# Development of the Method for the Appropriate Selection of the Successor by Applying Metadata to the Standardization Reports and Members

Isaac Okada[1], Minoru Saito[1], Yoshiaki Oida[1], Hiroyuki Yamato[2], Kazuo Hiekata[2], and Shinya Miura[3]

[1] System Engineering Knowledge Improvement div., System Engineering Technology Unit, Fujitsu Limited., Tokyo, Japan
{isaac-okada,saito.minoru,oida.yoshiaki}@jp.fujitsu.com
[2] Graduate School of Frontier Sciences, The University of Tokyo, Chiba, Japan
{yamato,hiekata}@k.u-tokyo.ac.jp
[3] Faculty of Engineering, The University of Tokyo, Tokyo, Japan
miura@is.k.u-tokyo.ac.jp

**Abstract.** In businesses and organizations, it is difficult to find the successor for various activities by considering a person's knowledge and actual experience. In this study, we find the successor to a member of a standardization activity. By assigning metadata to profiles and annual activity reports of members engaged in standardization activities, the relationship between the profiles and the annual activity reports is described as an RDF graph and visualized with nodes and links. This paper has two objectives. Objective-1 is the development and evaluation of a method to design the best combination of search queries to discover an appropriate successor. Objective-2 is the proposal and evaluation of an easy and understandable visualization method of the successor search results obtained in objective-1. The proposed procedure nominates candidates for the successor effectively and the results are visualized in the case study.

**Keywords:** metadata, RDF, semantic technology.

## 1 Introduction

Many companies have a huge and growing amount of resources in databases (DBs). In general, each resource is managed individually in accordance with the resource type (e.g., persons, goods, documents), but the relationships among such resources are not managed.

In official standardization activities performed by Fujitsu Limited for ISO, IEC, IEEE and etc., the “Name List DB” listing the members engaged in official standardization activities and the “Annual Activity Report DB” made by the standardization activity members have been individually accumulated and managed.

Some studies have covered the linkage among different data sources. Kashima [1] addressed link prediction by analyzing a network structure (i.e., link mining), and Matsumura et al. [2] addressed information utilization by linking museum information

H. Takeda et al. (Eds.): JIST 2012, LNCS 7774, pp. 255–266, 2013.

and local community information. Also, “VIVO” (http://vivoweb.org) visualizes relationships of researchers. Our research differs from these studies in that we aim to obtain effective results from the actual data in companies by focusing on limited business needs.

When choosing a successor for a standardization activity, the selected successor typically has knowledge, experiences, and a human network similar to or common to the predecessor. If we are able to support this member selection process through a rational and effective way by organizing the relationships among DB data, we can enhance the successor’s appropriateness, validity, and adequacy in an objective manner. In this paper, we have two objectives. Objective-1 is the development and evaluation of a method to design the best combination of search queries to discover an appropriate successor. Objective-2 is the proposal and evaluation of an easy and understandable visualization method of the successor search results obtained in objective-1.

## 2 Proposed Methods

### 2.1 Overview

To find effective methods to create superior search queries, resources are linked to each other using metadata technology and the relationship is visualized (Figure 1).

The proposed methods are based on the assumption that the “visualization” of linkages among resource elements will facilitate our searching and identifying similar relationships from the resources. Specifically, the proposed methods assign metadata to the two resources, “Name List DB” for the official standardization activities (persons) and “Activity Report DB” (documents), and "visualize" the relationship by which the same metadata link various DBs.

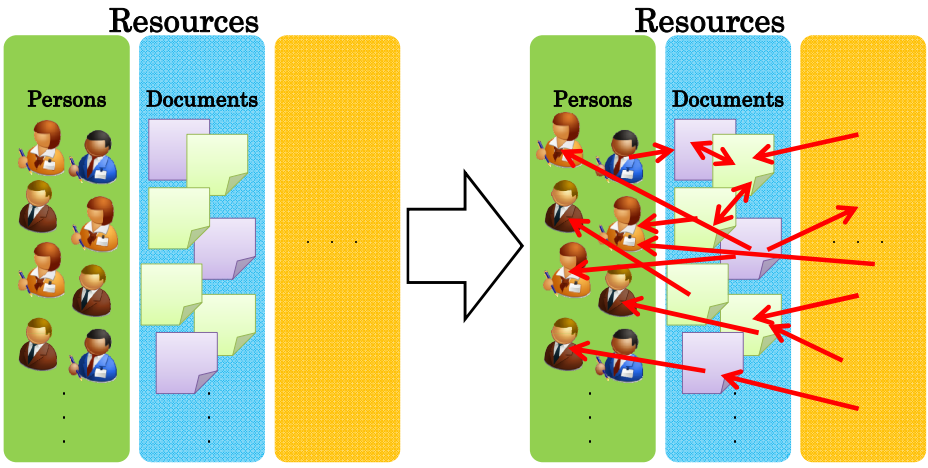


**Fig. 1.** Linking among resources using the metadata technology

## 2.2 Verification of Targeted Sample Data

This verification targeted the sample data of the following two DBs.

(1) "Name List DB" for members engaged in the official standardization activity (persons):

DB Content: Name, Department, Standardization activity field, Organization, and other information

(2) "Annual Official Standardization Activity Report DB" (documents):

DB Content: Documents (Word/PDF; Writer Name, Organization, and other information)

Conventionally, both the physical location and the management of these DBs have been independently controlled, without coordination between them.

(Note: In this research, the real names of persons were converted into aliases.)

## 2.3 Metadata Design

Metadata were designed by using "FOAF" (The Friend of a Friend project: http://www.foaf-project.org) for the Name List (persons), and "Dublin Core" for the Annual Activity Report (documents). Metadata unique to this system were additionally designed as "fkw". The unique metadata included "fkw:belongsTo" (standardization organizations to which the members belong) and "fkw:similarTo" (similar documents based on keywords) (Table 1).

**Table 1.** List of designed metadata

Meta-data for persons
(from the Name List of the members engaging in the standardization activity)

| Meta-data field (Property) | Details | Domain (rdfs:domain) | Range (rdfs:range) | Example of meta-data value |
|---|---|---|---|---|
| foaf:familyName | Last name | foaf:Person | rdfs:Literal | Fuji |
| foaf:firstName | First name | foaf:Person | rdfs:Literal | Taro |
| fkw:belongsTo | Standardization organization to which the member belongs | foaf:Person | fkw:StdOrganization | <http://intap.org> |
| fkw:stdOrgOfficialPosition | Post assigned by the standardization organization | foaf:Person | rdfs:Literal | Member |
| fkw:company | Name of Company to which the member belongs to | foaf:Person | rdfs:Literal | FUJITSU LIMITED |
| fkw:country | Nationality | foaf:Person | rdfs:Literal | Japan |

Meta-data for documents (from the Standardization Activity Reports)

| Meta-data field (Property) | Details | Domain (rdfs:domain) | Range (rdfs:range) | Example of meta-data value |
|---|---|---|---|---|
| fkw:similarTo | Similar document based on keywords | fkw:Document | fkw:Document | <http://id103.doc> |
| dcterms:creator | Writer | fkw:Document | foaf:Person | <http://id96.per> |
| dcterms:subject | Theme of the standardization activity | fkw:Document | fkw:StdOrganization | <http://intap.org> |
| fkw:fileName | File name | fkw:Document | rdfs:Literal | ActRepo.pdf |
| fkw:publishedyear | Name of Company to which the member belongs to | fkw:Document | rdfs:Literal | FUJITSU LIMITED |

Meta-data for standardization activity organizations
(from the Name List of the members engaging in the standardization activity)

| Meta-data field (Property) | Details | Domain (rdfs:domain) | Range (rdfs:range) | Example of meta-data value |
|---|---|---|---|---|
| fkw:orgName | Organization name | fkw:StdOrganization | rdfs:Literal | INTAP |

### 2.4 Automatic Assignment of Metadata

In designing metadata, explicit information in the data, such as the names of persons engaged in a standardization activity, is automatically set as the values of the metadata.

For the Annual Standardization Activity Report DB (documents), the proposed method is based on the assumption that an efficient metadata application can further consolidate the relationship between the Name List and the Activity Report, and so the two types of information were added to the data through the following two methodologies.

**(1) Evaluation of Similarities among Activity Reports**

Writers of documents with many similarities can have knowledge and experiences similar to each other. Therefore, a keyword extraction method using statistical information of word co-occurrence [3] was combined with the "inverse document frequency."

By this combined method, listed below, 10 keywords were extracted from each activity report as a set. Then, similarities among these sets were evaluated by Simpson's Coefficient, which was used for the evaluation of the similarity between sets. If the value calculated by the coefficient was greater than the threshold value 0.5, the relevant reports were determined to be similar to each other.

*[Step 1] Extract up to 10 keywords from each document.*

1. Calculate the expected co-occurrence frequency of each term in the document [3]. The expected co-occurrence frequency of term w and term g is shown as the product of $n_w$ and $p_g$.

$n_w$: The total number of terms in a sentence where the term $w$ appears

$$p_g: \frac{\text{The total number of terms in a sentence where the term } g \text{ appears}}{\text{The total number of terms in the whole document}}$$

2. Calculate the $\chi^2$ value which indicates the bias of the term co-occurrence [3].

$$\chi^2(w) = \sum_{g \in G} \frac{(freq(w,g) - n_w p_g)^2}{n_w p_g} \tag{1}$$

where *freq(w,g)* is the co-occurrence frequency of term $w$ and term $g$. The inverse document frequency (IDF) is calculated by the following equation.

3. Calculate the IDF

$$\mathrm{IDF}(t) = 1 + \ln\left(\frac{N_{all}}{N_t}\right) \tag{2}$$

Where

$N_{all}$: The total number of documents
$N_t$: The number of documents where the term *t* appears

4. Calculate IDF value $\times \chi^2$ for each term, and select 10 terms having the highest results from the top as keywords.

*[Step 2] Calculate the similarity among documents in accordance with keywords common to the documents.*

1. Calculate the similarity by the Simpson's Coefficient

$$\text{Simpson's Coefficient} = \frac{|\boldsymbol{X} \cap \boldsymbol{Y}|}{\min(|\boldsymbol{X}|, |\boldsymbol{Y}|)} \tag{3}$$

where
$\boldsymbol{X}$ : Set of keywords of document 1
$\boldsymbol{Y}$ : Set of keywords of document 2

2. As documents with results more than a certain threshold value are considered to be similar to each other, each document name is set to the other's metadata field fkw: similarTo.

**(2) Setting of Standardization Organization Name on Which an Activity Report Focuses.**
The Name List was compiled as a dictionary of Standardization Organization Names and the top three organization names were set as metadata on the activity report data to facilitate finding a standardization activity or similar activities on which the activity report focused.

Once metadata were applied to the Name List DB, the DB was referred to as the "Profile DB".

### 2.5 Metadata Description in RDF

By describing the applied metadata in the Resource Description Framework (RDF) (Figure 2), it was possible to plot the members engaging in standardization activities and the annual standardization activity reports in RDF graphs to visualize the inter-DB relationships (Figure 3).

**NameSpace**

foaf: <http://xmlns.com/foaf/0.1/>(Metadata relevant to persons)
dcterms: <http://purl.org/dc/terms/>(Metadata relevant to Web content)
fkw: <http://know.who.org/2011/06/stdorg#>(Metadata specific to this system)

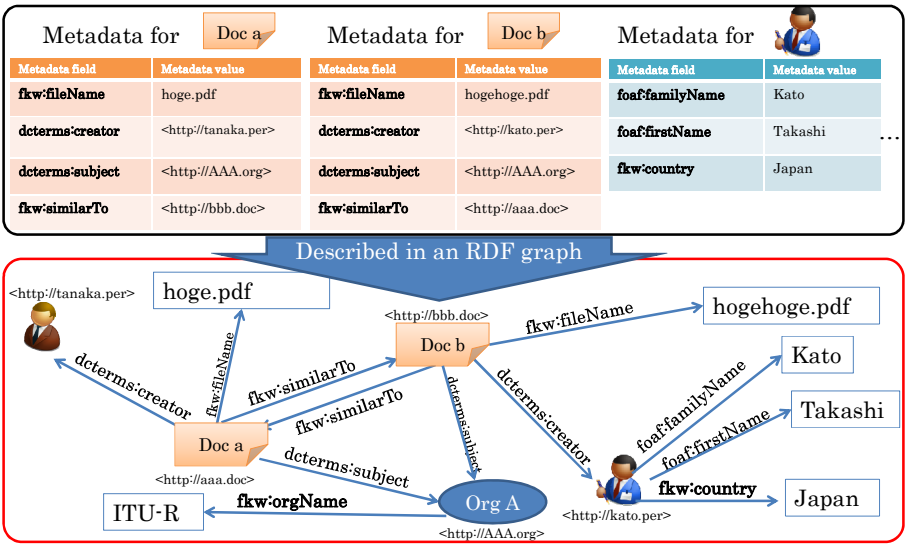


**Fig. 2.** Applying metadata and RDF description

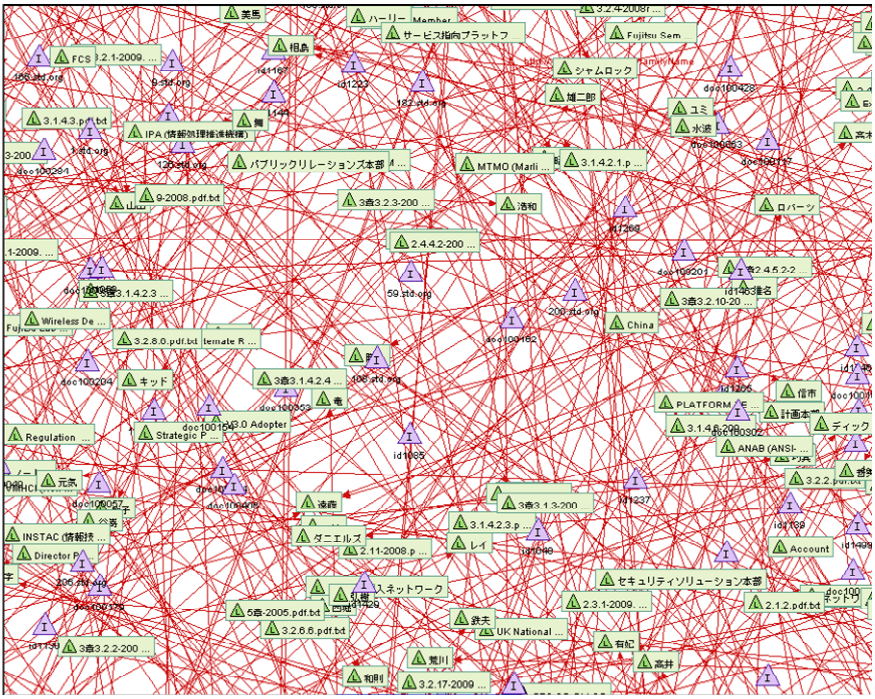

**Fig. 3.** Visualization using RDF Gravity (RDF Graph Visualization Tool)

By executing SPARQL queries on these RDF graphs (i.e. a group of on-tologized elements in both DBs), the same graph patterns were easily extracted.

### 2.6 Assumption of Successor Selection Logic for the Replacement of Members Engaging in a Standardization Activity

When the replacement of a member engaging in a standardization activity is needed because of reassignment or other reasons, customarily a colleague of the predecessor is selected as the successor. Accordingly, the successor has the knowledge and a human network similar to those of the predecessor, and thus such selection is considered as an appropriate decision.

The proposed method is based on the assumption that the conventional successor selection logic consists of five queries, as shown below (Table 2), on the basis of interviews from the relevant persons. The queries are described in SPARQL.

**Table 2.** Queries for researching persons who have the similar engineering expertise

| Query | Details |
|---|---|
| Q1 | Belonging to the same standardization organization as the predecessor. |
| Q2 | Belonging to a standardization organization on which the predecessor wrote the activity report as a theme. |
| Q2' | Writing (or has written) an activity report on a standardization organization to which the predecessor belongs. |
| Q3-1 | Writing (or has written) a report similar* to the report written by the predecessor. |
| Q3-2 | Writing (or has written) a report similar** to the report written by the predecessor. |

* This "similar to" means that the reports have been written on the same subject (dcterms:subject).
** This "similar to" means that the reports are linked with (fkw:similarTo).

### 2.7 Selection Procedure of Best Query Combination

The following procedure was adopted to select the best combination of queries.

1. Select the correct cases for training data.
2. Search by the assumed successor selection logic queries and combinations of these queries.
3. Calculate the information content ($I_{ij}$) of query combinations as follows:

$$\mathrm{I}_{ij} = \begin{cases} -\log_2 \dfrac{N_{ij}}{N_{all}} & (N_{ij} \text{ includes a successor}) \\ 0 & (N_{ij} \text{ includes no successor}) \end{cases} \tag{4}$$

where
$N_{ij}$: The number of successor candidates as matching results
$N_{\mathrm{all}}$: The number of all successor candidates

4. Select the best query combinations that maximize $I_i$ .

$$\mathrm{I}_i = \sum_j \mathrm{I}_{ij} \tag{5}$$

### 2.8 Visualization

The visualization method used by this system is as follows. Visualization of the resource link by the network helps the user's understanding of the relationship between resources. First, resources such as standards organizations and members are set as nodes. Second, the metadata link the resources. Then, the network is generated and visualized. A metadata value about the name of the resource is shown as a node label.

## 3 Case Studies for Verifying the Availability of Linked Data

This system was verified by whether it could find a successor in a similar manner to the conventional method for actual replacement cases.

Specifically, the following metadata were used to select the cases in which members engaging in standardization activities were replaced by the same process described in Section 2:

(1) The 5 queries mentioned above,
(2) Data of approximately 550 persons in the “Profile DB” (persons) for members engaging in the Official Standardization Activity,
(3) Reports of the past 5 years in the “Annual Standardization Activity Report DB” (documents).

Approximately 200 official standardization organizations were extracted. The system was verified as to whether it could reproduce the same results as conventional selections.

### 3.1 Selection of Correct Cases for Training Data

Table 3 shows the past 5 years’ actual replacements of members for standardization activity reports.

**Table 3.** Five persons who took over themes for standardization activity reports

| No. | Predecessor | Successor | Report theme | Year |
|---|---|---|---|---|
| 1 | Ishikawa | Kinoshita | OMA | 2005 → 2006 |
| 2 | Ishikawa | Murayama, Hata | ITU-R | 2005 → 2006 |
| 3 | Kinoshita | Ozaki | OMA | 2006 → 2007 |
| 4 | Kikuchi | Yokosuka | INTAP | 2007 → 2008 |
| 5 | Hata | Endo | ITU-R | 2008 → 2009 |

### 3.2 Searches by the Assumed Successor Selection Logic Queries

After conducting searches by the abovementioned 5 queries through the whole candidate DB (approximately 550 persons), the following was determined:

(1) How many persons were refined from the total candidates,
(2) Whether the actual successor really existed within the refined candidates.

### 3.3 Results

The 5 queries in 15 different combinations for the data of the targeted 550 persons were executed. Table 4 shows the results of the selection from the candidate DB to refine the number of candidates to 10 or less. If a case matches "Q1" and "Q2" separately, it is not certain that the case matches the combination of "Q1" and "Q2". The information content in Table 5 was calculated using equation (4).

**Table 4.** Query combinations for the 5 predecessors and results refined from approximately 550 candidates

| No | Predecessor | Q1 | Q2 | Q2' | Q3-1 | Q3-2 | Q1 ∩ Q2 | Q1 ∩ Q2' | Q1 ∩ Q3-1 | Q1 ∩ Q3-2 | Q2 ∩ Q2' | Q2 ∩ Q3-1 | Q2 ∩ Q3-2 | Q2' ∩ Q3-1 | Q2' ∩ Q3-2 | Q3-1 ∩ Q3-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ishikawa | 46 | 44 | 19 | 37 | 12 | 27 | 4 | 3 | 5 | 3 | 6 | 5 | 17 | 6 | 9 |
| 2 | Ishikawa | 46 | 44 | 19 | 37 | 12 | 27 | 4 | 3 | 5 | 3 | 6 | 5 | 17 | 6 | 9 |
| 3 | Kinoshita | 98 | 81 | 24 | 38 | 11 | 48 | 6 | 3 | 6 | 3 | 8 | 6 | 18 | 6 | 9 |
| 4 | Kikuchi | 166 | 138 | 18 | 55 | 21 | 71 | 7 | 1 | 13 | 1 | 5 | 3 | 8 | 6 | 13 |
| 5 | Hata | 72 | 32 | 19 | 30 | 7 | 10 | 5 | 1 | 5 | 1 | 6 | 3 | 6 | 5 | 7 |

**Table 5.** Infromation content

| No | Predecessor | Q1 | Q2 | Q2' | Q3-1 | Q3-2 | Q1 ∩ Q2 | Q1 ∩ Q2' | Q1 ∩ Q3-1 | Q1 ∩ Q3-2 | Q2 ∩ Q2' | Q2 ∩ Q3-1 | Q2 ∩ Q3-2 | Q2' ∩ Q3-1 | Q2' ∩ Q3-2 | Q3-1 ∩ Q3-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Ishikawa | 3.58 | 3.64 | 4.85 | 3.89 | 5.52 | 4.35 | 7.10 | 7.52 | 6.78 | 7.52 | 6.52 | 6.78 | 5.01 | 6.52 | 5.93 |
| 2 | Ishikawa | 0.00 | 3.64 | 0.00 | 3.89 | 5.52 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 5.93 |
| 3 | Kinoshita | 2.49 | 2.76 | 4.52 | 3.85 | 0.00 | 3.52 | 0.00 | 0.00 | 6.52 | 0.00 | 6.10 | 6.52 | 4.93 | 0.00 | 0.00 |
| 4 | Kikuchi | 0.00 | 0.00 | 4.93 | 3.32 | 4.71 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.10 | 6.52 | 0.00 |
| 5 | Hata | 2.93 | 4.10 | 4.85 | 4.19 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 6.52 | 0.00 | 6.52 | 0.00 | 0.00 |

Design method: Select a query combination $i$ so that the equation (5) gives the greatest value for $I_i$. Evaluate the design method using a cross-validation technique.

Conduct cross-validations for the 5 correct data sets by using 4 of them as training data and one of them as test data. The cross-validation results are given in the table below. The values marked in yellow are the maximum in each validation.

**Table 6.** Selection of best query combinations (5 cases)

| Training data | Q1 | Q2 | Q2' | Q3-1 | Q3-2 | Q1 ∩ Q2 | Q1 ∩ Q2' | Q1 ∩ Q3-1 | Q1 ∩ Q3-2 | Q2 ∩ Q2' | Q2 ∩ Q3-1 | Q2 ∩ Q3-2 | Q2' ∩ Q3-1 | Q2' ∩ Q3-2 | Q3-1 ∩ Q3-2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1,2,3,4 | 1.52 | 2.51 | 3.57 | 3.74 | 3.93 | 1.97 | 1.78 | 1.88 | 3.32 | 1.88 | 3.15 | 3.32 | ***4.01*** | 3.26 | 2.97 |
| 1,2,3,5 | 2.25 | 3.54 | 3.56 | 3.96 | 2.76 | 1.97 | 1.78 | 1.88 | 3.32 | 1.88 | ***4.78*** | 3.32 | 4.11 | 1.63 | 2.97 |
| 1,2,4,5 | 1.63 | 2.85 | 3.66 | 3.82 | 3.93 | 1.09 | 1.78 | 1.88 | 1.69 | 1.88 | 3.26 | 1.69 | ***4.41*** | 3.26 | 2.97 |
| 1,3,4,5 | 2.25 | 2.63 | 4.79 | 3.81 | 2.56 | 1.97 | 1.78 | 1.88 | 3.32 | 1.88 | 4.78 | 3.32 | ***5.64*** | 3.26 | 1.48 |
| 2,3,4,5 | 1.35 | 2.63 | 3.57 | 3.81 | 2.56 | 0.88 | 0.00 | 0.00 | 1.63 | 0.00 | 3.15 | 1.63 | ***4.39*** | 1.63 | 1.48 |

**Table 7.** Evalation of selection results

| | Q2 ∩ Q3-1 | Q2' ∩ Q3-2 | Q2' ∩ Q3-1 |
|---|---|---|---|
| training data:1,2,3,4 | | | 6.52 |
| training data:1,2,3,5 | 0 | | |
| training data:1,2,4,5 | | | 4.93 |
| training data:1,3,4,5 | | | 0.00 |
| training data:2,3,4,5 | | | 5.01 |

The average information amount obtained from the results of the selected queries was 3.29. The result is better than 2.67, which was the average of the information amount of all the queries.

### 3.4 Visualization

Furthermore, visualizing the data relationships also made it possible to find other successor candidates besides the actual successors.

The example shown in Figure 4 indicates that “Endo” and “Ozaki”, who have similar relationships, can also be candidates for successors for the former “Ishikawa”. “Kinoshita” and “Hata” were actually selected.

As shown in Table 3, “Ozaki” was selected as a successor to “Kinoshita” at a later date and “Endo” was also selected as a successor to “Hata”, which shows that the possibilities shown in Figure 4 are reasonable and appropriate.

Thus, visualization of the relationship between profiles (persons) and activity reports (documents) made it possible to propose other options for member nomination.

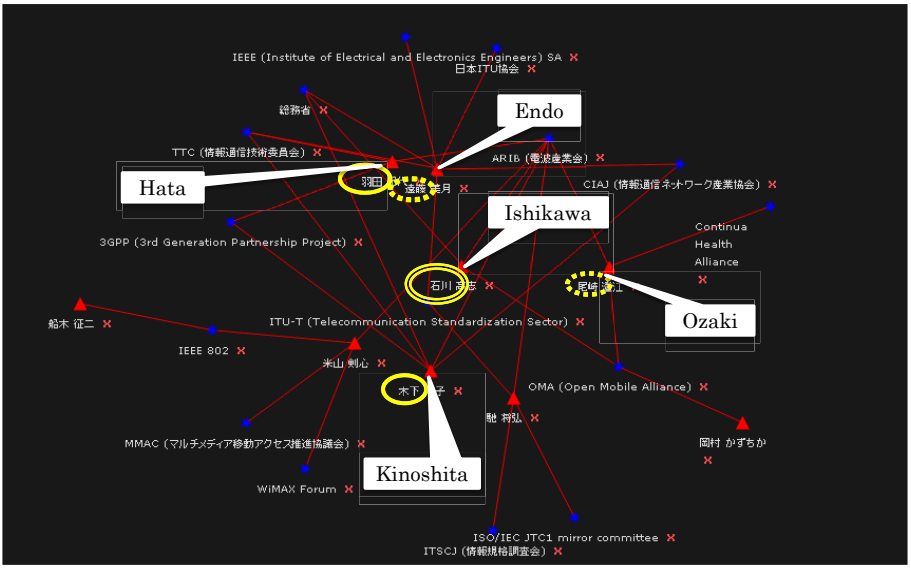


**Fig. 4.** Possibilities of the other successors

## 4 Discussion

Though the proposed design procedure gives better results for the successors than does random selection from the candidates in the lists, the accuracy is not good enough. However, the selection results are justified more clearly than the current practice because the results are based on clear rules and queries described in the RDF metadata and the SPARQL queries. The criteria can be clarified in this method and this is one of the contributions of this research.

The proposed design procedure is evaluated for a very small number of case studies and data, so the findings of this research must be confirmed continuously in the real field. Data on human resources are very difficult to manage in a company and compliance and privacy problems might be considered for further research. Though the findings of the case studies might be revealed by other approaches, the number of metadata fields is getting bigger and so this approach works more effectively when employed for search processes in the workplace.

As for the visualization, this manuscript proposes a method for rendering the relationships between resources, such as humans and documents. All the information in the graphs can be shown to the persons in charge to discover new findings. The relationships described in metadata, such as similarity, could be illustrated using varying font sizes and other visualization techniques.

## 5 Conclusion

A procedure for the selection of queries to find an appropriate successor to an activity member is proposed in this paper. The proposed procedure nominates candidates for the successor and the list of nominees is more accurate than that obtained by random selection. The proposed procedure effectively limits the number of qualified candidates and thus fulfills objective-1.

For objective-2, a visualization method for selection of the successor is proposed and the results of the visualization are shown. The links explain the similarities and commonalities of the data. Discovering an appropriate successor from the visualization results is also discussed in the paper. Visualization is one of the crucial approaches for extracting the background of human affairs. To promote more appropriate and efficient searches for a successor, visualizing data links is essential.

## 6 Future Research

In order to promote more efficient information search by the procedure for the selection of the queries to find an appropriate successor is proposed in the paper, it is essential to popularize this approach on a broad range of public data.

## References

1. Kashima, H.: Survey of Network Structure Prediction Methods. Journal of the Japanese Society for Artificial Intelligence 22(3), 344–351 (2007)
2. Matsumura, F., Kobayashi, I., Kamura, T., Kato, F., Takahashi, T., Ueda, H., Ohmukai, I., Takeda, H.: Collaboration between Linked Open Data of Museum Information and Regional Information. Information Processing Society of Japan, Computers and the Humanities Symposium, "JINMONKON 2011" Journal 2011(8), 403–408 (2011)
3. Matsuo, Y., Ishizuka, M.: Keyword Extraction from a Document Using Word Co-occurrence Statistical Information. Journal of Japanese Society for Artificial Intelligence 17(3), 217–227 (2002)