Assigning Vocation-Related Information to Person Clusters for Web People Search Results

Hiroshi Ueda[1)] Harumi Murakami[2)] Shoji Tatsumi[1)]
[1)]*Graduate School of Engineering, Osaka City University*
[2)]*Graduate School for Creative Cities, Osaka City University*

Abstract

Distinguishing people with identical names is becoming more and more important in searches on the Web. This research assigns useful labels to help users select person clusters that are separated into different people from the result of person searches on the Web. We proposed a method to label person clusters with vocation-related information (VRI). VRI includes broader terms that may not be considered vocations as well as terms that are useful to infer vocations, not only those rigorously defined as vocations. Our method is comprised of two processes: (a) extraction of VRI candidates using HTML structure and heuristics, and (b) VRI generation using term frequency, clustering synonyms, and calculation using a Web search engine. Experimental results revealed the usefulness of our proposed method.

1. Introduction

Finding information about people on the Web is one of the most popular search activities. According to [1], 30% of queries in Web searches include person names. Person name disambiguation, or distinguishing people with identical names, is becoming more and more important in Web searches. Most research of person name disambiguation concentrates on automatically separating Web pages for different people using clustering algorithms. However, if the list of search results is merely "person 1, person 2, …, and so on," users have difficulty determining which person clusters they should select.

This research assigns useful labels to help users select person clusters that are separated into different people from the result of person searches on the Web.

Our hypothesis argues that information about vocation is important to identify different people. We propose a method to label person clusters with vocation-related information (VRI) that includes broader terms that may not be considered vocations and terms useful to infer vocations, not only those rigorously defined as vocations. In addition, detailed information is more useful than a single vocation term. For example, *Professor of Graduate School of Engineering, Osaka City University* is often more useful than just a single vocation term (*researcher*) to identify a person. The examples in this paper were translated from Japanese into English for publication.

Table 1: People search results for "Suguru Egawa."

Person Clusters	Number of Web Pages	Generated VRI by our Method
Suguru Egawa 1	589	baseball player
Suguru Egawa 2	126	scholar of Russian literature
Suguru Egawa 3	1	commentator

Table 1 shows the example results for the person clusters for query "Suguru Egawa (Suguru is a given Japanese name and Egawa is a family name)" using our method. For the first cluster (Suguru Egawa 1), which includes 589 Web pages, VRI "baseball player " was generated.

Below, in Section 2 we explain our method. The experimental results are described in Section 3. We discuss our method's usefulness and related work in Section 4.

2. Method

We propose a method to extract vocation-related information (VRI) from Web pages in person clusters. VRI is information related to vocation, which is more useful for identifying people. In this paper, a person's name is composed of surname and given names.

We selected three types of VRI: (1) vocation, (2) organization and position, and (3) publication title and role. Examples of vocation types are *researcher*, *lawyer*, and *doctor*. Examples of organization and position types are *Professor of Graduate School of Engineering, Osaka City University* and *Vice president of ABC Corporation*. Examples of publication title and

978-0-7695-3571-5/09 $25.00 © 2009 IEEE
DOI 10.1109/GCIS.2009.254

role types are *author of Dragon Ball* and *translator of Harry Potter and the Philosopher's Stone*.

The method is comprised of two processes: (a) extraction of VRI candidates using HTML structure and heuristics, and (b) VRI generation using term frequency, clustering synonyms, and calculation using a Web search engine.

An overview of our method is displayed in Figure 1.

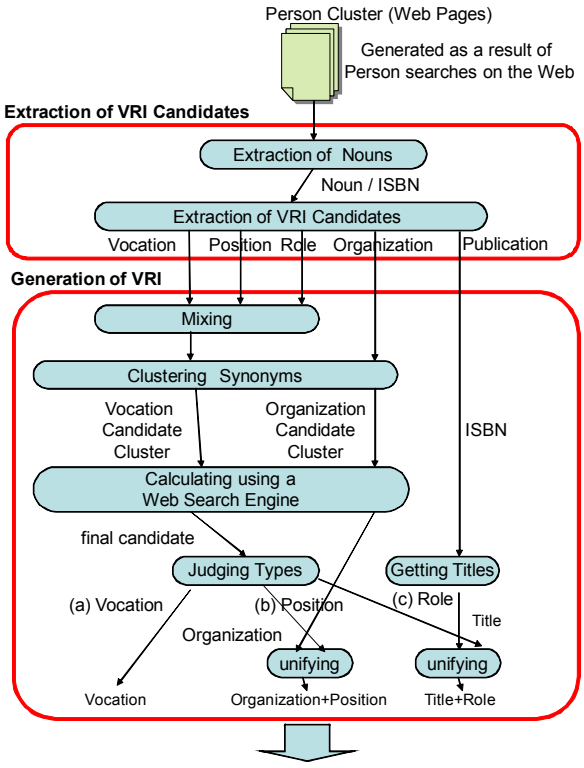


Figure 1: Overview of method.

2.1. Extraction of VRI candidates

This process identifies nouns that are related to persons from Web pages using HTML structures and candidates from these nouns using heuristics.

2.1.1. Extraction of nouns. First, our method extracts nouns when person names are contained in <p>, <tr>, first or second rows in <table>, <title>, <h1>, <h2>, and <h3>.

Second, our method concentrates on the location of person names and extracts the following: (a) nouns in brackets () that appear immediately after person names and (b) nouns that appear immediately before or after person names.

Finally, ISBN numbers are extracted using regular expressions and added as nouns to extract publication title term candidates in Section 2.1.2.

2.1.2. Extraction of VRI Candidates. Our method judges nouns to identify VRI candidates using heuristics. Since this process concentrates on Japanese, it must be modified for other languages.

(1) Vocation term candidates

17 heuristics were selected that concentrate on noun suffixes, including the following examples: a noun that ends with "SHI" (suffix that expresses professionals) is a vocation term candidate; a noun comprised of more than three characters and ends with "dash" (suffix that expresses professionals) is a vocation term candidate.

(2) Organization term candidates

Organization term candidates are judged by two steps. In the first step, NExT (a Japanese NE tool) is used. A noun (a) which ends with terms which are composed of more than one character defined as "organization" in NExT and (b) whose number of characters is more than "the number of terms identified in process (a) plus one." In the second step, 14 heuristics were described. An example of heuristics is as follows: a noun which starts with "KABU" (abbreviation of corporation) and comprises of more than five characters is an organization term candidate.

(3) Position term candidates

15 heuristics were selected that concentrate on noun suffixes, including the following examples: a noun that ends with "IN" (suffix that expresses being a member of) is a position term; a noun comprised of more than three characters and ends with "professor" is an position term candidate.

(4) Publication title term candidates

ISBN numbers are treated as publication title term candidates.

(5) Role term candidates

Three heuristics are created: nouns that end with "author", "editor", or "translator" are identified as role term candidates.

2.2. VRI generation

The process of VRI generation engenders VRIs from candidates. Our method is based on term frequency, clustering synonyms, and calculation using a Web search engine.

2.2.1. Clustering synonyms. Our method mixes vocation candidate terms, position candidate terms, and role candidate terms and clusters synonyms into vocation candidate clusters.

It also clusters synonyms into organization candidate clusters from organization term candidates.

Figure 2 displays the process of generating vocation candidate clusters.

The process of generating organization candidate clusters is only different in Step 2 from that of vocation candidate clusters. Suffix is changed to prefix.

> Step 1. Select the shortest candidate and generate a cluster from it.
> Step 2. Compare suffixes of candidates that do not belong to any cluster by pattern-matching. If matched, the candidate is included in the cluster.
> Step 3. Repeat Steps 1 and 2 until all candidates are processed.

Figure 2: Generating vocation candidate clusters.

2.2.2. Calculation using a web search engine. Firstly, our method selects clusters in which the number of term frequency is maximum from vocation candidate clusters. If there is one candidate term in the clusters, the candidate is selected as the final candidate. If there is more than one candidate term, one candidate term is selected from the following process. We use a Web search engine to calculate the weight of the candidate based on Mori's method using a Jaccard index [2]. The weight of candidate J(n,v) between person name n and candidate v is calculated as follows:

$$J(n,v) = \frac{|N \cap V|}{|N| + |V| - |N \cap V|} \qquad (1)$$

|N| is the number of Web search results when a query is n. |V| is the number of Web search results when a query is v. |N ∩ V| is the number of Web search results when a query is n and v.

The highest candidate is selected as the final candidate.

Second, we identify the type of final candidate, perform a different process based on the type, and generate a VRI. There are three types in the final candidate: vocation term, position term, and role term.

If the final candidate is a vocation term, it directly becomes a VRI.

If it is a position term, whether it includes organizational information must be determined using a method described in Section 2.1.2. If the position term already includes organizations like *professor of Osaka City University*, it becomes a VRI. If it does not contain organizational information, the subsequent process is done.

Our method obtains an organization term candidate from organization candidate clusters and selects clusters in which the number of term frequency is maximum from organization candidate clusters. If there is one organization term candidate in the clusters, it is unified with the final candidate (position term) and becomes a VRI. If there is more than one organization term candidate, one organization term candidate is selected by the following process. We use a Web search engine to calculate the weight of the candidate based on Mori's method using a Jaccard index [2]. The weight of candidate J(n,o,v) between person name n, organization candidate o, and position candidate v is calculated as follows:

$$J(n,o,v) = \frac{|N \cap O \cap V|}{|N \cap O| + |V| - |N \cap O \cap V|} \qquad (2)$$

|N ∩ V| is the number of Web search results when a query is n and v. |O| is the number of Web search results when a query is o. |N ∩ O ∩ V| is the number of Web search results when a query is n and o and v.

The highest candidate is selected as an organization candidate, unified with the position candidate, and becomes a VRI.

If the final candidate is a role term, our method selects one publication title term from the publication title term candidates. We select the ISBN number that is closest to the person's name and get the title of the ISBN number using Amazon Web Service (AWS). The publication title and final candidate (role term) are combined and become a VRI.

2.2.3. Example. Figure 3 illustrates the process of VRI generation when a person is a *Cultural Affairs Department journalist, Tozai newspaper.*

If we take maximal frequencies from the process results of extracting VRI candidates (vocation, position, and role term candidates), *voice actor* would become the answer because its frequency is maximal (6). However, that is wrong.

Our method of clustering synonyms creates vocation candidate clusters. A vocation candidate cluster whose seed is *journalist* and with seven candidates becomes the maximal frequency (11). Seven vocation candidate terms are examined by calculation using a Web search engine, and *Cultural Affairs Department journalist* becomes the final candidate. Since it is judged to be a position, we need to get an organization term.

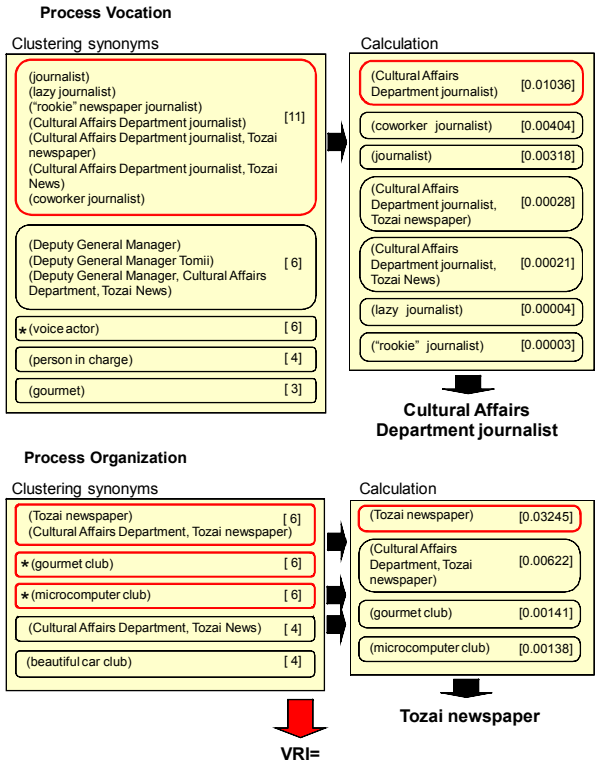
If we take maximal frequencies from the process results of the extracted VRI candidates (organization term candidates), *gourmet club* or *microcomputer club* would become the answer since their frequency is maximal (6). However, they are wrong.

Our method of clustering synonyms creates organization candidate clusters. Three organization

candidate clusters become the maximal frequency. Four organization term candidates are examined by calculating with a Web search engine, and *Tozai newspaper* becomes the organization term.

Finally, *Tozai newspaper* and *Cultural Affairs Department journalist* are combined and *Cultural Affairs Department journalist, Tozai newspaper* is generated.

The advantage of our method is the ability to generate correct and detailed VRIs, including *Cultural Affairs Department journalist, Tozai newspaper* instead of such wrong and simple information as *voice actor*.



Note: *: VRI candidates whose term frequencies are maximal without clustering synonyms. []: the number of frequency of VRI candidates in a cluster.

Figure 3: Examples of VRI generation.

3. Experiment

3.1. Dataset

Twelve person names with different people out of 20 person names used in [5] were selected as queries. 230 people were found in Web search results (google Web APIs) for twelve person name queries. The average number of different people was nineteen and the average number of web pages was 475.

3.2. Experiment 1

3.2.1. Method. Three person names were selected for queries from the dataset. They are researchers and the number of web pages is intermediate. The above three methods (a, b, c) were applied to the search results of the three queries. For 32 people, VRIs were generated with all three methods.

Three subjects evaluated whether the generated VRIs expressed the vocations of the people as follows.

- First, they manually extracted vocation terms or VRIs by checking all Web pages divided into 32 people.
- Second, they evaluated the related value to the generated VRIs by referring to the above manual results as follows: 2 (The term is the vocation of the designated person), 1 (The term is not the vocation, but is related to the designated person's vocation), or 0 (The term is not related to the vocation of the designated person at all).

We judged the data with the related values of 2 and 1 to calculate precision.

Table 2: Results of experiment 1.

	(a) Our method	(b) Web pages	(c) TF only
Related value	1.19	0.36	1.09
Precision	71.1%	24.4%	68.9%

3.2.2. Results and discussion. The results are shown in Table 2. For the average related values, our method was highest (1.19) compared to Web pages (0.36) and TF only (1.09). For the average precision, our method was 71.1% (64/90), Web pages was 24.4% (22/90), and TF only was 68.9% (62/90); our method was highest.

Our method is the best among the three methods in related values and precision. The comparison between our method and Web pages shows the usefulness of our method of extracting VRI candidates using HTML structure and heuristics. The results of our result and TF only were similar. When the number of VRI candidates was small, both methods tended to generate the same terms. When the number of VRI candidates was large, they tended to generate different terms.

3.3. Experiment 2

Experiment 2 investigated the usefulness of our method compared to related work. We compared (a) our method with methods (b) and (c). (b) extracts titles by pattern-matching. In the sentence, "David Lee is a painter," painter is extracted as a title. The most frequent title is treated as a vocation. (b) is a simulation of [3]. (c) is a manual method. The subject searches for the profile pages for the designated person and extracts the most suitable term as a VRI from the profile pages by checking Web pages. Hereafter we call (b) Title and (c) Profile.

3.3.1. Method. All twelve person names were used for the queries. Three subjects evaluated the VRIs as in Experiment 1. Related values and precision were calculated as in Experiment 1. Next, recall was defined and examined as follows:

$$recall = \frac{number\ of\ persons\ whose\ VRIs\ were\ relevant}{numbers\ of\ persons\ whose\ relevant\ VRIs\ were\ extracted\ manually} \quad (3)$$

3.3.2. Results and discussion. The results are shown in Table 3. For the average related values, our method was 1.41, Title was 1.17, and Profile was 1.76. For precision, our method was 84.7% (94/111), Title was 66.7% (4/6), and Profile was 90.5% (19/21). Concerning recall, our method was 66.7% (94/141), Title was 2.8% (4/141), and Profile was 13.5% (19/141).

Table 3: Results of experiment 2.

	(a) Our method	(b) Title	(c) Profile
Related value	1.41	1.17	1.76
Precision	84.7%	66.7%	90.5%
Recall	66.7%	2.8%	13.5%

Overall, our method outperformed Title. Although our method was inferior to Profile in related values and precision, it was much better in recall.

The above results suggest the usefulness of our method as an automatic method. It also outperformed Profile in terms of recall.

4. Related work and discussion

This research assigns useful labels to help users select person clusters that are separated into different people from the result of person searches on the Web. Wan assigns titles to person clusters using a method similar to Title in Experiment 2 [3]. We assign VRIs, which are more useful than titles.

Much work (e.g., [4, 5, 6, 7]) separates Web pages into person clusters [8], however, it seldom assigns labels to person clusters.

Technically, our idea is based on [2] in the sense that it uses not only term frequency but also a Web search engine to calculate terms.

Our work is also related to such clustering search engines as Vivisimo, which usually assign one label to a web page cluster to help users select a cluster based on such information as term frequency and URLs. TSUBAKI [9] clusters synonyms like our work, but it does not use Web search engines to calculate weights.

The main advantage of our method is to label VRIs to person clusters based on the context, not vocations based on term frequencies. For example, if a person is on a faculty staff as both a researcher and a teacher, our method outputs *Osaka City University professor*, or *computer science researcher* based on the context in which the person appears on the Web, instead of displaying a simple vocation word like *researcher* or *teaching staff*. In addition, no special vocation dictionary is needed, which is another benefit of our method.

The experimental results revealed the following. (1) The precision and recall of our method were 84.7%, and 66.7%, respectively. (2) Our method has the best related values, precision, and recall among the automatic methods. (3) Our method features much better recall against a manual method using profile pages. The above results suggest our method's usefulness.

Future work includes the following. First, we need to improve the extraction of VRI candidates. Second, when the number of Web pages is small, new algorithms should be included. Third, our method sometimes generates former vocations, which may be acceptable if the person is famous; however, users generally want current vocations. We need to explore algorithms to identify times and evaluate which is better to present to users.

5. Conlusion

We proposed a method to label person clusters with vocation-related information.

The main advantage of our method is to label VRIs to person clusters based on the context, not vocations based on term frequencies. In addition, no special vocation dictionary is needed. This is another great benefit of our method.

The experimental results revealed the following. (1) The precision and recall of our method were 84.7%, and 66.7%, respectively. (2) Our method has the best related values, precision, and recall among the automatic methods. (3) Our method features much better recall against a manual method using profile pages. The above results suggest our method's usefulness.

References

[1] R. Guha, and A. Garg, "Disambiguating People in Search", Stanford University, 2004.

[2] J. Mori, Y. Matsuo, and M. Ishizuka, "Personal Keyword Extraction from the Web", *Transactions of the Japanese Society for Artificial Intelligence*, 2005, Vol.20, No.5, pp. 337-345.

[3] X. Wan, J. Gao, M. Li, and B. Ding, "Person Resolution in Person Search Results: WebHawk", *CIKM2005, Proceedings of the Fourteenth ACM Conference on Information and Knowledge Management*, 2005, pp. 163-170.

[4] R. Bekkerman, A. and McCallum, "Disambiguating Web Appearances of People in a Social Network", *WWW2005, Proceedings of the Fourteenth World Wide Web Conference*, 2005, pp. 463-470.

[5] S. Sato, K. Kazama, and K. Fukuda, "Distinguishing between People on the Web with the Same First and Last Name by Real-world Oriented Web Mining", *IPSJ Transactions on Databases*, 2005, Vol.46, No. 8, pp. 26-36.

[6] I. Bhattacharya, and L. Getoor, "Collective Entity Resolution in Relational Data", *ACM Transactions on Knowledge Discovery from Data*, 2007, Vol. 1, Issue 1, Article No. 5.

[7] Z. Kozareva, R. Moraliyski, and G. Dias, "Web People Search with Domain Ranking", *Text, Speech, and Dialogue, Lecture Notes in Computer Science*, 2008, pp.133-140.

[8] J. Artiles, J. Gonzalo, and S. Sekine, "The SemEval-2007 WePS Evaluation: Establishing a Benchmark for the Web People Search Task", *Proceedings of the Fourth International Workshop on Semantic Evaluations*, 2007, pp. 64-69.

[9] Y. Baba, K. Shinzato, and K. Kurohashi, "Development of a Large-scale Web Page Clustering System using an Open Search Engine Infrastructure TSUBAKI", *IPSJ Sig Technical Report*, 2008, Vol. 2008, No. 4, pp. 67-74.