

# An Analysis of Researcher Network Evolution on the Web

Yutaka Matsuo<sup>1</sup>, Yuki Yasuda<sup>2</sup>

<sup>1</sup> National Institute of AIST, Aomi 2-41-6, Tokyo 135-0064, JAPAN

<sup>2</sup> University of Tokyo, Hongo 7-3-1, Tokyo 113-8656, JAPAN

**Abstract.** Social relation plays an important role in a real community. This paper overviews our project to automatically obtain a social network, especially a collaboration network of researchers, of a community from the Web. We operated our system at JSAI2003 and JSAI2004, and will operate at JSAI2005. Extracted networks in these three years are compared and analyzed to show the evolution of the researcher network.

## 1 Introduction

Social relation plays an important role in our daily life. People makes communications and shares information through social relations with others. Recently, a social network receives much attention on Web technology. For example, social networking sites (SNS) such as Friendster<sup>3</sup> and Orkut<sup>4</sup> grow rapidly and now over 200 SNSs exist. Information sharing on a social network is also proposed [4]. In the context of Semantic Web, social network is important to realize the web of trust, which enables to estimate information credibility and trustworthiness.

There are several ways to obtain social relation. One approach is to make a user describe his/her relation to others. In the social science, a network questionnaires are often made to obtain a social network e.g., asking “Please name your four closest friends.” Current SNSs realize such procedure online. On the other hand, automatic detection of relation is also possible from various sources of information such as e-mail archives, schedule data, and Web citation information [1, 8]. However, there is sometimes a concern of privacy: people do not want e-mail data to be analyzed, therefore the results are not easily utilized in other systems.

In the middle 90’s, Kautz and Selman develops a social network extraction system from the Web, called *referral web* [5]. This pioneering work focuses on co-occurrence of names on Web pages using a search engine. It estimates the strength of relevance of two persons X and Y by putting a query “X and Y” to a search engine: If X and Y are in strong relation, we can find a lot of evidences such as their homepages, lists of coauthors in technical papers and citations of papers, and organization charts. It has less problems about privacy because it uses publicly available information. Interestingly, a path from a person to a

<sup>3</sup> <http://www.friendster.com/>

<sup>4</sup> <http://www.orkut.com/>

person (e.g., from Henry Kautz to Marvin Minsky) is automatically obtained by the system.

Afterwards, with the development of WWW, more information on our daily activities becomes available online. Automatic extraction of social relation has much more potential and demand now compared to when referral web is first developed. This paper overviews advanced social network extraction from the Web, especially targeting researchers. Our algorithm can extract the strength of the relationship and also judge the type of relationships such as coauthorship, same laboratory, same project and co-attendance to a conference by text processing and machine learning.

Our system served at 17th and 18th Annual Conference of Japan Society of Artificial Intelligence (JSAI2003 and JSAI2004) and also will serve at JSAI2005. Social network is displayed at the conference site to show community overview and promote communication among participants. It is incorporated with scheduling support system and location information display system [10] in the ubiquitous computing environment.

In this paper, we take the JSAI case as an example. We analyze how the social network extracted from the Web evolves over time. Because Web information reflects the activity of researchers, analyzing the network brings us the insight how researcher's social relation changes and how Web technology can detect that.

This paper is organized as follows. The next section overviews how to obtain a social network from the Web. Section 3 addresses the analysis of social network evolution. Related works are described in Section 5. Finally, we conclude this paper.

## 2 Social Network Extraction

A social network is extracted by two steps. First we set nodes, then we add edges. In our approach, nodes in a social network, which represent persons, are given. In JSAI case for example, we collect authors and coauthors at the past five JSAI conferences and put them as nodes. We also collect affiliation (organization) with each name.

Next, edges between nodes are added using a search engine. For example, assume we are to measure the strength of relation between two names "Yutaka Matsuo" and "Mitsuru Ishizuka," we put a query

"Yutaka Matsuo" AND "Mitsuru Ishizuka"

to a search engine. In this case, we get 156 hits <sup>5</sup>, while we get only 7 hits if we put another query "Yutaka Matsuo" AND "Riichiro Mizoguchi". Because "Mitsuru Ishizuka" itself results in 1120 hits and "Riichiro Mizoguchi" results in 1130 hits, the difference of the hits by two names shows the bias of co-occurrence

---

<sup>5</sup> At the time point of January 8, 2004 by Google search engine. The query was in Japanese.

**Table 1.** Error rate of edge labels, precision and recall

class	error rate*	precision	recall**
Coauthor	4.1%	91.8% (90/98)	97.8% (90/92)
Lab	25.7%	70.9% (73/103)	86.9% (73/84)
Proj	5.8%	74.4% (67/90)	91.8% (67/73)
Conf	11.2%	89.7% (87/97)	67.4% (87/129)

\*: error rate of five-fold cross validation for 225 training data

\*\* : precision and recall for different 200 correct data.

of the two names: “Yutaka Matsuo” is likely to appear in Web pages with “Mitsuru Ishizuka” than “Riichiro Mizoguchi.” We can guess that Yutaka Matsuo has stronger relationship with Mitsuru Ishizuka. Actually in this example, Mitsuru Ishizuka was a supervisor of Yutaka Matsuo in his Ph.D. course. In this paper, we use “co-occur” to mean “appear in the same Web page.”

Our approach estimates the strength of relation by co-occurrence of two names. If the strength of relation is above a threshold, we add an edge between the corresponding two nodes. We use Simpson coefficient (or overlap coefficient) with a certain threshold based on several preliminary experiments.

$$R(X, Y) = \begin{cases} \frac{|X \cap Y|}{\min(|X|, |Y|)} & \text{if } |X| > k \text{ and } |Y| > k, \\ 0 & \text{otherwise} \end{cases}$$

Next, we judge types of relationships by applying text processing. Four classes of relationship among researchers are defined as follows:

- Coauthor: coauthors of a technical paper
- Lab: members of the same laboratory or research institute
- Proj: members of the same project or committee
- Conf: participants of the same conference or workshop

Each edge may have multi-labels. For example,  $X$  and  $Y$  have the relations of both “Coauthor,” and “Lab.”

We first fetch the top five pages retrieved by the query “ $X$  AND  $Y$ .” Then we extract features from the content of each page. In order to automatically judge classes of relationship, we take a machine learning approach. We apply C4.5 to derive classification rules for 275 pages which we manually assigned the correct labels. Obtained rules are shown in [6]. For example, the rule for Coauthor is simple: if two names co-occur in the same line, they are classified as coauthors. However, the Lab relationship is more complicated.

Table 1 shows error rate of five-fold cross validation for 275 training data. Though error rate for Lab is high, others have about or less than 10% error rate. After obtaining data, precision and recall is measured by manually labeling another 200 Web pages.<sup>6</sup> Coauthor class gives high precision and recall although the rule is very simple. However, Lab class gives low recall assumably because

<sup>6</sup> We do not use these 200 pages for making discriminant rules.

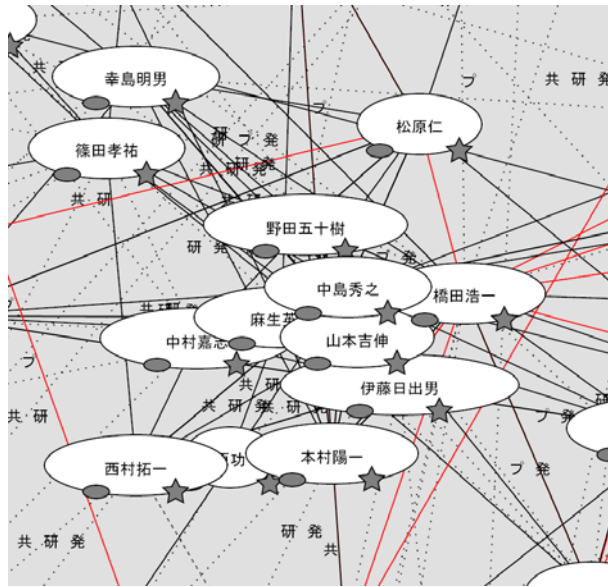


Fig. 1. Social network at JSAI2003 (zoomed)

laboratory pages has larger degree of variety. Proj and Conf classes gives higher error rate applied to another 200 data than 275 training data. It means that Web pages of these classes has a variety of styles, obtained rules are not enough.

We served our system at JSAI2003 and JSAI2004 on information kiosks and on the Web. Figure ?? shows a information kiosk where the extracted social network is displayed. Figure 1 is a part of the network with 266 nodes and 690 edges.

### 3 Analysis of Network Evolution

We have crawled the Japanese AI researcher network in 2003, 2004, and 2005. This section describes analysis of the network evolution.

Each year, about 500 persons attend the JSAI annual conference. We select 90 people who attended (and will attend) the conferences through 2003 to 2005 as a core group of the AI researchers. Table 2 shows the average Simpson value of whole participants and that of the 90 persons. Average co-occurrence numbers are also shown. We can see that the averages varies over time: Because the detail processing of a search engine varies over time, the obtained relation also varies over time. To eliminate the effect of the search engine factor, we normalize the values so that the total sum of the co-occurrences is a constant, and get the networks characterized as Table 3.

Figures 2, 3 and 4 show degree distribution on the obtained coauthorship network, the laboratory-relation network, and the Simpson coefficient network respectively. Coauthor and Laboratory network make smooth plots, though Project

**Table 2.** Average co-occurrence and Simpson value.

Year	Ave. Simpson of whole	Ave. Simpson of 90	Ave. cooc of whole	Ave. cooc of 90
2003	0.0665	0.0908	2.76	3.59
2004	0.00964	0.0471	1.17	2.31
2005	0.0347	0.164	1.73	4.10

**Table 3.** Number of edges and density for the 90 core members.

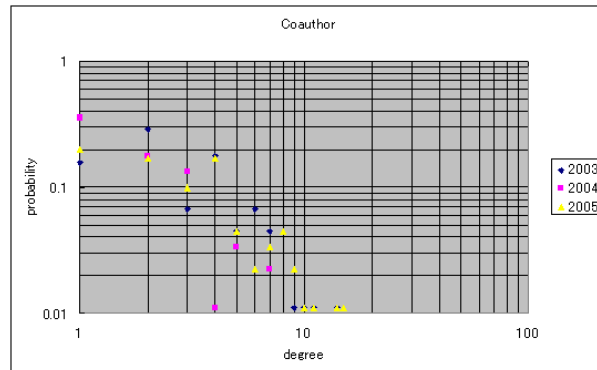
Year	Number of edges	Density
2003	189	0.0472
2004	419	0.105
2005	850	0.212

and Conference relation network do not. On the other hand, Simpson coefficient network brings almost linear plots on the log-log chart. The gamma of the power-law function is 1.50 in 2003, 1.01 in 2004 and 0.62 in 2005. That means some persons become to have more and more edges, while others reduce their degree.

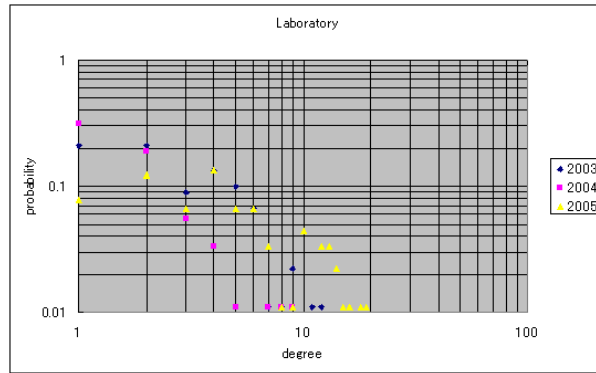
We conduct another kind of analysis: we measure the correlation among several centrality measures of three years. Interestingly, we discover strong correlation between the betweenness centrality of 2003 and the eigenvector centralities of 2004 and 2005. The betweenness centrality of 2004 also have high correlation with the eigenvector centrality of 2005. This suggests that a researcher who have diversified ties will get more central position later, which is consistent with the concept of structural holes [3].

## 4 Related work

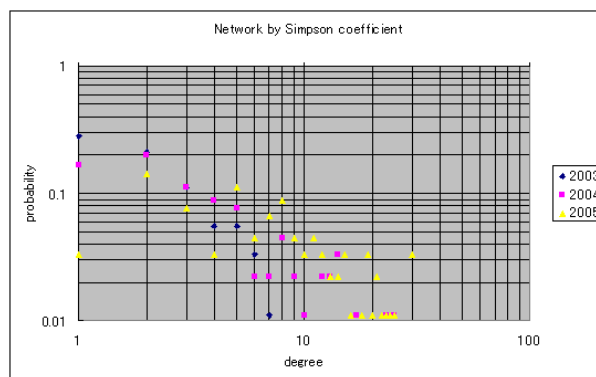
Kautz et. al develops a social network extraction system, called Referral Web, using co-occurrence of names on the Web [5]. Mika takes the similar approach



**Fig. 2.** Degree distribution of Coauthor network



**Fig. 3.** Degree distribution of Laboratory network



**Fig. 4.** Degree distribution of Simpson coefficient network

to ours [7] and try to extract Semantic Web community. Both studies employ Jaccard coefficient for co-occurrence index. Though the basic idea is similar to our approach, we further develop the mining algorithm: We apply text processing and machine learning to determine the class of relation. Furthermore, network analysis on the extracted network is also made [6]. It shows the applicability of calculating the trust of each person.

We develop a researcher mining and retrieval system, called Polyphonet<sup>7</sup>. The objective of the system is to provide search function based on the relation of researchers and promote efficient collaboration. For this purpose, we propose a series of mining approaches using a search engine: a keyword extraction algorithm from Web information to represent research relevant keyword such as research themes, project names, and organization names [9], and researcher clustering

<sup>7</sup> The system can be used at <http://www.carc.aist.go.jp/ponet/wai/>.

algorithm based on co-occurrence matrix of researcher's names and research topics [2].

## 5 Conclusion

This paper overviews an advanced Web mining approach to extract a social network of researchers using a search engine. Though we target researchers because of their abundant information on the Web, our approach is not limited to researchers. More and more information on ordinary people online makes our approach feasible in various domains.

Though we could not fully describe the analysis due to the space limitation, the evolution of the social network is suggestive enough and seems promising. Because much new information is available on the Web, monitoring and detection of a social network change may have potentials for some applications.

## References

1. Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.
2. Yohei Asada, Yutaka Matsuo, and Mitsuru Ishizuka. A method to automatically find foaf:group based on the cooccurrence of people with keywords in the web. In *Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, pages 34–37, 2004.
3. Ronald S. Burt. *The Social Capital of Structural Holes*, pages 149–92. Russel Sage Foundation, 2002.
4. Jeremy Goecks and Elizabeth D. Mynatt. Leveraging social networks for information sharing. In *Proc. 2004 ACM conference on CSCW*, pages 328–331, 2004.
5. H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magazine*, 18(2):27–35, 1997.
6. Yutaka Matsuo, Hironori Tomobe, Koiti Hasida, and Mitsuru Ishizuka. Finding social network for trust calculation. In *Proc. 16th European Conference on Artificial Intelligence (ECAI2004)*, pages 510–514, 2004.
7. Peter Mika. Bootstrapping the FOAF-Web: An experiment in social networking mining. In *Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, 2004.
8. Takeru Miki, Saeko Nomura, and Toru Ishida. Semantic web link analysis to discover social relationship in academic communities. In *Proc. SAINT 2005*, 2005.
9. Junichiro Mori, Yutaka Matsuo, Mitsuru Ishizuka, and Boi Faltings. Keyword extraction from the web for foaf metadata. In *Proc. 1st Workshop on Friend of a Friend, Social Networking and the Semantic Web*, pages 1–8, 2004.
10. Takuichi Nishimura, Yoshiyuki Nakamura, Hideo Itoh, and Hideyuki Nakamura. System design of event space information support utilizing CoBITs. In *Proc. IEEE ICDCS2004*, pages 384–387, 2004.