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

Masao Takaku†, Yasumasa Baba‡‡, Akiko Aizawa†††

† National Institute for Materials Science
‡‡ Research Organization of Information and Systems / The Institute of Statistical Mathematics
††† National Institute of Informatics
Motivation

• There are many types of information on the Web
• Entity-based search
  – Categories for types of information
  – Events, place, person, etc.
• Searching for a person
  – Not sufficient to search Google with the name of the person
  – E.g. Expert search
  – Through the service of
    • SNS
    • Web search engine
Objective

• Name identification
  – We integrated persons from different resources.
  – For identifying a different person with the same name, it is necessary to use several resources.
  – Estimating the costs

• Researcher’s information:
  – KAKEN Researcher ID (for researchers in Japan) → We developed “Researcher Information Server”.
• We have made a dataset for statistics related researchers.
  – We intend to help evaluating and developing identification data and search system for researchers.
  – Data: Member lists of three academic associates for statistics related field in Japan
Researcher Information Server

- **Data taken from KAKEN DB**
  - # of projects (Funded by KAKEN): 247,745
    - Annual reports of the projects from 1989 to 2004
  - # of researchers (project leader and co-researchers): 133,067
    - Individual researcher information

- **Basic information for researcher**
  - Researcher name, affiliation name, and position (job title).

- **Visualization**
  - Timeline: # of KAKEN projects and publications
  - Co-researcher network:
    - Past KAKEN project’s co-researcher
    - Using Google Maps
Statistics related field researchers (in three academic societies)

- We used member lists of academic societies (total: 3,031)
  - The Japan Statistical Society (2003): 1,545
  - The Behaviormetric Society of Japan (2005): 1,070

- Registered information: name, affiliation name, job title, address, zip-code
  - Only name and affiliation name were used.
Identification (Step 1)  
Name matching

• Matching the researcher with the same name researcher in Research Information Server (KAKEN-DB)
  – A total of 1,400 candidates:
    • The Japan Statistical Society : 697 (45%)
    • The Behaviormetric Society of Japan : 519 (49%)
    • The Japanese Society of Computational Statistics: 184 (44%)
  – After removing duplicates: 1,307 candidates
    • Candidates of having exactly the same name and affiliation were considered as an identical researcher.

• We manually identify the following data to KAKEN researchers ID.
### Identification (Step 2)

**Manual judgments**

- Adding a KAKEN ID for each member manually.

<table>
<thead>
<tr>
<th>Members of societies</th>
<th>KAKEN IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSJ0004</td>
<td>20024581</td>
</tr>
<tr>
<td>BSJ0005</td>
<td>50305313</td>
</tr>
<tr>
<td>JSS0012</td>
<td>70303047 09246528</td>
</tr>
<tr>
<td>BSJ0007</td>
<td>90184332</td>
</tr>
<tr>
<td>JSS0014</td>
<td>90184332</td>
</tr>
<tr>
<td>JSS0015</td>
<td>90132696</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Members of societies</th>
<th>KAKEN IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>BSJ0004</td>
<td>Different</td>
</tr>
<tr>
<td>BSJ0005</td>
<td>Identical</td>
</tr>
<tr>
<td>JSS0012</td>
<td>Different</td>
</tr>
<tr>
<td>BSJ0007</td>
<td>Identical</td>
</tr>
<tr>
<td>JSS0014</td>
<td>Identical</td>
</tr>
<tr>
<td>JSS0015</td>
<td>Different</td>
</tr>
</tbody>
</table>

A member of multiple societies

Judged by manual
Judgment system for identification
Results

- 266 candidates were marked as “duplicates” (members of multiple societies)
- 1008 candidates were identified with a researcher within KAKEN-DB.
Researchers having the same name

• (a) Within KAKEN-DB (for positives)
  – A member’s name matched with KAKEN ID, and a different researcher having the same names within KAKEN-DB.
  → 105 candidates (10%)

• (b) Concerning outside of KAKEN-DB (for negatives)
  – A member’s name matched with KAKEN ID, but that represented a different researchers.
  → 126 candidates

• (a) + (b) → 231 candidates (20%)
Duplicates (overlaps among the Japanese academic societies in Statistics-related field)

- JSCS: 46
- BSJ: 58
- JSS: 66
- Total: 504
Discussion: Criteria for identification

• How to determine two researchers are identical or not?
  – Researcher Information Server
    • Research area
    • Affiliations
    • Timeline based info.
  – Other resources:
    • JST ReaD (researcher directory)
    • On the Web
      – Organization website the researcher affiliates
      – Homepage provided by the researcher him/herself
  – Biography information provided as author introduction in his/her publications

• For statistics related field, left-side info are useful.
  – (Of course) His/her research fields are related to Statistics. It is a main hints for

• But, Sometime there is a difficult case to determine
  – By lack of information
  – It depends on the available resources and costs.
    • Need to clarify to utilize after the identification

• (Another possibility) Saving criteria and process for identification for the further analysis (afterwards).
Discussion: Visualizations

• Visualization of co-researchers network
  – Provides summary of researcher’s community.
  – Connects researchers with having joint-research in the past.
  – Shows one more hop into the other researchers.

• Applications (It will be helpful for …)
  – Judgment of Identification
  – Summarization of researchers network (community)
Visualization

• JSS
• BSJ
• JSCS
Conclusions

• Identification of Japanese researchers in the field of Statistics
  – Identification between academic societies and KAKEN-DB
  – Identification of about 3000 Statistics-related researchers with about 1,300,000 database leads to about 1000 identified persons.

• Future works
  – Semi-automatic process for identification (by using machine learning techniques)
  – Another dataset and resources
    • Other databases, transcriptions, and so on.
  – Comparison with other scientific fields
  – Represent a domain knowledge from researcher community
  – Entity-based search engine by identification results