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 concept 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 achieve 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, word-embeddings, topic modelling, citation-context

1 INTRODUCTION

Discovering relevant research publications from the 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, meta-data information of documents such as *title* [19], *abstract* [16], *citation-counts* [2] have been used extensively as features for the scholarly recommendations. However, use of meta-data information only, may not be enough as information, such as title and abstract, are sometimes written in a style to draw attention rather than to comprehensively describe a piece of work [1]. In comparison, full-text has not been as widely used as meta-data. 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, [14] used *citation-position*, which is the position of the cited document within the document and [11] used *citation-context*, which is the content around the citation.

Recommendations based purely on citation-based methods may suffer from *topic drifting* [11]. Topic drifting can be defined as moving away from the main concept. For instance, citations in the *Introduction* section are likely to introduce the domain and focus of the work such as 'Machine Learning (ML)' and 'Image Classification' respectively. Whereas citations in the *Related work* section are to criticise or compare one's work with others', which may include different methods to classify images and may include citations for underlying mathematics, as ML relies on Mathematics too. Therefore, if recommendations are based on citations only, this could result from treating all citations as equal and can recommend papers on mathematics when searching for image classification. We propose to combine *citation-position* and *citation-context* features to generate research paper recommendations using topic modelling [3] and word embedding [20]. We believe that by knowing the context behind citing a particular reference, we can improve the performance of *citation-proximity* based recommender systems.

To the best of our knowledge, this is the first study carried out to combine citation *context* and *position* for recommendation. The main contributions of this paper are as follows:

(1) A novel method combining citation position with textual information to map the semantic relationship between cited documents.
(2) A publicly available large scale citation-context based dataset for research.
(3) A qualitative evaluation using domain-knowledgeable participants for a specific domain.

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

2 RELATED WORKS

We focus on citation-context and pure citation-based related works. [11] used citation-context to find related papers using a Vector Space Model (VSM) model focusing on the exact terms matching on the content. Matching exact terms has the disadvantage that it may discard recommendations which are similar and related. In addition, their dataset is comparatively small containing only 1,273 research papers. Similarly, Bradshaw [4] used citation-context

RecSys2018, October 2018, Vancouver, Canada

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$ —*This is a pre-print version—
https://doi.org/10.1145/nnnnnnn.nnnnnnn

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. Doslu et al. [6] used the similar concept as [4], expanding their methods to search for similar terms as well.

Most recently, [14] have shown that it is possible to improve recommendation performance using co-citation information such as distance. Similarly, [22] analysed Wikipedia links using citation-based approaches (CPA and Co-citation) along-with a text-based technique called MoreLikeThis (MLT). They claimed that citation-based approaches are complementary to text-based methods.

3 CITATION-CONTEXT DATASET (C2D)

Datasets containing full-text research publications are limited in literature and their size is small, typically in thousands. We present a new large scale dataset called Citation-Context Dataset (C2D) containing 53 million unique citation-based records. This dataset is created from two million full-text research publications in Portable Document Format (PDF) format, provided by *source yet to be provided*[1] and extracted 1, 715, 459 documents in Text Encoding Initiatives (TEI) format. We then extracted various information like *title*, *abstract*, *authors*, *published date* and *citation-context* from each document. In the following subsection, we explain our assumptions, extraction of citation-context and creation of the dataset.

3.1 Citation-Context

We extracted the position of citation mentions including text around the cited documents; we term this information '*citation-context*'. For our purpose, we created citation-context using three sentences adopting a similar assumption to [12]; the sentence where the reference has been cited, the preceding, and the following sentence. At the start or end of a paragraph, the preceding or following sentence is not extracted respectively. Another way of creating *citation-context* is using a fixed window size of N words. Researchers [4, 9] adopted a fixed window size of 100 words. We believe a fixed window may not always provide a meaningful explanation. For example, if a sentence is cut-short at a random point, its meaning could be different or add noise to the dataset.

There have been several works carried out using citation-context such as [4, 6, 12, 21]; however, as best we know, a large-scale dataset containing citation-context information is yet to be published publicly. Although, [6] claimed to work on the CiteSeerX dataset[2] containing 1.8 million scientific articles and 41.5 million citation-contexts. CiteSeerX is a publicly accessible digital library so we investigated for collecting such data, but the citation-context feature is no longer an active service from CiteSeerX[3]. C2D is provided in the Tab Separated Value (TSV) format and can be downloaded from *link yet to be provided*[4] for research. Each record in the dataset is assumed to be a document with features such as ReferenceID, SourceID, SentenceNumber, ParagraphNumber, ChapterNumber, Title, PublishedDate, Authors, TextBeforeRefMention, TextWhereRefMention and TextAfterRefMention. Once the features are extracted, we further clean the data using various Natural Language Processing (NLP) techniques such as tokenisation and stop-word removal. We illustrate the above-mentioned process for feature extraction and dataset creation in Algorithm 1.

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

4 CITATION PROXIMITY-CONTEXT BASED METHOD

The proposed semantically enhanced method delivers relevant recommendations for a query document in a two-stage process: first, we employed Citation Proximity Analysis (CPA) to generate a set of relevant 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 recommended list generated in the first stage and compare it to that of the query document. For this, each topic is projected into multi-dimensional continuous-valued vectors to generate semantically similar topics. The pseudo-code of the process is presented in Algorithm 2 and described in Sections 4.1 to 4.4.

Algorithm 2: Pseudo-code for generating recommendations.

Input: $\mathbf{D}$, $\mathbf{X}$ and query document $\vec{q}$
Output: n recommendations

Run LDA on $\mathbf{D}$ 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 $\{sw_1, ..., sw_m\}$ semantically similar words of $\{w_1, ..., w_m\}$ using Glove's Wikipedia corpus.
    Find vector representations $\vec{v}$ of top one semantically similar word from $\{sw_1, ..., sw_m\}$
    Calculate Cosine similarity between $\vec{q}$ and $\vec{v}$ using Equation (1)
end
Reorder $\mathbf{R}$ into descending order of cosine similarity
Output top n items in $\mathbf{R}$

[1]To preserve anonymity, source is not provided
[2]http://csxstatic.ist.psu.edu/about/data
[3]http://csxstatic.ist.psu.edu/about
[4]To preserve anonymity, the link to the dataset is not provided but a sample of the dataset has supplied.

4.1 Citation Proximity Analysis (CPA)

We used CPA [8] 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 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 the number of journals and 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 a widely used concept to infer latent topics from a corpus of documents. According to [3], documents are considered as random mixtures over latent topics and each topic is characterised by a distribution over all the words. We followed the generative process provided by [3] to discover latent topics from the documents in the corpus **D**. The idea of 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, as finding different mentions for the same idea helps cluster the meaningful analysis of the research domain [24].

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 [10] to alleviate the short-texts issue. [10] 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. In our case, we used the corpus **D** for inferring topics and assigned those topics to relevant citation-contexts. We believe researchers tend to cite document for a specific reason. So, we consider only one topic for each citation-context. Furthermore, 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 Natural Language Processing (NLP) domain. However, to the best of our knowledge, word-embedding in the research publication recommendation domain is in its infancy. The reason can be the restriction on the public availability of full-text content. The success of distributed representation of words with semantic meaning can broaden the coverage of relevant and recommendable documents, whereas the systems which use the feature of exact content matching or pure citation based can have comparably limited recommendable items. Additionally, word-embedding can capture the subtle semantic relationships between terms in the corpus. For example, Capital and France are related to Paris (i.e. $\vec{France} + \vec{Capital} \approx \vec{Paris}$) [18]. Taking this idea as an inspiration, the penultimate stage of our model projects words to 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 [20] which focuses on global vectors. Prediction is done using statistical calculation rather than a probabilistic method and is better than the state-of-the-art word-embeddings model; Word2Vec [20]. We used the pre-trained publicly available vector representation of the Wikipedia corpus[5] provided by [20]. The immense and diverse range of enriched topics embedded in Wikipedia motivated us to choose this corpus. We considered T_i has 5 words i.e.$(w_1, ..., w_5)$ and used them as positive input, assuming that vector addition of positive terms can produce meaningful results. We then obtain a single vector representation v_i for each topic T_i.

4.4 Final recommendations

We use 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 between q and x; where $x \in \mathbf{R}$.

5 EXPERIMENTAL EVALUATION

We conducted an intrinsic evaluation on Citation-Context Dataset (C2D) by creating a user-based survey where 14 domain-knowledgeable participants took part to evaluate our proposed method and baseline systems listed inTable 1. We chose the domain of *Computer Science* specific to *Machine learning and Data Mining* and selected five query documents randomly. We then generated five recommendations for each query document based on each algorithm. The users were asked to rate each recommendation on a Likert scale [15]. 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) method, top-ranked items on the list are the most important and relevant item to the query item, as users are most likely to scan the top few recommendations. So, as an evaluation metric, we used Normalised Discounted Cumulative Gain (nDCG) which is increasingly adopted with machine learning techniques for ranking [17]. This method is well suited to evaluate recommendations of a non-binary judgement of relevance and it rewards recommended items at the top of the list more than the lower rank. Typically users may be interested in the $top - N$ ranked recommendations so we chose $nDCG@N$ where N is the number papers recommended by our proposed method, and the chosen values of $N = 3, 5$ for evaluation. Due to space limits, 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. According to nDCG results, our proposed algorithm $CPA_{ContextEmbed}$ performed better than baseline algorithms. However, results from both proximity-based citation analysis (CPA and $CPA_{MeanProx}$)

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

Method Name	Formula	Description
$Co-Citation$ [23]	$cocit^{ab} = \|Doc\|_{a \in Doc \wedge b \in Doc}$	where references a and b co-cited.
CPA [8]	$cpa^{ab} = \frac{\sum_{i=1}^{n}(w_i^{ab})}{n}$	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 value of a and b
$CPA_{MeanProx}$ [14]	$cpa_{Mean}^{ab} = \frac{\|Doc\|_{a \in Doc \wedge b \in Doc}}{\log(mean\{d_1^{ab}, ..., d_n^{ab}\})}$	where d_n is the last distance between the co-cited documents a and b.
$TF-IDF$ [13]	$W_{t,d} = tf_{t,d} * \log(\frac{N}{df_t})$	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 total number of documents

Table 1: This table illustrates the list of baseline methods to compare our proposed method $CPA_{ContextEmbed}$ illustrated in Algorithm 2

are surprising; these performed worst, with nDCG@5 values of 0.688 and 0.782 respectively in comparison to other methods. According to [8, 14, 22], the performance of CPA is higher in comparison to co-citation so we investigated our evaluation dataset and we believe that length of documents has a higher impact on the proximity-based approach; however further experiments are required to support this theory. On top of that, the size of our evaluation dataset should be increased and we should conduct an online evaluation using features like Click-Through Rate (CTR), number of downloads & co-downloads along-with user purpose in selecting particular recommendations. We believe the latter may give the subjective view of participants. Currently, we do not hold such information of the choice of participants apart from the topical relevance between the query and recommended documents.

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

We also compared $nDCG@3$ and $nDCG@5$ for $CPA_{ContextEmbed}$ with other baseline methods using $t-test$ [7]. We achieved over 95% confidence of significant positive differences with $p-value << 0.01$. Similarly, to check the homogeneity in the ratings of participants, we performed an inter-rater reliability check using Cronbanch's alpha [5] 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 scholarly paper recommendation. In particular, we used citation-context by combining it with pure citation-based features to alleviate topic drift. The use of techniques

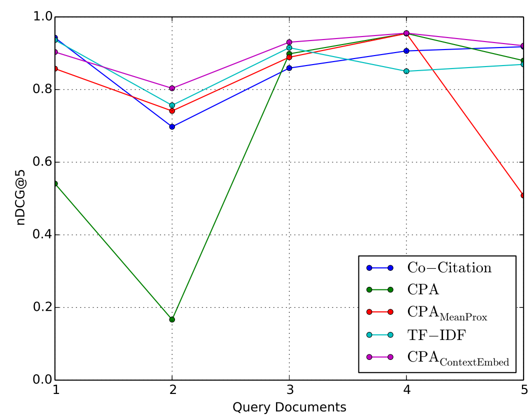

Figure 1: nDCG@5 for the proposed and baseline methods. (Best viewed in colour)

like topic modelling with word-embedding helps to find semantically similar concepts. Our results show that by incorporating two features, the performance of recommendations increased by 20% in comparison to the original CPA based method. In our study, the evaluators were knowledgeable in the domain so the credential of judgement is qualitatively valuable. Additionally, we believe that public availability of our dataset will fuel other citation-context based methods.

As future work, we intend to evaluate our algorithms by collecting users' activities such as CTR, download and co-downloads of recommendations. As discussed in Section 5.1, our foremost future work will be investigating the impact of the length of documents. On top of that, we will explore the inclusion of time-series and recency, as most of our participants provided feedback that, although the recommended results are topically relevant, the lists include dated papers. Finally, the current model heavily relies on offline computation and is not suitable for real-time recommendation. Therefore, we would like to focus on optimising the model in order to make it real-time operable.

REFERENCES

[1] 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

[2] 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

[3] 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

[4] 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.

[5] 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

[6] 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

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

[8] Bela Gipp and Joeran Beel. 2009. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. (07 2009).

[9] 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

[10] 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

[11] 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

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

[13] 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

[14] Petr Knoth and Anita Khadka. 2017. Can we do better than co-citations? - Bringing citation proximity analysis from idea to practice in research article recommendation. (08 2017), 14-25 pages.

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

[16] 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

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

[18] 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

[19] 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

[20] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

[21] 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

[22] 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.

[23] Henry Small. 1973. Co-Citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents. 24 (07 1973), 265 – 269.

[24] Chen-Tse Tsai, Gourab Kundu, and Dan Roth. 2013. Concept-based Analysis of Scientific Literature. In *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management (CIKM '13)*. ACM, New York, NY, USA, 1733–1738. https://doi.org/10.1145/2505515.2505613