Finding Social Network for Trust Calculation

Yutaka Matsuo[1] and **Hironori Tomobe**[2] and **Kôiti Hasida**[3] and **Mitsuru Ishizuka**[4]

Abstract. Trust is a necessary concept to realize the Semantic Web. But how can we build a "Web of Trust"? We first argue that a small "Web of Trust" for each community is very essential to realize a huge "Web of Trust." Then, we focus on an academic community as a microcosm of a "Web of Trust" and show a web mining approach to generate a social network automatically. Each edge is added using the number of retrieved pages by a search engine which includes both persons' names. Moreover, each edge is given a label, such as "co-authors" or "members of the same project," by applying classification rules to the page content. The relation of persons such as "coauthor" or "same laboratory" can be described by an RDF format. Finally, using the social network, we calculate authoritativeness of a node as a social trust and an individual trust.

1 Introduction

The Semantic Web is what we will get if we perform the same globalization process to Knowledge Representation that the Web initially did to Hypertext [2]. The vast amount of Web information is now changing knowledge processing and influencing how knowledge is represented and utilized in our daily lives, that has been for a long time an issue of AI. The Semantic Web can enable more advanced knowledge processing and there are a lot of space that AI can contribute.

Tim Berners-Lee developed the famous "layer cake" to illustrate languages for the Semantic Web. Languages for metadata, ontologies, rules, proofs, and logics sit a top of one another, and make more of the information on the web machine-readable. On top of the layer cake, we have a trust layer. Anyone can say anything on the Web. Therefore without trust we can not deterimine whom to believe. Trust is a necessary factor to fully utilize a semantic web.

For realizing trust on the network, some studies specifically address authentication, access control and delegation by digital signature. Using a digital signature for RDF statements, we can verify that a certain person wrote them. However, even if we can be sure that he wrote them, how can we determine that he himself is reliable? For example, if one is going to make a travel schedule using information on the Web, how can one know that the information is written by a reliable person?

Therefore, it is important to argue whether the source of information is reliable and credible, aside from authentication techniques. Aaron Swartz and James Hendler describe "Web of Trust" as follows [14]:

> Now it's highly unlikely that you'll trust enough people to make use of most of the things on the Web. That's where the "Web of Trust" comes in. You tell your computer that you trust your best friend, Robert. Robert happens to be a rather popular guy on the Net, and trusts quite a number of people. And of course, all the people he trusts, trust another set of people. Each of these measures of trust is to a certain degree (Robert can trust Wendy a whole lot, but Sally only a little).

Based on the trust network, the computer can decide how trustworthy persons, resources, and pieces of information are.

The physical world already offers a "web of trust"; it is a kind of social network. I trust one of my friends, therefore I also trust a person introduced by that friend. I trust a company by the reason that one of my companies is dealing with that company. In this way, our social network works well to assess trustworthiness. Such a mechanism is likely to work well on the Semantic Web, too. Especially, the trustworthiness of persons is important because web resources are usually created by a person or group. Usually, if a person is reliable, what she writes is also reliable.

However, a person usually has many friends, partners, and acquaintances. According to social scientists, a person can name 200 to 5000 people without aid [1]. It is overwhelmingly demanding to write down all the relations that one has. To make matters worse, such relations are dynamic. New relations appear every day and old relations weaken gradually. The degree of the relation will change over time.

Two means can resolve this problem:

- Focus only on important relations: For example, to grant permission to access a very confidential file, we have only to show your computer a couple of your best friends. However, this network will be so sparse that it might not work well to judge the reliability of a person and a resource.
- Alleviate the cost to write down relations: If everyday software (e.g., mailers, browsers, schedulers, and groupware) are equipped with the detector of relation to others, we can automatically generate a list of persons we may trust. Alternatively, if we can extract a social network from the Web by a web mining approach, it can be used as a surrogate for a "Web of Trust."

This study employs the latter option, especially a web mining approach.

How does a "Web of Trust" appear and finally cover all the Web? Some may think as follows: At the beginning, a person or an organization will trust some acquaintances. A trust network appears locally and grows gradually by adding new nodes and edges. Two local networks sometimes connect; at some time point, a "Web of Trust" emerges as a huge connected network which covers almost the entire Web.

However, our view is different: At the beginning, a local community will develop a small "Web of Trust" within the community. Such a community includes an academic society of a certain field, an interest group of a certain topic, and a cooperation group of a certain industry. The small "Web of Trust" in a local community is helpful for judging the reliability of a person, an organization, or a piece

[1] National Institute of Advance Industrial Science and Technology (AIST), Japan email: y.matsuo@carc.aist.go.jp
[2] University of Nagoya, Japan email: tomobe@nagao.nuie.nagoya-u.ac.jp
[3] AIST, Japan email: hasida.k@aist.go.jp
[4] University of Tokyo, Japan ishizuka@miv.t.u-tokyo.ac.jp

of information. Some nodes have a high degree of trust edges; thus they are considered reliable. A newcomer can gain trust by somehow tying himself to a trusted nodes. The small "Web of Trust" has its *raison d'etre* within the community. Then, small "Webs of Trust" will appear one by one in different communities. These local "Webs of Trust" will be superposed one by one because a person or an organization belong to several communities at the same time. Finally, they will come to comprise huge "Web of Trust" over the entire Web, encompassing many local trust networks.

Because each local trust network is concerned with a certain topic or field, trustworthiness will not be transmitted straightforwardly among different networks. For example, trustworthiness in a community of cartoon films will not be used to measure the trustworthiness of a scientific statement. However, trustworthiness of trade in a certain industry can be transmitted into other industries. This view is consistent with our daily lives; a cartoon critic is reliable about cartoons, but not reliable about scientific research unless the critic also has that expertise.

This paper shows a new algorithm that automatically generates a social network in an academic society. It is one approach to establish the local trust network in a community. However, in different communities, there might be other appropriate ways to establish trust networks: sometimes, it is appropriate for users to explicitly write relations or to use groupware to elucidate those relations.

The remainder of this paper is organized as follows. In the next section, we describe our web mining approach to extract a social network from the Web. We show an example and evaluate the approach in Section 3. Section 4 describes trust calculation based on the social network. We conclude the paper after discussion in Section 6.

2 Social Network Extraction

There are many communities in the physical world and online: students at a university, workers at a corporation, members in an academic society, members in interest groups, and so on. This study specifically addresses an academic society: Japanese Society of Artificial Intelligence (JSAI). The reason why we choose JSAI is the availability of information about it on the Web. The information of an 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 our algorithm well.

2.1 Invention of Nodes and Edges

An academic society retains member profiles (name, affiliation, qualification, contact address, and so on), but never places such information before the public. However, we can obtain a list of contributors to conferences. In the case of JSAI, it has a regular annual conference. We first choose the contributors to the last four annual conferences (JSAI99, JSAI2000, JSAI2001, and JSAI2002) as active members of the JSAI community. Each active member of JSAI is considered to be a node in a social network. A node is labeled by the name of its corresponding person.

Next, edges between nodes are added utilizing Web information. The most simple approach is to measure 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 address a query "X and Y" to a search engine and get a documents including those words in the text. Also, we make a query "X or Y", and get b matched documents. The relevance of "Yutaka Matsuo" and "Hironori Tomobe" is approximated by the Jaccard coefficient

$$rel(x,y) = Jaccard(X,Y) = \frac{\#(X \cap Y)}{\#(X \cup Y)},$$

say a divided by b. The $rel(x,y)$ represents the relevance of node x and y.

If X and Y have a strong relation, the retrieved documents might include X's and Y's homepages, their publication pages, a laboratory's member list page, a conference program page, an expert committee's page, and so on. Therefore, $\#(X \cap Y)$ becomes large compared to $\#(X \cup Y)$. If relevance of a node pair is larger than the given threshold, an edge is added with its weight equal to the relevance.

Some modifications are necessary:

- 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. To keep the coverage as high as possible, we make a query "X and (A or B or ...)" 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" (the abbreviated name of the institute), and C is "Cyber Assist Research Center" (a department of the institute).
- The Jaccard coefficient generally gives a famous person few edges because denominator b is very large in comparison to the numerator a. We can modify denominator b to $\min(\#X, \#Y)$, which places too much weight on a person with few edges. Therefore, we finally employ this formula[5]:

$$rel(x,y) = \begin{cases} \frac{\#(A \wedge B)}{\min(\#(A),\#(B))} & if \#(A) > k \; and \; \#(B) > k, \\ 0 & otherwise. \end{cases}$$

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

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

We define labels (i.e., classes) for each edge 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 has multi-labels. For example, X and Y have the relations of both "Coauthor," and "Lab."

We first fetch the top three pages retrieved by the 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. Table 1 shows attributes and values for each page content. We currently use manually-selected word groups to characterize pages, as shown in Table 2. [6]

Classification rules are obtained as follows: We first checked 275 pages manually and assigned labels to each page. These pages (feature values) and correct labels are used as training data. We employ

[5] This formula can be considered as threshold-based Simpson coefficient.
[6] These word groups can also be learned automatically in the future.

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 cooccur 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 lines	yes or no (for each group)

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 2. Word groups (translated from Japanese).

Group	Words
A	publication, paper, presentation, activity, theme, award, authors etc.
B	member, lab, group, laboratory, institute, team, etc.
C	project, committee
D	workshop, conference, seminar, meeting, sponsor, symposium, etc.
E	association, program, national, journal, session, etc.
F	professor, major, graduate student, lecturer, etc.

C4.5 [13] to derive classification rules because of their ease of interpretability. Some of the obtained rules are shown in Table 3: For example, if two names cooccur in the same line, they are classified as coauthors. If the number of cooccurrences is more than one, and the title does not include word group D, but the first five line includes word groups A and E, then the relation is classified as "members of the same laboratory."

2.3 Example and Evaluation

Figure 1 is a part of the social network of JSAI community. A node is labeled as the corresponding participant name (in Japanese), and an edge is labeled as "Coauthor", "Lab", "Proj", or "Conf". The whole network is shown in Fig. 2 (in SVG format). We have more than 1500 people in the community from which we choose about 150 members to illustrate this network.

Table 4 is the error rate of edge label classification based on five-fold cross validation of 275 training data. The error rate of "Lab" label is high, but the performance is generally good. However, we should note that there might be relations that are almost impossible to infer from the Web, e.g., coauthors of a forthcoming paper or members of the same laboratory ten years ago. More detailed evaluation is in progress.

Table 5 is evaluation by questionnaires. We conducted a Web-based questionnaire to 141 persons who registered to JSAI2003 Web system. 82 persons (58%) answered the questions. We asked several questions on the relation to 20 persons for each person, such as "Does or did this person belong to the same laboratory?" The answers are considered as correct samples and compared with obtained labels from the Web. Though recall is low, precision is around 80%. Low recall is due to several reasons: we do not use all the retrieved pages for mining the relations, and there is not all the information on the Web.

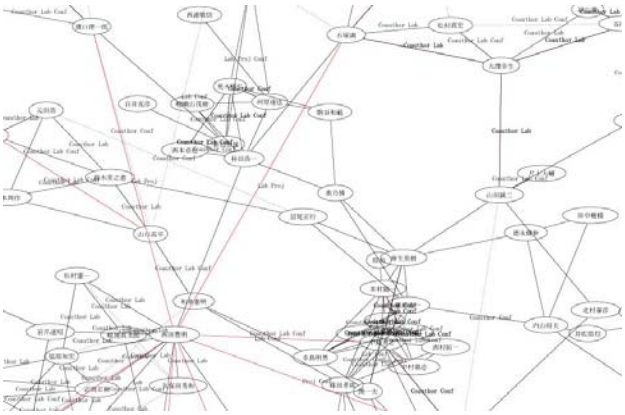

Figure 1. A part of the JSAI social network.

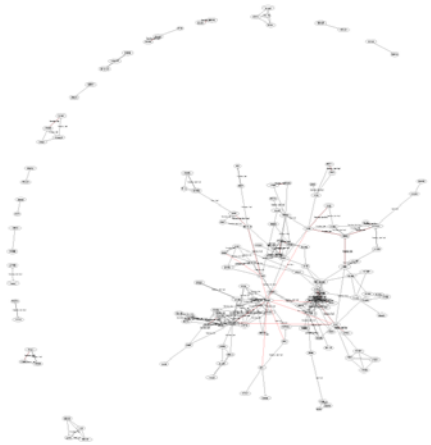

Figure 2. Social network of JSAI.

Table 6. Result of Authority Propagation

	Name	Activation	Freq	Comment
1	西田豊明 (Toyoaki Nishida)	5.53	624	Former commissioner of JSAI, Prof.
2	石田亨 (Toru Ishida)	4.98	574	Former commissioner of JSAI, Prof.
3	中島秀之 (Hideyuki Nakashima)	4.52	278	Former Commissioner of JSAI
4	橋田浩一 (Koiti Hasida)	4.499	345	Commissioner of JSAI
5	石塚満 (Mitsuru Ishizuka)	4.24	377	Commissioner of JSAI, Prof.
6	奥乃博 (Hiroshi Okuno)	3.89	242	Commissioner of JSAI, Prof.
7	溝口理一郎 (Riichiro Mizoguchi)	3.60	404	Commissioner of JSAI, Prof.
8	山田誠二 (Seiji Yamada)	3.35	168	Associate Prof.
9	武田英明 (Hideaki Takeda)	3.22	435	Associate Prof.
10	山口高平 (Takahira Yamaguchi)	3.10	236	Prof.
11	大澤幸生 (Yukio Ohsawa)	2.98	185	Associate Prof.
12	田中穂積 (Hozumi Tanaka)	2.90	465	Chairperson of JSAI, Prof.
13	徳永健伸 (Takenobu Tokunaga)	2.89	302	Associate Prof.
14	古川康一 (Koichi Furukawa)	2.77	141	Former Commissioner of JSAI, Prof.
15	河原達也 (Tatsuya Kawahara)	2.74	440	Prof.

Table 7. Result of Authority Propagation from Yutaka Matsuo

	Name	Activation	Freq	Comment
1	松尾豊 (Yutaka Matsuo)	230.6	136	myself
2	石塚満 (Mitsuru Ishizuka)	28.7	377	my former supervisor, co-author
3	大澤幸生 (Yukio Ohsawa)	19.5	185	my former project leader, co-author
4	西田豊明 (Toyoaki Nishida)	14.5	624	professor of lecture at university
5	松村真宏 (Naohiro Matumura)	13.5	82	my former colleague, co-author
6	山田誠二 (Seiji Yamada)	12.7	168	acquaintance
7	高間康史 (Takafumi Takama)	12.3	16	former researcher of my former laboratory
8	石田亨 (Toru Ishida)	12.1	574	advisary board of current research center
9	山口高平 (Takahira Yamaguchi)	11.5	236	acquaintance
10	田中英彦 (Hidehiko Tanaka)	11.3	842	professor at university

Table 4. Error rate of edge labels by five-fold cross validation.

Label	Coauthor	Lab	Proj	Conf
Error rate	4.1%	25.7%	5.8%	11.2%

Table 5. Evaluation of edge labels by inquiry for JSAI2003 participants.

Label	Precision	Recall
Coauthor	89.0% (81/91)	32.1% (81/252)
Lab	78.3% (72/92)	18.7% (72/385)
Proj	50.0% (9/18)	3.0% (9/300)
Conf	79.5% (35/44)	6.5% (35/538)

```
<rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/
22-rdf-syntax-ns#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/"
 xmlns:foaf="http://www.carc.aist.go.jp/
~y.matsuo/acsn/0.1/">

<foaf:Person>
  <foaf:name>Yutaka Matsuo</foaf:name>
  <foaf:mbox>y.matsuo@aist.go.jp</foaf:name>
  <foaf:workplacehomepage
     rdf:resource="http://www.carc.aist.go.jp/">
  <acsn:Coauthor>
    <foaf:Person>
       <foaf:name>Mitsuru Ishizuka</foaf:name>
    </foaf:Person>
  </acsn:Coauthor>
</foaf:Person>
```

Figure 3. Sample code of FOAF made from a mined relation from the Web.

3 Trust Calculation

Using the social network, we can obtain the “authoritativeness” of a node. It can be considered as reliability, or in other words, social trust. On the other hand, the network is used to calculate trust that can be accorded to that person; in other words, individual trust. This section mentions these two kinds of trust.

3.1 Social Trust

So far, numerous studies have been addressed for network analysis. The small world phenomenon was inaugurated as an area of experimental study in the social sciences by Stanley Milgram in the 1960s [11, 15]. The importance of weak ties, which is a short-cut between clusters of people, was mentioned 30 years ago [7].

Freeman proposes a number of ways to measure “centrality” of a node [5]. Considering an actors’ social network, the simplest is to count the number of others with whom an actor maintains relations. The actor with the most connections, i.e., the highest *degree*, is most central. Another measure is *closeness*, which calculates the distance from each person in the network to each other person based on the connections among all members of the network. Central actors are closer to all others than are other actors. A third measure is *betweenness*, which examines the extent to which an actor is situated between others in the network, i.e., the extent to which information must pass through them to get to others, and thus the extent to which they will

be exposed to information circulating in the network. Matsuo et al. develops a measure called *contribution*, which calculates the difference of closeness of all nodes with and without a certain node. It measures a node's contribution to the whole (small world) structure by temporarily eliminating the node [10].

The Google[7] search engine uses the link structure for ranking Web pages, called PageRank [3]. A page has a high rank if the sum of the ranks of its backlinks is high. The rank of a page is divided among its forward links evenly to contribute to the ranks of the pages they point to. PageRank is a global ranking of all Web pages and is known to perform very well.

We employ here a PageRank-like model to measure authoritativeness of each member. Each node v has an authority value $A_n(v)$ on iteration n. The authority value propagates to neighboring nodes in proportion to the relevance to the node:

$$A_{n+1}(v') = c \sum_{v' \in Neighbor(v)} \frac{rel(v, v')}{rel\_sum(v')} A_n(v') + cE(v) \quad (1)$$

and

$$rel\_sum(v) = \sum_{v'' \in Neighbor(v)} rel(v, v''),$$

where $Neighbor(v)$ represents a set of nodes each of which is connected to node v, c is a constant, and E represents a source of authority value. We set E as uniform over all nodes. For mathematical details, see [3].

Table 6 shows a result applied to the JSAI community extracted from the Web. Among 1509 people in the community, these people have high authority value $A(v)$ after 1000 iterations. Present or former commissioners of JSAI comprise 9 of 15 people. Others are younger; they are not yet commissioners, but very active researchers who are mainly working in JSAI.

The top listed people by this algorithm are authoritative and reliable in the JSAI community. However, authoritative people are not always listed highly by our approach. For example, currently JSAI has 20 commissioners (including a chairperson and two vice chairpersons), but we can extract only 5 current commissioners of the top 15. In other words, our approach seems to have high precision, but low recall. This is attributable to the lack of information online. Especially, elder authorities tend to have made many publications before the WWW came to daily use.

We don not show a comparison to other measures because of space limitation, but this PageRank-like algorithm performs better than other measures including degree, closeness, and the number of retrieved pages. For example, by measuring the number of retrieved pages, a famous person tends to be ranked highly regardless of the contribution to the community. This measure is insufficient for our purposes because we want to know the authoritativeness in the target community. Alternatively, we can measure the topic-sensitive PageRank [8] of one's homepage as that person's authoritativeness. However, the connection between authority of a person and authority of a homepage is not clear; some pages have high PageRanks because their contents are popular, not because they are written by authorities.

3.2 Individual Trust

If we set a certain node v_{target} as a source of authority value, the result can be interpreted as showing authority for the node, in other words, individual trust. We set the initial authority as follows.

$$E(v) = \begin{cases} 1.0 & if\ v = v_{target}, \\ 0.0 & otherwise \end{cases}$$

[7] `http://google.com`

Then, we propagate the authority following Eq. (1) for 300 iterations. Table 7 is the result obtained by setting v_{target} as node "Yutaka Matsuo."

The familiar persons for the first author, e.g., a supervisor, a project leader, colleages, and co-authors are ranked as high. This can be used to approximate individual trust. For example, if a person is judged as very familiar to me, then she can automatically have permission to access my work libraries. Otherwise, she must ask my permission.

4 Related Works and Conclusion

Referral Web [9] is a project to discover a social chain from an individual to the target person from the Web; however, in our case, we first extract a list of members in the community, and try to determine their social network. Murata attempts to discover the relation between Web pages using the number of retrieved documents [12]. We also use the number of retrieved documents, but we also use the contents of the retrieved documents to classify the relation into four categories. Dan Brickley and Libby Miller invented an RDF vocabulary called FOAF (Friend-of-a-Friend) to create a social network. A user creates one or more FOAF files on a Web server and shares the URLs so software can use the file information [4, 16]. But there work is on how to represent human relation and not on how to automatically obtain the relation.

In this paper, we argue how local trust networks will finally constitute a huge "Web of Trust." We focus on the academic community and show an algorithm to mine the social network using a search engine and machine learning. Furthermore, the relation is utilized to measure the authoritativeness of a member as social trust or individual trust. There are many research issues that should be investigated to realize a "Web of Trust" on the Semantic Web.

REFERENCES

[1] Albert-Lászió Barabási, *LINKED: The New Science of Networks*, Perseus Publishing, Cambride, MA, 2002.

[2] Tim Berners-Lee. http://www.w3.org/DesignIssues/RDFnot.html, 1998.

[3] S. Brin and L. Page, 'The anatomy of a large-scale hypertextual web search engine', in *Proc. 7th WWW Conf.*, (1998).

[4] FOAF: the 'friend of a friend' vocabulary. http://xmlns.com/foaf/0.1/.

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

[6] Jennifer Golbeck, James Hendler, and Bijan Parsia, 'Trust networks on the semantic web', in *Proc. WWW 2003*, (2003).

[7] M. Granovetter, 'Strength of weak ties', *American Journal of Sociology*, **78**, 1360–1380, (1973).

[8] Taher Haveliwala, 'Topic-sensitive PageRank', in *Proc. WWW2002*, (2002).

[9] H. Kautz, B. Selman, and M. Shah, 'Referral web: Combining social networks and collaborative filtering', *Communications of the ACM*, **40**(3), 63–65, (1997).

[10] Yutaka Matsuo, Yukio Ohsawa, and Mitsuru Ishizuka, 'Document as a small world', in *New Frontiers in Artificial Intelligence, Joint JSAI 2001 Workshop Post-Proceedings, LNAI 2253*, pp. 444–448, (2001).

[11] S. Milgram, 'The small-world problem', *Psychology Today*, **2**, 60–67, (1967).

[12] Tsuyoshi Murata, 'Finding related web pages based on connectivity information from a search engine', in *Proc. 10th WWW Conf.*, (2001).

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

[14] Aaron Swartz and James Hendler, 'The Semantic Web: A network of content for the digital city', in *Proc. Second Annual Digital Cities Workshop*, (2001).

[15] D. Watts and S. Strogatz, 'Collective dynamics of small-world networks', *Nature*, **393**, 440–442, (1998).

[16] XML watch: Finding friends with xml and rdf. http://www-106.ibm.com/developerworks/xml/library/x-foaf.html, 2002.