Using Citation-Context to Reduce Topic Drifting on Pure Citation-Based Recommendation

Anita Khadka
The Open University
Milton Keynes, UK
anita.khadka@open.ac.uk

Petr Knoth
The Open University
Milton Keynes, UK
petr.knoth@open.ac.uk

ABSTRACT

Recent works in the area of academic recommender systems have demonstrated the effectiveness of co-citation and citation closeness in related-document recommendations. However, documents recommended from such systems may drift away from the main theme of the query document. In this work, we investigate whether incorporating the textual information in close proximity to a citation as well as the citation position could reduce such drifting and further increase the performance of the recommender system. To investigate this, we run experiments with several recommendation methods on a newly created and now publicly available dataset containing 53 million unique citation-based records. We then conduct a user-based evaluation with domain-knowledgeable participants. Our results show that a new method based on the combination of Citation Proximity Analysis (CPA), topic modelling and word embeddings achieves more than 20% improvement in Normalised Discounted Cumulative Gain (nDCG) compared to CPA.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Document topic models*; *Information retrieval*;

KEYWORDS

Recommender systems; citation proximity analysis; citation-context; word embeddings; topic modelling

1 INTRODUCTION

Discovering relevant research publications from a huge corpora of digital libraries is a challenging problem. A recommender system is a valuable tool that can sift through the corpus and suggest the most relevant articles. Over the years, metadata information of documents such as *title* [28], *abstract* [25], *citation-counts* [4] have been used extensively as features for recommending research publications. However, metadata information only may not be sufficient to provide accurate recommendations. Titles and abstracts are sometimes written in a style intended to draw attention rather than to comprehensively describe a piece of work [3]. In comparison, full-text has not been as widely used as metadata. A major reason behind this may be the limited availability of the full-text documents. However, thanks to the open access movement, more full-texts have now become publicly available. Consequently, several recent studies examined full-text features to improve the quality of recommendations. For example, [20] used *citation-position*, which is the offset of the citation mention as used in the body of the document and [15] used *citation-context*, which is the text around the citation mention.

Recommendations based purely on citation-based methods may suffer from *topic drifting* [15]. Topic drifting can be defined as moving away from the main theme of the work. For instance, citations in the *Introduction* section are likely to introduce the domain and focus of the work. For example, a given paper may introduce 'Machine Learning' as domain and 'Image Classification' as focus, whereas citations in the *Related work* section are more focus on criticising or comparing one's work with others'. In our previous example, this may include different methods to classify images and may include citations for any underlying mathematics. If recommendations are based on citations only, and all citations are treated as equal then slightly off-topic papers can be recommended. For instance, papers focused on mathematical methods may be recommended when searching for image classification papers. We propose to combine *citation-position* and *citation-context* features to generate research paper recommendations using topic modelling [5] and word embeddings [29]. We believe that by knowing the context behind citing a particular reference, the performance of *citation-proximity* based recommender systems can be improved.

While previous work has discussed the idea of exploring citation-context for academic paper recommendations [2], to the best of our knowledge, this is the first work that provides a full implementation and evaluation of the idea of combining citation *context* and *position* using 53 million unique citation based records. The main contributions of this paper are as follows:

- A novel academic paper recommendation method combining citation position with textual information
- A publicly available citation-context based dataset containing 53 million unique records [18].
- A qualitative evaluation using domain-knowledgeable participants for a specific domain i.e. *Computer Science.*

The rest of the paper is organised as follows. We review related works in the next section. Then we present our dataset in Section 3. Section 4 describes the proposed method and Section 5 reports the qualitative results obtained from the user-study. Finally, Section 6 concludes the paper with ideas for future works.

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5901-6/18/10…$15.00
https://doi.org/10.1145/3240323.3240379

2 RELATED WORK

In this section, we provide an overview of those works that have applied or discussed citation-context and/or citation-position for recommendation tasks. [15] used citation-context to find topic *sensitive* related papers using a Vector Space Model (VSM) model focusing on exact terms matching on the content, which has the disadvantage that it may discard recommendations which are similar and related. In addition, their dataset contains only 1,273 papers which is significantly smaller than our dataset. Similarly, [6] used citation-context to index cited documents using a fixed length of 100 words with 50 words on either side of the citation mention. They used exact terms mentioned in the citation-context as a *reference* to index and find similar documents, assuming authors use meaningful terms to describe cited documents. [9] used a similar concept to [6], expanding their methods to search for similar terms.

Since its conceptualisation, Citation Proximity Analysis (CPA) [11] has successfully been used to compute the similarity of documents [7, 22], web pages [12] and authors [19]. More recently, CPA has been used to enhance the recommendation of academic papers by using co-citation information such as distance [20]. A similar idea was explored in [31] to provide linked-based recommendations in Wikipedia. Both citation-position and citation-context have been researched but separately, either for reducing topic-drift or increasing the performance of systems and mostly done in comparatively smaller dataset to our dataset. We propose to combine both features to get the benefits and perform experiment on a citation-based dataset containing 53 million records.

3 CITATION-CONTEXT DATASET (C2D)

Datasets containing full-text research publications are limited in the literature and their size is small, typically in the thousands. We present a dataset called Citation-Context Dataset (C2D) containing 53 million unique citation-based records. This dataset is created from two million full-text research publications provided by CORE [1]. First, we parsed the Portable Document Format (PDF) documents using the GROBID parser [24] and extracted 1, 715, 459 documents in the Text Encoding Initiatives (TEI) format. We then extracted *title*, *abstract*, *authors*, *published date* and *citation-context* from each document. For the *citation-context*, we extracted the position of citation mentions and the text around the cited documents. For our purpose, we created citation-context using three sentences; the sentence where the reference has been cited, the preceding, and the following sentence [16]. If the citation is located at the start or end of a paragraph, the preceding or following sentence is not extracted respectively. Researchers [6, 13] used a slightly different approach to create *citation-context*: they adopted a fixed window size of 100 words. However, we believe a fixed window may not always provide a meaningful explanation. For example, a sentence may be cut-short at a random point, and so can add noise to the dataset.

Several works have investigated the use of citation-context. These works either make use of relatively small datasets [6, 9, 16, 30], or the citation-context feature they used are not publicly available. [9] claimed to work on the CiteSeerX dataset[1] containing 1.8 million scientific articles and 41.5 million citation-contexts, however, the citation-context feature is no longer an active service from CiteSeerX[2]. Therefore, to the best of our knowledge, C2D is the first of its kind and provides the largest dataset to date. This dataset is available for academic purposes [18].

Algorithm 1: Pseudocode for C2D creation

```
Input: Corpus $\mathbf{D}$ containing full text documents
Output: Set $\mathbf{X}$ containing citation features of all the
        documents in $\mathbf{D}$
Initialise $\mathbf{X} = \emptyset$.
for each document $\mathbf{d} \in \mathbf{D}$ do
    for each citation $\mathbf{c} \in \mathbf{d}$ do
        Extract $\vec{x} = [f_1, ..., f_n]$, where $f_1$= Title, $f_2$= Author
          name, $f_3$= Citation-Context and so on.
        Add $\vec{x}$ to $\mathbf{X}$
    end
end
```

Each record in the dataset consists of a ReferenceID, SourceID, ChapterNumber, ParagraphNumber, SentenceNumber, Title, PublishedDate, Authors, TextBeforeRefMention, TextWhereRefMention and TextAfterRefMention where ReferenceID is the unique identifier of the paper being cited, sourceID is the identifier of a citing document. ChapterNumber, ParagraphNumber and SentenceNumber are the chapter, paragraph and sentence numbers respectively of the citing document where the ReferenceID is cited. Similarly, Title, PublicationDate and Authors are the title, published date and authors of the cited document. Finally TextWhereRefMention, TextBeforeRefMention and TextAfterRefMention indicate the sentences where the reference has been cited, the preceding and the following sentences respectively, further details on the dataset can be found in [18]. Once these features are extracted, we pre-process the data using Natural Language Processing (NLP) techniques such as tokenisation and stop-word removal. We describe the feature extraction and dataset creation processes in Algorithm 1.

4 CITATION PROXIMITY-CONTEXT BASED METHOD

The proposed method delivers recommendations for a query document in a two-stage process: first, we employed CPA to generate a set of documents which are cited in close proximities and ranked on the basis of higher weighted average values of Citation Proximity Index (CPI). In the second stage, we infer topics from each recommendation generated in the first stage and compare it to that of the query document. For this, each topic is projected into multidimensional continuous-valued vectors to generate semantically similar topics. The pseudocode of the process is presented in Algorithm 2 and explained in Sections 4.1 to 4.4.

4.1 Citation Proximity Analysis (CPA)

We used CPA [11] to produce an initial set of relevant articles (**R**). In this method, co-cited documents are strongly or weakly related to each other based on their locations. For example, if two citations appear in the same sentence, this method assumes a stronger relation between them than a pair of citations appearing in different sentences or paragraphs. The strength of relationships between

[1] http://csxstatic.ist.psu.edu/about/data

[2] http://csxstatic.ist.psu.edu/about

Algorithm 2: Pseudocode for generating recommendations.

```
Input: D, X and query document $\vec{q}$
Output: $n$ recommendations
Run LDA on D to generate model $L$ and topic set $T$
Create $\mathbf{R} \subseteq \mathbf{X}$ using CPA
for $each\ \vec{x} \in \mathbf{R}$ do
    Assign topic $t$ to $\vec{x}$ using $L$.
    Find $\{sw\}$ semantically similar top word from the topic $t$
      using Glove's Wikipedia corpus.
    Find vector representations $\vec{v}$ of $\{sw\}$
    Calculate Cosine similarity between $\vec{q}$ and $\vec{v}$
end
Reorder R into descending order of cosine similarity
Output top $n$ items in R
```

citations at different levels appearing in same sentence, paragraph, chapter, journal, same journal but different versions are 1, $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ respectively. However, we focus only up-to chapter level strength because different versions of journals are minimal in C2D. Once the CPI values of each pair of co-cited documents are computed, the strength of relationships is calculated by computing the weighted average of those values. Based on this strength, the recommendations are ranked.

4.2 Topic inference from citation-context

Topic models are widely used statistical models to infer latent topics from a corpus of documents. According to [5], each document is considered as a collection of words and is represented as a random mixtures over latent topics and each topic is characterised by a distribution over all words. We followed the generative process provided by [5] to discover latent topics from the documents in the corpus **D**. The idea behind applying topic modelling on citation-context is to cluster the documents which are focused on the same concept but portrayed in different ways by different authors. We believe by inferring latent topics from documents and clustering them according to similar themes will help reduce topic drifting.

Inferring latent topics from short texts is a tricky task due to the lack of word co-occurrences which can also result in an incoherent analysis of results. Therefore, we took inspiration from [14] to alleviate the short texts issue. [14] performed an empirical study on topic generations using a Twitter dataset, by aggregating tweets from a user and creating a long text as a document. We used corpus **D** for inferring topics and assigned those topics to relevant citation-context. We believe researchers tend to cite documents for a specific reason. So, we consider only one topic for each citation-context and each topic is described by a number of words.

4.3 Topic to word embeddings

The concept of word embedding in a vector space has widely been used in the NLP domain. However, to the best of our knowledge, the use of word embeddings in the research publication recommendation domain is still in its infancy. The reason for this may be the restriction on the public availability of full-text content. We assume that the distributed representation of words with semantic mappings can broaden the coverage of relevant and recommendable documents, whereas the systems which use the exact content matching or pure citation-based methods can have comparably limited recommendable items. Additionally, word embeddings can capture the subtle semantic relationships between terms in the corpus. For example, Capital and France are related to Paris (i.e. $\vec{France} + \vec{Capital} \approx \vec{Paris}$) [27]. Taking this idea as an inspiration, the penultimate stage of our model projects words to a vector space. Each topic T_i of i^{th} citation-context is a set of words $\{w_1, w_2, ..., w_m\}$ where m is the number of words assigned for each topic discussed in Section 4.2.

Then, we used a statistical model '*Glove*' introduced by [29] which focuses on global vectors, where prediction is done using a statistical method rather than a probabilistic method. *Glove* has been shown to perform better than the state-of-the-art word embeddings model (i.e. Word2Vec[27]) [29]. We used the pre-trained publicly available vector representation of the Wikipedia corpus[3] provided by [29]. The immense and diverse range of enriched topics embedded in Wikipedia motivated us to choose it. We considered topic T_i has 5 words i.e.($w_1, ..., w_5$) and used them as positive input to the model where we computed a mean of the projection weight vectors of the given words. We then obtained a single vector representation v_i for topic T_i.

4.4 Final recommendations

We use a cosine similarity metric between the vector representations to measure the similarity between query documents and the initial list of recommendations (**R**).

$$d(v_q, v_x) = \frac{v_q.v_x}{\|v_q\|\|v_x\|} \tag{1}$$

where v_q and v_x are the vector representations of the query and document recommended respectively. Finally, the recommended documents are ranked based on decreasing cosine similarity value between q and x; where $x \in \mathbf{R}$.

5 EXPERIMENTAL EVALUATION

We conducted an intrinsic evaluation of the C2D by creating a user-based survey where 14 domain-knowledge-able participants took part to evaluate our proposed method and the baseline systems listed in Table 1. We chose the domain of *Computer Science*, specifically *Machine learning and Data Mining*, and selected five query documents randomly. We then generated five recommendations for each query document based on five different methods, including our proposed method (see Table 1). The users were asked to rate each recommendation on a Likert scale [21]. The scale has four options to chose from, namely, *Extremely Relevant*, *Very Relevant*, *Somewhat Relevant* and *Not Relevant*.

5.1 Results And Discussion

According to Information Retrieval (IR), top-ranked items are the most important and relevant to the query. Typically users are most likely to scan the top few recommendations. So, as an evaluation metric, we used Normalised Discounted Cumulative Gain (nDCG). This metric is well suited to evaluate recommendations on a non-binary judgement of relevance and rewards recommended items at the top of the list more than the lower rank [26].

[3] https://nlp.stanford.edu/projects/glove/

Method Name	Formula	Description
$Co-Citation$ [32]	$cocit^{ab} = \|Doc\|_{a \in Doc \wedge b \in Doc}$	where references a and b are co-cited documents in a citing document Doc.
CPA [11]	$cpa^{ab} = \frac{\sum_{i=1}^{n}(w_i^{ab})}{n}$	where w_i is the i^{th} value of CPI (weighted) between the co-cited documents a and b; n is the number of cpi values of a and b
$CPA_{MeanProx}$ [20]	$cpa_{Mean}^{ab} = \frac{\|Doc\|_{a \in Doc \wedge b \in Doc}}{\log(mean\{d_1^{ab}, ..., d_n^{ab}\})}$	where d_1^{ab} and d_n^{ab} are the distances between the co-cited documents a and b.
$TF-IDF$ [17]	$W_{t,d} = tf_{t,d} * \log(\frac{N}{df_t})$	where $W_{t,d}$ is the weight for a term t in a document d, $tf_{t,d}$ is number of occurrences of t in d and df_t is number of documents containing t. N is the total number of documents

Table 1: List of baseline methods to compare our proposed method $CPA_{ContextEmbed}$ illustrated in Algorithm 2

Method Name	nDCG@3	nDCG@5	*p*-value against $CPA_{ContextEmbed}$
$Co-Citation$	0.717	0.864	<< 0.01
CPA	0.575	0.688	<< 0.01
$CPA_{MeanProx}$	0.659	0.782	<< 0.01
$TF-IDF$	0.764	0.865	<< 0.01
$CPA_{ContextEmbed}$	**0.838**	**0.902**	–

Table 2: nDCG results at 3^{rd} and 5^{th} positions of recommendations for proposed and baseline methods including *p*-value from *t*-test

As explained earlier, users may be interested in the *top−N* ranked recommendations so we chose *nDCG@N*, where N (= 3, 5) is the number of papers recommended by baselines and our proposed method. Due to space limitations, we have only illustrated a graph of nDCG@5 in Figure 1. However, Table 2 shows nDCG results at both 3^{rd} and 5^{th} positions. The results show that our proposed algorithm $CPA_{ContextEmbed}$ performed better than the baseline algorithms. However, results from both proximity-based citation analysis (CPA and $CPA_{MeanProx}$) are surprising; these performed worst, with nDCG@5 values of 0.688 and 0.782 respectively in comparison to other methods. According to [11, 20, 31], the performance of CPA is higher in comparison to co-citation. An investigation of the produced recommendations and evaluation data suggests that length of the documents has a higher impact on the proximity-based approach. This would suggest that looking into ways of normalising for document length might be a plausible avenue for future research. We also compared *nDCG@*3 and *nDCG@*5 for $CPA_{ContextEmbed}$ with other baseline methods using $t-test$ [10]. We achieved over 95% confidence of significant positive differences with $p-value << 0.01$. And, to check the homogeneity in the ratings of participants, we performed an inter-rater reliability check using Cronbanch's alpha [8] and obtained the value of 0.904, which signifies participants are in agreement.

6 CONCLUSION AND FUTURE WORK

In this work, we explored a novel method using citation information to improve the performance of research paper recommendations. In particular, we used citation-context by combining it with citation-proximity features to alleviate topic drift by using techniques like topic modelling with word embeddings to find semantically similar topics. Our results show that by incorporating two features, the performance of recommendations increased by 20% in comparison to the original CPA based method. As our evaluators were knowledgeable in the domain, the credential of the judgement is qualitatively valuable. Additionally, we believe that the dataset generated as part of this work will fuel other citation-context based research.

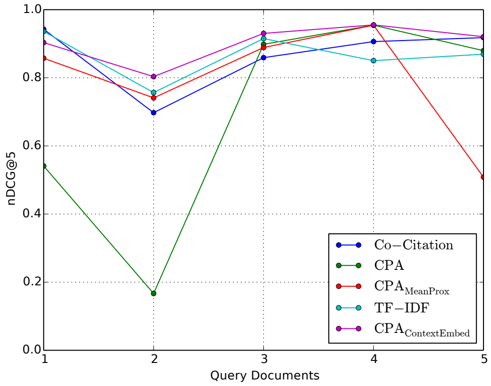

Figure 1: nDCG@5 for the proposed and baseline methods. (Best viewed in colour)

In the future, we intend to build and incorporate user profiles with papers' *external* characteristics such as the number of citations, type of publication, and the prestige of the publication venue into the recommendation process. We believe it will boost the performance of systems for finding relevant papers based on users' interests. Therefore, we aim to explore the impact of these characteristics including recency. The reason behind including recency is that according to most of our participants, although the recommendations were topically relevant, some recommendations were outdated. Additionally, we intend to evaluate our algorithms further by collecting users' activities such as Click-Through Rate (CTR), download and co-downloads including the purpose of selecting particular recommendations. We will compare our methods with other additional baseline approaches such as [13, 23]. Finally, as discussed in Section 5.1, our foremost priority will be investigating the impact of the length of documents.

REFERENCES

[1] CORE Admin. 2018. CORE reaches a new milestone: 75 million metadata and 6 million full text. (2018). https://blog.core.ac.uk/2017/05/16/core-reaches-a-new-milestone-75-million-metadata-and-6-million-full-text

[2] Joeran Beel. 2017. Virtual Citation Proximity (VCP): Calculating Co-Citation-Proximity-Based Document Relatedness for Uncited Documents with Machine Learning [Proposal]. (10 2017). https://doi.org/10.13140/RG.2.2.18759.39842

[3] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (01 Nov 2016), 305–338. https://doi.org/10.1007/s00799-015-0156-0

[4] Steven Bethard and Dan Jurafsky. 2010. Who Should I Cite: Learning Literature Search Models from Citation Behavior. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM '10)*. ACM, New York, NY, USA, 609–618. https://doi.org/10.1145/1871437.1871517

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *J. Mach. Learn. Res.* 3 (March 2003), 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937

[6] Shannon Bradshaw. 2003. Reference Directed Indexing: Redeeming Relevance for Subject Search in Citation Indexes. In *Research and Advanced Technology for Digital Libraries*, Traugott Koch and Ingeborg Torvik Sølvberg (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 499–510.

[7] Giovanni Colavizza, Kevin W. Boyack, Nees Jan van Eck, and Ludo Waltman. 2018. The Closer the Better: Similarity of Publication Pairs at Different Cocitation Levels. *Journal of the Association for Information Science and Technology* 69, 4 (2018), 600–609. https://doi.org/10.1002/asi.23981

[8] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 3 (01 Sep 1951), 297–334. https://doi.org/10.1007/BF02310555

[9] Metin Doslu and Haluk O. Bingol. 2016. Context sensitive article ranking with citation context analysis. *Scientometrics* 108, 2 (01 Aug 2016), 653–671. https://doi.org/10.1007/s11192-016-1982-6

[10] John E. Freund. 1992. *Mathematical Statistics*. Prentice-Hall International, Inc, Englewood Cliffs, New Jersey, USA.

[11] Bela Gipp and Jöran Beel. 2009. Citation Proximity Analysis (CPA) : A New Approach for Identifying Related Work Based on Co-Citation Analysis. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics, vol. 1*, Birger Larsen (Ed.). BIREME/PANO/WHO, São Paulo, 571–575. http://www.sciplore.org/wp-content/papercite-data/pdf/gipp09a.pdf

[12] Bela Gipp, Adriana Taylor, and Jöran Beel. 2010. Link Proximity Analysis - Clustering Websites by Examining Link Proximity. In *Research and Advanced Technology for Digital Libraries*, Mounia Lalmas, Joemon Jose, Andreas Rauber, Fabrizio Sebastiani, and Ingo Frommholz (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 449–452.

[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 (WWW '10)*. ACM, New York, NY, USA, 421–430. https://doi.org/10.1145/1772690.1772734

[14] Liangjie Hong and Brian D. Davison. 2010. Empirical Study of Topic Modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics (SOMA '10)*. ACM, New York, NY, USA, 80–88. https://doi.org/10.1145/1964858.1964870

[15] Shen Huang, Gui-Rong Xue, Ben-Yu Zhang, Zheng Chen, Yong Yu, and Wei-Ying Ma. 2004. TSSP: A Reinforcement Algorithm to Find Related Papers. In *Web Intelligence, 2004. WI 2004. Proceedings. IEEE/WIC/ACM International Conference on.* 117–123. https://doi.org/10.1109/WI.2004.10038

[16] Wenyi Huang, Saurabh Kataria, Cornelia Caragea, Prasenjit Mitra, C. Lee Giles, and Lior Rokach. 2012. Recommending citations: translating papers into references. In *CIKM*.

[17] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–21. https://doi.org/10.1108/eb026526 arXiv:https://doi.org/10.1108/eb026526

[18] Anita Khadka. 2018. Citation-Context Dataset (C2D). (8 2018). https://doi.org/10.21954/ou.rd.6865298

[19] Ha Jin Kim, Yoo Kyung Jeong, and Min Song. 2016. Content- and proximity-based author co-citation analysis using citation sentences. *Journal of Informetrics* 10, 4 (2016), 954 – 966. https://doi.org/10.1016/j.joi.2016.07.007

[20] Petr Knoth and Anita Khadka. 2017. Can we do better than co-citations? Bringing Citation Proximity Analysis from idea to practice in research articles recommendation. In *CEUR Workshop Proceedings*, Vol. 1888. 14–25.

[21] RA Likert. 1932. A Technique for Measurement of Attitudes. *Archives of Psychology* 22 (01 1932), 1–55.

[22] Shengbo Liu and Chaomei Chen. 2012. The proximity of co-citation. *Scientometrics* 91, 2 (01 May 2012), 495–511. https://doi.org/10.1007/s11192-011-0575-7

[23] Xiaozhong Liu, Yingying Yu, Chun Guo, and Yizhou Sun. 2014. Meta-Path-Based Ranking with Pseudo Relevance Feedback on Heterogeneous Graph for Citation Recommendation. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM '14)*. ACM, New York, NY, USA, 121–130. https://doi.org/10.1145/2661829.2661965

[24] Patrice Lopez. 2009. GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction for Scholarship Publications. In *Research and Advanced Technology for Digital Libraries*, Maristella Agosti, José Borbinha, Sarantos Kapidakis, Christos Papatheodorou, and Giannis Tsakonas (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 473–474.

[25] Yang Lu, Jing He, Dongdong Shan, and Hongfei Yan. 2011. Recommending Citations with Translation Model. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management (CIKM '11)*. ACM, New York, NY, USA, 2017–2020. https://doi.org/10.1145/2063576.2063879

[26] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[27] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2 (NIPS'13)*. Curran Associates Inc., USA, 3111–3119. http://dl.acm.org/citation.cfm?id=2999792.2999959

[28] Cristiano Nascimento, Alberto H.F. Laender, Altigran S. da Silva, and Marcos André Gonçalves. 2011. A Source Independent Framework for Research Paper Recommendation. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*. ACM, New York, NY, USA, 297–306. https://doi.org/10.1145/1998076.1998132

[29] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[30] Anna Ritchie, Simone Teufel, and Stephen Robertson. 2006. How to Find Better Index Terms Through Citations. In *Proceedings of the Workshop on How Can Computational Linguistics Improve Information Retrieval? (CLIIR '06)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 25–32. http://dl.acm.org/citation.cfm?id=1629808.1629813

[31] M. Schwarzer, M. Schubotz, N. Meuschke, C. Breitinger, V. Markl, and B. Gipp. 2016. Evaluating link-based recommendations for Wikipedia. In *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. 191–200.

[32] Henry Small. 1973. Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. *Journal of the American Society for Information Science* 24 (07 1973), 265 – 269.