Bela Gipp and Jöran Beel. Citation Proximity Analysis (CPA) - A new approach for identifying related work based on Co-Citation Analysis. In Birger Larsen and Jacqueline Leta, editors, *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 2, pages 571–575, Rio de Janeiro (Brazil), July 2009. International Society for Scientometrics and Informetrics. ISSN 2175-1935. Downloaded from www.sciplore.org.

# Citation Proximity Analysis (CPA) – A new approach for identifying related work based on Co-Citation Analysis

Bela Gipp[1] and Jöran Beel[2]

[1] *Bela@Gipp.com*, [2] *Joeran@Beel.org*
Otto-von-Guericke University, Dept. of Computer Science, Magdeburg, Germany

## Abstract
This paper presents an approach for identifying similar documents that can be used to assist scientists in finding related work. The approach called *Citation Proximity Analysis* (CPA) is a further development of co-citation analysis, but in addition, 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, the more likely it is that they are related. In comparison to existing approaches, such as bibliographic coupling, co-citation analysis or keyword based approaches the advantages of CPA are a higher precision and the possibility to identify related sections within documents. Moreover, CPA allows a more precise automatic document classification. CPA is used as the primary approach to analyse the similarity and to classify the 1.2 million publications contained in the research paper recommender system Scienstein.org.

## Introduction and Motivation
The search for related scientific work can be tedious, and often important documents are missed out. Difficulties are caused by an increasing number of publications, growing exponentially at a yearly rate of 3.7 %, unclear nomenclature, synonyms and numerous other factors [1]. In practice, most searches for related work start with some initial papers and navigating the citation web nearest to those papers. However, even the more advanced approaches for identifying related work based on co-word analysis, collaborative filtering, Subject-Action-Object (SAO) structures or citation analysis do often not deliver satisfying results [2-8]. Therefore, we developed a new approach to determine the similarity of documents, which we name *Citation Proximity Analysis* (CPA). The approach is based on co-citation analysis and improves precision by considering the position of citations. The presented approach was developed for the research paper recommender Scienstein[1] to assist researchers in finding related work.

The first part of this paper gives an overview about existing methods to identify similar documents, whereas the focus lies on the most popular citation analysis approaches and their strengths and weaknesses. The second part explains how the CPA can be used to measure similarity and the steps necessary to calculate a new metric that we call Citation Proximity Index (CPI). Afterwards, first results from an empirical study comparing the performance of co-citation analysis and CPA are presented. Finally, an outlook on further implications and how the CPA could be used in other fields is given.

[1] www.scienstein.org is a research paper recommender focusing on identifying related work developed by the authors

## Related Work

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 CiteSeer[2] [9].

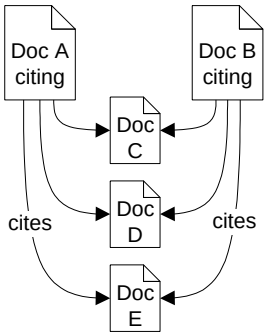


Figure 1: Bibliographic coupling

Documents are bibliographically coupled if they cite one or more documents in common. Figure 1 illustrates this approach: Papers A and B are related because they both cite papers C, D and E.

In contrast, two documents are "co-cited" when at least one paper cites both. This approach is illustrated in Figure 2: 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].

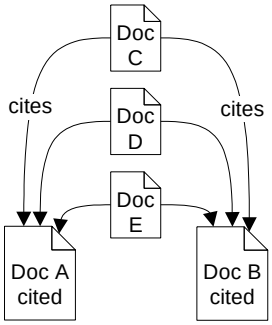


Figure 2: Co-citation analysis

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 analysing the constellation of citations. Since these approaches are system-inherent, it is also not possible to determine in which part of a related document the content of interest can be found.

## Citation Proximity Analysis (CPA)

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 document[3].

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

[3] 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.

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 same sentence the probability that they are very similar is higher (CPI = 1) as if they were only in the same paragraph (CPI = 1/2). See Figure 3.

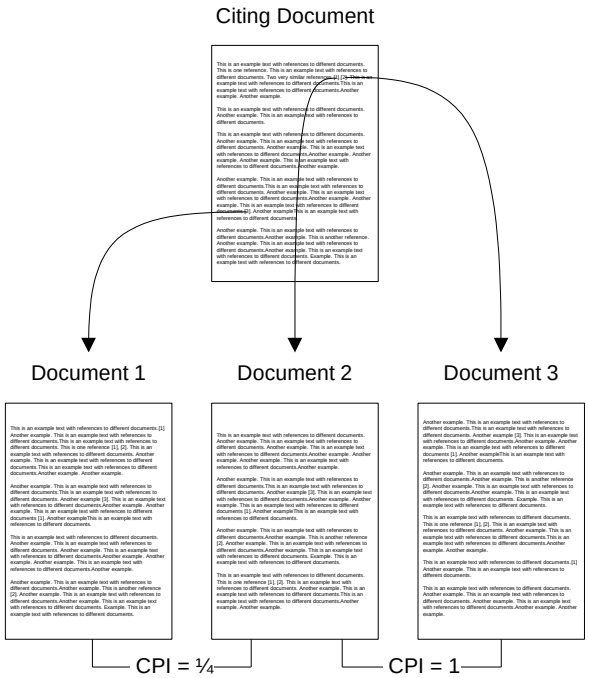


Figure 3: Illustration of Citation Proximity Analysis

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 analysing 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 Scienstein database. The results show that in comparison to the results delivered by co-citation analysis, CPA delivers considerably better results in identifying similar documents.

## Empirical Comparison of Co-Citation Analysis and CPA

In the following, first results of a study examining the suitability of CPA to identify related work are presented. The complete study will be published separately. As it would be unfeasible to compare the results with every known approach, the focus laid on a comparison with Co-citation analysis as this approach usually delivers the best results. The 21 study participants have been asked to select three similar documents from the Scienstein.org database and then six "related work recommendations" have been provided. Three of them were chosen based on co-citation strength and three based on CPA without indicating the used approach. The results show that the CPA performs significantly better in identifying related work than the commonly-used Co-citation analysis.

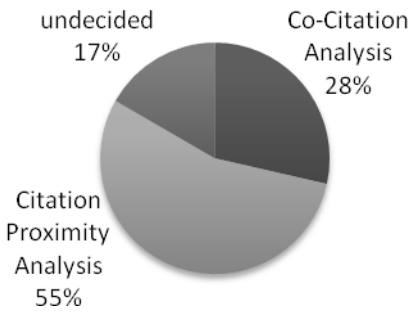


Figure 4: Comparison of CPA and Co-citation analysis

As the pie chart indicates, nearly twice as many study participants obtained more suitable documents when the CPA was used in comparison to the documents obtained by co-citation analysis. Not surprisingly, the study also substantiated the assumption that especially for documents with extensive bibliography or documents that have not been referenced frequently, CPA delivers superior results.

Taking into consideration that CPA essentially works like co-citation analysis with the distinctive difference that the proximity among citations is analysed and therefore additional information about relatedness is gathered, it is not surprising that CPA outperforms Co-citation analysis in every examined scenario[4].

**Outlook & Conclusion**

Besides identifying related work, the authors currently apply the idea behind CPA for automatic document classification for the research paper recommender Scienstein [11]. The aim is to automatically analyse the topics within documents by analysing 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 in contrast to a previous study [reference B]...*" So by analysing the words between two references it is often possible to automatically analyse in which relationship these two references stand to each other.

It is also often 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 gained information can be used to classify further documents and to develop a more sophisticated '*Web of Science*'[5]. 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.

As shown, the CPA offers substantial advantages in identifying related documents in comparison to existing approaches. However, it should also be taken into account that the effort to calculate the CPA 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

[4] A detailed description of the study and its results will be published seperatly.
[5] http://www.garfield.library.upenn.edu/papers/mapsciworld.html

abstract and the bibliography can be accessed, documents cannot be parsed as OCR[6] 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 the CPA delivers superior results, it cannot completely replace co-citation analysis.

## References


[1] May, R. M. 1997. The Scientific Wealth of Nations, Science, vol. 275, no. 5301, pp. 793-796.

[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.


[6] Optical Character Recognition