Identification of Scholarly Papers and Authors

Baba, Kensuke
Research and Development Division, Kyushu University Library

Mori, Masao
Institutional Research Office, Kyushu University

Ito, Eisuke
Research Institute for IT, Kyushu University

https://hdl.handle.net/2324/18910

出版情報：2011. Springer
バージョン：
権利関係：

Identification of Scholarly Papers and Authors

Kensuke Baba* Masao Mori Eisuke Ito

Abstract

Repositories are being popular as places for publication of research outputs. To make more efficient use of scholarly information on the internet, repositories are required to cooperate with other databases. One of the essential processes of the cooperation is identification of scholarly papers and their authors. The straightforward approach is string matching of the title and authors' name, however this approach cannot always solve the difficulties by basic clerical errors and same names. This paper proposes a method to compensate for the inaccuracy of the identification by connecting different databases. The main idea of the method is that different metadata of a scholarly paper is linked by the authors themselves, therefore the correspondence is guaranteed by the authors. The authors of this paper are developing a system based on the idea on the repository and the researcher database in their university.

Keywords: Institutional repository, Web database, identification of scholarly papers, identification of authors.

1 Introduction

The number of digital contents on the internet is rapidly increasing. Especially, for scholarly information, electronic journals and repositories [7] are being popular as places for publication of research outputs. The metadata of a scholarly paper (that is, the information about the title, the author(s), and so on) is usually registered in plural databases severally, and therefore the metadata has some variations. For some papers, in addition to the metadata, the full-text is archived in plural databases and it has some versions such as the pre-/post-print. The scholarly papers should be organized to make more efficient use for the users of the information.

In order to organize scholarly papers, it is not practical but ideal that an authority should manage all the papers. A feasible solution is cooperation of databases and advanced search functions thereby. For the solution, we have to make clear the relation on scholarly papers. The first step is "identification of scholarly papers", that is, to link the variations of the metadata of each paper. The versions of each paper can be managed by processing this step in detail. As the second step, one of the simplest organizations is classification with respect to the authors. The classification requires "identification of authors". As the result of these identifications, the metadata should have IDs which correspond to real papers and authors.

The straightforward approach of the identification of scholarly papers and authors is string matching by the title and authors' name. Some variations of the title can be identified by approximate string matching [6]. As for authors, the accuracy can be improved by matching of extra information such as the affiliation. However, this approach cannot always solve the difficulties by basic clerical errors and same names. Another approach of a different quality is confirmation by the authors themselves. For example, the problems are solved by adding unified IDs for scholarly papers and authors (such as DOI and the ID for membership of an association) to the metadata when the paper is registered. However, it is difficult to popularize unified IDs in advance, and moreover this solution cannot be applied to the papers which are already archived. The main idea of our solution is that the confirmation by the authors is realized by a cooperation of databases.

In this paper, we are trying to solve two problems in practical systems as a case study. Kyushu University has the researcher database DHJS (Kyushu University Academic Staff Educational and Research Activities Database, "Daigaku Hyoka Joho System" in Japanese) [1] and the repository QIR (Kyu(Q)shu University Institutional Repository) [2]. One of the problems we tackle is about identification of scholarly papers. DHJS has the metadata of scholarly papers which are produced by the researchers in the university. The number of the registered metadata is about 70,000, however it is estimated that at most about 20% is duplicate data. The other problem is about identification of authors. In QIR, a search of an author is operated by the naive string matching on the metadata, therefore the search cannot recognize any same name. The previous problems are solved by the following cooperation of the systems. By connecting the metadata in DHJS to the full-text in QIR,

- the first problem is solved since the identification of any paper is operated in QIR by handwork,

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

- the second problem is solved since a user authentication is required in DHJS for registration of metadata.

The number of the institutions who have own repository in the world is about 1,600 as of September 2011 [3], and most of the institutions are considered to have the same problem. In this paper, the situation of the practical systems in Kyushu University are shown in detail, and the problem and solution are described formally. Therefore, the proposed idea is applicable to other institutions.

2 Problem

This section describes the current situation of two databases, DHJS and QIR, and then formalize the problems we tackle.

2.1 DHJS

DHJS is the researcher database of Kyushu University. DHJS has various kinds of data about 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, and any researcher has a duty to register their research activities includes the metadata of scholarly papers into DHJS. 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. Fig. 1 is an example of the list of the metadata of scholarly papers in DHJS. The icons in the figure are mentioned in the following section.

The number of the metadata of papers registered in DHJS is about 70,000 as of January 2011. If a paper was written by plural authors in Kyushu University, the metadata of the paper might be registered by each authors severally. We practically estimated the ratio of duplicate data in DHJS by calculating the edit distance [8] between the titles of the papers in Kyushu University. Fig. 2 is the result of the calculation for a department with about 15,000 pieces of metadata. The horizontal axis shows the number of pairs and the vertical axis the edit distance which is formalized by the length of the longer title. The number of the pairs whose edit distance is less than 0.1 is about 3,000, that is, the number of the duplicate data is at most about 3,000 (and 1,500 if we assume that 4 pieces of metadata are registered for a single paper on average)[1]. Therefore, at most about 20% of the metadata are estimated to be duplicate. There was no significant difference of the ratio for every departments. By identification of the duplicate data, we can make more efficient use of the database, for example, we should be able to refer co-authors' site in DHJS from the metadata.

2.2 QIR

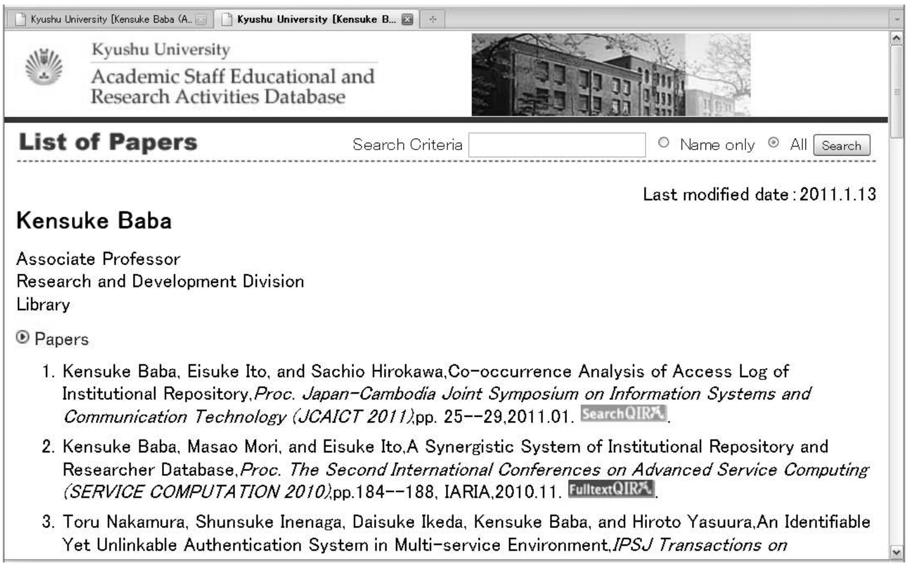
QIR is the institutional repository operated by Kyushu University Library. Generally, institutional repository archives the full-text of each paper in addition to its metadata. The total number of the items (papers, slides, and so on) in QIR is about 17,000 as of January 2011. The registration of items to QIR are operated by stuff in Kyushu University Library, and therefore the confliction of items are checked by handwork at the time. Fig. 3 is an example of the metadata of an item in QIR. The name of each author is linked to the profile page of the author, however the page is just the result of the naive string matching of the name for the items in QIR.

2.3 Formalization

The problems in the previous subsections are formalized in this subsection. Since we are focusing scholarly papers and their authors, we consider the set P of the real scholarly papers and the set A of the real authors. We define the *metadata* of a scholarly paper to be a pair of a string (the *title*) and a non-empty set of strings (the *author(s)*). The problems are defined to be, for a given set M of metadata, to find functions $f : M \to P$ and $g : M \to 2^A$, respectively. In other words, it is to put indexes in P and A on the title and the author(s) of each metadata.

In this paper, the input sets of metadata are M_D and M_Q for the metadata in DHJS and QIR, respectively. Then, the following assumptions are given for the problem by the current situations in the previous subsections. As mentioned in Subsection 2.1, the metadata in DHJS is registered by one of the authors and the registration requires a user authentication. Therefore, it is regarded that we have the function $g_1 : M_D \to A$ such that $g_1(d) \in g(d)$ for any $d \in M_D$. As to Subsection 2.2, the correspondence between the metadata in QIR and the full-text is guaranteed by the check of the stuff in Kyushu University Library. Therefore, we have the function $f_1 : M_Q \to P$ such that $f_1(q) = f(q)$ for any $q \in M_Q$.

[1] The number of the duplicate data is defined to be the gap between the number of the metadata and the number of the distinct papers. If we assume that the duplicate data is made by n authors for each paper, then ${}_nC_2$ pairs are counted for each paper and the number of the duplicate data is $n - 1$ for each paper. Therefore, the number of the duplicate data is $2m/n$ for the number m of the counted pairs.


Kyushu University [Kensuke Baba (A...
Kyushu University [Kensuke B...
Search

Kyushu University
Academic Staff Educational and
Research Activities Database

List of Papers Search Criteria ○ Name only ● All

Last modified date : 2011.1.13

Kensuke Baba

Associate Professor
Research and Development Division
Library

Papers

1. Kensuke Baba, Eisuke Ito, and Sachio Hirokawa,Co-occurrence Analysis of Access Log of Institutional Repository,*Proc. Japan-Cambodia Joint Symposium on Information Systems and Communication Technology (JCAICT 2011)*,pp. 25--29,2011.01. SearchQIR.
2. Kensuke Baba, Masao Mori, and Eisuke Ito,A Synergistic System of Institutional Repository and Researcher Database,*Proc. The Second International Conferences on Advanced Service Computing (SERVICE COMPUTATION 2010)*,pp.184--188, IARIA,2010.11. FulltextQIR.
3. Toru Nakamura, Shunsuke Inenaga, Daisuke Ikeda, Kensuke Baba, and Hiroto Yasuura,An Identifiable Yet Unlinkable Authentication System in Multi-service Environment,*IPSJ Transactions on*

Figure 1: The Web image of the list of scholarly papers in DHJS.

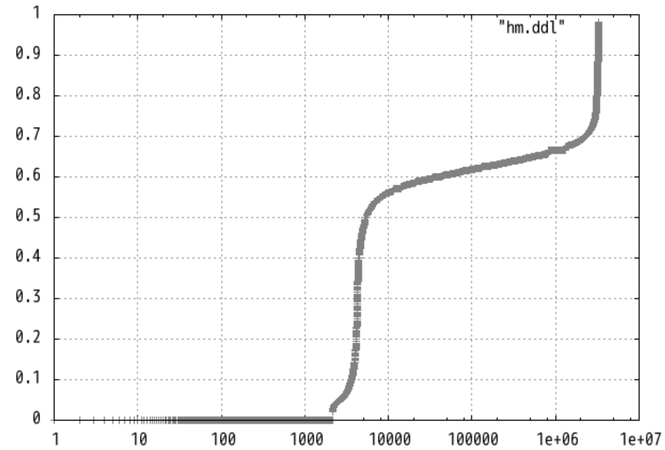


Figure 2: The edit distances between the all possible pairs of the 14,599 titles in DHJS for a department.

3 Solution

3.1 Main Idea

The main idea of our solution is, in terms of the formalization in Subsection 2.3, to find the function $h : M_D \to M_Q$ such that

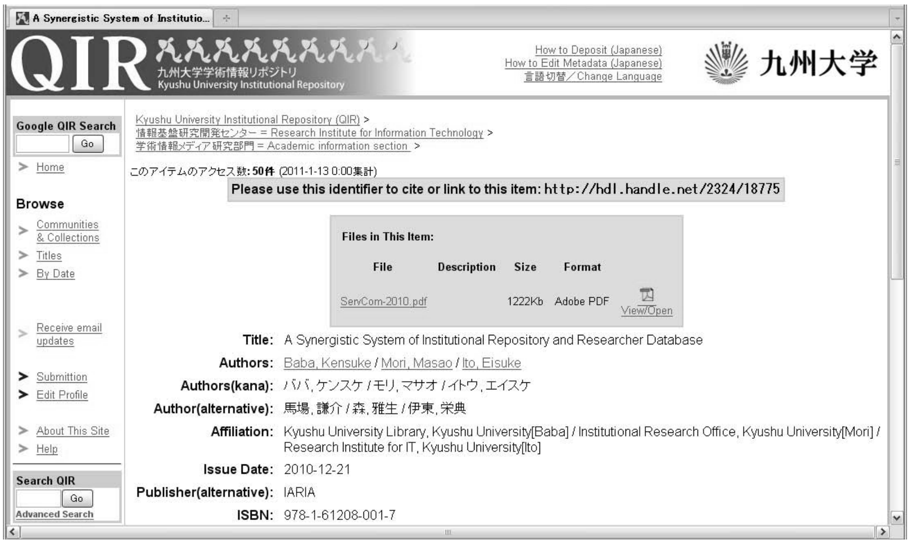
- $f_1(h(d)) = f(d)$ for any $d \in M_D$,
- $g_1(h^{-1}(q)) = g(q)$ for any $q \in M_Q$, where $h^{-1}(q) = \{d \in M_D \mid h(d) = q\}$ and $g_1(S) = \{g_1(s) \mid s \in S\}$ for a set $S \subseteq M_D$, and
- $g(q) = g(d)$ for any $d \in h^{-1}(q)$.

This situation is illustrated in Fig. 4. In this example, $P = \{p_1, p_2\}$, $A = \{a_1, a_2, a_3\}$, $M_Q = \{q_1, q_2\}$, and $M_D = \{d_1, d_2, d_3\}$. The functions f and g are as the following table.

x	q_1	q_2	d_1	d_2	d_3
$f(x)$	p_1	p_2	p_1	p_2	p_2
$g(x)$	$\{a_1\}$	$\{a_2, a_3\}$	$\{a_1\}$	$\{a_2, a_3\}$	$\{a_2, a_3\}$

As the assumptions, we have the functions f_1 and g_1 such that $f_1(q_i) = p_i$ for $i = 1, 2$ and $g_1(d_i) = a_i$ for $i = 1, 2, 3$. Then, h should be $h(d_1) = q_1$, $h(d_2) = q_2$, and $h(d_3) = q_2$. Therefore, by h, we have $h \circ f_1$ and $h^{-1} \circ g_1$ such that $h \circ f_1(d_1) = p_1$, $h \circ f_1(d_2) = p_2$, $h \circ f_1(d_3) = p_2$, $h^{-1} \circ g_1(q_1) = \{a_1\}$, and $h^{-1} \circ g_1(q_2) = \{a_2, a_3\}$. Additionally, we have $h \circ h^{-1} \circ g_1$ such that $h \circ h^{-1} \circ g_1(d_1) = \{a_1\}$, $h \circ h^{-1} \circ g_1(d_2) = \{a_2, a_3\}$, and $h \circ h^{-1} \circ g_1(d_3) = \{a_2, a_3\}$.

In the sense of the practical system, the first condition is clearly realized by linking the metadata of a paper in DHJS to the metadata of the paper in QIR. The second one is realized by the previous link if any author of the papers in QIR register the metadata of



A Synergistic System of Institutio...

QIR 九州大学学術情報リポジトリ Kyushu University Institutional Repository

How to Deposit (Japanese)
How to Edit Metadata (Japanese)
言語切替／Change Language

九州大学

Google QIR Search [Go]

> Home

Browse

> Communities & Collections
> Titles
> By Date

> Receive email updates

> Submittion
> Edit Profile

> About This Site
> Help

Search QIR [Go]
Advanced Search

Kyushu University Institutional Repository (QIR) >
情報基盤研究開発センター = Research Institute for Information Technology >
学術情報メディア研究部門 = Academic information section >

このアイテムのアクセス数: **50件** (2011-1-13 0:00集計)

Please use this identifier to cite or link to this item: http://hdl.handle.net/2324/18775

Files in This Item:

File	Description	Size	Format	
ServCom-2010.pdf		1222Kb	Adobe PDF	View/Open

Title:	A Synergistic System of Institutional Repository and Researcher Database
Authors:	Baba, Kensuke / Mori, Masao / Ito, Eisuke
Authors(kana):	ババ, ケンスケ / モリ, マサオ / イトウ, エイスケ
Author(alternative):	馬場, 謙介 / 森, 雅生 / 伊東, 栄典
Affiliation:	Kyushu University Library, Kyushu University[Baba] / Institutional Research Office, Kyushu University[Mori] / Research Institute for IT, Kyushu University[Ito]
Issue Date:	2010-12-21
Publisher(alternative):	IARIA
ISBN:	978-1-61208-001-7

Figure 3: The Web image of the metadata of an item in QIR.

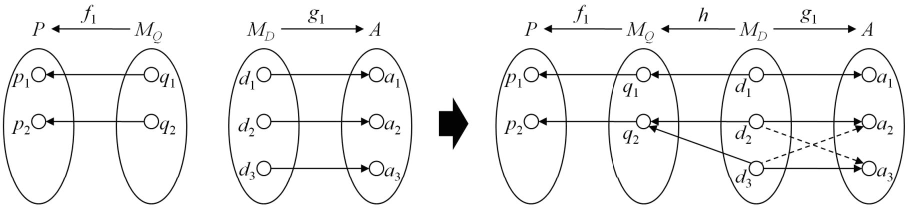


Figure 4: The relation between f_1, g_1, and h for an example (P, A, M_Q, M_D). By h, we can compensate the relation of the dotted arrows.

the papers. It is also clear that the last one is realized by the link since the author(s) of a paper in DHJS is same as the paper in QIR.

3.2 Implementation

We have already developed a system which links the metadata in DHJS to the full-text in QIR [5]. In Fig. 1, the colored icon "fulltextQIR" is connected to the corresponding full-text in QIR. Researchers put icons on the list in the data-entry system in DHJS, and link them to the full-text by themselves. The other icon "SearchQIR" means that the metadata is not linked yet. Therefore, the correspondence between the metadata in DHJS and the metadata in QIR is guaranteed by check of the author instead of string matching. Namely, this link system realizes the function h. The ID of any paper in QIR is attached to the metadata in DHJS by this link. Additionally, since the metadata in DHJS has the ID of the author who registered the metadata, we can put the ID to the corresponding metadata in QIR by the link. By returning the author IDs from QIR to DHJS after the IDs for all the authors are attached, also the metadata in DHJS can have the IDs of the authors.

One of the problems in the implementation is that the number of the metadata (full-text) in QIR is small compared with the number of metadata in DHJS. As mentioned in Section 2, the number of the metadata in QIR is 17,000 while the number in DHJS is at least about 56,000. To verify the effectiveness of our solution, the number of metadata in QIR is required to be large to complete the correspondence with the metadata in DHJS. We are developing a system to encourage researchers to register their papers to QIR by showing the result of access log in QIR [4].

4 Conclusion and Future Work

A method to compensate for the inaccuracy of identification of scholarly papers and authors on the metadata in separated databases was proposed. We formalized the problem in the practical systems and proposed the solution in terms of the formalization. Moreover, we showed the outline of the implementation based on the idea of the proposed solution.

One of our future work is the implementation of our solution along with the outline. Then, we are going to examine the accuracy of the identification, and observe the number of the access and the registered papers of

QIR.

References

[1] Kyushu University Academic Staff Educational and Research Activities Database. http://hyoka.ofc.kyushu-u.ac.jp/search/index_e.html, [accessed 2010-11-02].

[2] QIR: Kyushu University Institutional Repository. https://qir.kyushu-u.ac.jp/dspace/, [accessed 2010-11-02].

[3] ROAR: Registry of Open Access Repositories. http://roar.eprints.org/, [accessed 2010-11-02].

[4] 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.

[5] 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.

[6] M. Crochemore and W. Rytter. *Text Algorithms*. Oxford University Press, 1994.

[7] P. Suber. Open access overview. Open Access News, 2007. http://www.earlham.edu/˜peters/fos/overview.htm, [accessed 2010-11-02].

[8] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1):168–173, 1974.