Name Disambiguation of Japanese Researchers: A Case Study with Statistics Research Community

Masao Takaku[1] Yasumasa Baba[2] Akiko Aizawa[3]

[1] Science Information Office, National Institute for Material Science
E-mail: TAKAKU.Masao@nims.go.jp
[2] Transdisciplinary Research Integration Center, Research Organization of Information and Systems
E-mail: baba@ism.ac.jp
[3] National Institute of Informatics
E-mail: aizawa@nii.ac.jp

Keywords: Name disambiguation, record linkage

1 Introduction

In the integration of large-scale database, record entity matching is essential task. In particular, entities having the same personal name appears in different databases, but those actually represent different entities. Those kinds of records need a human intervention to interpret whether they are the same entity or not. It takes much cost for large-scale databases. For example, social number can identify a person uniquely. In the case of lack of such identification methods, we will be only able to use literal expression of his/her name and some of the attributions.

In this paper, we report the case study in record matching of researcher names. We construct a sample dataset of Japanese researchers in the field of statistics, and then manually identify their names with *Kakenhi* database which includes approximately 15,000 researchers in Japan.

2 Dataset

We used *Kakenhi* database as a large-scale researcher database, and collected member lists of three statistics related societies in Japan. The Grants-in-aid for Scientific Research (*Kakenhi*) is one of the largest research funds in Japan. In this paper, we used the database records from 1982 to 2004 in Kakenhi DB, and the database included approximately 150,000 unique researchers in Japan.

Our target dataset for disambiguation of researcher names is member list of statistics societies in Japan, shown in Fig. 1.

Table 1: Dataset for disambiguation of researcher names: member lists of statistics societies in Japan

society	number of persons	matched with Kakenhi DB
Japan Statistical Society (as of 2003)	1,545	697 (45%)
Behaviormetric Society of Japan (as of 2005)	1,070	519 (49%)
Japanese Society of Computational Statistics (as of 2002)	416	184 (44%)
(Total)	3,031	1,400 (46%)

Nearly half of researcher's names in those societies were appeared in Kakenhi DB. And then we classified their correspondence by manual.

3 Results

Figure 2 shows result of disambiguation among researcher records in Kakenhi DB and society members. We will discuss further details at the conference session.

Table 2: Disambiguation of researcher records between Kakenhi DB and statistics member lists

appeared in DB	number of researchers	needs for disambiguation or not	
not appeared	126 (11.1%)	-	
appeared	1,008 (88.9%)	(no records with the same name)	903 (89.6%)
		(having records with the same name)	105 (10.4%)