biased random walk is set to 80. The remaining parameters setting follows other baseline methods in these experiments. The results of three types of CSCR based methods compared with the Link-PLSA-LDA, PWR, and PWMP methods are shown as Table 1. With the length of recommendation lists increasing gradually, the values of *Recall* and *NDCG* of these three methods increase as well since more papers are recommended as n increases.

It's not difficult to observe that with the recommendation list get longer gradually, the *Recall* and *NDCG* performances enhanced on all methods since a longer of recommendation list means more recommended papers, which leads to more matched papers. Further, the following two observations are obtained.

Observation 1: Table 1 reveals a significant conclusion: three types of CSCR based method—-CSCR-DeepWalk, CSCR-GraRep and CSCR-LINE outperform Link-PLSA-LDA, PWR, and PWMP in terms of *Recall* and *NDCG*. It can be concluded that the link and content incorporating approach employed by the CSCR method is superior to the link and content incorporating approaches in conventional methods.

Observation 2: Another significant observation is that under the same conditions, CSCR-DeepWalk obtains more accurate recommendation results than CSCR-GraRep and CSCR-LINE in terms of *Recall* and *NDCG*. It can be concluded that in the CSCR framework, the DeepWalk-based distributed graph feature representation approach is more suitable than GraRep and LINE based approaches. That is to say, the DeepWalk-based paper node embedding approach can capture more structure feature information in the citation network representation task. And it also revealed that DeepWalk is more suitable than GraRep and LINE in the CSCR-based citation recommendation task.

5.4 Experimental parameters analysis

In this section, two parameters inside our model are analyzed: the size of Top N Similarity Set *TNSS* and Embedding size *d*.

(a) Top N similarity set *TNSS*

In our proposed method, the different *n* settings for Top N Similarity Set *TNSS* determine different supplementary link scale among papers, and inevitably different paper embedding results would be generated. It can be derived that the different *n* settings will have different effects on the recommendation results. For the value setting of *n*, there is no standard. Thus, to evaluate the impact of different sizes of Top N Similarity Set *TNSS* and find out the optimized values in our proposed method, massive experiments with different sizes of recommendation results would be conducted repeatedly. *CSCR* is performed with various settings of $n \in (1, 9)$, where the step length is 1. Fig. 2 shows *Recall* and *NDCG* performance of *CSCR* assigned different *TNSS* values respectively. Further, the following observations are obtained.

As shown in sub-figures 2a and b, *Recall@75* and *Recall@100* show an upward trend as *n* increases from 1 to 2, and then reveal a downward trend when *n* reachs 3. The better *Recall* values are achieved when *n* is 2. Similarly, as shown in sub-figures 2c and d, when *n* increased from 1 to 3, the *NDCG@75* and *NDCG@100* also showed an upward trend. After *n* is 3, *Recall@75* and *Recall@100* show a volatile downward trend. The reason for the downward trend is: since we calculated the similarity for incorporating the auxiliary relationships into the citation network and get the value of *TNSS*, a larger value of *n* will contain some information that irrelevant to the target paper and it will affect the accuracy of the recommendation results.

It is also can be observed that when *n* is set to 2, the best values of *Recall* were obtained, whereas the results are the worst when *n* is set to 1. When *n* is set to 3, the values of *NDCG* are the best, whereas the results were the worst when *n* is set to 9. In a word, when the setting of *n* is too large or too small, the performances will deteriorate. These related experimental results indicated that the effect of the recommendation results is closely related to the value of *n*. By comprehensive consideration on the *Recall* and *NDCG* performance, we finally set 2 as preferable value for *n* in our experiment.

(b) Embedding size *d*

We also conducted extensive experiments to analyze the effect of the embedding size *d*. To setting the value of *d*, there is also no standard. Thus, to evaluate the impact of different embedding size *d* and find out the optimized values in our proposed method, massive of recommendation experiments with different embedding size *d* would be conducted repeatedly. Specifically, *CSCR* is performed with eight candidate values of *d* respectively, i.e. 15, 25, 50, 65, 75, 100, 125, 150, 175. Fig. 3 shows *Recall* and *NDCG* performance of *CSCR* assigned different *d* values respectively. Further, the following observations are obtained.

As shown in sub-figures 3a and b, both *Recall@75* and *Recall@100* show an overall upward trend as *d* increases from 15 to 75. Although there is a slight drop when *d* increases from 50 to 65, *Recall@75* and *Recall@100* show a significant increase when *d* increases from 65 to 75. Then, both *Recall@75* and *Recall@100* tend to decrease as *d* increases from 75 to 175. As shown in sub-figures 3d and d, both *NDCG@75* and *NDCG@100* show an upward trend as *d* increases from 15 to 75, and then tend to decrease as *d* increases from 75 to 175. The reason for the downward trend is: since value of *d* control the dimensionality in the embedding learning procedure, too high dimensionality is easy to cause overfit that impacts the learning performance.

It is also can be observed that when *d* is set to 75, the best values of both *Recall* and *NDCG* are obtained. When the setting of *d* is too large or too small, the performances will deteriorate. These related experimental results indicated that

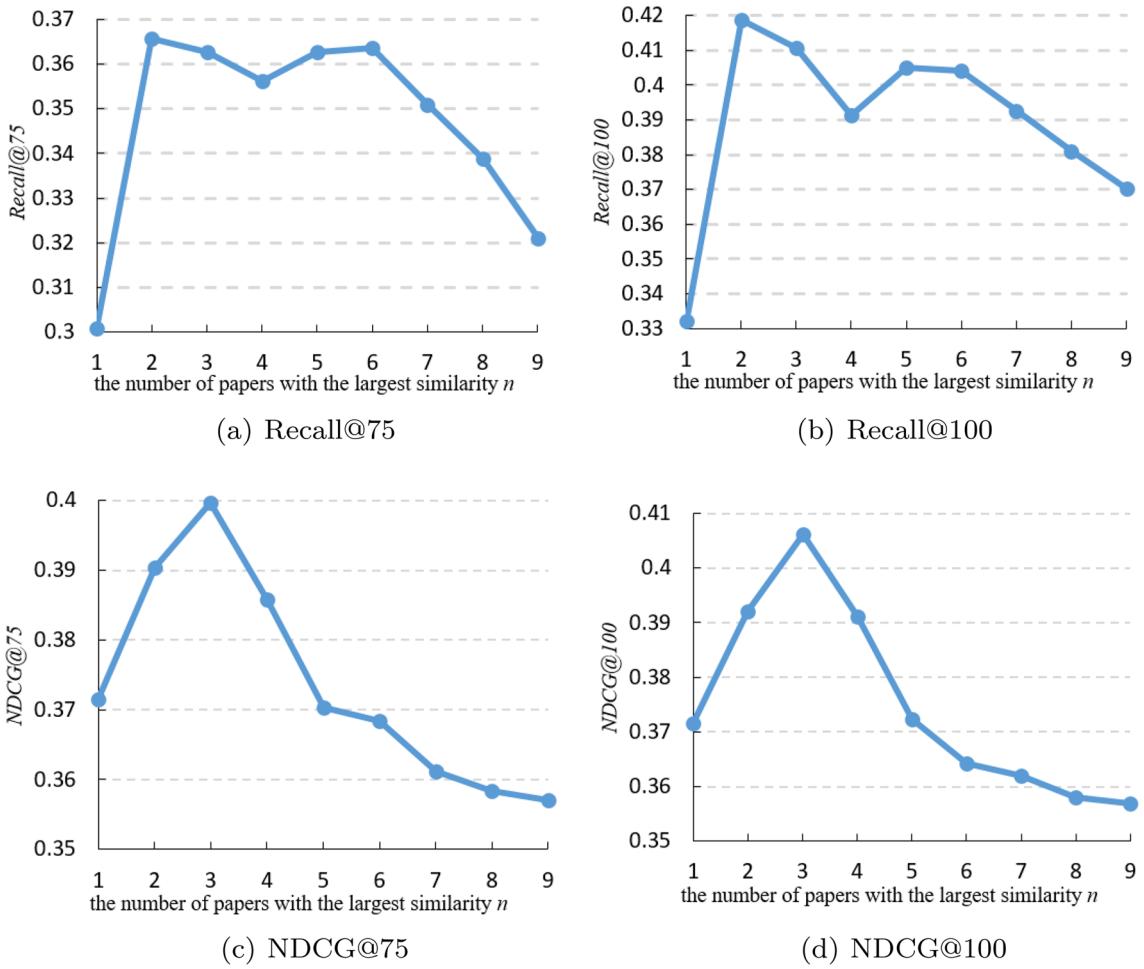

Fig. 2 *Recall* and *NDCG* for different sizes of TNSS (*n*). **a** *Recall*@75. **b** *Recall*@100. **c***NDCG*@75. **d** *NDCG*@100

the effect of the recommendation results is closely related to the value of *d*. By comprehensive consideration on the *Recall* and *NDCG* performance, we finally select 75 as preferable value for *d*.

Generally, both *Recall* and *NDCG* began to decrease when *d* and *n* reached certain values. And when the setting of *n* and *d* is too large or too small, the performances will deteriorate. By further comparing the Figs. 2 and 3, it can be demonstrated that the impact of changes in the parameter *n* was more obvious than that caused by *d*. All curves decreased slowly in Fig. 3 but quickly in Fig. 2, which indicate that *n* has a larger impact than *d* on the recommendation results. These phenomena illustrate that the feature information that DeepWalk obtains from the citation is limited by increasement of *d*, due to inaccurate calculations of paper content similarity. Moreover, an increase in *n* somehow added useless or even wrong information to the feature vector, due to overfitting on embedding learning procedure.

6 Conclusions

The existing approaches face the issues of data sparsity and high-dimensional in large-scale bibliographic network representation, which hinder the citation recommendation performance. To address these problems, we proposed a Content-Sensitive citation representation approach for Citation Recommendation, named CSCR. We linked some

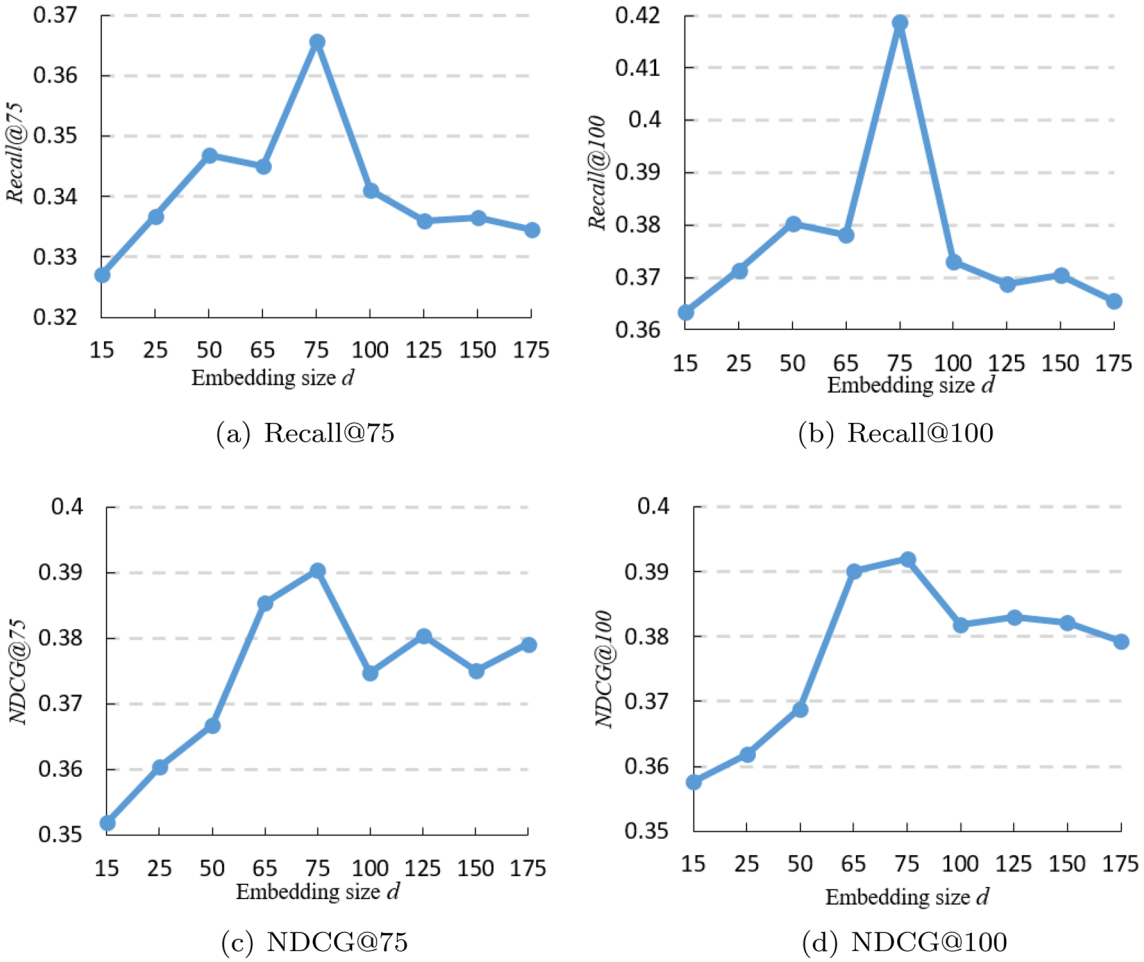

Fig. 3 *Recall* and *NDCG* for different sizes of embedding dimension *d*. **a** *Recall*@75. **b** *Recall*@100. **c** *NDCG*@75. **d** *NDCG*@100

papers without citation relationships but with high content relevances. So, the data sparsity problem can be alleviated as well as without increasing the dimensionality of the network. Moreover, papers with high content relevances can be found more easily during the recommendation process, which can promote citation recommendation performance. In the future, we will explore more effective feature for appending the citation auxiliary links. And we will find more powerful network representation method to achieve citation network representation.

Acknowledgements This work was supported in part by the Natural Science Basic Research Plan in Shaanxi Province of China (Program No. 2020JQ-214), and the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement (Program No. 840922).

References

Ali Z, Qi G, Kefalas P, Abro WA, Ali B (2020) A graph-based taxonomy of citation recommendation models. Artif Intell Rev 53:5217–5260

Amami M, Pasi G, Stella F, Faiz R (2016) An lda-based approach to scientific paper recommendation. International conference on applications of natural language to information systems. Springer, Berlin, pp 200–210

Cai X, Han J, Li W, Zhang R, Pan S, Yang L (2018) A three-layered mutually reinforced model for personalized citation recommendation. IEEE Trans Neural Netw Learn Syst 29(12):6026–6037

Cao S, Lu W, Xu Q (2015) Grarep: learning graph representations with global structural information. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, pp 891–900

Chen X, Hj Z, Zhao S, Chen J, Yp Z (2019) Citation recommendation based on citation tendency. Scientometrics 121(2):937–956

Dai T, Zhu L, Cai X, Pan S, Yuan S (2018) Explore semantic topics and author communities for citation recommendation in bipartite bibliographic network. J Ambient Intell Hum Comput 9(4):957–975

Dai T, Zhu L, Wang Y, Carley KM (2019) Attentive stacked denoising autoencoder with bi-lstm for personalized context-aware

citation recommendation. IEEE/ACM Trans Audio Speech Lang Process 28:553–568
Ding Y, Zhang G, Chambers T, Song M, Wang X, Zhai C (2014) Content-based citation analysis: the next generation of citation analysis. J Assoc Inform Sci Technol 65(9):1820–1833
Grover A, Leskovec J (2016) node2vec: scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 855–864
Guo L, Cai X, Hao F, Mu D, Fang C, Yang L (2017) Exploiting fine-grained co-authorship for personalized citation recommendation. IEEE Access 5:12714–12725
Guo L, Cai X, Qin H, Guo Y, Li F, Tian G (2019) Citation recommendation with a content-sensitive deepwalk based approach. In: 2019 International Conference on Data Mining Workshops (ICDMW), IEEE, pp 538–543
He Q, Pei J, Kifer D, Mitra P, Giles L (2010) Context-aware citation recommendation. In: Proceedings of the 19th International Conference on World Wide Web, ACM, pp 421–430
Hu Y, Xiong F, Pan S, Xiong X, Wang L, Chen H (2021) Bayesian personalized ranking based on multiple-layer neighborhoods. Inform Sci 542:156–176
Ko Y (2012) A study of term weighting schemes using class information for text classification. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 1029–1030
Lau JH, Baldwin T (2016) An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368
Li Y, Wang S, Ma Y, Pan Q, Cambria E (2020) Popularity prediction on vacation rental websites. Neurocomputing 412:372–380
Liu Q, Chen E, Xiong H, Ding CH, Chen J (2011) Enhancing collaborative filtering by user interest expansion via personalized ranking. IEEE Trans Syst Man Cybern 42(1):218–233
Liu H, Kong X, Bai X, Wang W, Bekele TM, Xia F (2015) Context-based collaborative filtering for citation recommendation. IEEE Access 3:1695–1703
Liu Z, Sun M, Lin Y, Xie R (2016) Knowledge representation learning: a review. J Comput Res Dev 53(2):247
Liu H, Wang Y, Peng Q, Wu F, Gan L, Pan L, Jiao P (2020) Hybrid neural recommendation with joint deep representation learning of ratings and reviews. Neurocomputing 374:77–85
Liu X, Yu Y, Guo C, Sun Y (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, pp 121–130
Livne A, Gokuladas V, Teevan J, Dumais ST, Adar E (2014) Citesight: supporting contextual citation recommendation using differential search. In: Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 807–816
Ma X, Wang R (2019) Personalized scientific paper recommendation based on heterogeneous graph representation. IEEE Access 7:79887–79894
Nallapati R, Cohen WW (2008) Link-plsa-lda: a new unsupervised model for topics and influence of blogs. In: Proceedings of the 12th International AAAI Conference on Web and Social Media, pp 84–92
Pan L, Dai X, Huang S, Chen J (2015) Academic paper recommendation based on heterogeneous graph. Chinese computational linguistics and natural language processing based on naturally annotated big data. Springer, Berlin, pp 381–392
Pan S, Hu R, Sf Fung, Long G, Jiang J, Zhang C (2019) Learning graph embedding with adversarial training methods. IEEE Trans Cyber 50(6):2475–2487
Perozzi B, Al-Rfou R, Skiena S (2014) Deepwalk: online learning of social representations. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp 701–710
Radev DR, Muthukrishnan P, Qazvinian V, Abu-Jbara A (2013) The ACL anthology network corpus. Lang Resour Eval 47(4):919–944
Ren X, Xea Y, Liu J (2014) Cluscite: effective citation recommendation by information network-based clustering. In: In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, pp 821–830
Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q (2015) Line: Large-scale information network embedding. In: Proceedings of the 24th International Conference on World Wide Web, pp 1067–1077
Torres R, McNee SM, Abel M, Konstan JA, Riedl J (2004) Enhancing digital libraries with techlens. In: Digital Libraries, 2004. Proceedings of the 2004 Joint ACM/IEEE Conference on, IEEE, pp 228–236
Totti LC, Mitra P, Ouzzani M, Zaki MJ (2016) A query-oriented approach for relevance in citation networks. In: Proceedings of the 25th International Conference Companion on World Wide Web, pp 401–406
Wang J, Zhu L, Dai T, Wang Y (2020a) Deep memory network with bi-lstm for personalized context-aware citation recommendation. Neurocomputing 410:103–113
Wang W, Tang T, Xia F, Gong Z, Chen Z, Liu H (2020b) Collaborative filtering with network representation learning for citation recommendation. IEEE Trans Big Data. https://doi.org/10.1109/TBDATA.2020.3034976
Wei X, Croft WB (2006) Lda-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp 178–185
West JD, Wesley-Smith I, Bergstrom CT (2016) A recommendation system based on hierarchical clustering of an article-level citation network. IEEE Trans Big Data 2(2):113–123
Xiong F, Shen W, Chen H, Pan S, Wang X, Yan Z (2019) Exploiting implicit influence from information propagation for social recommendation. IEEE Trans Cybern 50(10):4186–4199
Yang Z, Wu B, Zheng K, Wang X, Lei L (2016) A survey of collaborative filtering-based recommender systems for mobile internet applications. IEEE Access 4:3273–3287
Yang L, Zhang Z, Cai X, Guo L (2019) Citation recommendation as edge prediction in heterogeneous bibliographic network: a network representation approach. IEEE Access 7:23232–23239
Zhang W, Yoshida T, Tang X (2011) A comparative study of tf* idf, lsi and multi-words for text classification. Expert Syst Appl 38(3):2758–2765
Zhang D, Yin J, Zhu X, Zhang C (2018) Network representation learning: a survey. IEEE Trans Big Data 6(1):3–28