This document is not a research article, but a proposal. I publish it to initiate discussions, as example for my students, and to encourage others to conduct the research as I do not always have the resources to pursue the proposed work myself. Please note that while I try to add as many relevant references as possible to the proposal, it is likely that I missed some. Feel free to contact me if you believe that someone else published the idea already, or an important reference is missing. If you are interested in a joint project relating to this proposal, please also contact me.

Virtual Citation Proximity (VCP): Calculating Co-Citation-Proximity-Based Document Relatedness for Uncited Documents with Machine Learning [Proposal]

Joeran Beel, Department of Computer Science and Statistics, Intelligent Systems, KDEG Group, ADAPT Centre, joeran.beel@adaptcentre.ie

Abstract. The relatedness of research articles, patents, legal documents, web pages, and other documents is often calculated with citation or hyperlink based approaches such as citation proximity analysis (CPA). In contrast to text-based document similarity, citation-based relatedness covers a broader range of relatedness. However, citation-based approaches suffer from the many documents that receive little or no citations, and for which document relatedness hence cannot be calculated. I propose to calculate a machine-learned 'virtual citation proximity' (or 'virtual hyperlink proximity') that could be calculated for all documents for which textual information (title, abstract …) and metadata (authors, journal name …) is available. The input to the machine learning algorithm would be a large corpus of documents, for which textual information, metadata *and* citation proximity is available. The citation proximity would serve as ground truth, and the machine-learning algorithm would infer, which textual features correspond to a high proximity of co-citations. After the training phase, the machine-learning algorithm could calculate a virtual citation proximity even for uncited documents. This virtual citation proximity would express in what proximity two documents would likely be cited, if they were cited. The virtual citation proximity then could be used in the same way as "real" citation proximity to calculate document relatedness, and would potentially cover a wider range of relatedness than text-based document relatedness.

Keywords: document relatedness, citation analysis, citation proximity analysis, digital libraries, recommender systems, search engines

1 INTRODUCTION

Retrieving a list of 'related documents' – e.g. web pages, patents, or research articles – for a given source document is a common feature of many applications, including recommender systems and search engines (Figure 1). Document relatedness is typically calculated based on documents' text (title, abstract, full-text) and metadata (authors, journal …), or based on citations/hyperlinks [1–10][1]. The intuition behind text-based relatedness measures is that two documents are more highly related the more terms they share (in place of terms, concepts, topics, n-grams, embeddings, etc. may also be used [11–17]). The intuition of citation-based relatedness is that authors cite documents because they consider them to be related to the manuscript they are currently writing. Consequently, two documents that are (co-) cited are both related to the citing document and to each other [18–24].

While text-based relatedness is about the *similarity* of documents, i.e. the proportion of terms they have in common, citation-based relatedness is broader. Two documents can be co-cited, and hence related, for many reasons [25–29]. For example, the two documents may use the same algorithm (to solve the same or different problems); the two documents may be written by the same author; or the two documents may be co-cited for less predictable reasons, for example if both are examples of well-written academic articles and the citing author is writing a book on academic writing. Today's text-based methods can hardly calculate such types of semantic relatedness.

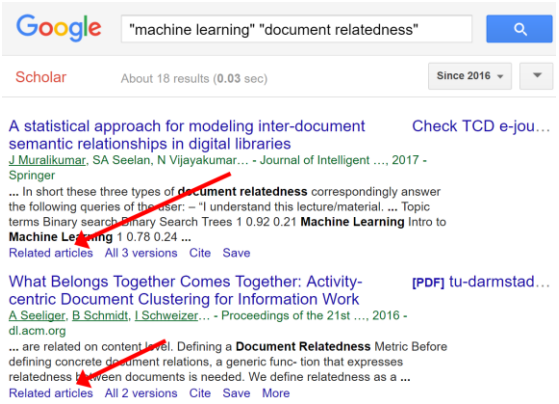

Figure 1: Google Scholar's "Related article" feature

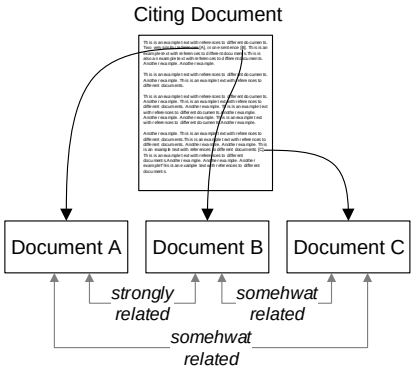

Figure 2: Citation-Proximity Analysis

One particularly effective citation-based approach is citation proximity analysis (CPA) (Figure 2) [30–36]. Its intuition is that

[1] In the current proposal, I focus on the relatedness of research articles based on citations. However, the work proposed here could easily be extended or shifted to calculating relatedness of patents, legal documents, and web pages based on citations or hyperlinks.

the closer the proximity in which two documents are cited in a text, the more highly related the two documents are. For instance, if two documents A and B are cited within the same sentence, they are more highly related than documents A and C that are cited in different paragraphs (Figure 2). Citation-proximity analysis has successfully been used to calculate relatedness of research articles [30–36], as well as the relatedness of Wikipedia articles [37], [38], web pages [39], and authors [40].

2 RESEARCH PROBLEM

Although citation-proximity may be very effective, it suffers from the same problem as other citation-based approaches: To calculate relatedness, documents need to be (frequently) co-cited. However, most research articles are never cited, and it usually takes a year or more before a document receives its first citation [41], [42]. Consequently, document relatedness based on citation-proximity can only be calculated for few documents in a corpus and tends to work best on older documents.

3 RESEARCH GOAL

The research goal is to develop a citation-proximity-alike relatedness method that works for less cited and uncited documents. The new method should reflect the variety of relatedness that may be expressed through co-citation proximity. This means the relatedness should not simply be based on the proportion of terms that two documents share and hence only express the degree of similarity.

To achieve the goal, I propose to calculate a 'virtual citation proximity' that is machine learned. The input to the machine learning algorithm would be a large document corpus including the documents' text (title, abstract, full-text, references/citations) and metadata (authors, journal …), and their citation proximities. Based on the co-citation proximity as ground truth and the textual features and metadata, the machine-learning algorithm will infer what features make a co-cited document pair related. Subsequently, the algorithm calculates a virtual citation proximity based only on documents' textual features and metadata. This virtual citation proximity would express in what proximity two documents would likely be cited, if they were cited. Once the machine learning algorithm is trained on a sufficiently large corpus, it could calculate the virtual proximity for any document pair for which textual information and metadata is available. Once the virtual citation proximity is calculated, it could be used in the same way as the normal citation-proximity to calculate document relatedness (e.g. calculate the citation-proximity index, CPI [31]).

Although virtual citation proximity is based on textual features and metadata, I hypothesise that it will produce similar results as real citation-proximity, since the machine learning model is based on real citation proximity as ground truth. With the recent advances in machine learning, particularly deep learning, I hypothesise that a (deep) machine-learning algorithm will be able to detect hidden layers in the text. These will allow determining what makes two documents related, more reliable than the typical assumption in text-based approaches that two documents are related when they share the same terms.

Virtual citation proximity will combine the best of both worlds, i.e. it can be calculated for every document (like today's text-based methods), yet provide a high variety of relatedness and be highly effectiveness (like today's citation-based approaches). Hence, virtual citation proximity has the potential to advance significantly related-document calculations for search engines and recommender systems. Related-document calculations will not only improve for academic documents but potentially for legal documents, patents, medical documents and web pages, too. In latter case, the method should be called 'Virtual Hyperlink Proximity' or 'Virtual Link Proximity'.

4 RELATED WORK

The method that is closest to using citation-proximity as ground truth is using expert judgements like the biomedical classification MeSH [43–45], the ACM classification [46], DMOZ [47], or ACL [48]. The MeSH classification represents the major fields in the biomedical domain and was created by medical experts (Figure 3). New biomedical publications are often classified with MeSH, i.e. they are assigned to one of the MeSH categories, and two documents in the same category are considered to be related. Machine learning algorithms can infer from the existing documents in a category, which textual features make a document likely to belong to a certain category. New documents can then automatically be classified.

Figure 3: Excerpt of MeSH

There are several disadvantages to using expert classifications like MeSH. First, they are one-dimensional, i.e. they provide only one type of relatedness (typically, the overall topic a research article is about). However, there may be many other dimensions of relatedness (e.g. a shared algorithm or shared methodology applied in different domains). Second, most classification schemes allow documents to be in one or two categories only. Especially with today's increasingly interdisciplinary work, this is often not enough to adequately find all related documents. Third, classification schemes typically have a limited number of categories (a few dozen or hundreds). This means, every category contains thousands of documents that might be somewhat related

but only at a broad level. Fourth, the classifications are static, i.e. articles are classified at the time of publication. If a classification scheme is changed, the papers usually cannot be updated. Finally, for many domains, expert classifications do not exist. Hence document relatedness in these domains is difficult to learn.

With citation proximity as ground truth, the mentioned problems could be overcome. (Virtual) citation proximity (1) covers many types of relatedness; (2) allow documents to be in unlimited numbers of co-citation clusters; (3) has no limitations for the number of clusters; (4) is dynamic; and (5) can be learned for any domain that uses citations.

5 PROPOSED METHDOLOGY

To achieve the research goal, a number of tasks need to be undertaken.

- A thorough literature review on citation-proximity analysis, and other methods to calculate document relatedness, both text-based and citation-based. This includes a review of currently used ground-truth such as MeSH but also related fields such as social-tagging [49], folksonomies [50], and query log analysis [51], which are sometimes used to train machine learning algorithms.
- Identification of promising machine-learning algorithms that could be capable of learning and calculating virtual citation proximity. This includes both 'normal' machine learning algorithms as well as deep learning algorithms.
- The compilation of a large dataset containing co-citation proximity information and text and metadata of the co-cited documents. The dataset ideally contains tens of millions of articles, and hundreds of millions of citations from different disciplines (e.g. biomedical, computer, and social sciences). Depending on the eventual project scope, the corpus could focus on research articles, legal documents (laws, rulings, …), patents, web pages, or massive content repositories such as Wikipedia.
- Evaluation and fine-tuning the different machine learning algorithms. The evaluation should be performed in two ways. First, using an offline dataset. This means, a part of the previously created dataset will not be used for training the machine learning algorithms, but to evaluate the effectiveness. Second, the algorithms could be evaluated in a live search engine or recommender system (for instance in Mr. DLib [52]).

6 OUTLOOK

It could further be interesting to use machine learning to understand *why* documents are cited in close proximity, and incorporate this knowledge into the relatedness-calculation process. Considering citation context [53–55], i.e. the sentences surrounding a citation, might further improve the calculation of virtual citation proximity. It might also be interesting to take bias and motivation to create citations into account [27], [56], [57], e.g. citations that were done for illegitimate reasons should be removed from the training corpus. Finally, the concept of virtual-citation proximity might also be used to improve related applications such as co-authorship analysis [58] and identifying related authors.

REFERENCES

[1] K. Balog, N. Takhirov, H. Ramampiaro, and K. Nørvåg, "Multi-step Classification Approaches to Cumulative Citation Recommendation," in *Proceedings of the OAIR'13*, 2013.

[2] T. Chakraborty, N. Modani, R. Narayanam, and S. Nagar, "Discern: a diversified citation recommendation system for scientific queries," in *2015 IEEE 31st International Conference on Data Engineering*, 2015, pp. 555–566.

[3] C. Caragea, A. Silvescu, P. Mitra, and C. L. Giles, "Can't See the Forest for the Trees? A Citation Recommendation System," in *iConference 2013 Proceedings*, 2013, pp. 849–851.

[4] D. Duma, M. Liakata, A. Clare, J. Ravenscroft, and E. Klein, "Rhetorical Classification of Anchor Text for Citation Recommendation," *D-Lib Magazine*, vol. 22, no. 9/10, 2016.

[5] B. Gipp and N. Meuschke, "Citation Pattern Matching Algorithms for Citation-based Plagiarism Detection: Greedy Citation Tiling, Citation Chunking and Longest Common Citation Sequence," in *Proceedings of the 11th ACM Symposium on Document Engineering*, 2011, pp. 249–258.

[6] A. Livne, V. Gokuladas, J. Teevan, S. T. Dumais, and E. Adar, "CiteSight: supporting contextual citation recommendation using differential search," *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, pp. 807–816, 2014.

[7] N. Meuschke, "Citation-based Plagiarism Detection for Scientific Documents," Dep. of Computer Science, Otto-von-Guericke-University Magdeburg, Germany, 2011.

[8] X. Ren, "Effective citation recommendation by information network-based clustering," 2016.

[9] K. Sugiyama and M.-Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation," in *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, 2013, pp. 153–162.

[10] F. Zarrinkalam and M. Kahani, "Using Semantic Relations to Improve Quality of a Citation Recommendation System," *Soft Computing Journal*, vol. 1, no. 2, pp. 36–45, 2013.

[11] M. S. Shaikh, "Applications of Machine learning to document classification and clustering," in *2017 International Conference on Innovations in Electrical Engineering and Computational Technologies (ICIEECT)*, 2017, pp. 1–1.

[12] G. Nikolentzos, P. Meladianos, F. Rousseau, Y. Stavrakas, and M. Vazirgiannis, "Shortest-Path Graph Kernels for Document Similarity," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 1891–1901.

[13] J. Muralikumar, S. A. Seelan, N. Vijayakumar, and V. Balasubramanian, "A statistical approach for modeling inter-document semantic relationships in digital libraries," *Journal of Intelligent Information Systems*, vol. 48, no. 3, pp. 477–498, Jun. 2017.

[14] S. Kim, W. J. Wilbur, and Z. Lu, "Bridging the gap: a semantic similarity measure between queries and documents," *arXiv preprint arXiv:1608.01972*, 2016.

[15] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[16] J. Mei, X. Kou, Z. Yao, A. Rau-Chaplin, A. Islam, A. Moh'd, and E. E. Milios, "Efficient Computation of Co-occurrence Based Word Relatedness," in *Proceedings of the 2015 ACM Symposium on Document Engineering*, 2015, pp. 43–46.

[17] J. Bian, B. Gao, and T.-Y. Liu, "Knowledge-powered deep learning for word embedding," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2014, pp. 132–148.

[18] P. Ahlgren and B. Jarneving, "Bibliographic Coupling, Common Abstract Stems and Clustering: A Comparison of Two Document-document Similarity Approaches in the Context of Science Mapping," *Scientometrics*, vol. 76, no. 2, pp. 273–290, 2008.

[19] K. W. Boyack and R. Klavans, "Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 12, pp. 2389–2404, 2010.

[20] E. Garfield, "From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography," in *speech deliveredat Drexel University, Philadelphia, PA, November*, 2001, vol. 27.

[21] O. Küçüktunç, K. Kaya, E. Saule, and U. V. Catalyürek, "Fast Recommendation on Bibliographic Networks with Sparse-Matrix Ordering and Partitioning," *Social Network Analysis and Mining*, vol. 3, no. 4, pp. 1097–1111, 2013.

[22] F. Osareh, "Bibliometrics, Citation Anatysis and Co-Citation Analysis: A Review of Literature II," *Libri*, vol. 46, no. 4, pp. 227–225, 1996.

[23] H. Small, "Co-citation in the scientific literature: A new measure of the relationship between two documents," *Journal of the American Society for information Science*, vol. 24, no. 4, pp. 265–269, 1973.

[24] H. Small, "Co-citation in the Scientific Literature: A New Measure of the Relationship Between Two Documents," *Journal of the American Society for Information Science*, vol. 24, pp. 265–269, 1973.

[25] L. Bornmann and H. D. Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[26] N. Kaplan, "The Norms of Citation Behavior: Prolegomena to the Footnote," *American Documentation*, vol. 16, no. 3, pp. 179–184, Jul. 1965.

[27] C. Thornley, A. Watkinson, D. Nicholas, R. Volentine, H. R. Jamali, E. Herman, S. Allard, K. J. Levine, and C. Tenopir, "The role of trust and authority in the citation behaviour of researchers.," *Information Research*, vol. 20, no. 3, pp. 1–17, 2015.

[28] T. A. Shah, S. Gul, and R. C. Gaur, "Authors self-citation behaviour in the field of Library and Information Science," *Aslib Journal of Information Management*, vol. 67, no. 4, pp. 458–468, 2015.

[29] P. Willett, "Readers' perceptions of authors' citation behaviour," *Journal of Documentation*, vol. 69, no. 1, pp. 145–156, 2013.

[30] A. Balaji, S. Sendhilkumar, and G. S. Mahalakshmi, "Finding Related Research Papers Using Semantic and Co-Citation Proximity Analysis," *Journal of Computational and Theoretical Nanoscience*, vol. 14, no. 6, pp. 2905–2909, 2017.

[31] B. Gipp and J. Beel, "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 (ISSI'09)*, 2009, vol. 2, pp. 571–575.

[32] P. Knoth, L. Anastasiou, A. Charalampous, M. Cancellieri, S. Pearce, N. Pontika, and V. Bayer, "Towards effective research recommender systems for repositories," in *Proceedings of the Open Repositories Conference*, 2017.

[33] S. Liu and C. Chen, "The Effects of Co-citation Proximity on Co-citation Analysis," in *Proceedings of the Conference of the International Society for Scientometrics and Informetrics*, 2011.

[34] S. Liu and C. Chen, "The Proximity of Co-citation," *Scientometrics*, vol. 91, no. 2, pp. 495–511, Dec. 2011.

[35] G. Colavizza, K. W. Boyack, N. J. van Eck, and L. Waltman, "The Closer the Better: Similarity of Publication Pairs at Different Co-Citation Levels," *arXiv preprint arXiv:1707.03076*, 2017.

[36] P. Knoth and A. Khadka, "Can we do better than Co-Citations?- Bringing Citation Proximity Analysis from idea to practice in research article recommendation," in *Proc. of the 2nd Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL), Tokyo, Japan, CEUR-WS. org*, 2017.

[37] M. Schwarzer, M. Schubotz, N. Meuschke, C. Breitinger, V. Markl, and B. Gipp, "Evaluating Link-based Recommendations for Wikipedia," in *Proceedings of the 16th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2016, pp. 191–200.

[38] M. Schwarzer, C. Breitinger, M. Schubotz, N. Meuschke, and B. Gipp, "Citolytics: A Link-based Recommender System for Wikipedia," in *Proceedings of the Eleventh ACM Conference on Recommender Systems*, 2017, pp. 360–361.

[39] B. Gipp, A. Taylor, and J. Beel, "Link Proximity Analysis - Clustering Websites by Examining Link Proximity," in *Proceedings of the 14th European Conference on Digital Libraries (ECDL'10): Research and Advanced Technology for Digital Libraries*, 2010, vol. 6273, pp. 449–452.

[40] H. J. Kim, Y. K. Jeong, and M. Song, "Content- and proximity-based author co-citation analysis using citation sentences," *Journal of Informetrics*, vol. 10, no. 4, pp. 954–966, 2016.

[41] M. Golosovsky, "Power-law citation distributions are not scale-free," *Phys. Rev. E*, vol. 96, no. 3, p. 032306, Sep. 2017.

[42] G. Abramo, C. A. D'Angelo, and A. Soldatenkova, "The dispersion of the citation distribution of top scientists' publications," *Scientometrics*, vol. 109, no. 3, pp. 1711–1724, Dec. 2016.

[43] M. Dıaz-Galiano, M. Garcıa-Cumbreras, M. Martın-Valdivia, A. Montejo-Ráez, and L. Urena-López, "Integrating mesh ontology to improve medical information retrieval," in *CLEF*, 2007, vol. 5152, pp. 601–606.

[44] S. Bloehdorn, P. Cimiano, and A. Hotho, "Learning Ontologies to Improve Text Clustering and Classification," in *From Data and Information Analysis to Knowledge Engineering: Proceedings of the 29th Annual Conference of the Gesellschaft für Klassifikation e.V. University of Magdeburg, March 9–11, 2005*, M. Spiliopoulou, R. Kruse, C. Borgelt, A. Nürnberger, and W. Gaul, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 334–341.

[45] F. Camous, S. Blott, and A. Smeaton, "Ontology-based MEDLINE document classification," *Bioinformatics Research and Development*, pp. 439–452, 2007.

[46] J. Broisin, M. Brut, V. Butoianu, F. Sedes, and P. Vidal, "A personalized recommendation framework based on cam and document annotations," *Procedia Computer Science*, vol. 1, no. 2, pp. 2839–2848, 2010.

[47] S. E. Middleton, D. C. De Roure, and N. R. Shadbolt, "Capturing knowledge of user preferences: ontologies in recommender systems," in *Proceedings of the 1st international conference on Knowledge capture*, 2001, pp. 100–107.

[48] M. Valenzuela, V. Ha, and O. Etzioni, "Identifying Meaningful Citations.," in *AAAI Workshop: Scholarly Big Data*, 2015.

[49] T. Bogers and V. Petras, "Tagging vs. Controlled Vocabulary: Which is More Helpful for Book Search?," *iConference Proceedings*, 2015.

[50] D. Benz, A. Hotho, R. Jäschke, B. Krause, and G. Stumme, "Query logs as folksonomies," *Datenbank-Spektrum*, vol. 10, no. 1, pp. 15–24, 2010.

[51] R. Baeza-Yates and A. Tiberi, "Extracting semantic relations from query logs," in *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2007, pp. 76–85.

[52] J. Beel, A. Aizawa, C. Breitinger, and B. Gipp, "Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia," in *Proceedings of the ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL)*, 2017, pp. 1–2.

[53] M. Eto, "Evaluations of context-based co-citation searching," *Scientometrics*, vol. 94, no. 2, pp. 651–673, 2013.

[54] S. R. Lawrence, K. D. Bollacker, and C. L. Giles, "Autonomous citation indexing and literature browsing using citation context," U.S. Patent US 6,738,780 B2Summer-2004.

[55] H. Small, "Citation Context Analysis," *Progress in Communication Sciences*, vol. 3, pp. 287–310, 1982.

[56] J. P. Ioannidis, "Statistical Biases in Science Communication: What We Know About Them and How They Can Be Addressed," *The Oxford Handbook of the Science of Science Communication*, p. 103, 2017.

[57] A. Zuccala, N. Robinson-Garcia, R. Repiso, and D. Torres-Salinas, "Using network centrality measures to improve national journal classification lists," in *arXiv preprint arXiv:1606.00240*, 2016.

[58] F. Momeni and P. Mayr, "Evaluating Co-authorship Networks in Author Name Disambiguation for Common Names," in *Research and Advanced Technology for Digital Libraries: 20th International Conference on Theory and Practice of Digital Libraries, TPDL 2016, Hannover, Germany, September 5–9, 2016, Proceedings*, N. Fuhr, L. Kovács, T. Risse, and W. Nejdl, Eds. Cham: Springer International Publishing, 2016, pp. 386–391.

DOCUMENT HISTORY

- 2017-10-27: First version published on arxiv