Recommending Research Articles Using Citation Data

André Vellino
(School of Information Studies
University of Ottawa
avellino@uottawa.ca)

Abstract: This study compares some of the characteristic differences among recommendations generated by a citation-based recommender and a user-based recommender for research articles. As with other application domains, the application of collaborative filtering methods for recommending items in a digital library suffers from a sparsity of usage data. One method for addressing this sparsity problem is to employ an article's references as a proxy for co-download information. Another method is to harvest large quantities of usage data from distributed OpenURL web logs. These different data sources produce significantly different recommendations, both at the individual and at the aggregate level. The experiments randomly selected about 9,000 articles from among 6.6 million articles as starting points for generating recommendations. Among the characteristics of the resulting recommendations are the semantic distance between the seed articles and the recommended ones, the coverage of the recommendations and the spread in publication dates between recommendations and the seed. A parallel experiment on a subset of these articles aimed to test the hypothesis that generating PageRank values for articles might be an effective method for weighting references to improve citation-based recommendations.

Key Words: recommender system, digital library, collaborative filtering, citation analysis
Category: H.3.3, H.3.7

1 Introduction

Digital libraries are an important domain for the application of recommender systems. The rapid growth of scholarly research, particularly with the advent of Open Access publishers and an increasing long tail of rarely cited articles, puts an onus on intelligent discovery tools to suggest literature that is not necessarily key-word related to the items that the user has already found with search and yet remains topically relevant.

On the surface, research articles bear a sufficient similarity to other kinds of items, such as books, DVDs and music that the problem of recommending new items of possible interest to users should be akin to if not the same as it is in e-commerce portals. Indeed a number of digital library recommenders have been deployed and studied since the year 2000, many of them employing either different kinds of data sources or a variety of combinations from which to generate recommendations: users' behaviour data from download logs, text-content in the articles, user-bookmarks and the network of article citations. These data can all be used to cluster users according to some measure of similarity: user-user or

item-item similarity based on usage patterns or content-based similarity or some hybrid of the two [Burke, 2002].

Collaborative filtering [Su and Khoshgoftaar, 2009] is a method for predicting user preferences and interests based on the collective data of a user community's past usage behaviour. To recommend an item using collaborative filtering, items must have some kind of "preference ratings", obtained either explicitly from the user or implicitly from an analysis of usage patterns (clickstream, downloads, etc.) [Pohl et al., 2007] or from citation data [McNee et al., 2002].

The effectiveness of a recommender depends as much on the data sources and on the specific algorithm used for measuring similarity as it does on the end-user's task [Herlocker et al., 2004]. In particular, it is known that user-based collaborative filtering does not produce useful recommendations if the usage data is sparse [Su and Khoshgoftaar, 2009], a particularly acute problem for items in digital libraries. However, for sparse matrices where the number of articles dominates the number of users it is known that user-based collaborative filtering recommenders have better accuracy than item-based ones [Bogers and van den Bosch, 2008].

In the context of digital library recommenders, two main strategies have been employed to address the data sparsity problem (i) harvesting usage data from a large number of distributed usage logs and (ii) taking advantage of the citation network of articles as a proxy for preferences. The aim of the present study, briefly reported on in [Vellino, 2010], is to empirically compare the properties of two specific recommenders that employ these two strategies: the 'bX' recommender by ExLibris that applies collaborative filtering to distributed usage data, and the Sarkanto recommender that applies collaborative filtering to article citations. The present study extends these initial results by sampling a larger collection of documents, a greater number of recommendations, and includes, in addition, an analysis of publication date distributions. A separate experiment also shows that the application of PageRank to the citation matrix does nothing improve recommendation accuracy.

The rest of the paper is organized as follows. Section 2 provides a brief survey of recommendation methods in digital libraries and section 3 presents the motivation and backround for this study. Section 4 discusses the methodology and results and section 5 describes an attempt to enhance the accuracy of citation-based recommendations with the use of PageRank.Section 6 concludes the paper and discusses further research directions.

2 Related Work

Recommending items in a digital library has been attempted using all three of the principal methods used in recommender systems: user-based collaborative filtering, content-based filtering and hybrids of the two.

Collaborative filtering is especially useful when the items to be recommended have few or no content-based features. Webster et al. [Webster et al., 2004] 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.

BibTip [Mönnich and Spiering, 2008, Franke et al., 2008], developed at Karlsruhe University, is an instance of a user-based collaborative filtering recommender for items in a library catalog that employs OPAC usage data and contains neither citation information nor full-text content. Users' session behaviour are mined to extract user-item preferences and in this respect, most closely resembles conventional commercial product recommenders, except for the absence of explicit ratings.

The recommender system that is deployed on the bibliographic social reference manager site CiteULike [Bogers and van den Bosch, 2008] operates in a similar fashion to BibTip except that the data source is the site users' collections of article bookmarks. Like OPAC usage data, bookmarked articles are implicit data and do not indicate the users' specific ratings for the articles.

In contrast with collaborative filtering, content-based recommenders need to measure item-item similarity according to some feature extracted from the items or from metadata [Balabanović and Shoham, 1997]. 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.

One well-known feature of content-based recommendations is that they rarely stray semantically from the content-clusters of previously rated items. One approach used to overcome this overspecialization of recommendations 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 [Zhang et al., 2002].

The open-source repository platform DSpace [Elliott and Erickson, 2008] is an instance of a content-based recommender that generates recommendations based on a user-selected set of examples that circumscribe the "research context" of the user. Recommendations from DSpace are generated by applying a content similarity measure on the metadata about the articles using a Jaccard similarity coefficient.

The first collaborative filtering recommender designed specifically for research papers was developed by the university of Minnesota's GroupLens team and later deployed by the University of Minnesota Library [McNee et al., 2002].

This system employed the strategy of using citations as a substitute for item ratings to address the problem of usage data sparsity. This is the same strategy that was implemented in the Sarkanto recommender studied in this paper. Its successor, TechLens+ [Torres et al., 2004] adopted a hybrid (citation-based and content-based) approach to improve on its predecessors' precision. The authors also studied users perceptions of paper recommendations generated by such a hybrid recommender.

Hybrid approaches take various forms, which are clearly summarized by [Burke, 2002] and [Adomavicius et al., 2005]. 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 [Shahabi et al., 2001]. For instance, experiments with the Recommendz system [Garden and Dudek, 2006] has 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. [Wang et al., 2006] describes 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.

3 Motivation and Background

The motivation for the initial study mentioned above was to help determine which of two collaborative filtering methods—usage-based or citation-based—would be better suited for deployment at the Canada Institute for Scientific and Technical Information (CISTI), Canada's national science library. The objective was to compare and contrast some of the recommendation result-sets and behavioural characteristics of 'bX' and Sarkanto with respect to prediction coverage (the percentage of the items for which the system is able to generate a recommendation) and serendipity (the extent to which the recommended items is unexpected to users).

However, evaluating serendipity as a function of user expectation would have required a user study beyond the scope of this research. I chose instead to consider that semantic distance is an approximation for serendipity.

3.1 Article Collection

The articles that formed the basis for both the previous and the present experiments were extracted from a collection of approximately 6.6M articles held by CISTI. The majority of these articles are in the fields of Science, Technology and Medicine (STM) and date from between 1995 and 2009. They were published in approximately 2400 journals and conference proceedings in a variety of fields including Medicine (671 journals or 16% of the total), Agricultural and Biological Sciences (9%), Engineering and Technology (8%), Biochemistry, Genetics, and Molecular Biology (6%), Chemistry (5%) and Computer Science (5%).

The full article collection is rich in reference metadata but only 1.8M articles among the 6.6M contain references to articles inside the collection. In that subset of 1.8M articles, the average number of references to other articles in the 6.6M collection was 6 per article. Although the full-text of these articles was available for analysis, this information was not used because many of the recommendations generated by 'bX' were not also articles that existed in this collection and hence could not be text-mined without an additional harvesting step.

3.2 Data Sparsity

If the sparsity of a user-item matrix is measured as the number of links between users and items (either ratings, or the occurrence of a download or citation) divided by the total number of possible links between users and items, then the sparsity of data used for typical collaborative filtering tasks, such as recommending movies with the Netflix data set [Bennett et al., 2007], is about 1%.

In a digital library these ratios are orders of magnitude smaller than in recommenders for commercial merchandise—on the order of tens of thousands of users per month for a collection of tens of millions of items. For example, the sparsity of the matrix of (scholars) to items (articles) in a substantial bibliographic portal such as Mendeley is $2.66x10^{-5}$, almost three orders of magnitude smaller than Netflix. In the data provided by Mendeley in response to the DataTEL challenge [Jack et al., 2010], out of the 3,652,286 unique articles, 3,055,546 (83.6%) were referenced by only 1 user and 378,114 were referenced by only 2 users. Less than 6% of the articles referenced were referenced by 3 or more users and the most frequently referenced article was referenced 19,450 times.

These extremely small user-item ratios are clearly insufficient for collaborative filtering to produce reliable recommendations. Hence the need for supplementary data, either distributed usage logs or citation data.

3.3 'bX'

'bX' is a commercially available web service from ExLibris [ExLibris, 2009] that recommends research articles using data obtained from OpenURL logs of users'

co-downloads. The recommender's design is based on research on the large scale usage of scholarly resources that permits the harvesting of inter-institutional aggregation of log data [Bollen and van de Sompel, 2006]. The quantity of data obtained with this method is sufficiently large to make it possible to apply collaborative filtering effectively: as OpenURL resolver logs grow over time, 'bX' recommendations reflect users' aggregate behaviour with increasing precision and accuracy [Herlocker et al., 2004].

3.4 Sarkanto

Another strategy for addressing the data sparsity problem, first used by TechLens+ [Torres et al., 2004], and re-implemented in Sarkanto is to take advantage of bibliographic citations in the articles as a proxy for user ratings. The idea is to consider an article as a "user" and the articles that it cites to be the article's "preferences" (or boolean ratings). Sarkanto is a user-based collaborative filtering recommender that implements k-nearest neighbour and cosine correlation in the Taste framework (now Mahout [Mahout, 2009]).

Since the document collection used for this experiment was static, the list of recommendations generated for each article (i.e. "user") was pre-computed rather than dynamically computed. However, recent advances in sparse matrix ordering and partitioning [Küçüktunç et al., 2013] make it possible to generate real-time recommendations efficiently from large, sparse citation networks.

An obvious limitation of this approach is that bibliographic references, while an indicator of relevance, are not necessarily an indication of *favourable* relevance in the mind of the author. Findings in [Case and Higgins, 2000] showed that authors were motivated to cite a work for a variety of reasons, including the fact that citing it might promote the authority of their own work or that the cited work deserved criticism. As early as 1965 Garfield identified fifteen such reasons for citing a work [Garfield, 1965]. Today, the Citation Typing Ontology [Peroni and Shotton, 2012] provides a rich machine-readable taxonomy for the characterization of bibliographic citations with almost ninety semantic relations such as *agrees with*, *corrects*, *supports*, and *uses conclusions from*.

As a first approximation to ranking the relative importance of referencer to the citing article, I also conducted an experiment that assigns weights according to the PageRank valuse of the references as calculated from the entire citation graph. The method and results of that experiment are discussed in section 5.

4 Methodology and Quality Comparison of Sarkanto and 'bX'

A typical method for assessing the effectiveness of a recommender algorithm is leave-one-out cross-validation [Herlocker et al., 2004]. For a recommender of scholarly articles that uses citation-based "ratings" such as Sarkanto, a sample of

test articles is selected and, for each article in that set, one reference is removed and the recommender is tested for whether it predicts the removed reference. If the removed reference ranks highest in the list of recommendations, it belongs to the Top-1 recommendations, if it ranks in the first five recommendations it belongs to the Top-5, etc.

Results from a previous study that compared citation-based recommendation with a method based on modeling cognitive memory (a model of human memory performance for cognitive tasks) show that the accuracy rate of Sarkanto for Top-10 predictions, using leave-one-out cross validation on a subset of the collection used in the current study, is close to 20% [Rutledge-Taylor et al., 2008]. Given the limitations of the significantly reduced citation data in that study, this is a respectable score.

However, the measures often used to evaluate the efficacy of algorithms such as Top-N or Mean Absolute Error (MAE) [Herlocker et al., 2004] are not applicable in this situation, principally because the absence of OpenURL log data does not enable the tester to determine which are the gold-standard recommendations. Hence, a meaningful comparative evaluation of the quality of recommendations generated by each recommender could only be provided by a human-subject expert that inspects the results and assesses the relevance of each recommendation [Gunawardana and Shani, 2009].

Therefore, instead of using any of the above measures, I compared other characteristics of the recommendations generated by each of these strategies: coverage, diversity, complementarity, and publication date.

4.1 A Priori Comparison

The Sarkanto and 'bX' recommenders each have *a priori* strengths and weaknesses in their respective approaches. For instance, while 'bX' can take advantage of a voluminous amount of globally distributed usage data, this data may not reflect, even in the aggregate, the interests of specialists in any given field. Usage data from OpenURL logs is indiscriminate between expert researchers and undergraduate university students. In addition, a dependence on usage information makes such a recommender unable to address the recommendation needs of users interested in the end of long tail of sparsely researched areas. One consequence of this is that the publication dates of 'bX' recommendations should typically be skewed towards the present.

On the other hand a recommender that uses bibliographic citations instead of usage data suffers from other limitations. One is that citations are static and citation-based recommendations don't reflect current usage trends. In addition, there is a lag period of about 2 years between the publication of an article for which there begins to develop co-downloading information and it being cited in other publications [Pohl et al., 2007]. Hence one would expect at least that much

of a difference in the publication dates of recommendations. Finally, as noted earlier, an article's references are not necessarily a signal of endorsement by the author, although there is no reason to believe that co-download information is any more of a signal of endorsement.

4.2 Experimental Comparison

This experiment is very similar to the experiments reported in [Vellino, 2010]. The experiment compares the semantic diversity of recommendations, the number of recommended articles the extent to which recommendations from these different sources overlap and the publication-date characteristics from each source. Figure 1 shows a sample of recommendations generated by both Sarkanto and 'bX'.

The semantic distance between the seed article and the recommendations generated by both 'bX' and Sarkanto was measured by examining the "journal diversity" among the recommended articles relative to the original article. It would have been preferable to use the semantic distance between the seed article and each of the recommended articles based on their full-text content. However, only the full-text in our corpus was available for mining and only some of the full-text was available for only some of the recommended articles generated by Sarkanto. Even fewer were available in full-text for those generated by 'bX'.

Hence I chose to use an aggregate measure from the data that underlies the semantic map produced by Newton et. al. [Newton et al., 2009] and from which the semantic distances among 2365 journals are calculated. For the purposes of this map, a 'journal' is considered to be the concatenation of the full-text of all the available articles in that journal. Each journal is represented by a coloured dot and the colours on this map correspond to the publisher-generated main subject category to which the journal belongs. This map is reproduced in Figure 2.

The semantic distances between each journal in this map were computed using Widdows' Semantic Vectors method [Widdows and Ferraro, 2008] on the full-text of a collection of 5.7 million articles, a curated subset of the collection described above. The average distance between a randomly selected pair of such journals is 0.79 where 1.0 is the distance between a journal and itself and 0.0 is the distance between two journals that have no terms in common. As a whole, this collection is relatively homogeneous in its subject matter. Thus, in this collection, the two most similar journals are at a distance of 0.998 and the two most dissimilar are at a distance of 0.2724.

From the collection of 1.8 million test articles that contain references, 9453 articles were randomly selected as seed articles for generating recommendations. For the citation-based recommender a seed article amounts to a collection of references that substitutes for the "user profile" from which a recommendation

The recommendations for this article:

- Snow avalanche hazard modelling of large areas using shallow water numerical methods and GIS
  U. Gruber, P. Bartelt (2007) *Environmental Modelling and Software* **22:** 1472.

Citation Based Recommendations	User Based Recommendations
1) Computing extreme avalanches (2004) *Cold Regions Science and Technology* **39** 161-180 2) Error in a USGS 30-meter digital elevation model and its impact on terrain modeling (2000) *Journal of Hydrology* **233** 154-173 3) Dry Granular Flow Modelling Including Erosion and Deposition (2003) *Surveys in Geophysics* **24** 569-585 4) An alternative form for the statistical distribution of extreme avalanche runout distances (2005) *Cold Regions Science and Technology* **42** 185-193 5) Calculating Internal Avalanche Velocities from Correlation with Error Analysis (2003) *Surveys in Geophysics* **24** 499-524 6) Experimental devices to determine snow avalanche basal friction and velocity profiles (2004) *Cold Regions Science and Technology* **38** 17-30 7) Optimization the basis of code-making and reliability verification (2000) *Structural Safety* **22** 27-60 8) On full-scale avalanche measurements at the Ryggfonn test site, Norway (2007) *Cold Regions Science and Technology* **49** 39-53 9) Dense snow avalanche modeling: flow, erosion, deposition and obstacle effects (2004) *Cold Regions Science and Technology* **39** 193-204 10) Error propagation of DEM-based surface derivatives (2005) *Computers and Geosciences* **31** 1015-1027	1) Snow avalanche hazard modelling of large areas using shallow water numerical methods and GIS Gruber U., (2007-10-01) *Environmental modelling & software* **22** 1472-1481 2) Characteristics and mitigation of the snow avalanche hazard in Kaghan Valley, Pakistan Himalaya De Scally F., (1994) *Natural hazards* **9** 197-213 3) Cartographic modeling of snow avalanche path location within Glacier National Park, Montana Walsh S., (1990-05-01) *Photogrammetric engineering and remote sensing* **56** 615-621 4) The influence of tree and branch fracture, overturning and debris entrainment on snow avalanche flow Bartelt P., (2001) *Annals of glaciology* **32** 209-216 5) Effects of release conditions uncertainty on avalanche hazard mapping Barbolini M., (2002) *Natural hazards* **25** 225-244 6) Avalanche climatology of the western United States, with an emphasis on Alta, Utah Mock C., (1992) *The Professional geographer* **44** 307-318 7) Altered streamflow and sediment entrainment in the Gunnison gorge Elliott JG., (1997) *Journal of the American Water Resources Association* **33** 1041-1054 8) On probability analysis in snow avalanche hazard zoning Harbitz C., (2001) *Annals of Glaciology* **32** 290-298 9) Mammoth Mountain, California Weaver T., (2008) *Skiing* **61** 66- 10) Regionalization and reconstruction of snow water equivalent in the upper Colorado River basin Timilsena J., (2008) *Journal of hydrology* **352** 94-106

Figure 1: Sample side-by side recommendation of articles using citation and usage data

is generated. For the user-based recommender 'bX', the metadata from the seed article (title, authors, publication date, etc.) are used to construct an OpenURL request a list of recommendations from the 'bX' recommender web service. The distribution of subject areas and publication venues for the sampled seed articles was roughly the same as that of the collection as a whole.

For each seed article selected at random from the text collection, we compared the recommendations generated by 'bX' and Sarkanto and counted:

1. the number of recommended articles
2. the semantic distance between the seed article and the recommended articles
3. the number of times that 'bX' and Sarkanto both recommended articles from a given seed article

Figure 2: Journal Semantic Distance Map in Newton, Callahan and Dumontier [Newton et al., 2009]. Different colours indicate journals in different fields.

4. for each instance where both 'bX' and Sarkanto produced a set of recommendations from the same seed, which one of 'bX' or Sarkanto had greater journal diversity
5. the average distance, in years, between the publication date of the seed article and the publication dates of the recommended articles.

Note that the variety of journals that can be recommended in the 'bX' system is significantly greater than the range available in Sarkanto, given the limited number of publishers (about 50) in the article collection used by Sarkanto.

4.3 Results

Out of the 9453 test runs, 2873 generated one or more recommendations using 'bX' compared with 2263 for Sarkanto (i.e. the recommendation coverage was 30% vs. 24% in Table 1). Sarkanto recommended an average of 9.7 articles per seed-article vs. 8.4 for 'bX', which had been configured to generate as many as possible.

The number of seed articles that produced recommendations with both Sarkanto and 'bX' was only 12% meaning that most of the time either one or the other recommender would produce a result, indicating a high degree of complementarity. Furthermore, within this 12% of articles for which both recommenders produced

	Seeds	Productive Seeds	Sarkanto	'bX'	Both
Number	9453	3998	2263	2873	1138
Percentage	100 %	42 %	24 %	30 %	12 %

Table 1: Summary Table of Citation and Usage-based coverage

a result, none of the recommended articles were the same, as illustrated in Figure 1. The results of this part of the experiment do not differ significantly from those reported in [Vellino, 2010].

The results for semantic diversity, however, differ significantly and surprisingly from the earlier study. In this experiment both 'bX' and Sarkanto yielded about the same semantic distance measure between the seed article and the recommended articles, namely 0.948 (for 'bX') and 0.956 (for Sarkanto). This is a significant discrepancy from previous results that can only be explained by the relatively greater random sample size in this experiment.

It is useful to have a baseline for these similarity measures. Given that the subject matter of the collection as a whole is relatively homogeneous and that a randomly selected article from the collection is likely (by this measure) to have a relatively high degree of similarity to the seed article, we also measured the semantic distance between the same seed articles and articles that were randomly generated from the collection to be 0.80.

The new result in this study centers on publication dates. We measured the average differences between the publication dates of the recommended articles and the seed articles. Both recommenders are able to and generally do recommend articles that were published either before or after the seed article's publication date. There is, however, an inherent bias in OpenURL usage-log data towards more recent content. Thus the average age of Sakanto recommendations was +7.6 years (prior) to the seed article's publication date whereas for 'bX' the average age is −0.6 years, i.e. forward looking, on average. As a baseline reference, for randomly selected articles, the average age distance from the seed article is +6.3 years. Citation based recommenders are therefore *heavily* weighted towards recommending older articles.

5 Enhancing Sarkanto with PageRank

In the absence of any further information about the various reasons for which an article may be cited one might expect that substituting the boolean ("cited" or "not-cited") with the PageRank value of the article, as computed from the citation graph, would improve the Top-N effectiveness of collaborative filtering recommendations.

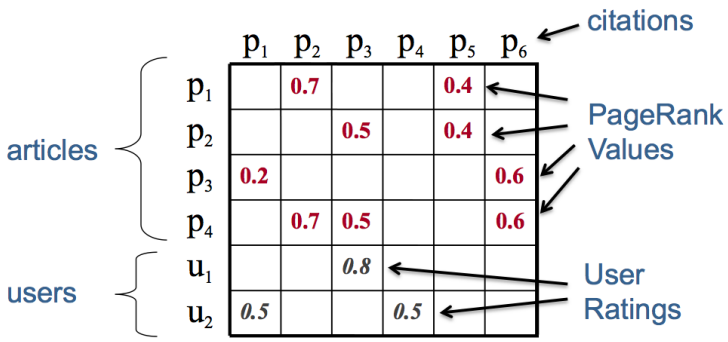

Figure 3: Ratings Matrix populated with PageRank and User Ratings

Previous successes in the use of PageRank as a measure of the "impact" of an article [Ježek and Steinberger, 2008, Ma et al., 2008] suggested that PageRank weights could be a reasonable proxy for the numeric rating that an article might give to an article it references. Furthermore, there exists a PageRank-based method (PaperRank) for predicting co-occurring references in an article [Gori and Pucci, 2006] that is, by itself, an adequate recommender method. This would seem to indicate that combining such a method with a collaborative filtering algorithm could yield superior results.

A Weighted PageRank algorithm [Xin and Ghorbani, 2004] was applied to the graph of 1,461,305 references contained in 369,470 articles in the collection, for an average of 3.9 references (in the collection) per article in the collection. For the purposes of this experiment, co-authorship information was ignored.

Once the PageRank value for each article was computed, this value was assigned to each occurrence of the article in the citation matrix and all occurrences of that article's "rating" have the same value (see Figure 3). Thus every "pseudo-user" (i.e. article) weights every occurrence of any given citation equally. This is clearly not a good analogue for the usual "user rating" notion used in collaborative filtering: two different articles containing the same reference would typically assign a different rating to the reference. However, this property (of assigning equal weights to every cited article) also holds in the case where no PageRank values are used (i.e. where the ratings matrix is boolean). In that respect the two rating methods are on par.

The off-line experiments were modeled after off-line experiments designed for TechLens+ which measured the recommender's effectiveness by computing Top-N recommendations [Torres et al., 2004]. The set of test articles was selected to have a significant number of references and, for each article in that set (the active article) each reference was removed one at a time and the recommender was

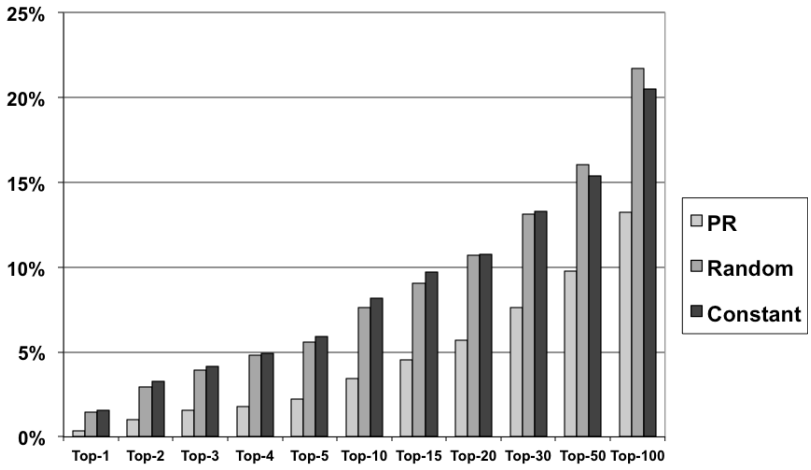

Figure 4: Top-N results for PageRank, Random and No Page Rank

tested for whether it predicts the removed reference. If the removed reference ranks highest in the list of recommendations, it belongs to the Top-1 recommendations, if it ranks in the first five recommendations it belongs to the Top-5, etc. The test set was selected from a self-contained subset of 369,470 articles from the large collection described. These articles contain 1,461,305 references to articles in the large collection.

Our tests differ from those in the TechLens+ study in that we chose to perform leave-one-out evaluations exhaustively on a sample of the articles biased towards those with the most references to items in the bibliographic collections. The equivalent strategy for evaluating a recommender for movies would be to pick the top K users who have the most movie ratings and attempt to predict each rating for all of the ratings in the top K users. We chose this method to increase the likelihood of picking a random article with significantly more than one reference. Leaving one reference out for each of the articles with the most references seemed more likely to produce a recommendation that was correct.

Another difference from the TechLens+ experiments is that recommendations that have a publication date that is later than that of the paper whose list of references are being used as preference ratings were not filtered out. The justification in TechLens+ for this filtering out of future papers is that the recommender should not recommend a paper that did not exist at the time the active paper was published. However, this methodological constraint on the experiment was not applied in the 'bX'-Sarkanto experiments and there was no need to add this constraint in this one either.

Each reference for an article was assigned one of two possible "preference"

values: a fixed constant for all articles and the PageRank value of the article. As a control experiment, the effect of using a random preference value was also measured.

The experimental results of the effect of PageRank are summarized in Figure 4. The results show that the use of PageRank has a markedly negative impact on the quality of recommendations. Using a constant instead of a PageRank value for the rating of a reference improves recommendations by a factor of 400% (for Top-1 recommendations) down to an increase of 55% for Top-100. Moreover, substituting a randomly generated value for the PageRank of each article is roughly equivalent to using a fixed constant or boolean value for all articles (i.e. not using PageRank).

While this result appears counter-intuitive, there is a plausible explanation. The PageRank value of an article reflects its overall status in the network. To the citing article, however, the value of the cited article is clearly much higher than the value attributed by the network as a whole. Thus assigning the PageRank value to the cited article disproportionately devalues the relevance of that article to the citing article. Thus, assigning a network based ranking to an individual article's reference unfairly penalizes that article's contribution to the recommendation process.

6 Discussion and Future Work

Initial experiments with these two recommenders suggested that they produced recommendations with significantly different journal-to-journal semantic diversity, with the citation-based recommender offering greater serendipity. A larger sampling indicates otherwise.

Certainly, both recommenders are topical. Inspection of the side-by-side recommendation lists generated by the 12% of seed-articles that generate results for both 'bX' and Sarkanto clearly shows this. For instance, if a source article was about "avalanche modelling", as in Figure 1, the recommended articles tended to also be about snow or computer modeling. Yet they are also complementary in coverage, if only in the span of date ranges that they each cover.

Using citations as a method for dealing with the recommender system data sparsity problem is not only an alternative to harvesting large amounts of usage data. It also serves to generate different kinds of recommendations than those from usage data. One depends on the domain of authors' relevance judgments and the other on readers' relevance judgments. The user-based method recommends "articles that other users also downloaded" whereas the citation-based one recommends "articles whose citation patterns are similar to this one".

The complementarity in coverage between these two methods suggests that it might be useful to combine them to form a hybrid recommender. There are,

however, several unresolved issues with hybridizing these methods, not least of which is how to compare their rankings. Furthermore, end users of recommenders need to understand the sources of data that are used to generate the recommendations if they are going to trust them. Explanations for such properties as topic diversity (when it arises), and publication date biases would help users choose which kind of recommender data-sources are most relevant for their information retrieval tasks. Thus, it would be preferable for a system that combined usage-based and citation-based recommendations not to hybridize them but to offer them as complementary alternatives with different explanations.

The significant disparity between the publication dates of recommendations generated by the two methods would benefit from further analysis. According to [Hajra and Sen, 2006] the age distribution of references made to a paper obeys a power law decay while the age distribution of references made by a paper has an exponential decay. Such a model of the aging characteristics of the citation network for this collection is needed to explain why citation-based recommendations are so much older than even the average distance between a given article and the rest of the collection.

The use of PageRank to weight the relative importance of a reference in an article has the disadvantage that this same weight applies to all other occurrences of that reference. It does not distinguish between one article's citation of that reference and another's. A more discriminating method for weighting references would be to detect the citation features of the references in the body of the text to determine which were the more important references in the bibliography, features like the number of times a reference is mentioned in the body of the citing paper [Zhu et al., 2013]. Such a method for weighting references, even if it ranked a citation's importance to the article on a 5 point scale, could be combined with a measure of co-citation proximity in the body of the article [Gipp and Beel, 2009] to improve the precision of citation-based recommendations.

7 Acknowledgments

I would like to thank the Canada Institute for Scientific and Technical Information for the research use of their digital collection and to Daniel Lemire, Peter Cashin, Jean Jervis and several anonymous reviewers for helpful comments and references to related work.

References

[Adomavicius et al., 2005] Adomavicius, G., Sankaranarayanan, R., Sen, S., and Tuzhilin, A. (2005). Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Trans. Inf. Syst.*, 23(1):103–145.

[Balabanović and Shoham, 1997] Balabanović, M. and Shoham, Y. (1997). Fab: Content-based, collaborative recommendation. *Commun. ACM*, 40(3):66–72.

[Bennett et al., 2007] Bennett, J., Elkan, C., Liu, B., Smyth, P., and Tikk, D. (2007). KDD Cup and workshop 2007. *SIGKDD Explor. Newsl.*, 9:51–52.

[Bogers and van den Bosch, 2008] Bogers, T. and van den Bosch, A. (2008). Recommending scientific articles using citeulike. In *RecSys '08: Proceedings of the 2008 ACM conference on Recommender systems*, pages 287–290, New York, NY, USA.

[Bollen and van de Sompel, 2006] Bollen, J. and van de Sompel, H. (2006). An architecture for the aggregation and analysis of scholarly usage data. In *Proceedings of the Third International Conference on Digital Information Management*, pages 87–92, New York, NY. JCDL '06, ACM.

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370.

[Case and Higgins, 2000] Case, D. O. and Higgins, G. M. (2000). How can we investigate citation behavior?: a study of reasons for citing literature in communication. *J. Am. Soc. Inf. Sci.*, 51(7):635–645.

[Elliott and Erickson, 2008] Elliott, D. Rutherford, J. and Erickson, J. (2008). A recommender system for the dspace open repository platform. *HP Labs Technical reports*, HPL-2008-21.

[ExLibris, 2009] ExLibris (2009). http://www.exlibrisgroup.com/.

[Franke et al., 2008] Franke, M., Geyer-Schulz, A., and Neumann, A. W. (2008). Recommender services in scientific digital libraries. In *Multimedia Services in Intelligent Environments*, pages 377–417. Springer.

[Garden and Dudek, 2006] Garden, M. and Dudek, G. (2006). 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.

[Garfield, 1965] Garfield, E. (1965). Can citation indexing be automated? In *Statistical association methods for mechanized documentation, Symposium proceedings*, ed. M. E. Stevens et al. (Washington: National Bureau of Standards, Miscellaneous Publication 269, 1965), pages 188–192.

[Gipp and Beel, 2009] Gipp, B. and Beel, J. (2009). Citation proximity analysis (cpa)-a new approach for identifying related work based on co-citation analysis. In *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, volume 2, pages 571–575.

[Gori and Pucci, 2006] Gori, M. and Pucci, A. (2006). Research paper recommender systems: A random-walk based approach. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM Int. Conf. on Web Intelligence*, pages 778–781, Washington, DC, USA. IEEE Computer Society.

[Gunawardana and Shani, 2009] Gunawardana, A. and Shani, G. (2009). A survey of accuracy evaluation metrics of recommendation tasks. *J. Machine Learning Research*, 10:2935–2962.

[Hajra and Sen, 2006] Hajra, K. B. and Sen, P. (2006). Modelling aging characteristics in citation networks. *Physica A: Statistical Mechanics and its Applications*, 368(2):575–582.

[Herlocker et al., 2004] Herlocker, J., Konstan, J., Terveen, L., and Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53.

[Jack et al., 2010] Jack, K., Hammerton, J., Harvey, D., Hoyt, J. J., Reichelt, J., and Henning, V. (2010). Mendeley's reply to the datatel challenge. *Procedia Computer Science*, 1(2):1–3.

[Ježek and Steinberger, 2008] Ježek, K. Fiala, D. and Steinberger, J. (2008). Exploration and evaluation of citation networks. *ELPUB2008. Openness in Digital Publishing: Awareness, Discovery and Access - Proc. of the 12th Int. Conf. on Electronic Publishing*, pages 351–362.

[Küçüktunç et al., 2013] Küçüktunç, O., Kaya, K., Saule, E., and Catalyürek, U. V. (2013). Fast recommendation on bibliographic networks with sparse-matrix ordering and partitioning. *Social Network Analysis and Mining*.

[Ma et al., 2008] Ma, N., Guan, J., and Zhao, Y. (2008). Bringing pagerank to the citation analysis. *Information Processing and Management*, 44(2):800–810.

[Mahout, 2009] Mahout (2009). http://lucene.apache.org/mahout.

[McNee et al., 2002] McNee, S., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S., Konstan, J., and Riedl, J. (2002). On the Recommending of Citations for Research Papers. *Proceedings of the 2002 ACM Conference on Computer Supported Cooperative work*, pages 116–125.

[Mönnich and Spiering, 2008] Mönnich, M. and Spiering, M. (2008). Adding value to the library catalog by implementing a recommendation system. *D-Lib Magazine*, 14(5/6).

[Newton et al., 2009] Newton, G., Callahan, A., and Dumontier, M. (2009). Semantic journal mapping for search visualization in a large scale article digital library. In *Second Workshop on Very Large Digital Libraries, ECDL 2009*, New York, NY.

[Peroni and Shotton, 2012] Peroni, S. and Shotton, D. (2012). FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17(0):33–43.

[Pohl et al., 2007] Pohl, S., Radlinski, F., and Joachims, T. (2007). Recommending related papers based on digital library access records. *ACM/IEEE Joint Conference on Digital Libraries, JCDL 2007*.

[Rutledge-Taylor et al., 2008] Rutledge-Taylor, M., Vellino, A., and West, R. (2008). A Holographic Associative Memory Recommender System. In *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*.

[Shahabi et al., 2001] Shahabi, C., Banaei-Kashani, F., Chen, Y., and McLeod, D. (2001). Yoda: An Accurate and Scalable Web-based Recommendation System. *Sixth International Conference on Cooperative Information Systems (CoopIS 2001), Trento, Italy, September*.

[Su and Khoshgoftaar, 2009] Su, X. and Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009:4.

[Torres et al., 2004] Torres, R., McNee, S., Abel, M., Konstan, J., and Riedl, J. (2004). Enhancing Digital Libraries with TechLens+. *Proceedings of the 2004 Joint ACM/IEEE Conference on Digital Libraries*, pages 228–236.

[Vellino, 2010] Vellino, A. (2010). A comparison between usage-based and citation-based methods for recommending scholarly research articles. In *Proceedings of the 73rd ASIS&T Annual Meeting*, volume 47 of *ASIS&T '10*.

[Wang et al., 2006] Wang, J., de Vries, A., and Reinders, M. (2006). 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.

[Webster et al., 2004] Webster, J., Jung, S., and Herlocker, J. (2004). Collaborative filtering: a new approach to searching digital libraries. *New Review of Information Networking*, 10(2):177–191.

[Widdows and Ferraro, 2008] Widdows, D. and Ferraro, K. (2008). Semantic vectors: A scalable open source package and online technology management application. In *Proceedings of the Sixth International Language Resources and Evaluation, LREC '08*, New York, NY. ECDL 2009, ACM.

[Xin and Ghorbani, 2004] Xin, W. and Ghorbani, A. (2004). Weighted pagerank algorithm. In *CNSR '04: Proceedings of the Second Annual Conference on Communication Networks and Services Research*, pages 305–314, Washington, DC, USA. IEEE Computer Society.

[Zhang et al., 2002] Zhang, Y., Callan, J., and Minka, T. (2002). Novelty and redundancy detection in adaptive filtering. In *Proceedings of the 25th ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 81–88, New York, NY.

[Zhu et al., 2013] Zhu, X., Turney, P., Lemire, D., and Vellino, A. (2013). Measuring academic influence: Not all citations are equal. *unpublished manuscript*.