A Feedback System on Institutional Repository

Kensuke Baba
Library
Kyushu University
Fukuoka, Japan
baba@lib.kyushu-u.ac.jp

Masao Mori
Institutional Research Office
Kyushu University
Fukuoka, Japan
mori@ir.kyushu-u.ac.jp

Eisuke Ito, Sachio Hirokawa
Research Institute for Information Technology
Kyushu University
Fukuoka, Japan
{*itou, hirokawa*}*@cc.kyushu-u.ac.jp*

***Abstract*—Repositories are playing an important role in the idea of open access to scholarly information. To increase the number of repositories and the contents in each repository, the effectiveness of repositories should be clear for researchers, that is, providers of the contents. This paper proposes a system which analyzes the access log to the contents in an institutional repository and returns the result to the authors as a feedback from readers. However, the results of detailed analyses with respect to a particular researcher tend to include a kind of individual data, therefore the accesses to the results must be controlled. The proposed system solves the problem by connecting with the researcher database in the institution.**

***Keywords*-Institutional repository; Web database; access log; co-occurrence; visualization.**

I. INTRODUCTION

"Open access [20]" to scholarly information provides free availability of research outputs such as scholarly papers. According to Registry of Open Access Repository (ROAR) [8], the number of research institutions who give the researchers a mandate to provide open access to their research outputs is increasing. Especially, for researchers funded by a public institution, the obligation seems to be the general situation. For example, in 2008 the National Institutes of Health (NIH) showed their policy which requires researchers funded by NIH to open their research outputs [9]. 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*.

According to ROAR, the number of the IRs in the world is about 2,000 as of January 2011. Since the number of the higher education institutions considered in Ranking Web of World Universities [7] is more than 20,000, there is yet room for increasing the number of IRs. Additionally, the number of the research outputs archived in the repositories is estimated to be small compared to the total number. For example, the ratio in the IR of Kyushu University [5] is at most about 30% [12], although the number of the items in the IR ranks 76th in Ranking Web of World Repositories [6] as of January 2011. Namely, most institutions are considered to have a large number of research outputs potentially. To encourage researchers to register their buried outputs (and prevent burying current outputs), we should show the effectiveness of IR for the researchers.

The distinguishing trait of repository is that the detailed situation of usage of the contents can be observed as its access log. For authors, that is, researchers who provide the contents in IR, some kinds of information obtained from the access log can be an incentive to register their research outputs to IR. Actually, some kinds of correlation between the simple total of the access to a paper and the number of the citations to the paper were shown, for some open access journals [18], [17], [21], and for a subject repository [14], [15]. As for IRs, there exist some researches of basic analysis [13], [19]. In addition to the basic analyses, more detailed analyses are required to squeeze useful information for authors from the access log. Some simple analyses (for example, counting the number of the access with respect to each item, author, and region of the referrer) can be operated by a standard function of DSpace [3] or Google Analytics [4]. However, as for advanced analyses, it is not clear what kind of analysis is suitable for authors.

We are developing a feedback system on the IR of Kyushu University. In addition to simple statistics, we analyzed co-occurrence on the access of the same reader [10]. In this paper, we introduce a system which returns the result of the analyses as a feedback from readers into the authors. One of the problems in the implementation is that some authors do not want the result of the analyses to be carried in a conspicuous place. Some IRs display the total number of the access to each item in the IR as a ranking. However, if we display a detailed ranking about authors, some authors may criticize the system (even if the access log is open). The feedback system solves the problem by connecting with the researcher database of Kyushu University [2]. The researcher database has an interface for any researcher in Kyushu University to register their research outputs, and the interface requires an identification to login. Therefore, we can control the access to the result of the analyses by displaying the result on the researcher database instead of the IR.

The main idea of the system is to increase the number

 ISBN: 978-1-61208-135-9

of the items in an IR by showing the result of access log analyses to authors. This paper is regarded as

- a case study of advanced analysis for access log and
- a case study of implementation of the feedback system.

As to the former, 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. As to the latter, this study solves the problem of access control to the result of log analyses by connecting an IR to a researcher database. Since most research institutions have its researcher database, the main idea can be applied to other institutions.

The rest of this paper is constructed as follows. Section II describes the basic information of the IR and the researcher database in Kyushu University to make the problem clear. Section III explains the purpose and the outline of the system we are developing. Section IV concludes this paper and introduce our future work.

II. DATABASES

This section describes the basic information of QIR, Kyu(Q)shu University Institutional Repository and DHJS, Academic Staff Educational and Research Activities Database in Kyushu University ("Daigaku Hyoka Joho System" in Japanese) to make clear the problems we tackle.

A. QIR

QIR is the IR based on DSpace and operated by Kyushu University Library. Generally, IR archives the full-text of each item in addition to its metadata such as the title and the author(s). The total number of the items in QIR is about 16,000 as of January 2011. Ranking Web of World Repositories is taking account of the number of the full-text files as an element of the ranking, then the number of QIR ranks 76th as of January 2011. Since the scope of the ranking is about 2,000 IRs, in most of the IRs the items are less than the number.

Figure 1 shows the number of access and the number of downloads on QIR from July 2008 to December 2009. There exists a month in which the number of the access is more than 200,000. We considered that the number of the access is enough for analyses to obtain some kinds of useful knowledge.

B. DHJS

DHJS is the researcher database of Kyushu University. DHJS has various kinds of data of the researchers in the university, for example, the posts, their research interests, and the scholarly papers they produced. The number of the researchers in the university is about 3,000 as of October 2010. DHJS consists of the two subsystems, the data-entry system and the viewer system. The data-entry system supports researchers to register their research activities to DHJS and equips a user (that is, a researcher) identification by a password. The viewer system shows the research activities registered in DHJS by the data-entry system.

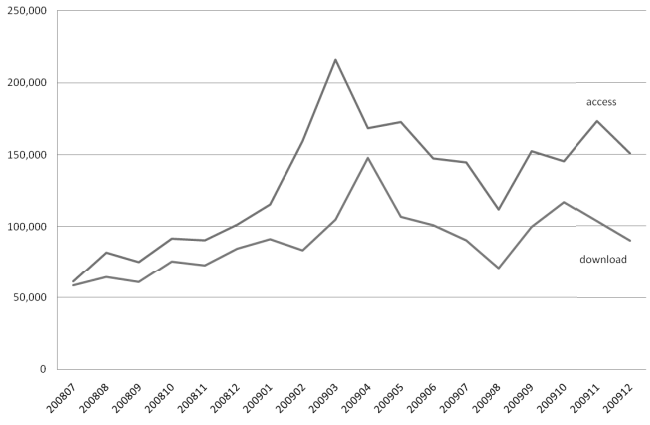

Figure 1. The number of the access and the number of the downloads on QIR from July 2008 to December 2009.

In Kyushu University, any researcher has a duty to register their research activities includes the metadata of scholarly papers into DHJS. Therefore, DHJS has the metadata of most research outputs which were produced in the university in recent years. The number of the "metadata" of scholarly papers registered in DHJS is about 70,000 as of January 2011. The ratio of duplicate data (that is, metadata for the same paper) is estimated at most about 20% [12]. On the other hand, QIR has only 16,000 "full-texts" as mentioned in the previous subsection. That is, potentially, there exists a large number of research outputs which are produced in Kyushu University but are not archived in QIR. Moreover, since the number of the items in QIR ranks 76th in the world, it is estimated that there exists a lot of buried papers in most of research institutions.

We already developed a system which links the metadata of each research output in DHJS to the full-text in QIR [11]. By the linking system, researchers can register the metadata and the full-text of their research outputs into QIR from the data-entry system of DHJS. Since the registration of metadata to DHJS is a duty for the researchers in Kyushu University, the linking system can reduce some efforts to register full-texts to QIR. Therefore, the linking system is another solution of the problem we tackle in this paper.

III. FEEDBACK SYSTEM

We are developing a feedback system on QIR connected with DHJS. This section explains the purpose and the outline of the system, and shows the interface of the system we developed.

A. Overview

According to the basic information in Section II, it is estimated that there exist a large number of unregistered

 ISBN: 978-1-61208-135-9

research outputs in Kyushu University, and most research institutions are in the same situation. A reason of the previous situation is that researchers have no incentive to register their research outputs to IR. Our solution is to analyze the access log of an IR and return the result to researchers as a feedback from the readers of their research outputs. Then, the researchers can obtain the knowledge of reader's interests, which is instructive for spotting a research trend.

Some basic analyses of access log can be applied by DSpace, Google Analytics, and so on. For example, we can count the total number of the access 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 access, and the keywords if the access comes from the result of a search engine. In addition to the basic analyses, we focused on co-occurrence of access [10].

A problem of implementation of the feedback system is that some analyses related to the authors make a kind of individual information. (Note that this problem is different from one for individual data of reader which can be obtained from the access log such as the IP-address.) For example, as to the ranking of the access and the keywords at the referrers for each researcher, some researchers do not want be open. Especially for the ranking, some researchers are worrying that the ranking would be used for assessment of the researchers, rather than the typical privacy problem. Actually, the simple total of the access in IR is not suitable as a criterion for papers or researchers at present, although there seems to be a correlation between the number of access and the number of citations.

To solve the problem, the access to the result of the analyses should be controlled. The system we are developing utilizes the identification function of DHJS. Although QIR also has an identification function of users, the number of the users who have the account of QIR is small. On the other hand, the registration to DHJS is a duty of any researcher in the university. Figure 2 is the outline of the system. As mentioned in Subsection II-B, we have already developed the system to register the metadata and full-text of research outputs to QIR from DHJS [11]. The system introduced in this paper is realizing the other arrow in Figure 2, that is, a feedback from readers of QIR to researchers.

B. Interface

The system applies basic analyses and a co-occurrence analysis to the access log of QIR. The target data is the log from June 2008 to December 2009 and the total number of the access is 23,847,393. We filtered noises by internet bots, and then the amount decreased to be 14,870,045.

1) Basic Analysis: The factors of the basic analyses are

- the total number of the access with respect to each author, and
- its ranks in the department and in Kyushu University.

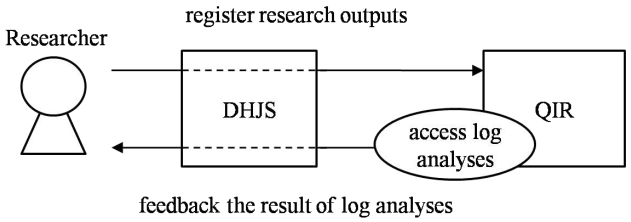

Figure 2. The outline of the feedback system of QIR connecting with DHJS.

Although the total number and the ranking are obtained by simple calculations, they are suitable examples that encourage researchers to register their papers but cause the problem mentioned in previous subsection.

Figure. 3 is an example of the Web image which shows the result of the basic analyses. As we mentioned in the previous subsection, this Web image is shown for a particular user only. The graph describes the number of the access to the items of the user and the top 10 user in the university. The horizontal axis shows the months and the vertical axis the number of the access. The user cannot know who the authors of the top 10 are but which line is for the user. The table is the ranks of the number in the department of the user and in the university for each month.

By the total number of the access, it is expected that the user can know the interest of readers. However, actually, the number depends on some unessential factors, hence it cannot be regarded as a criterion of a research trend or a quality of the paper. This situation is considered to improve by an increase of the number of access and a strict filtering of the noises by bots. We are going to extend the analysis to more detailed results, for example, classifications with respect to each item, the region of the referrer, and so on.

2) Co-occurrence Analysis: We consider "the combination of items which the same user accessed" in addition to "the number of the access" to obtain more meaningful knowledge from the access log.

For the co-occurrence analysis, we adapted a hypothesis that the access form the same address in the same day represents one reader. On the hypothesis, 88,464 readers were regarded to access to more than two items for the access log of QIR. Figure 4 is an example of the result of the co-occurrence analysis. In the graph, a node shows an item, and the two integers in a node the number of the access and the identifier of the item, respectively. An arrow means that the item which corresponds to the end node is accessed with the item of the start node by the same reader. For example, the sub-graph of the top in Figure 4

$$(19\ {}^{*}2961) \rightarrow (2\ 10851)$$

means that the number of the access to the item 2961 is 19, and two readers who read the item 2961 also read the item 10851. The initial nodes to construct the graph are decided

 ISBN: 978-1-61208-135-9

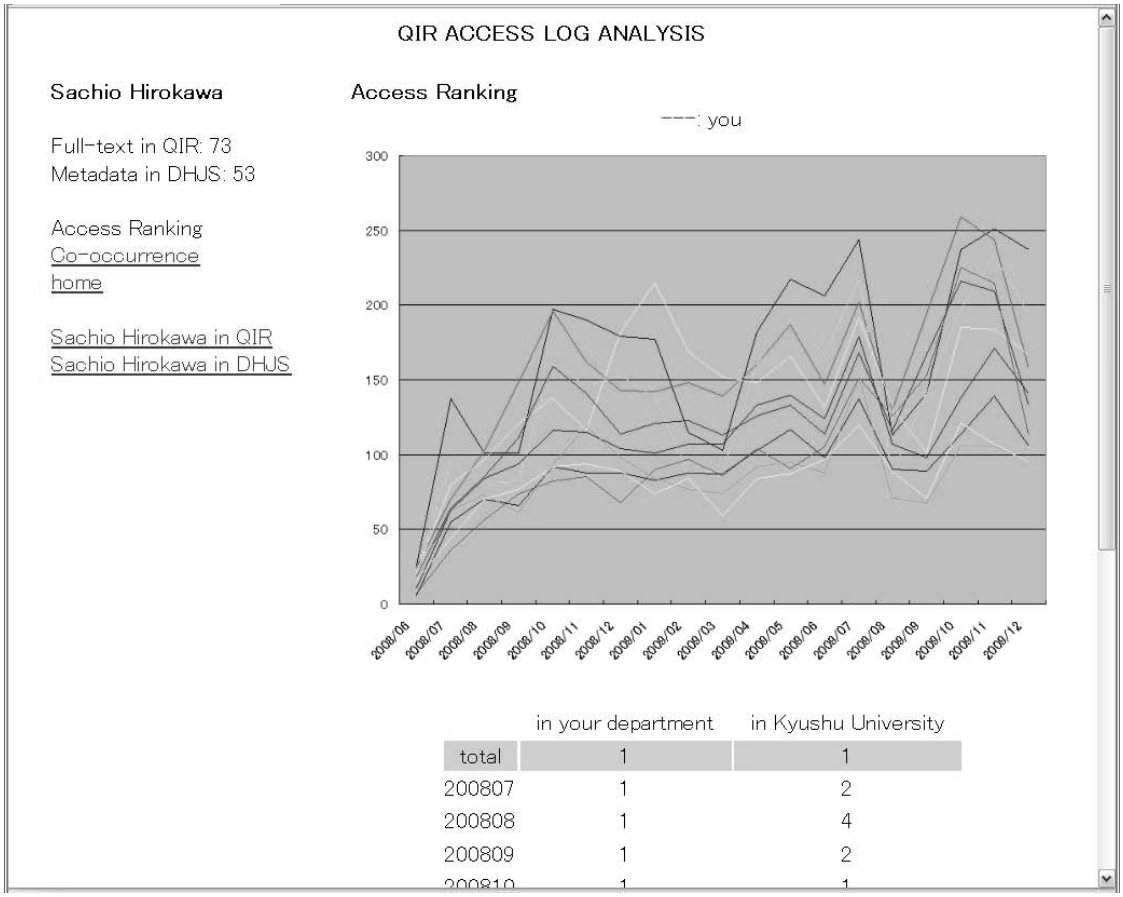

Figure 3. The result of the total number of the access to the items of an author and the ranking.

as the result of a search by a query, and the initial nodes have "∗" in the node.

By choosing some papers related to a research area as the initial papers for construction of the graph, this analysis might be able to find other papers of the area or a nontrivial relation between the area and other areas. As a consideration of the graph, we found that the shape of the graph tends to be classified roughly in two types: one is spreading to some nodes from an initial node (as the left-hand in Figure 5), and the other is making a line by some nodes (as the right-hand in Figure 5). Compared with the former, the latter is expected to be indicating a kind of typical papers in a research topic. On the other hand, the former is considered to be a result of access from results of search engines.

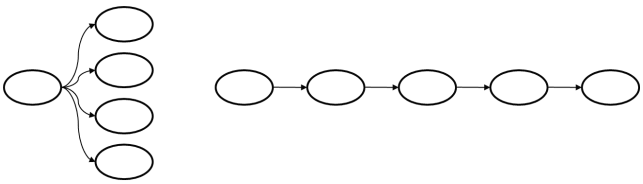

Figure 5. Two types of the shape of graphs for the co-occurrence analysis.

IV. Conclusion and Future Work

In this paper, we introduced a system which analyzes the access log of an institutional repository and returns the result to the authors as a feedback from the readers of their research outputs. The feedback system realizes an access control to the result of the analyses by connecting a researcher database. The main idea of the system, to connect a researcher database, is applicable to other research institutions.

One of our future work is the improvement of the user interface. In addition to the selection of the factors of the analysis, the layout shall be refined. Another one is the verification of the effectiveness of the feedback system. We are going to observe the number of the registration and access in the period from the implementation of the system to verify the effect of the system.

Acknowledgment

We thank the anonymous referees for their helpful comments to improve this work.

 ISBN: 978-1-61208-135-9

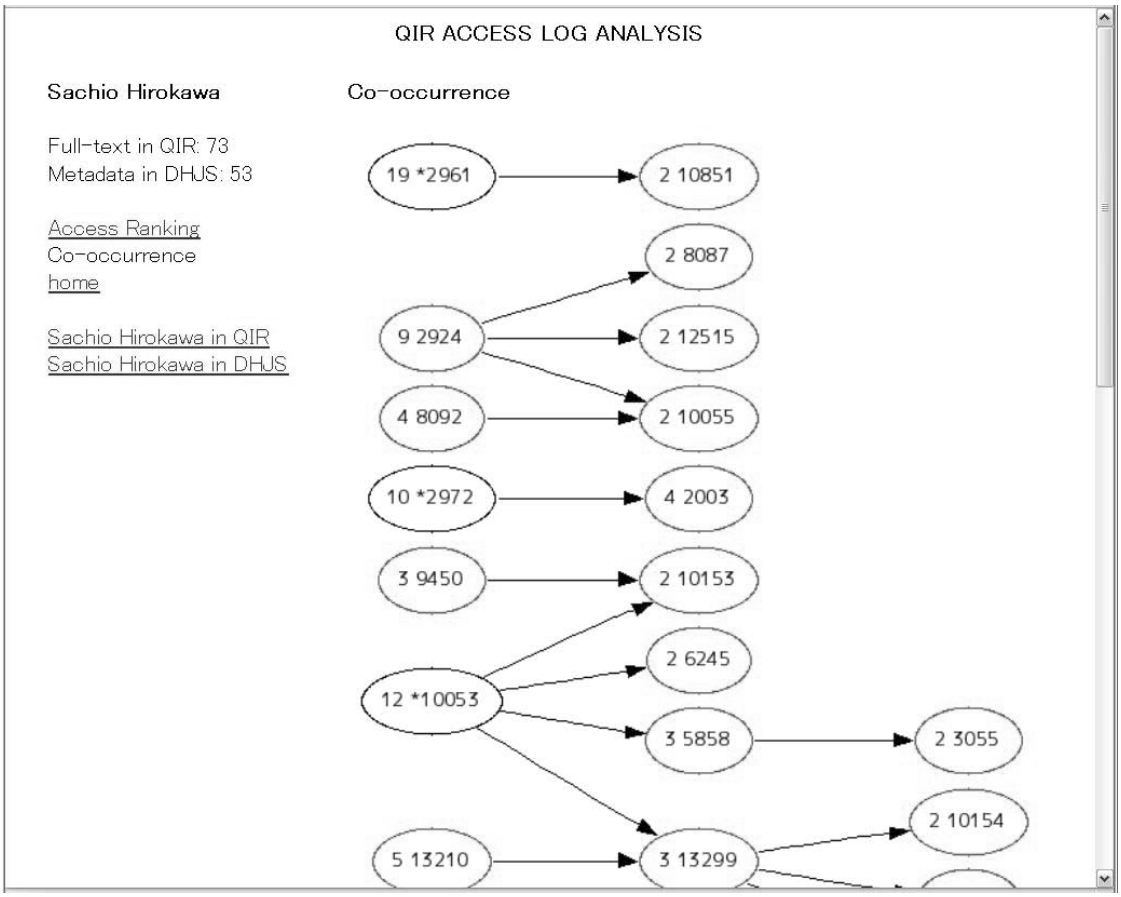

Figure 4. An example of the result of the co-occurrence analysis.

REFERENCES

[1] arXiv. http://arxiv.org/, [accessed 11 Mar, 2011].

[2] DHJS: Kyushu University Academic Staff Educational and Research Activities Database. http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html, [accessed 11 Mar, 2011].

[3] DSpace. http://www.dspace.org/, [accessed 11 Mar, 2011].

[4] Google Analytics. http://www.google.com/intl/en/analytics/, [accessed 11 Mar, 2011].

[5] QIR: Kyushu University Institutional Repository. https://qir.kyushu-u.ac.jp/dspace/, [accessed 11 Mar, 2011].

[6] Ranking Web of World Repositories. http://repositories.webometrics.info/, [accessed 11 Mar, 2011].

[7] Ranking Web of World Universities. http://www.webometrics.info/, [accessed 11 Mar, 2011].

[8] ROAR: Registry of Open Access Repositories. http://roar.eprints.org/, [accessed 11 Mar, 2011].

[9] 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 11 Mar, 2011].

[10] K. Baba, E. Ito, and S. Hirokawa. Co-occurrence analysis of access log of institutional repository. In *Proceedings of Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011)*, pages 25–29, 2011.

[11] 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. IARIA, 2010.

[12] K. Baba, M. Mori, and E. Ito. Identification of scholarly papers and authors. In *Proceedings of the Third International Conference on 'Networked Digital Technologies' (NDT 2011)*, 2011.

[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.

 ISBN: 978-1-61208-135-9

[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] P. Suber. Open access overview. Open Access News, 2007. http://www.earlham.edu/˜peters/fos/overview.htm, [accessed 11 Mar, 2011].

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

 ISBN: 978-1-61208-135-9