Co-occurrence Analysis of Access Log of Institutional Repository

Baba, Kensuke
Kyushu University Library, Kyushu University

Ito, Eisuke
Research Institute for IT, Kyushu University

Hirokawa, Sachio
Research Institute for IT, Kyushu University

出版情報：Proc. Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011), pp.25-29, 2011-01
バージョン：
権利関係：

Co-occurrence Analysis of Access Log of Institutional Repository*

Kensuke Baba† Eisuke Ito Sachio Hirokawa

Abstract

Institutional repository is playing an important role to guarantee open access to research outputs by self archiving. However, the number of the items in most institutional repositories is extremely fewer than that of the total research outputs produced in the institute. One of the reasons is that most researchers have no incentive to register their research outputs, simply because the effectiveness of registration to institutional repository is not clear. The authors are constructing a feedback system for researchers who register their research outputs to institutional repository. In this paper, they focus on access log analysis to discover meaningful knowledge on when, how, and why the items are accessed. The knowledge from the access log can utilized also for recommendation of items for users (readers) of the institutional repository. This paper shows some results of co-occurrence analysis for access log of the institutional repository of Kyushu University, and shows some ideas of advanced analysis to obtain meaningful knowledge.

keyword: Institutional repository, access log, recommendation, co-occurrence, visualization, formal concept analysis.

1 Introduction

"Open access [21]" to scholarly information provides free availability of research outputs such as scholarly papers. The number of research institutions who adopt a mandate to provide open access to their research outputs is increasing [9]. Especially, for researches founded by a public institution, it seems to be the general situation that the outputs should be open. For example, the National Institutes of Health (NIH) shows their policy which requires researchers founded by NIH to open their research outputs [10]. One of the vehicles for delivering open access is "self archiving" [16], and then a *repository* is a system to archive and open research outputs. A repository for outputs in an institution is called an *institutional repository* (*IR*) and one for outputs on a particular research area (for example, arXiv [1]) a *subject repository*.

The number of the research institutions who have their own IR is about 1,600 as of September 2010 [8]. Since the number of the higher education institutions in the world is more than 20,000 [7], there is yet room for increasing the number of IRs. Additionally, the ratio of the research outputs archived in the repositories to the total number is estimated to be small, for example, the ratio in the IR of Kyushu University (QIR) [5] is about 20% [12] while the number of the items in QIR ranks 126th in the world [6]. To improve this situation, in addition to making the ideas of open access and IR universally known, the effectiveness of IRs should be clear with practical data.

The distinguishing trait of repository is that the detailed situation of usage of the contents can be observed as its access log. If some kinds of meaningful knowledge are mined from the access log, it can be regarded as an effectiveness of the IR. Some basic analyses (for example, counting the number of the accesses with respect to each item, author, region of the referrer, and so on) can be operated by a standard function of DSpace [2] or Google Analytics [3]. Although there exist some researches of quantitative analysis of accesses on IR [13, 19], more detailed and qualitative analyses are required to show an effectiveness of IR. As for some open access journals [18, 11, 17, 22] and a subject repository [14, 15], some kinds of correlation are shown between the number of the citations to a paper and the number of the accesses to the paper. However, a straightforward application of a similar analysis to some IRs does not find any meaningful correlation [20].

A reason of the difficulty on log analysis for IR is that the number of the archived items and the number of the accesses to the items are not enough as samples to estimate some properties statistically. Moreover, compared with open access journals and subject repositories, IRs archive research outputs of various research areas, hence the number of the accesses for each item tends to be small. Therefore, we focused on co-occurrence in accesses of each user. Namely, by considering "the combination of items which the same user accessed" in addition to "the number of the accesses", it is expected that we can obtain meaningful knowledge from access log even in the unfavorable situation for IR. We apply co-occurrence analysis to

*An edited version of this report was published in: *Proc. Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011), pp. 25–29, Jan, 2011.*

†Research and Development Division, Kyushu University Library, *baba@lib.kyushu-u.ac.jp*

Table 1: The number of the contents in QIR as of October 4, 2010.

Type of item	♯ items
Journal Article	1,275
Thesis or Dissertation	108
Departmental Bulletin Paper	11,596
Conference Paper	995
Presentation	238
Book, Chapter	124
Technical Report	544
Research Paper	139
Article	290
Preprint	164
Learning Material	35
Others	531
Total	16,039

the practical access log of QIR. In this paper, as a preparation for the analysis, we count the number of the accesses to each item and the number of the co-occurrences in the accesses of the same user, thereby we show that the access log is worth analyzing.

The main idea of this paper is to encourage authors to register their research outputs to an IR by showing the result of access log analyses. This paper is regarded as a case study of advanced analysis for access log. This work is the first step to study what kind of analysis is useful for authors. Based on this study, various kinds of analysis can be verified from the viewpoint of the incentive for authors to register their research outputs.

2 Motivation

This section describes the basic information of QIR, Kyu(Q)shu University Institutional Repository [5], and the purpose of our analysis of the access log.

2.1 QIR

QIR is the IR based on DSpace and operated by Kyushu University Library since April 2006. The total number of the items in QIR is 16,039 as of October 4, 2010. Table 1 shows the number of each type of items in QIR. The largest type is "Departmental Bulletin Paper" and its ratio is about 72%. Most of QIR items are original (that is, unpublished), therefore these are not suitable for the analysis of the relation with the number of citations [20].

Generally, IR archives the full-text of an item in addition to its metadata such as the title and the author(s). Figure 1 is the Web interface of QIR. The page is the profile page of a researcher and the list is the result of a search of the name in the author fields. The third column is the title of each item and linked to the site of detailed information of the item which includes the full-text. The rightmost column is the number of the accesses to each item, and the number is counted by a standard function of DSpace. In addition to the search of the author, it is possible to search by general keywords in the fields of title, abstract, and so on.

The number of the (metadata of) distinct papers registered in the researcher database of Kyushu University [4] is about 70,000, while the number of the full-text in QIR is about 1,400 as of March 2010 [12]. That is, potentially, there exists a large number of research outputs which are produced in Kyushu University but are not archived in QIR. If the effectiveness of registration to QIR is made clear, it is expected that some researchers register their buried papers. We have developed a linking system between the researcher database and QIR to improve the interface of registration to QIR [12]. The system is another solution of the problem we tackle in this paper.

2.2 Purpose of Analysis

The purpose of this work is to obtain a meaningful knowledge from the access log of IR. Especially, we try to estimate what users want to know by using items in IR. By the knowledge of user's interests, IR can provide some recommendations of items for users, which improve a convenience of IR. For researchers who register their research outputs to IR, IR can feedback the knowledge of user's interests, which is instructive for spotting a research trend.

Some basic analyses of access log can be applied by DSpace [2], Google Analytics [3], and so on. For example, we can count the total number of the accesses for each item and show the ranking on the IR by some basic functions on DSpace. Google Analytics can collect statistics about the region of the referrers of accesses, and the keywords if the access comes from the result of a search engine. However, more detailed and complex analyses are necessary to obtain the user's interests. Especially, for IR in which the number of accesses is small, the analyses are required to mine a knowledge from a limited number of samples.

We focused on the co-occurrence of accesses in addition to the simple total of the number of accesses. In this paper, we practically apply a simple co-occurrence analysis into the access log of QIR to confirm that some meaningful knowledge can be found by co-occurrence analysis. As a preparation for the experiment, first we count the number of the accesses for the items of each author for a single month, and thereby we confirm that there exists an enough num-

Figure 1: The web image of a list of items in QIR. This example is the result of a search of "Kensuke Baba" in the author fields.

ber of accesses in the log data for co-occurrence analysis. Next we count the number of the co-occurrences in the accesses of the same user for a day to confirm there exists co-occurrences on the access log, which is a necessary condition of the effectiveness of the analysis.

3 Experimental Results

The target data of the experiment is the access log of QIR from June 2008 to December 2009. The total number of the accesses is 23,847,393.

3.1 Preprocessing of Log Data

First, we filtered noises by internet bots such as Web crawlers of search engines as a preprocessing of the log data. In the data, the accesses of which address contains the substring "bot" were deleted. By the preprocessing the amount of the data decreased to be 14,870,045 which is about 62% of the original data (Table 2).

For the filtered data of the previous subsection, we counted the total number of the accesses to the items of each author for each month. The result with respect to the top 10 authors is shown in Figure 2. In the graph, the horizontal axis shows the months and the vertical axis the number of the accesses. The number of the accesses for each month is almost from 50 to 250, therefore the number is enough for more precise analyses even after the filtering. The change of the number seems to be reflecting the vacations in our social life, which proves that the accesses are made by human.

Table 2: The number of the accesses by internet bots on QIR from June 2008 to December 2009.

	♯ total access	♯ access by bots	Ratio
2008	7,388,562	2,787,763	37.7%
2009	16,458,831	6,189,585	37.6%
Total	23,847,393	8,977,348	37.6%

3.2 Number of Co-occurrences

We adapted a hypothesis that the access form the same address in the same day represents one user. Table 3 shows the distribution of the access frequency and the number of users. We analyzed 87,628 users, which appears with "∗" in Table 3, whose access count in a day is between 1 and 50.

We removed the users with very frequent accesses or with single access. Most of the frequent users are internet bots. There seems to be a few number of real human users with very frequent access. However, their access targets are so diffused that we removed them from our analysis. We removed the user with single access, since these information gives no hints for co-occurrence analysis. As the result, we found

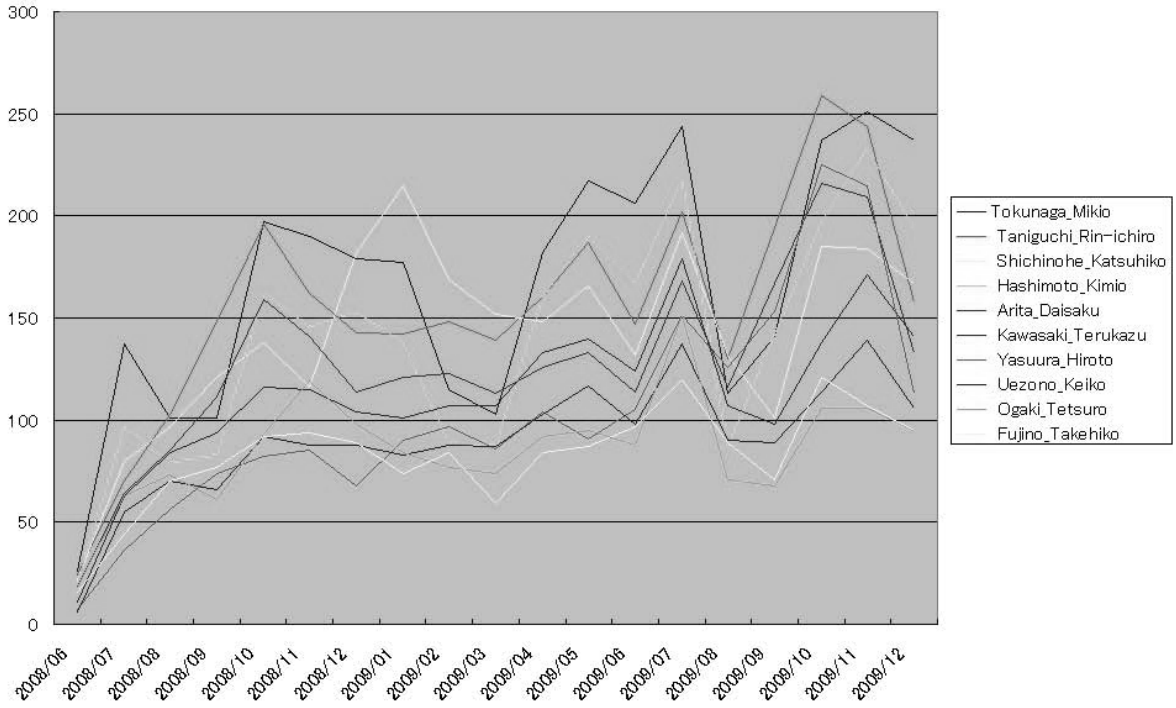

Figure 2: The number of the accesses to the items of the top 10 authors.

Table 3: The number of access count from the same IP.

♯ access (n)	♯ users	Ratio to a	Ratio to b
$1 \leq n$	(^{a})852,346	100.0%	—
$n = 1$	763,882	89.6%	—
$2 \leq n$	(^{b})88,464	10.4%	100.0%
$2 \leq n \leq 50$	(*)87,628	10.3%	99.1%
$50 \leq n$	836	—	0.9%

that there exists an enough number of co-occurrences of accesses in QIR even after the filtering of bots.

3.3 Co-occurrence Analysis

Figure 3 is the result of a co-occurrence analysis for the items in QIR. In the graph, a node shows an item, and the two integers in a node the number of the accesses and the identifier of the item, respectively. Then, a round node shows an item and a square one a list set. An arrow means that the item which corresponds to the end node is used with the item of the start node by the same user. For example, the sub-graph of the top in the figure means that the number of the accesses to the item 2961 is 19, and two users who used the item 2961 also used the item 10651. The initial nodes to construct the graph are decided as the result of a search by a query, and the initial nodes have "$*$" in the node.

4 Conclusion and Future Work

To confirm the effectiveness of co-occurrence analysis, a simple co-occurrence analysis was applied to the practical access log data of the IR of Kyushu University. The number of the accesses to the items of each author for a month and the number of the co-occurrences on accesses by the same user were counted. As the result of the analysis, it was shown that an enough number of co-occurrences exist in the data even after filtering noises by internet bots.

We are going to apply more detailed analysis of co-occurrence to pick meaningful knowledge out of the access log. Moreover, we are going to verify the versatility by applying the analysis to the access log of other IRs. Kyushu University Library is archiving the access log for the electronic journals which are available in Kyushu University, then applying the analysis to the log data is one of our future work.

References

[1] arXiv. http://arxiv.org/, [Accessed Nov. 30,

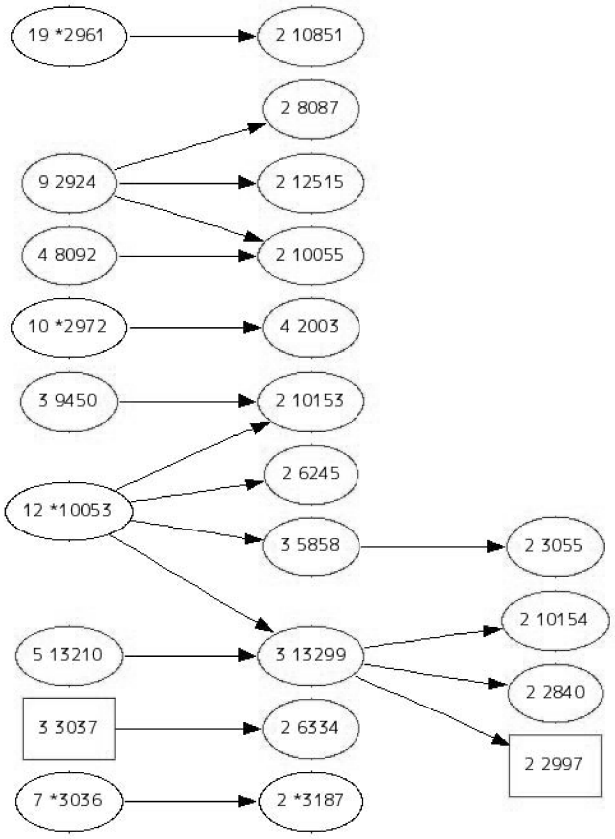

Figure 3: A graph of the result of co-occurrence analysis.

2010].

[2] DSpace. http://www.dspace.org/, [Accessed Nov. 30, 2010].

[3] Google Analytics. http://www.google.com/intl/en/analytics/, [Accessed Nov. 30, 2010].

[4] Kyushu University Academic Staff Educational and Research Activities Database. http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html, [Accessed Nov. 30, 2010].

[5] QIR: Kyu(Q)shu University Institutional Repository. https://qir.kyushu-u.ac.jp/dspace/, [Accessed Nov. 30, 2010].

[6] Ranking Web of World Repositories. http://repositories.webometrics.info/, [Accessed Nov. 30, 2010].

[7] Ranking Web of World Universities. http://www.webometrics.info/, [Accessed Nov. 30, 2010].

[8] ROAR: Registry of Open Access Repositories. http://roar.eprints.org/, [Accessed Nov. 30, 2010].

[9] ROARMAP: Registry of Open Access Repository Material Archiving Policies. http://www.eprints.org/openaccess/policysignup/, [Accessed Nov. 30, 2010].

[10] Analysis of comments and implementation of the NIH public access policy. The National Institutes of Health, 2008. http://publicaccess.nih.gov/analysis_of_comments_nih_public_access_policy.pdf, [Accessed Nov. 30, 2010].

[11] Deciphering citation statistics. *Nature Neuroscience*, 11(6):619, 2008.

[12] K. Baba, M. Mori, and E. Ito. A synergistic system of institutional repository and researcher database. In *Proceedings of the Second International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2010)*, pages 184–188. IAIRA, 2010.

[13] A. I. Bonilla-Calero. Scientometric analysis of a sample of physics-related research output held in the institutional repository strathprints (2000–2005). *Library Review*, 57(9):700–721, 2008.

[14] T. Brody, S. Harnad, and L. Carr. Earlier web usage statistics as predictors of later citation impact. *Journal of the American Society for Information Science and Technology*, 57(8):1060–1072, 2006.

[15] P. M. Davis and M. J. Fromerth. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics*, 71(2):6203–215, 2007.

[16] S. Harnad, T. Brody, F. Vallieres, L. Carr, S. Hitchcock, Y. Gingras, C. Oppenheim, H. Stamerjohanns, and E. Hilf. The access/impact problem and the green and gold roads to open access. *Serials Review*, 30(4):310–314, 2004.

[17] D. E. O'Leary. The relationship between citations and number of downloads in decision support systems. *Decision Support Systems*, 45(4):972–980, 2008.

[18] T. V. Perneger. Relation between online "hit counts" and subsequent citations: Prospective study of research papers in the bmj. *BMJ*, 329:546–547, 2004.

[19] P. Royster. Publishing original content in an institutional repository. *Serials Review*, 34(1):27–30, 2008.

[20] S. Sato, H. Tomimoto, and H. Itsumura. The relationship between citations and number of downloads in institutional repositories. (in Japanese), http://www.tulips.tsukuba.ac.jp/dspace/handle/2241/104229, [Accessed Nov. 30, 2010].

[21] P. Suber. Open access overview. Open Access News, 2007. http://www.earlham.edu/~peters/fos/overview.htm, [Accessed Nov. 30, 2010].

[22] B. A. Watson. Comparing citations and downloads for individual articles. *Journal of scientific research on biological vision*, 9(4):1–4, 2009.