Erschienen in: Bulletin of IEEE Technical Committee on Digital Libraries ; 7 (2011), 1

Identifying Related Work and Plagiarism by Citation Analysis

Bela Gipp

OvGU, Germany / UC Berkeley, California, USA
gipp@berkeley.edu

Abstract
This updated and revised paper gives an overview of my PhD research. It focuses on two newly developed approaches. Citation Proximity Analysis (CPA) allows the identification of related work by analyzing the co-occurrence of citations within documents. In contrast to co-citation analysis various factors, such as the proximity of citations to each other, are taken into account. The second approach is called Citation based Plagiarism Detection (CbPD). In comparison to the currently used text-based plagiarism detection approaches this citation- analyzing approach enables a better detection rate in identifying plagiarism forms such as paraphrasing, translations and idea plagiarism.

Keywords: Document Similarity, Relatedness, Clustering, Plagiarism Detection, Duplicate Detection, Citation Analysis, Citation Proximity Analysis, Citation Order Analysis, Language Independent

1 Introduction & Motivation

The search for related work is such a time-consuming procedure, that even when performed by experienced scientists, it often leads to unsatisfying results. To alleviate the problem, search engines such as Google Scholar and Citeseer offer to display 'related documents'.

The best results are usually achieved by hybrid research paper recommender systems. By combining techniques such as citation analysis, co-word analysis, collaborative filtering, and Subject-Action-Object (SAO) structures, recommendations can be given. However, these approaches are only suitable to a limited extent for identifying related work [1, 10, 2, 8, 11, 6, 13]. According to our examinations, for scientific documents, the best results can usually be achieved by applying the citation-based bibliographic coupling and co-citation analysis.

The aim is to develop new citation-based approaches in order to identify related documents and plagiarism. So far, two new approaches have been developed, called Citation Proximity Analysis (CPA) and Citation based Plagiarism Detection (CbPD). CPA is a further development of co-citation analysis, whereas CbPD is based on bibliographic coupling, but in addition, analyzes the order of citations.

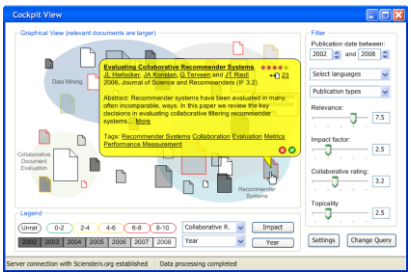

Figure 1: GUI SciPlore – clustering related documents

In the research paper recommender system SciPlore.org, these approaches are mainly used for two purposes. First, to identify related documents as shown in Figure 1; and secondly, to give recommendations for related documents based on one or more documents the user has been interested in, as shown in Figure 2.

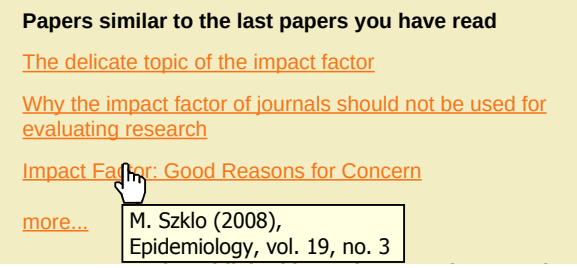

Figure 2: Recommendation of related papers

Throughout this document two types of semantic relatedness are distinguished. I adopt the perspective of Resnik and consider '*similarity*' as a special case of semantic relatedness [9]. Two documents, for instance, are *related* if they address the same research question. Two documents are *related* and *similar* if they are, for instance, duplicates, plagiarized or translated.

In the first part of this paper, related work is presented and currently applied citation analysis approaches discussed. In the next section the research design is presented. Afterwards, the CPA and CbPD are introduced and compared in regard to their suitability for Academic Recommender Systems. The paper concludes with a summary and an outlook.

2 Proposed Research & Related Work

The usefulness of a research paper recommender system depends to a large extent on its ability to automatically determine related documents to one or more documents. Various approaches exist to measure the degree of relatedness 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 [1, 2, 8]. Many citation analysis approaches exist and they all have their own strengths and weaknesses for identifying related 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 relevant that were referenced by the input document. Better results can usually be obtained by bibliographic coupling and co-citation analysis, which allow calculating the coupling strength [11]. These approaches, which were already invented in the 60s and 70s, are used by scientists and by academic search engines like CiteSeer[1] [3].

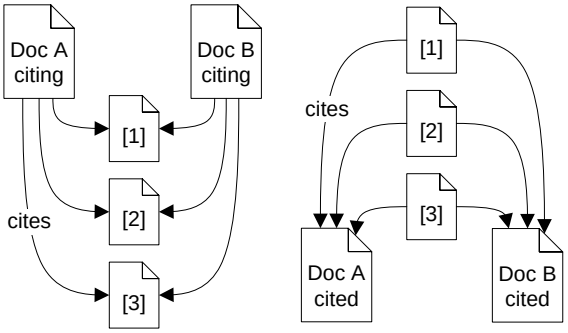

Figure 3: Bibliographic coupling (left) and Co-citation (right)

Documents are bibliographically coupled if they cite one or more documents in common. Figure 3 (left) illustrates this approach: Papers A and B are related because they both cite papers 1, 2 and 3.

In contrast, two documents are 'co-cited' when at least one paper cites both. This approach is illustrated in Figure 3 on the right: Papers A and B are related because they are both cited by papers 1, 2 and 3. The more co-citations two papers receive, the more related they are [11].

Although both approaches are suitable to identify related papers, they serve different purposes. Whereas bibliographic coupling is retrospective, co-citation is essentially a forward-looking perspective [3]. 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.

3 Research Questions

I want to answer the following three research questions in order to improve Academic Paper Recommender Systems.

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

- What are the strengths and weaknesses of the currently used approaches in order to measure semantic relatedness? (whether it be citation-, text- or user behavior-based)
- Is there a better way to automatically measure semantic relatedness?
- How do these new approaches perform in comparison to the currently used approaches?

4 Methodology

The methodology follows six steps. Currently, the empirical study is in progress. .

1. Literature review and evaluation of existing approaches
    - Text mining (bag of words, etc.)
    - Citation analysis (bibliographic coupling, co-citation analysis)
    - Community based approaches (tagging, annotating etc.)
    - Further aspects like ranking algorithms, collaborative document evaluation, mind maps, etc.
2. Development of two new approaches to alleviate the shortcomings of existing approaches
    - Citation / Quotation Proximity Analysis (CPA)
    - Citation based Plagiarism Detection (CbPD)
3. Implementation of existing and new approaches in prototype
    - see www.SciPlore.org
4. Empirical comparison and analysis of suitability (qualitative and quantitative)
    - Quality of results
    - Performance
5. Extension and optimization of new approaches
    - Combination with existing approaches
    - Adjustment of parameters
6. Development of a procedure model that considers the document type
    - Scientific publications containing citations and a clear structure such as abstract, related work, findings etc.
    - Websites, patent applications, technical documents, etc.

5 First Results

Two new approaches called Citation Proximity Analysis (CPA) and Citation based Plagiarism Detection (CbPD) have been developed. 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 CbPD, the pattern, order, co-occurrence etc. of citations is analyzed, allowing the identification of a text that has been translated from language A to language B, as the citations remain in a similar or even identical order.

Citation Proximity Analysis (CPA)

Instead of just using the bibliography, in CPA 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[2].
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 related is higher (CPI = 1) than if they are cited only within the same paragraph (CPI = ¼). See Figure 4.

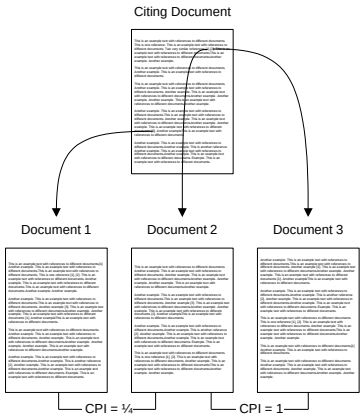

Figure 4: Illustration CPA

However, further research needs to be performed to identify the appropriate weighting of the CPI values according to their occurrence, which also seems to

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

depend on the publication's research field or type. It seems, for instance, that for analyzing a technical report or patent specification, different weightings seem more suitable than for a research article.

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 related documents [4].

The same principle can be applied to links on websites or to quotations instead of citations (Quotation Proximity Analysis). If passages of two documents are quoted by a third document, the quoted documents are likely to be related. The closer the quotations are within the text of the quoting document, the higher the assumed relatedness as illustrated in the following figure.

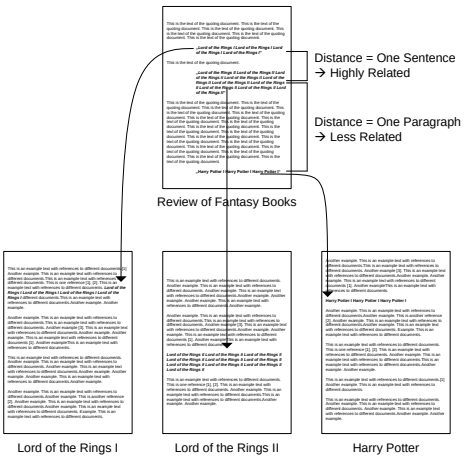

Figure 5: Quotation Proximity Analysis

The 'Review of Fantasy' book quotes passages from two different editions of 'Lord of the Rings' and of 'Harry Potter.' Between the quotes of the different 'Lord of the Rings' volumes only one sentence occurs. Therefore, a relatively high relatedness of these two quotes/quoted books can be assumed. In contrast, the distance between the quote from 'Harry Potter' and the 'Lord of the Rings' is larger. Therefore, the relatedness of these quotes and the quoted books can be assumed to be lower, but still higher as if they would not appear at all in the same document. A modification of the approach also allows classifying unknown documents based on containing quotes. In the example, the 'Review of Fantasy Books' could be classified automatically if at least one of the quoted books has already been classified. This is especially useful for documents not containing references or quotes as for instance in novels.

Citation based Plagiarism Detection

Similar to the idea of CPA is another approach, which I call Citation based Plagiarism Detection (CbPD) or Citation Order Analysis [5]. Hundreds of papers have been published covering sophisticated approaches to detect plagiarism, and dozens of applications have been developed. All of them use more or less sophisticated approaches to analyze the text, but ignore the used citations [7, 12]. These approaches deliver good results in detecting copied text passages, but fail if text has been paraphrased or translated as shown in Figure 6.

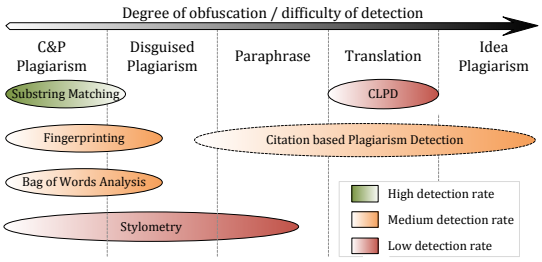

Figure 6: Comparison of Plagiarism Detection Systems

In contrast to CPA, in CbPD mainly uses factors such as citation order and pattern analysis. The main advantage in comparison to the usually applied text-analysis approaches is that even if documents were translated or paraphrased they can still be identified as similar. Figure 7 and Figure 8 illustrate the concept.

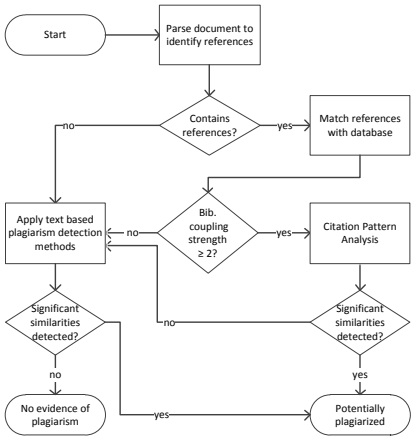

Figure 7: Citation Pattern Analysis

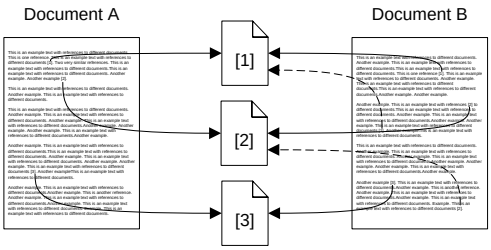

Figure 8: Illustration of Citation Order Analysis

A comparison with the existing approaches is problematic, as both approaches have their own strengths. Whereas text-based approaches detect local similarity, like copied sentences, this citation-based approach analyzes global similarity. The interpretation, for instance, of a precision and recall value only makes sense when compared to other approaches. Since no other approaches exist for paraphrased and translated scientific text, such a comparison is not feasible. The test sets, like the PAN-PC-10 that was used at the Competition on Plagiarism Detection in 2010, are tailored to compare the performance of classical plagiarism detection systems, but are unsuitable to test this new approach, because citations were ignored.

To evaluate our approach, we ran a test on 0.8 million scientific publications from open access repositories and hid among them 20 specially-designed plagiarized documents. To create a more realistic test scenario, we deleted some citations, added new ones, changed the order slightly, and changed the citation style. The outlined approach identified 19 of the test documents, along with hundreds that contained at least some plagiarized sections. One very short document was not identified; it cited five sources, of which we deleted two. Figure 9 shows how the CbPD compares to the currently used text-based detection approaches. It also indicates that the performance is best if the text-based and the citation-based approach are combined.

above average = 1 average = 2 below average = 3	existing				new	
	Fingerprinting	Vector Space Retrival	Substring Matching	Intrinsic	Citation Pattern Based	Combined (Text&Citation)
Copy&Paste (c&p)	1	1	1	2	1	1
Shake&Paste (s&p)	1	1	1	2	2	1
Expansive	2	2	3	3	2	1
Contractive	1	1	2	3	2	1
Mosaic	2	2	2	3	3	2
Technical disguise	3	3	3	3	1	1
Undue paraphrase	3	3	3	3	1	1
Translated	3	3	3	3	1	1
Idea plagiarism	3	3	3	3	2	2
Self-plagiarism	1	1	1	3	1	1

Figure 9: Comparison of Detection Quality

By lowering the threshold, not only can plagiarism be detected, but also documents which have not been cited, that were involved in the creation process. Figure 10 shows an example. Document A was probably read by the author of Document B, but Doc A was not cited. This is not usually considered plagiarism, but knowledge concerning which papers were involved in the creation process can be of interest.

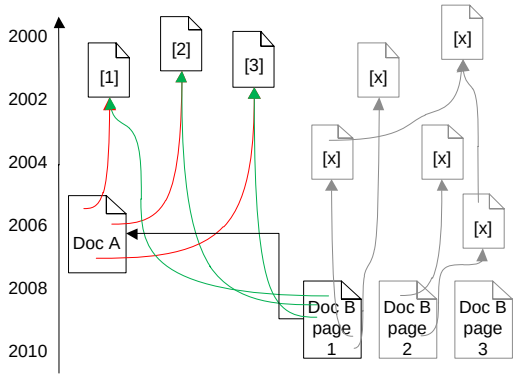

Figure 10: Identification of non-cited documents

6 Conclusion

This paper gave an overview of my PhD research project, which addresses the difficulty of measuring document relatedness in order to e.g. improve Academic Recommendation Systems and to identify plagiarism. Two approaches were presented and their advantages and disadvantages discussed. For more in-depth information please consult my publications.

References

[1] Joeran Beel and Bela Gipp. The Potential of Collaborative Document Evaluation for Science. In George Buchanan, Masood Masoodian, and Sally Jo Cunningham, editors, *11th International Conference on Asia-Pacific Digital Libraries (ICADL'08) Proceedings*, volume 5362 of *Lecture Notes in Computer Science (LNCS)*, pages 375–378, Heidelberg (Germany), December 2008. Springer. Also available on http://www.sciplore.org.

[2] RM Fano. Information theory and the retrieval of recorded information. In *Documentation in Action: Based on 1956 Conference on Documentation at Western Reserve University*, page 238. Reinhold Publishing Corp., 1956.

[3] E. Garfield. From bibliographic coupling to co-citation analysis via algorithmic historio-bibliography. volume 27, 2001.

[4] Bela Gipp and Joeran 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. Also available on http://www.sciplore.org.

[5] Bela Gipp and Joeran Beel. Citation Based Plagiarism Detection - A New Approach to Identify Plagiarized Work Language Independently. In *Proceedings of the 21st ACM Conference on Hyptertext and Hypermedia (HT'10)*, pages 273–274, New York, NY, USA, June 2010. ACM.

[6] R. Klavans and K.W. Boyack. Identifying a better measure of relatedness for mapping science. *Journal of the American Society for Information Science and Technology*, 57(2):251–263, 2006.

[7] R. Lukashenko, V. Graudina, and J. Grundspenkis. Computer-based plagiarism detection methods and tools: An overview. In *Proceedings of the 2007 international conference on Computer systems and technologies*, page 40. ACM, 2007.

[8] IV Marshakova. System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6(2):3–8, 1973.

[9] P. Resnik et al. Using information content to evaluate semantic similarity in a taxonomy. In *International Joint Conference on Artificial Intelligence*, volume 14, pages 448–453, 1995.

[10] A. Rip and J.P. Courtial. Co-word maps of biotechnology: An example of cognitive scientometrics. *Scientometrics*, 6(6):381–400, 1984.

[11] 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*, 24:265–269, 1973.
[12] Benno Stein, Paolo Rosso, Efstathios Stamatatos, Moshe Koppel, and Agirre Eneko, editors. *Proceedings of the 3rd PAN Workshop. Uncovering Plagiarism, Authorship and Social Software Misuse*, 2009.
[13] C. Sternitzke and I. Bergmann. Similarity measures for document mapping: a comparative study on the level of an individual scientist. *Scientometrics*, 78(1):113–130, 2009.