Web Mining Approach for a User-centered Semantic Web

Junichiro Mori[1,2], Yutaka Matsuo[2], Koichi Hashida[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 (AIST), Japan
y.matsuo@carc.aist.go.jp, hashida.k@aist.go.jp

Abstract. In this paper, we propose a Web mining approach for the Semantic Web. The approach uses a search engine and the traditional web as a source of information to produce semantically rich information. In particular, we assess one community and obtain the social network and related information from the Web. As an example, we extract the social network of an academic society and show that extracted information can be incorporated into FOAF representation and utilized to measure the authoritativeness of a member in terms of social trust or individual trust. To demonstrate our Web mining approach in the real application, we show a researcher mining and retrieval system. Finally, we discuss the manner in which the Web mining approach contributes to availability to users of the Semantic Web.

1 Introduction

The Semantic Web [2] is designed to let users make explicit statements about any resource, and maintain that data themselves in an open and distributed manner. Several standards such as the Resource Description Framework (RDF) [18] and Web Ontology Language (OWL) [19] have been developed to realize the layer cake of the Semantic Web.

From the viewpoint of end users, expressing semantics about people and their relationships has garnered considerable interest. The Friend of a Friend (FOAF) project [4] is an extremely popular ontology of the Semantic Web [6]. It is essentially a vocabulary for describing people and whom they know. The FOAF ontology is not 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 are defined with the local name "person" [1]. In fact, many vocabularies for user semantics have been developed [20, 5, 12].

Supported by these user-side ontologies, users are gradually coming to adopt Semantic Web technologies both explicitly and implicitly. For example, in Weblogs, which are diary-like sites, users attach a FOAF profile to a Weblog and publish various contents by the RDF site summary (RSS). Some social networking sites that allow users to maintain an online network of friends associates for social or business purposes publish their users' social network data in FOAF format. Approaching the top of the Semantic

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

Web layers, calculation of a "Web of Trust" on a FOAF-based network is also proposed [10].

Users are beginning to accept FOAF and its extensions as something of a standardized ontology for representing user semantics on the Semantic Web. While some users are explicitly authoring their FOAF files, others use FOAF file that systems automatically create using their Web pages. In fact, considering the personal information that the FOAF vocabulary expresses, we find that much information is contained in the traditional Web. For example, imagine a researcher: that researcher's information might be in an affiliation page, a conference page, an online paper, or even in a Weblog. A method that can process the vast amount of information on the current "non-semantic" Web and can thereafter produce semantic information would facilitate and accelerate the use of the Semantic Web. For example, reusing existing sources of information on the Web would solve semantic annotation problems by helping users to create their metadata.

In this paper, we propose a Web mining approach for the Semantic Web. The approach uses a search engine and the traditional web as an information resource to produce semantically rich information. In particular, we examine one community and extract its social network and related information from the Web. As an example, we infer the social network of an academic society and show that extracted information can be incorporated in FOAF representation. It can then be used to measure the authoritativeness of a member as social trust or individual trust. To demonstrate our Web mining approach in an actual application, we show a researcher mining and retrieval system. Finally, we discuss how the Web mining approach contributes to user aspects in the Semantic Web.

The remainder of this paper is organized as follows: section 2 describes the proposed Web mining method and its application. Section 3 presents discussion of the Web mining approach for user aspects in the Semantic Web. Section 4 shows a comparison of our method with related works. Finally, we conclude this paper in section 5.

2 Web Mining Approach for the Semantic Web

This study specifically addresses one community and obtains the social network and related information from the Web. One reason for focusing on a community is that we believe that a huge "Web of Trust" over the entire Web comprises the superposed local "Webs of Trust" in each community to which a person or an organization belongs to.

Numerous communities exist in the physical world and online. We specifically examine an academic society: Japanese Society of Artificial Intelligence (JSAI). We choose JSAI because of its inherent availability of related information on the Web. Information related to this academic society in computer science is available online to a great degree. Another reason is that we are actually working mainly in JSAI so we can evaluate the extracted information. The following sections show how to automatically obtain JSAI members' social networks and related information from the Web.

2.1 Social Network Extraction

Before extracting the social network, we choose the participants to the last four annual JSAI conferences as active members of the JSAI community. Each active member of JSAI is represented as a node in a social network. A node is labeled with the name of its corresponding person.

Next, edges between nodes are added using Web information. A simple approach to measure the relevance of two nodes is to use word co-occurrence information. Herein, 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, assume we are to measure the relevance of two names "Junichiro Mori"(denoted x) and "Yutaka Matsuo" (denoted y). We first address two names $n1$, $n2$ as a query "$n1$ and $n2$" to a search engine and get $|N1 \cap N2|$ documents including those words in the text. Therein, N denotes a Web page set that includes a name n. Additionally, we make another query "$n1$ or $n2$" and obtain $|N1 \cup N2|$ matched documents. The relevance between $n1$ and $n2$ is approximated by the Jaccard coefficient $|N1 \cap N2|/|N1 \cup N2|$. If $n1$ and $n2$ have a strong relation, the retrieved documents might include $n1$'s and $n2$'s homepages, their publication pages, a laboratory's member list page, a conference program page and so on. In that case, $|N1 \cap N2|$ becomes large compared to $|N1 \cup N2|$. However, the Jaccard coefficient generally gives a famous person few edges because the denominator $|N1 \cup N2|$ is very large in comparison to $|N1 \cap N2|$. We can modify denominator $|N1 \cup N2|$ to $min(|N1|, |N2|)$, which places too much weight on a person with few edges. Therefore, the relevance of node $n1$ and $n2$ is represented by the following threshold-based Simpson coefficient:

$$R(n1, n2) = \begin{cases} \frac{|N1 \cap N2|}{min(|N1|, |N2|)} & if |N1| > k \; and \; |N2| > k, \\ 0 & otherwise \end{cases}$$

We set $k = 30$ for JSAI case. If we wish to estimate the co-occurrence more precisely to a person with small hits, we can pursue other alternatives to calculate statistical reliability. If relevance $R(n1, n2)$ of a node pair is larger than the given threshold, an edge is added with its weight equal to the relevance.

In the same manner as with the edge relation extraction, we can extract information of each node by considering the co-occurrence between the name and the term. For example, the search result of a query "Tim Berners-Lee and Semantic Web" returns about 76500 documents while about 9850 documents are returned for the query "Tim Berners-Lee and Software engineering". In this manner, we can infer that "Semantic Web" is more relevant to "Tim Berners-Lee" than "Software engineering". The term set of each node is acquired by retrieving the person's name that represents the node. Among the set, the term that often co-occurs with a person's name is chosen as his or her node keyword [2].

It is more useful to assign each edge a "label" for the relationship between two persons. For example, two nodes have the relation of "colleagues of the same research

[2] As a measure of co-occurrence, we use the Jaccard coefficient.