Keyword Extraction from the Web for Creation of Person Metadata

Junichiro Mori[1,3], Yutaka Matsuo[2], Mitsuru Ishizuka[1], and Boi Faltings[3]

[1] University of Tokyo, Japan,

{jmori, ishizuka}@miv.t.u-tokyo.ac.jp

[2]National Institute of Advanced Industrial Science and Technology, Tokyo, Japan

y.matsuo@carc.aist.go.jp

[3]Ecole Polytechnique Federal Lausanne, Switzerland

{junichiro.mori, boi.faltings}@epfl.ch

Abstract

As one emerging metadata standard for the Semantic Web, FOAF defines an RDF vocabulary for expressing metadata about people and the relation among them. This paper proposes a novel keyword extraction method to extract FOAF metadata from the Web. The proposed method is based on co-occurrence information of words. Our method extracts relevant keywords depending on the context of a person. Our experimental results show that extracted keywords offer strong potential for FOAF metadata extraction from the Web.

1 Introduction

With the currently growing interest in the Semantic Web, metadata is coming to play an important role in the Web. As one emerging metadata standard for the Semantic Web, FOAF[Brickley and Miller, 2004], "Friend of a Friend", defines an RDF vocabulary for expressing metadata about people and the relation among them. FOAF provides a way to create machine-readable documents on the Web, and process them easily through merging and aggregating them.

As major hurdle of the Semantic Web is the creation of metadata, FOAF also needs methods that facilitate and accelerate metadata creation so that FOAF metadata is being widely used and exchanged. One way to solve the problem of metadata creation is to extract metadata from the Web. Considering the personal information that the FOAF vocabulary expresses, we notice that a lot of information is contained in the Web pages. For example imagine a researcher: that researcher's information can be in an affiliation page, a conference page, an online paper, or even in a blog. In fact, we can expect that these pages contain a lot of FOAF metadata even including information that we would not expect to find.

One of our research goals is to extract metadata of people in their various contexts. The extracted metadata would be expressed in the FOAF vocabulary and its extension. As a preliminary report to achieve this goal, we propose a novel keyword extraction method to extract personal information from the Web. Our hypothesis is that some FOAF metadata of a person is included in the extracted keywords. We analyze the extracted keywords to verify this hypothesis, and explore the possibility of FOAF metadata extraction from the Web. In particular we analyze the following points.

- What kind of FOAF metadata are included in the extracted keywords?
- What kind of Web page contains FOAF metadata?

2 Keyword Extraction

As an experimental attempt, we extracted the keywords of FOAF project members (25 persons in total). First, we put each person's full name to a search engine, and retrieve documents related to each person. From the search result, we used the top 10 documents per person as the initial documents that might contain personal keywords.

The html files, at to a maximum 10 files per person, are acquired from the initial documents of each person. They are pre-processed with html-tag deletion and part-of-speech tagging(POS). Then, the term set for keyword extraction is extracted from pre-processed html files using the term extraction tool, Termex [Nakagawa *et al.*, 2003]. After the whole procedure of extracting the term set, we acquired about 1000 terms per person on the average. The relevant keyword of each person is chosen from these terms.

Because the term set includes both relevant and irrelevant terms for personal information, we need to evaluate the relevance of term as a personal keyword. The simplest approach to measure term relevance as a personal keyword is to use co-occurrence. In this paper, we define co-occurrence of two terms as term appearance in the same Web page. If two terms co-occur in many pages, we can say that those two have a strong relation and one term is relevant for another term. This co-occurrence information is acquired by the number of retrieved results of a search engine. For example, assume we are to measure the relevance of name N(e.g. "Dan Brickley") and term w(e.g "Semantic Web"). We first put a query, "N and w", to a search engine and obtain the number of retrieved documents that is denoted by $|N \text{ and } w|$. We continuously apply a query, "N" and "w", and get the number of matched documents for each, $|N|$ and $|w|$. Then, the relevance between the name N and the term w, denoted by $r(N, w)$, is approximated by the following Jaccard coefficient.

$$r(N, w) = \frac{|N \text{ and } w|}{|N| + |w| - |N \text{ and } w|}$$

This Jaccard coefficient captures the degree of co-occurrence of two terms by their mutual degree of overlap.

Because different Web pages usually reflect different contexts of a person, we introduce the notion of a context to the keyword extraction Then, the extended rele-

vance of person N and term w in context C, denoted by $score(N, C, w)$, is calculated as the following.

$$score(N, C, w) = \frac{r(N, w)}{MAX(r(N, w))} + \alpha \frac{r(C, w)}{MAX(r(C, w))}$$

Therein, α denotes the relevance between the person and the context. The term w with the higher $score(N, C, w)$ is considered to be a more relevant keyword for person N in context C.

3 Keyword Analysis for FOAF Metadata Extraction

As an example of extracted keywords, Table.1 shows higher-ranked keywords of "Dan Brickley", one of FOAF founders. As explained in the previous section, the context is considered in the keyword extraction. In this experiment, we used "FOAF" as the context term. With this context, keywords are chosen in relation to FOAF.

Among the higher-ranked keywords, we can find various personal information such as personal name, organization, project, and related URL. To analyze what kind of FOAF metadata are included in the extracted keywords, we annotated one of 8 property labels (Name, Term, Project, Organization, Event, URL, Position, Community) to higher-ranked keywords of each person. These property labels are prepared considering a correspondence to FOAF metadata. Thereby we acquired 1392 labeled keywords in total (about 55 keywords per person on average).

Table.2 shows the distribution of property labels. Nearly half of higher-ranked keywords are occupied with technical terms. Notwithstanding, it is noteworthy that other properties such as personal names, organizations, and projects also appear to a certain degree. In particular, as shown on the right side column, the properties for each person are distributed in a balanced manner. This distribution means that if we extract about 55 higher-ranked keywords of one person, in average we can obtain about 20 names of his acquaintance, 2 or 3 related organization, and 1 or 2 projects. These numbers nearly match our activity and show the possibility of using the keywords for FOAF metadata extraction.

Furthermore, we analyzed which Web pages include a higher-ranked keyword. First, we classified all 250 web pages (10 per a person) that were used to extract the initial term set. Thereby, we produced the 17 categories such as "Personal page", "blog", "Online paper". As a result, we find that "Personal page" is a good source for personal information such as names, organizations, and projects. Online bibliography page can provide the information that is related to personal research activities such as coauthors, projects, and events(e.g. conferences and workshops). Interesting point is that blog pages offer the potential to be a resource for various personal information. One reason is that they contain more personal information. Another reason is that they track the change of information through frequent page updates.

4 Discussion

In applying the extracted keywords to FOAF metadata, one problem is that we are not sure that the keyword is really relevant metadata of a person. Although "Dan Brickley" and "ILRT" co-occur in many pages, they might be no relation between them in the real life. Therefore, someone should evaluate the propriety of a keyword as actual metadata. One approach to solve this problem would be an interactive system of FOAF file creation. Reusing and modifying a keyword as a candidate of FOAF metadata, a user can easily create his or her own or another person's FOAF file.

Another problem of using keywords as FOAF metadata is to decide a certain keyword property. In our experiment the property label was given manually to each keyword. However it is not efficient to put a property to numerous extracted keywords. One approach to automatically decide the property of a keyword is to use the machine learning technique that is often used in the entity extraction research.

Table 1: Higher-ranked keywords of "Dan Brickley"

Dan Brickley	Dan Connolly
Libby Miller	Jan Grant
FOAF	RDF Interest Grop
Semantic Web	xmlns.com/foaf
Dave Beckett	RDF
RDFWeb	Eric Miller
ILRT	FOAF Explorer

Table 2: The distribution of property labeled to higher-ranked keywords

Property	The number	per person
Technical term	695(49.9%)	27.8
Name	476(34.1%)	19.04
Organization	71(5.1%)	2.84
URL	65(4.6%)	2.6
Project	35(2.5%)	1.4
Event	31(2.2%)	1.24
Community	10(0.7%)	0.4
Position	9(0.6%)	0.36
Total	1392	

5 Conclusion

The Web holds much personal information that can be used as FOAF metadata. This paper proposes a novel keyword extraction method to extract personal information from the Web. Our results show the important possibility of using extracted keywords as FOAF metadata. Importantly, our method can capture the personal information in a different context. This allows us to obtain various person-related metadata.

Because the Web is such a large information resource, its information runs the gamut from useful to trivial. It presents the limitation that it must be publicly available on the Web. For further improvement of the proposed method, we must analyze "what" information of "who" in the Web, and its reliability . In this regard, blog and social network sites are noteworthy subjects for the future.

References

[Brickley and Miller, 2004] http://xmlns.com/foaf/0.1/

[Nakagawa *et al.*, 2003] http://gensen.dl.itc.u-tokyo.ac.jp/win.html