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

Recently, the idea of an entity-based search engine has become popular beyond a usual search engine, where its target is just looking for a topically relevant page or site. An entity-based search engine aims at identifying some sorts of entities, and treats them as a first target for its search collection. In order to provide an entity-based search engine, we need not only just to retrieve a relevant page or site, but also to specify what an entity is and to identify the entity. And then, some of the domain knowledge for the entity is required to be obtained. Additionally, an identification method for the entity has to be established. The authors have attempted to develop such identification methodologies for searching scientists in academia.

For person identification, we need to obtain a reference point for information on people. There are two options for that. One is to make use of a kind of a list of identified persons. The other one is to (semi-)automatically identify persons with machine learning techniques. In this paper, we take the former approach since we are able to take advantage of a personal identification database built by manual. Note that a name of a person is not always sufficient for identifying that person, because it is possible to have the same name on different persons. The database we used has a unique key for each person, not on his/her name, so we can use it as a reference point for identifying that person.

We used KAKEN-DB[1], which is provided by National Institute of Informatics, and identified Japanese academic scientists in it by using unique KAKEN-ID. In this paper, we will report results on identification of Japanese scientists of statistics-related field and discuss its application for that dataset.

2 A Dataset on Japanese Scientists of Statistics

We used member lists of three Japanese societies in statistics-related field: Japan Society of Statistics (JSS); Behavior Statistics of Japan (BSJ); and Japanese Computational Statistics Society (JCSS). These member lists were taken of members as of 2003 for JSS, 2005 for BSJ, and 2002 for JCSS.

Table 1: Dataset on Japanese Scientists of statistics-related field

Society	Number of members
JSS (2003)	1,545
BSJ (2005)	1,070
JCSS (2002)	416

Table 1 shows the number of members in each society. For identification process, first, we extracted candidate persons for identification from these members by matching their names with records in KAKEN-DB. We got 1,400 records of candidates for Statistics scientists, where 697 (45%) members of JSS, 519 (49%) of BSJ, and 184 (44%) of JCSS. Approximately 45–50% of the members in the dataset could be a candidate for identification with KAKEN-DB. Then, the duplicate records in the candidate list were excluded by manual. We got 1,134 researchers as candidates for identification.

[1] http://seika.nii.ac.jp/

3 Identification for researchers

For the identification process, we filtered researchers by matching their names and treated the matched researchers as candidates, and then confirmed by manual whether each pair of records is the same person or not. Note that we treated a different person when each of records does not have the same name in their name fields.

We built a simple tool for identification, and used it for manual processing (shown in Fig.1).

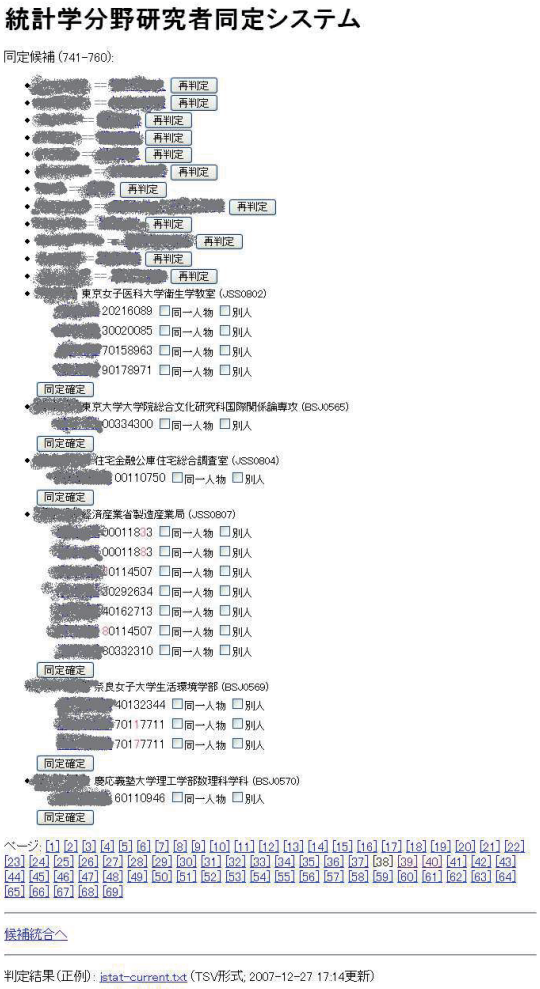

Figure 1: UI of the identification tool

Figure 1 shows our identification tool. In the tool, twenty candidate researchers is displayed at one time, and can be identified whether each person corresponds to a candidate or not, by submitting the decision through the checkbox UI. Information on each researcher is displayed in a list. Ex. his/her name, affiliation, and society which belongs to that person. Some of the researcher names matched toward several records of KAKEN-DB. In such cases, we need name disambiguation. When disambiguation needed, the names of multiple records were shown on a list of candidates. Identification of a researcher was recorded with KAKEN-ID and the researcher's unique key on the member list by submitting through checkboxes whether he/she could be regarded the same person or not. Additionally, the tool supports to detect and merge duplicate records among the member lists.

4 Result

As a result of identification, the candidates were classified into three categories: marked as duplicates, positive samples, and negative samples. The positive sample means a candidate who can be identified

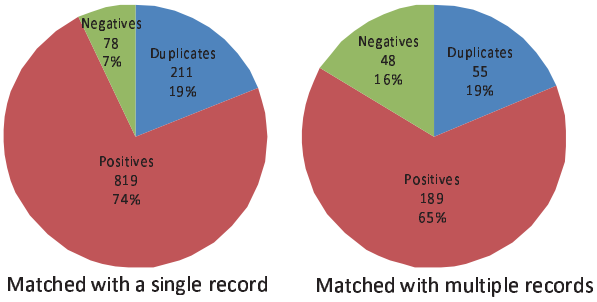

Figure 2: Result of identification for candidates matched for a single record and for multiple records

for one or more records in KAKEN-DB. And the negative sample means opposite, where a candidate identified to no records in KAKEN-DB. We identified 266 (19%) candidates as duplicates, 1008 (72%) candidates as positives, and 126 (9%) candidates as negatives.

Figure 2 shows detailed results for each fraction of candidates who matched multiple records and who matched a single record in KAKEN-DB. Both of the types of candidates show similar trends. Positive candidates were 90.8% out of the single matched candidates, and 83.1% out of the multiple matched candidates.

For the multiple records matched candidates, 82 candidates had at least two or more records with the same name, but different persons. For the single matched candidates, 130 candidates out of all (15.8%) including negative samples had at least one or more records with the same name, but different person.

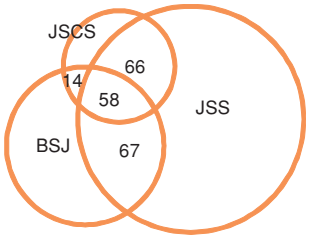

Figure 3: Duplicate members among societies

Figure 3 shows the number of the researchers which belong to each society. There are 58 researchers joined all three societies, 66 researchers for JSS and JCSS, 67 researchers for BSJ and JSS, and 14 researchers for BSJ and JCSS.

5 Discussion

5.1 Criteria on identification

For the identification process, we need a criteria on whether a pair of personal information pointed to the same person or not, whatever the process made by manual or semi-automatically. For the results reported here, we used information on name and affiliation included in the both sources. So, we were able to use the affiliation information as primary information for identification. And then since the field of the researchers was also limited to the statistics-related, we were able to make decision based on the researcher's project information within KAKEN-DB. By making use of these information, we can easily identify a pair of information if their names and affiliations are the same exactly and the field of the researches in KAKEN-DB is the statistics related one. On the other hand, if this information does not match exactly and/or the project information in KAKEN-DB is not sufficient, the identification could be difficult to judge. For that case, we looked for further information on the Web. There are a several resources, ReaD researcher database[2] provided by JST, the official site of the person's affiliation, and then we can find information through general search engines. In some cases, we used an author profile page on

[2] http://read.jst.go.jp/

a publisher site which published the book/article written by that researcher. And in other some cases, we are able to find person information by searching web search engine with the keyword of the name of the person and the affiliation, and then in several cases, we could find the person moved his/her position to another affiliation. For disambiguating a candidate with different researchers, we searched ReaD database if they appeared as different entries, and we checked their affiliation histories on the database. In these identification processes, a candidate matched with a single record can be easily identified, but it is not the case for multiple matched candidates. It takes time for confirming whether the multiple records matched with the candidate are different one or not. The problem is not only for time to confirm, but also criteria on what needed for the identification. In our case, we mainly used web resources, but information on the Web is not always accurate and reliable, so we have to balance between precise identification and costs for identification.

The trade-off between reliability and costs of identification should be considered with the purpose of identification. Additionally, the identification process takes a few minute per candidate for easily identifiable case, but for more difficult one it takes much time, roughly 30–60 minutes per candidate. For larger scale identification, it is needed to estimate time to do identification in total.

5.2 Visualization of co-researcher network

The dataset we used is subject to statistics related field, so with a visualization technique we are possible to illustrate some characteristics of community of Japanese statistics related researchers.

Figure 4 shows a co-researcher network which illustrates how researchers have conducted their activity under the KAKEN research funds. The figure was created by GraphViz[3], a network visualization tool. The nodes in the Fig.4 represent each researcher and those researchers are placed as long as his/her co-researchers placed together with weight of their co-occurrence within the research projects in the past. The color of nodes represent the researcher's society. The red nodes means the member of JSS, blue one for JSCS, and green for BSJ. And duplicate members are represented with the additive complementary colors. Ex. purple for JSS and JSCS, light blue for BSJ and JSCS, yellow for JSCS and JSS, and black for all three societies. From our prior analysis, the co-researcher network in KAKEN-DB is generally dense and widely connected, and Fig.4 shows very similar trends even in the field of statistics.

This kind of visualization takes advantage into two applications. One is for contribution to the identification process. As described later, the identification process itself is possible to use some of semi-automatic techniques. Our visualization of co-researcher network seems to be treated as a feature for automating the process. In such kind of automatic identification, a densely connected node, which is a sort of "core researcher", can be estimated with a high possibility to be identified with other dataset. In other words, a core researcher in a field-specific network is estimated as an identifiable one than others. Once some of candidates are identified, network measurements like centeredness or connectivity for each node can be calculated and used as a feature for identification of future process. This remains as a future work.

The other application of visualization is used for understanding of the field itself. The visualization indicates a possibility to illustrate a key person on that field. Hook and Böner (2005) proposed such kind of techniques as "knowledge domain visualization", and it is able to help learners to understand its structure and knowledge well.

5.3 Towards automatic identification

As described in Section 4, most of the candidates in our dataset were positive one, and it accounts for approximately 90% of the candidates after excluding duplicates. From this result, some sorts of identification could be done by (semi-)automatically way. Especially in the case that both of name and affiliation matched exactly, most of all was positive one. Additionally, most of candidates who had the same name within the dataset were simply judged as duplicates, except only two cases. Since we have got the recorded positive and negative samples from our identification process, those information could be usable for machine learning techniques. This also remains as future work.

[3] http://www.graphviz.org/

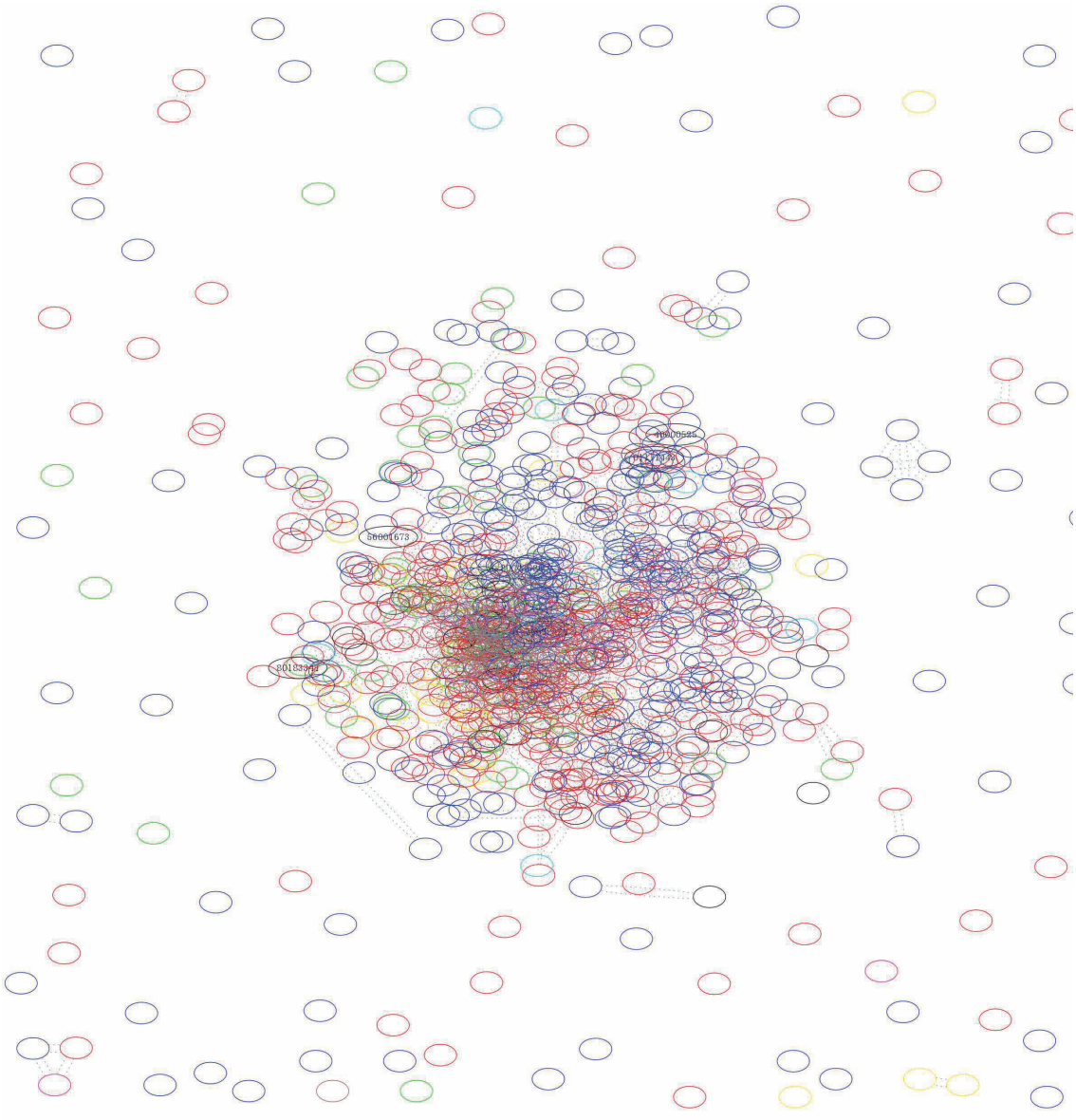

Figure 4: Co-researcher network of Japanese statistics related researchers (among societies)

6 Conclusion

In this paper, we reported on identification results of researchers from Japanese statistics-related societies, and its insight into further case. Identification of Web site/page related to a person as well as some of automation and visualization issues are remained as future work.

References

Hook, P. and Börner, K. (2005). Educational knowledge domain visualizations: tools to navigate, understand, and internalize the structure of scholarly knowledge and expertise. Spink, A. and Cole, C. (eds), *New Directions in Cognitive Information Retrieval*, 187–208, Springer.