Finding User Semantics on the Web using Word Co-occurrence Information

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

[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

Abstract. With the currently growing interest in the Semantic Web, describing user semantics to model users and their social relationships is coming to play an important role. This paper proposes a novel keyword extraction method to extract user semantics from the Web. Based on co-occurrence information of words, the proposed method extracts relevant keywords depending on the context of a person. Our evaluation shows better performance to $tfidf$-based keyword extraction. We also discuss application of our method in the Semantic Web.

1 Introduction

With currently growing interest in the Semantic Web [2] and new standards for metadata description such as the Resource Description Framework (RDF) [15], metadata has gradually been becoming popular in the Web. Another recent trend in the Web is that the user is gradually coming to play a central role in Web contents. For example, in Weblog variety of contents is created by a user. And several Social Networking sites through which users can maintain an online network of friends or associates for social or business purposes have been launched recently. Therein, data about millions of people and their connections is publicly available on the Web.

With these recent Web trends, expressing semantics about people and their relationships has been gained interest. The Friend of a Friend (FOAF) project [3] is one of the Semantic Web's largest and most popular ontologies [6]. It is essentially a vocabulary for describing people and whom they know. The FOAF ontology isn't the only one people use to publish social information on the Web. For example, it is reported that more than 360 RDF Schema or OWL classes defined with the local name "person" [1]. In fact, many vocabularies and frameworks for user semantics have being developed [16][5][11].

Users are begining to accept FOAF and its extensions as something of a standardized ontology for representing user semantics on the Semantic Web. However, as a major problem of the Semantic Web is in metadata annotation, metadata for users must also overcome the problem so that every user can easily annotate his or her data. The key clue to facilitate and accelerate metadata generation is to reuse much information which already has existed on the Web. In fact, while some FOAF files are from users

[1] http://swoogle.umbc.edu

who have authored their own data, others are from Web sites that publish data from their databases using the FOAF ontology. For example, imagine a researcher: that researcher's information can be found in an affiliation page, a conference page, an online paper, or in a Weblog.

One of our research goals is to find user semantics which already have been on the Web, and apply Semantic Web technologies to them. Therein, question is how we can find user's relevant information. In this paper, we propose a novel keyword extraction method to extract personal information from the Web. The proposed method is based on the statistical feature of word co-occurrence. The basic idea is a following: if a word co-occurs with a person's name in many Web pages, the word might be a relevant keyword about his or her information. Importantly, our method extracts relevant keywords depending on the context of a person.

The remainder of this paper is organized as follows: section 2 describes the proposed keyword extraction method. In section 3, we evaluate the method. In section 4, we discuss the limitation and application of our method in the Semantic Web. In section 5, we compare our method with related works. Finally, we conclude this paper in section 6.

2 Keyword Extraction

2.1 Basic Idea

The simple approach to find someone's keyword is to use word co-occurrence information. Here, we define co-occurrence of two words as word appearance in the same Web page. If two words co-occur in many pages, it is assumed that those two have a strong relation. The co-occurrence information is acquired by the number of retrieved documents of a search engine result. For example the search result of a query "Alfred Kobsa and User Modeling" returns about 3100 documents while about 450 documents for a query "Alfred Kobsa and Software engineering". In this manner, we can guess that "User Modeling" is more relevant to "Alfred Kobsa" than "Software engineering". Our first hypothesis that:

Hypothesis1: The word that co-occurs with a person's name in many Web pages could be his or her keyword.

Although we can find many Web pages that contain a person's name, each page may contain personal information in different contexts. For example, imagine that one person who is both a researcher and an 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. We define a context word as word that describes someone's context. For example, "Art" and "Research" can be respectively context words for his art activities and research activities. Our second hypothesis that:

Hypothesis2: The word that co-occurs with a context word in many Web pages could be the keyword in the context.

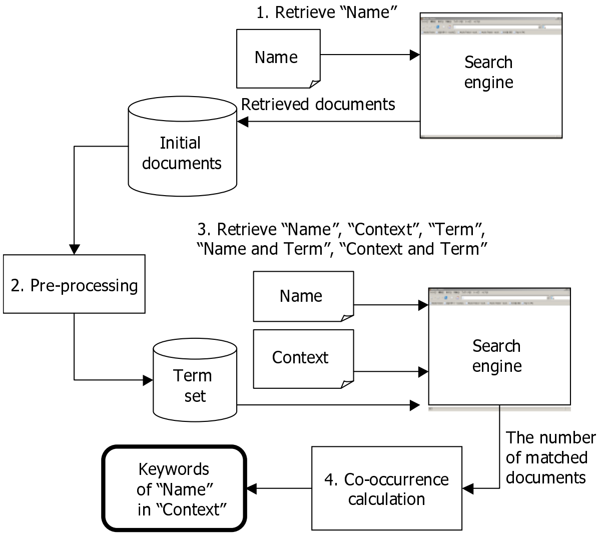
1. Retrieve "Name"
Name
Search engine
Retrieved documents
Initial documents
2. Pre-processing
Term set
3. Retrieve "Name", "Context", "Term", "Name and Term", "Context and Term"
Name
Context
Search engine
The number of matched documents
4. Co-occurrence calculation
Keywords of "Name" in "Context"

Fig. 1. Procedure of keyword extraction

2.2 Scoring Keywords based on Word Co-occurrence

Figure 1 shows procedures of the proposed keyword extraction. The proposed method has two main steps: (1) First step is to extract words that co-occur with a person's name in Web pages. (2) Second step is to give a score to each word using the degree of word co-occurrence in Web pages.

First, in order to extract words that co-occur with a person's name, we put his or her full name to a search engine [2]. As a search engine, we used Google [3] which currently addresses data from more than 8 billion Web pages. From the search result, we used the top 10 html files [4] as initial documents. The initial documents are pre-processed with html-tag deletion and part-of-speech (POS) tagging . Then, using the term extraction tool, Termex [5], we extract terms from pre-processed html files. Termex extracts terms from POS data based on statistical information of conjunctions between parts of speech. It can also extract nominal phrases that include more than two nouns such as "User Modeling". After the whole procedure of extraction, we extract about 1000 terms per person.

Based on the previous basic idea, the relevant keyword for a person is chosen based on word co-occurrence information. As a measure of co-occurrence, we use Jaccard coefficient that captures the degree of co-occurrence of two terms by their mutual degree

[2] Because of the same-name problem that is discussed in Section 4, we add a person's affiliation to the query when search results are taken up by another person of the same name.

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

[4] Currently, we don't use other file types such as .pdf, .doc

[5] http://gensen.dl.itc.u-tokyo.ac.jp/win.html

of overlap. Jaccard coefficient is often used to evaluate tie strength between two objects [9]. Assume we are to measure the relevance of name n and term w. We first put a query, "n and w", to a search engine and obtain the number of retrieved documents that is denoted by $|N \cap W|$. Therein, N denotes a Web page set that includes n and W denotes a Web page set that includes 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 name n and term w, denoted by $J(n, w)$, is approximated by the following Jaccard coefficient.

$$J(n, w) = \frac{|N \cap W|}{|N \cup W|} = \frac{|N \cap W|}{|N| + |W| - |N \cap W|}$$

To extract the keyword in relation to a certain context, we need to estimate the relevance between the term and the context. If we replace the name n with the context c in the relevance, $J(n, w)$, we can obtain the relevance between context c and term w, $J(c, w)$, 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.

$$Score(n, c, w) = J(n, w) + \alpha J(c, w)$$

Therein, α denotes the relevance between the person and the context. We define threshold k for $J(n, c)$ to exclude terms that are not relevant for a person, but that have strong relation to the context. α and k are currently decided based on a heuristic method [6]. The term w with the higher $Score(n, c, w)$ is considered to be a more relevant keyword for person n in context c.

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 relation between persons as follows:

$$RScore(n1, n2, c, w) = Score(n1, n2, w) + \beta\ J(c, w)$$

Therein, $n1$ and $n2$ denote each person's names in the relation. Context c can be considered in the relation between persons. β is the parameter of relevance between the persons and the context. This formula shows the term relevance of the relation between person $n1$ and $n2$ in the context c.

As an example of extracted keywords, Table 1 shows higher-ranked keywords of "Mitsuru Ishizuka" [7] who is a co-author of this paper. Each column in the table shows higher-ranked keywords based on $tfidf$, co-occurrence without the context, co-occurrence with the context "Artificial Intelligence", respectively, from the left column. Table 2 shows higher-ranked keywords with the context "University". Note that depending on the context word, context-related words (in bold type) come to appear in higher-ranked keywords. The order of higher-ranked keywords also changes in relation to the context. As an example of the relation keywords, Table 3 shows higher-ranked keywords between "Mitsuru Ishizuka" and "Yutaka Matsuo".

[6] For keywords in Table 1-3, we used as $\alpha = avg(J(n, w))/(3 * avg(J(c, w)))$, $k = 0.001$

[7] He is a chairperson of the Japanese Artificial Intelligence society

Table 1. Higher-ranked keywords of "Mitsuru Ishizuka" using $tfidf$ and co-occurrence based method

$tfidf$	Co-Occurrence (without the context)	Co-Occurrence (with the context "Artificial Intelligence")
University of Tokyo	Yutaka Matsuo	**AI society**
University	Hiroshi Dohi	Yutaka Matsuo
JAVA application	Character Agent	**Natural Language**
Character	Koichi Hashida	Koichi Hashida
Scenario Emergence	Life-like Interface	Hiroshi Dohi
Research Institute	Naoaki Okazaki	Character Agent
Electronics	University of Tokyo	Life-Like Interface
Microsoft	Life-like Agent	Naoaki Okazaki
Iba laboratory	Hypothetical Reasoning	University of Tokyo
Yukio Osawa	Sadao Kurohashi	Life-like Agent
Program Committee	Life-like Internface	**AI journal**

Table 2. Higher-ranked keywords of "Mitsuru Ishizuka" with the context "University"

Co-occurrence with the context "University"
Yutaka Matsuo
Graduate School of Engineering
Hiroshi Dohi
Character Agent
Life-Like Interface
Artificial Intelligence
University of Tokyo
Faculty of Engineering
Life-life agent

Table 3. Higher-ranked keywords of the relation between "Mitsuru Ishizuka" and "Yutaka Matsuo"

Co-occurrence with the context "Artificial Intelligence"
National Institute of Advanced–
–Industrial Science and Technology
Artificial Intelligence
Ishizuka Laboratory
Naoaki Okazaki
Hiroshi Dohi
Yukio Osawa
Koishi Hashida
Naohiro Matsumura

3 Evaluation

To evaluate the proposed method and validate our hypotheses, we extracted keywords of 10 Artificial Intelligence researchers. For each subject, we showed keywords that are extracted from the Web by tf (term freqequcy), $tfidf$ (term frequency inverse document frequency), $co\text{-}occur$ (co-occurrence without the context), and our method (co-occurrence with the context). $tfidf$ is a method widely used by many keyword extraction systems to score individual words within text documents in order to select concepts that accurately represent the content of the document. $tfidf$ score of a word can be calculated by looking at the number of times the word appears in a document and multiplying that number by the log of the total number of documents (corpora) divided by the number of documents that the word resides in. As corpora, we used 3981 html files which are collected from the search results of 567 Japanese AI researchers'name.

Table 4. Precision, Coverage, Context Precision for 6 subjects

Method	tf	$tfidf$	*co-occur*	**ours**
precision	0.13	0.18	0.60	**0.63**
coverage	0.20	0.24	0.48	**0.56**
context precision	0.05	0.04	0.15	**0.19**

The idf is defined by $log(D/df(w))$+1, where D is the number of all documents and $df(w)$ is the number of documents including word w. In co-$occur$, keywords were extracted based on only co-occurrence between the name and term. In our method, we used "Artificial Intelligence" as the context word .

Using each method we extracted and shuffled the higher-ranked 20 terms derived each method. Then, the subjects were asked following three instructions:

I1 Check terms that are relevant to your research activities.
I2 Choose five terms that are indispensable for your research activities.
I3 Check terms that are relevant to your research activities from the viewpoint of Artificial Intelligence.

Precision was calculated by the ratio of the checked terms to 20 terms derived by each method (I1). Coverage of each method was calculated by taking the ratio of the indispensable terms included in the 20 terms to all the indispensable terms (I2). It is desirable to have the indispensable term list beforehand. However, it is very demanding for subjects to provide a keyword list without seeing a term list. In our experiment, we allowed subjects to add any terms to the indispensable term list even if they were not derived by any of the methods. Context precision is an evaluation criterion to measure how well context-related keywords are extracted. It is calculated by the ratio of the checked terms to 20 terms derived by each method (I3). Results are shown in Table 4. Compared with tf and $tfidf$, co-occurrence based methods exceed both in precision and coverage. tf and $tfidf$ select terms that appear frequently in the document (although $tfidf$ considers frequencies in other documents). On the other hand, co-occurrence based methods extract keywords in relation to another term even if they do not appear frequently. This leads to better performance of co-occurrence based methods. With regard to context precision, our method that considers the context performs better than other methods. This means that our method can extract keywords in relation to the context (in this case, "Artificial Intelligence") better than other methods.

4 Discussion

4.1 Limitation

One problem of retrieving a person's name in a search engine is the case of two or more people having the same full name. One way to alleviate this same-name problem is to add a person's affiliation to the query. However, this degrades the coverage of search results. In particular, this makes the search focus more on one's activity in relation to

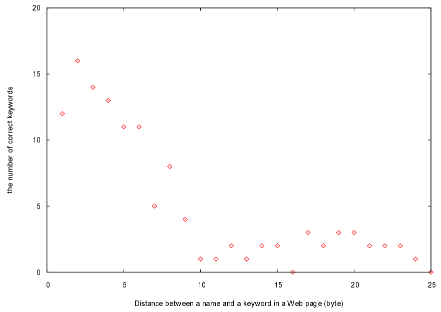

Fig. 2. Distance between a name and a keyword vs. the number of correct keywords

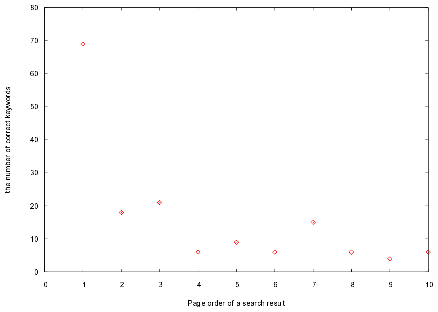

Fig. 3. Page order of a search result vs. the number of correct keywords

the affiliation. It also excludes other contexts. It is necessary to solve the same-name problem without losing various contexts of people.

While it is easy to obtain the information about researchers, ordinary people hardly expose their information in the Web. For further improvement of the proposed method, we must analyze what information is available about who in the Web, and its reliability. In this regard, Weblog and Social Network sites where people write variety of information are noteworthy subjects for the future. On the other hand, we should take care not to intrude on a user's privacy even in information extracted from the Web. A person sometimes does not know that his or her information is extracted from the Web only by name. We must clarify the use of information only for useful services for a user.

The more we use Web pages and select kewywords from whole pages, the greater the number of queries must be posted to a search engine later. For that reason, we used the top 10 documents of search results to reduce the load of using a search engine. However, it is arguable how many documents of search results are to be used and whether the distance between a person's name and a keyword in a document is taken into account. Figure 3 shows the graph with the y-axis as the number of correct keywords that users chose in the experiment and the x-axis as the distance between a person'name and a keyword in a Web page. and Figure 4 shows the relation between the number of correct keywords and page-rank order of a search result. While most of keywords appear around a person's name and are contained in a higher-ranked page, some keywords are acquired indepedently of the distance and page-rank. We must examine these optimal parameters to extract adequate keywords.

4.2 Application to User Modeling in the Semantic Web

As shown Tables 1-3, keywords include various personal information such as person's name, organization, research project, and research interest. These are easily incorporated in personal metadata, for example, FOAF properties:

foaf:knows, foaf:currentorganization, foaf:currentproject, foaf:interest

Currently, we are developing a method to automatically classify properties of keywords and generate personal metadata [10]. Figure 5 shows a FOAF file which is generated

```
<foaf:Person>
<foaf:mbox rdf:resource=””/>
<foaf:name>Mitsuru Ishizuka</foaf:name>
<foaf:interest rdfs:label=”Character agent” rdf:resource=””/>
<foaf:currentProject rdfs:label =”Life-like interface” rdf:resource=””/>
<foaf:workplaceHomepage rdfs:label=”University of Tokyo” rdf:resource=””/>
<foaf:knows>
<foaf:Person>
<foaf:mbox rdf:resource=””/>
<foaf:name>Yutaka Matsuo</foaf:name>
......
```

Fig. 4. An example of FOAF file based on extracted keywords.

based on keywords. Using the keywords, we can facilitate the creation of a personal metadata file. We can also apply the metadata to partially annotate Web pages where keywords are extracted.

Once personal metadata or annotated Web pages is acquired, it can be very useful for a user profile in the Semantic Web. User adaptive system can use the profile for service such as recommendation and personalization. For example, the system might adapt to following user requests: Who knows this person? Who is involved in this project? Who knows this research topic well? Which pages include this person's information?

We addressed the importance of a person's context in keyword extraction. The context often defines kinds of properties. Currently, there is no FOAF vocabulary or its extensions to define a 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 user semantics, we need further consideration to make it expressive and usable.

5 Related works

Aiming at extracting keywords, our method is regarded as an IE (Information Extraction) method. Up to now, many IE methods have relied on predefined templates and linguistic rules or machine learning techniques to identify certain entities in text documents [14]. For example, some previous IE researches have addressed the extraction of personal information. In [1], the authors propose a method to extract artist information from Web pages, such as name and date of birth, and automatically generate his or her biography. In [7], they address the extraction of personal information such as name, project, publication in a specific department using unsupervised information extraction. These methods usually define properties, domains, or ontology beforehand. In contrast, we extract various information based on statistical word co-occurrence using only a name and a search engine without any predefined restrictions.

Many keyword extraction methods for documents such as newspapers and scientific papers have been studied. In contrast to those documents, Web pages are too diverse and heterogeneous to apply the previous methods since they include free text and unstructured data, lack regular sentences. It is also difficult to apply probabilistic co-occurrence measures such as mutual information [4] and Log-Likelihood [8] since it is hard to estimate relevant population (the total number of Web pages and words) on the Web.

Some researches have focused on using a search engine to measure the stregth of relation between words [9, 12, 13]. They focus on extracting user's relationships or social network from the Web or the domain-specific terms. In our method, we can capture the various aspects of personal information from different Web pages using the notion of a context.

6 Conclusions

As users are gradually coming to play a central role in the Web contents, eliciting and representing personal information will increasingly be important in the user modeling research. In particular, with the currently growing trend toward the Semantic Web, expressing user semantics about people and the relations among them has been gained interest. This paper proposed a novel method to extract part of user semantics as keywords from the Web. Our evaluation showed better performance to $tfidf$-based keyword extraction.

Importantly, we use the Web as huge database and a search engine as its interface to obtain personal information in different contexts. While plenty of information is getting available on the Web, reusing and integrating online information of users will have significantly impact on personalization in the Semantic Web.

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. D. Brickley and L. Miller. FOAF: the 'friend of a friend' vocabulary. http://xmlns.com/foaf/0.1/, 2004.
4. K. W. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography, *Proc. of ACL '89, Association of Computational Linguistics*, 1989.
5. I. Davis and E. Vitiello Jr. RELATIONSHIP: A vocabulary for describing relationships between people, http://vocab.org/relationship/, 2004.
6. L. Ding, L. Zhou, T. Finin, A. Joshi. How the Semantic Web Is Being Used An Analysis of FOAF Documents. *In Proc. of the 38th Ann. Hawaii International Conference System Sciences*, 2005.
7. 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.
8. E. T. Dunning. Accurate methods for the statics of surprise and coincidence, *Computational Linguistics*, vol.19, No.1, pp.61–74, 1993.
9. H. Kautz, B. Selman and M. Shah. The Hidden Web. *AI Magazine*, Vo.18, No.2, pp.27–36, 1997.

10. J. Mori, Y. Matsuo, M. Ishizuka, and B. Faltings. Keyword Extraction from the Web for FOAF Metadata. *In Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and Semantic Web*, 2004.
11. Y. Matsuo, M. Hamasaki, J. Mori, H. Takeda and K. Hasida. Ontological Consideration on Human Relationship Vocabulary for FOAF. *In Proceedings of the 1st Workshop on Friend of a Friend, Social Networking and Semantic Web*, 2004.
12. Yutaka Matsuo, Hironori Tomobe, Koiti Hasida, Mitsuru Ishizuka. Mining Social Network of Conference Participants from the Web. *In Proceedings of the International Conference on Web Intelligence*, pp.190–194, 2003.
13. , Peter Mika, Bootstrapping the FOAF-Web: an experiment in social networking network mining, *Proc. of 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, 2004.
14. 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.
15. Resource Description Framework(RDF) Schema Specification. In *W3C Recommendation*, 2000.
16. Representing vCard Objects in RDF/XML. http://www.w3.org/TR/2001/NOTE-vcard-rdf-20010222/