Extracting Inter-Firm Networks from World Wide Web
Using General-Purpose Search Engine

Yingzi Jin¹, Yutaka Matsuo², and Mitsuru Ishizuka¹
¹University of Tokyo, Hongo 7-3-1, Tokyo 113-8656, Japan
eiko-kin@mi.ci.i.u-tokyo.ac.jp, ishizuka@i.u-tokyo.ac.jp
²National Institute of Advanced Industrial Science and Technology y.matsuo@aist.go.jp

Abstract
Social relations play an important role in a real community. Interaction patterns reveal
relations among actors (such as persons, groups, firms), which can be merged to produce
valuable information as a network structure. This paper presents a new approach to extract
inter-firm networks from the web for further analysis. Extraction of relations between a pair of
firms is realized using a search engine and text processing. Because names of firms co-appear
coincidentally on the web, we propose an advanced algorithm, which is characterized by
addition of keywords (relation keywords) to a query. The relation keywords are obtained from
the web using a Jaccard coefficient. As an example, we extract a network of 60 firms in Japan
including IT, communication, broadcasting, and electronics firms from the Web and show
comprehensive evaluations of our approach. Our method is a first attempt to extract inter-firm
networks from the web using a search engine. The approach is also applicable to other actors,
such as famous persons, organizations or other multiple relational entities.

Keywords
World wide web, Social network, Information extraction, Search query, Relation extraction.

Article Type
Research paper

INTRODUCTION
Various relationships exist among firms such as mergers, acquisitions and partnerships.
Together, these relationships define a network between firms. Such networks are useful in
analyzing firms' competitiveness and helps in determining its marketing strategy. Furthermore,
overall network features can assist us in analyzing the stability and growth of the industry.
Numerous studies of social network analyses have been conducted in the fields of economics
and other social sciences (Bengtsson and Kock, 1999; Papakryiazis and Boudourides, 2001;
Battiston, 2004; Yeung, 2005).

Many studies have investigated methods to extract social networks from the web while
targeting people (particularly researchers or students). For example, using social networking
services (SNS), aggregating Friend-of-a-Friend (FOAF) documents (Finin et al., 2005; Miki and
Nomura and Ishida, 2005). Particularly, several studies have been undertaken to use a search
engine to extract social networks (Kautz and Selman and Shah, 1997; Mika, 2005; Matsuo et
al., 2006a; 2006b). Co-occurrence of names on the web is commonly used as proof of
relational strength. However, the co-occurrence methods can not apply directly for some
entities such as famous people, organization or firms, which have multiple relations, and
relational information on the web affected by media effect. Many economic analyses of inter-
firm networks have been obtained relational data only from the stock market or shareholding
information in business magazines that are much less diverse (Battiston, 2004; Souma and
Fujiwara and Aoyama, 2005).

Many relations among firms are published on the web in news articles or news releases
(Figure 1). Our work emphasizes the investigation of such published relations on the web to
address the relation extraction problem. Given a list of firms V={v₁, v₂, ...}, our goal is to
retrieve and extract relations among them to construct inter-firm networks G=(V, E), in which
each edge $e=(v_1, v_2) \in E$ represents a relationship between $v_1$ and $v_2$. We specifically seek to develop methods that acquire relationships from the web, the largest available resource that deals with all firms. For each pair of firms $(v_1, v_2)$, our system address two problems: (a) collecting information about target relations, such as "Company v1 merged with Company v2"; and (b) relation extraction, such as extract capital alliance (merge) from above sentence. For collecting information from entire web, we use a general-purpose search engine. Query expansion and modification techniques are applicable in this case (Glover et al., 2001; Oyama and Kokubo and Ishida, 2004). Research on relation extraction has been promoted by Message Understanding Conferences (MUCs) and Automatic Content Extraction (ACE) programs. Numerous techniques to address this task have been proposed in the literature, such as pattern matching (Brin and Page, 1998), kernel methods (Zhao and Grishman, 2005), and logistic regression (Kambhatla, 2004). For the firm case, our extraction task is to detect relations among same types (i.e., COM type) of entities.

In this study, we use a search engine to collect target relational pages from the web. Since names of firms co-appear coincidentally on the web, we propose to add additional words (call relation keyword) to name pairs of firms as a query. We then apply a simple pattern-based approach to extract the relations. We extract alliance relations as a positive relation and lawsuit relations as a negative relation. Much of this daily information is obtainable from the web. Examination of daily changing and complex social relationships is important for analyzing social trends, understanding social structures, and for formulating new industrial activities. Our method is a first attempt to extract inter-firm networks from the web using a search engine. Our approach is applicable to other entities, such as famous persons, or other multiple relational entities.

The remainder of this paper is organized as follows. We show related work in next section, and then we describe our proposed idea and methods. Then we also show experiments and evaluations before we conclude this paper.

**RELATED WORKS**

Many studies have used search engines to extract social networks automatically from the Web (Kautz and Selma and Shah, 1997; Mika, 2005; Matsuo et al., 2006a; 2006b). Co-occurrence of names on the Web is commonly used as evidence of relational strength (Kautz and Selma and Shah, 1997). Related to the Semantic Web community, P. Mika developed a system called Flink, which extracts relational information from web pages, e-mail messages, publications, and self-created FOAF profiles (Mika, 2005). The web mining component of the system uses a search engine to measure the strength of relations among researchers. Comparably, Y. Matsuo and his colleagues developed a system called POLYPHÔNET, mainly for use by Japan's AI community (Matsuo et al., 2006b). However, the co-occurrence-based methods become ineffective when two target entities co-occur universally on many Web pages. We take two persons to explore this problem: Bill Gates and George Bush. Those two names “coincidentally” co-occur on the Web very often: They may be on the same news pages just because they made some statements on the same day. They may be on the pages that list “the most famous persons in the world”. For that reason, it is not a good idea to measure the strength of relations simply through the use of co-occurrence measures. This problem is
commonly confronted for firms: a firm name is similar to a famous person's name, and they often co-occur for various reasons, even though no formal relations exist among them. When the relation between firms attracts attention by media services (such as a lawsuit relation), many pages describe and comment on it; in contrast, only a few pages exist on the web if the relation gets no attention. Considering that media effects influence the number of web pages that appear, co-occurrence of names on the web is not always useful to represent the actual relations linking two firms.

Web search by query modification and expansion is described in (Glover et al., 2001); they extracted query modification rules for finding personal homepages and calls for papers. For information retrieval and query expansion, S. Oyama's work is more closely related to ours (Oyama and Kokubo and Ishida, 2004). They added keywords called "keyword spices" to the user's input query with a Boolean expression for a domain-specific web search. They sampled web pages using initial keywords and classified them manually as either relevant or irrelevant, thereby producing a training corpus. Subsequently, they applied a decision-tree learning algorithm to discover keyword spices. Our system sets relation keywords that are added by query as combinations of one or two terms. Therefore, a Jaccard coefficient is used simply to measure the scores. Other studies such as Flink use a phrase "Semantic Web OR Ontology", POLYPHONET adds affiliation information together with a name for disambiguation. To extract characteristic key phrases for a person automatically, D. Bollegara clusters web pages that are related by each name into several groups using text similarity (Bollegara and Matsuo and Ishizuka, 2006).

Battiston et al. extract shareholding relationships from stock market information (MIB, NYSE and NASDAQ) to analyze characteristics of market structure (Battiston, 2004). Souma et al. extract data published by Tokyo Keizai Inc. to construct Japanese shareholding networks to analyze features of Japanese companies' growth (Souma and Fujiwara and Aoyama, 2005). Our work specifically addresses alliance and lawsuit relations among firms from published resources on the web. Consequently, we can obtain relations easily and can track down daily changing and social trends. Dealing with time series changes of relations is one of our interests for future work.

Name disambiguation is an important problem for social network mining. To date, several studies have produced attempts at personal name disambiguation on the Web (Bekkerman and McCallum, 2005; Bollegara and Matsuo and Ishizuka, 2006; Li and Morie and Roth, 2005). However, ambiguity in firm (or organization) names is lesser compared to personal names. We intend to explore ambiguities in company names in our future work.

**SOCIAL NETWORK EXTRACTION FOR FIRMS**

**Basic Concept**

In social sciences, the definition of a weak or strong tie might vary among contexts (Marsden, 1984). For example, the frequency or degree of relations affects that strength; multiple relations between two actors also can imply a stronger tie. In the firm case, the types of relations define the strength: For example, a capital alliance relation is stronger than a business alliance relation. Consequently, to present a tie among firms, it is appropriate that we identify the concrete relations of firms.

For using a search engine to retrieve and extract relations, a proper query is necessary. The intuitive query is the names of the two firms. For example, we issue a query such as "Matsushita AND JustSystem"1 to discover data containing their relationships. Thereby, we obtain as many as 425,000 pages. Many top-ranked pages are lawsuit-relation pages2 which drew much attention during the last year. Therefore, analyzing these pages, we were able to identify lawsuit relations among them. However, two companies showed a collaboration relation in knowledge management in 2001, which pages are in lower ranks of 124th, on

---

1 Both are names of famous Japanese corporations.
2 http://pc.watch.impress.co.jp/docs/2005/0201/just2.htm
account of a relation related to a firm several years ago, it might be lost. Of course, we can
download and analyze all the returned pages from a search engine to find all possible relations,
but that is both time consuming and costly.

As a solution, we can add some word or combination of words (called relation keyword) to
a search query and apply text processing to confirm the existence of fact. Using this strategy,
we can efficiently identify relations among firms. For example, when we wish to extract lawsuit
relations, we add a term “lawsuit”. We issue a query “Matsushita AND JustSystem AND
lawsuit” so that the search engine will return the lawsuit pages that are associated with the
two firms. Then we can conduct text processing to these pages to validate the relation’s
existence. This idea is similar to keyword spices (Oyama and Kokubo and Ishida, 2004), which
extend queries for domain-specific web searches. Question-answering systems also construct
elaborate queries for using a search engine (Ramakrishnan et al., 2004). Requirements of
relation keywords are identifying the relations more precisely and reducing the leakage of
relation pages if they exist. Therefore, both precision and recall are important for relation
keywords.

Our system has two major procedures: an online procedure and an offline procedure. In
the offline, relation keywords for each relation types are obtained beforehand using our
proposed method. In the online, a list of firms and specific relation types are given as an input
and the output is a social network of firms. In the following, we will first consider relation types
described in our study; then we propose relation keyword extraction. Finally, we will describe
online processes of our system. The entire system is depicted in Figure 2.

Relation type

Relationships among firms are various. For example, capital combinations such as mergers,
acquisitions, joint ventures, and business partnerships, such as business alliances, co-
development, service provision, and dispatching personnel, competition, lawsuit, etc. It is
considerable that pairs of firms have multiple relations. For example, two firms have alliance
and lawsuit relations. Each relation is typed in a more detailed manner. Alliance relations
between firms include capital alliances and business alliances, where the former usually
represents a stronger relation than the latter. A lawsuit relation has multiple stages; at some
time, the dispute will be settled by mutual accommodation or by final judgment. Therefore, the
relation can be typed as either on in a claim phase or in an accommodation phase. For
dynamic and complex relational networks, it is important to distinguish such typical and
temporal relations for detailed analyses of social networks (Marsden, 1984; Scott, 2000).
In this study, we address an alliance and lawsuit relation, which respectively represent positive and negative relations. The alliance relation is distinguished by business alliances and capital alliances; in addition, the lawsuit relation is separated into a claim phase and an accommodation phase. In the following, we designate these separate relations as detail relations.

**Relation Keyword Extraction**

In this section, we describe relation keyword extraction methods which are useful to collect relation pages from the web, and which are useful for the relation extraction procedure. Good relation keywords should satisfy a proper balance between specificity and generality.

The intuitive method for finding relation keywords is to select terms that appear often in the target pages (where the target relation is described) and which do not appear in other pages. Therefore, we need to collect annotated web pages of specific relations of the firms as a training corpus. Then we estimate the classification features of each word and word combination. We simply measure the $F$-value for each word (or word combination) $w$ to see how the training documents can be classified correctly. However, collecting and annotating the training corpus requires many hours of tedious work.

In our study, we propose to use a search engine to extract relation keywords. This method is identical to that of Mori’s work (Mori et al., 2005), in which a specific word $w_c$ is assigned, which can represent the relation most precisely. In our work, we regarded $w_c$ as seeds of relation keywords. If we want to retrieve an alliance relation, we add $w_c$ such as “alliance” to a search query; words that co-occur frequently with it also become good clues to discern the relation. We use the Jaccard coefficient, to measure the relevance of word $w$ to word $w_c$.

$$J_{wc}(w) = \frac{|w_c \cap w|}{|w_c \cup w|} \quad (1)$$

Where, $|w_c \cap w|$ represents the number of hits yielded by the query $w_c$ AND $w$ and $|w_c \cup w|$ represents the number of hits by the query $w_c$ OR $w$. Words $w$ with large Jaccard coefficients are also used as relation keywords aside from $w_c$. It would save costs of annotating training data with relevance or non-relevance manually. For choosing candidate words, it is necessary to prepare some target pages. However, they are easily obtainable from several news articles such as Yahoo! News, about target relations.

As preprocessing, we first eliminate all html tags and scripts from these web pages. Then we extract the body text of pages and apply a part-of-speech tagger Chasen\(^3\) to choose nouns and verbs (except stop words). Then we select the top N words with highest $tf*idf$ score\(^4\). These words are candidates of relation keywords. We also use two-word combinations as candidates. We measure the score of each candidate word / phrase by calculating the Jaccard coefficient with a seed of relation keywords $w_c$ (We determine these words through preliminary experiments. For instance, we used alliance AND corporate as $w_c$ for alliance relations.). Candidates with the highest scores are recognized as relation keywords. Choosing the relation keywords can be treated as feature selection for classifying relation pages, but a combination of complex queries does not work well for a search engine. Therefore,

\(^3\)http://chasen.naist.jp/hiki/Chasen/
\(^4\)Here, $tf*idf = tf(w) \times log(N/|w|)$, where $tf(w)$ is the number of occurrences in news articles containing $w$. In addition, $N$ is the total number of Web documents, and $|w|$ is the number of web pages containing $w$. 

---

**Figure 3** A procedures to extract relations using text processing.

```bash
function RELATIONEXTRACTION(D, x, y, W)
    scoreₓᵧ = 0
    S ← GetSentence(D, x, y)
    for each s ∈ S do
        if s contain “x” and s contains “y” then
            scoreₓᵧ ← |wₓ∩wᵧ| / (|wₓ| + |wᵧ|)
        if scoreₓᵧ > scoreₓᵧ then
            scoreₓᵧ ← scoreₓᵧ
        done
    done
```

---

3 http://chasen.naist.jp/hiki/Chasen/
4 Here, $tf*idf = tf(w) \times log(N/|w|)$, where $tf(w)$ is the number of occurrences in news articles containing $w$. In addition, $N$ is the total number of Web documents, and $|w|$ is the number of web pages containing $w$. 

we simply consider words or combinations of words as relation keyword candidates. It is explicit that the weight of \( w \) varies according to the relation type \( r \). Once we find the relation keyword, we can extract the relations among many firms. For detailed relations, it is necessary to prepare relation keywords for each detailed relation, but extraction methods for relation keywords are similar.

**Relation Extraction**

Online, a list of firms and specific relation types are given as an input and the output is a social network of firms. Three steps exist: making queries, Google search, and network construction. First, we make queries by adding relation keywords to each pair of firms. Then, we put these queries into the Google search engine to collect top- \( n_p \) web pages. For this experiment, we set \( n_p=2 \) and \( n_p=5 \). Finally, for each downloaded document \( D \), we conduct text processing to judge whether or not the relation actually exists. A simple pattern-based heuristic (as described in Figure 3) has been useful in our experience. We first select all sentences \( S \) that include the two firms' names (\( x \) and \( y \)) and assign each sentence the sum of relation keyword scores \( t_w \) in the sentence. The score of firms \( x \) and \( y \) is the maximum of the sentence scores. An edge is invented between the two firms if score\(_{xy} \) is greater than a certain threshold, i.e., if the two firms seem to have the target relation with high reliability.

**Experiments**

A network of 60 firms in Japan including IT, communication, broadcasting, and electronics firms, is extracted. For the dataset, we manually created a dataset for these 60 firms. The annotators decided the relations among the firms based only using the information available on the web. In our experiments, we will first show the extracted relations and networks about alliance and lawsuit (and detail relations) among these firms, and indicate the overall performance of our system. Then we will represent extracted relation keywords and show their effectiveness. Finally, we will show the application of our system to Semantic Web.

**Extracting Relation Keywords**

To extract relation keywords for each concrete relation, we gathered 456 pages and 165 pages, respectively, for alliance and lawsuit relations from Nikkei Net and IP News sites. After preprocessing and scoring, we obtained the highest scores as relation keywords. Table I shows the top five relation keywords and their Jaccard scores denoted as \( t_w \).

To evaluate the effectiveness of relation keywords, we compared information contained in retrieved pages merely by putting a pair of names as a search query to adding relation keywords to the query. We compared five methods as follows:

- **noW**: A firm pair (without relation keywords) is used as a query.
- **W1**: A firm pair and the top-weighted relation keyword \( (w_1) \) are used as a query.
- **W2**: A firm pair and the second-weighted relation keyword \( (w_2) \) are used as a query.
- **W1+ W2**: It generates two queries: W1 and W2.
- **W1+W2+noW**: It generates three queries: W1, W2, and noW.

---

5 Nikkei Net (http://release.nikkei.co.jp/) is a famous online business newspaper. IP News (http://news.braina.com/judge.html) is an online news archive of intellectual property issues.

6 For our experiment, we mainly used web pages in Japanese. Therefore, relation keywords are translated from Japanese.
Table I  Relation keywords extracted from the Web using a Jaccard coefficient

<table>
<thead>
<tr>
<th>Alliance relation</th>
<th>$t_w$</th>
<th>Capital alliance</th>
<th>$t_w$</th>
<th>Business alliance</th>
<th>$t_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>alliance AND corporate</td>
<td>1.000</td>
<td>operation AND capital</td>
<td>1.000</td>
<td>alliance AND business</td>
<td>1.000</td>
</tr>
<tr>
<td>alliance AND stock</td>
<td>0.878</td>
<td>capital AND operate</td>
<td>0.553</td>
<td>alliance AND corporation</td>
<td>0.475</td>
</tr>
<tr>
<td>alliance AND company</td>
<td>0.704</td>
<td>capital AND company</td>
<td>0.548</td>
<td>alliance AND operation</td>
<td>0.459</td>
</tr>
<tr>
<td>alliance AND system</td>
<td>0.565</td>
<td>capital</td>
<td>0.543</td>
<td>alliance AND develop</td>
<td>0.437</td>
</tr>
<tr>
<td>alliance AND business</td>
<td>0.534</td>
<td>capital AND manage</td>
<td>0.533</td>
<td>alliance AND company</td>
<td>0.432</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lawsuit relation</th>
<th>$t_w$</th>
<th>Claim phase</th>
<th>$t_w$</th>
<th>Accommodation phase</th>
<th>$t_w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>violate AND lawsuit</td>
<td>1.000</td>
<td>violate AND sue</td>
<td>1.000</td>
<td>lawsuit AND accommodate</td>
<td>1.000</td>
</tr>
<tr>
<td>violate AND claim</td>
<td>0.514</td>
<td>parent AND sue</td>
<td>0.533</td>
<td>accommodate AND company</td>
<td>0.648</td>
</tr>
<tr>
<td>violate AND judge</td>
<td>0.490</td>
<td>sue AND technology</td>
<td>0.486</td>
<td>accommodate AND announce</td>
<td>0.646</td>
</tr>
<tr>
<td>violate AND court</td>
<td>0.458</td>
<td>sue AND develop</td>
<td>0.483</td>
<td>accommodate AND develop</td>
<td>0.641</td>
</tr>
<tr>
<td>violate AND indemnity</td>
<td>0.444</td>
<td>sue AND relevance</td>
<td>0.469</td>
<td>accommodate AND product</td>
<td>0.640</td>
</tr>
</tbody>
</table>

The noW query is considered as the existing method (i.e. Mika and Matsuo’s method) as baseline of this evaluation; the others are proposed method variations. In all cases, we downloaded the same number of web pages. The other conditions are all same. For instance, one of the variations of our method $W1+W2+noW$ generates three queries W1, W2, noW, and downloads 10 pages in total for the three queries. For example, using W1 as the query we download 3 pages, 4 for W2, and 3 for noW.

Figure 4 shows the results. Overall, the proposed methods perform better than the existing method noW with respect to precision. The precision and recall are 65.7% / 95.0%, respectively, if we do not use relation keywords at all. Relation keywords improve the precision using the same number of downloaded documents. By integrating multiple queries (as $W1+W2+noW$ case), we can achieve the highest precision as 71.9% while maintaining a high recall (92.5%).

Extracting Relations and Networks
The obtained network for 60 firms in Japan is shown in Figure 5. Black lines represent alliances (bold lines show capital alliances and thin ones show business alliances) and red lines represent lawsuits (bold lines show the claim phase and thin ones show the accommodation phase).

Using our system described in previous section, we extract relationships among 60 firms. The precision and recall of our system are shown in Table II. For $60C_2 = 1770$ pairs of firms, 113 pairs actually show alliance relations. Our system correctly extracted 70 pairs. There were actually 21 and 100 pairs of capital and business alliances; our system extracted 9 and 60, respectively. Compared to alliances, the lawsuit relations show higher recall, probably because lawsuit relations are described in rather common formats using words such as judgment, lawsuit, or accommodate.
The simple pattern-based rule can extract relations between firms efficiently. Sometimes, it is unable to deal with complex meanings of sentences. Applying advanced relation extraction approaches, such as conversion of sentences into syntactic tree, might improve future results.

Although they are not comparable technically, we compared the data set against Nikkei Net and IP News, using the search functionality provided in these sites. We collected all alliance and lawsuit relations from each firm’s news articles appeared in these sites (Table III), and compared those relations to our results. The precision values at these sites are 100%, but the respective recall rates of alliance and lawsuit relations among 60 firms are low, at 22.8% and 68.8%, respectively, because these sites deal little with information related to small companies and foreign corporations. The alliance and lawsuit relations are easily obtainable from the web using our algorithm.

**Application**

The obtained network is useful in several ways. We might find a cluster of firms and characterize a firm by its cluster. Business experts often make such inferences based on firm relations and firm groups. For that reason, the firm network might enhance inferential abilities on the business domain. Also we might using obtained networks to recommend business partners based on structural advantages. As a related work, F. Gandon et al. build a Semantic
Web server that maintains annotations about the industrial organization of Telecom Valley to partnerships and collaboration (Gandon et al., 2005).

We present a prototypical example of applications using a network of firms. We calculate the centrality, which is a measure of the structural importance of a node in the network, for each firm on the extracted network (on alliance relations). Table IV shows the top ten firms by eigenvector and betweenness centrality. These firms have remained large and reliable corporations in Japan for decades. Interestingly, IBM, Livedoor, and Cisco are on the list. These firms might bridge two or more clusters of firms: IBM and Cisco are United States firms and form alliances with firms in multiple clusters; Livedoor is famous for its aggressive M & A strategy in Japan. Such information can only be inferred after extracting a network. There seem to be many potential applications that can make use of social networks in various analyses.

**Table IV Centrality**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Eigenvector</th>
<th>Betweenness</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Name</td>
<td>Value</td>
</tr>
<tr>
<td>1</td>
<td>Matsushita</td>
<td>0.366</td>
</tr>
<tr>
<td>2</td>
<td>Hitachi</td>
<td>0.351</td>
</tr>
<tr>
<td>3</td>
<td>NEC</td>
<td>0.289</td>
</tr>
<tr>
<td>4</td>
<td>Fujitsu</td>
<td>0.275</td>
</tr>
<tr>
<td>5</td>
<td>Toshiba</td>
<td>0.263</td>
</tr>
<tr>
<td>6</td>
<td>Rakuten</td>
<td>0.257</td>
</tr>
<tr>
<td>7</td>
<td>JustSystem</td>
<td>0.241</td>
</tr>
<tr>
<td>8</td>
<td>KDDI</td>
<td>0.208</td>
</tr>
<tr>
<td>9</td>
<td>Tokyo Electric</td>
<td>0.207</td>
</tr>
<tr>
<td>10</td>
<td>Seiko Epson</td>
<td>0.204</td>
</tr>
</tbody>
</table>

**DISCUSSION**

There are various important relationships among firms other than the alliance relation and lawsuit relation. For example, the mutual stockholding relation, capital combination, trade relation, personal relation (i.e. mutual dispatch of officials), rival and competitive relation. This paper deals with two types of relations, namely, alliance and lawsuit. The alliance relation is further distinguished by two detail relations: business alliance which includes contacting for new product development, service providing; and capital alliance which includes integration or transfer of business, merger and acquisitions. The lawsuit relation is distinguished by claim phase relation and accommodation phase relation. These relations are published by news articles or by news releases which might easily obtained from the Web. The rival and competitive relations can also be found from sites of product comparison, but different extraction method should be proposed, and our approach does not cover the area. Also, mutual stockholding, and personal relations are usually might be partially published on the Web, therefore they are not dealt with. We also plan to extend our algorithm to extract more kinds of relations as well as achieve higher performance in the future. For example, to modify queries using OR or NOT options so that we can retrieve more detail relations, to apply advanced text processing tools such as converting sentences into syntactic tree to improve the precision, or to deal with tabular data etc.

**CONCLUSION**

This paper described a method to extract inter-firm networks from the web. Given a list of names of firms, our system uses a search engine to collect target pages from the web, and applies text processing to construct a network of firms. To retrieve target pages we append the query with keywords indicating the relation. Moreover, we proposed an automatic method to extract such keywords from the web. Although we focused on alliance and lawsuit relations, in future we plan to extend the proposed method to other types of relations between firms.

**REFERENCES**

RBekkerman, R and McCallum, A. (2005), "Disambiguating web appearances of people in a social network", in Proceeding of the 14th World Wide Web Conference, Edinburgh, Scotland.


Bollegara, D. and Matsuo, Y. and Ishizuka, M. (2006), "Disambiguating personal names on the web using automatically extracted key phrases", in Proceeding of the 17th European Conference on Artificial Intelligence, Riva del Garda, Italy.


Papakyriazis N. V. and Boudourides, M. A. (2001), "Electronic weak ties in network organisations", in 4th German Online Research Conference, Germany.


Zhao, S. and Grishman, R. (2005), "Extracting relations with integrated information using kernel methods", in Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Michigan, USA.