Co-usage of documents in a large digital library*

submitted to Fourth Delos Workshop

2002–04–14

José Manuel Barrueco Cruz
Biblioteca de Ciències Socials "Gregori Maians"
Campus dels Tarongers s/n
Universitat de València
46071 València
Spain
jose.barrueco@uv.es
http://www.uv.es/~barrueco

Thomas Krichel
Palmer School of Library and Information Science
Long Island University
720 Northern Boulevard, Brookville
New York 11548–1300
USA
Krichel@openlib.org
http://openlib.org/home/krichel

Abstract

The RePEc Economics library offers the largest distributed source of freely downloadable scientific research reports in the world. WoPEc is a user services of that library. It operates on the Internet since 1993. It has a well-established user community, and a relatively narrow subject coverage.

In this paper, we wish to find out which papers in the collection are similar through usage. The idea is that if different users request a couple of papers consistently together, then these papers are likely to correspond to the same information needs. They are similar in this sense. We present a theoretical discussion of these relationships and an empirical assessment. We introduce a measure of co-usage and estimate results for the WoPEc user service.

This paper is available online at http://openlib.org/home/krichel/papers/kumegawa.html. However, that version does not contain mathematical expressions and is provided for evaluation purposes only. The full paper is available in PDF for A4 paper, and for letter size paper.)

*The work discussed here has received financial support by the Joint Information Systems Committee of the UK Higher Education Funding Councils through its Electronic Library Programme.

1 Introduction

Creating relationships between similar documents is an activity that has a long distinguished history in Information Retrieval. In his textbook on the subject Korfhage (1997) writes

> Probably the single key concept behind information storage and retrieval is document similarity.

A practical motivation for the search for similarity is to improve information retrieval. When users have found one document they can be directed to similar documents that may be worth their attention. A more theoretical motivation is the discovery of patterns of thought within a body of knowledge that is dispersed across a collection of documents.

There are, of course, many ways in which similarity between documents can be assessed. One way is to use the judgment of an expert. This is typically done by classifying documents according to a subject classification. This is a tried and trusted method. However, it requires human expertise and labor. These do not come cheap these days. Therefore many collections have no classification data. Within the collection of documents that we study in this paper, three in four documents do not have a subject classification. Other classic measures of relationship between academic papers are the strengths of co-citation and bibliographic coupling. Such an analysis presumes that a citation index for the papers is available. Producing one manually is even more expensive than subject classification. It is possible to use recently available software for autonomous citation indexing. Barrueco Cruz and Krichel (2002) report on on recent effort s using the same digital library that we study here.

In this paper, we introduce a completely new concept of similarity that is user-centered. We start with the idea that two documents are similar if they respond to the same information need. Thus roughly speaking we can think of two document as similar, if many users will withdraw the papers together while they are working on the system. In Section 2, we refine this idea further. Provided that we have a digital library with a long log of usage, we can find which papers have frequently been used together. This is what we will refer to as "co-usage" in the following.

We are not aware of a study that has developed and tested this concept within a digital library context. We are, of course, aware that commercial systems use a related concept. Amazon.com, for example, suggests to a customer who is considering to purchase an item x that other people who bought x also bought item y. But since users are anonymous in our digital library, within the scope of our collection, the evidence on common usage is more difficult to gather. In Section 3 we present some analytical results. Section 4 concludes.

2 Measuring co-usage

Our dataset comes from logs of the usage of the RePEc digital library. We have built this library over many years. RePEc is the largest free distributed academic digital library in the world. It is a pioneering effort in three respects. First, RePEc pioneered the business model of the Open Archives Initiative distinction between data providers and service providers. RePEc has grown over 200 archives that are used in 10 different user services. Second, RePEc is more than a collection of data about documents. RePEc offers a comprehensive picture of academic output by describing collection of documents, persons that are involved as authors of documents or editors of collections, and the institutions to which they belong in a relational database scheme. Third, RePEc is essentially volunteer driven and has no owner. It is a public good much like the Internet.

The data that we examine comes from the WoPEc user service. This service, founded in 1993 is available on the web at three locations since 1994. These are Hitotsubashi University in Tokyo, Japan, Manchester Computing in the UK, and Washington University in St. Louis in the USA. The web site has static pages for each paper. In addition, there are a number of pages that link different papers together. There are pages that list papers that have been issued within same series, and there are pages that group papers with the same subject classification. The log data also contains indication about what searches users have been making. However, as far as we are concerned, only the access to the description of the individual papers is relevant.

To begin with, recall that WoPEc is a non-commercial service. WoPEc does not enter into a contractual arrangement with a customer, with or without an exchange of money. Contractual arrangements with customers are expensive to arrange because they involve complicated negotiations, privacy issues etc. On the other hand contractual arrangement—even without exchange of money—generate customer data. The customer data itself has considerable value, see Shapiro and Varian (1999) for an illuminating discussion. In the period that we examine, cookies have not been used. Therefore we need to wade through the log to find traces of usage by the same persons. All that the log offers is a trace to the Internet address (i.e. IP number) of the client machine, a file or action performed, and a time when this happened. Over recent years, the Internet is more and more accessed through personal computers rather than through multi-user systems. However, we can not be completely sure that an IP number corresponds to the machine of an individual user, because there are caching sites. Therefore we must be watchful of IP addresses that appear particularly frequently.

Let us for the moment make the assumption that the usage data from one single machine comes from one single user. We are then in the happy situation that we have identified the user. However, to be able to claim that papers that are appearing together are joined together because they correspond to the same user need, we still have to assume that the information need of the user remains the same throughout the period of observation. This is problematic. Typically users will regard WoPEc as a reference source, and therefore naturally turn to it with different information needs. Therefore it is likely that the log contains material that corresponds to different information needs, even if they are expressed by the same user.

Therefore, we have chosen to think of users using the system within sessions. We typically think of a session as

an academic, using a web browser on a desktop machine, going to the WoPEc site, make a few searches and look at a few results. Or we can think of it as a student, walking into a computer lab, uses a search engine, finds a page on the WoPEc service and navigates it for further data. Technically, we define a session as a subset of the total log that comes from the same IP address within a certain time frame. In fact we assume that there must not be more than one hour between two consecutive log entries for them to qualify to be within the same session. In addition, we do not consider repeat access to the same paper as two different accesses to that paper. Such access is likely to occur mainly for technical reasons, such as that the user did not have time to read the article at the first discovery. Under these restrictions it appears to be convenient to model sessions as sets of papers. We can now proceed to a formal definition of co-usage.

Let D be the set documents in the library and S be the set of sessions. Each session s contains a subset of the set of documents, i.e. $s \subset D$. Introduce $|\cdot|$ as as notation for the number of elements in a set. Let there be two documents d_i and d_j. Let $u(d_i, d_j)$ be the set of all sessions that contain both d_i and d_j. For convenience of notation, write $u(d_i) = u(d_i, d_i)$. Then we propose to call the co-usage coefficient $c(d_i, d_j)$ as follows

$$c(d_i, d_j) = \frac{|u(d_i, d_j)|^2}{|u(d_i)||u(d_j)|}.$$

The motivation for introducing the product in the denominator is to ensure that we do not have situations where one paper appears frequently with another by the sheer size of its own usage. It also leads $c(d_i, d_j)$ to gain the attractive properties that $0 \leq c(d_i, d_j) \leq 1$ and that $c(d_i, d_i) = 1$.

It is clear that the measure that we propose here is an intuitive one. It is not a scientifically developed metric of the likelihood that, whatever size the session is, we will see the two papers appearing together in the session. It is possible to evaluate that probability. However that calculation is not attempted here. It is left for future work. Even if we can theoretically ascertain how to calculate the probability, we can not in practice calculate it for all possible document couples, because of the amount of calculations that would have to be made within a large digital library as ours.

3 Results

We ran a program to search for co-usage on a log for three years 1999–2001. There are 102551867 lines in the log. We discarded 24836221 requests for images, 401614 requests for CSS files, and 9315 contained another error. 3164598 requests contained a 404 error and were omitted. Most of them went to `robots.txt`.

3966837 distinct hosts accessed the system. We exclude all accesses from hosts that accessed `robots.txt`. In addition, there are cache and proxy sites that we would like to remove as well. This is a more difficult terrain. The rule that we chose was to look at each day's log, and detect all hosts that issued more than 100 requests. In total, we detected 23068 such hosts, and discarded 17484961 lines because they were made by these hosts. This left us with 23413849 to analyzed. We detected 5251337 sessions. We limited ourselves to sessions that had more than 3 accesses and less than 100, and that included at least one search. There were 69179 sessions that remained for analysis.

We can not make the results available here, for obvious reasons of space. Therefore we have placed them at http://openlib.org/home/krichel/kumegawa/report.htm. We print the top couples that have the highest co-usage, up to a rounded coefficient of 50%. The following results appear when we look through the data.

First, papers in the same series seem to be much more prone to co-usage than papers from different series. It is often papers by the same authors, over different periods of time, available in the same working paper series or journal that appeared to be co-used a lot. This observation suggests that the original split of the data in different series does not only provide an organizational layer to the data, but also a semantic one.

Second, from reading through the related couples, it appears that a key concept that appears in the title immediately explains the co-usage. Thus, it should be possible to gain insights into key concepts that are of user interest by looking at the title data of papers that are co-used, and find the common words. This would be an interesting subject for further research.

4 Conclusions

This paper has introduced co-usage as a means to make document to document relationships. The concept of co-usage has been found to lead to excellent empirical results. Our results suggests that the concept of co-usage is a valid and interesting one. Any remaining problems that we are experiencing to make it work in practice are consequences of poor measurement of usage, rather than flaws in the underlying theoretical reasoning. Just in the case of co-citation and bibliographic coupling, there will be a debate on how useful co-usage analysis really is. We have opened that debate with this paper.

There are two further consequences from this paper. In the short run, we intend to incorporate co-usage similarity into future developments of WoPEc, to build links from the description of one paper to the description of another as suggested by co-usage analysis. In the longer run, we believe that our work can be seen as simple, but pioneering step towards information retrieval systems that are learning. Instead of having a system that has a fixed procedure that is applicable whatever the query an the collection, we can imagine that future information systems will learn from past user behavior to better serve the present user. Co-usage appears as a key concept to move things forward in that direction.

References

Barrueco Cruz, José Manuel and Thomas Krichel (2002). Automated extraction of citation data in a large digital library. available at http://openlib.org/home/krichel/papers/ciudadreal.pdf.

Korfhage, Robert R. (1997). *Information Storage and Retrieval*. John Wiley and Sons.

Shapiro, Carl and Hal Varian (1999). *Information Rules*.

Harvard Business School Press.