Link Analysis in Mind Maps: A New Approach to Determining Document Relatedness

Jöran Beel
Otto-von-Guericke University
FIN / ITI / VLBA-Lab
Germany
beel@sciplore.org

Bela Gipp
Otto-von-Guericke University
FIN / ITI / VLBA-Lab
Germany
gipp@sciplore.org

ABSTRACT
In a previous paper we presented various ideas on how information retrieval on mind maps could enhance applications such as expert systems, search engines and recommender systems. In this paper we present the first research results. In a brief experiment we researched link analysis respectively citation analysis, if applied to mind maps, is suitable to calculate document relatedness. The basic idea is that if two documents A and B are linked by the same mind map, these documents are likely to be related. This information could be used by item-based document recommender systems. In the example, document B could be recommended to those users interested in document A. In addition, we propose that those documents linked in high proximity within a mind map are more closely related than those documents linked in lower proximity. The results of our experiment support our ideas. It seems that link analysis applied to mind maps can be used for determining the relatedness of documents and therefore for improving document recommender systems.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering, retrieval models, search process*

H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *system issues, user issues*

General Terms
Algorithms, Measurement, Documentation.

Keywords
mind maps, recommender systems, research paper recommender, document recommender, metrics, citation analysis, link analysis

Proceedings of the 4th International Conference on Ubiquitous Information Management and Communication'10, January 14–15, 2010, Suwon, Korea.

1. INTRODUCTION
Mind mapping is a common method to structure and visualize ideas, manage electronic literature and to draft documents. Some users do link in their mind map to external documents such as PDFs or websites. Some even cite scholarly literature, for instance by adding BibTeX keys to a mind map's node (see Figure 1 for an example). In a recent paper we proposed to analyze these links and references to determine the relatedness of those documents that are linked in the mind map [1][1].

The basic idea is that two documents are related if they are both linked by a mind map. In addition, it was assumed that the closer the links occur in the mind map, the higher related the linked documents are. If the assumption proves to be right, *Link Analysis in Mind Maps* (LAMM) could be used to enhance search engines and document recommender systems since these systems often present related documents to their users.

We conducted a brief experiment to test the proposed idea and present the results in this paper. The focus of this paper lies on calculating the relatedness of scholarly literature and on enhancing research paper recommender systems as we plan to integrate LAMM into our academic search engine and research paper recommender system *SciPlore*[2]. However, it's highly probable that the results would be similar for other kind of documents linked by a mind map such as websites.

In the next section, related work about research paper recommender systems and citation analysis is presented. It is then followed by a section showing the methodology which has been used to evaluate LAMM. Finally, the results, a discussion, and an outlook towards future work conclude.

2. RELATED WORK
Several attempts have been made to establish research paper recommender systems [2-7]. Some of them use citation analysis to determine the degree of relatedness between two papers. An overview of different citation analysis approaches for determining the relatedness of research papers is given in [8]. At

[1] In this paper we do not distinguish between linking files and referencing scholarly literature, for instance with a BibTeX key. Citations, links to files on the user's hard drive and hyperlinks to websites are all considered as 'link'.

[2] http://www.sciplore.org

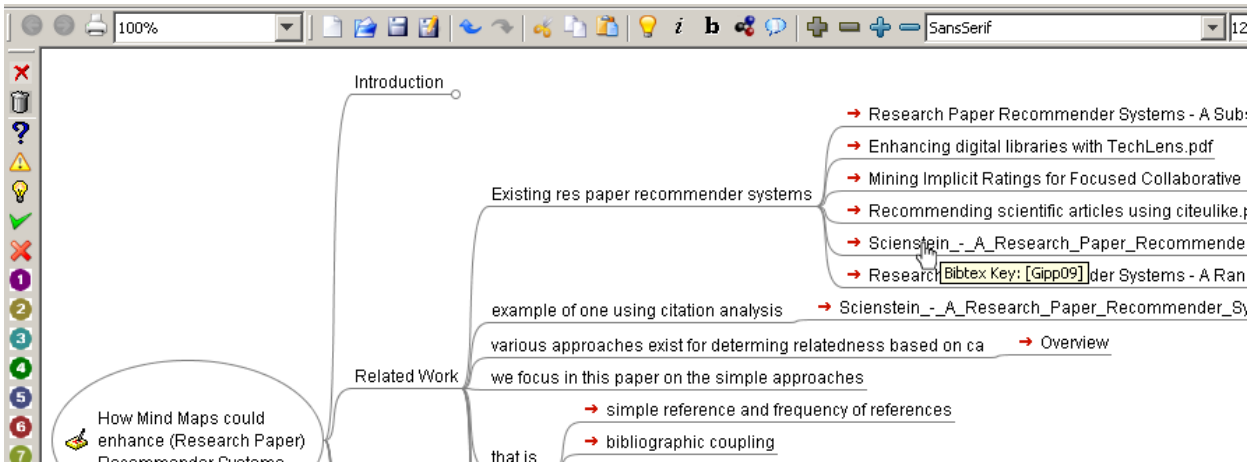

Figure 1: Mind map draft of a paper (arrows indicate a link to a PDF file; the tooltip displays a BibTeX key)

this time, our research focuses on *co-citation analysis* [9] and its extension *citation proximity analysis* [10].

According to co-citation analysis, two papers A and B are related if a third paper C references both. If more than one paper reference paper A and B together, their relatedness is supposed to be even higher. Citation proximity analysis additionally considers the location of citations in the full text: Two papers A and B are supposed to be more highly related when they are closely referenced by a third paper C in the text. For instance, if paper C references paper A and B in the same sentence, A and B are likely to be highly related. If paper C references paper A in the beginning of a 100-page document and paper B at the end, their relatedness is probably not nearly as high.

Co-citation analysis and citation proximity analysis can be used by research paper recommender systems to make item-based recommendations: If paper A and B are related, paper B may be recommended to those users interested in paper A (but not knowing paper B yet).

However, co-citation analysis and citation proximity analysis have to cope with some drawbacks.

1. **Availability of Data:** Co-citation analysis and citation proximity analysis cannot be applied to all research papers due to a lack of (correct) data [11, 12]: many research papers are not cited at all; citation databases such as ISI Web of Knowledge do not cover all available publications; and due to technical difficulties, citations are not always recognized correctly, which in turn leads to incorrect data in citation databases.
2. **Robustness of Data:** Citations are often considered as biased because authors do cite papers they should not cite and do not cite papers they should cite [12]. Accordingly, citation based recommender systems might provide irrelevant recommendations.
3. **Timeliness of Data:** Publishing scientific articles is a slow process and it takes months or even years before they are published and citations are received. Accordingly, documents recommended based on citation analysis are, at the very least, several months old.
4. **Metrics:** There exist metrics for measuring the relatedness of research papers based on citation analysis (for instance, *coupling strength* [13] or the *citation proximity index* [10]). However, to our knowledge, each metric focuses solely on one citation analysis approach and no combining metric exists yet. Consequently, relatedness of research papers based on citations cannot be measured and expressed thoroughly.

Summarized, citation analysis applied to scholarly literature can do a good job in identifying related articles, but there is room for improvement.

3. METHODOLOGY

Our intention was to conduct an experiment to obtain first indications if *Link Analysis in Mind Maps* (LAMM) might be suitable for determining research paper relatedness. Two assumptions were researched:

1. Two research papers A and B are related if at least one mind map links them both
2. Two research papers A and B are more highly related the more closely they are linked within a mind map

As part of the experiment, five mind maps were analyzed which were originally created for drafting research papers, respectively Masters Theses[3]. That means each of the mind maps links at least to a few PDF files representing academic articles. From each mind map, links (respectively citations) to three articles were extracted and pairs were built (see Figure 2 for illustration). The first pair was built from the first and second link in a mind map. Since the distance between them was low, we expected this pair to be 'highly related'. The second pair was built from the first and last link in the mind maps. Here, the distance between the links was high. Accordingly, we expected the corresponding articles to be less closely related.

[3] Two mind maps represented drafts of our own papers and three mind maps were created by some of our students for their Masters' theses.

To test our assumptions, titles and abstracts of the linked PDFs were extracted. Since five mind maps were analyzed, five pairs with low distance (expected relatedness = (very) high) and five pairs with high distance (expected relatedness =low) existed. In addition, five 'control pairs' of papers were created. We created these pairs in a way that they should appear as not being related to each other at all[4].

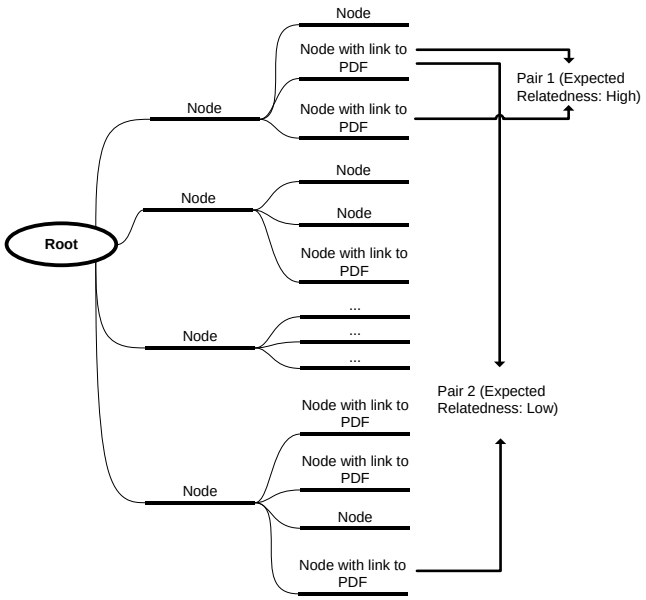

Figure 2: Link Extraction from the Mind Maps (Illustration)

To build a pair of papers with a potentially high relatedness, the very first and the second link of a mind map were taken, so the proximity between the two papers was high. To obtain a pair of papers with rather low relatedness, the very first and the very last link of a mind map were taken.

All pairs were shown to five participants[5] and the participants had to rate the relatedness of the pairs on a scale from 1 to 5 (1 = not related, 5 = highly related) . For evaluation, ratings were painted in a scatter plot for each participant as well as the overall rating (mean and median). A more detailed statistical analysis was not considered necessary, since the graphs showed quite clear results and the amount of data was too little for extensive statistic analyses.

4. RESULTS

Figure 3 shows the results. On average (mean), those pairs linked closely together in the mind maps were considered significantly more often (highly) related than those pairs not linked closely together. The control pairs, which were not linked by any mind map, were all rated as not related, on average.

Some outliers exist: On average, pair 2 in mind map 2 was considered higher related than pair 1 in mind map 2. In addition, pair 2 of mind map 3 and pair 1 of mind map 5 were rated as almost not related. However, this is not surprising since mind maps are usually used for drafting a paper and therefore variances are to be expected.

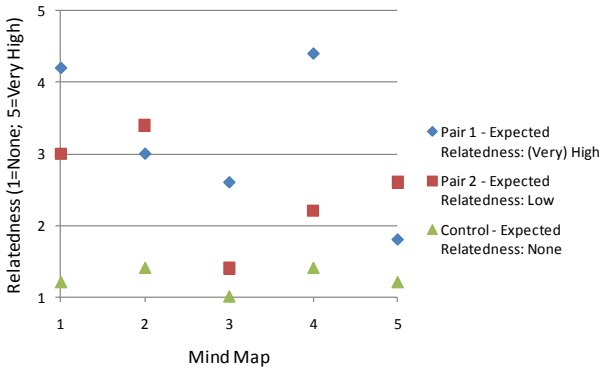

Figure 3: Relatedness of Pairs in Mind Maps (Mean)

5. DISCUSSION

Overall, the results are a first indication that mind maps can be used to calculate research papers' relatedness. However, it needs to be emphasized that all five mind maps were created by our students and ourselves and hence came from the same 'school of thought'. It's very possible that other researchers use mind maps in a different way, which would then lead to variations in the results.

In addition, similar problems as for classic ciation analysis are to be expected for *Link Analysis in Mind Maps*. These problems are related to data availability, robustness, timeliness and metrics and are discussed in the following sections.

5.1 Availability of Data

Data availability seems to be the main challenge LAMM will have to face. It is unknown how many researchers use mind maps and how many are willing to share their data. It could be that the number is rather low. Nevertheless, mind mapping is a popular application. For instance, the mind mapping tool *FreeMind* is downloaded over a 150,000 times a month [14], more than 1.5 million people use *MindManager* [15] and there exist dozens of tools more [16, 17]. Even platforms for sharing mind maps exist already[6]. On our website sciplore.org we also offer a special mind mapping software for researchers which will enable us to collect mind maps [18].

Overall, we are confident, that sufficient data can be collected that makes LAMM worth researching. Certainly, it will never replace citation analysis in scholarly literature or hyperlink analysis on websites but LAMM could serve as a complement for both.

[4] The papers were taken from the SciPlore database, were not linked by any of the mind maps and did not cite each other.

[5] None of the participants were involved in creating the mind maps. The pairs of papers were distributed to the participants without their knowledge of the pairs being linked by a mind map or not. Each participant was shown all 15 pairs at once.

[6] For instance, http://www.mappio.com, http://share.xmind.net, and http://www.mindmeister.com/maps/public/

Technical problems (in terms of identifying references) should be equal or even less for LAMM than for classic citation analysis. If users link to a unique identifier such as a BibTeX key, the corresponding metadata should be easily extractable from the user's bibliographic database. If the user links a PDF file, at least the title should be easily identifiable from the PDF, in most cases[7].

5.2 Robustness of Data

All social media platforms do have to cope with spam and fraud as soon as they become successful. There is no reason to assume this would be different if mind maps were used for calculating relatedness of documents. However, most social media platforms also find a way to cope with fraud and spam. If only mind maps of 'trusted' users were used, serious spam and fraud could probably be prevented successfully. Trustworthiness of users probably could be determined in cooperation with social networks, other community websites or by usage data of mind mapping software.

5.3 Timeliness of Data

With LAMM, timeliness has a clear advantage over classic citation analysis. Mind maps do not need to be published in journals or at conferences. They could be analyzed the moment they are created. This would enable research paper recommender systems to recommend new publications faster than with classic citation based approaches.

5.4 Appropriate Metrics

LAMM could use the same metrics that are used for citation analysis. Perhaps slight modifications would have to be made, but overall, metrics should be very similar (and so the advantages and disadvantages of citation based metrics).

6. SUMMARY & FUTURE RESEARCH

In this paper we presented L*ink Analysis in Mind Maps* (LAMM). LAMM is an approach for determining the relatedness of documents by applying methods from hyperlink and citation analysis to mind maps. The basic idea is: If two documents A and B are linked or referenced by a mind map, these articles are likely to be related. Consequently, a recommender system could recommend document B to those users liking document A. In addition, we proposed that two documents are higher related when their proximity in the mind map is higher. In a small study (five mind maps and five participants) we obtained first indications that our assumptions could be true. The participants rated research articles that were linked in high proximity in the mind map, as more highly related than those articles linked within low proximity. Advantages and problems of LAMM in comparison to classic citation analysis were also discussed in this paper. Especially in respect to timeliness, MMCA seems likely to outperform classic citation analysis. On the other hand, data availability is likely to be a much larger problem than it is for citation analysis.

Overall, LAMM might prove to be a promising field of research having the chance to complement classic citation analysis and enhance research paper recommender systems in the long run. However, there is a need for more research since many questions remain unanswered:

- How many researchers are using mind maps?
- How many are willing to share them?
- How can spam and fraud be prevented?
- Which metrics should be used to measure relatedness?
- How should these metrics be combined with existing ones based on citations and other techniques (for instance, based on text mining and collaborative filtering)?

While this paper focuses on determining relatedness of scholarly literature, LAMM could be applied equally well to other document types such as web pages.

[7] We developed a tool for extracting titles from PDFs. First tests are promising.

7. REFERENCES

[1] Jöran Beel, Bela Gipp, and Jan Olaf Stiller. Information Retrieval on Mind Maps – What could it be good for? In *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom'09)*, Washington (USA), November 2009. IEEE.

[2] N. Agarwal, E. Haque, H. Liu, and L. Parsons. Research Paper Recommender Systems: A Subspace Clustering Approach. *Lecture notes in computer science*, 3739: 475, 2005.

[3] R. Torres, S.M. McNee, M. Abel, J.A. Konstan, and J. Riedl. Enhancing digital libraries with TechLens. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, pages 228–236. ACM New York, NY, USA, 2004.

[4] T.Y. Tang and G. McCalla. Mining implicit ratings for focused collaborative filtering for paper recommendations. In *UM 2003, Workshop on User and Group Models for Web-based Adaptive Collaborative Environments*, 2003.

[5] T. Bogers and A. van den Bosch. Recommending scientific articles using citeulike. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 287–290. ACM New York, NY, USA, 2008.

[6] M. Gori and A. Pucci. Research paper recommender systems: A random-walk based approach. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 778–781. IEEE Computer Society Washington, DC, USA, 2006.

[7] Bela Gipp, Jöran Beel, and Christian Hentschel. Scienstein: A Research Paper Recommender System. In *Proceedings of the International Conference on Emerging Trends in Computing (ICETiC'09)*, pages 309–315, Virudhunagar (India), January 2009. Kamaraj College of Engineering and Technology India, IEEE. Available on http://www.sciplore.org.

[8] W. Lu, J. Janssen, E. Milios, N. Japkowicz, and Y. Zhang. Node similarity in the citation graph. *Knowledge and Information Systems*, 11 (1): 105–129, 2007.

[9] I V Marshakova. System of document connections based on references. *Nauchno-Tekhnicheskaya Informatsiya*, 2 (6): 3–8, 1973.

[10] 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. Available on http://www.sciplore.org.

[11] D. Lee, K. Jaewoo, M. Prasenjit, L. Giles, and O. Byung-Won. Are your citations clean? *Communications of the ACM*, 50: 33–38, 2007.

[12] M.H. MacRoberts and B. MacRoberts. Problems of Citation Analysis. *Scientometrics*, 36: 435–444, 1996.

[13] 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.

[14] SourceForge. SourceForge.net: Project Statistics for FreeMind. Website, 2008. URL http://sourceforge.net/project/-stats/-detail.php?group_id=7118&ugn=freemind&type=prdownload&mode=year&year=2008&package_ihttp://sourceforge.net/project/-stats/-detail.php?group_id=7118&ugn=freemind&type=prdownload&mode=year&year=2008&package_id=0.

[15] MindJet. MindJet: About MindJet. Website, Juli 2009. URL http://www.mindjet.com/about/.

[16] Open Directory Project DMOZ. Open Directory - Reference: Knowledge Management: Knowledge Creation: Mind Mapping: Software. Website, Juli 2009. URL http://www.dmoz.org/-Reference/Knowledge_Management/Knowledge_Creation/-Mind_Mapping/Software/.

[17] Mind-Mapping.org. Software for mindmapping and information organisation. Website, Juli 2009. URL http://-www.mind-mapping.org/mind-mapping-software/-35?selectedCategories[]=mind%20maps&selectedOSes[]=all%20operating%20systems&pastOrPresent[]=current&datePicker2=&filterData=Show+selected+items.

[18] Jöran Beel, Bela Gipp, and Christoph Müller. 'SciPlore MindMapping' – A Tool for Creating Mind Maps Combined with PDF and Reference Management. *D-Lib Magazine*, 15 (11), November 2009. doi: 10.1045/november2009-inbrief. URL http://www.dlib.org/dlib/november09/11inbrief.html. Brief Online Article. Alternatively available on http://www.sciplore.org.