Mining Social Network of Conference Participants from the Web

Yutaka Matsuo
National Institute of Advanced
Industrial Science and Technology (AIST)
Aomi 2-41-6, Tokyo 135-0064, Japan
y.matsuo@carc.aist.go.jp

Hironori Tomobe
University of Tokyo
Hongo 7-3-1, Tokyo 113-8656, Japan
tomobe@miv.t.u-tokyo.ac.jp

Kôiti Hasida
AIST
Aomi 2-41-6, Tokyo 135-0064, Japan
hasida.k@carc.aist.go.jp

Mitsuru Ishizuka
University of Tokyo
Hongo 7-3-1, Tokyo 113-8656, Japan
ishizuka@miv.t.u-tokyo.ac.jp

Abstract

In a ubiquitous computing environment, it is desirable to provide a user with information depending on a user's situation, such as time, location, user behavior, and social context. At conventions, such as academic conferences and exhibitions, where participants must register in advance, the social context of participants can be extracted from the Web using their names and affiliations without asking the participants many questions. In this paper, we attempt to extract the social network of participants from the Web, where a node represents a participant and an edge represents the relationship of two participants. Each edge is added using the number of pages retrieved by a search engine which include both participants names. Moreover, each edge has a label such as "co-authors" and "members of the same project" by applying classification rules to the page content. We show an example of the extracted network and make a preliminery evaluation. This network can be used in many information services, such as finding an appropriate introducer or negotiater, and who one should talk to in order to efficiently expand his/her network.

1. Introduction

In a ubiquitous computing environment [8], much information regarding users' behavior can be obtained by a sensor network. We seek to provide users with personalized information depending on the situation: time, location, and user behavior. Especially at conventions such as academic conferences, the social context of each user is very important because the participants gather to experience new encounters and exchange knowledge face-to-face.

Assume a participant at a conference wants to make friends with researchers with similar interests near his current location. A future ubiquitous environment might detect the user's location and recommend that the user talk to a certain person. However, without background knowledge about the social network, the system may recommend the user's colleague or supervisor because they share the same interests. To make information services more "smart", such knowledge is indispensable.

By utilizing the knowledge about participants' social network, many potential applications can be considered. Assume a user wants to talk to a certain person and wants someone to introduce her. With the help of social network knowledge, the system can determine who is appropriate to introduce her. Conversely, one can find the path from herself to anyone with whom she might be talking. Another example might be efficient networking. A weak tie, which in social network theory is a connection between groups that don't ordinarily interact, plays an important role in getting valuable information [3]. The system can suggest who may be a candidate for this weak tie, that is, one who shares similar interests, but is in a different social group. Also, if one wishes, one could find who he should make a tie with in order to become more centered in the network [2].

At academic conferences such as WI2003, a participant must register a profile (at least name and affiliation) prior to the conference. In such cases, it is reasonable to assume that we have a list of participants and time to gather information about those participants from the Web. Referral Web [4] is a project to discover a social chain from an individual to the target person from the Web; however, in our case, fortunately we have a list of names in advance, and try to discover the entire network structure among participants from the Web. Digital services for social events are

Table 1. Attributes and possible values.

Attribute		Values
NumCo	The number of cooccurrences of X and Y	zero, one, or more than one
SameLine	Whether the names co-occur at least once in the same line	yes, or no
FreqX	Frequency of occurrence of X	zero, one, or more than two
FreqY	Frequency of occurrence of Y	zero, one, or more than two
GroTitle	Whether any of a word group (A-F) appears in the title	yes or no (for each group)
GroFFive	Whether any of a word group (A-F) appears in the first five line	yes or no (for each group)

not rare [7]. Many systems are developed for context-aware mobile services [1]. However, our approach is unique in that the system is conscious of the social network generated from information on the Web. Our system will serve at the 16th Annual Conference of the Japanese Society for Artificial Intelligence (JSAI 2003), installed as a location-based information support system [5].

The rest of this paper is organized as follows: In the next section, we describe how to extract a social network from the Web. In Section 3, we show an example and make evaluations for edge labels. After the discussion in Section 4, we conclude the paper.

Table 2. Word groups (translated from Japanese).

Group	Words
A	publication, papers, presentation, activities, themes, awards, authors, etc.
B	members, lab, group, laboratory, institute, team, etc.
C	project, committee
D	workshop, conference, seminer, meeting, sponsor, symposium, etc.
E	association, program, national, journal, session, etc.
F	professor, major, graduate student, lecturer, etc.

2 Social Network Extraction

2.1 Invention of Nodes and Edges

We assume that the names and affiliations of participants are given beforehand. [1] Therefore, nodes of the social network are invented first. Next, edges between nodes are added utilizing Web information. The most simple approach is to measure the relevance of two nodes based on the number of retrieved results by a search engine. For example, assume we are to measure the relevance of two names ‘Yutaka Matsuo” (denoted X) and “Hironori Tomobe” (denoted Y). We first put a query “X and Y” to a search engine and get a documents including those words in the text. (We only need the number of matched documents, not the whole contents of the matched documents.) Also, we put a query “X or Y”, and get b matched documents. The relevance of “Yutaka Matsuo” and “Hironori Tomobe” is approximated by the Jaccard coefficient $\#(X \cap Y)/\#(X \cup Y)$, say a divided by b.

If the Jaccard coefficient of a node pair is larger than the given threshold, an edge is added with its weight equal to the Jaccard coefficient. Some modifications are:

- There can be more than one person with the same family and given name. Adding affiliation to the query will alleviate this problem, but degrade the coverage. In order to keep the coverage as high as possible, we make a query “X and (A or B or $\ldots$)” instead of “X” where A and B are affiliations of X. For example, X is “Yutaka Matsuo,” and A is “National Institute of Advanced Industrial Science and Technology”, B is “AIST” (short for the institute), and C is “Cyber Assist Research Center” (a department of the institute).
- The Jaccard coefficient generally gives a famous person few number of edges because denominator b is very large compared to numerator a. Therefore we modify denominator b to $\min(\#X, \#Y)$. [2] We also add edges measured by the frequency of X and Y, i.e., $\#(X \cap Y)$, on top of the edges by the Jaccard coefficent.

However, even if few pages include both X and Y, if one of the pages is the laboratory member list, X and Y have a strong relation; i.e., they are the members of the same laboratory. Moreover, in order to discriminate several kinds of relationships, we need content analysis as described in the next section.

2.2 Extraction of Edge Label

It is more useful if each edge has a “label” for the relationship between two persons. For example, two nodes have the relation of “colleagues of the same research institute,” “professor – student,” “members of the same commit-

---

[1] Although some participants appear without prior registration, we currently ignore these cases.

[2] In some cases, it is more appropriate to employ a directed network representation instead of an undirected network, assigning conditional probability $\#(X \cap Y)/\#X$ to each directed edge.

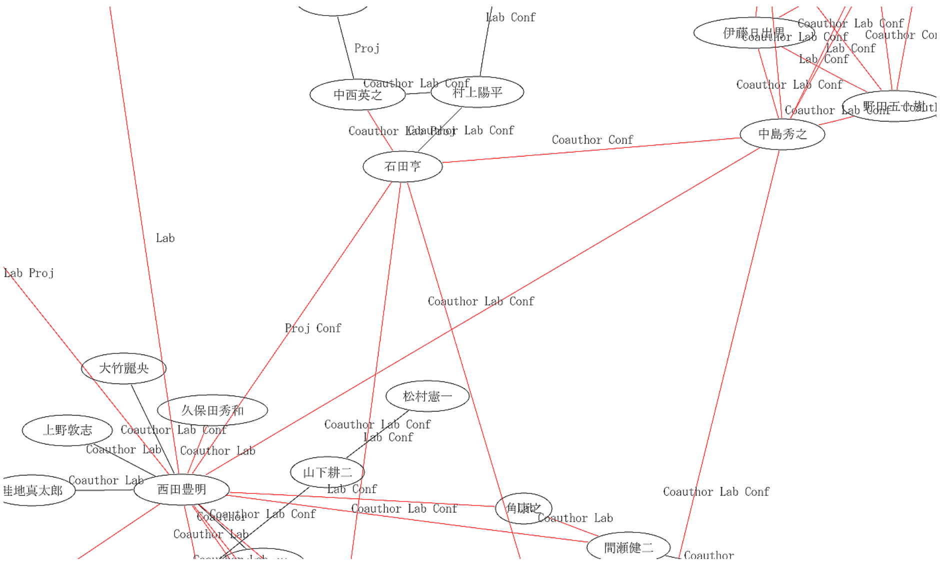

Figure 1. A part of social network of contributors in JSAI 2002.

tee," and so on. We discriminate the relationship by consulting retrieved page contents and applying classification rules. These rules are obtained by a machine learning approach.

We define labels (i.e., classes) for each edge as follows:

- Coauthor: Co-authors 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 has multi-labels. For example, X and Y have the relation of both "Coauthor," and "Lab."

We first fetch top three pages retrieved by a query "X and Y." Then we extract some features from the content of each page. We apply classification rules to the features and get labels of the relation between X and Y. Attributes and values for each page content are shown in Table 1. We currently use manually-selected word groups to characterize pages, shown in Table 2. [3]

Classification rules are obtained as follows: We first checked 300 pages manually and assigned labels to each page. These pages (feature values) and correct labels are used as training data. We employ C4.5 [6] to derive classification rules because of its ease of interpretability. Some of the obtained rules are shown in Table 3: For example, if two names co-occur in the same line, they are classified as coauthors. If the number of cooccurrences is more than one, and the title doesn't includes word group D but the first five lines include word groups A and E, then the relation is classified as members of the same laboratory.

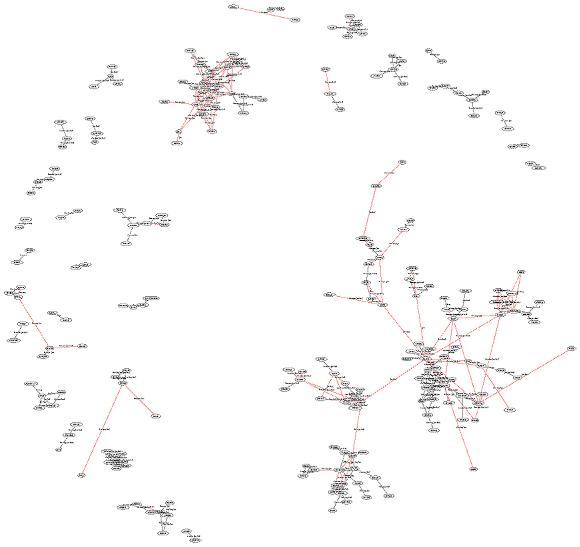

Figure 2. Social network of contributors in JSAI 2002.

3 Example and Preliminery Evaluation

Figure 1 is a part of the social network among contributors to JSAI2002. A node is labeled as the corresponding participant name (in Japanese), and an edge is labeled as "Coauthor", "Lab", "Proj", or "Conf". Around 470 people

---

[3] These word groups can also be automatically learned in the future.

Table 3. Obtained rules.

Class	Rule
Coauthor	SameLine=yes
Lab	(NumCo = more_than_one & GroTitle(D)=no & GroFFive(A) = yes & GroFFive(E) = yes)
	or (FreqX = more_than_two & FreqY = more_than_two & GroFFive(A) = yes & GroFFive(D)=no) or ...
Proj	(SameLine=no & GroTitle(A)=no & GroFFive(F)=yes) or ...
Conf	(GroTitle(A)=no & GroFFive(B)=no & GroFFive(D)= yes) or ...

Table 4. Evaluation of edge labels.

Label	Precision	Recall
Coauthor	25/26 (96.2%)	25/26 (96.2%)
Lab	20/29 (68.9%)	20/23 (87.0%)
Proj	2/2 (100%)	2/16 (12.5%)
Conf	24/24 (100%)	24/57 (42.1%)

contributed to the conference, therefore we have 470 nodes in the network. The entire network is shown in Figure 2.

Table 4 is the evaluation for 58 edges in the network, which are manually assigned correct labels. Precision is relatively high; however, recalls for Proj or Conf are low. This means that the extracted relations are reliable, but there might be overlooked relations. Furthermore, we should note that there might be relations that might not be able to be infered from the Web, e.g., coauthors of a forthcoming paper or members of the same laboratory ten years ago.

4 Discussion

If we look at Figure 2 in detail, we can see each participant's location in the network: clusters of participants and connecters of clusters. In the future, we may be able to show the distribution of research topics superimposed on the social network, how one can efficiently expand one's network, and the difference and characterization of conferences from the viewpoint of a social networks.

Various information services aware of the social network are possible, such as recommendation of a person who is distant from you in the network but has similar interests, and finding an appropriate introducer or negotiator. Furthermore, location-based and text-based applications will be enhanced with the help of social network knowledge, e.g., by refraining from suggesting a social network "neighbor" or associate because they already know each other well.

The objective of our research is to show the effectiveness and limits of extracting a social network in a closed community from the Web. However, there are some privacy issues related to extracting a social network. Generally, a participant doesn't know that her social context are extracted only by her name and affiliation. We should take care not to intrude on a user's privacy and to use the information only for useful services for a user.

Another potential problem is copyrights. Some Web pages prevent the utilization of their contents for revision or manipulation especially for commercial objectives. We should pay attention to this problem. (Fortunately, it seems that Web pages of universities, academic societies, and conferences are not as strict as corporate sites.)

5 Conclusion

Japanese people are relatively sensitive to their hierarchical social relations with others. Therefore, background knowledge of the social network is essential for providing more personalized information. However, asking questions about one's relationship with others is very intrusive. Our approach seems promising in that the social network can be obtained from the Web. We will attempt to further develop an information system conscious of an individual's social context.

References

[1] G. Chen and D. Kotz. A survey of context-aware mobile computing research. Technical Report TR2000-381, Dartmouth College, 2000.

[2] L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1:215–239, 1979.

[3] M. Granovetter. Strength of weak ties. *American Journal of Sociology*, 78:1360–1380, 1973.

[4] H. Kautz, B. Selman, and M. Shah. Referral web: Combining social networks and collaborative filtering. *Communications of the ACM*, 30, 1997.

[5] T. Nishimura, H. Itoh, Y. Yamamoto, and H. Nakashima. A compact battery-less information terminal (CoBIT) for location-based support systems. In *Proceedings SPIE*, number 4863B-12, 2002.

[6] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, California, 1993.

[7] Y. Sumi and K. Mase. Digital assistant for supporting conference participants: An attempt to combine mobile, ubiquitous and web computing. In *Proceedings UBICOMP 2001 (LNCS 2201)*, pages 156–175, 2001.

[8] M. Weiser. The computer for the twenty-first century. *Scientific American*, 268(3):94–104, 1991.