

# Retrieving Data from Mind Maps to Enhance Search Applications

Jöran Beel<sup>1</sup>

<sup>1</sup> Otto-von-Guericke University, Computer Science/ITI/VLBA-Lab  
Magdeburg, Germany

and

University of California, Berkeley, School of Information  
Berkeley, CA, United States

j.beel@sciplore.org

**Abstract.** Web search, academic search and expert search engines often have difficulties in classifying and ranking objects and generating summaries for them. In this paper I propose that data retrieved from mind maps may enhance these search applications. For instance, similar to anchor text analysis, documents, i.e. web pages or academic articles, linked in a mind map might be classified with text from the linking nodes. This idea, along with additional ideas, related work, the indented methodology, and first research results are summarized in this paper.

**Keywords:** mind maps, information retrieval, search applications, digital libraries, document classification, document relatedness, expert finding

## 1 Introduction & Motivation

Search engines need to classify objects such as websites, books, academic papers or people, and make them retrievable, usually via keyword search. This task includes the need for good ranking algorithms to determine the order of search results. Search engines also need to summarize the contents of indexed objects and display these summaries on the result page. For a website, a summary could be the website's title or extracts of those parts of the website in which the search term occurs. Many search engines also offer a function to display related objects. For instance, when a user finds an interesting web page with Google he can click on the "similar" link to retrieve a list of similar web pages.

However, neither classifying objects, nor ranking them, nor creating summaries, nor identifying related objects is a trivial task. Synonyms, unclear nomenclature, and different needs of users are only some of the difficulties search engines have to cope with.

In this paper I present my PhD project which aims at enhancing the classification and ranking of objects, summarization of objects and identification of related objects.

An *object* can be any item search engines index such as web pages (web search), books (book search), academic articles (academic search), mp3s (music search) or a person's skills and interests (expert search). The focus of my research is on enhancing these search applications by utilizing data retrieved from mind maps. Mind maps are special kinds of documents that have not been utilized for information retrieval before, to the best of my knowledge.

In the following, related work is presented about mind mapping and problems that search engines have to cope with. Then, the main research question and some ideas are presented how data retrieved from mind maps could help in solving or at least alleviate the stated problems. In the fourth section the intended methodology is presented, followed by first research results that were obtained during the last few months of my study.

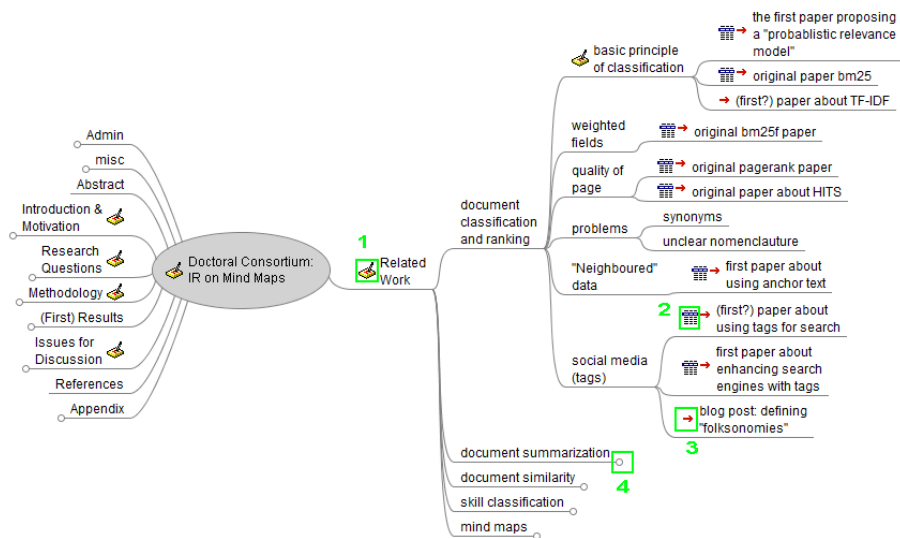


Figure 1: Mind Map of this Paper

## 2 Proposed Research & Related Work

In the following I describe some initial ideas how mind maps and data retrieved from them could be used to enhance search applications. In addition, related work is presented. All ideas are based on the assumption that the text in a mind map may describe the mind map's author as well as objects linked in the mind map.

## 2.1 Mind Maps

Mind maps were originally invented by Tony Buzan in the 1970s [1] and are nowadays used by millions of people for brainstorming, note taking, project planning, decision making, and document drafting. A mind map is a graphical representation of ideas and always has a central node which represents the main idea the mind map is about. From the central node child-nodes cover sub-topics. Figure 1 shows an example – the mind map I created as draft of this paper. Hundreds of books and research articles were published about how to create mind maps and about evaluating mind maps’ effectiveness, for instance, in the field of education.

Over 100 software tools exist to support the creation of mind maps [2]. The most popular ones are *MindManager* with about 1.5 million users [3] and *FreeMind* with about 150,000 downloads a month [4]. Most software tools for creating mind maps support the following key features: the addition of notes to a node (see ‘1’ in Figure 1), adding attributes to a node, e.g. a BibTeX key (see ‘2’ in Figure 1), linking a node with a file, e.g. a PDF (see ‘3’ in Figure 1), and unfolding branches (see ‘4’ in Figure 1: all branches with the little circle can be unfolded which would reveal more nodes).

As there are millions of mind mapping users, there are presumably tens of millions of mind maps on the users’ computers. In addition, there are mind map galleries on the internet on which users can offer their mind maps for download. All these mind maps contain lots of information about the users who created the mind maps as well as about objects linked in the mind maps (e.g. websites). Nevertheless, this rich source of information has not yet been used in information retrieval.

## 2.2 Document Classification and Ranking

### Related Work

In keyword search, a document is generally considered more relevant the more often the search term occurs in the document (in relation to the overall word count) [5]. Common algorithms based on term frequency are TF-IDF [6] and BM25 [7]. For some years, not only word frequency but document structure has been taken into account: a word occurring in the title is weighted more heavily than a word occurring, for instance, in the body text. A popular algorithm in this field is BM25f [8]. Problematic with such approaches solely based on term frequency and document structure is that a document will only be found for specific words the author has used in the document, but not for any synonyms.

Therefore, some search engines classify documents with terms contained in ‘neighbored’ documents. That means if document A and document B are somehow related, terms occurring in document B might be used to classify document A. Relatedness usually is assumed when two documents are connected via hyperlinks (in case of websites) or references (in case of scholarly literature). A common approach in this field is analyzing anchor text of links [9]. That means terms from the descriptive text of a hyperlink are used to classify the linked document.

An extension of anchor text analysis is the analysis of social tags or so called “folksonomies” [10] to enhance search engines. Social tags can be used to classify documents *instead* of analyzing the document’s text [11] or as *addition* [12]. In the

latter case a social tag could be seen as additional text field with a certain weighting comparable, for instance, to the occurrence of the term in the title or abstract.

To improve ranking, some search engines consider the quality or popularity of a document. The basic assumption is that the more popular a document, the more likely other users will like it. Accordingly, popular documents are ranked higher than less popular documents. Some of the most applied algorithms for determining document popularity are Google's PageRank [13] and HITS [14].

Overall, classification and ranking of documents works quite well. However, every search engine user knows that there is still room for improvement.

### **Proposed Research**

I propose that mind maps may be seen as neighboring documents to those documents linked in the mind map. Accordingly, approaches such as anchor text analysis could be applied to mind maps. In the example (Figure 1) one of the nodes, called "original pagerank paper", links to a PDF file on my hard drive. This linked PDF is the article in which S. Brin and L. Page introduce the PageRank algorithm [13].

The text of the node "original pagerank paper" could be considered as anchor text and the linked paper could be classified with it. In addition, not only text from the linking node itself could be taken but from parental nodes, too. In the example, the parent nodes of "original pagerank paper", i.e. "quality of page" and "document classification and ranking", describe Brin and Page's paper quite well, too.

However, the challenge would be to identify the appropriate nodes. For instance, the parent-node "related work" would not appropriately describe the linked paper. In addition, some nodes do not contain a few keywords but a long text. In this case, it would be necessary to extract the most relevant keywords from the text.

## **2.3 Document Summarization**

### **Related Work**

Search engines display summarizing data for a search result. This may be the document title, URL, or an extract of the document's text. Alternatively to extracts, researchers automatically created abstracts [15] and created summaries based on user generated data such as hyperlinks [16], and social annotations [17]. However, it is debatable whether current summarization approaches deliver satisfying results. For instance, the summary of the search result in Figure 2 does not seem to be very informative as there is lots of repetition.

[Using \*\*Mind Maps\*\* To Teach Social Problems Analysis.](#)

AR Peterson, PJ Snyder - 1998 - [eric.ed.gov](#)

... ED424882 - Using **Mind Maps** To Teach Social Problems Analysis. ... ERIC #: ED424882.

Title: Using **Mind Maps** To Teach Social Problems Analysis. ...

[Cited by 10](#) - [Related articles](#) - [Cached](#) - [Import into BibTeX](#)

**Figure 2: Example of Summary Data on Google Scholar**

### Proposed Research

Mind maps could be used to complement summarization of documents. A node's text, and the text of parent nodes, could be seen as a summary for the linked document. In the example (Figure 1) a summary for Brin and Page's paper [13] could be

*document classification and ranking – quality & popularity of page – original pagerank paper*

This summary could be displayed instead or in addition to text extracts (see Figure 3 for an example).

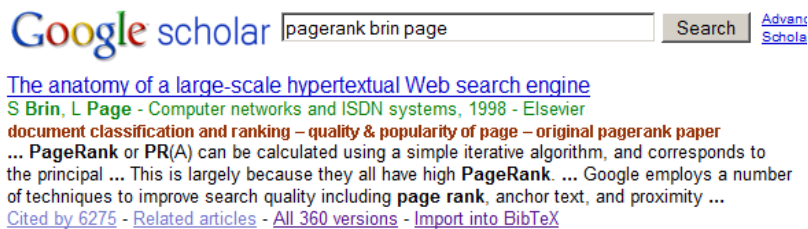


Figure 3: Additional Document Summary

## 2.4 Document Relatedness

### Related Work

To determine related documents, search engines often apply text analysis. Simply said, the more words two documents have in common, the more related they probably are ([18] provides a good overview on this topic). Collaborative filtering is another approach for determining related documents. It includes the process of clustering documents (or other objects) e.g. based on ratings of users ([19] provides an overview on collaborative filtering based recommender systems). Finally, citation analysis, more precisely bibliographic coupling [20] and co-citation analysis [21], are approaches to determine document relatedness, foremost in the field of scientific documents.

Text based approaches again have to cope with synonyms and unclear nomenclature and the other approaches suffer from the fact that usually only a fraction of items in a collection are cited/linked or rated.

### Proposed Research

Common link and citation analysis approaches could be applied on mind maps to calculate the relatedness of linked documents. The basic idea, analog to co-citation analysis, would be: two documents linked in the same mind map are related. This could be extended analog to Citation Proximity Analysis [22]: The higher the proximity of linked documents within the mind map, the higher their relatedness may be assumed.

In the example (Figure 1) the “original pagerank paper” [13] and “original paper about HITS” [14] are linked in direct proximity - the distance between them is just 1 (only the parental node divides them). In contrast, the “original bm25 paper” [7] is linked in a different branch of the mind map. Accordingly, the HITS and PageRank paper would be assumed to be more closely related than the PageRank and BM25 paper (which probably would be considered true by most people).

## **2.5 Skill Classification**

### **Related Work**

Finding the right experts in a big company e.g. for a certain project is a difficult endeavor. In first attempts, databases were used and employees could enter their skills manually [23].

In the last decade much research has been performed on automatically creating skill profiles. Probably the most promising approach is analyzing documents. For instance, if a researcher has published many documents containing the word ‘information retrieval’, she probably has some expertise in the field of information retrieval. Typical documents being analyzed are visited websites [24], emails [25], and other documents someone has written [26].

### **Proposed Research**

The ideas presented so far require that a user links external documents such as PDFs or websites in a mind map. However, even if no documents are linked in a mind map, interesting data can be retrieved – not about documents but about the mind map's author.

I propose that a mind map describes very well the skills and interests of its author. Therefore, expert search systems could be enhanced with data retrieved from mind maps. Similar to analyzing emails or other documents authored by a user, each node in a mind map may be taken as a description of the user's skills or interests. For instance, if the mind map in the example (Figure 1) was utilized by the expert search system my company employs, my boss might be presented with my name if he was searching for someone knowledgeable in the field of mind maps, information retrieval, PageRank, etc.

Knowing the skills and interests of someone is not only interesting for expert search, but also for advertising. If a user creates a mind map, for instance, about his upcoming holidays, relevant advertisement for hotels, rental cars, etc. could be displayed.

## **3 Research Question**

Although millions of people are creating mind maps and presumably tens of millions of mind maps must be stored on peoples' computers, no one (to my knowledge) has ever attempted to retrieve information from these mind maps to enhance other applications.

In my PhD project I will answer the research question:

*“How can data retrieved from mind maps be used to enhance search applications?”*

## **4 Methodology**

### **4.1 Data Collection**

The most critical issue in my PhD project is the availability of data. While millions of web pages or PDF files are publicly available, mind maps usually are not. However, there is around a dozen websites on which thousands of users offer their mind maps for public download (e.g. [www.mappio.com](http://www.mappio.com) or [www.xmind.net/share](http://www.xmind.net/share)). These mind maps should be sufficient for my PhD project. I am also developing my own mind mapping software *SciPlore MindMapping* ([www.sciplore.org](http://www.sciplore.org)) which allows me to analyze mind maps created by *SciPlore MindMapping*'s users.

### **4.2 Development**

After collecting a sufficient number of mind maps I will apply existing approaches from citation analysis and text mining on mind maps. Most likely, these approaches cannot be applied one to one. Therefore, I will have to modify them to suit the needs of mind mapping. I will also attempt to develop unique approaches solely focusing on mind maps. Furthermore, I will have to develop additional tools to analyze and extract data. For instance, I require a tool for identifying PDF files that are linked in the mind maps.

### **4.3 Evaluation**

Once the algorithms for classifying documents, creating summaries, determining similarities and creating skill profiles are developed, the data needs to be evaluated. This will be done via the academic search engine [SciPlore.org](http://SciPlore.org), which I am developing, and a web service.

[SciPlore.org](http://SciPlore.org) is an academic search engine which I will use for evaluating my results. So far, [SciPlore.org](http://SciPlore.org) classifies documents only with BM25 based on the full text of a document. For evaluation, results will be displayed as a mix of classifications based on BM25 only and classifications based on BM25 *and* data retrieved from mind maps. If results based on BM25 *and* mind maps are clicked more often by users than results based on BM25 only, I would consider my approach as superior. The same evaluation would be done for the other approaches I proposed in this paper: Based on click-through rates, the performance of my approaches is evaluated.

For my data I will offer a web service. This way, other websites (possibly CiteSeer, CiteULike, etc.) could access the data and enhance their services. Based on their click-through data, my approaches could then be evaluated, too.

## 5 First Results

In the past months some results were already obtained.

A tool for extracting titles from PDF files (from the full text; not the PDF's metadata) was developed which achieves descent results [27].

A mind mapping software was developed. The beta version can be downloaded at [www.sciplore.org](http://www.sciplore.org). A video is available demonstrating the features<sup>1</sup>.

A pilot study was conducted [28]. In this study the relatedness of documents was calculated based on mind maps. The participants rated those documents linked in high proximity in a mind map in most cases, as highly related. In contrast, the participants rated a control pair of papers significantly less often as related. However, the study had only 5 participants and can therefore only be seen as a very first indication that mind maps might be used to determine document similarity.

## 6 Summary & Outlook

In this paper several ideas were presented how data retrieved from mind maps could enhance search applications. It was proposed that mind maps may enhance classification of documents, when approaches such as anchor text analysis would be applied to mind maps. Further, mind maps might also enhance expert search, document summarization and document recommender systems. However, several challenges exist. Among others, data availability and the risk of fraud might become an issue.

## References

- [1] Toni Buzan. *Making the Most of your Mind*. Pan Books, 1977.
- [2] Mind-Mapping.org. Software for mindmapping and information organisation. Website, Juli 2009. URL [http://www.mind-mapping.org/?selectedCategories\[\]=all+categories&pastOrPresent\[\]=current](http://www.mind-mapping.org/?selectedCategories[]=all+categories&pastOrPresent[]=current).
- [3] MindJet. MindJet: About MindJet. Website, Juli 2009. URL <http://www.mindjet.com/about/>.
- [4] 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](http://sourceforge.net/project/stats/-detail.php?group_id=7118&ugn=freemind&type=prdownload&mode=year&year=2008).

---

<sup>1</sup> <http://youtube.com/watch?v=jRHqLktIMWw>

- [5] S.E. Robertson and K.S. Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science and Technology*, 27 (3): 129–146, 1976.
- [6] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24 (5): 513–523, 1988.
- [7] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. Okapi at TREC-3. In *Proceedings of the Third Text REtrieval Conference (TREC'94)*, 1994.
- [8] S. Robertson, H. Zaragoza, and M. Taylor. Simple BM25 extension to multiple weighted fields. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, pages 42–49. ACM New York, NY, USA, 2004.
- [9] O.A. McBryan. GENVL and WWW: Tools for Taming the Web. In *Proceedings of the First International World Wide Web Conference*, volume 341, 1994.
- [10] G. Smith. Folksonomy: Social Classification, August 2004. URL [http://-atomiq.org/archives/2004/08/folksonomy\\_social\\_classification.html](http://-atomiq.org/archives/2004/08/folksonomy_social_classification.html). Blog Post.
- [11] X. Wu, L. Zhang, and Y. Yu. Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, page 426. ACM, 2006.
- [12] Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, page 116. ACM, 2007.
- [13] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer networks and ISDN systems*, 30 (1-7): 107–117, 1998.
- [14] J.M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46 (5): 604–632, 1999.
- [15] HP Luhn. The automatic creation of literature abstracts. *Advances in Automatic Text Summarization*. MIT Press, Cambridge, MA, pages 58–63, 1956.
- [16] J.Y. Delort, B. Bouchon-Meunier, and M. Rifqi. Enhanced web document summarization using hyperlinks. In *Proceedings of the fourteenth ACM conference on Hypertext and hypermedia*, pages 208–215. ACM New York, NY, USA, 2003.
- [17] H. Zhang, Z.C.W. Ma, and Q. Cai. A study for documents summarization based on personal annotation. In *Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5*, pages 41–48. Association for Computational Linguistics Morristown, NJ, USA, 2003.
- [18] M.D. Lee, B. Pincombe, and M. Welsh. An empirical evaluation of models of text document similarity. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, pages 1254–1259, 2005.
- [19] J.B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. Collaborative filtering recommender systems. *Lecture Notes In Computer Science*, 4321: 291, 2007.
- [20] M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14: 10–25, 1963.
- [21] 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.

- [22] 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. Also available on <http://www.sciplore.org>.
- [23] ME Maron, S. Curry, and P. Thompson. An inductive search system: Theory, design, and implementation. *IEEE Transactions on Systems, Man and Cybernetics*, 16 (1): 21–28, 1986.
- [24] J. Wang, Z. Chen, L. Tao, W.Y. Ma, and L. Wenyin. Ranking user's relevance to a topic through link analysis on web logs. In *Proceedings of the 4th international workshop on Web information and data management*, pages 49–54. ACM New York, NY, USA, 2002.
- [25] C.S. Campbell, P.P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531. ACM New York, NY, USA, 2003.
- [26] M. Maybury, R. D'Amore, and D. House. Expert finding for collaborative virtual environments. 2001.
- [27] Jöran Beel, Bela Gipp, Ammar Shaker, and Nick Friedrich. SciPlore Xtract: Extracting Titles from Scientific PDF Documents by Analyzing Style Information (Font Size). In J. Jose, M. Lalmas, A. Rauber, F. Sebastiani, and I. Frommholz, editors, *Proceedings of the 14th European Conference on Digital Libraries (ECDL'10)*, volume 6273 of *Lecture Notes of Computer Science (LNCS)*, pages 407–410. Springer, September 2010. Also available on <http://www.sciplore.org>.
- [28] Jöran Beel and Bela Gipp. Link Analysis in Mind Maps: A New Approach To Determine Document Relatedness. In *Proceedings of the Fourth International Conference on Ubiquitous Information Management and Communication (ICUIMC'10)*. ACM, January 2010. Also available on <http://www.sciplore.org>.