

# Recommending Related Papers Based on Digital Library Access Records

Stefan Pohl  
Cornell University  
Ithaca, NY, USA  
sp424@cs.cornell.edu

Filip Radlinski  
Cornell University  
Ithaca, NY, USA  
filip@cs.cornell.edu

Thorsten Joachims  
Cornell University  
Ithaca, NY, USA  
tj@cs.cornell.edu

## ABSTRACT

An important goal for digital libraries is to enable researchers to more easily explore related work. While citation data is often used as an indicator of relatedness, in this paper we demonstrate that digital access records (e.g. http-server logs) can be used as indicators as well. In particular, we show that measures based on co-access provide better coverage than co-citation, that they are available much sooner, and that they are more accurate for recent papers.

## Categories and Subject Descriptors

H.3.7 [Information Storage and Retrieval]: Digital Libraries; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval

**General Terms:** Algorithms, Experimentation

**Keywords:** Recommendations, co-citation, co-download, http access logs

## 1. INTRODUCTION

In scientific literature, citation information is a key source of information about relationships between documents. Citations are used to measure impact of documents and journals [3], to identify related papers via co-citation and bibliographic coupling [9, 5], and to improve ranking in keyword-based search [6]. Unfortunately, there are at least two problems with citation data. First, extracting citations from academic articles requires manual curation or sophisticated natural language processing. This makes such data costly and time-consuming to obtain. Second, it takes considerable time for a newly published article to gather a sufficient number of citations for meaningful statistical analysis.

We demonstrate that access data of the form “user X downloaded document Y” can be used as a substitute for citation data. Access data does not suffer from the drawbacks of citation data, since it is available sooner and can easily be extracted from digital library access logs.

In this paper, we focus on using access data to identify related papers. Treating access data as a bi-partite graph of users and documents analogous to item-to-item recommendation systems (see e.g. [8]), we explore an access-based measure to quantify the degree to which pairs of articles are

related. We evaluate how well this measure predicts future co-citations on the arXiv e-Print archive [1]. Our results show that access-based measures have vastly larger coverage and are more accurate at finding related work than co-citation for recently published papers. Additional and more detailed results can be found in [7].

## 2. ACCESS DATA

The arXiv collection recorded over 650 million accesses to over 350,000 scientific documents between 1994 and July 2006. For each access, we extracted the time and date, source IP address and document accessed. After filtering proxies and crawlers, we segmented accesses into 30 minute sessions from each IP, assuming these to define individual users. This gives, for each session, a set of documents downloaded by the user. Analogous to the co-citation measure of relatedness [9, 5], we then transform these sets into counts of how often each pair of documents was co-downloaded.

When dealing with access data, care has to be taken to avoid presentation biases [4]. In particular, we found that access data is influenced by publication date. First, older papers tend to be less accessed. Second, papers published at the same time are often presented on the same web page and tend to be co-accessed more often. For arXiv, this is especially visible during the first month after publication: Many users subscribe to announcements listing all new articles. In contrast, we expect most valuable access data to result from searches of users for a specific topic. Hence we ignored co-downloads of documents appearing together that occurred during the first month after publication.

## 3. EVALUATION

We compare co-citation and co-download in terms of coverage and recommendation quality.

### 3.1 Coverage of Co-access Data

Figure 1 shows the maximum number of times each paper in arXiv was co-cited and co-downloaded with any one other paper. The papers are sorted by this count independently along the horizontal axis. We see that using co-download, for almost every paper in arXiv we have data and are therefore able to make recommendations. In contrast, about two thirds of the papers have no known co-citations, and those that do often are co-cited only once or twice. Hence co-citation cannot make any recommendations for most papers.

More detail about the number of recommendations that co-download and co-citation can make is given in Figure 2. The graph shows the mean number of recommendations (i.e. papers with non-zero co-downloads or co-citations) as a function of time after publication of a paper. For efficiency, we limit ourselves to 100 recommendations per paper. We

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.

JCDL'07, June 17–22, 2007, Vancouver, British Columbia, Canada.  
Copyright 2007 ACM 978-1-59593-644-8/07/0006 ...\$5.00.

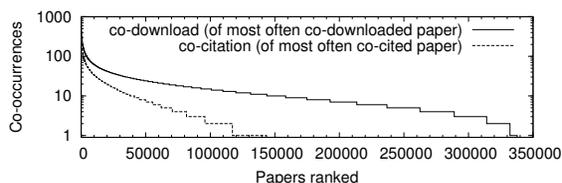


Figure 1: Co-citation and co-download coverage

see that the number of recommendations from co-downloads quickly reaches the maximum. In contrast, the average number of recommendations from co-citations grows much more slowly. Thus, building a citation based recommendation system takes considerably more time than using access data.

### 3.2 Recommendation Quality

While Figure 1 shows that the number of co-downloads is much larger than the number of co-citations, the former might be more noisy. An author’s decision to cite a paper is likely to mean more than a download by an anonymous user, who might never even look at the document. To examine if repeated measurement compensates for such noise, we now evaluate the quality of recommendations.

We use the number of co-downloads before 2005 as a measure of similarity between two papers. In particular, for a given paper, we recommend related papers by sorting all other papers by this similarity. As ground truth, we took the publications after 2005 and assumed that papers cited together are related, and all other papers are unrelated. The citation data came from manual processing of roughly 200,000 arXiv submissions administered in the SLAC/SPIRES database. To estimate the quality of the recommendations, we take one paper D from a set of references in a 2005 paper, calculate the Mean Average Precision (MAP) [2] for the recommendations made by co-download, and aggregate the MAP scores as a function of the time after publication of D. We report average results over 7,500 papers. We also performed the same experiment using co-citation instead of co-download.

Figure 3 shows the MAP of the recommendation lists. We see that using co-download results in higher MAP than co-citation for the first two years after publication of a paper. Hence, to find related documents to recent papers, co-download is more informative. In this collection, two years after publication there are usually sufficiently many citations for co-citation to catch up.

While co-download performance slightly decreases beginning two years after publication, we believe that this is an artifact of the experiment design: the reference lists of papers citing older papers tend to contain other older papers. While co-citation benefits from this, co-download is penalized as it often recommends more recent related papers. Note that the experiment is further biased in favor of the co-citation measure, since (future) co-citations are used as ground truth.

Finally, the low absolute MAP values result from us considering only a single set of papers in the same references list as related for each evaluation. This means that when evaluating, there are often only about 10 other papers considered related out of the entire collection.

## 4. CONCLUSIONS

We have demonstrated that digital library access records are a valuable resource, containing implicit information about

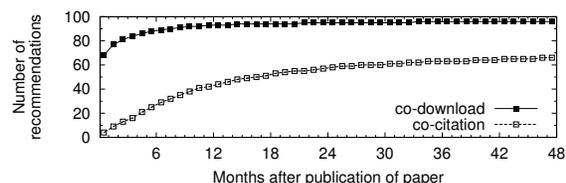


Figure 2: Count of recommendations over paper age

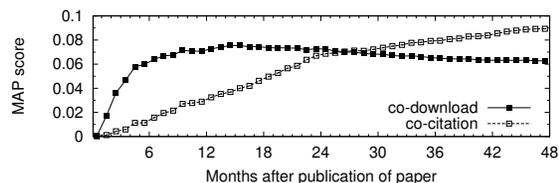


Figure 3: Mean average precision over paper age

the relatedness of pairs of documents. We found that co-download is able to outperform citation-based recommendations on recently published papers in the arXiv collection. Furthermore, in contrast to co-citation, recommendations from co-download are available for practically all documents in the collection in a timely fashion and without need for expensive extraction and curation. We conclude that access records can be used effectively in recommender systems for digital libraries, especially when citations are not available (e.g. for images, audio, video, documents without citations), where citation indexing is difficult, or where citations are rare or unlikely to be resolved.

We thank Paul Ginsparg and Simeon Warner (arXiv.org) for valuable discussions and for the data that enabled this research. We also thank Travis Brooks (SLAC/SPIRES). This research was supported by a Google gift and NSF CAREER Award IIS-0237381. The second author was partly supported by a Microsoft Research Fellowship.

## 5. REFERENCES

- [1] arXiv e-Print archive, <http://arxiv.org/>.
- [2] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.
- [3] E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.
- [4] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR*, 2005.
- [5] S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan and J. Riedl. On the recommending of citations for research papers. In *CSCW*, 2002.
- [6] L. Page, S. Brin, R. Motwani and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Technical report, Stanford Digital Library Technologies Project*, 1998.
- [7] S. Pohl. Using Access Data for Paper Recommendations on ArXiv.org. Master’s thesis, Technische Universität Darmstadt, Germany, December 2006.
- [8] B. Sarwar, G. Karypis, J. Konstan and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *WWW*, 2001.
- [9] H. G. Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *JASIST*, 24(4):265–269, 1973.