Introducing Docear's Research Paper Recommender System

Joeran Beel
Docear
Magdeburg, Germany
beel@docear.org

Stefan Langer
Docear
Magdeburg, Germany
langer@docear.org

Marcel Genzmehr
Docear
Magdeburg, Germany
genzmehr@docear.org

Andreas Nürnberger
OvGU, FIN, DKE Group
Magdeburg, Germany
andreas.nuernberger@ovgu.de

ABSTRACT
In this demo paper we present Docear's research paper recommender system. Docear is an academic literature suite to search, organize, and create research articles. The users' data (papers, references, annotations, etc.) is managed in mind maps and these mind maps are utilized for the recommendations. Using content based filtering methods, Docear's recommender achieves click through rates around 6%, in some scenarios even over 10%.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *information filtering*

General Terms
Management, Design

Keywords
recommender systems, user model, mind map, mind mapping, research paper recommender system, content based filtering

1. INTRODUCTION
Literature management, i.e. searching, organizing and creating literature, is important for researchers and students. Especially the search for relevant literature is challenging due to the millions of articles and books being published every year and the fact that most search services such as the *ACM Digital Library* focus only on publications of selected publishers (e.g. *ACM*). In addition, full-texts are often not freely accessible and need to be paid unless ones university or library has a subscription for the publisher.

Our open source tool *Docear* (www.docear.org) supports researchers with literature management by bundling several applications that help in searching, organizing, and creating academic literature [2]. Docear has the unique feature of utilizing mind maps for information management. That means users organize their data in a tree-like data structure and not in a table or with social tags. Figure 1 shows an example of a mind map we created as draft for this paper. The mind map we created outlines the skeleton of this paper and the nodes (i.e. the "entities" with the words) link to articles in which the information was originally found. More information on Docear can be found in [2].

In this paper we present the research paper recommender system which we developed for Docear.

JCDL'13, July 22–26, 2013, Indianapolis, Indiana, USA.
ACM 978-1-4503-2077-1/13/07.

2. RELATED WORK
One of the earliest research paper recommender systems was TechLens [8]. TechLens used the two most popular recommendation approaches – content based filtering (CBF) and collaborative filtering (CF). In CBF, the words of a user's documents are taken to build a user model, and documents that contain the same words as the user model are recommended. In CF, similar users are determined by comparing how they rated items (the more often two users rated items alike, the more similar they are assumed to be). Then, items in the collection of the similar user are recommended to the other user.

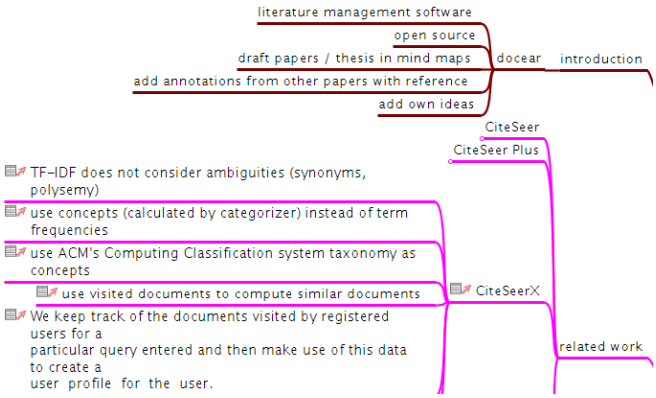

Figure 1: Draft of this paper as mind map

For the academic search engine CiteSeer there have been several different recommender systems proposed – some from the CiteSeer developers [4] and some from third parties [5]. Among others, they used citations instead of words to find similar scientific articles. Other research paper recommender systems include CiteULike's recommender system [3], Claper [9], and SCuBA [1].

All these recommender systems suffer from one problem: They have rather limited information about their users. For instance, if a recommender system analyses the papers a researcher has published [5], there are maybe a few dozens of papers to analyze. In contrast, recommender systems in other domains often have access to much more information about their users. For instance, on *Last.fm* a typical user listens to a few dozens of songs – on a single day.

3. DOCEAR'S RECOMMENDER SYSTEM
Docear has access to quite diverse, and quite a lot, information about its users. With Docear, users search for literature, they organize their literature, and they draft their own literature. That means Docear knows what a researcher is currently looking for, which articles a researcher already knows, which ones he is currently reading and on which new papers he is currently working. In addition, Docear allows its users to make annotations in PDF files, i.e. adding comments or highlighting passages they consider

important. Consequently, Docear does not only know which papers a researcher currently is reading but also which parts of an article he considers particularly important. This allows highly personalized recommendations.

We are experimenting with different variations of content based filtering for Docear. All variations have in common that from a user's mind map collection a user model is build and the user model is matched with Docear's Digital Library containing around 1.8 million research articles from various disciplines (Figure 2). The Docear desktop software synchronizes all mind maps of a user every couple of minutes with Docear's server. When a user requests recommendations, Docear sends the request to Docear's Digital Library. This creates a user model and returns ten recommendations which are all accessible in full-text. Some users have reservations against having their mind maps transferred to Docear's server and having them analyzed. Therefore, Docear allows users to turn recommendation off which about 2/3 of users do (when users turn recommendations off, they still can use Docear to manage their literature).

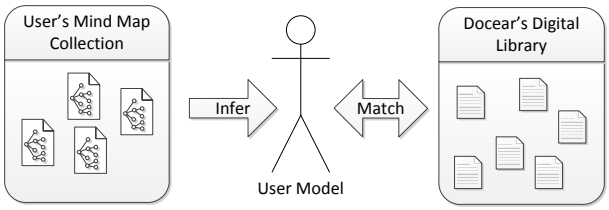

Figure 2: Basic recommendation model

Currently, we are experimenting with different variations of content based filtering. For instance, we are varying the number of mind maps and nodes that are analyzed for creating a user model (e.g. all mind maps the user created vs. only the last edited mind map), the user model size and the type of TF-IDF being used. Based on an evaluation with 938 users and 32,790 delivered recommendations, the average click through rate (CTR) is 6.03%. However, for specific variations such as when stop-words are removed, TF-IDF is applied instead of term frequency only, *and* only user requested recommendations are analyzed, click through rates increases to 10.31% (Figure 4), and in some specific scenarios even more. In comparison, a typical click through rate in advertisement is 0.5% [6]. Research paper recommender systems typically achieve CTRs of around 8% [7]. Further details on the recommendation algorithms will be published in a forthcoming paper.

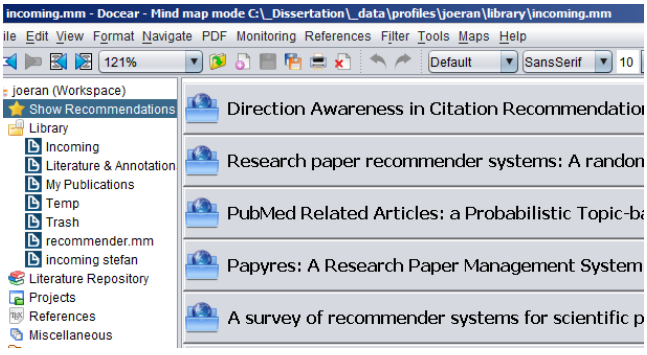

Figure 3: Recommendations in Docear (screenshot)

4. OUTLOOK

Although Docear's recommender system is in a rather early development stage, it already performs quite well. We are confident that the performance can still be improved. For instance, as soon as user numbers of Docear rises, collaborative filtering might become an interesting addition to content based filtering (right now, there are too few similar users to perform CF). An extension of Docear's digital library should also improve the recommender's performance. With 1.8 million articles available, there are only few relevant articles to choose from for many researchers. When more papers are added to the digital library, there will be more potentially relevant recommendations. Finally, we plan on extending the recommender to not only recommend research papers but other items relevant to researchers. For instance, journals and conferences, research grants, and for students we would like to recommend university programs.

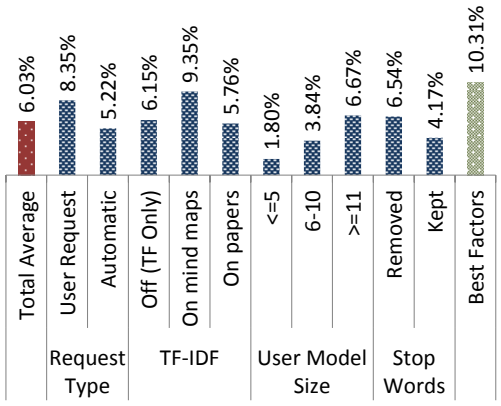

Figure 4: Results of content based filtering variations

5. REFERENCES

[1] Agarwal, N., Haque, E., Liu, H. and Parsons, L. 2006. A subspace clustering framework for research group collaboration. *International Journal of Information Technology and Web Engineering*. 1, (2006), 35–58.

[2] Beel, J., Gipp, B., Langer, S. and Genzmehr, M. 2011. Docear: An Academic Literature Suite for Searching, Organizing and Creating Academic Literature. *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries* (2011), 465–466.

[3] Bogers, T. and Bosch, A. van den 2008. Recommending scientific articles using citeulike. *Proceedings of the 2008 ACM conference on Recommender systems* (2008), 287–290.

[4] Bollacker, K.D., Lawrence, S. and Giles, C.L. 1998. CiteSeer: An autonomous web agent for automatic retrieval and identification of interesting publications. *Proceedings of the 2nd international conference on Autonomous agents* (1998).

[5] Chandrasekaran, K., Gauch, S., Lakkaraju, P. and Luong, H. 2008. Concept-based document recommendations for citeseer authors. *Adaptive Hypermedia and Adaptive Web-Based Systems* (2008), 83–92.

[6] Manchanda, P., Dubé, J.P., Goh, K.Y. and Chintagunta, P.K. 2006. The effect of banner advertising on internet purchasing. *Journal of Marketing Research*. 43, (2006), 98–108.

[7] Pohl, S., Radlinski, F. and Joachims, T. 2007. Recommending Related Papers Based on Digital Library Access Records. *Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries* (Vancouver, BC, Canada, 2007), 417–418.

[8] Torres, R., McNee, S.M., Abel, M., Konstan, J.A. and Riedl, J. 2004. Enhancing Digital Libraries with Techlens+. *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries* (Tuscon, AZ, USA, 2004), 228–236.

[9] Wang, Y., Zhai, E., Hu, J. and Chen, Z. 2010. Claper: Recommend classical papers to beginners. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery* (2010), 2777–2781.