

# A Hybrid, Multi-Dimensional Recommender for Journal Articles in a Scientific Digital Library

André Vellino  
CISTI, National Research Council  
Ottawa, Ontario K1A 0R6  
andre.vellino@nrc.ca

David Zeber  
Cornell University, Dept. of Statistics  
Ithaca, NY 14853-3801  
dsz5@cornell.edu

## Abstract

*A recommender system for scientific scholarly articles that is both hybrid (content and collaborative filtering based) and multi-dimensional (across metadata categories such as subject hierarchies, journal clusters and keyphrases) can improve scientists' ability to discover new knowledge from a digital library. Providing users with an interface which enables the filtering of recommendations across these multiple dimensions can simultaneously provide explanations for the recommendations and increase the user's control over how the recommender behaves.*

## 1 Introduction

The success of scientific digital libraries depends increasingly on their ability to provide differentiating features and value-added services. Recommender systems, for instance, can help to alleviate users' information overload problems and increase their satisfaction with the information retrieval experience. As Torres et al. [27] point out, the rate of growth in scientific research papers cries out for more and better information retrieval tools besides full-text indexing engines.

In a similar vein, Smeaton and Callan [23] argue that digital libraries, which provide only conventional search and browsing capabilities, will eventually be frustrating to users. Instead, they argue, library portals must provide information proactively and tailor their services to individuals and communities. Avancini et al. [4] claim that digital libraries should not be merely anonymous information resources but rather community-based services which require personalized service offerings such as alerting and group opinion sharing.

Recommender systems for scientific digital libraries that have been the subject of experiments in recent years [13, 18, 23, 27, 29] have used corpora that are primarily in the

field of computer science. However, designing an effective recommender system for journal articles in a broader Scientific, Technical and Medical (STM) digital library poses special challenges and presents unique opportunities.

First, as has been noted with recommender systems for corpora such as CiteSeer and the ACM Digital Library, the problem of data sparsity is chronic [27]. In a STM digital library which includes disciplines such as biology, chemistry and medicine, the ratio of users to items is likely to be even lower and the sparsity problem correspondingly exacerbated.

Second, individual users of STM libraries have very particular information needs, such as seeking answers to highly specific scientific questions [14]. At the same time, a recommender system for a STM digital library has to cater to a wide range of such needs that stem both from the large numbers of scientific specializations and from the substantially different kinds of objectives that scientific researchers have when using a library portal.

There are, on the other hand, opportunities to be exploited by the highly specific nature of articles in a STM library. In addition to the copious amount of metadata available for its items – citation indexes, keyphrases and catalog data, for instance – there are increasing quantities of semantic tags in scientific documents [24] and domain-specific machine-learning tools for content extraction [2].

We believe these characteristics of STM digital collections present an opportunity for creating best-of-breed recommenders that leverage each other's strengths. We propose that the hybrid recommenders such as TechLens+ described in Torres et al. [27] and McNee et al. [18] be extended with multi-dimensional context information in a manner similar to the one described by Adomavicius et al. [1]. Our hypothesis is that the addition of multiple context dimensions to article recommendations will enable an explanation-based user interface for filtering out irrelevant recommendations that offers more control to the user and creates greater user trust.

## 2 Recommender Strategies

Recommending an item such as a scientific journal article to a user is typically done either by clustering similar items according to some characteristic of the item (content-based recommendation) or by profiling the users' behaviour and clustering users according to some measure of similarity among them (collaborative filtering) or by combining the two (hybrid recommenders.)

Content-based recommenders need to measure the similarity of the item with all the other items, where the salient content features of the items to be recommended can be either extracted from the items or obtained from metadata. For text items in a library, this could be the feature vectors obtained from the text, or, for items with no text content (e.g. scanned images), salient features could be provided by metadata such as bibliographic categories, authors, title, abstract, etc.

However, the recommendations generated by content-based recommenders will rarely stray far from the content-clusters of the previously rated items. One approach used to overcome the overspecialization of recommendations [30] is either to introduce randomness in the recommendation and to filter out items that are *too* similar or to complement them with collaborative filtering systems, which provide a source of naturally occurring serendipity from user behaviour.

### 2.1 Collaborative Filtering

Collaborative filtering is a method for predicting user preferences and interests based on the collective data of a user community's past usage behaviour. That such a technique can be successful at making predictions rests on the assumption that people who exhibited similar behaviour in the past will tend to exhibit similar behaviour in the future. To recommend an item using collaborative filtering, items must have preference ratings, obtained either explicitly from the user or implicitly from an analysis of usage patterns (clickstream, downloads, etc.) [19] or from citation data [17].

Collaborative filtering techniques come in two main flavours: memory-based and model-based [7]. Memory-based algorithms use all the data collected from all users to make individual predictions, whereas model-based algorithms first construct a statistical model of the users and then use that model to make predictions. Thus memory-based algorithms are generally less efficient and more resource intensive whereas model-based algorithms are generally less accurate because of the greater degree of indirection. Nevertheless, computationally efficient model-based methods now compare favourably with memory-based methods [15].

Collaborative filtering is especially useful when the

items to be recommended have few or no content-based features. Webster et al. [29] point out that since many traditional library resources, such as catalogues, contain only metadata about the items in a collection (i.e. there is no full text to index), traditional search techniques are of limited usefulness. In such situations, collaborative filtering can help induce links between library objects for which there are no syntactic clues for relatedness.

### 2.2 Hybrid Systems

Hybrid approaches take various forms, which are neatly summarized by Burke [9] and Adomavicius [1]. One approach uses content-based methods for developing user models and clustering users by a content-based similarity measure in order to make collaborative recommendations. This enables recommendations to be made either by matching the item's content with the user's profile or by using other users' profiles [21]. For instance, experiments with the Recommendz system [11] have shown that usage data may be usefully combined with full-text information and semantic metadata to provide recommendations. Alternatively, the results of two separate recommenders may be either averaged or given a fair vote depending on the context.

Hybrids between item-based and user-based collaborative filtering systems also exist. Wang et al. [28] describe item-item, user-item, and user-user collaborative filtering in combination with content-based methods both to cluster items and to cluster users. These experiments show that hybrid methods go some way toward alleviating the data sparsity problem and also provide higher quality recommendations.

### 2.3 Multi-Dimensional Recommendations

Recommender systems can be extended to produce recommendations across a broader range of dimensions than simply user and item. For example, Adomavicius et al. [1] consider the time, place and movie-viewing companion as categories with which to augment a recommender like Movielens with contextual dimensions. With the corresponding interface, this enables the user to navigate the space of possible recommendations, much in the same way as OLAP applications allow users to navigate through multi-dimensional data-cubes.

In the domain of music recommendation, Anderson et al. [3] describe a recommender mediated by a RuleML-based engine which filters recommendations that best match user queries based on the dimensions of overall impression, lyrics, music, originality, and quality of performance. By analogy, if there is metadata that can be associated with either the item (e.g. bibliographic metadata such as subject hierarchies, journal categories and semantically related

keyphrases) or with the user (e.g. demographic data or interest profiles) or if collections of items or users can be classified into multiple classes (e.g. using an unsupervised categorizer), then recommendations can also be made along these axes as well. Furthermore, these dimensions could also be used for explaining recommendations.

The user model for the multi-dimensional hybrid system that we envisage for our experiment is similar to the one described by Symeonidis [25]. User models will be defined by combinations of explicitly specified interests and competencies, feature-vectors implicitly obtained from the text-content of retrieved or viewed articles as well as the collaborative filtering information obtained from usage data.

As McNee et al. [17] have shown, a citation graph can be used to seed a collaborative filtering recommender for journal articles. For recently published papers which don't yet have many citations, co-downloading data can be used to complement citation information [19]. In our experiment, we will use PageRank [8] to weight the citation-based "ratings" that papers give one another. One experiment we intend to perform is to determine whether the core citation paths [6, 10] would be an effective way to weight the text content of the citations included in the content-based user profile.

Yet another source of information from which to draw when generating recommendations from a hybrid recommender are the search queries and the clickstream [26] to create user profiles for clustering but also to help rank recommendation results. Moreover, queries themselves may be considered as items for the recommender to present to the user [5].

### 3 Explaining Recommendations

A critical element for the user acceptance of recommenders is the trust that the user develops in the systems' ability to reliably predict items of interest. As Sinha and Swearingen [22] have shown, one way to develop this trust is to offer transparent explanations for the recommenders' behaviour.

Herlocker et al. [12] offer a model for explanations of recommendations based on the user's conceptual model of the recommendation process. They show that providing explanations for a collaborative filtering system's recommendations – e.g. a histogram that maps the ratings for an item by the user's "neighbours" – can improve its acceptance. McCarthy et al. [16] go further and show how explanations can help users understand the remaining recommendation opportunities if the current recommendation doesn't match their interest.

Studies that explored several different explanation interfaces [20] show that breaking down recommendation sets into labeled clusters that categorize recommendations by

combinations of reasons is considerably more effective than one that simply lists all the criteria for recommending items.

For a recommender of scientific research articles which uses a hybrid strategy similar to TechLens+ [27], a multi-dimensional explanation-based interface could not only improve user acceptance and confidence in the recommendations but also provide an interface for guiding the recommender's criteria in selecting item or user neighbourhoods.

With the help of metadata, such as subject catalogue information and article keyphrases, explanations for recommendations or clusters of recommendations could be organized in multiple dimensions to satisfy information seeking needs that complement the analytical model described in [14]. Offering such explanations to an exacting user (such as a scientific researcher) could go some way toward mitigating the "looking stupid" effect of poor recommendations observed by McNee [18].

### 4 Future Work

The hypothesis of this research is that the extension of a TechLens-like recommender for scientific research articles with multi-dimensional explanation and navigation features will help both to enhance the user's trust in the recommender and to improve the quality of recommendations.

If the additional dimensions are derived from the article metadata, one challenge will be to devise meaningful scores for items or classes of items. Another challenge will be to measure the incremental changes in user benefit that result from these specific extensions to the recommender. We intend to obtain this data primarily from query and clickstream logs. We will also use questionnaires that assess the added value of both explanatory and multi-dimensional filters for the recommender.

To perform our experiments we will extend the Taste recommender system<sup>1</sup> with a multi-dimensional hybrid recommender that will be integrated into an experimental digital library's search and browse interface to a corpus from the NRC Research Press<sup>2</sup>. The PageRank-mediated citation-data from this corpus will serve to seed the recommender which will be the subject of experimentation with scientific scholars on CISTI Lab<sup>3</sup>.

The measure of success for a recommender of scholarly articles is, after all, whether it can help scientists to advance knowledge.

### References

- [1] G. Adomavicius, R. Sankaranarayanan, S. Sen, and A. Tuzhilin. Incorporating Contextual Information in Rec-

<sup>1</sup><http://taste.sourceforge.net>

<sup>2</sup><http://pubs.nrc-cnrc.gc.ca>

<sup>3</sup><http://lab.cisti-icist.nrc-cnrc.gc.ca>

- ommender Systems Using a Multidimensional Approach. *ACM Trans. Inf. Syst.*, 23(1):103–145, 2005.
- [2] S. Ananiadou and E. John McNaught. *Text Mining for Biology and Biomedicine*. Artech House, Boston, 2006.
- [3] M. Anderson, M. Ball, H. Boley, S. Greene, N. Howse, D. Lemire, and S. McGrath. RACOFI: a Rule-Appling Collaborative Filtering System. In *Proceedings of COLA'03*. IEEE/WIC, October 2003.
- [4] H. Avancini, L. Candela, and U. Straccia. Recommenders In A Personalized, Collaborative Digital Library Environment. *Journal of Intelligent Information Systems*, 27(1), 2007.
- [5] R. Baeza-Yates, C. Hurtado, and M. Mendoza. Query Recommendation Using Query Logs in Search Engines. *International Workshop on Clustering Information over the Web (ClustWeb, in conjunction with EDBT)*, Crete, Greece, March (to flapper in LNCS), 2004.
- [6] V. Batagelj. Efficient Algorithms for Citation Network Analysis. *Arxiv preprint cs.DL/0309023*, 2003.
- [7] J. Breese, D. Heckerman, C. Kadie, et al. Empirical Analysis of Predictive Algorithms for Collaborative Filtering. *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, 461, 1998.
- [8] S. Brin and L. Page. The anatomy of a large-scale hyper-textual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
- [9] R. Burke. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, 2002.
- [10] B. Dumouchel and J. Demaine. Knowledge Discovery in the Digital Library: Access Tools for Mining Science. *Information Services and Use*, 26(1):39–44, 2006.
- [11] M. Garden and G. Dudek. Mixed Collaborative and Content-Based Filtering with User-Contributed Semantic Features. In *Proceedings of the 21st AAAI National Conference on Artificial Intelligence (AAAI'06)*, Boston, Massachusetts, July 2006.
- [12] J. L. Herlocker, J. A. Konstan, and J. Riedl. Explaining Collaborative Filtering Recommendations. In *CSCW '00: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 241–250, New York, NY, USA, 2000. ACM Press.
- [13] S. Hwang. A Prototype WWW Literature Recommendation System for Digital Libraries San-Yih Hwang, Wen-Chiang Hsiung, Wan-Shiou Yang The Authors. *Online Information Review*, 27(3):169–182, 2003.
- [14] J. Konstan, S. McNee, C.-N. Ziegler, R. Torres, N. Kapoor, and J. Riedl. Lessons on Applying Automated Recommender Systems to Information-Seeking Tasks. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, Boston, MA, USA, July 2006.
- [15] D. Lemire and A. Maclachlan. Slope One Predictors for Online Rating-Based Collaborative Filtering. In *Proceedings of SIAM Data Mining (SDM'05)*, 2005.
- [16] K. McCarthy, J. Reilly, L. McGinty, and B. Smyth. Thinking Positively-Explanatory Feedback for Conversational Recommender Systems. *Proceedings of the European Conference on Case-Based Reasoning (ECCBR-04) Explanation Workshop*, pages 115–124, 2004.
- [17] S. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. Lam, J. Konstan, and J. Riedl. On the Recommending of Citations for Research Papers. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative work*, pages 116–125, 2002.
- [18] S. M. McNee, N. Kapoor, and J. A. Konstan. Don't Look Stupid: Avoiding Pitfalls when Recommending Research Papers. In *CSCW '06: Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, pages 171–180, New York, NY, USA, 2006. ACM Press.
- [19] S. Pohl, F. Radlinski, and T. Joachims. Recommending Related Papers Based on Digital Library Access Records. *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007, Vancouver, B.C., Canada, June 18-23, 2007*.
- [20] P. Pu and L. Chen. Trust Building with Explanation Interfaces. In *IUI '06: Proceedings of the 11th International Conference On Intelligent User Interfaces*, pages 93–100, New York, NY, USA, 2006. ACM Press.
- [21] C. Shahabi, F. Banaei-Kashani, Y. Chen, and D. McLeod. Yoda: An Accurate and Scalable Web-based Recommendation System. *Sixth International Conference on Cooperative Information Systems (CoopIS 2001)*, Trento, Italy, September, 2001.
- [22] R. Sinha and K. Swearingen. The Role of Transparency in Recommender Systems. *Conference on Human Factors in Computing Systems*, pages 830–831, 2002.
- [23] A. F. Smeaton and J. Callan. Personalisation and Recommender Systems in Digital Libraries. *International Journal on Digital Libraries*, V5(4):299–308, 2005.
- [24] M. Smith. Scientific Research Communication: The Promise and Current Realities of Enhanced Publications. [http://www.spatial.maine.edu/icfs/Smith white paper 2006.pdf](http://www.spatial.maine.edu/icfs/Smith%20white%20paper%202006.pdf), 2006.
- [25] M. Y. Symeonidis P., Nanopoulos A. Feature-weighted user model for recommender systems. In *Proceedings of the UM 2007 conference*, pages 97–107, Corfu, 2007.
- [26] Thorsten Joachims and Laura Granka and Bing Pan and Helene Hembrooke and Filip Radlinski and Geri Gay. Evaluating The Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web search. *ACM Trans. Inf. Syst.*, 25(2):7, 2007.
- [27] R. Torres, S. McNee, M. Abel, J. Konstan, and J. Riedl. Enhancing Digital Libraries with TechLens+. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, pages 228–236, 2004.
- [28] J. Wang, A. de Vries, and M. Reinders. Unifying User-Based and Item-Based Collaborative Filtering Approaches by Similarity Fusion. *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 501–508, 2006.
- [29] J. Webster, S. Jung, and J. Herlocker. Collaborative Filtering: a New Approach to Searching Digital Libraries. *New Review of Information Networking*, 10(2):177–191, 2004.
- [30] Y. Zhang, J. Callan, and T. Minka. Novelty and Redundancy Detection in Adaptive Filtering. In *SIGIR '02: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 81–88, New York, NY, USA, 2002. ACM Press.