A Systematic Review of Citation Recommendation Over the Past Two Decades

Yicong Liang, Hong Kong Metropolitan University, Hong Kong*

Lap-Kei Lee, Hong Kong Metropolitan University, Hong Kong

ABSTRACT

A citation is a reference to the source of information used in an article. Citations are very useful for students and researchers to locate relevant information on a topic. Proper citation is also important in the academic ethics of article writing. Due to the rapid growth of scientific works published each year, how to automatically recommend citations to students and researchers has become an interesting but challenging research problem. In particular, a citation recommendation system can assist students to identify relevant papers and literature for academic writing. Citation recommendation can be classified into local and global citation recommendation depending on whether a specific local citation context is given; e.g., the text surrounding a citation placeholder. This article provides a systematic review on global citation recommendation models and compares the reviewed methods from the traditional topic-based models to the recent models embedded with deep neural networks, aiming to summarize this field to facilitate researchers working on citation recommendation.

KEYWORDS

Citation Recommendation, Recommender System, Systematic Review

INTRODUCTION

The volume of scientific articles has increased so dramatically in the past decade; it is impossible for a researcher or student to digest all the new information available in the scientific repository. Various online academic service providers have allowed users to access papers through their search engines, such as Google Scholar, ACM Digital Library, ScienceDirect, IEEE Xplore, and Semantic Scholar, where users can input queries to search for relevant articles in their database. Paper recommender systems can complement the search engine, and the possible recommendation scenarios can be grouped into three categories based on the recommendation timing, i.e., before, during, or after a search session.

DOI: 10.4018/IJSWIS.324071

*Corresponding Author

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.
(Li et al., 2019). In particular, citation recommendation is provided during a search session, in which related paper recommendations can be displayed beside the content that the user is currently browsing. While citation recommendation falls into one of the scenarios of paper recommendation, there are fundamental differences between citation and paper recommendation. Paper recommendation focuses on providing users with articles that are worthwhile to read and examine in the context of a research topic. Furthermore, a paper recommender system gives personalized article recommendations based on modeling the user’s profile from their behavior history, e.g., clicked/bookmarked/written documents. However, citation recommendation refers to the task of recommending appropriate citations for a text passage within a document (Färber & Jatowt, 2020). While there are some published surveys on paper recommendation (Beel et al., 2016; Bai et al., 2019), the present survey focuses on citation recommendation.

The traditional process of finding relevant citations requires much tedious and manual work. A researcher needs to record a document collection on their own, and the chance of identifying papers for citing depends on whether the researcher already knows the candidates. Another option for acquiring citation candidates is to rely on a bibliographic database, such as Google Scholar, or domain-specific platforms such as DBLP in computer science or PubMed in biomedical and the life sciences domain. However, finding the appropriate query keywords to search for cited papers is skillful work and requires considerable time and effort (Färber & Jatowt, 2020). The motivation for citation recommendation is to provide an efficient method for the citation process. The user provides the written text to the recommender system, and then the system presents a snippet with citations. For instance, given the citation context “FM has a uniform weight in feature interactions; the researchers introduce attention mechanism to enable features differently for link prediction” within a document, the citation recommendation system might produce two citations as follows: “FM (Rendle, 2010) has a uniform weight in feature interactions; the researchers introduce attention mechanism (Vaswani et al., 2017) to enable features that contribute differently for link prediction.”, where the two corresponding references are added respectively to (1) a publication introducing Factorization Machine (FM) which is a popular technique in modeling feature interaction, and (2) a publication backing up the statement that the proposed method introduces an augmented component for the task of link prediction.

Two user groups can benefit from a citation recommender system, as shown in the following examples:

- **Expert Group:** An experienced researcher familiar with their research field is going to conduct a survey on a research project related to AI education systems. Citation recommendation can be beneficial, as such a user might not be familiar with publications about educational psychology.
- **Non-expert Group:** Newcomers, such as early-stage master’s students and Ph.D. students, might get frustrated when facing the tremendous amount of research publications, and they are unaware of most of the relevant literature in their research areas. Citation recommender systems help those novice researchers to find cite-worthy publications to write their research proposals.

Given the significance and popularity of paper recommendation research, four survey papers related to research-paper recommendation were published in recent years (Beel et al., 2016; Bai et al., 2019; Ali et al., 2020; Färber & Jatowt, 2020). The following brief discussion reviews the main differences between these works and the present paper. Beel et al. (2016) and Bai et al. (2019) mainly focus on models related to paper recommendations, and these models are classified into three major categories, namely, collaborative filtering (CF), content-based filtering (CBF), and a hybrid of CF and CBF. Modeling user profiles is an essential component of a paper recommendation framework since the recommended documents need to match the targeting user’s profile. But the present survey aims to review citation recommendation models, and the reviewed algorithms in this work include representation learning and deep learning-based novel methods in recent years. The problem of citation recommendation can be further split into local (context-aware) and global (non-context-aware) citation
recommendation (He et al., 2010) (see Section 2 for more details). Ali et al. (2020) surveyed citation recommendation models only focusing on deep neural networks and classified the methods into six criteria. Färber and Jatowt (2020) surveyed papers in local citation recommendation models. The present survey mainly focuses on global citation recommendation models, including both traditional machine learning methods and recent deep learning techniques.

Since scientific papers are the objects to be recommended, some recommender systems also recommend “citations” and the difference between recommending papers and citations is sometimes marginal (Beel et al., 2016). To identify relevant literature for this survey, we conducted a two-step literature search on Google Scholar, ACM Digital Library, ScienceDirect, and IEEE Xplore. First, we searched for and downloaded all articles that had relevance for research paper/citation recommendation systems. Second, we filtered the articles to keep those focusing on the task of citation recommendation instead of paper recommendation. Finally, 27 approaches in total were chosen for an in-depth analysis as shown in Table A1 and Table A2 in the Appendix section.

This survey makes the following three contributions:

- This study surveyed 27 global citation recommendation models published over the past two decades. To characterize and distinguish different models, this work systematically classified global citation approaches into five categories and investigated various ranking score functions for generating citation candidates.
- The possible datasets and evaluation metrics for evaluating citation recommendation performance were identified. Bibliometric analysis in the field of citation recommendation was also performed to identify productive researchers and their affiliations.
- The challenges in the area of citation recommender systems were outlined and discussed in terms of evaluation framework selection, user profile construction, and learning network representation.

### Comparison of Citation Recommendation Approaches

This paper first introduces the definitions and properties of the citation recommendation problem. The main notations are listed in Table 1.

The problem of citation recommendation can be classified into local citation and global citation recommendations according to the range of context (He et al., 2010). Specifically, given a document $q$, the global context is the title and abstract of $q$; the local context is the text surrounding a citation or placeholder; the out-link context corresponds to the citing document; and the in-link context corresponds to the cited document. Their problem definitions are as follows:

<table>
<thead>
<tr>
<th>Notations</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q$</td>
<td>an input query manuscript</td>
</tr>
<tr>
<td>$p$</td>
<td>a candidate cited paper</td>
</tr>
<tr>
<td>$\varphi(\cdot)$</td>
<td>a representation function projecting the input (e.g., paper) to an embedding latent space</td>
</tr>
<tr>
<td>$s(q, p)$</td>
<td>score function measures how likely query manuscript $q$ is to cite paper $p$</td>
</tr>
<tr>
<td>$G, E$</td>
<td>citation graph $G$, where a vertex in $V$ denotes paper, and an edge in $E$ denotes a citation relation between the citing paper and the cited paper</td>
</tr>
</tbody>
</table>
Global Citation Recommendation: Given a query manuscript \( q \) without a bibliography, a global citation recommendation is a ranked list of citations in a corpus that are recommended as candidates for the bibliography of \( q \).

Local Citation Recommendation: Given an out-link context \( c \) with respect to a document \( d \), a local recommendation is a ranked list of citations in a corpus that are recommended as candidates for the placeholder associated with \( c \).

According to He et al. (2010), the input to the global recommender system is only a manuscript containing a bag of words without references. However, from the reviewed articles, we found that most work, especially graph-based methods (Liu et al., 2014b; Jiang et al., 2015; Cai et al., 2018b), adopted a relaxed definition. The proposed methods treat the input manuscript as an unfinished project, and the research topic (i.e., title, abstract, or even full text) and part of references from the project can help to find the other relevant citation documents. This paper reviewed the models in existing works with such a relaxed definition of global citation recommendation. The citation recommendation problem can be cast into the problem of learning a recommendation score function on the input \((q, p) : Q \times P \rightarrow R\), where \( q \in Q \) is a query manuscript and \( p \in P \) is a target paper based on the bibliographic network.

Collaborative Filtering Method

The collaborative filtering (CF) technique has been widely and successfully applied in many recommender systems. The intuition behind CF is that the collaborative behaviors of all users can be leveraged for predicting the behavior of a target user. CF-based recommendation systems (Koren et al., 2022) need to create a user rating matrix relating two fundamentally different entities, namely, users (rows in the matrix) and items (columns in the matrix); the entry \( r_{ij} \) in the matrix represents the rating of item \( j \) given by user \( i \). There are two main techniques of CF: the neighborhood approaches (user-based and item-based) and latent factor models.

In the context of citation recommendation, McNee et al. (2002) first introduced the CF framework by transferring the citation network into the user-item rating matrix. Specifically, given a citation relation in the network, the citing paper is treated as a user while the cited paper is considered as an item. McNee et al. (2002) compared four CF algorithms in the experiments and found that user-item and item-item CF achieved promising performance in offline experiments. The CF framework has the advantage of lightweight architecture since only the ID information and interaction history are utilized. However, the major drawback of applying the CF technique in citation recommendation is that the “user” (citing paper) will never add more votes (cited papers) after the citation network is constructed.

Caragea et al. (2013) addressed the citation recommendation problem by proposing a singular value decomposition (SVD) approach (Deerwester et al., 1990) that performed better than traditional item-based CF methods in their experiment evaluated on the Citeseer digital library. Specifically, both citing and cited papers are mapped into a joint \( k \)-dimensional latent factor space, and the paper-pair correlation is modeled as an inner product in this space. Each paper is represented as a vector in \( R^k \). Let \( \varphi(q_u) \in R^k \) and \( \varphi(p_i) \in R^k \) denote the vectors of citing paper \( q_u \) and cited paper \( p_i \), respectively. The correlation between \( q_u \) and \( p_i \) can be computed by their inner product as follows:

\[
    r_{ui} = \varphi(q_u)^T \varphi(p_i)
\]

The final citation recommendation is based on the score computed from the vectors of the query \( q_u \) and the candidate paper \( p_i \).
Feature-Based Method

Strohman et al. (2007) integrated text features with citation features to rank scientific articles for given a query paper, and their experimental results showed that text similarity alone has a poor performance in finding relevant citations while citation features play a major role in finding high-quality articles. Expanding on the use of text features and simple citation features in paper ranking models, Bethard and Jurafsky (2010) introduced sixteen features including different citation behaviors in the document retrieval model, and proposed an iterative training algorithm to learn the combined weights of features in the scoring function. Finally, the retrieval model scores each candidate paper \( p \) against the query paper \( q \) using a weighted sum of feature scores:

\[
s(q, p) = \sum_i \omega_i \times f_i(q, p),
\]

(2)

where \( \omega_i \) and \( f_i(q, p) \) are the weight and the feature score of the \( i \)-th feature.

Paper citations should be organized into different groups and each group should have its own behavior pattern to identify information of interest (Ren et al., 2014). Ren et al. (2014) noted that citations tend to be clustered into different interest groups in a heterogeneous bibliographic network and mining such interesting citation patterns were good features for the task of citation prediction. Their proposed model ClusCite defined the score function \( s(q, p) \) by capturing the relatedness between input query \( q \) and target paper \( p \) according to the interest group and relative importance of \( p \) within each interest group. In order to capture the semantic relevance between papers and interest groups, various meta-paths are incorporated into the scoring function.

Gupta and Varma (2017) incorporated content with graph structure to learn the representation of papers for citation recommendation. They applied doc2vec (Le & Mikolov, 2014) to capture semantic information at the document level and DeepWalk (Perozzi et al., 2014) to capture topology properties at the node level. Two representations from different modalities were transformed into a common space by using the Canonical Correlation Analysis (CCA) method (Hardoon et al., 2004), and a linear combination of the resulting projections was treated as the final representation of each paper. The final representations of two papers were input into the system for the task of binary prediction of a citation link.

Topic Model-Based Method

Textual information (i.e., topic) and citation relation are two important properties of academic research papers that should be explored in relation to the paper recommendation. Nallapati et al. (2008) argued that the explicit document dependency (i.e., citation relation) better captured the topic distributions of documents which could improve the performance of citation link prediction. They proposed the model Link-PLSA-LDA by jointly modeling text and citations in the topic-modeling framework. Inspired by protein-protein interactions, they combined the LDA (Blei et al., 2003) model with the Mixed Membership Stochastic Block (MMSB) (Airoldi et al., 2008) model to capture the semantics of citation links. Their experiments treated the citation recommendation task as a link prediction problem and the ranking score was based on the conditional probability \( Pr(p \mid q) \) of a cited candidate \( p \) given a query paper \( q \).

As academic research papers are quite different from other items in a recommender system, citation recommendation is more sensitive to the publication date, and users’ (researchers’) information needs may change over time. Jiang et al. (2014) proposed a chronological recommendation model DTCIM for users to generate citation recommendations shown in different time slices. DTCIM models “information need shifting” on the topic level where the topic can be treated as three kinds of time-series distributions (i.e., document-topic, topic-word, and document-topic influence distributions)
rather than a static term distribution. Both document dynamic topic-model and time-decay factor contribute to defining the ranking score function that helps to generate ranking lists in different time slices.

Researchers usually prefer to cite topic-relevant papers and it remains challenging to model the implicit correlations between topics and citations. To leverage content information and link structure in the bibliographic network, Dai et al. (2018a) proposed an LDA-based probabilistic topic model TMAILCCite that combined topic-model and matrix factorization to extract semantic topics and qualified author communities. In TMAILCCite, topics were generated by LDA, and the author community was detected by matrix factorization, where such integration improved the performance of citation recommendation. Dai et al. (2018a) chose the exponential mean function as the indicator for ranking score based on the work of Chang and Blei (2009), where the mean function had a good prediction performance in the experiment. In another work, Dai et al. (2018b) proposed the method TopicCite by jointly combining feature regression with topic models for citation recommendation and topic extraction. Five types of citation features with 29 citation features in total, e.g., text similarity, citation count, author similarity, author history, venue relevancy, and meta-path similarity, were extracted from bibliographic data. Feature regression and topic modeling mutually reinforced the learning of the feature weights and topic distributions. Dai et al. (2018b) defined the ranking score function between the query and a candidate paper based on the topic distribution similarity and a linear combination of citation feature similarity.

Graph-Based Method

Academic papers are intrinsically connected in a citation network via their citation relations; the random walk process plays an important role for ranking purposes in graph-based methods. The task of ranking vertices in a graph can be formulated as a function \( f : V \rightarrow R^+ \). A random walk defines a Markov chain in a given network, where each vertex represents a state and a walk transits from one state to another based on a transition probability \( p(u, v) \). In the context of a citation graph, a random walk on a graph is defined by a transition probability function: \( p : V \times V \rightarrow [0, 1] \). Let \( p_T(v) \) denote the probability that the walker is at the cited paper \( u \) at time \( T \). A standard random walk can be defined as:

\[
p_T(v) = \sum_{(u,v) \in E} p(u, v) p_{T-1}(u)
\]

where \( p_T(v) \) can be used for getting citation recommendation results. PageRank technique, which is a special case of random walk, has been widely used in node prestige computation in a network. Gori et al. (2006) applied a biased version of PageRank to the paper recommendation system based on a citation graph. The proposed random walk-based scoring algorithm found relevant papers according to a small set of user-selected articles.

A citation network with one single type of vertex (paper) is considered a homogeneous network, but some works find that extracting semantics in heterogeneous information networks (HIN) including multiple types of entity (e.g., papers, authors, venues) and relations (paper-paper citation relation and author-author collaboration relation) helps to improve recommendation performance. The concept of meta-paths has been introduced in measuring node similarity (PathSim) (Sun et al., 2011) in complex network structures, and various meta-paths address different hypotheses for ranking. In general, the meta-path-based similarity for the object \( x \) and object \( y \) can be computed as:

\[
s(x, y) = \sum_{p \in P} g(p)
\]
where \( g(p) \) denotes some particular measure (e.g., random walk) defined on the path instance \( p \) (in the meta path \( P \)) between \( x \) and \( y \). Meta-paths plus random walk from starting vertices to candidate paper vertices in HIN is a popular combination for ranking score function (Ren et al., 2014; Liu et al., 2014a; Liu et al., 2014b).

Conventional methods oversimplify the characteristics of the citation relationship; in other words, all that matters is whether \( \text{paper}_1 \) cites \( \text{paper}_2 \), regardless of importance, sentiment, or motivation. Liu et al. (2014b) proposed a context-rich heterogeneous network approach by characterizing the importance of citation relationships and topical citation motivation. Specifically, they extracted the full-text information in each paper to find different citation contexts for each cited paper and treated each citation as a new node rather than an edge in HIN so that there could be multiple paths between two citing and cited papers. To enrich the information of input query, given a user’s initial text information need, Liu et al. (2014b) obtained a set of seed nodes in terms of papers by using text search (language model). In order to quantify the ranking score of candidates relevant to the seeds, a meta-path-based random walk measure was proposed for the score function. Following the seed nodes concept for ranking, Liu et al. (2014a) used pseudo-relevant feedback (PRF) to extract seed nodes from HIN, which more global scholarly information can be obtained for citation recommendation tasks. Liu et al. (2014a) argued that different metapaths including various citing patterns and meta-paths with restriction (i.e., nodes in the path related to a specified keyword topic) could enhance the accuracy of the ranking function. Their experimental results showed that various meta-paths with author-centric, citation-centric, and venue-centric perspectives facilitated users to find relevant cited candidates. The heterogeneous network in Liu et al. (2014a) and Liu et al. (2014b) was built by using full-text data rather than scholar metadata (e.g., venue, author). Furthermore, a citation is a vertex instead of an edge in the graph, and it is connected to the keyword nodes, which indicates the citation’s topical motivation information.

Many existing techniques for citation recommendation only focus on suggesting prestigious and well-cited articles. Chakraborty et al. (2015) were the first to address the diversity problem, and the proposed model DiSCern finds relevant and diversified citations in response to a search query. To make a diversified recommendation, DiSCern expanded the query in order to cover as much space of related topics as possible by leveraging a keyword-keyword graph (Chakraborty et al., 2015). Articles corresponding to the expanded query were collected to determine a subgraph from the original citation network, and a reinforced random walk-based algorithm was adopted on the subgraph to generate the citation recommendation. Jiang et al. (2015) employed a supervised document influence model to extract the content time-varying dynamics and applied multiple meta-paths for ranking recommended articles. Three time-series topic-based distributions explore the dynamic content on the topic level in terms of dynamic topic mixture proportion, dynamic topic evolution, and dynamic topical influence. Jiang et al. (2015) adopted a meta-path-based random walk method in the heterogeneous network as the score ranking function. To mine the information-need shifting, they learned the evolving feature weights in different stages.

Due to language barriers, junior researchers or graduate students are unable to efficiently locate publications hosted in a foreign language repository. Jiang et al. (2018) addressed the problem of cross-language citation recommendation (CCR) that, given a query paper in a source language, the system output a list of ranked papers in the target language—i.e., recommending English candidate cited papers for a given Chinese publication. They proposed the Hierarchical Representation Learning on Heterogeneous Graph (HRLHG) model to learn a representation function by mapping the multilingual repositories to a joint embedding space and applied the embedding method for CCR. The scoring ranking function for CCR is defined as a conditional probability \( \Pr(p_c | p_q) \), i.e., the probability of the candidate paper \( p_c \) in the target language given a particular query paper \( p_q \) in the source language:
Pr(p_e | p_q) = \Delta(\varphi(p_q), \varphi(p_e))

(5)

where $\Delta$ is a probability scoring function (e.g., cosine similarity with Relu function) based on the learned representation vector $\varphi(p_q)$ and $\varphi(p_e)$.

To enrich the query information, personalized citations can be made via random walk-based methods (Mu et al., 2017; Cai et al., 2018b), network representation-based methods (Guo et al., 2017; Cai et al., 2018c; Yang et al., 2019), or deep learning methods (Cai et al., 2018a; Zhang, Yang et al., 2018). Mu et al. (2017) argued a common drawback of existing approaches was that the searcher’s query context was overlooked (e.g., ignoring the specific requirement of the user or diversified link information). Their proposed model MMRQ incorporated the query information into a multi-layered graph and designed mutual reinforcement rules to establish a ranking framework. Particularly, they exploited all available multiple types of relations involving both intra- and inter-network information among author, article, and keyword entities extracted from literature collections to build a multi-layered graph. To recommend personalized citation, the query information was incorporated into the multi-layered graph to obtain a query-focused and mutually reinforced recommendation (Mu et al., 2017).

To make citation recommendations, co-authorships can provide helpful information for analyzing the citation behavior of researchers based on the assumption that researchers may cite similar papers if they have collaborated in the past. Guo et al. (2017) argued that binary modeling of co-authorship may introduce information loss in terms of missing strong or weak relationships between specific research topics. In other words, the importance weight of the co-authorship link should be considered (i.e., collaboration influence within topics). They proposed using K-means clustering (Wu et al., 2016), topic model, and random walk to capture the authors’ collaboration influence. A biased random walk is applied on the heterogeneous graph to update the probability of visiting each node, and citation ranking scores are based on the probability value when the iterative updated process converges.

Computation towards the entire graph might introduce a high computational complexity due to the huge graph size. To reduce the high computation complexity in the process of finding candidates, Cai et al. (2018b) applied a three-layered (i.e., paper layer, author layer, and venue layer) interactive clustering approach to group related nodes (i.e., papers, researchers, and venues) in a subgraph where citation recommendations were generated by clusters associated with researchers’ needs. For personalized citation recommendation, they integrated the identity of the researcher and the query text information into a random walk process on the constructed subgraph. The proposed mutual reinforcement model not only recommended relevant papers for citation, but also recommended researchers and venues to a given user. The ranking score was based on the vertex importance of the subgraph after finishing an iterative relevance propagation process.

Instead of learning representation from an original citation network, the VOPRec proposed by Kong et al. (2018) learned vertex representation from an extended citation graph, where some similar nodes were artificially linked to each vertex even though they did not have citation relations. In other words, in the citation network, papers are connected due to their existing citation relations; however, in an extended citation network, papers are also connected due to sharing some common features (e.g., text content or structure topology). Kong et al. (2018) utilized Doc2Vec (Le & Mikolov, 2014) to find similar papers in textual content and Struct2vec (Ribeiro et al., 2017) to find other similar papers in structural property to reconstruct the network. A biased random walk was used on the extended weighted citation network to learn the integrated representation for each paper and the top-N papers were recommended according to the cosine similarity of the two node vectors:

$$s(q, p) = \frac{\varphi(q) \cdot \varphi(p)}{\varphi(q) \varphi(p)}$$

(6)
Cai et al. (2018c) proposed the integration of intra-vertex relationships, inter-vertex relationships, and vertex-content correlation to model a bibliographic network representation. A biased random walk was used to generate a path sequence and the skip-gram framework was adopted to train the model. Finding relevant citation papers can be treated as a link prediction task, and Yang et al. (2019) developed a network representation-based edge prediction (NREP) model that learned knowledge from existing citation links and predictive representation for citation recommendation. The heterogeneous representation for each node was learned from tripartite information (i.e., vertex-content correlation, neighbor context, and node pair within or without edge) and the ranking results were generated based on the similarity of learned vectors (Yang et al., 2019).

Chen et al. (2019) argued that if an article was chosen as a reference, it satisfied some particular conditions considered as citation tendency. Their proposed algorithm CIRec is based on heterogeneous network modeled citation tendency as the probability of entity-to-entity migration. Specifically, they applied weighted random walk on the weighted heterogeneous information network to promote papers’ walk to their cited papers and used a skip-gram model for learning vertex representation under various characteristics and features. The citation ranking score was based on the cosine similarity of the learned vectors.

Deep Learning-Based Method

Traditional language models (e.g., word2vec) (Mikolov et al., 2013) use bag-of-words as a feature to train a low-dimensional vector representation of each word for downstream tasks. To provide personalized citation recommendations, deep learning techniques have been adopted in some works. Cai et al. (2018a) proposed the model GAN-HBNR, which exploited a generative adversarial network (GAN) to obtain low dimensional embedding of different objects in a heterogeneous bibliographic network and the obtained network representations. For a personalized recommendation, they formulated the input as the manuscript author and manuscript text, and the similarity scores were calculated between the representation of the candidate from the trained corpus and representations of the author and text.

Many existing citation recommendation systems need meta information (e.g., author, publication venues) to locate relevant papers. Bhagavatutula et al. (2018) proposed a content-based method for citation recommendation, where metadata was optional input. Their experiment results suggested that adding metadata favors self-citations which were of less help in the citation recommendation scenario. For a query document, it is computationally expensive to score each document in the corpus as a candidate reference. Bhagavatutula et al. (2018) recommended citations in two phases in terms of candidate selection and reranking candidates in an efficient manner. In phase 1, they trained a model NNSelect to project each paper into a latent vector space using a weighted sum from the word embedding learned by optimizing the per-instance triplet loss (Wang et al., 2014), and they extracted the query document’s nearest neighbors as candidates. In phase 2, they trained another model NNRank which could estimate the probability that a candidate document should be cited in the query document. The ranking score function was based on the probability from the sigmoid output in NNRank.

Zhang et al. (2018) proposed a GAN-based model method for personalized citation recommendation. Their proposed model VCGAN leveraged the principle of adversarial learning (Goodfellow et al., 2014) to learn two network representations in terms of content-based and author-based graph representation. In particular, they used a graph convolution network (GCN) (Kipf & Welling, 2016) to extract the node representation encoding both graph structure and node features and applied a GAN framework to obtain the graph embedding matrix to generate citations for a given query manuscript. Specifically, the corresponding embedding vectors of query paper \(q\) and recommended candidate paper \(p\) were looked up from the embedding matrix, and the ranking score was computed by the cosine similarity of these two representation vectors.

Recommender systems not only recommend suitable referred articles to users, but also help users to manage their reading literature (i.e., recommending labels or tags). Galke et al. (2018) utilized
adversarial autoencoders for multi-modal recommendation tasks in terms of recommending citations and tags. Multi-layer perceptron (MLP) was used to build the components in the encoder and decoder parts, and multi-modal input (i.e., user-rating matrix and title of documents) was incorporated with the adversarial autoencoder block to conduct the two recommendation tasks. Their experiments explored some interesting interactions between input modalities and the task and found that multi-modal variants achieved a better performance than solely content-based variants when item co-occurrence resembled relatedness (Galke et al., 2018).

Graph-based methods are based on an adjacency matrix and suffer from data sparsity problems in large-scale bibliographic networks. Some approaches (Cai et al., 2018b; Guo et al., 2017) expand the network by integrating different types of vertices (e.g., papers, authors, venues, keywords) and relationships (e.g., paper-paper, paper-author, and paper-venue), which would increase the dimension of the bibliographic network and lead to high computation complexity for learning network representation. Guo et al. (2022) proposed their model CSCR to artificially add links between two papers if they had similar content, which can alleviate the data sparsity problem but did not increase the dimension of the adjacency matrix. They utilized doc2vec (Le & Mikolov, 2014) to find similar papers in order to reconstruct the network and adopted DeepWalk (Perozzi et al., 2014) to learn node representation from the extended citation network.

Summary of Graph-Based and Deep Learning Models

In general, graph-based models can be further separated into different groups according to the ranking score function. The random walk-based scoring function takes the query manuscript $q$ as input, the walker traverses starting from $q$ to other nodes based on the constructed graph. Random walk scores induce the sorting of papers according to their expected liking for the input. The difference between random walk-based methods and graph-based models mainly comes from the constructed graph. For example, some graph-based models, like that of Gori and Pucci (2006), only used the citation network, while others incorporated the citation network with the co-author network, paper-keyword network (Guo et al., 2017; Mu et al., 2017), or venue-paper network (Cai et al., 2018b; Cai et al., 2018c) in building the graph.

Meta-path-based ranking functions rely on the random walk-based algorithm, but the meta-path model is defined on the heterogeneous network and the walker is restricted under some predefined meta-path schemes. The difference among these meta-path graph-based models mainly comes from the constructed heterogeneous graph and the setting of meta-path schemes. For example, the citation relation in Liu et al. (2014) was considered as a node rather than an edge in the graph. For generating chronological recommendations, time-related topic features were integrated into meta-path settings and the historical topic label was one type of vertex in the constructed graph (Jiang et al., 2015).

Regarding graph-based models using the cosine similarity (cos-sim) ranking function, these models focused on extracting features of entities in the graph and finally projecting each paper to a common latent vector space. The ranking score for the recommendation was based on the cosine similarity between the input query and candidate paper embeddings. These cos-sim graph-based models mainly differed in their feature extractors and constructed graphs. For example, to make cross-language citation recommendations, the model proposed by Jiang et al. (2018) learned a representation function that can handle a multilingual environment. The models proposed by Kong et al. (2018), Yang et al. (2019), Chen et al. (2019), and Guo et al. (2002) learned network representation based on the extension of the skip-gram algorithm, and their major difference was the graph construction.

VOPREC (Kong et al., 2018; Guo et al., 2022) reconstructed the citation network by linking the nearest text-based and structure-based neighbors. Chen et al. (2019) used heterogeneous information networks including paper-paper, paper-author, and paper-term networks, and Yang et al. (2019) learned network representation from heterogeneous bibliographic networks containing nodes of papers, authors, and venues.
For the task of global citation recommendation, deep learning methods were applied in building feature extractors (Cai et al., 2018a; Zhang et al., 2018; Galkeet et al., 2018), candidate selection, and ranking modules (Bhagavatula et al., 2018). Cai et al. (2018a) and Zhang et al. (2018) utilized a GAN framework to train and obtained the embeddings of network representation, where the ranking results were based on the similarity of query manuscript and candidate vectors. Galke et al. (2018) used MLP to model the latent vector for the title of the paper and adversarial autoencoder to obtain a multi-modal representation. Bhagavatula et al. (2018) used a supervised neural model to retrieve document embedding and adopted contrast learning to select candidates in phase 1, and they reranked the candidates according to the predicted citing probability by the proposed NNRank module implemented in a three-layer feedforward neural network.

EVALUATION

In the context of the research paper recommendation, there are three main approaches based on different scenarios, namely, offline evaluation, online evaluation, and user studies (Beel et al., 2016). Compared to online evaluation and user studies, it is convenient to implement an offline evaluation framework that measures the accuracy of a recommender system based on a predefined ground truth. Even though criticism has been raised in offline evaluation (e.g., a strongly pruned dataset causes bias in the evaluation (Beel et al., 2016)), offline evaluation still dominated the experiment performance evaluation in citation recommendation, while only a few works (McNee et al., 2002; Gori & Pucci, 2006) adopted human subject evaluation to conduct their experiment. The reason is that every published research paper has its own reference list and such citations provided by the authors of a query document can be treated as the ground truth.

To evaluate the performance of citation recommendation, the paper’s reference list is treated as ground truth during the offline evaluation. Specifically, for a testing citing paper, a subset of its cited papers are randomly selected for holdout and are treated as the ground truth for comparison. The proposed approach is evaluated by assessing which of the citations that have been recommended by the system are also in the original publications. This evaluation method is called “citation re-prediction” (Färber & Jatowt, 2020). The popular evaluation metrics in the field of recommender systems can also be transferred into evaluating citation re-prediction tasks, including precision, recall, $F_1$ score, normalized discounted cumulative gain (NDCG), mean average precision (MAP), and mean reciprocal rank (MRR). These six evaluation metrics focus on past citing behavior. To keep the evaluation more scalable, some relaxed citation prediction evaluation metrics are adopted (e.g., relative co-cited probability) (He et al., 2010; Chakraborty et al., 2015).

Datasets

Both paper recommendation and citation recommendation share some sources of datasets in common. For paper recommendation, Beel et al. (2016) found that CiteSeer and CiteUlike are the top two datasets in offline evaluations. Paper recommendation is a personalized recommendation service, and both CiteSeer and CiteUlike provide personal settings (e.g., personal collections, and document bookmarking). The datasets in these reviewed papers are shown in Table A3 and A4 in the Appendix section. The data from ACL Anthology Network (AAN) was used in 12 works for evaluation (44%). DBLP and ACM Digital Library were two other popular datasets used in 9 and 6 citation recommendation experiments, respectively (33% and 22%). These three datasets provide citation network information which makes it convenient for researchers to set up the ground truth for offline evaluations. Also, it contains meta-information that helps to build representation for each paper (e.g., titles, abstracts, authors, and venues). Wangfang is the only one non-English dataset for citation recommendation and it is adopted to evaluate a cross-lingual citation recommendation model (Jiang et al., 2018) between Chinese and English.
Citation recommendation models employ different data factors and features including title, abstract, keywords, author information, venue information, and citation network. For the surveyed papers, information used from the datasets is listed in Table A3 and Table A4. As researchers want the system to recommend articles with content close to their research interests, the paper text content in terms of title and abstract are leveraged in modeling content representation. A citation network is used in all surveyed citation recommendation approaches. Exploring the structural property of papers in the citation network helps to find relevant citation candidates. The features of the author and venue are combined in heterogeneous network-based models where multiple types of vertices and relations are leveraged in the network representation.

**Evaluation Metrics**

Precision and recall are two commonly used evaluation metrics in the field of information retrieval. Precision@N measures the proportion of the recommended citations that is relevant to the ground truth in the top-N recommendation list. Recall@N measures the rate of real citations over the total original citations in the top-N results.

\[
\text{Precision} = \frac{\sum_{d \in Q(D)} |R(d) \cap T(d)|}{\sum_{d \in Q(D)} |T(d)|}
\]

\[
\text{Recall} = \frac{\sum_{d \in Q(D)} |R(d) \cap T(d)|}{\sum_{d \in Q(D)} |R(d)|}
\]

where \(Q(D)\) is the set of test papers, \(R(d)\) denotes the recommended citation result for a test paper \(d\), and \(T(d)\) denotes the ground truth in \(d\) (i.e., original references). \(F_1\) score is a balance between precision and recall and gives the harmonic mean of precision and recall.

\[
F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}
\]

To better evaluate the quality of the ranked sequence of documents generated by the recommender system, mean average precision is widely used as another evaluation metric. Specifically, given a test paper \(d\), a precision-recall curve (PRC) can be plotted by computing a pair of precision and recall at every position in the ranked list of documents. Average precision (AP) can be calculated under the PRC (Zhu, 2004).

\[
\text{AP} = \sum_{k=1}^{n(d)} P(k) \Delta r(k)
\]

where \(k\) is the rank in the sequence of recommended citations. \(P(k)\) is the precision at cut-off \(k\) in the list, and \(\Delta r(k)\) denotes the change in recall from \(k-1\) to \(k\). MAP for a set of test papers is the mean of AP scores for each test query paper (Zhu, 2004).
\[
\text{MAP} = \frac{\sum_{d \in Q(D)} AP(d)}{|Q(D)|}
\]

MRR is the inverse position of the first relevant document in the result list and it evaluates the performance on how far away the first relevant reference paper occurs from the top position.

\[
\text{MRR} = \frac{1}{|T(d)|} \sum_{d \in T(d)} \frac{1}{r_d}
\]

where \( r_d \) is the rank of the first relevant citation in the result for test paper \( d \). Discounted cumulative gain (DCG) adopts a graded relevance scale of documents from the result to evaluate the gain of a document based on its position in the result list.

\[
\text{DCG}_k = \sum_{i=1}^{k} \frac{\text{rel}_i}{\log_2 (i + 1)}
\]

where \( \text{rel}_i \) is an indicator function equaling 1 if the item at rank \( i \) is a relevant document. The normalized DCG (NDCG) values for all test papers can be averaged to evaluate the performance of a ranking algorithm.

\[
\text{NDCG}_k = \frac{\text{DCG}_k}{\max \{ \text{DCG}_k \}}
\]

The main drawback of setting removed original citations as ground truth is that the relevant citations are always in the past. Some relevant or even better recommendations other than the original ones among the top-\( N \) results cannot be captured by traditional metric-like precision (He et al., 2010). For each pair of papers \( d_i, d_j \) where \( d_i \) is an original citation and \( d_j \) is a recommended one, the co-cited probability (CP) (He et al., 2010) is defined as:

\[
\text{CP} = \frac{\text{count}(d_i \cap d_j)}{\text{count}(d_i)}
\]

where \( \text{count}(d_i \cap d_j) \) denotes the number of papers citing both \( d_i \) and \( d_j \), and \( \text{count}(d_i) \) is the number of papers citing \( d_i \).

The evaluation metrics used in the reviewed papers are listed in Table A3 and Table A4 in the Appendix section. Recall, NDCG, and MAP were the top 3 popular metrics for evaluating citation recommendation performance. Both recall and NDCG were used in 14 evaluations (52%); MAP in 13 evaluations (48%); precision in 9 evaluations (33%); MRR in 8 evaluations (30%); \( F_1 \) and co-cited probability in 2 and 1 evaluations. Most of the reviewed articles utilized at least two metrics for evaluation and the combination of recall and NDCG was the favorite choice. The reason is that...
recall focuses on the fraction of relevant retrieved instances, and in addition, NDCG considers the ranking order of relevant instances. If the proposed model obtains a promising score in both recall and NDCG, it shows that the recommended citations are relevant and appear in the top-ranking position. In our surveyed articles, DiSCern (Chakraborty et al., 2015) was the only model concerning diversified citation recommendation; co-cited probability was used in evaluating the effectiveness of retrieving diversification.

Bibliometric Analysis

This section presents the bibliometric analysis of the surveyed articles. The researchers found that the reviewed papers are authored by 79 researchers from 32 affiliations. The most productive authors and affiliations in the citation recommendation field are shown in Table 2 and Table 3. The most productive authors are Xiaoyan Cai and her co-author Libin Yang from Northwestern PolyU. And the most productive research groups are from Northwestern PolyU, University of Minnesota, Xi’an Jiaotong University, and UIUC. These researchers and affiliations are mainly from China (55%) and the USA (31%). However, 80% of the authors published no more than one paper on citation recommender systems and the result is consistent with the findings of Beel et al. (2016) that most of the authors published no more than one work on research paper recommendation. All of the reviewed papers were authored by multiple authors: the majority of articles had four (30%) or five (22%) authors.

Table 2. Most productive authors in citation recommendation field

<table>
<thead>
<tr>
<th>Author</th>
<th>Papers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Xiaoyan Cai</td>
<td>10</td>
</tr>
<tr>
<td>Libin Yang</td>
<td>6</td>
</tr>
<tr>
<td>Lantian Guo</td>
<td>5</td>
</tr>
<tr>
<td>Xiaozhong Liu</td>
<td>5</td>
</tr>
<tr>
<td>Liangcai Gao</td>
<td>4</td>
</tr>
<tr>
<td>Fei Hao</td>
<td>3</td>
</tr>
<tr>
<td>Tao Dai</td>
<td>3</td>
</tr>
<tr>
<td>Zhuoren Jiang</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 3. Most productive affiliations in citation recommendation field

<table>
<thead>
<tr>
<th>Affiliation</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Northwestern Polytechnical Univ</td>
<td>29</td>
</tr>
<tr>
<td>Univ of Minnesota</td>
<td>8</td>
</tr>
<tr>
<td>XJTU</td>
<td>8</td>
</tr>
<tr>
<td>UIUC</td>
<td>7</td>
</tr>
<tr>
<td>Indiana Univ</td>
<td>6</td>
</tr>
<tr>
<td>Anhui Univ</td>
<td>5</td>
</tr>
<tr>
<td>Dalian Maritime Univ</td>
<td>5</td>
</tr>
<tr>
<td>Dalian Univ of Technology</td>
<td>5</td>
</tr>
</tbody>
</table>
Challenges

Recommender systems have evolved into an important tool to help people make more informative, efficient, and effective decisions in almost every aspect of their daily lives. There are plenty of mature applications-related recommendation systems in various fields, such as e-commerce (e.g., Amazon, Taobao), entertainment (e.g., Youtube, Tiktok), and social networking (e.g., Facebook, Linkedin). However, to the best of the authors’ knowledge, few applications related to academic citation recommendation are accessible to the public. A similar version of the citation recommendation system is to find related papers (Lo et al., 2019). Specifically, finding related papers is such a problem that, given an article, during this browsing session, a list of relevant articles is displayed beside the article that the user is currently reading. Some major challenges for citation recommendation systems are discussed in the following sections.

Evaluation Framework Selection

Most of the reviewed citation recommendation models conducted offline evaluations that original references of papers were considered as ground truth. The major drawback of this strategy is that the published date of the original references is earlier than the citing paper, which guides the system to find “old” candidates. Due to the rapid growth of scientific publications each year, researchers might prefer citing those papers published recently. Besides offline evaluation, online evaluation needs to be incorporated. Online evaluations in terms of click-through rate (CTR) is widely used in e-commerce recommender system (Beel et al., 2016). In the context of citation recommendation, the system should record the ratio of co-clicked or downloaded articles to the number of articles displayed.

User Profile Construction

Authors of citing papers may differ in their knowledge background and expertise when selecting citations. How to give a personalized citation recommendation is a promising trend in the future. It is easy to build a user profile on some recommender systems (e.g., Facebook) where users have a strong motivation to give self-presentation (Nadkarni & Hofmann, 2012). However, it is difficult to model user profiles in citation recommendation systems due to the lack of explicit interactions between users and the system. Some work, e.g., the work of Li et al. (2019), tried to alleviate this problem by sending out paper recommendations in email newsletters based on the users’ browsing history from the academic search engine and creating a continuous feedback loop. In the context of citation recommendation, the system needs to help users to create their personal profiles in terms of some short self-introduction (e.g., affiliation, research topics, available publications, and bookmarking).

Learning Network Representation

The mainstream of the citation recommendation mechanism is twofold: learning document representations from the training corpus and measuring the relatedness of the embeddings between query manuscript and recommended candidate. As academic papers lie in a heterogeneous information network, the quality of network representation learning including article embedding is crucial to find relevant citation candidates. Deep learning techniques including MLP, CNN, and RNN, which are capable of capturing the semantic representations and associated contextual information of research papers, have been applied in citation recommendations (Ali et al., 2020). The deep neural network model Transformer (Vaswani et al., 2017) has achieved state-of-the-art status in some tasks related to natural language processing (NLP) and computer vision (CV), but Transformer is still not applied to citation recommendation problems. The major problem is how to transfer citation data into sequential data (e.g., extracting sequences with citation behavior patterns). With such training data, Transformer with a self-attention mechanism can be applied to learn the network representation.
CONCLUSION

Citation recommendation plays an important role in the context of scientific big data to assist students and researchers to identify relevant papers and literature for academic writing. This paper presented a brief and systematic review related to global citation recommendation models over the past two decades. The authors discussed the benefits of citation recommender systems and the problem formulation of citation recommendation. In order to characterize and distinguish the different models, this reviewed paper systematically classified global citation approaches into five categories—namely, collaborative filtering, feature-based, topic model-based, graph-based, and deep learning-based methods. In addition, the researchers investigated the ranking framework toward citation candidates according to different ranking score functions. Furthermore, this paper examined the possible datasets and evaluation metrics for evaluating citation recommendation performance from the surveyed articles and completed some bibliometric analysis to find productive researchers and affiliations in this field. Finally, the threefold challenges were outlined in the field of citation recommender systems.

ACKNOWLEDGMENT

The work described in this paper was supported by the Katie Shu Sui Pui Charitable Trust — Academic Publication Fellowship (Project Reference No. KSPF/2022/02).
REFERENCES


## APPENDICES

Table A1. Overview of global citation recommendation methods. Part-A.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Venue</th>
<th>Approach</th>
<th>Score function</th>
</tr>
</thead>
<tbody>
<tr>
<td>(McNee et al., 2002)</td>
<td>CSCW</td>
<td>CF-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Gori &amp; Pucci, 2006)</td>
<td>WI</td>
<td>Graph-based</td>
<td>Random walk</td>
</tr>
<tr>
<td>(Strohman et al., 2007)</td>
<td>SIGIR</td>
<td>Feature-based</td>
<td>Weighted linear combination</td>
</tr>
<tr>
<td>(Nallapati et al., 2008)</td>
<td>KDD</td>
<td>Topic model</td>
<td>Conditional probability</td>
</tr>
<tr>
<td>(Bethard &amp; Jurafsky, 2010)</td>
<td>CIKM</td>
<td>Feature-based</td>
<td>Weighted sum of feature scores</td>
</tr>
<tr>
<td>(Ren et al., 2014)</td>
<td>KDD</td>
<td>Feature-based</td>
<td>Weighted sum of feature scores</td>
</tr>
<tr>
<td>(Liu et al., 2014b)</td>
<td>JCDL</td>
<td>Graph-based</td>
<td>meta-path-based random walk</td>
</tr>
<tr>
<td>(Liu et al., 2014a)</td>
<td>CIKM</td>
<td>Graph-based</td>
<td>meta-path-based random walk</td>
</tr>
<tr>
<td>(Jiang et al., 2014)</td>
<td>Web-KR</td>
<td>Topic model</td>
<td>weighted sum of topic relevance scores</td>
</tr>
<tr>
<td>(Jiang et al., 2015)</td>
<td>CIKM</td>
<td>Graph-based</td>
<td>meta-path-based random walk</td>
</tr>
<tr>
<td>(Chakraborty et al., 2015)</td>
<td>ICDE</td>
<td>Graph-based</td>
<td>reinforced random walk</td>
</tr>
<tr>
<td>(Gupta &amp; Varma, 2017)</td>
<td>WWW</td>
<td>Feature-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Guo et al., 2017)</td>
<td>IEEE Access</td>
<td>Graph-based</td>
<td>Random walk</td>
</tr>
</tbody>
</table>

Table A2. Overview of global citation recommendation methods. Part-B.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Venue</th>
<th>Approach</th>
<th>Score function</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cai et al., 2018a)</td>
<td>AAAI</td>
<td>Graph-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Bhagavatula et al., 2018)</td>
<td>NAACL</td>
<td>Deep learning</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Zhang et al., 2018)</td>
<td>ISMIS</td>
<td>Deep learning</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Cai et al., 2018b)</td>
<td>IEEE TRANS NNLS</td>
<td>Graph-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Dai et al., 2018)</td>
<td>JAIHC</td>
<td>Topic model</td>
<td>Exponential mean function</td>
</tr>
<tr>
<td>(Galke et al., 2018)</td>
<td>UMAP</td>
<td>Deep learning</td>
<td>Decoder output</td>
</tr>
<tr>
<td>(Jiang et al., 2018)</td>
<td>SIGIR</td>
<td>Graph-based</td>
<td>Cos-sim with Relu</td>
</tr>
<tr>
<td>(Mu et al., 2017)</td>
<td>IEEE Access</td>
<td>Graph-based</td>
<td>Random walk</td>
</tr>
<tr>
<td>(Dai et al., 2018b)</td>
<td>IEEE Access</td>
<td>Topic model</td>
<td>weighted sum of feature scores</td>
</tr>
<tr>
<td>(Cai et al., 2018c)</td>
<td>IEEE Access</td>
<td>Graph-based</td>
<td>Random walk</td>
</tr>
<tr>
<td>(Kong et al., 2018)</td>
<td>IEEE TRANS ETC</td>
<td>Graph-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Yang et al., 2019)</td>
<td>IEEE Access</td>
<td>Graph-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Chen et al., 2019)</td>
<td>Scientometric</td>
<td>Graph-based</td>
<td>Cos-sim</td>
</tr>
<tr>
<td>(Guo et al., 2022)</td>
<td>JAIHC</td>
<td>Graph-based</td>
<td>Cos-sim</td>
</tr>
</tbody>
</table>
### Table A3. Datasets and evaluation metrics. Part-A.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Datasets</th>
<th>Evaluation metric</th>
<th>Information used</th>
</tr>
</thead>
<tbody>
<tr>
<td>(McNee et al., 2002)</td>
<td>ResearchIndex</td>
<td>precision</td>
<td>Citation network</td>
</tr>
<tr>
<td>(Gori &amp; Pucci, 2006)</td>
<td>ACM digital library</td>
<td>precision</td>
<td>Citation network</td>
</tr>
<tr>
<td>(Strohman et al., 2007)</td>
<td>Rexa</td>
<td>MAP</td>
<td>Title, author, citation network</td>
</tr>
<tr>
<td>(Nallapati et al., 2008)</td>
<td>Citeseer</td>
<td>recall</td>
<td>Title, abstract, citation network</td>
</tr>
<tr>
<td>(Bethard &amp; Jurafsky, 2010)</td>
<td>AAN</td>
<td>MAP</td>
<td>Title, abstract, author, venue, citation network</td>
</tr>
<tr>
<td>(Ren et al., 2014)</td>
<td>DBLP, PubMed</td>
<td>precision, recall, MRR</td>
<td>Title, abstract, author, venue, citation network</td>
</tr>
<tr>
<td>(Liu et al., 2014b)</td>
<td>ACM Digital library</td>
<td>MAP, NDCG</td>
<td>Keyword, author, venue, citation network</td>
</tr>
<tr>
<td>(Liu et al., 2014a)</td>
<td>ACM Digital library</td>
<td>MAP, NDCG</td>
<td>Keyword, author, venue, citation network</td>
</tr>
<tr>
<td>(Jiang et al., 2014)</td>
<td>ACM Digital library</td>
<td>MAP, NDCG</td>
<td>Title, abstract, citation network</td>
</tr>
<tr>
<td>(Jiang et al., 2015)</td>
<td>ACM Digital library</td>
<td>MAP, NDCG</td>
<td>Title, abstract, citation network</td>
</tr>
<tr>
<td>(Chakraborty et al., 2015)</td>
<td>High-Energy Physics</td>
<td>recall, MAP, co-cited probability</td>
<td>Keyword, citation network</td>
</tr>
<tr>
<td>(Gupta &amp; Varma, 2017)</td>
<td>Arnetminer</td>
<td>precision, MAP, NDCG</td>
<td>Title, abstract, citation network</td>
</tr>
<tr>
<td>(Guo et al., 2017)</td>
<td>AAN</td>
<td>recall, NDCG</td>
<td>Keyword, author, citation network</td>
</tr>
</tbody>
</table>

### Table A4. Datasets and evaluation metrics. Part-B.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Datasets</th>
<th>Evaluation metric</th>
<th>Information used</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Cai et al., 2018a)</td>
<td>AAN, DBLP</td>
<td>recall, MAP, MRR</td>
<td>Title, abstract, citation network</td>
</tr>
<tr>
<td>(Bhagavatula et al., 2018)</td>
<td>DBLP, PubMed, OpenCorpus</td>
<td>MRR, F₁</td>
<td>Title, abstract, keyword, author, venue, citation network</td>
</tr>
<tr>
<td>(Zhang et al., 2018)</td>
<td>AAN</td>
<td>recall, NDCG</td>
<td>Title, abstract, author, citation network</td>
</tr>
<tr>
<td>(Cai, et al., 2018b)</td>
<td>AAN, DBLP, Citeseer</td>
<td>recall, MAP, MRR</td>
<td>Author, venue, citation network</td>
</tr>
<tr>
<td>(Dai et al., 2018a)</td>
<td>AAN, DBLP</td>
<td>precision, recall, MRR</td>
<td>Title, abstract, author, venue, citation network</td>
</tr>
<tr>
<td>(Galke et al., 2018)</td>
<td>CITREC PubMed</td>
<td>MRR</td>
<td>Title, citation network</td>
</tr>
<tr>
<td>(Jiang et al., 2018)</td>
<td>Wangfang, ACM digital library</td>
<td>MAP, NDCG, precision, MRR</td>
<td>Keyword, citation network</td>
</tr>
<tr>
<td>(Mu et al., 2017)</td>
<td>AAN</td>
<td>recall, NDCG</td>
<td>Keyword, author, citation network</td>
</tr>
<tr>
<td>(Dai et al., 2018b)</td>
<td>AAN, DBLP</td>
<td>precision, recall, NDCG</td>
<td>Title, abstract, keyword, author, venue, citation network</td>
</tr>
<tr>
<td>(Cai et al., 2018c)</td>
<td>AAN, DBLP</td>
<td>recall, MAP, MRR</td>
<td>Title, abstract, author, venue, citation network</td>
</tr>
<tr>
<td>(Kong et al., 2018)</td>
<td>Physical Review A</td>
<td>precision, recall, F₁, NDCG</td>
<td>Title, abstract, citation network</td>
</tr>
<tr>
<td>(Yang et al., 2019)</td>
<td>AAN, DBLP</td>
<td>recall, MAP, NDCG</td>
<td>Title, abstract, author, venue, citation network</td>
</tr>
<tr>
<td>(Chen et al., 2019)</td>
<td>AAN, DBLP</td>
<td>recall, NDCG, precision</td>
<td>Title, abstract, author, citation network</td>
</tr>
<tr>
<td>(Guo et al., 2022)</td>
<td>AAN</td>
<td>recall, NDCG</td>
<td>Title, abstract, citation network</td>
</tr>
</tbody>
</table>
Yicong Liang received his Bachelor of Engineering in School of Software from Sun Yat-sen University, Guangzhou, China. He is currently working toward a Master of Philosophy degree in School of Science and Technology in Hong Kong Metropolitan University. His research interests include recommender system, artificial intelligence in education.

Lap-Kei Lee is an Assistant Professor at the School of Science and Technology of Hong Kong Metropolitan University. He received his Bachelor of Engineering in Computer Engineering and Doctor of Philosophy in Computer Science from the University of Hong Kong. His research interests include the design and analysis of algorithms (especially in online job scheduling and data stream algorithms), natural language processing, algorithm engineering, and educational technology.