ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

A Survey on Citation Recommendation System

Muhsina V P[1], Naseer C[2]

PG Student, Dept of Computer Science & Engineering, Government Engineering College, Palakkad, Kerala, India[1]

Assistant Professor, Department of Computer Science and Engineering, Government Engineering College, Palakkad, Kerala, India[2]

Abstract: Citation recommendation is challenging and a meaningful research problem and which is applicable in research paper publication. Simply it means that, it helps the researchers to find related work of their manuscript while authoring a paper. Various methodologies are proposed for the recommending citation from content of paper. The purpose of this paper is to survey various contemporary techniques for citation recommendation. Respective motivations of these approaches are discussed and their advantages and limitations are compared in this survey.

Keywords: Citation, Citation recommendation, Global recommendation, Local recommendation

I. INTRODUCTION

Citation recommendation is task of recommending citation to the researcher while authoring a paper. In order to understand the citation recommendation first you need to know about citation and its importance. Citations are important in academic dissemination and it is the manner in which you tell your readers that specific material in your work originated from another source. Citations are helpful to anyone who wants to find out more about your ideas and where they came. Elements included in citation are author of paper, year of publication etc.

Citation recommendation is import ant in the area of research paper publication. Due to many more papers are published day by day in digital library, it become challenging for conducting comprehensive scientific literature review. Citation recommendation can help to improve the quality and efficiency of this problem by recommending published documents as citations for a query manuscript. During peer review process or early stage of research life cycle, researcher doesn't have knowledge about what works are done earlier related to this. So researchers seek out the related works while authoring a paper. This citation recommendation system is extremely helpful for the researchers to understand their topic deeply and to compare their ideas. It helps researchers to find related work and also to check the completeness of citations while authoring a paper.

When beginning a work in a new research subject, a researcher usually wants to have a quick understanding of the leaving literacy works in this field, including which papers are the most relevant papers and what sub-topics are presented in these papers. Search for related work is an important part of writing papers. At the point when paper are written, ordinarily the author wants to make a few citation at a place however he isn't sure which papers to cite. Since the number of research paper published is exponentially growing, the filtering process is generally tedious and time consuming. Traditional literature search engines like Google Scholar, CiteSeer [1], can retrieve relevant papers based on giving certain keywords as query. Researchers need to experience them manually to discover works that should be reoffered to. Due to huge development of research article in previous decade and introduction of new terminologies and new knowledge in research area, the process of searching for existing work becomes difficult for both junior and experienced researcher. What we might want is a system that can suggest citation for a manuscript written by researcher without compelling for large portion of work. Such system can improve efficiency of searching related work.

Almost all citation recommendation system can be divided in to two categories, global and local citation recommendation. Global citation recommendation [2], [3], [4] and [5] recommends list of references based on taking whole scholarly article as query document and local citation recommendation [6] and [7] recommends citations based on local context of input sentences and an optional placeholders for candidate citation.

This survey aims to discuss about the different methodologies for citation recommendation and highlight the advantages and disadvantages of each of them. Such a comparative study is very much relevant because of the wide

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE
International Journal of Advanced Research in Computer and Communication Engineering
PCON-2019
National Conference and Seminar on Innovations in Engineering & Technology
Govt Polytechnic College, Palakkad
Vol. 8, Special Issue 1, January 2019

range of applications that makes use of citation recommendation. This survey will provide some insights for choosing the right methodology for developing a citation recommendation system for a given domain based on its requirements.

This paper is organized as follows: Section II describes various citation recommendation system. Section III discusses critical analysis of different systems. Section IV explained about various datasets used for citation recommendation. Section V gives brief concluding comments.

II. CITATION RECOMMENDATION SYSTEM

Citation recommendation system recommends citation to the researcher while authoring a paper. It finds prior work related to the topic under investigation and to find missing relevant citations. There various recommendation system for citation based on manuscript or content of paper. Fig. 1 shows the types of citation recommendation system.

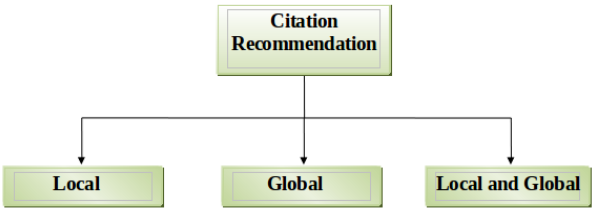

Fig. 1. Types of citation recommendation system

A. Local citation recommendation system

A local citation recommendation system takes a couple of sentences and an optional placeholder for the candidate citation as input and output is recommending citations based on context of input sentences. Some of local citation recommendation systems are introduced in [6], [8], [7].

W. Huang et al. [6] suggests a neural probabilistic model that learns the probability of citing a paper given a citation context based on distributed representations of words and documents using multilayer neural network. It helps knowledgeable researchers to give an accurate citation context for a cited paper or to find the right paper to cite given context.

This method focus on the local (or context based) citation recommendation. This neural model learns the distributed semantic representations of the words and the cited documents. By using the distributed representations of words and documents, train a neural network model that estimates the probability of citing a paper given a citation context. The neural network model will tune the distributed representations of words and documents so that the semantic similarity between the citation context and the cited paper will be high. Neural network architecture for learning context word and cited document are shows in Fig. 2. In a neural probabilistic model, the conditional probability can be defined using a softmax function. The skip-gram model is able to learn high quality representations of words. At the bottom of the Fig. 2 shows an example that demonstrates how one pair of citation context and cited document is processed by the neural network.

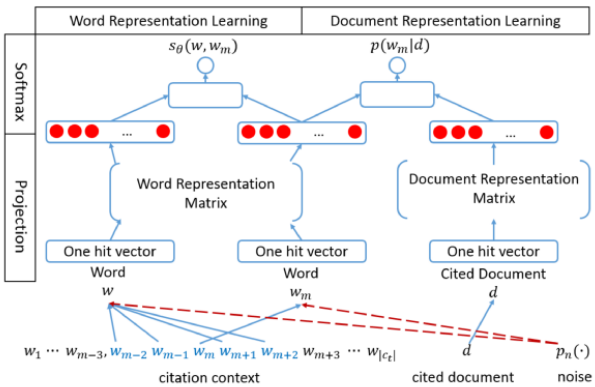

Fig. 2. Neural network architecture for word and document representation learning [6]

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

B. Global recommendation System

A global citation recommendation system takes whole manuscript as input and then recommending citations for the paper. Some works like [2], [4] are example for global citation recommendation system.

Ren et al. [4], introduce citation recommendation problem as graph based method. It examine problem with regards to heterogeneous bibliographic network and propose a novel cluster-based citation recommendation frame work, called ClusCite. Based on multiple type of relationship in a network, citations are softly clustered in to interest groups. Those interested groups are used for predicting each query citations having its own model for paper authority and relevance. In this work [4], defines citation recommendation problem as:

Given a heterogeneous bibliographic network G, and the terms, authors and target venues for a query manuscript $q \in Q$, aim to build a recommendation model specifically for q, and recommend a small subset of target papers $p \in P$ as high quality references for q, by ranking the papers with the score functions (q, p).

This method [4] presents a cluster-based citation recommendation framework and it has two steps:

1) By solving a joint optimization problem, learning the model parameters based on known citations
2) Based on learned ClusCite model, making paper specific recommendation.

Group memberships for attribute objects, feature weights for interest groups, and object relative authority within each interest group are the three parameters to learn the model. The joint optimization problem carries out graph regularized co-clustering to learn the model parameters, which reduces prediction error and graph regularization. By doing so, it guarantee the learned model can yield good performance on training data and it can cluster attribute objects in terms of their citation interests. Here, ClusCite minimization algorithm is designed to iterate between co-clustering and authority propagation.

The method in [4], given a heterogeneous bibliographic network and the terms, authors and target venues for a query manuscript, it aims to build a recommendation model for query manuscript, and recommend a small subset of target papers as high quality references for query manuscript, by ranking the papers with the score function. This ClusCite citation recommendation framework requires metadata (i.e., authors, venues, key phrases, etc...).

C. Bhagavatula et al. [2] introduce a global citation recommendation system, which takes entirely scholarly research article as input and recommend appropriate citation. This is a content-based method for citation recommendation in an academic draft.

In this work [2], citation recommendation is divided in to two steps: candidate selection and reranking candidate. By encoding textual content of each document, a neural model is made to embed all available documents into a vector space. Then select the nearest neighbors of a query document as candidates by K-nearest neighbouring method and rerank the candidates using another neural model that trained to distinguish between observed and unobserved citation. In the case of new publication, this system can embed new document in the same vector space to identify candidate citation by preventing retraining the models.

In candidate selection phase, it computes dense embedding for query document using dense embedding model. It also computes feature vector of title and abstract of query document. Then parameters from document embedding model are trained to predict cosine similarity. Using per instance triplet loss, this model trained to predict high cosine similarity for document cited in query document(i.e., Positive document) and low cosine similarity for the document not cited in query document(i.e. negative document). Negative samples like random, negative nearest sample and citation-of-citation are selected in an efficient manner. Output of this phase is list of candidate citation with corresponding similarity scores. In re-ranking phase, it takes the output of first phase as input to the model. Here it uses another neural model trained to distinguish between observed and unobserved citation. Input features used in this phase are dense features of title, abstract of query document and documents in the corpus, sum of scalar weight of word types that occur in both documents, number of times candidate document has been cited in the corpus and cosine similarity between the documents are concatenated. Then it given into two dense exponential linear unit layer and one sigmoid layer. The output of this phase is probability that query document, d_q cites candidate document, d_i .

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

During peer review process or early stage of research paper writing, metadata such as author name, year of publication, venue, etc. are not available. This content based method [2] for citation recommendation which remains robust when metadata are missing for query documents. This system introduces a model for recommending citation called citeomatic. This is helpful for the researchers to find the existing work of research topic. The input to this model can be query document or URl or details of paper which includes metadata.

C. Both local and global citation recommendation system

The combined local and global citation recommendation takes advantages of both to provide efficient citation suggestion. Here individually computing the global recommendation and local recommendation. After the recommendation results are generated, the topic based global recommendation results will be used to filter out irrelevant results from local recommendation. W. Huang et al. [5], presents citation recommender system called Refseer, which automatically suggests candidate citation

W. Huang et al. [5], presents citation recommender system called Refseer, which automatically suggests candidate citation based on query manuscript. It helps to check the completeness of citation while authoring a paper. RefSeer presents both topic based global recommendation and also citation-context based local recommendation. This system uses all paper metadata provided by CiteSeer [1] which contains scholarly research articles. Papers content are parsed and used for training topic based model.

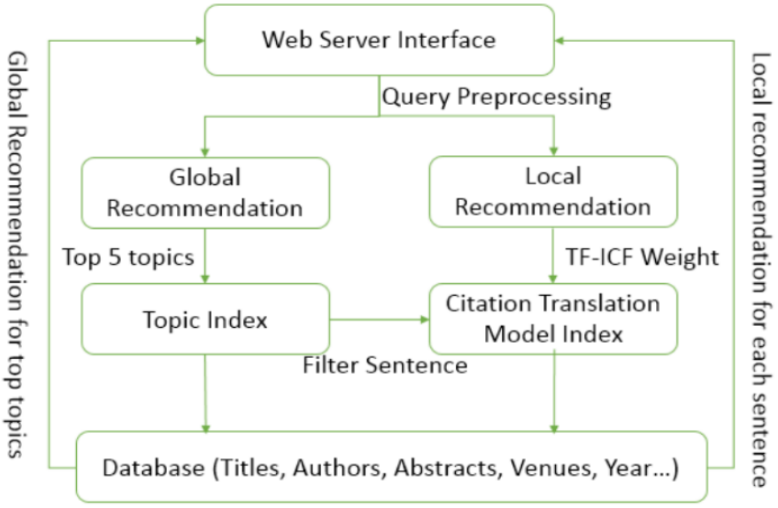

Fig. 3. Overview of Refseer infrastructure [5]

The Fig. 3 shows overview of Refseer infrastructure. First pre-process the query for recommendation. In query processing, given a query, and then split it into list of sentences by using sentence parser and remove stop words in the sentence. For global recommendation model, to infer the topics of model the entire query will be used. For local recommendation model, process each sentence as separate query and recommended list of related papers for each sentences. After query pre-processing the entire information flow divided in to two parts. Then local and global citation recommendations are individually computed.

For Global recommendation, Cite-PLSA-LDA [9] is used for computing topical composition of each paper. Cite-PLSA-LDA model is extension of LDA model, where words in the citation context related to the topics of citing document as well as topics of cited document. Use word-topic distribution Refseer infer top 5 topics from input query and use topic citation distribution for each topic to recommend a list of citation. For local recommendation, to learn the translation likelihood of citing a document given a word, RefSeer uses Citation Translation Model [10]. This translation model consider citation context as the source language and reference list are the target language. It uses term-frequency-inverse-context frequency(TF-ICF) to measure the probability of a citation need. Where, TF_t is defined as the number of times a given word t appears in query Q and ICF gives a measure of whether or not the word is common or rare across all citation contexts. Since not all sentences in query doesn't require citation, use topic distribution of the input query inferred in the phase of global recommendations to rule out sentences.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

III. CRITICAL ANALYSIS

Analysis of the above discussed approaches for citation recommendation is done as follows by considering a particular contribution of each.

In work [6] propose a neural probabilistic model that learns the chance of citing a paper given a citation context based on distributed representation. It helps the knowledgeable researchers to give an accurate citation context for a cited paper or to seek out the proper paper to cite. The representations of words and documents are learnt simultaneously from citation context and cited document pairs. Semantic embeddings of context and cited documents are helpful to get better result. This work [6] may be the first work that evaluates citation recommendation on the large scale dataset.

One problem of this method [6] is, it develop a document representation based on its constituent words only. It learns an explicit representation of each training document separately that is not a deterministic function of the documents words. Since a never-before-seen document doesn't have a readymade representation this makes model effectively transductive. Another problem of this method is that it needs a candidate document to have minimum one in-coming citation to be eligible for citation this disadvantages recently published documents.

The system by [4], introduce a novel citation recommendation framework to catch citation behaviours for each query document, in light of both paper relevance and importance. By clustering citations into different interest groups, it aims to study the significance of different relevance features for each interest group, and derive paper relative authority within each group. This method formulates a joint optimization problem to learn model parameters by taking advantage of multiple relationships in the network, and develop an efficient algorithm to solve it called ClusCite.

Coming to the limitation of the method of [4], it requires various information of query document such as author name and publication venue (i.e it requires metadata for citation suggestion). This may not be available in the early stage of research project. Since it is graph based method, the training complexity of the ClusCite algorithm is cubic in the number of edges within the graph of authors, venues and terms. This can be prohibitively cost extensive datasets.

This paper [2] presents a system for recommending citations based on the content of query document. It helps the researcher to find related work while authoring a paper. Based on this work, it introduces a scalable web-based literature review tool called Citeomatic. The main advantage of this method is it does not need information like author names which may be missing, e.g., during the peer review process. The limitation of system by [2] is Citeomatic model is limited to semantic scholar corpus. So papers outside from the scholar corpus give irrelevant citation as suggestion.

Introduce a citation recommendation system called RefSeer by [5], which automatically suggests citation for input query. It presents both topic based global recommendation and also citation-context based local recommendation. Advantage of RefSeer is it designed to deal with long queries. Queries are ranging from a sentence to an entire manuscript. It can be used to check the completeness of citations while authoring a paper. The complexities of training and recommending are efficient and scalable. Drawback of the system [5] is it requires metadata of query document to find citations. During early stage of writing process metadata may not be available. So it might be difficult for researcher with new ideas, to find related work.

IV. DISCUSSION OF DATASETS USED

There are different datasets available for analysing different citation recommendation system.

A snapshot of CiteSeer paper and citation database was obtained at Oct. 2013 are used in system [6]. The dataset is divided into two parts: papers crawled before 2011 (included) as the training set and papers crawled after 2011 as the testing set. Citations are extracted along with their citation contexts. One citation context consists of the sentence wherever a citation seems, as well as the sentences that seem before and once. Two different bibliographic datasets: the DBLP dataset [11] and the PubMed dataset re used in [4]. Citation information are extracted and built a DBLP citation dataset. Then generated a subset of the aforementioned dataset by filtering out papers with incomplete metadata information or less than 5 citations. Keywords and key phrases are extracted from paper titles and abstracts using the TF-IDF measure and the TextBlob noun phrase extractor. To process the PubMed Central dataset, the same method as described above is used to generate a subset.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

The DBLP dataset and PubMed datasets are also used in [2] to differentiate with existing work on recommending citation. Here, the DBLP Dataset contains over 50K scientific articles in the computer science domain, with an average of 5 citations for each article. The PubMed dataset contains over 45K scientific articles in the medical domains, with an average of 17 citations for each article.

In both datasets, contain documents with title, abstract, venue (i.e. journal or conference where the document was published), authors, citations (i.e. other documents in the corpus that are referenced in the given document) and key phrases. This method [2] also introduces OpenCorpus, a new dataset of 7 million scientific articles primarily drawn from the computer science and neuroscience domain. Due to licensing constraints, documents in the corpus only includes the title, abstract, year, author, venue, key phrases and citation information but do not include the full text of the scientific articles.

CiteSeer digital library [1], CiteULike are the two datsets used in [5] for citation recommendation. The dataset obtained from CiteULike from November 2005 to January 2008. Besides these two small dataset, also trained global and local recommendation models on the whole CiteSeer repository.

Different datasets are used for different citation recommendation system. Since the datasets and evaluation criteria used in the above discussed methodologies for citation recommendation is different, a comparative study based on the results is not possible. Table 1 shows the contribution of the methodologies discussed and datasets used by them in a nutshell

Table I

Type of Citation Recommendation System	**Model**	**Highlights**	**Dataset**
Local	W. Huang et al. [5]	Neural probabilistic model, Skip-gram model, Multi-layer neural network	A snapshot of CiteSeer paper, Citation database
Global	C. Bhagavatula et al. [2]	Neural model for candidate selection and reranking candidate, Does not require metadata	DBLP, PubMed, OpenCorpus
	Ren et al. [8]	Graph based, ClusCite algorithm, Joint optimisation problem	DBLP, PubMed
Both local and global	W. Huang et al. [4]	Topic based and context based recommendation	CiteSeer digital library, CiteULike

CONCLUSION

Citation recommendation helps the researchers to find related work of their idea while authoring a paper. Here recommending citations based content of paper, it can be local or global or both. All of these systems of citation recommendation can be included in either of these recommendations. Hence survey on citation recommendation may be helpful in order to choose the best technique. There is wide range of methodologies for recommending citation given query document or manuscript. Different authors choose different techniques for suggesting citations based on content of query document. We have discussed few approaches on citation recommendation and a comparative study is performed on them.

REFERENCES

[1]. C. L. Giles, K. D. Bollacker, and S. Lawrence, "Citeseer: An automatic citation indexing system," in Proceedings of the third ACM conference on Digital libraries, pp. 89–98, ACM, 1998..
[2]. C. Bhagavatula, S. Feldman, R. Power, and W. Ammar, "Content-based citation recommendation," arXiv preprint arXiv:1802.08301, 2018.
[3]. S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl, "On the recommending of citations for research papers," in Proceedings of the 2002 ACM conference on Computer supported cooperative work, pp. 116–125, ACM, 2002.
[4]. X. Ren, J. Liu, X. Yu, U. Khandelwal, Q. Gu, L. Wang, and J. Han, "Cluscite: Effective citation recommendation by information networkbased clustering," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 821– 830, ACM, 2014.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

IJARCCE

International Journal of Advanced Research in Computer and Communication Engineering

PCON-2019

National Conference and Seminar on Innovations in Engineering & Technology

Govt Polytechnic College, Palakkad

Vol. 8, Special Issue 1, January 2019

[5]. W. Huang, Z. Wu, P. Mitra, and C. L. Giles, "Refseer: A citation recommendation system," in Digital Libraries (JCDL), 2014 IEEE/ACM Joint Conference on, pp. 371–374, IEEE, 2014.
[6]. W. Huang, Z. Wu, L. Chen, P. Mitra, and C. L. Giles, "A neural probabilistic model for context based citation recommendation.," in AAAI, pp. 2404–2410, 2015.
[7]. Q. He, J. Pei, D. Kifer, P. Mitra, and L. Giles, "Context-aware citation recommendation," in Proceedings of the 19th international conference on World wide web, pp. 421–430, ACM, 2010.
[8]. Q. He, D. Kifer, J. Pei, P. Mitra, and C. L. Giles, "Citation recommendation without author supervision," in Proceedings of the fourth ACM international conference on Web search and data mining, pp. 755–764, ACM, 2011.
[9]. S. Kataria, P. Mitra, and S. Bhatia, "Utilizing context in generative bayesian models for linked corpus.," in AAAI, vol. 10, p. 1, 2010.
[10]. W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles, and L. Rokach, "Recommending citations: translating papers into references," in Proceedings of the 21st ACM international conference on Information and knowledge management, pp. 1910–1914, ACM, 2012.
[11]. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 990–998, ACM, 2008.
[12]. J. Beel, B. Gipp, S. Langer, and C. Breitinger, "paper recommender systems: a literature survey," International Journal on Digital Libraries, vol. 17, no. 4, pp. 305–338, 2016.
[13]. S. Gupta and V. Varma, "Scientific article recommendation by using distributed representations of text and graph," in Proceedings of the 26th International Conference on World Wide Web Companion, pp. 1267– 1268, International World Wide Web Conferences Steering Committee, 2017.
[14]. H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele, and F. Xia, "Context based collaborative filtering for citation recommendation.," IEEE Access, vol. 3, no. 1, 2015.
[15]. T. Strohman, W. B. Croft, and D. Jensen, "Recommending citations for academic papers," in Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 705–706, ACM, 2007
[16]. X. Yu, Q. Gu, M. Zhou, and J. Han, "Citation prediction in heterogeneous bibliographic networks," in Proceedings of the 2012 SIAM International Conference on Data Mining, pp. 1119–1130, SIAM, 2012.