Keyword Extraction from the Web for Personal Metadata Annotation

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, Japan
y.matsuo@carc.aist.go.jp
[3] École Polythechnique Fédérale de Lausanne, Switzerland
junichiro.mori, boi.faltings@epfl.ch

Abstract. With the currently growing interest in the Semantic Web and Social Networking, personal metadata is coming to play an important role in the Web. This paper proposes a novel keyword extraction method to extract personal 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 are useful for personal metadata creation. We also discuss the annotation of personal metadata and application to the Semantic Web.

1 Introduction

The Semantic Web[2] is a new paradigm which brings "structure" to the meaningful content of the Web. With currently growing interest in the Semantic Web and new standards for metadata description such as the Resource Description Framework (RDF)[13], metadata is gradually gaining popularity in the Web.

Another recent trend in Web development is "Social Networking"[7]. Social Networking sites are community sites through which users can maintain an online network of friends or associates for social or business purposes. Numerous Social Networking sites have been launched recently.

As seen in Social Networking, a user itself is gradually coming to play a central role in the Web contents (e.g. In "Weblog", variety of contents is created by a user). With these recent Web trends, expressing metadata about people and the relations among them is recently gaining interest. In fact, some vocabularies and frameworks for personal metadata description have being developed [5][9][15][16].

Using these vocabularies, a user is gradually creating his or her personal metadata. However, as a major problem of the Semantic Web is the metadata annotation, personal metadata must also overcome the problem and need methods that facilitate and accelerate metadata annotation [8][10]. Although there are some supporting tools to create personal metadata such as Foaf-a-Matic[4], this tool facilitates only basic descriptions.

[4] http://www.ldodds.com/foaf/foaf-a-matic.html

Considering personal metadata, 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 Weblog. In fact, we can expect that these pages contain a lot of personal metadata even including information that we would not expect to find. Therein, questions are:

- What kind of personal metadata are in the Web?
- What kind of Web page contains personal metadata?
- How are extracted metadata applied to semantic annotation?

Considering these points, one of our research goals is to extract personal metadata from the Web and apply them to semantic annotation. As a preliminary report to achieve this goal, we propose a novel keyword extraction method to extract personal information from the Web.

The remainder of this paper is organized as follows: section 2 describes the proposed keyword extraction method using an actual example. In section 3, we show the extracted keywords and analyze them. In section 4, we discuss the annotation of personal metadata. Section 5 contains related works. Finally, we address future works and conclude this paper in section 7.

2 Keyword Extraction

2.1 Extraction of the Initial Term Set for Keyword

As an experimental attempt, we extracted the keywords of Program Committee members of SemAnnot 2004 Workshop (There are 28 members including chair persons). First, we need to acquire Web pages that contain information of respective committee members and their mutual relationships. A simple way of acquiring those Web pages is to use a search engine. It is reasonable to use a search engine because it can search many Web pages in less than a few seconds. It also tracks the temporal variance of the Web. In this experiment, we used Google[5], which currently addresses data from 4 billion Web pages.

We first put each person's full name to a search engine (name is quoted with double quotation such as "Siegfried Handschuh") 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 search result documents include not only html files but also other file types such as .pdf, .doc, .xls, .ppt. In this experiment, we used only html files. Furthermore, we did not use metadata indicators in an html file such as META tags and RDF. In the future, we are planning to use other file types along with html files that already have been attached metadata.

The html files, at to a maximum of 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 [14]. Termex extracts terms

[5] http://www.google.com

from POS data based on statistical information of conjunctions between parts of speech. Termex[6] can also extract nominal phrases that include more than two nouns such as "Annotation tool". After the whole procedure of extracting the term set, we extracted about 1000 terms per person on the average. The relevant keyword of each person is chosen from these terms. Figure 1 shows steps of the proposed keyword extraction.

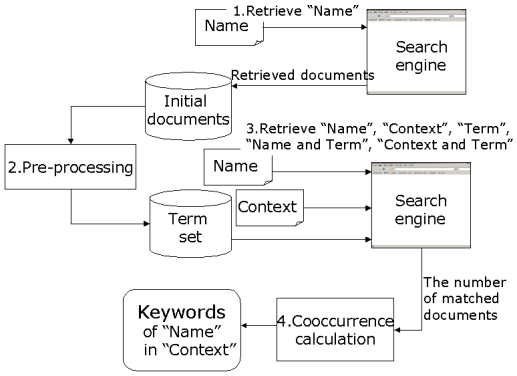

Fig. 1. Procedure of keyword extraction

2.2 Keyword Extraction Using Co-Occurrence Information

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. This subsection explains the scoring method that gives relevance as a personal keyword to the term.

Term relevance based on Co-Occurrence The simple 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 documents of a search engine result. For example, assume we are to measure the relevance of name N (e.g. "Siegfried Handschuh") and term w (e.g. "Annotation"). Here, w is the term in the term set W extracted from the initial documents of the person named "N" . 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 obtain the number of retrieved documents for each, $|N|$ and $|w|$. Then, the relevance between the name N and the term w, denoted by $r(N, w)$, is

[6] Termex can be used for both Japanese and English POS data

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.

Keyword of person As described in a previous subsection, the term set of a person is extracted from various Web pages. Although the Web pages contain a person's name in the text, each page may contain personal information in different contexts. For example, imagine that one person, named "Tom", is both a researcher and a artist, we can expect that his name may appear not only in academic-related pages, but also in other pages related to his art activities. Even among his academic-related pages, there might be different pages depending on his acquaintances, affiliations, and projects. In this way, different Web pages reflect different contexts of a person. Here, we introduce the notion of a context to extract the keyword that captures the context of a person.

To extract the keyword in relation to a certain context, we must estimate the relevance between the term and the context. If we replace the name N with the context C in the relevance, $r(N, w)$, we can obtain the relevance between context C and term w, $r(C, w)$, in the same manner. Then, the relevance of person N and term w in the context C, denoted by $score(N, C, w)$, is calculated as the following.

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

$$(\frac{r(N, w)}{MAX(r(N, w)} > threshold)$$

Therein, α denotes the relevance between the person and the context. For example, we can use $r(N, C)$ as α. $MAX(r(X, Y))$ is the maximum value of the Jaccard coefficient in the term set W. We define the $threshold$ for $r(N, C)$ to exclude terms that are not relevant for a person, but that have strong relation to the context. $threshold$ is decided based on heuristic method. The term w with the higher $score(N, C, w)$ is considered to be a more relevant keyword for person N in context C.

Regarding the "Tom" example, if we set "Art" as the context, we can get keywords related to his art activities. Alternatively, if we include his research project name as the context, keywords related to his project would be acquired.

Keywords showing a relation between persons If we consider the relation between two persons in terms of their contexts, one person can be regarded as a part of the context of another person. Hence, we can apply the previous formula to keyword extraction of the relations among persons as follows:

$$score(N1, N2, W) = \frac{r(N1, w)}{MAX(r(N1, w))} + \beta \frac{r(N2, w)}{MAX(r(N2, w))}$$

$$(\frac{r(N1, w)}{MAX(r(N1, w)}, \frac{r(N2, w)}{MAX(r(N2, w)} > threshold)$$

Therein, $N1$ and $N2$ denote each person's names in the relation. β is the parameter of relevance between persons, such as $r(N1, N2)$. This formula shows the relevance of person $N1$'s term w in relation to person $N2$.

As there are many contexts of a person, the relations among persons also have a variety of contexts. For example, the relation of two persons in the academic field might be coauthors, have the same affiliation, the same project; they may even be friends. The relevance of person $N1$'s term w in relation to person $N2$ in the context C, $score(N1, N2, C, w)$, is given as follows:

$$score(N1, N2, C, w) = score(N1, N2, w) + \gamma \frac{r(C, w)}{MAX(r(C, w))}$$

Therein, γ is the parameter of relevance between the persons and the context, such as $r(N1 and N2, C)$.

3 Keyword Analysis for Personal Metadata

3.1 Personal Metadata in Keywords

As an example of extracted keywords, Table. 1 shows higher-ranked extracted keywords of "Siegfried Handschuh". Each column in the table shows higher-ranked keywords based on Term Frequency Inverse Document Frequency (TFIDF), co-occurrence without the context, and co-occurrence with the context, respectively, from the left column.

In TFIDF-based keywords, we can find keywords that are related to the person such as "annotation" and "semantic". Nevertheless, there are many irrelevant words including general words. Because TFIDF is based on the frequency of word appearances in a text, it is difficult for a word to become higher-ranked in terms of relevance with another word. On the other hand, in co-occurrence-based keywords, general words are excluded and relevant words of each person appear in the rank list.

As explained in the previous section, the context can be considered in the keyword extraction. In this experiment, we used "Semantic Web" as the context. With this context, keywords are chosen in relation to one's activity about the Semantic Web. In the column of "Co-Occurrence with the context", we can find that context-related keywords come to appear in the rank list. The order of higher-ranked keywords also changes in relation to the context.

The column at the right side shows a property label for each keyword in "Co-Occurrence with the context". Considering a correspondence to existing personal metadata vocabularies such as FOAF, we have defined six property labels: Name (N), Technical term (T), Event (E), Organization (O), Project (P), URL. In order to analyze what kind of property is included in keywords, we annotated a property label to higher-ranked keywords of each person. Thereby, we acquired 1646 labeled keywords in total (about 60 keywords per person on average).

Table. 2 shows the distribution of property labels. Nearly half of higher-ranked keywords are occupied with names. Notwithstanding, it is noteworthy that other properties such as 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 indicates that if we extract about 60 higher-ranked

Table 1. Higher-ranked keywords of “Siegfried Handschuh” using TFIDF and co-occurance-based method

TFIDF	Co-Occurrence (without the context)	Co-Occurrence (with the context “Semantic Web”)	Property
Semantic	Siegfried Handschuh	Siegfried Handschuh	N
Siegfried Handschuh	Ljiljana Stojanovic	Ljiljana Stojanovic	N
Office	Nenad Stojanovic	Nenad Stojanovic	N
annotation	Marc Ehrig	Steffen Staab	N
Person	Julien Tane	Marc Ehrig	N
Web	Steffen Staab	Julien Tane	N
Karlsruhe	Daniel Oberle	Daniel Oberle	N
Konstanz	Valentin Zacharias	Valentin Zacharias	N
E223	Andreas Hotho	Andreas Hotho	N
CREAM	relational metadata	**Semantic Web**	T
karlsruhe.de	annotation of web pages	relational metadata	T
message	Knowledge Markup	annotation of web pages	T
Inf.wiss	Large Scale Semantic Web	Knowledge Markup	T
knowledge	automatic CREAtion of Metadata	Large Scale Semantic Web	T
Webmaster	Annotation Workshop	Knowledge Markup Workshop	E
Appointment	Knowledge Markup Workshop	**International Semantic Web Conference**	E
AIFB	KCAP	KCAP	E
Katarina Stanoevska	AIFB	AIFB	O
Beat Schmid	University of Karlsruhe	University of Karlsruhe	O
Alexander Maedche	OntoAgents	OntoAgents	P

keywords of one person, on average we can obtain about 30 names of his acquaintance, 2 or 3 related organizations, and 1 or 2 projects. These numbers nearly match our research activity and show the possibility of using keywords for personal metadata. In this analysis, we took many keywords together as “technical terms”. If we classify each keyword more precisely, we could discover other personal metadata in keywords.

3.2 Personal Metadata in the Web

To further explore the possibility of personal metadata extraction from the Web, we analyzed which Web pages include a higher-ranked keyword. First, we classified all 280 Web pages (10 per person) that were used to extract the initial term set. Thereby, we prepared the 11 categories shown in Table. 3. “Personal page” includes personal Web pages of the affiliation or one’s own domain. “Other page” includes uncategorized pages and non-html pages such as .pdf and .ppt files. “Event page” includes conference, workshop, and meeting pages. ML log is the email exchanged in a mailing list. DBLP [7] is the online bibliography of Computer Science papers. As seen in the table, “Personal page” is the most dominant type of Web page. Because a person’s name was used as a query, it is natural that we obtain a personal page in a search result.

[7] http://www.informatik.uni-trier.de/~ley/db/

Table 2. Distribution of properties labeled to higher-ranked keywords

Property	Number	Per person
Name	767 (46.5%)	27.3
Technical term	613 (37.2%)	21.8
Event	105 (6.3%)	3.7
Organization	73 (4.3%)	2.6
Project	48 (2.5%)	1.7
URL	40 (2.4%)	1.4
Total	1646	

Table 3. Classification of the Web page type

Web Page	Number
Personal page	73 (26.0%)
Other page	42 (15.0%)
Event	32 (11.4%)
ML log	27 (9.6%)
Online paper	26 (9.2%)
DBLP	22 (7.8%)
Organization	17 (6.0%)
Project	16 (5.7%)
Book	11 (3.9%)
Publication list	8 (2.8%)
Weblog	6 (2.1%)
Total	280

Table 4. Distribution of each keyword property to each Web page type

Web page	Name	Technical Term	Event	Organization	Project	URL
Personal page	**234 (19.3%)**	**199 (24.0%)**	**31 (24.0%)**	**35 (36.8%)**	**30 (44.1%)**	**14 (25.9%)**
Other page	42 (3.4%)	15 (1.8%)	4 (3.1.%)	3 (3.1%)	1 (1.4.%)	2 (3.7%)
Event	**223 (18.3%)**	**171 (20.6%)**	**29 (22.4%)**	**25 (26.3%)**	1 (1.4%)	**14 (25.9%)**
ML log	165 (13.6%)	122 (14.7%)	11 (8.5%)	**16 (16.8%)**	**8 (11.7%)**	**8 (14.8%)**
Online paper	12 (0.9%)	33 (3.9%)	4 (3.1%)	1 (1.0%)	1 (1.4%)	2 (3.7%)
DBLP	**314 (25.9%)**	**189 (22.8%)**	**38 (29.4%)**	0	**11 (16.1%)**	0
Organization	66 (5.4%)	45 (5.4%)	4 (3.1%)	8 (8.4%)	5 (7.3%)	9 (16.6%)
Project	46 (3.7%)	13 (1.5%)	0	5 (5.2%)	5 (7.3%)	5 (9.2%)
Book	18 (1.4%)	7 (0.8%)	1 (0.7%)	0	0	0
Publication list	85 (7.0.%)	24 (2.8%)	6 (4.6%)	1 (1.0%)	1 (1.4%)	0
Weblog	7 (0.5%)	10 (1.2%)	1 (0.7%)	1 (1.0%)	5 (7.3%)	0
Total	1212	828	129	95	68	54

Table 4 shows which category of Web page a higher-ranked keyword belongs in (a keyword may appear in more than one category). Specifically examining each column, we find which kind of Web page each property is included in. Moving the focus to a row in the table, we can find what kind of property each Web page category includes.

Although name entities can be acquired most from “Personal page”, DBLP is also a good information resource to extract a name entity. DBLP contains coauthor information of a paper. Therefore, the extracted name is related to one’s acquaintance in a research activity. “Event page”, such as conference, workshop, is a information resource of various personal information. However, because the Event page is not specified to a certain person, “Personal page” gives more accurate information about each person.

Overall, “Personal page” is a good information source for personal metadata such as names, organizations, and projects. Event page and DBLP provide metadata that are

related to personal research activities such as coauthors, projects, and events including conferences and workshops.

4 Annotation of Personal Metadata

Our keyword extraction method can be applied to semantic annotation in following ways.

- **Annotation for Web page :** As our analysis showed, our personal keyword extraction method offers strong potential for personal metadata extraction from the Web. Extracted personal metadata can be applied to partially annotate the Web pages using metadata description framework such as the RDF[13]. Because metadata are given the relevancy in relation to a person, annotated Web pages can be used in many applications such as Information retrieval and Information integration. For example, using annotated Web pages, the search engine that supports the Semantic Web could answer to following question:
  - Who knows this person?
  - Who is involved in this project?
  - Who knows this research topic well?
  - Which pages include this person's information?
- **Annotation for Personal Metadata File :** Extracted personal metadata is used not only for annotating a Web page, but also for annotating a personal metadata file. As one emerging personal metadata standard, "Friend of a Friend", FOAF[5], defines an RDF vocabulary for expressing metadata about people, the relation among them, and the things they create and do. FOAF provides a way to create machine-readable personal documents on the Web, and to process them easily through merging and aggregating them. Because extracted metadata are easily incorporated in FOAF, we can facilitate the creation of FOAF documents.

This paper presents discussion of the importance of a person's context in keyword extraction. The context often defines the properties. Currently, there is no FOAF vocabulary to define a context. In addition to FOAF, there are many vocabularies and framework for personal metadata such as Topicmaps [9], RDF-vCard [16], Person class of DAML+OIL [15]. However, none of them address the notion of a personal context. One way to introduce a personal context to those metadata frameworks is to prepare schema that corresponds to respective contexts. Regarding the expression of personal metadata, we need further consideration to make the metadata expressive and usable.

5 Related works

Aiming at extracting and annotating personal metadta, our method is regarded as one of Information Extraction(IE) methods supporting a semantic annotation. Up to now, many IE methods rely on predefined templates and linguistic rules or machine learning techniques to identify certain entities in text documents[12]. Furthermore, they usually define properties, domains, or ontology beforehand. However, because we try to extract various information from different Web pages, we don't use predefined restrictions in the extraction.

Some previous IE researches have addressed the extraction and annotation of personal metadata. In [1], they proposes the method to extract a artist information, such as name and date of birth, from documents and automatically generate his or her biography. They attempt to identify entity relationships, metadata triples (subject-relation-object), using ontology-relation declarations and lexical information. However, Web pages often include free texts and unstructured data. Thereby, capturing entity relationships becomes infeasible because of lacking regular sentences. Rather than focusing on the entity relationship, we find the entity in the Web pages based on the relevance in relation to a person.

In [6], they address the extraction of personal information such as name, project, publication in a specific department using unsupervised information extraction. It learns to automatically annotate domain-specific information from large repositories such as the Web with minimum user intervention. Although they extract various personal metadata, they don't consider the relevance of extracted metadata. Because extracted metadata in our method have the relevance, they can be used as reliable initial seeds for bootstrap learning for automatic annotation in their method.

Although the aim is not extracting personal metadata, in [11], they proposes the method to extract a domain terminology from available documents such as the Web pages. This method is similar to our one in terms of that terminology are extracted based on the scoring measure. However, their measure is based not on the co-occurrence but on the frequency. Furthermore, they focus on the domain-specific terms rather than personal metadata and the method is domain dependent. In our method, we can capture the various aspects of personal metadata even from different domain resources using the notion of a context.

6 Future works and Conclusion

To apply our keyword extraction methods to personal metadata annotation, we must consider and solve following points in the future.

- **Evaluation of personal metadata :** One problem is that we are not sure that the extracted metadata are true. Although two terms co-occur in many Web pages, they might not have any relation. Therefore, someone should evaluate the propriety of a keyword as actual metadata. One approach to solve this problem would be an interactive annotation system[3]. Reusing and modifying a keyword as a candidate of personal metadata, a user can easily annotate personal metadata.
- **Entity recognition of keywords :** Another critical problem 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 techniques in the entity recognition research[4].
- **Privacy problem of information extraction from the Web :** A person sometimes does not know that his or her information is extracted from the Web only by name. Therefore, we should take care not to intrude on a user's privacy even in information extracted from the Web. We must clarify the use of the information only for useful services for a user.

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

References

1. H. Alani et al. Automatic Extraction of Knowledge from Web Documents. *In Workshop of Human Language Technology for the Semantic Web and Web Services, 2nd International Semantic Web Conference*, Sanibel Island, Florida, USA, 2003.
2. T. Berners-Lee, J. Hender, O. Lassila. The Semantic Web. Scientic American, 2001.
3. F. Ciravegna, A. Dingli, D. Petrelli, and Y. Wilks. User-system cooperation in document annotation based on information extraction. *In Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, EKAW02*, Springer Verlag, 2002.
4. H. Cunningham et al. GATE:A Framework and Graphical Development Environment for Robust NLP Tools and Application. *In Proceedings of the 40th Anniversary Meeting Assoc. for Computational Linguistics(ACL2002)*, East Stroudsburg, Pa., 2002.
5. Dan Brickley and Libby Miller. FOAF: the 'friend of a friend' vocabulary. http://xmlns.com/foaf/0.1/, 2004.
6. A. Dingli, F. Ciravegna, D. Guthrie, Y. Wilks. Mining Web Sites Using Unsupervised Adaptive Information Extraction. *In Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, 2003.
7. L. Garton, C. Haythornthwaite, and B. Wellman. Studying online social networks. In Doing Internet Research, S. Jones, Ed. Sage, Thousand Oaks, CA, pp. 75–105, 1999.
8. J. Kahan and M. R. Koivunen. Annotea: An open rdf infrastructure for shared web annotations. *In Proceedings of the 10th International WWW Conference*, pp.623–632, 2001.
9. Lars Marius Garshol. Living with topic maps and RDF. http://www.ontopia.net/topicmaps/materials/tmrdf.html, 2003.
10. S. Staab, A. Maedche, and S. Handschuh. An Annotation Framework for the Semantic Web. *In Proceedings of 1st International Workshop MultiMedia Annotation,* 2001.
11. P. Velardi, M. Missikoff, R. Basili. Identification of relevant terms to support the construction of Domain Ontologies. *In ACL-EACL Workshop on Human Language Technologies*, Toulouse, France, 2001.
12. R. Yangarber and R. Grishman. Machine Learning of Extraction Patterns from Unanotated Corpora: Position Statement. *Workshop Machine Learning for Information Extraction*, IOS Press, Amsterdam, pp.76–83, 2000.
13. Resource Description Framework(RDF) Schema Specification. In *W3C Recommendation*, 2000.
14. http://gensen.dl.itc.u-tokyo.ac.jp/win.html
15. DAML Ontology Library. http://www.daml.org/ontologies/
16. Representing vCard Objects in RDF/XML. http://www.w3.org/TR/2001/NOTE-vcard-rdf-20010222/