

# Recommending Journal Articles with PageRank Ratings

André Vellino  
Canada Institute for Scientific  
and Technical Information  
National Research Council  
Ottawa, Canada, K1A 0R6  
andre.vellino@nrc.ca

## ABSTRACT

The TechLens+ strategy for addressing the recommender cold-start problem in a scholarly digital library is to seed the preference matrix with article references. However, this method generates boolean ratings rather than ratings on a numerical scale, as is more typical with recommender systems for commodity products. One strategy for generating a better preference matrix for collaborative filtering recommendations is to compute the PageRank values for the articles in the citation graph of the article collection and to substitute the boolean ratings with PageRank “ratings”. There is a significant amount of prior research which suggests that this strategy should generate better Top-N recommendations. However, the experimental results described in this paper show that PageRank ratings are inferior to both boolean ratings and random (but consistent) ratings.

## Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval, Information Filtering

## General Terms

Algorithms, Experimentation

## Keywords

PageRank, Recommender System, Citations, Digital Library

## 1. INTRODUCTION

Recommending journal articles in a digital library with collaborative filtering is more difficult than recommending other kinds of items (songs, movies, merchandise, etc.) in part because of the greater sparsity of ratings data. Not only is the ratio of users to items in a digital library typically one or two orders of magnitude smaller than in recommenders for commercial merchandise – on the order of tens of thousands

of users per month on a collection of tens of millions of items – but also the average number of item ratings per user in a digital library is likely to be much smaller, thus exacerbating the data sparsity problem.

One remedy that has been used to address this problem (e.g. by TechLens+ [15]) is to leverage the bibliographic citations in the journal articles as a substitute for user ratings. However, this solution is partial at best. One reason 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 [5] 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 and that the cited work deserved criticism.

Another reason that citations are a poor substitute for ratings is that they only provide boolean rather than numeric ratings, as is usually the case with movies and music. One approach to measuring the importance of a bibliographic citation in an article is to measure the frequency with which the cited article is mentioned within the citing article. Such a measure would be some indication of the relative relevance of citations in the bibliography. Another indicator of importance could be the distribution in various sections of the article where the cited article is mentioned. However, such methods would require a detailed analysis of the citation occurrences in the text that is not readily available.

A different way to assign a numerical rating to a bibliographic citation is to give it the value obtained from applying a link analysis algorithm – such as HITS [10] or PageRank [4] – to the graph of article citations.

There are several reasons for expecting that ratings defined by PageRank values could improve collaborative filtering recommendations. First, the use of PageRank for evaluating the influence of scholarly journals as a whole has successfully been applied by Eigenfactor [2]. In that study journals titles obtain high impact scores if they are often cited by other high impact journals as measured by an eigenvector centrality method [11] equivalent to PageRank. Hence it seemed likely that applying this technique to individual articles would also be effective.

Second, previous successes in the use of PageRank as a measure of the “impact factor” of an article [9, 12] suggests that PageRank weights could be a proxy for the numeric rating that an article might give to another article. Furthermore, a recent study [16], which addressed the question of whether Web links are analogous to citations, showed that there was a significant correlation (57%) between web-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Recommender Systems 2009 New York, New York USA  
Copyright 200X ACM X-XXXXX-XX-X/XX/XX ...\$10.00.

citations and bibliographic citations (ISI’s Journal Impact Factor), further reinforcing the belief that measures applied to one also apply to the other.

Third, it has been shown [8] that a PageRank-based method (PaperRank) for predicting co-occurring references in an article, given an initial subset of references, 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.

Finally, we know that collaborative filtering data can be used by a PageRank algorithm to improve rankings in search results [13]. One might expect the converse effect – applying PageRank data to a collaborative filtering algorithm – would yield improved recommendations.

This paper describes the effect of applying a simplified Weighted PageRank algorithm on a citation graph and using the resulting rankings as preference-scores from which to generate item-based recommendations. Contrary to expectations, experimental results (section 4.1) show that PageRank significantly *decreases* the quality of recommendations based on Top-N measures. Section 5 discusses some of the possible reasons for this counter-intuitive result.

## 2. PAGERANK ON CITATIONS

Consider the graph of references to articles in a collection as the raw data to which the PageRank algorithm is applied. For the purposes of this study co-authorship information was ignored (although it has recently been shown that including information about co-authorship enhances the PageRank values [6]) and used only the citation network to establish a PageRank value for each article using a Weighted PageRank algorithm [17].

Weighted PageRank assigns larger values to more frequently cited articles by computing a function on both the number of referencing articles and the number of referenced articles on the nodes in the graph. Thus if article  $u$  references article  $v$  then the link between  $u$  and  $v$  is weighted by both the inbound links on  $u$  and  $v$  (denoted by  $W_{(u,v)}^{in}$ ) and the outbound links on  $u$  and  $v$  ( $W_{(u,v)}^{out}$ ) where

$$W_{(u,v)}^{in} = \frac{I_v}{\sum_{p \in R(u)} I_p}$$

and

$$W_{(u,v)}^{out} = \frac{O_u}{\sum_{p \in R(u)} O_p}$$

$R(u)$  - represents the list of references in article  $u$  and  $I_k$  ( $O_k$ ) is the number of references to (references from) article  $k$ .

Hence the page rank value is given by

$$PR(v) = (1 - d) + d \sum_{u \in B(v)} PR(u) W_{(u,v)}^{in} W_{(u,v)}^{out}$$

where  $d=0.8$ , which was determined to be an optimal value in the application of PageRank to improve the relevance ranking of results obtained from a Lucene search on the Cystic Fibrosis Reference Collection [7].

Typically the graph of citations in a collection of journal articles is not necessarily fully connected. Depending on

	p <sub>1</sub>	p <sub>2</sub>	p <sub>3</sub>	p <sub>4</sub>	p <sub>5</sub>	p <sub>6</sub>	← citations
articles	p <sub>1</sub>	0.2				0.4	← PageRank Values
	p <sub>2</sub>			0.5		0.4	← PageRank Values
	p <sub>3</sub>	0.2			0.3		← PageRank Values
	p <sub>4</sub>		0.7	0.5			← PageRank Values
users	u <sub>1</sub>			0.8			← User Ratings
	u <sub>2</sub>	0.5			0.5		← User Ratings

Figure 1: Ratings Matrix populated with PageRank and User Ratings

the collection, the graph may contain one or more disconnected subgraphs. Hence, before applying the algorithm, an artificial root-node was created to which all the disjoint subgraphs are forced to connect and then applied the PageRank algorithm to the fully connected graph. This process yields a consistent, normalized PageRank value in the range [0,1] for each article.

## 3. RECOMMENDER

The experiments described below are the result of applying a user-based collaborative filtering recommender that implements k-nearest neighbour and cosine correlation in the Taste framework (now part of the Apache-Lucene machine learning library Mahout [1].) Each item (article) that has references to other items (articles) is considered as a “user” and is given a preference-weighting (rating) for each reference that is equal to the PageRank value for that article. We chose user-based collaborative filtering because it has been shown that for sparse matrices where the number of items dominates the number of users, user-based article recommenders have better accuracy than item-based collaborative filtering [3].

Once the PageRank value for each article is computed, this value is assigned to each occurrence of the article in the citation matrix and all occurrences of that article’s “rating” will have the same value (see Figure 1). 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 citation would typically assign the citation a different rating. 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.

In the absence of any further information about the various reasons for which an article may be cited – such as “coverage” in the “related work” section or the self-reference of prior work for which the current article is the continuation – one might expect that substituting the boolean (“cited” or “not-cited”) with the PageRank value of the article would improve the Top-N effectiveness of collaborative filtering recommendations.

## 4. EXPERIMENTS

The off-line experiments were modeled after off-line experiments designed for TechLens+ which measured the rec-

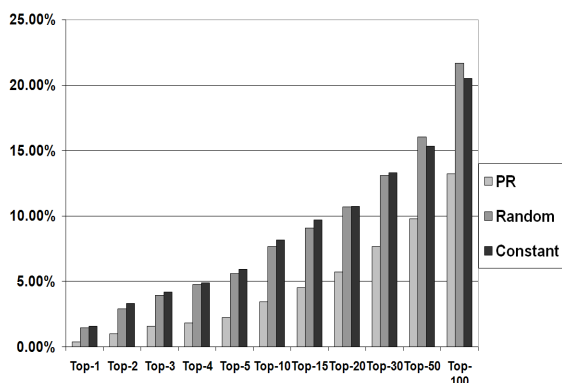


Figure 2: Top-N results for PageRank, Random and No Page Rank on a large data set

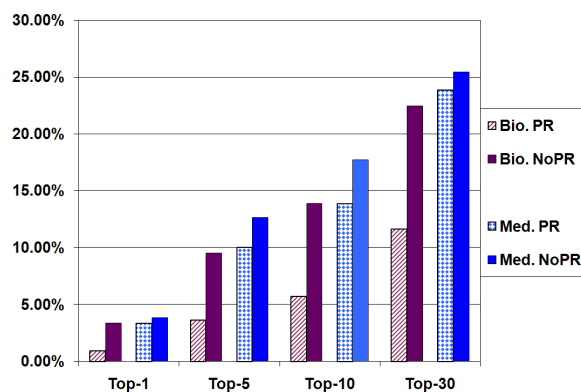
ommender’s effectiveness by computing Top-N recommendations [15]. The set of test articles was selected to have a significant number of references and, for each article in that set – the active article – each reference was removed one at a time and the recommender was tested for whether it predicts the 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.

We compared the Top-N effectiveness of PageRank-weighted references against boolean (constant) weighted references on three related bibliographic collections. The large collection of biomedical articles subsumes a small collection taken from medical journals and a medium-sized collection was extracted from journals on biology. In each case, we considered only the articles for which there were references in the collection.

The large collection contains 369,470 articles and 1,461,305 references to those articles. The small subset contained 7,495 articles which contained only 41,00 references to articles in that subset (for an average of only 0.55 references per article.) The medium-sized subset contained 38,667 and also had a small average number of references (1.15 per article) to articles in that subset. The “connectivity” of the article collections — measured as the number of references in an article plus the number of articles that cite it — was slightly above 1 for the small collection, slightly above 2.3 for the larger collection and over 3.9 for the entire collection (compared to 14 in the CiteSeer collection used in TechLens+).

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



[htp]

Figure 3: Top-N results of PageRank and No Page Rank on 2 data sets

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 further narrows the number of usable references that can be used for recommendation purposes, and on such relatively small collections, we deemed this constraint unnecessary.

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.

## 4.1 Results

The experimental results of the effect of PageRank are summarized in Figure 2. The results show that 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).

A striking, albeit unimportant aspect of the results summarized in Figure 3 is that the smallest dataset produced the best Top-N results. This is an artifact of experimental constraint that recommended items are only permitted to be the elements of the set of articles to which references are made within the article set.

More importantly, the graph in Figure 3 shows that *not* using PageRank dramatically improves the effectiveness of recommendations. Yet we expected that PageRank values from a citation graph would at least be an approximate substitute for ratings on citations.

## 5. DISCUSSION

It is not clear how best to interpret this result, particularly in view of the success of applying this kind of technique for measuring the impact factor of scholarly journals as a whole [2]. One cause may be that the PageRank value of an article is a measure of the overall citation network’s “rank-

ing” of that article based on the entire citation graph. Thus for a (relatively) rarely cited article, the application of the PageRank value of that article as a measure of its preference weighting from a citing article puts a considerable bias against it (too close to “0”) compared with a purely boolean rating. In those instances, a constant (boolean) weight, or even a random rating is preferable. On the other hand, for frequently cited articles, one would expect that assigning them high page-rank value instead of a “1” to have little or no effect.

This effect is going to be even more pronounced in cases where an article’s low PageRank value is due to the short length of time that an article has been published. It is known that recent articles (especially in the first 2 years after the publication date) have significantly lower citation rates than older ones [14]. Although that bias could easily be attenuated by eliminating articles that are less than two years old, this too may have had an impact on the results for this experiment.

We noted in section 2 that every cited article is assigned the same PageRank value and that this value is itself a poor substitute for a true rating, in part because of the wide range of motivations that authors have for citing an article. For instance, survey articles may have a large number of references simply because they cover a wide field. It might be that a better measure of an article’s “preference” for another article might be to weight each article’s PageRank rating by a function of the density of citations in the article. Thus a short survey article with many references would have lower preference weights and a long focused article with fewer references would have higher preference weights.

## 6. FUTURE WORK

While this experiment shows a negative result for a particular implementation of PageRank, it would be imprudent to conclude that applying link analysis techniques on citations is useless input to the collaborative filtering recommendation of journal articles. Different damping factors ( $d$ ) could have a significant positive effect as well as a more refined implementation of PageRank that uses co-authorship information [6]. Also, it is possible that alternative link-analysis algorithms such as HITS could generate better recommendation results than PageRank.

## 7. REFERENCES

- [1] <http://lucene.apache.org/mahout>.
- [2] W. J. Bergstrom, C.T. and M. Wiseman. The Eigenfactor Metrics. *Journal of Neuroscience*, 28(45):11433–11434, 2008.
- [3] T. Bogers and A. van den Bosch. Recommending Scientific Articles using CiteULike. In *Proceedings of the Second Annual Conference on Recommender Systems*, pages 287–290, Lausanne Switzerland, October 2008.
- [4] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Computer Networks and ISDN Systems*, pages 107–117. Elsevier Science Publishers B. V., 1998.
- [5] D. O. Case and G. M. Higgins. How can we investigate citation behavior?: a study of reasons for citing literature in communication. *J. Am. Soc. Inf. Sci.*, 51(7):635–645, May 2000.
- [6] D. Fiala, F. Roussetot and K. Ježek. PageRank for Bibliographic Networks. *Scientometrics*, 76(1):135–158, 2008.
- [7] D.Inkpen, A. Constantinescu, and G. Newton. Ranking scientific articles using a citation-based pagerank algorithm. In *NLPIR4DL 09: Workshop on text and citation analysis for scholarly digital libraries*. ACL, 2009.
- [8] M. Gori and A. Pucci. Research paper recommender systems: A random-walk based approach. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 778–781, Washington, DC, USA, 2006. IEEE Computer Society.
- [9] Ježek, K. Fiala, D. and Steinberger, J. 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, 2008.
- [10] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46:668–677, 1998.
- [11] M. E. J. Newman. Mathematics of networks. In *The New Palgrave Encyclopedia of Economics, 2nd edition*, L. E. Blume and S. N. Durlauf (eds.), Basingstoke, UK, 2008. Palgrave Macmillan.
- [12] N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Information Processing and Management*, 44(2):800–810, 2008.
- [13] S.-T. Park and D. M. Pennock. Applying collaborative filtering techniques to movie search for better ranking and browsing. In *KDD '07: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 550–559, New York, NY, USA, 2007. ACM.
- [14] 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*.
- [15] 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.
- [16] L. Vaughan and D. Shaw. Bibliographic and web citations: What is the difference? *Journal of the American Society for Information Science and Technology*, 54(14):1313–1322, July 2003.
- [17] W. Xin and A. Ghorbani. Weighted pagerank algorithm. In *CNSR '04: Proceedings of the Second Annual Conference on Communication Networks and Services Research*, pages 305–314, Washington, DC, USA, 2004. IEEE Computer Society.