C-Rex: A Comprehensive System for Recommending In-Text Citations with Explanations

Michael Färber
Karlsruhe Institute of Technology
Karlsruhe, Germany
michael.faerber@kit.edu

Vinzenz Zinecker
Karlsruhe Institute of Technology
Karlsruhe, Germany
ucgvm@student.kit.edu

Isabela Bragaglia
Karlsruhe Institute of Technology
Karlsruhe, Germany
ufoue@student.kit.edu

Sebastian Celis
Karlsruhe Institute of Technology
Karlsruhe, Germany
uurkz@student.kit.edu

Maria Duma
University of Freiburg
Freiburg, Germany
maria.moloci@gmail.com

ABSTRACT

Finding suitable citations for scientific publications can be challenging and time-consuming. To this end, context-aware citation recommendation approaches that recommend publications as candidates for in-text citations have been developed. In this paper, we present C-Rex, a web-based demonstration system available at http://c-rex.org for context-aware citation recommendation based on the Neural Citation Network [5] and millions of publications from the Microsoft Academic Graph. Our system is one of the first online context-aware citation recommendation systems and the first to incorporate not only a deep learning recommendation approach, but also explanation components to help users better understand why papers were recommended. In our offline evaluation, our model performs similarly to the one presented in the original paper and can serve as a basic framework for further implementations. In our online evaluation, we found that the explanations of recommendations increased users' satisfaction.

KEYWORDS

recommender system, citations, scholarly data, neural network, information retrieval

1 MOTIVATION

Citing is an important concept of scientific writing, as it allows writers to give reference and to support mentioned concepts or claims. Due to the increasing amount of literature published every year [17], it can be hard to keep an overview of the literature on the
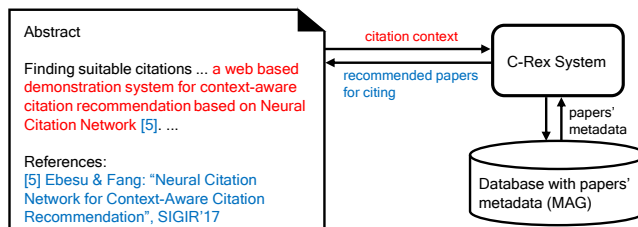
Figure 1: Our C-Rex system for citation recommendation.

topic at hand when writing publications.[1] This is where citation recommendation can help [1, 7]. The concept of context-aware citation recommendation, as considered in this paper, is that users input text (e.g., sentences) for which they require a citation, and the system recommends publications that may fit the text as citations. Figure 1 illustrates this concept. Context-aware citation recommendation requires obtaining a sufficient understanding of the citation context as input text; on the other hand, it requires identifying the top k relevant publications (e.g., $k = 5$) from a collection of up to millions of publications.

Several approaches to context-aware citation recommendation have been proposed [7]. However, existing running systems have not been trained, evaluated, or deployed based on the Microsoft Academic Graph (MAG) [22]. The MAG distinguishes itself from other publications' metadata collections in that it is particularly large – as of December 2020, modeling more than 120 million scientific publications and 240 million publications in total – and it also covers very recent publications in contrast to the widely used, rather small publication corpora [7] such as the ACL Anthology Network (ACL-AAN) [21] or the ACL Anthology Reference Corpus (ACL-ARC) [2]. Thus, we train, evaluate, and deploy the current state-of-the-art recommendation approach *Neural Citation Network* (NCN) [5] on the MAG data set. Second, almost no approach exists that has been deployed online as a running demonstration system. To the best of our knowledge, the only published context-aware citation recommendation systems are RefSeer [3] and CITEWERTs [9]. However, both systems lack explanations of the recommended publications, making it difficult for users to understand why a publication is recommended. Thus,

---

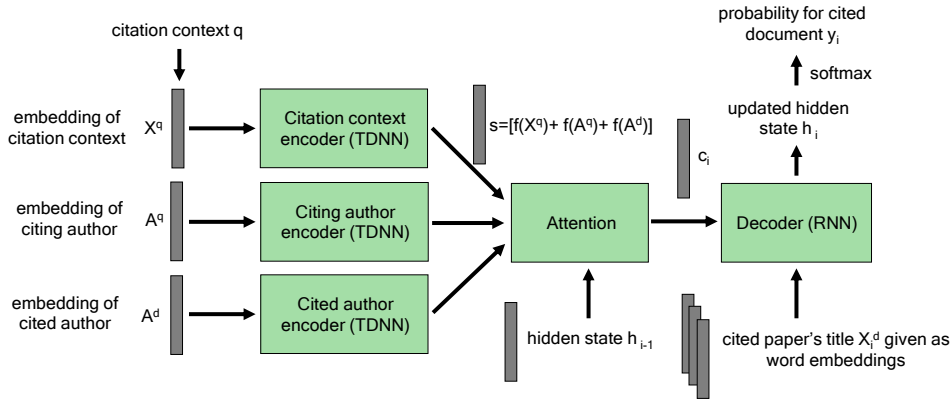[1] In this paper, "paper" and "publication" are used synonymously.

Figure 2: The neural citation network (NCN) architecture.

we implement and deploy an online demonstration system based on the NCN model that highlights decisive information despite the usage of a deep learning-based recommendation model. Our demonstration system *C-Rex*[2] is available at **http://c-rex.org/**.

Overall, we make the following contributions:

(1) We incorporate explainability components to the state-of-the-art NCN based on deep learning approach and indicate decisive words for each recommendation to the user.

(2) We adapt the implementation of the NCN to the MAG as and underlying data set, utilizing millions of up-to-date publications' metadata.

(3) We develop a UI/UX design that allows parallel requests and a semi-automatic pipeline that ensures updates so that recently published data can also be considered and recommended.

(4) We evaluate our citation recommendation system, C-Rex, both offline and online.

Frontend and backend source code is provided at https://github.com/sebastiancelis98/CitationRexApp, while information about the training process is available at https://github.com/michaelfaerber/NCN4MAG.

Our paper is structured as follows. In Section 2, we give an overview of our approach. In Section 3, we describe the evaluations performed offline and online. After outlining related work in Section 4, we conclude in Section 5.

2 APPROACH

The system's back end consists of a Flask server running a neural network in Python 3, as well as a PostgreSQL database containing the publications' metadata. The front end consists of a JavaScript web application running on the flutter framework. In the following, we describe the recommendation method and the data processing.

2.1 Recommendation Method

The starting point of our citation recommendation system is the Neural Citation Network (NCN) [5]. As depicted in Figure 2, this state-of-the-art approach for context-aware citation

recommendation relies on an encoder-decoder architecture: By analyzing the link between a citation context and a cited publication in a given publication collection, the model learns to "re-predict" the link for new but similar citation contexts. This is accomplished by using a time-delay neural network (TDNN), a CNN variant, to encode the words given in the citation context. The output is weighted by an attention layer and then decoded by a gated recurrent unit (GRU) to return the cited publication's title. Our system does not require the citing author's name as input.

The NCN has shown state-of-the-art performance given large training data sets [5]. Given the reimplementation of Färber et al. [8], we adapt the citation recommendation approach to the data of the Microsoft Academic Graph (MAG). Furthermore, to provide our system as an online demonstration system, we implement a corresponding user interface and interfaces between the front end and the back end. We use Flask and PyTorch as the leading frameworks for the implementation. To extract words from the input text that were decisive for the recommendation, we utilize the attention layer of the NCN, which lets us extract weights for each word in the titles of the recommended publications.

2.2 Dataset

The model is trained and tested based on context-citation pairs from existing publications. To this end, we use a current version (i.e., as of 2020-10) of the MAG [22]. In total, the MAG contains the metadata of more than 120 million scientific publications and more than 1 billion context–citation pairs, and thereby provides a solid and realistic basis for recommending suitable publications for the purpose of citing. Since the NCN has not been trained on data from the MAG, we need to adapt its implementation to MAG data. For our online demonstration system, we train the NCN model based on all *computer science publications in the MAG* (24.1 million publications). For the offline evaluation, we used the MAG computer science publications published between 2014 and 2019.

Furthermore, we deploy a semi-automatic process, including MAG data retrieval and neural network training, that allows us to update the neural network model based on fresh MAG data.[3] In this way, our system is kept up to date.

---

[2]*C-Rex* stands for *Comprehensive System for Recommending In-Text Citations with Explanations.*
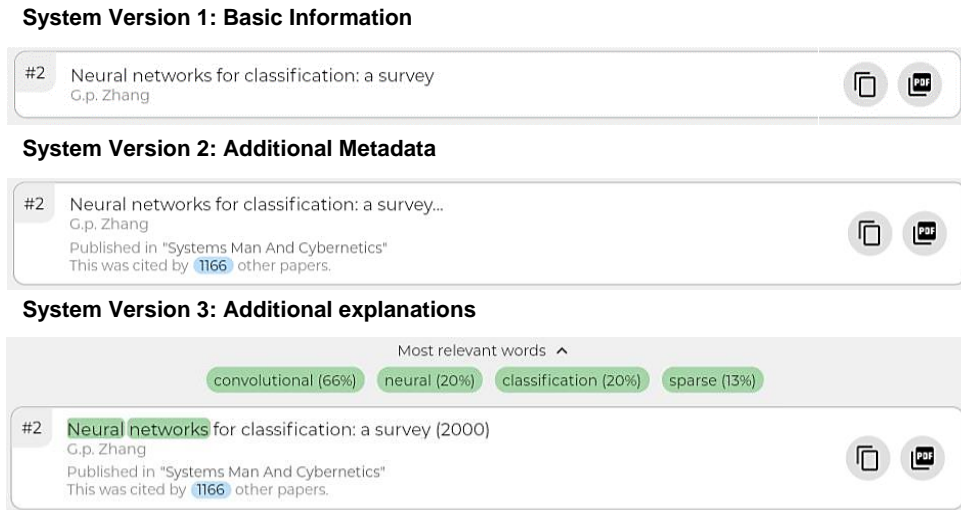
[3]See our documentation at https://github.com/michaelfaerber/NCN4MAG.

System Version 1: Basic Information

System Version 2: Additional Metadata

System Version 3: Additional explanations

Figure 3: Citation recommendation with different system setups.

Table 1: Results of the offline evaluation.

Model	Recall	MAP	MRR	NDCG
Neural Citation Network	0.1762	0.1050	0.1080	0.1279
TDNN-to-RNN	0.1496	0.0747	0.0767	0.1046

3 EVALUATION

In the following, we present an offline evaluation of our recommendation approach, as well as an online evaluation based on the running citation recommendation system.

3.1 Offline Evaluation

Evaluation Setting. Following the NCN model proposed by Ebesu et al. [5], we conducted an offline evaluation using English computer science publications in the MAG dataset published between 2014 and 2019. The NCN model relies on citation contexts and author information. After preprocessing the data, we obtained a set of 4.2 million context-citation pairs. We divided the data by year into three parts, taking 2017 as a split year: publications published before 2017 formed a training set with 3,576,180 citation-context pairs, publications published in 2017 were a validation set with 317,906 pairs, and those after 2017 were in the test set with 398,747 pairs. The preprocessing phases included tokenization, lemmatizing the data, removing stopwords, cutting citation contexts to a maximum of 100 words and citing titles to a maximum of 30 words. We considered the first 5 authors for each publication and replaced missing authors with the <UNK> token. We numericalized the citation contexts, citation titles, and authors using a vocabulary size of 20,000 terms. Further, we generated word embeddings for the numericalized data using a PyTorch embedding layer. The embeddings are initialized randomly from a standard normal distribution and learned during the training process. For the decoder part, we preselected candidate citation titles using the BM25 ranking function, similar to Ebesu et al. [5].

The hyperparameters were determined according to a set of experiments with varying hyperparametric settings. We varied all the parameters in the network and chose the best performing ones. The network capacity, which defines the embedding dimension, the number of convolutional filters and the hidden size of the RNN, was set to 256. The batch size was 64, and the number of recurrent layers in the RNN decoder was two. For the NCN context encoder, the region size filters were [4,4,5]; for the author encoder, the region size filters were [1,2]. During the training phase, we used gradient clipping at 5, dropout probability of 0.2 and Adam optimizer [10] for a total of 5 training iterations.

Evaluation Results. We evaluated two context-aware citation recommendation models, following the implementations proposed by Ebesu et al. [5]. The NCN is the original model with a TDNN-based encoder for citation contexts together with author information and a GRU-based decoder for cited titles of candidate documents. The second model, called TDNN-to-RNN, followed the NCN model but did not use author information. We report the models' performances in terms of recall, MAP, MRR and nDCG in Table 1.

The NCN model outperformed the TDNN-to-RNN model, showing a slight advantage of incorporating the author information in the recommendation model. We could not reproduce the NCN model's performance as reported in the original paper. We achieved a recall@10 of 0.176, which is less than the recall@10 of 0.29 reported by Ebesu et al [5]. Looking at the TDNN-to-RNN model, we obtained a recall@10 of 0.1496, which was similar to the originally reported recall@10 of 0.1579. The author information seems to have played an important role in Ebesu et al.'s study, since the recall@10 increased from 0.1579 to 0.291 when using author metadata. In our case, the author's information had no strong influence on the model performance. This can be explained by the fact that we used a different database than Ebensu et al. The author metadata in our MAG dataset is noisy to some extent, and author names may have been incomplete or duplicated [6]. Cleaning the MAG data and thus providing more accurate recommendations is considered

Table 2: Explanation goals.

Explanation goal	Description
Effectiveness	Help users make good decisions
Efficiency	Help users make decisions faster
Persuasiveness	Convince users to try or buy
Satisfaction	Increase the ease of use or enjoyment
Transparency	Explain how the system works
Trust	Increase users' confidence in the system

Table 3: Evaluation of the explanation goals.

	Basic Info.	Metadata	Explanation Components
Effectiveness	2.7	3.2	3.9
Efficiency	2.8	3.5	4.1
Persuasiveness	3.0	3.7	3.7
Satisfaction	2.8	4.0	4.1
Transparency	2.6	3.0	4.1
Trust	2.5	3.3	3.7

a future goal. Nevertheless, given previous research [5, 8], our recommendation model performs as expected when considering no author information. In light of the latest research on language models, we suggest considering neural network architectures based on Transformers and BERT to solve the local citation recommendation task.

3.2 Online Evaluation

Evaluation Setting. We designed the online evaluation to determine the explanatory capabilities of our recommendation system. To evaluate the extent to which the explanations succeeded in meeting the widely used explanation goals [27], we created three different versions of our system (see Figure 3):

- The **basic system** provides basic publications' metadata (title, author's name, and URL).
- The **metadata system** provides additional metadata, such as the publication year, venue, citation count, and publisher.
- The **explanation components system** provides metadata as well as highlighted words explaining the recommendations.

We divided the system's users into two groups: *experts*, who had written four or more scientific publications, and *non-experts*, who had written fewer than four scientific publications. Of the 15 participants, two were experts and 13 non-experts. They used the three systems for the citation recommendation and gave feedback based on effectiveness, efficiency, persuasiveness, satisfaction, and transparency [26] with descriptions in Table 2. All answers were given on a scale from one to five.

Evaluation Results. The results for each explanation goal are shown in Table 3. System 1 (basic system) obtained the lowest ratings across all explanation goals. In contrast, System 3, with explanation components, obtained the highest scores for all six explanation goals. It achieved a considerably higher rating regarding efficiency, satisfaction, and transparency. This means that our system with explanations helps users (1) to make faster decisions, (2) to solve the task of citing more satisfactorily, and (3) to understand what each recommendation is based on better than the other two systems.

4 RELATED WORK

Citation recommendation approaches. A thorough overview of citation recommendation approaches and datasets is provided by Färber et al. [7]. Citation recommendation approaches differ in their settings. If only a fragment of an input text document is used

as the citation context (e.g., a sentence [11, 14] or a window of 50 words), we call it *local citation recommendation* or *context-aware citation recommendation*. If there is no specific citation context, but instead the whole input text document or the document's abstract is used for the recommendation (see, e.g., [18, 20, 23, 24]), we call it *global citation recommendation* or *non-context-aware citation recommendation* (following He et al. [13]). Such differences in the setup make it particularly difficult to compare the performance of citation recommendation approaches [7].

McNee [19] in 2002 and Strohman et al. [23] in 2007 published the first global citation recommendation papers, while local citation recommendation was first introduced by He et al. [13] in 2010. He et al. expanded their model in [12]. Huang et al. [14] built upon the idea by translating specific keywords in the contexts (source language) into cited documents (target language), thereby creating a de facto machine translation system for citation recommendation.

Tang et al. [25] introduced embedding-based approaches to the field of context-aware citation recommendation. Jiang et al. [15, 16] also used embeddings in the context of cross-language global citation recommendation. Similar works were carried out by Cai et al. [4] and Zhang et al. [28] in 2018.

In this paper, we based our approach on the NCN approach by Ebesu et al. [5], as their approach has yielded state-of-the-art results and is currently widely used in the scientific community [8].

Citation recommendation demonstration systems. To the best of our knowledge, the only published context-aware citation recommendation demonstration systems are the RefSeer system [3] and the CITEWERTs system [9]. Both systems are based on traditional information retrieval techniques (e.g., LSI) and do not provide explanations of the recommended publications in their user interfaces.

5 CONCLUSION

With CɪᴛᴇRᴇx, we proposed a large-scale context-aware citation recommendation system to the public. Our system combines a state-of-the-art recommendation approach with the Microsoft Academic Graph as a very large and up-to-date dataset about publications, and is designed to present not only recommendations, but also indicate explanations. In the offline evaluation, we achieved similar results to Ebesu et al. [5] when considering no author information. Our online evaluation showed that explanation components enhance the users' interactions with the system. In future work, we will optimize the system with regard to runtime and expand the domains of publications used.

REFERENCES

[1] Jöran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *Int. J. Digit. Libr.* 17, 4 (2016), 305–338.

[2] Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan R. Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation* (Marrakech, Morocco) *(LREC'08)*.

[3] Kurt D Bollacker, Steve Lawrence, and C Lee Giles. 1998. CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. In *Proceedings of the 2nd International Conference on Autonomous Agents* (St. Paul, Minnepolis, USA) *(AGENTS'18)*. 116–123.

[4] Xiaoyan Cai, Junwei Han, and Libin Yang. 2018. Generative Adversarial Network Based Heterogeneous Bibliographic Network Representation for Personalized Citation Recommendation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)* (New Orleans, Louisiana, USA). 5747–5754.

[5] Travis Ebesu and Yi Fang. 2017. Neural Citation Network for Context-Aware Citation Recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Shinjuku, Tokyo, Japan) *(SIGIR '17)*. 1093–1096.

[6] M. Färber. 2019. The Microsoft Academic Knowledge Graph: A Linked Data Source with 8 Billion Triples of Scholarly Data. In *Proceedings of the 18th International Semantic Web Conference* (Auckland, New Zealand) *(ISWC'19)*. 113–129.

[7] Michael Färber and Adam Jatowt. 2020. Citation Recommendation: Approaches and Datasets. *Int. J. Digit. Libr.* 21, 4 (2020), 375–405.

[8] Michael Färber, Timo Klein, and Joan Sigloch. 2020. Neural Citation Recommendation: A Reproducibility Study. In *Proceedings of the 10th International Workshop on Bibliometric-enhanced Information Retrieval co-located with 42nd European Conference on Information Retrieval* (Lisbon, Portugal) *(BIR@ECIR'20)*. 66–74.

[9] Michael Färber, Alexander Thiemann, and Adam Jatowt. 2018. CITEWERTs: A System Combining Cite-Worthiness with Citation Recommendation. In *Proceedings of the 40th European Conference on IR Research* (Grenoble, France) *(ECIR'18)*. 815–819.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning.* MIT Press. http://www.deeplearningbook.org.

[11] Jing He, Jian-Yun Nie, Yang Lu, and Wayne Xin Zhao. 2012. Position-Aligned Translation Model for Citation Recommendation. In *Proceedings of the 19th International Symposium on String Processing and Information Retrieval* (Cartagena de Indias, Colombia) *(SPIRE'12)*. 251–263.

[12] Qi He, Daniel Kifer, Jian Pei, Prasenjit Mitra, and C. Lee Giles. 2011. Citation Recommendation Without Author Supervision. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining* (Hong Kong, China) *(WSDM '11)*. 755–764.

[13] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web* (Raleigh, NC, USA) *(WWW '10)*. 421–430.

[14] Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles, and Lior Rokach. 2012. Recommending Citations: Translating Papers into References. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management* (Maui, HI, USA) *(CIKM'12)*. 1910–1914.

[15] Zhuoren Jiang, Yao Lu, and Xiaozhong Liu. 2018. Cross-language Citation Recommendation via Publication Content and Citation Representation Fusion. In *Proceedings of the 18th ACM/IEEE Joint Conference on Digital Libraries* (Fort Worth, TX, USA) *(JCDL'18)*. 347–348.

[16] Zhuoren Jiang, Yue Yin, Liangcai Gao, Yao Lu, and Xiaozhong Liu. 2018. Cross-language Citation Recommendation via Hierarchical Representation Learning on Heterogeneous Graph. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (Ann Arbor, MI, USA) *(SIGIR'18)*. 635–644.

[17] R Johnson, A Watkinson, and M Mabe. 2018. *The STM Report: An Overview of Scientific and Scholarly Publishing.* Technical Report.

[18] Saurabh Kataria, Prasenjit Mitra, and Sumit Bhatia. 2010. Utilizing Context in Generative Bayesian Models for Linked Corpus. In *Proceedings of the 24th AAAI Conference on Artificial Intelligence* (Atlanta, Georgia, USA) *(AAAI'10)*. 1340–1345.

[19] Sean M. McNee, Istvan Albert, Dan Cosley, Prateep Gopalkrishnan, Shyong K. Lam, Al Mamunur Rashid, Joseph A. Konstan, and John Riedl. 2002. On the Recommending of Citations for Research Papers. In *Proceedings of the ACM 2002 Conference on Computer Supported Cooperative* (New Orleans, Louisiana, USA) *(CSCW'02)*. 116–125.

[20] Ramesh Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. 2008. Joint Latent Topic Models for Text and Citations. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD'08)*. 542–550.

[21] Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2013. The ACL anthology network corpus. *Language Resources and Evaluation* 47, 4 (2013), 919–944.

[22] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Paul Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web Companion* (Florence, Italy) *(WWW'15)*. 243–246.

[23] Trevor Strohman, W. Bruce Croft, and David D. Jensen. 2007. Recommending citations for academic papers. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands) *(SIGIR'07)*. 705–706.

[24] Jie Tang and Jing Zhang. 2009. A Discriminative Approach to Topic-Based Citation Recommendation. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining* (Bangkok, Thailand) *(PAKDD'09)*. 572–579.

[25] Xuewei Tang, Xiaojun Wan, and Xun Zhang. 2014. Cross-language Context-aware Citation Recommendation in Scientific Articles. In *Proceedings of the 37th International Conference on Research and Development in Information Retrieval* (Gold Coast , QLD, Australia) *(SIGIR '14)*. 817–826.

[26] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction* 22, 4-5 (2012), 399–439.

[27] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*, Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, 353–382.

[28] Ye Zhang, Libin Yang, Xiaoyan Cai, and Hang Dai. 2018. A Novel Personalized Citation Recommendation Approach Based on GAN. In *Proceedings of the 24th International Symposium on Foundations of Intelligent Systems* (Limassol, Cyprus) *(ISMIS'18)*. 268–278.