Identifying related Documents for Research Paper Recommender by CPA and COA

Bela Gipp and Jöran Beel

***Abstract*—This work-in-progress paper introduces two new approaches called Citation Proximity Analysis (CPA) and Citation Order Analysis (COA). They can be applied to identify related documents for the purpose of research paper recommender systems. CPA is a variant of co-citation analysis that additionally considers the proximity of citations to each other within an article's full-text. The underlying idea is that the closer citations are to each other in a document, the more likely it is that the cited documents are related. For example, citations listed in the same sentence are more likely to express related thoughts than citations listed only in the same section. In COA, the order of citations are considered, allowing the identification of a text similar to one that has been translated from language A to language B, as the citations would still occur in the same order. However, it is also shown that CPA and COA cannot replace text analysis and existing citation analysis approaches for research paper recommender systems since they all have their own strengths and weaknesses.**

***Index Terms*—Bibliometrics, citation proximity analysis, citation order analysis, related documents, research paper recommender**

I. INTRODUCTION

The search for related work is a time-consuming procedure that even if performed by experienced scientists often leads to unsatisfying results. To alleviate the problem, search engines such as Google Scholar and Citeseer offer to display "similar" documents based on text and citation analysis.

Superior results are usually achieved by hybrid research paper recommender systems. By combining further techniques such as co-word analysis, collaborative filtering, Subject-Action-Object (SAO) structures, etc., more precise recommendations can be given. However, these approaches are only suitable to a limited extent for identifying related work [2-8].

Taking everything into account, our examination suggests that in the case of scientific documents, usually the best results can be achieved by applying co-citation analysis. Citation proximity analysis is a further development of co-citation analysis.

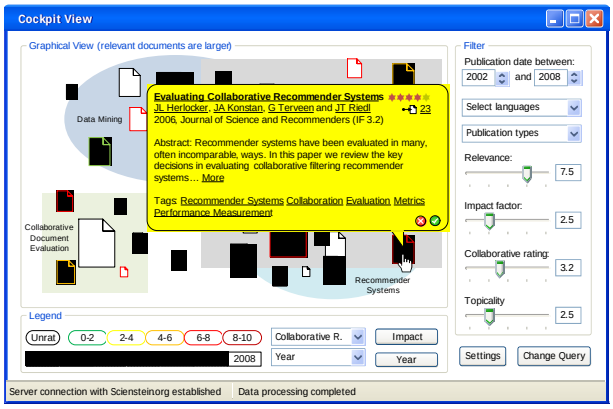

Figure 1: GUI SciPlore – clustering similar documents

In the research paper recommender SciPlore.org this approach is mainly used for two purposes. First, to cluster similar documents as shown in Figure 1; and secondly, to give recommendations for further related documents based on one or more documents the user has been interested in, as shown in Figure 2.

In the first part of this paper related work is presented and the commonly applied citation analysis approaches discussed with the focus on co-citation analysis. In the following section the CPA approach is introduced. Afterwards, the existing citation analysis approaches are compared to CPA and their suitability for research paper systems examined. The paper concludes with a summary and an outlook which includes how this new approach is going to be integrated in the research paper recommender SciPlore.org.

Bela Gipp and Jöran Beel are with the Otto-von-Guericke University Magdeburg, Department of Computer Science, ITI and SciPlore.org (gipp|beel@sciplore.org).

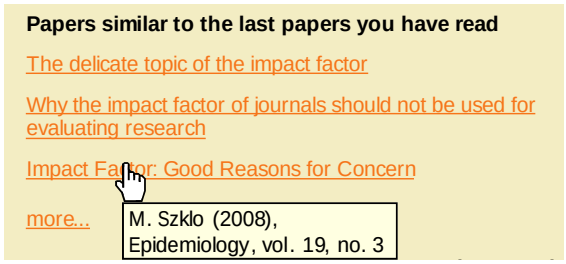

Figure 2: Similar paper recommendation

II. RELATED WORK

The usefulness of a research paper recommender system depends to a large extent on its ability to automatically determine related work to one or more documents. Various approaches exist to determine the degree of similarity of documents in order to identify related work.

Whereas text-mining approaches are used in cases in which references are not stated, citation analysis approaches usually deliver superior results as e.g. synonyms and unclear nomenclature do not lead to misleading results [3, 4, 5]. Many citation analysis approaches exist and they all have their own strengths and weaknesses for identifying similar documents. Among the most widely used are the easily applicable 'cited by' approach, which considers papers as relevant that cite the same input document and the 'reference list' approach, which considers papers as relevant that were referenced by the input document. The best results can usually be obtained by bibliographic coupling and co-citation analysis, which allow calculating the coupling strength [6]. These approaches, which were already invented in the 60s and 70s, are used by scientists and on academic search engine websites like CiteSeer1 [9].

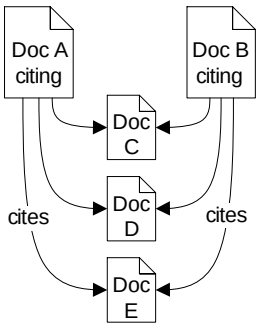

Figure 3: Bibliographic coupling

Documents are bibliographically coupled if they cite one or more documents in common. Figure 3 illustrates this approach: Papers A and B are related because they both cite papers C, D and E.

[1] http://citeseer.ist.psu.edu

In contrast, two documents are "co-cited" when at least one paper cites both. This approach is illustrated in Figure 4: Papers A and B are related because they are both cited by papers C, D and E. The more co-citations two papers receive, the more related they are [6].

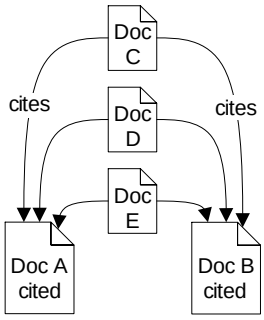

Figure 4: Co-citation analysis

Although both approaches are suitable to identify similar papers, they serve different purposes. Whereas bibliographic coupling is retrospective, co-citation is essentially a forward-looking perspective [9]. However, both approaches often deliver unsatisfying results, since they only make use of the bibliography at the end of the document without analyzing the constellation of citations. Therefore it is not possible to determine in which part of a related document the content of interest can be found.

III. CITATION PROXIMITY ANALYSIS AND CITATION ORDER ANALYSIS

Instead of just using the bibliography, in CPA the information derived from the proximity of the citations to each other in the full-text is used to calculate the Citation Proximity Index (CPI) in three steps.

1. The document is parsed and a series of heuristics are used to process the citations including their position within the document2.

2. The citations are assigned to their corresponding items in the bibliography. The overall margin of error with the system we have developed equals nearly three percent for the first and second step.

3. In the third step the proximity among each citation-pair is examined. The underlying assumption is that the closer the citations are to each other, the more likely it is that they are related. Based on this proximity analysis, the CPI is calculated. If for example two citations are given in the

[2] The citations were parsed using a modified version of parsCit (http://wing.comp.nus.edu.sg/parsCit) in combination with exclusively developed software, which is available upon request from the authors.

same sentence the probability that they are very similar is higher (CPI = 1) as if they were only in the same paragraph (CPI = ¼). See Figure 5.

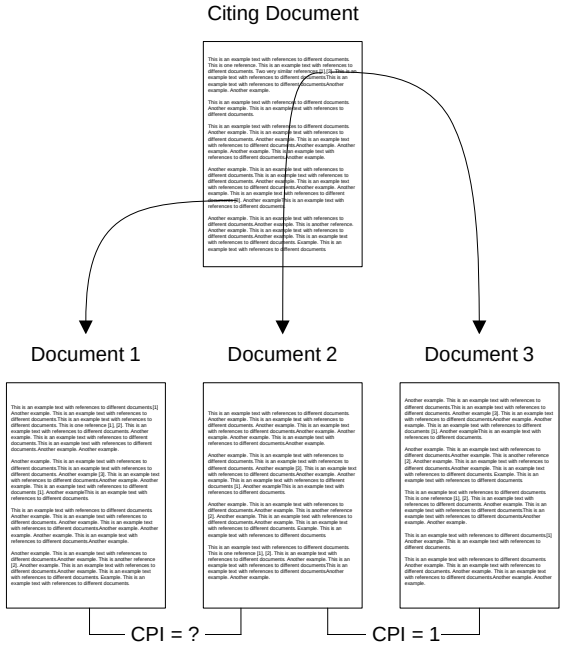

Figure 5: Illustration CPA

However, further research needs to be performed to identify the appropriate weighting of the CPI values according to their occurrence, which also seems to depend on the publication's research field and publication's research type. For example, it seems that for analyzing a technical report or patent specification, different weightings seem suitable. First empirical evaluations have lead to the values shown in Table 1 for calculating the CPI.

Table 1: CPI values

Occurrence	CPI value
Sentence	1
Paragraph	1/2
Chapter	1/4
Same journal / same book	1/8
Same journal but different edition	1/16

The results delivered by CPA can be improved by evaluating as many sources as possible. This can be the case due to multiple occurrences of the same citation and due to multiple documents citing a certain document. In our series of tests we experienced the best results by calculating the weighted average of the CPIs. By automating the process described above, we have calculated the CPI for publications contained in the SciPlore database. The results show that in comparison to the results delivered by co-citation analysis, CPA delivers considerably better results in identifying similar documents [1].

Similar to the idea of CPA is another approach currently under development, that we call Citation Order Analysis (COA). In contrast to CPA, in COA, only the order of citations is considered. The main advantage in comparison to the usually applied text analysis approaches is that even if documents are translated or paraphrased they can still be identified as similar. Depending on the level of tolerance even if citations were omitted, summarized documents can be identified. This way a digital fingerprint of documents can be created that can, besides for recommender systems, also be used to identify plagiarized work. In some regard, this approach is similar to bibliographic coupling. However, by additionally considering the order of citations, this approach is more precise and robust. Figure 6 illustrates the concept.

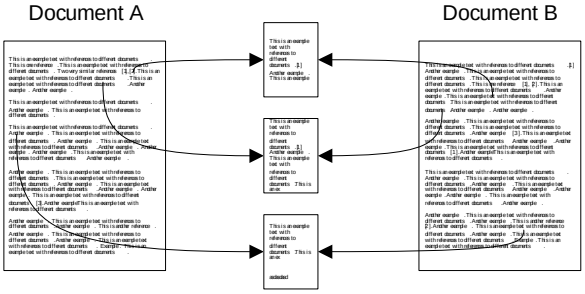

Figure 6: Illustration Citation Order Analysis

IV. OUTLOOK

Besides identifying related work, the authors work on applying the idea behind CPA for automatic document classification for the research paper recommender SciPlore [11]. The aim is to automatically analyze the topics within documents by analyzing the distribution of references within research papers. So instead of knowing, for instance, that a certain publication focuses on the relativity theory, the CPA makes it possible to identify the document sections focusing for example, on '*Time dilation*', '*Length contraction*' or '*Mass-energy equivalence*' and then to give specific recommendations within documents or books.

Moreover, it is possible to combine the CPA with text mining algorithms in order to automatically detect e.g. contradicting studies. "*The author A has shown in his recent study [reference A] that <u>in contrast</u> to a previous study [reference B]...*" So by analyzing the words between

two references it is often possible to automatically analyze the exact relationship between these two references and how they compare to each other.

Oftentimes it is possible by knowing the position of each citation within a document, to draw conclusions about the document type e.g. state-of-the art publications, etc. The gathered information can be used to classify further documents and to develop a more sophisticated '*Web of Science*'. We believe that these technologies, in combination with collaborative filtering, will be the future for identifying related work and will open the doors for powerful research paper recommender systems.

V. DISCUSSION & CONCLUSION

As shown, the CPA and COA offer substantial advantages in identifying related documents in comparison to existing approaches. However, it should also be taken into account that the effort is considerable. It is not sufficient to evaluate the bibliography of documents, but it is necessary to process the complete document, identify each reference and map it to the corresponding entry in the bibliography, which is in practice not always possible, and leads in ca. 3% of cases to mismatches. This is because sometimes only an abstract and the bibliography can be accessed, documents cannot be parsed as OCR fails, or a reference style is used that makes it unfeasible to automatically link references to the corresponding items in the bibliography. This leads to the conclusion that although these new approaches deliver superior results, they cannot completely replace the already existing approaches, but should be used in combination.

REFERENCES

[1] Gipp, B. & Beel, J. (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, pp. 571-575.

[2] Rip, A., & Courtial, J. (1984). Co-Word Maps of Biotechnology: An Example of Cognitive Scientometrics. Scientometrics, 6(6), 381-400.

[3] Fano, R. M. 1956. Information theory and the retrieval of recorded information, in Documentation in Action, Shera, J. H. Kent, A. Perry, J. W. (Edts), New York: Reinhold Publ. Co., pp. 238–244.

[4] Marshakova, I. V. 1973. System of document connections based on references, Nauchno-Tekhnicheskaya Informatsiya, vol. 2, no. 6, pp. 3–8.

[5] Beel, J. & Gipp, B. 2008, The Potential of Collaborative Document Evaluation for Science, the 11th International Conference on Digital Asian Libraries (ICADL 2008), December 2 - 5, Kuta, Indonesia, published in G. Buchanan, M. Masoodian & S. Cunningham (Eds.), Digital Libraries: Universal and Ubiquitous Access to Information of Lecture Notes in Computer Science, vol. 5362, DOI 10.1007/978-3-540-89533-6, ISSN 0302-9743, pp. 375-378, Springer-Verlag Berlin Heidelberg.

[6] Small, H. 1973. 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.

[7] Klavans, R., & Boyack, K. (2006). Identifying a better measure of relatedness for mapping science, Journal of the American Society for Information Science and Technology, Vol. 57, No. 2, pp. 251-263.

[8] Sternitzke, C. Bergmann, I. (2009), Similarity measures for document mapping: A comparative study on the level of an individual scientist, Scientometrics, Vol. 78, No. 1, pp. 113-130.

[9] Garfield, E. (2001, November 27, 2001). From Bibliographic Coupling to Co-CitationAnalysis Via Algorithmic Historio-Bibliography: A Citationist's Tribute to BelverC. Griffith. Paper presented at the Drexel University, Philadelphia, PA.

[10] Giles, C. L. Bollacker, K. D. And Lawrence, S. 1998. CiteSeer: an automatic citation indexing system, In Digital Libraries 98 - The Third ACM Conference on Digital Libraries, pp. 89-98.

[11] Gipp, B. Beel, J. & Hentschel, C. (2009), Scienstein - A Research Paper Recommender System, in Proceedings of IEEE International Conference on Emerging Trends in Computing. Tamil Nadu, India.