博士学位論文

Extraction and Application of Social Networks from the World Wide Web
(Webからの社会ネットワークの抽出および応用)

A DISSERTATION SUBMITTED TO THE
GRADUATE SCHOOL OF INFORMATION SCIENCE &
TECHNOLOGY,
THE UNIVERSITY OF TOKYO

Fulfillment of the Requirements for the Degree of Doctor of Philosophy

Supervisor: Professor Mitsuru Ishizuka

Yingzi JIN（金 英子）

December 2008

Acknowledgments

I would like to thank all of the people who have assisted and inspired me during my doctoral study. First and foremost, I would like to thank my advisor Prof. Dr. Mitsuru Ishizuka for his kind and wise advice and the numerous opportunities that he has provided to me over the years. Prof. Ishizuka has always been accessible and willing to help his students with their research. For that reason, research life has been smooth and rewarding for me. I also thank the members of my review committee, Prof. Dr. Masaru Kitsuregawa, Prof. Dr. Jun Adachi, Prof. Dr. Takashi Chikayama, Associate Prof. Dr. Masashi Toyoda, and Associate Prof. Yutaka Matsuo for their helpful comments and insights related to in this thesis.

Special appreciation goes to Associate Prof. Dr. Yutaka Matsuo for his advice and encouragement on matters related to this thesis. He always encouraged me and provided insights, suggestions, and ideas that improved all of my graduate work; he also read each of my papers and commented helpfully on them. I would like to thank Dr. Naoaki Okazaki for advising me and providing me with the background knowledge in my research field. I also acknowledge the students and other staff members of Ishizuka Laboratory. Assistant Professor Hiroshi Dohi and secretary Tameko have given all students a favorable environment for doing research and have helped in many official procedures. In addition, I would like to thank Danushka Bollegala, Tadanobu Furukawa, Jie Yang, Yulan Yan, Haibo Li, Mostafa Al Masum, Aruturo Nakasone, Alena Neviarouskaya, and all other members of the laboratory, we have studied and grown up together while sharing knowledge and learning from one another.

As a research fellow of the Japan Society for the Promotion of Science (JSPS), I would like to thank JSPS for a grant that made it possible to complete this study. Furthermore, I would like to thank the JGC-S Scholarship Foundation for their assistance and encouragement of my research work.

Last but not least, I would like to express my full gratitude and love to my family. Special appreciation is extended to my parents and my sister who always encouraged me to follow my dreams and to assume all responsibilities arising from this. My parents-in-law have also sustained me substantially, treating me like their own daughter. I am especially grateful to my husband Li for his patience, understanding, support, and love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Social Network Analysis

Social network analysis has been pursued in social sciences since the 1930s. It characterizes social relations in terms of nodes and ties. Nodes are individual actors within networks (such as persons, companies, and organizations), and ties are the relations between actors (such as friendship, collaboration, and alliance). In the terms of theory used in the field social relations are emphasized over the attributes of individuals. Interaction patterns reveal relations among actors, which can be merged to produce valuable information as a network structure. Different from conventional data which apply specifically to actors and attributes, network data apply specifically to actors and relations. Therefore, network data are usually represented as matrices and graphs: matrices represent the adjacency of each actor to every other actor in a network; a graph (sometimes called a sociogram) comprises nodes (i.e. actors) connected by edges (i.e. relations), which are used for visualization and navigation of relations on the network.

The major emphases of network analysis are seeing how the individuals are embedded within a structure and how the whole pattern of individual choices gives rise to more holistic patterns. Many network properties such as degree, distance, centrality, and various kinds of positional and equivalence are analyzed in social network analyses. The following are noteworthy examples. The degree (in-degree and out-degree) of an actor informs us about the extent to which an actor might be constrained by, or constrain others. The extent to which an actor can reach others in the network might be useful in describing an actor's opportunity. The local connections of actors are important for understanding the social behavior of the whole population, as well as for understanding each. Several centrality measures (e.g., degree centrality, betweenness centrality, and closeness centrality) are used to identify the prominence or importance of an individual actor embedded in a network, which measures often engender distinct results with different perspectives of "actor location" i.e., local (e.g. degree) and global (e.g. eigenvector) locations, in a social network [107].

The power of social network analyses has become apparent in its use as an orienting idea and as a specific body of methods [91]. The Japan Society for Software Science and Technology (JSSST) has launched a panel–the Special Interest Group on Emergent Intelligence on Network (SIG-EIN)–to facilitate the study of social networks. The International Network for Social Network Analysis (INSNA) has held a Sunbelt Conference every year. The journal of "Social Networks" has published both theoretical and substantive papers. Social network analysis has emerged as a key technique for analyses undertaken in modern sociology, social psychology, information science, communication studies, and economics.

Application of Social Networks

Social networks are useful for analyzing social phenomena as well as business strategy. Regarding the first, "six degrees of separation" has been popularized by a famous experiment: as a sample, US individuals were asked to contact a particular target person by passing a message along a chain of acquaintances. The average length of successful chains turned out to be about five intermediaries or six separation steps, which underscored the small world phenomenon in US human society [56]. Subsequently, many researchers have described small world phenomena from various real-world networks such as small world on the Web [2, 49], small world from human language [34], and small world phenomena on e-mail of college students[108].

Organization or company networks can be used to enhance inferential abilities on the business domain and recommend business partners based on structural advantages. Gandon et al. build a Semantic Web server that maintains annotations about the industrial organization of Telecom Valley to partnerships and collaboration [37]. Battiston et al. extract shareholding relations from stock market information (MIB, NYSE and NASDAQ) to analyze characteristics of market structure [10]. Souma et al. extract data published by Tokyo Keizai Inc. to construct Japanese shareholding networks to analyze features of Japanese companies' growth [95].

In the context of the Semantic Web, social networks are crucial to realize a Web of trust that facilitates estimation of information's credibility and its provider's trustworthiness [41, 68]. Ontology construction is also related to social networks [74]: for example, if many people share two concepts, the two concepts might be related. Information sharing and recommendation [78, 39] on social networks are other applications that are served by the Semantic Web. Our lives are influenced strongly by social networks without our knowledge of their implications. For that reason, myriad applications are relevant to social networks

[97].

New trends on the World Wide Web

The World Wide Web (commonly shortened to the Web) was begun in 1989 by Tim Berners-Lee as a system of interlinked hypertext documents accessed via the Internet. The Web allowed for the spread of information over the Internet using an easy-to-use and flexible format. Recently, the trends of "Web 2.0" in the computer industry have cast the Internet as a platform that is intended to enhance the users' creativity, communications, information sharing, collaboration, and functionality of the Web. For instance, Social Network Service (SNSs) such as MySpace (http://www.myspace.com), Facebook (http://www.facebook.com/), Friendster (http://www.friendster.com/), and mixi (http://mixi.jp/) specifically build online communications of people who share interests and activities. The Semantic Web as an extension of the Web, specifically emphasizes the semantics of information and services on the Web, making it possible for machines to understand and use Web contents. A set of principles such as Resource Description Framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL), as a core of Semantic Web, are intended to provide a formal description of concepts, terms, and relations within a given knowledge domain. A popular application of the Semantic Web is Friend of a Friend (FOAF), which describes relations among people and others in terms of an RDF. Semantic Wave 2008 Report (http://www.project10x.com/) described the innovation of the Web from the "Web" (connect information) to the "Social Web" (connect people), "Semantic Web" (connect knowledge), and the "Ubiquitous Web" (connect intelligence) in view of the increasing social connectivity and connectivity & reasoning. New trends and innovations such as "Web 3.0" and "Web 4.0" etc. are progressing on the Web. As a consequence, the Web has grown to encompass immense amounts of widely distributed, interconnected, rich, and dynamic information.

Mining the Web

The current development of Internet environments such as broadband and wireless networks enable users to access the Web conveniently. Nearly 210 million people [1] in China and 88.1

[1] According to a report released by the China Internet Network Information Center（CNNIC) in 2007.

million people [2] in Japan are currently using the Internet. Moreover, the current development of Web applications such as Blogs and Wikis enable users to create their Web contents easily. With the rapid growth of contents on the Web, the quantity of information is becoming more important in the Web.

Mining the Web to discover knowledge stored in billions of Web pages is an important issue to preserve and develop the heritage and legacy of humankind. Web search engines such as Google (http://www.google.com/), Yahoo! Search (http://search.yahoo.com/), Baidu (http://www.baidu.com), and MSN Search (http://www.live.com/), which are designed to search for information related to the Web, serve as entrances to the Internet. The engine returns a listing of best-matching Web pages according to its criteria with the number of results when a user enters a query into a search engine. Users can specify the query using the Boolean operators AND, OR, and NOT. Many search engines provide a Web API (e.g., Yahoo! Search BOSS), which enables us to access to the search engine and obtain free search results supplied in the program. Using a search engine (via API) one can collect and download relative contents from the Web and we can measure the global popularity of a query on the entire Web by the hit number that is provided along with search results.

1.2 Contributions of the Thesis

The contributions of this thesis are summarized as follows:

- We expand social network mining from the Web so that is applicable to various domains. Two major improvements are proposed and described—*relation identification* and *threshold tuning*—which respectively examine complex and inhomogeneous communities on the Web. Because of those improvements, social network extraction becomes more generally applicable to various entities. We introduce general social network extraction, which can support existing studies using social networks in the Semantic Web in chapter 6.

- Because our method can extract relations from among entities, it can output machine-processable knowledge about the relations automatically from the information related

[2] Based on a survey of Japan Ministry of Internat Affairs and Communications in 2007

to current Web. Although some approaches exist to generate RDF statements by Web mining, our study provides an alternative.

- We show examples and evaluations for companies' and artists' networks. The social network of companies constructed by relation identification approach from the Web yield an overview of characteristics of companies' relational structural in an industry; the centrality of companies on the network reflects business activities on their strategies. Additionally, it is noteworthy that our system was operated on the Web site for the International Triennale for Contemporary Arts (Yokohama Triennale 2005), a famous exhibition of modern art, to navigate users using the extracted social network of artists. We briefly present an overview of that site.

- We further provide an example of advanced utilization of a social network mined from the Web. Based on the intuition that relations and structural embeddedness of actors are influential to predict features of entities, we constructed a ranking learning model from social networks to predict the ranking of other entities. The results emphasize the usefulness of our approach, by which we can understand the important relations as well as the important structural embeddedness to predict features of entities. We extract various networks from the Web to construct multi-relational networks to construct ranking models that are more suitable to explain real-world phenomena than single-relational networks. The proposed ranking learning model combines various network features; the model can be combined with a conventional attribute-based approach.

- Through social network extraction and application of social networks on the Web, this thesis presents a bridge between relation extraction and ranking learning for advanced knowledge acquisition for Web Intelligence.

1.3 Organization of the Thesis

This thesis is presented as follows. Part 1 specifically examines the first topic of social network extraction from Web using a general search engine. First, chapter 2 presents background knowledge and existing studies. chapter 3 presents definitions of the problems of social network extraction from the Web, and identifies important assumptions and shortcomings from previous approaches. Then chapter 4 and 5 introduce our proposed approaches respectively, which specifically address a complex and inhomogeneous community, and which use companies and artists as examples. Then chapter 6 proposes a general model of social network extraction and addresses our ideas to obtain various social networks from the Web. Part 2 specifically examines the second topic of application of a social network. It provides an example of advanced utilization of social networks mined from the Web. chapter 7 presents ranking of learning approaches based on extracted social networks. Finally, we describe salient conclusions reached through this study and areas that are promising for future work in chapter 8.

Chapter 2

Background and Existing Studies

As background knowledge, Web mining and Information Extraction are introduced first. Basic tasks of this field and introduce several recent studies are described. An introduction of fundamental ideas and indices in social network analysis is provided next, followed by presentation of some studies of social network extraction.

2.1 Web Mining and Information Extraction

Basic Tasks of Web Mining and Information Extraction

Concomitantly with the aggregation of the huge, diverse, and dynamic information available on the Web, many people confronted information overload (from the so-called information explosion) during the last decade [65, 45]. Therefore, Web mining research is of substantial importance in our lives for discovery of information and knowledge from the huge warehouse of information that is the Web. Four tasks have been assigned to Web mining research: finding resources, selecting information, discovering valuable patterns, and analyzing patterns. Finding resources means the process of retrieving the data from the text sources available on the Web such as electronic newsletters, news groups, blogs, and event information. Usually researchers perform crawling or use search engines to find resources on the Web. Selecting information is transforming collected resources by pre-processing such as removing stop words, and stemming for obtaining the desired representation such as finding phrases in the training corpus, and transforming the representation to relational form. Automatically discovering valuable patterns is an important development for additional machine learning and data mining techniques. By analyzing validation and interpretation of the mined patterns, we can discover and create knowledges. It implicitly covers the standard process of knowledge discovery in database (KDD) [33]

Web mining research is classifiable into three categories [57, 62]: Web usage mining, Web structure mining, and Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web usage data includes data from Web server access logs, proxy server logs, browser logs, user profiles, user sessions or transactions, cookies, user queries, bookmark data, and any other data. Web structure mining is undertaken to elucidate a model or useful knowledge underlying the link structures of the Web. Web content mining describes the discovery of useful information from the Web contents, data, and documents. Fundamentally, Web contents comprise data of several types such as textual, image, audio, video, and metadata, in addition to hyperlinks. Several studies specifically examine mining rich media data [53, 48, 88], but many studies are undertaken to examine text or hypertext contents. Data mining and text mining are related to, but yet different from, Web content mining because Web data are mainly semi-structured or unstructured, whereas data mining deals primarily with structured data, and text mining address only unstructured texts.

Information Extraction (IE) aims at extraction of relevant facts from the documents. Aiming at extracting relations and networks from the Web, our method can be regarded as Web mining and IE. A typical task of IE is to scan a set of documents written in a natural language and populate a database with the extracted information. Current approaches to IE use natural language processing techniques that specifically examine very restricted domains. For example, Message Understanding Conference (MUC) is a competition-based conference that specifically examines a predefined domain (e.g., MUC-1 and MUC-2 focused on naval operations messages, MUC-6 focused on news articles on management changes). Based on different data that IE might be focused upon, IE can be considered as two types: IE from unstructured text and IE from semi-structured data. Structural IE research usually uses the meta-information that is available inside the semi-structured data [94, 79]. The IE tasks from unstructured data typically use a rather basic to slightly deeper linguistic pre-processing such as syntactic analysis, semantic analysis, and discourse analysis before performing data mining [28, 22, 94]. For the research fields of information extraction from Web data, we can say that Web mining is a part of the IE field. Some studies in IE specifically investigate specific Web sites such as Wikipedia, homepages, SNSs sites. Some research efforts have used machine learning or data mining techniques to learn extraction patterns or rules for Web documents semi-automatically or automatically.

Recent Studies of Web Mining and Information Extraction

The new generations of the Web such as “Web 2.0” and “Semantic Web” in the computer industry have come to characterize the Internet as a platform that is intended to enhance users’ creativity, communications, information sharing, collaboration, and functionality of the Web. Mining the Web to discover knowledge stored in billions of Web pages is an important issue to preserve and develop the heritage and legacy of humankind.

In the following, we will introduce two main trends of recent studies of the Web mining and information extraction. Trends of studies are collecting, extracting, and mining user-generated data from SNSs, Blogs, and Wikis on the Web to discover knowledge. With the success and popularity of social network services (SNSs) such as MySpace (http://www.myspace.com), Facebook (http://www.facebook.com/), Friendster (http://www.friendster.com/), and mixi (http://mixi.jp/) on the Web, groups of people connect through the Internet with common interests. Many studies have been designated to analyze these user-generated data for social interest discovery, knowledge sharing, information recommendation, community discovery, etc. to serve social, educational, political, and business purposes [3, 111, 115, 93, 31, 35, 58, 61]. Adamic et al. [3] seek to understand Yahoo Answers (YA)'s knowledge sharing activity by analyzing the forum categories and clustering them according to content characteristics and patterns of interaction among the users. Yang et al. [111] examine the behavior of users on a big Witkey Web sites in China, Taskcn.com to observe several characteristics in users' activity over time for knowledge sharing. Zhou et al. [115] sample documents from CiteSeer and two other sites to construct multiple graphs (i.e., citation graph, author graph, and venue graph), and combine these graphs to measure document similarity for document recommendations. Singla et al. [93] collect chat-relations from MSN Messenger, and apply data mining techniques to analyze the relation between communication and personal behavior on the Web. Ding et al. [31] and Finin et al. [35] observe how social networks and the semantic Web are embodied in FOAF and how FOAF documents might be used to support Web-based information system based on a large collection (over 1.5 million) of real world Friend-of-a-Friend (FOAF) documents harvested from the Web. Li et al. [58] discover social interests based on user-generated tags. Blogs are important Web contents generated by users; they provide commentary or news on a particular subject and support users who want to write personal online diaries that often contain users' true voices, valuable opinions, and comments. Modeling online reviews [101], integrating opinions [63], and detecting informative and affective articles [80] from these contents are also hot topics in current Web mining and IE field research [4, 120]. In addition, numerous wiki Websites described by simplified markup language such as Wikipedia (http://wikipedia.org/wiki/), Citizendium (http://en.citizendium.org/wiki/), and TWiki (http://twiki.org/) are supported and maintained through collaboration of users. Many researchers mine and extract useful knowledge from wiki sites (particularly Wikipedia) [46, 110, 100].

Another trend of studies is the application and use of Web search engines to perform Web mining and information extraction. A Web search engine is a tool designed to search for information related to the Web. One can specify the query with Boolean operators—AND, OR, and NOT—to extract more specific information. Others can use global popularity of a query on entire Web by its hit number of results. And others can download top hit pages or snippets to improve analyses. Oyama et al. [83] build a domain-specific search engine by adding domain-specific keywords (called "keyword spices") to the user's input query and forwarding it to a general-purpose Web search engine. Question-answering systems also construct elaborate queries for using search engines [87]. Cimiano et al. [27] proposed Pattern-based Annotation through Knowledge on the Web *PANKOW*, which is a method employing an unsupervised, pattern-based approach to categorize instances with regard to an ontology. They composed it with candidate concept (e.g., "Country", "Hotel") to generate hypothesis phrases (e.g., "South Africa is a country","South Africa is a hotel") to categorize candidate proper noun (e.g., "South Africa") into a concept. They put these hypothetical phrases as a query to a search engine (e.g. Google) to obtain the number of hits, and to sum up the query results to a total for each instance-concept pair. Therefore, they categorize the candidate proper noun into their highest ranked concepts. The *PANKOW* used only the hit number, whereas the more advanced system *C-PANKOW* (Context-driven *PANKOW*) [27] is downloading and processing abstracts of the *n* first hits to avoid generation of numerous linguistic patterns and correspondingly large number of Google queries. The main idea of *PANKOW* and *C-PANKOW* is to approximate semantics by considering information about the statistical distribution of certain syntactic structures over the Web. Many natural language-processing applications use search engines to locate numerous Web documents or to compute the statistics over the Web corpus [20, 103, 32, 104]. Etzioni et al. [32] introduce a system called *KNOWITALL* that extracts facts, concepts, and relations from the Web. In fact, *KNOWITALL* formulates queries automatically based on its extraction rules to compose a search query. For example, it issues the query "cities such as" to a search engine, downloads each of the pages named in the engine's results; it then appropriates sentences on each downloaded page. Turney et al. [103, 32, 104] use a search engine to measure word co-occurrence probabilities for the purpose of word sense disambiguation, and Bollegala et al. [16] measure the semantic similarity between words using the search engines. On the other hand, many studies have extracted applied search engine to support computation of relations and similarities for people, words, etc. [51, 55, 73, 69, 70]. Kautz et al. developed a social network extraction system called the *Referral Web* [51]. The system uses a search engine to retrieve Web documents that include a given personal name. [55] Knees et al. classify artists into genres using co-occurrence of names and keywords of music in the top 50 pages retrieved using a search engine. Mika et al. developed *Flink*, a system for extraction, aggregation, and visualization of online social networks for the Semantic Web community [73]. A social network of 608 researchers from both academia and industry is extracted and analyzed. The Web-mining component of Flink, similarly to that used in Kautz's work, employs co-occurrence analysis. The strength of relevance of two persons *X* and *Y* is estimated by putting a query *X* AND *Y* to a search engine. If *X* and *Y* share a strong relation, then we can usually find additional evidence on the Web such as links found on home pages, lists of co-authors in technical papers, and organizational charts. Matsuo et al. developed a system called *Polyphonet*, which also uses a search engine to measure the co-occurrence of names [69, 70]. Our method of social network extraction can be characterized as one such approach that uses search engine results to extract and construct the social network for more various entities.

2.2 Social Network Analysis and Extraction

Basic Concept of Networks and Actors

Social network analysis use *graphs* and *matrices* to represent information about patterns of ties among social actors. Graphs (sometimes called a sociograms) have been widely used in social network analysis as a mode for formally representing social relations and quantifying important social structural properties, beginning with Moreno [77]. We would begin by setting each actor as a "node" with a label, and connecting them according to their relations using links. A graph might represent only one type of tie or relation (e.g. "friendship"), or more than one kind of relation (e.g., "friendship" and "kinship"). A graph that represents a single kind of relation is called a *simplex graph*, by contrast, multiple and various ties exist among actors might be illustrated in *multiple graphs* with the actors in the same locations in each. In a graph, each tie or relation might be directed, or it might be a tie that represents co-occurrence, co-presence, or a bonded-tie between the pair of actors. The "directed" ties (which can be binary, signed, ordered, or valued) are represented with arrows that have arrowheads, indicating who is directing the tie toward whom. The "co-occurrence" or "co-presence" graphs use the convention of connecting the pair of actors involved in the relation with a simple line segment (no arrowhead). The strength of ties among actors in graph might be nominal or binary, signed, ordinal, or valued. The nominal or binary tie represents the presence or absence of a tie, the signed tie represents a negative tie, a positive tie, or no tie, the ordinal tie represents whether the tie is the strongest, next strongest, etc. and the valued tie measured on an interval or ratio level.

Graphs are very useful to present an overview of social networks. However, for social networks that contain actors and/or relations of many kinds, it becomes visually complicated and difficult to see patterns. It is also possible to represent social networks in the form of matrices. The most common form of matrix in social network analysis is an "adjacency matrix" (sometimes called a sociomatix) comprising of many rows and columns and where the elements represent ties between the actors. Most simply, each element is binary. That is, if a tie is present, a one is entered in a cell; if no tie exists, a zero is entered. An adjacency matrix is often convenient to refer characteristics of relations of *x*. For example, when all the elements of a row of *x* are taken, they show who *x* chose as friends; when all of elements of column of *x* are taken, they show who chose *x* as a friend. Furthermore, if we summed the elements of the column vectors, it would be measuring how "popular" the *x* in the network; and if we summed the elements of the row of *x* that means who "active" the *x* in the network. Sometimes, it is helpful to rearrange the rows and columns of a matrix (i.e., "permutation" of the matrix) so that patterns are more distinct. The patterns and grouping of cells are useful to understand how some sets of actors are "embedded" in social roles or in larger entities. Social network analysis uses several other mathematical operations that can be performed on matrices for various purposes, such as matrix addition and subtraction, transposes, inverse, and matrix multiplication [107].

Many indices in social network analysis are useful to elucidate properties of network structures and embeddedness of actors. Local connections of actors are important for understanding their social behavior. The *network size* is usually indexed simply by counting of nodes. Because fully saturated networks are empirically rare, the *density* of the ties is usually examined to observe how close a network is, by calculating the population of all ties that could be present, compared to those that actually are present. The degree of an actor (in-degree and out-degree if the network is directed) tells us that the actor might be constrained by, or constrain others. The extent to which an actor can reach others in the network might be useful in describing an actor's opportunity structure. A commonly used approach to indexing the distances between actors is the geodesic. The *geodesic* is useful for describing the minimum distance between actors. The geodesic distances between pairs of actors are commonly used to measure closeness. The average geodesic distance for an actor to all others, the variation in these distances, and the number of geodesic distances to other actors might all describe important similarities and differences between actors in how and how closely they are connected to their entire population.

Common Indices in Social Network Analysis

Network analysis often describes the way in which an actor is embedded in a relational network as imposing constraints on the actor and offering the actor opportunities. Actors that face fewer constraints, and who have more opportunities than others, are in favorable structural positions. Social network analysis provides several different approaches to the notion of the power and centrality that attaches to positions in structured of social relations. Here, we review some basic measures of the "centrality" of individual positions. The *degree centrality* [82, 107] shows whether an actor has an advantaged position. When the actors have more ties to other actors, they have many ties. Therefore, they might have alternative ways to satisfy needs. Consequently, they are less dependent on other individuals. For companies as an example, if they have many ties, they are often third parties and deal markets through exchanges among others. They are able to benefit from this brokerage. Consequently, the degree centrality is an indicative measure of an actor's centrality and power potential in the network. The degree centrality measures only reflect the immediate ties that an actor has, rather than indirect ties to all others. One actor might be tied to numerous others, but those others might be rather disconnected from the network overall. In such a case, the actor could be quite central, but only in a local neighborhood. *Closeness centrality* approaches [15, 11] emphasize the distance of an actor to all others in the network by particularly addressing the geodesic distance from each actor to all others. One could consider either directed or undirected geodesic distances among actors. Simply in our thesis, we examine undirected ties. The sum of these geodesic distances for each actors is the "farness" of the actor from all others. We can convert this into a measure of closeness centrality by taking the reciprocal and norming it relative to the most central actor. *Betweenness centrality* [36] views an actor as being in a favored position if that the actor falls on the geodesic paths linking other pairs of actors in the network. That is, the more people depend on actor *x* to make connections with other people, the more power *x* has. If however, if two actors are connected by more than one geodesic path, and *x* is not on all of them, then the *x* loses some power, or rather, must necessarily share that power. The *eigenvector approach* [18] is an effort to find the most central actors in terms of the "global" structure of the network, and to pay less attention to patterns that are more "local". Other centrality indices such as "flow centrality", "information centrality" [98] are proposed for more various perspectives of importance of actors in network.

A common interesting aspect of social structures is in the *sub-structure* in terms of groupings or cliques. The number, size, and connections among the sub-groupings in a network can indicate quite a lot about the likely behavior of the network as a whole. Numerous useful algorithms have been developed for network analysis—*cliques*, *n-cliques*, *n-clans*, and *k-plexes*—to identify how larger structures are compounded from smaller ones. As the most common concept, a clique in a graph is a maximal complete subgraph of three or more nodes [64]. For relaxing the strong assumptions of clique, *n*-clique is defined as sub-structures where *n* stands for the length of the path allowed to make a connection to all members [6], *n*-clans is *n*-cliques with an additional condition limiting the maximum path length within a clique [76], and *k*-plexes allows that actors might be members of a clique even if they have ties to all but *k* other members [92]. Division of actors into cliques or "sub-groups" can be important for understanding how the network as a whole is likely to behave. For example, if the actors in one network form two non-overlapping cliques, the mobilization and diffusion might spread rapidly across the entire network. In contrast, if the actors form groups that do not overlap, traits might occur in one group and not diffuse to the other. Knowing how an individual is embedded in the structure of groups might also be extremely important for understanding that person's behavior. Some people might act as "bridges" between groups, others might be isolates; some actors might be cosmopolites, and others might be locals in terms of their group affiliations. All of these aspects of sub-group structure can be relevant to predicting the behavior of the network as a whole.

Discussing social roles and social positions in ways that are quite useful for social network analysis. The social position refers to a collection of actors, whereas the social role refers to the ways in which occupants of a position relate to the occupants of other positions [43]. At least two different meanings of "similar" positions of actors are used to indicate "structural equivalence," and "regular equivalence" [90]. Two actors are said to be exactly *structurally equivalent* if they have identical relations to all other nodes. Two actors are said to be *regularly equivalent* if they have the same profile of ties with members of other sets of actors that are also regularly equivalent. The structural equivalence is the oldest and currently the most widely used definition of equivalence for positional analysis of social networks. Actors who are structurally equivalent face nearly the same matrix of constraints and opportunities in their social relations. For examining structural equivalence, or similarity of network positions among actors, the Pearson Correlation, the Euclidean Distance, the proportion of matches are commonly used.

Social Network Extraction

Originally in the social sciences, sociologists conducted personal interviews and long term observation to collect network data. The typical approach of network questionnaire surveys was often performed to obtain social networks, e.g., asking "please indicate which persons you would regard as your friend." However, regularly posing such questions to many people entailed huge costs; responses were time-consuming and often difficult to obtain.

With the spread of information technology, many data are standardized and digitized into electric data. Many researchers have collected relational data from these electric data to construct social networks to analyze. Some examples are the following. Collection of "citation," "co-citation," "co-author," "co-present" relations among technical papers with authors from electric library or digital bibliography & library project (DBLP) is suggested in some reports [25, 7]; Extract of "shareholding," "owned-by" relations among companies from stock market information (e.g., MIB, NYSE, and NASDAQ) and the electric financial press (e.g. Tokyo Keizai Inc.) is described in other reports [10, 95]; Choosing "in conversation with" relations from archives of e-mail exchange, and telephone conversations is described in other papers [1, 72, 105]. Although the extraction method is simple and the confidence of the analysis result is high, however, the data are costly, lacking diversity, and their use is often hinderd by privacy concerns.

In contrast, social networking services (SNSs) provide various ways for users to interact (such as e-mail and instant messaging services) and to connect with friends (usually with self-descriptions), in addition to providing recommender systems linked to trust. Several studies have been undertaken to infer social networks from SNSs by analyzing relations between communication and personal behavior from MSN Messenger network [93], detecting conflicts of interest (COI) among potential reviewers and authors of scientific papers using "knows" relations from FOAF and "co-author" relations from DBLP. Current SNSs realize network questionnaires online. Nevertheless, the obtained relations are sometimes inconsistent: users do not name some of their friends merely because they are not in the SNSs or perhaps the user has merely forgotten them. Some name hundreds of friends, whereas others name only a few. Therefore, deliberate control of sampling and inquiry are necessary to obtain high-quality social networks on SNSs.

Another stream of studies treat the entire Web as a corpus from which to obtain social networks using a search engine. Simply put, they query a search engine about two names, then show how the two people are related. Co-occurrence of names on the Web is commonly used as proof of relation strength [51]. Related to the Semantic Web community, Mika developed a system called *Flink*, which extracts personal information from Web pages, emails, publication archives, and FOAF profiles [73]. The system uses a search engine to mine the strength of relations among researchers. Comparably, Matsuo and his colleagues developed a system called *Polyphonet*, mainly for use by the AI community in Japan [71, 70]. Its Web mining function extracts a social network of researchers using a search engine by identifying types of relations such as "co-authorship," "same-laboratory," "co-project," and "co-attendance" relations. Using search engines as an entrance to the Web, we can obtain social networks from structured or unstructured data, and obtain information about whether actors belong to the same community or individually appeared on the Web.

However, it is noteworthy that most studies of this genre target researchers or students. The reasons might lie in the fact that researchers are familiar to the researchers themselves and the relational evidence of researchers is readily obtainable from various online data sources. Admitting that the researcher domain is a useful test-bed because intuitive evaluation is crucially important for research and development, the next steps should be taken in domains other than those of researchers from both technical and commercialization perspectives. If we already possess methods to extract social networks for researchers, why not expand them to examine human relations in other professions, to analyze non-human social entities such as organizations and groups? This study is designed to expand current social network mining from the Web to apply it to other groups of entities other than researchers.

Part I

Social Network Extraction from the Web

Abstract

Social networks have recently attracted much attention for their importance to the Semantic Web. Several methods exist to extract social networks for people (particularly researchers) from the Web using a search engine. Our goal is to expand existing techniques to obtain social networks among various entities. In this part, we first introduce problems and assumptions in previous studies in the social network extraction field in chapter 3. Then we propose two improvements—*relation identification* in chapter 4 and *threshold tuning* in chapter 5—which enable us to address complex and inhomogeneous communities, respectively. Social networks among companies and artists (of contemporary) are extracted as examples: Results of several evaluations emphasize the effectiveness of these methods. Our system was used at the International Triennale of Contemporary Art (Yokohama Triennale 2005) to facilitate navigation of artists' information. This study contributes to the Semantic Web in that we increase the applicability of social network extraction for several studies.

Chapter 3

Problem Definition

3.1 Introduction

Social networks explicitly exhibit relations (called *ties* in social sciences) among individuals and groups (called *actors*). They have been studied in social sciences since the 1930s. To date, vastly numerous studies using social network analysis have been conducted [91]. In the context of the Semantic Web, social networks are crucial to realize a Web of trust that facilitates estimation of information's credibility and its provider's trustworthiness [42]. Ontology construction is also related to social networks: P. Mika presents discussion of the relation between the community and emergent ontology from a social network perspective [74]. Information sharing and recommendation [78, 39] on social networks are other applications that are served by the Semantic Web. Our lives are influenced strongly by social networks without our knowledge of their implications. For that reason, many applications are relevant to social networks [97].

Social networks are obtained from various sources, such as e-mail archives, FOAF documents, and DBLP. For example, Finin et al. extract a social network from the Web by collecting FOAF documents [35]. Particularly, several studies have been undertaken to use a search engine to extract social networks from the entire Web [51, 70, 73]. Co-occurrences of names on the Web, which are basically obtained by posing a query including two names to a search engine, is commonly used as proof of relational strength. Using a search engine to recognize the relation of two entities (or two words) has increasingly gained attention in the field of natural language processing [26, 52, 102].

This study is intended to expand current social-network mining techniques using a search engine to obtain a social network among various entities. Specifically in this part, two improvements are proposed to apply our method to complex and inhomogeneous communities: *relation identification* and *threshold tuning*. We extract two social networks as examples: artists of contemporary art, and famous companies in Japan. We must identify the relation types such as alliances and lawsuits; consequently, we can make elaborate queries and apply text processing to extract a social network among companies. Our algorithm adds a *relation keyword* to the search query to emphasize a specific relation. Extracting a social network of artists, on the other hand, requires adaptive tuning of thresholds because the appearance of each artist on the Web is completely different. Optimal thresholds are sought to invent appropriate edges between entities.

3.2 Related Works

Social Network Extraction from the Web

Numerous studies have obtained and analyzed social networks on the Web: Adamic collects relations among students from Web link structure and text information, and characterizes the social networks among Stanford students and MIT students [1]. T. Finin describes a large collection of FOAF documents (over 1.5 million) from the Web and analyzes the structure of friendship networks in the Semantic Web [35]. Trust calculation [42] is a major application of social networks. Some studies seek other applications: A. McCallum and his group present an end-to-end system that integrates both e-mail and Web content automatically to help users maintain large contact databases [29]. Aleman-Meza et al. use relational data from both FOAF and DBLP to detect relations among potential reviewers and authors of scientific papers [7].

Several studies have particularly addressed the use of a search engine for social network extraction. In the mid-1990s, H. Kautz and B. Selman developed a social network extraction system called the *Referral Web* [51]. The system uses a search engine to retrieve Web documents that include a given personal name. Recently, P. Mika developed *Flink*, a system for extraction, aggregation, and visualization of online social networks for the Semantic

Web community [73]. A social network of 608 researchers from both academia and industry is extracted and analyzed. The Web-mining component of Flink, similarly to that used in Kautz's work, employs co-occurrence analysis. The strength of relevance of two persons, *X* and *Y*, is estimated by putting a query *X* AND *Y* to a search engine. If *X* and *Y* share a strong relation, further evidence of the relation is usually available on the Web, such as links found on home pages, lists of co-authors in technical papers, and organizational charts. In Flink, the strength of relations among individuals is calculated using the Jaccard coefficient $n_{X \cap Y}/n_{X \cup Y}$, where $n_{X \cap Y}$ represents the number of hits yielded by the query *X* AND *Y* and $n_{X \cup Y}$ represents the number of hits by the query *X* OR *Y*. The two researchers are considered to share a relation if the value is greater than a certain threshold. The term "*Semantic Web* OR *ontology*" is added to the query for name disambiguation.

Matsuo et al. developed a system called *Polyphonet*, which also uses a search engine to measure the co-occurrence of names [69, 70]. In their study, several co-occurrence measures [66] have been compared, including the matching coefficient ($n_{X \cap Y}$), mutual information, Dice coefficient, Jaccard coefficient, and overlap coefficient. The overlap coefficient $n_{X \cap Y}/\min(n_X, n_Y)$ performs best according to the experiments. In addition, *Polyphonet* was operated at several AI conferences in Japan and a couple of international conferences to promote participants' communication. For disambiguating personal names, key phrases such as affiliations are added to queries.

We regard the two studies by Mika and Matsuo as relevant precedent studies, and propose some improvements to increase the applicability of that approach.

3.3 Problem of Existing Methods

The fundamental idea underlying the existing studies by Mika and Matsuo is that *the strength of a relation between two entities can be estimated according to the co-occurrence of their names on the Web*. The criteria to recognize a relation, such as the measure of co-occurrence and a threshold, are determined beforehand. An edge will be invented when the relation strength by the co-occurrence measure is higher than the predefined threshold. Although the approach is effective for extracting a social network of researchers, our preliminary study indicates that it does not perform well for various entities on the Web.

As the first reason, co-occurrence-based methods become ineffective when two entities co-occur universally on numerous Web pages. For example, when we want to infer two companies' relations from the Web, we submit a query "*Matsushita* AND *JustSystem*" [1] to a search engine. Consequently, we have designated as as many as 425,000 pages, for which the Jaccard coefficient is 0.031. However, this figure is unreliable considering the media effect on the Web. Regarding the domain of companies, many relations are published in news reports and on news releases that are distributed on the Web. Many Web pages describe and comment on the relation if the news is given attention by media services or people. Conversely, if it were not given attention, only a few pages would describe the relations. Considering that media effects influence the number of Web pages, co-occurrence of names on the Web is not always available to represent the relational strength of two entities.

For the second reason, co-occurrence-based methods function ineffectively when applied to *inhomogeneous* communities. An inhomogeneous community means, in this paper, a community that includes people in different fields, different nations, or different cultures, where a relation is difficult to obtain using a single criterion. Researchers' communities (of the same research field) present a homogeneous character; for that reason, using a single criterion to calculate the relation works well. In contrast, the international artist community is more inhomogeneous. For example, two Japanese artists, "*Taisuke Abe*" and "*Jun Oenoki*", have no prior relations, but their Jaccard coefficient is high: 0.024. Two international artists "*Beat Streuli*" from Switzerland and "*Nari Ward*" from Jamaica have co-participated in several exhibitions, but their coefficient is low: 0.0009. This happens because the community comprises many people from different contexts. For that reason, it is difficult to recognize the relation precisely using a single criterion.

We consider that the precedent studies of the research domain implicitly use the following two assumptions:

Assumption 1 Generally, Web pages are created according to results of two actors' co-participation in events. Therefore, the number of Web pages is assumed to show a useful correlation to the strength of two actors.

Assumption 2 A community to be extracted as a social network is assumed to be homogeneous.

[1] Both are names of famous Japanese corporations.

In the following section, we will introduce our ideas of improvements, *relation identification* and *threshold tuning* , which respectively mitigate violations of these assumptions. To emphasize the effectiveness of our methods, we apply each method to our case studies: Extracting social networks of companies in chapter 4 and artists in chapter 5. A general extraction model bundling these different extraction methods will be described in chapter 6.

3.4 Proposed Approach

3.4.1 Relation Identification for Complex Relations

In social sciences, the definition of a weak or strong tie might vary among contexts [67]. For example, the frequency or degree of relations affects that strength; multiple relations between two actors can also imply a stronger tie. In the company case, the type of relation defines the strength: For example, a capital alliance relation is stronger than a business alliance relation. Consequently, to present a tie among companies, it is appropriate that we identify the concrete relations of companies. As a solution, we add some word or combination of words to a search query. Using this strategy, we can identify relations among companies efficiently. 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 companies. Then we can conduct text processing to these pages to validate the relation's existence. This idea is similar to keyword spices [83], which extend queries for domain-specific Web searches. Question-answering systems also construct elaborate queries for using search engines [87].

We designate such a keyword to be added as a *relation keyword*. By adding relation keywords, we can extract particular relations among entities, which can be a solution for validation of **Assumption 1**. Below, we explain some issues about relation types and extraction of relation keywords.

Relation Types

A pair of entities is considered to have multiple relations. For example, two companies share alliance and lawsuit relations. Each relation is typed in a more detailed manner. Alliance relations between companies 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, it will be settled according to mutual accommodation or by final judgment. Consequently, the relation can be typed into the claim phase and the accommodation phase. For dynamic and complex relational networks, it is important to distinguish such typical and temporal relations for detailed analyses of social networks [67, 91].

Relation Keyword Extraction

We need some relation keywords to extract particular types of relations between companies. The intuitive method for finding relation keywords is to select terms that appear often in target pages (where the target relation is described) and which do not appear in other pages. Therefore, as a training corpus, we must collect annotated Web pages that describe specific relations of the companies. Once we identify appropriate relation keywords, we can extract relations among many companies.

Collecting and annotating the training corpus requires many hours of tedious work. For this study, we also try to use a search engine to extract relation keywords. This method is identical to that of Mori's work [78], in which a specific word w_c is assigned, which can represent the relation most precisely. If we want to retrieve an alliance relation, we add "*alliance*" (denoted as w_c) to a search query; words that co-occur frequently with it become good clues to discern the relation. We use the Jaccard coefficient $n_{w_c \cap w}/n_{w_c \cup w}$ to measure the relevance of word w to word w_c. The words w with large Jaccard coefficients are also used as relation keywords in addition to w_c. Use of those words would save costs of annotating training data with relevance or non-relevance manually.

3.4.2 Threshold Tuning for Inhomogeneous Communities

Commonly in studies of social network analysis, network questionnaire studies have been conducted. Typically, participants are asked "Please name your four closest friends." The respondents would then list the relations that are personally important. In other words, the relation is recognized using a subjective criterion for each participant. We propose to use this subjective criterion for the solution against **Assumption 2**. For example, even if the relation between "*Beat Streuli*" and "*Nari Ward*" is weaker than the objective standard, it is important to "*Beat Streuli*" if no other person has a stronger relation. Consequently, we might add an edge between them.

We employ two criteria that correspond to objective and subjective importance of relations for actors. We first invent edges using objective criteria with a consistent threshold T. Then we invent edges using subjective criteria for actors who have no certain number M of edges. This procedure alleviates the problem of some nodes having too many edges and some nodes being isolated. The combination of two criteria enables more exhaustive extraction for every node than the previous method, although it sometimes yields low precision. For that reason, we must find the appropriate parameters so that the target network is extracted as precisely as possible.

Setting Parameters for Each Community

Parameters vary according to the domain of a community. For example, T in the researcher community might be higher than that in artist community, simply because researchers' names are more likely appear on the Web than artists' names. Therefore, some training data are necessary for learning the appropriate values for each target community. Simply, the parameters are tuned so that the performance of relation identification is maximized: We maximize the F-value. Methods that are more effective to determine the parameters are bootstrapping or user interaction. Using the bootstrapping method, we can repeat the sampling and estimation process to determine parameters. Using the user interaction method, we can use the users' feedback to reconstruct the network dynamically with the best parameters that can maximize the F-value.

We apply each idea of method to our case studies: Extracting social networks of companies and artists, which are representative data of complex and inhomogeneous communities respectively. Details of the proposed methods are given in chapter 4 and chapter 5.

Chapter 4

Social Network Extraction for Complex Relations

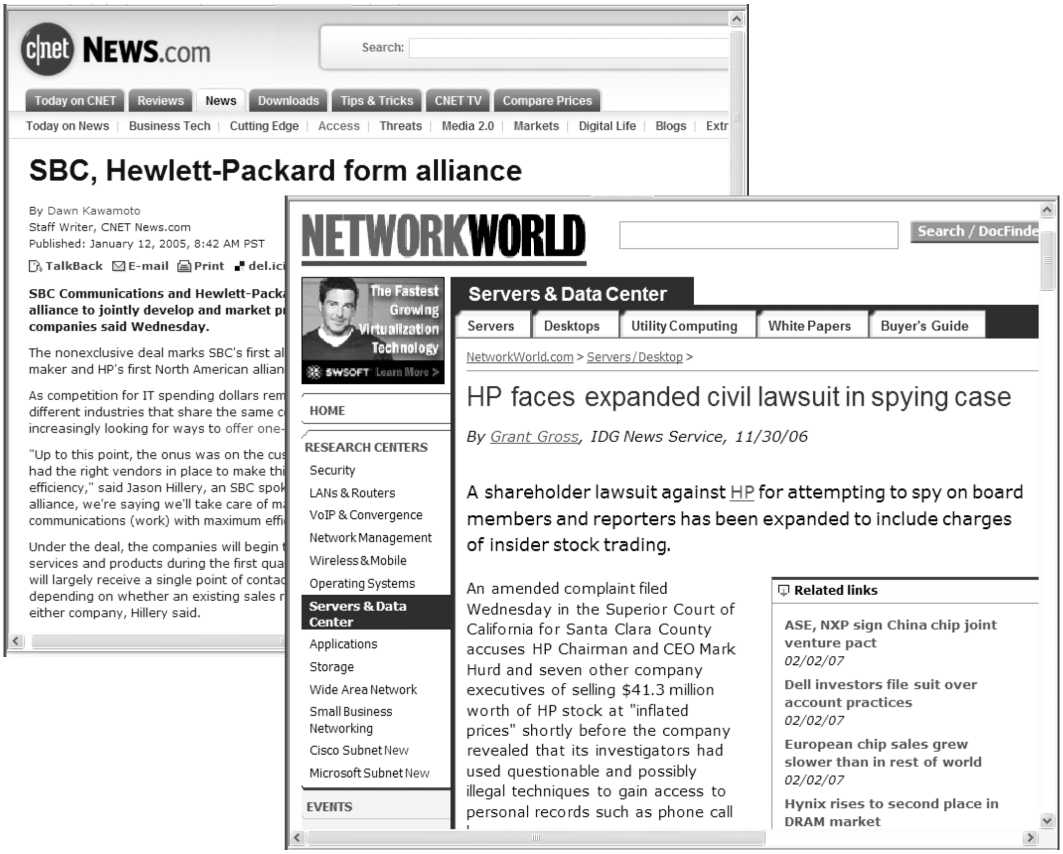
4.1 Introduction

Various relations exist among companies such as mergers, acquisitions, and partnerships. Together, these relations define a network among companies. Such networks are useful in analyzing a companies' competitiveness; they also help in determining 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 [13, 85, 10, 112].

Many researchers have examined methods to extract social networks from the Web while targeting people (particularly researchers or students). Some common examples are using social networking services (SNSs) and aggregating Friend-of-a-Friend (FOAF) documents [35, 75]. Particularly, several studies have been undertaken to use a search engine to extract social networks [51, 73, 69, 70]. Co-occurrence of names on the Web is used widely as proof of relational strength. However, the co-occurrence methods can not apply directly for some entities such as famous people, organization or companies, which have multiple relations, and relational information related to the Web affected by media effects. Many economic analyses of inter-company networks have been obtained using only relational data from the stock market or shareholding information available in business magazines, which are much less diverse [10, 95].

Many relations among companies are published on the Web in news articles or news releases (Fig. 4.1). Our work emphasizes the investigation of such published relations on the Web to address the relation extraction problem. Given a list of companies V={$v_1, v_2, ...$}, our goal is to retrieve and extract relations among them to construct inter-company networks $G(V, E)$, in which each edge e=(v_1, v_2) $\in$ E represents a relation between v_1 and v_2. We specifically seek to develop methods that acquire relations from the Web, the largest available resource that deals with all companies. For each pair of companies (v_1, v_2), our system addresses two problems: (a) collecting information about target relations, such as "Company v_1 merged with Company v_2"; and (b) relation extraction, such as extract capital alliance (*merge*) from the 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 [40, 83]. Research on relation extraction has been promoted at 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 [19], kernel methods [113], and logistic regression [50]. For the company case, our extraction task is to detect relations among same types (i.e., *COM* type) of entities.

For this study, we use a search engine to collect target relational pages from the Web. Because names of companies co-appear coincidentally on the Web, we propose to add additional words (call *relation keyword*) to name pairs of companies 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 relations is important for analyzing social trends, understanding social structures, and for formulating new industrial activities. Our method is a first attempt to extract inter-company networks from the Web using a search engine. Our approach is applicable to other entities such as famous persons and other multiple-relational entities.



SBC, Hewlett-Packard form alliance

By Dawn Kawamoto
Staff Writer, CNET News.com
Published: January 12, 2005, 8:42 AM PST

SBC Communications and Hewlett-Packa
alliance to jointly develop and market p
companies said Wednesday.

The nonexclusive deal marks SBC's first al
maker and HP's first North American allian

As competition for IT spending dollars rem
different industries that share the same c
increasingly looking for ways to offer one-

"Up to this point, the onus was on the cus
had the right vendors in place to make thi
efficiency," said Jason Hillery, an SBC spok
alliance, we're saying we'll take care of m
communications (work) with maximum effi

Under the deal, the companies will begin t
services and products during the first qua
will largely receive a single point of contac
depending on whether an existing sales r
either company, Hillery said.

HP faces expanded civil lawsuit in spying case

By Grant Gross, IDG News Service, 11/30/06

A shareholder lawsuit against HP for attempting to spy on board members and reporters has been expanded to include charges of insider stock trading.

An amended complaint filed Wednesday in the Superior Court of California for Santa Clara County accuses HP Chairman and CEO Mark Hurd and seven other company executives of selling $41.3 million worth of HP stock at "inflated prices" shortly before the company revealed that its investigators had used questionable and possibly illegal techniques to gain access to personal records such as phone call

ASE, NXP sign China chip joint venture pact
02/02/07
Dell investors file suit over account practices
02/02/07
European chip sales grew slower than in rest of world
02/02/07
Hynix rises to second place in DRAM market

Figure 4.1: News about companies' relations on the Web

4.2 Social Network Extraction for Companies

4.2.1 Basic Concept

In social sciences, the definition of a weak or strong tie might vary among contexts [67]. For example, the frequency or degree of relations affects that strength; multiple relations between two actors can imply a stronger tie. In the company case, the type of relation defines the strength: For example, a capital alliance relation is stronger than a business alliance relation. Consequently, to present a tie among companies, it is appropriate that we identify the concrete relations of companies.

For using a search engine to retrieve and extract relations, a proper query is necessary. The intuitive query is the names of the two companies. For example, we issue a query such as ”*Matsushita* AND *JustSystem*”[1] to discover data that are helpful to define their relations. Thereby, we obtain as many as 425,000 pages. Many top-ranked pages are lawsuit-relation pages[2], which drew much attention during the last year. Therefore, analyzing these pages, we were able to identify lawsuit relations among them. However, the two companies exhibited a collaboration relation in knowledge management in 2001, for which pages are in lower ranks of 124^{th}; on account of the collaboration relation that prevailed 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 a *relation keyword*) to a search query and apply text processing to confirm the existence of that fact. Using this strategy, we can identify relations among companies efficiently. For example, when we wish to extract lawsuit relations, we merely 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 companies. Then we can conduct text processing to these pages to validate the relation’s existence. This idea is similar to keyword spices [83], which extend queries for domain-specific Web searches. Question-answering systems also construct elaborate queries for using a search engine [87]. 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 companies and specific relation types are given as an input and the output is a social network of companies. 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 Fig. 4.2.

[1] Both are names of famous Japanese corporations.

[2] http://pc.watch.impress.co.jp/docs/2005/0201/just2.htm

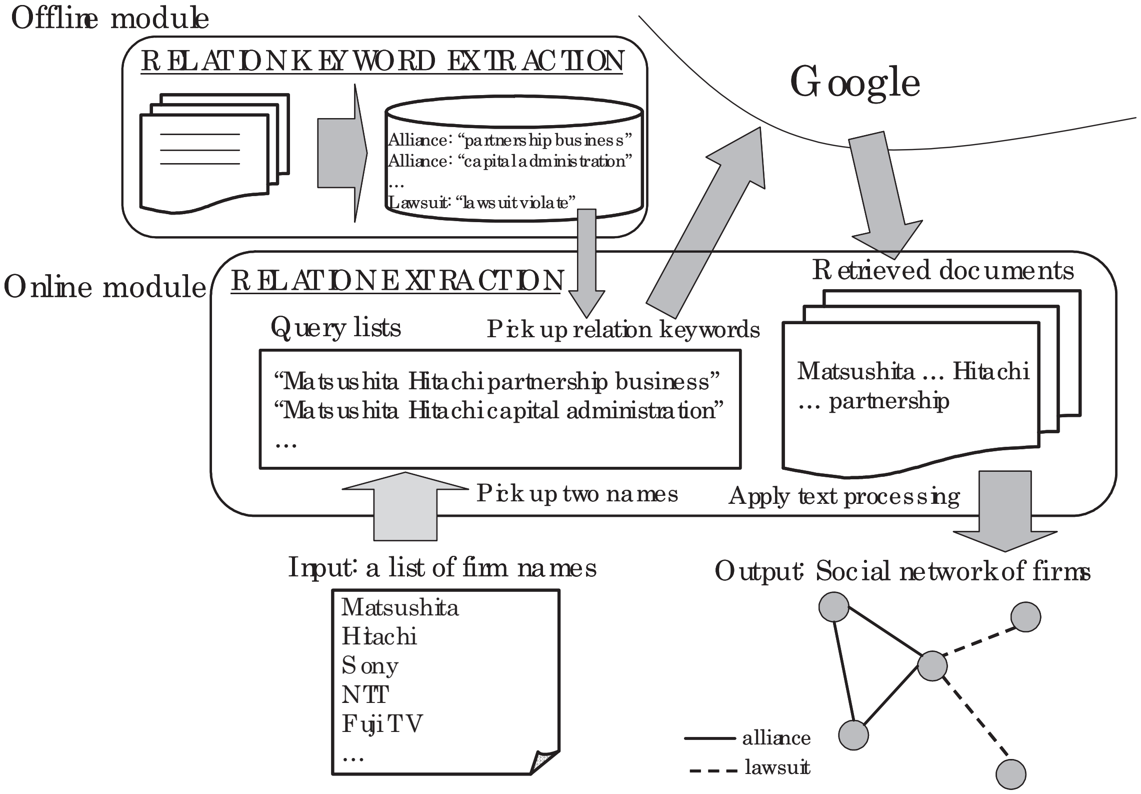

Offline module
RELATION KEYWORD EXTRACTION
Alliance: "partnership business"
Alliance: "capital administration"
...
Lawsuit: "lawsuit violate"
Google
Online module
RELATION EXTRACTION
Retrieved documents
Query lists
Pick up relation keywords
"Matsushita Hitachi partnership business"
"Matsushita Hitachi capital administration"
...
Matsushita ... Hitachi
... partnership
Pick up two names
Apply text processing
Input: a list of firm names
Matsushita
Hitachi
Sony
NTT
Fuji TV
...
Output: Social network of firms
alliance
lawsuit

Figure 4.2: System flow to extract a company network.

4.2.2 Relation type

Relations among companies are various: capital combinations such as mergers, acquisitions, joint ventures, and business partnerships, such as business alliances, co-development, service provision, and dispatching personnel, competition, and lawsuit. It is considered that pairs of companies have multiple relations. For example, two companies have alliance and lawsuit relations. Each relation is typed in a more detailed manner. Alliance relations between companies 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 being 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 [67, 91].

4.2.3 Relation Keyword Extraction

In this section, we describe relation keyword extraction methods that are useful to collect relation pages from the Web, and that are useful for the relation extraction procedure. Good relation keywords are expected to 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 must collect annotated Web pages of specific relations of the companies 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 [78], 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.

$$J_{w_c}(w) = |w_c \cap w|/|w_c \cup w|, \tag{4.1}$$

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 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 readily obtainable from several news articles from sources such as Yahoo! News for 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 extract nouns

[3] http://chasen.naist.jp/hiki/ChaSen/

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 used *alliance* AND *corporate* as w_c for alliance relations. Additionally, we use the word that appeared in the first lines in Table 4.1 as w_c for each relation: We determine these words through preliminary experiments.). 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, 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 types r. Once we find the relation keyword, we can extract the relations among many companies. For detailed relations, it is necessary to prepare relation keywords for each detailed relation, but extraction methods for relation keywords are similar.

4.2.4 Relation Extraction

Online, a list of companies and specific relation types is given as an input; the output is a social network of companies. Three steps are used: making queries, Google search, and network construction. First, we make queries by adding relation keywords to each pair of companies. We use the top n_q relation keywords from Table 4.1. Then, we put these queries into the Google search engine to collect top-n_p Web pages. For this experiment, we set $n_q = 2$ 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 portrayed in Fig. 4.3) has been useful in our experience. We first select all sentences S that include the two companies' names (x and y) and assign each sentence the sum of relation keyword scores t_w in the sentence. The score of companies x and y is the maximum of the sentence scores. An edge is invented between the two companies if $score_{xy}$ is greater than a certain threshold, i.e., if the two companies seem to have the target relation with high reliability.

[4] Here, $tf * idf = tf(w) * log(N/|w|)$, 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

```
**function** $R_{ELATION}E_{XTRACTION}(D, x, y, W)$ $score_{xy} \leftarrow 0$
$S \leftarrow$ GetSentences($D$,$x$,$y$)
**for each** $s \in S$ **do**
    **if** $s$ **contains** "$x$" *and* $s$ **contains** "$y$" **then**
    $score_s \leftarrow \sum_{w_i(\in W)\ \textbf{contained in}\ s} t_{w_i}$
    **if** $score_s > score_{xy}$ **then**
        $score_{xy} \leftarrow score_s$
**done**
**if** $score_{xy} > score_{thre}$ **then**
    **do** set an edge between $x$ and $y$ in $G$
**done**
```

Figure 4.3: A procedure to extract relations using text processing.

4.3 Experiments

A network of 60 companies in Japan including IT, communication, broadcasting, and electronics companies is extracted. For the dataset, we manually created a dataset for these 60 companies. The annotators decided the relations among the companies using only the information available on the Web. These experiments first show the extracted relations and networks for alliance and lawsuit (and detail) relations among these companies. Results will also be useful to assess the overall performance of the system. Subsequently, the relation keywords are extracted and their effectiveness is evaluated. Finally, the application of this system to the Semantic Web will be demonstrated.

4.3.1 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 [5]. After preprocessing and scoring, we obtained the highest scores as relation keywords. Table 4.1 portrays the top five relation keywords and their Jaccard scores denoted as t_w [6].

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 to evaluate the effectiveness of the relation keywords. We compared five methods as follows:

- **noW:** A company pair (without relation keywords) is used as a query.
- **W1:** A company pair and the top-weighted relation keyword (w_1) are used as a query.
- **W2:** A company 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.

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 identical. For instance, one variation of our method **W1+W2+noW** generates three queries W1, W2, noW, and download 10 pages in all for the three queries. For example, using W1 as the query we download 3 pages, 4 for W2, and 3 for noW.

Fig. 4.4 and Fig. 4.5 depict 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

[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 4.1: Relation keywords extracted from the Web using a Jaccard coefficient.

Alliance relation	t_w	**Capital alliance**	t_w	**Business alliance**	t_w
alliance AND *corporate*	1.000	*operation* AND *capital*	1.000	*alliance* AND *business*	1.000
alliance AND *stock*	0.878	*capital* AND *operate*	0.553	*alliance* AND *corporation*	0.475
alliance AND *company*	0.704	*capital* AND *company*	0.548	*alliance* AND *operation*	0.459
alliance AND *system*	0.565	*capital*	0.543	*alliance* AND *develop*	0.437
alliance AND *business*	0.534	*capital* AND *manage*	0.533	*alliance* AND *company*	0.432

Lawsuit relation	t_w	**Claim phase**	t_w	**Accommodation phase**	t_w
violate AND *lawsuit*	1.000	*violate* AND *sue*	1.000	*lawsuit* AND *accommodate*	1.000
violate AND *claim*	0.514	*patent* AND *sue*	0.533	*accommodate* AND *company*	0.648
violate AND *judge*	0.490	*sue* AND *technology*	0.486	*accommodate* AND *announce*	0.646
violate AND *court*	0.458	*sue* AND *develop*	0.483	*accommodate* AND *develop*	0.641
violate AND *indemnify*	0.444	*sue* AND *relevance*	0.469	*accommodate* AND *product*	0.640

multiple queries (as **W1+W2+noW** case), we can achieve the highest precision as 71.9% while maintaining a high recall (92.5%).

4.3.2 Extracting Relations and Networks

The obtained network for 60 companies in Japan is portrayed in Fig. 4.6. Bold lines represent capital alliances, thin lines are business alliances, dashed lines represent the claim phases in lawsuit relations and dotted lines are accommodation phases in the lawsuits.

Using our system, as described in Section 4, we extract relations among 60 companies. The precision and recall of our system are presented in Table 4.2. For ${}_{60}C_2 = 1770$ pairs of companies, 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*.

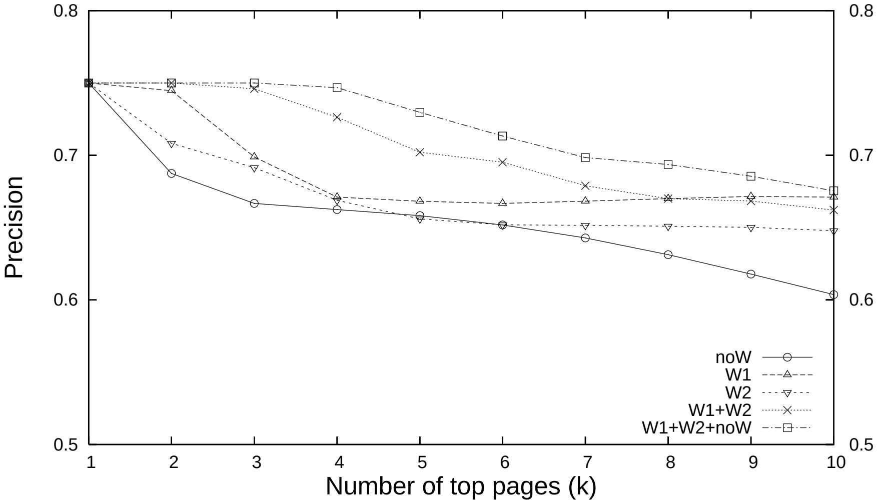


Figure 4.4: Precision of retrieved pages.

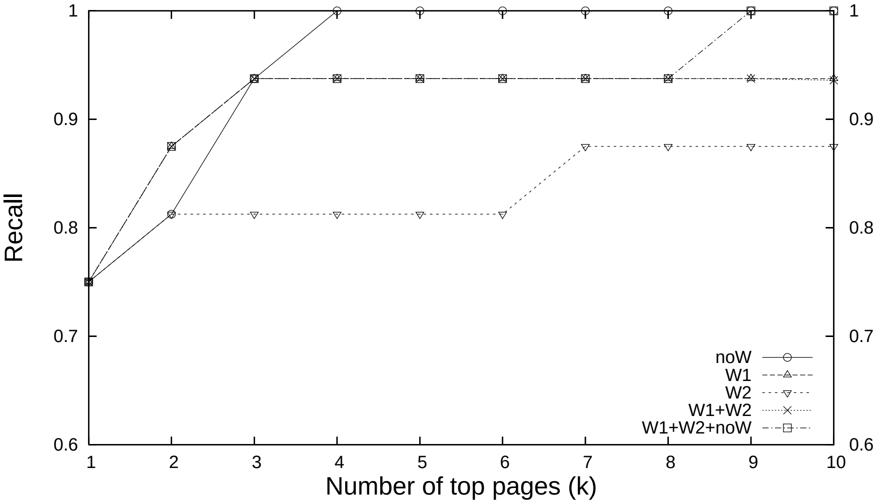


Figure 4.5: Recall of retrieved pages.

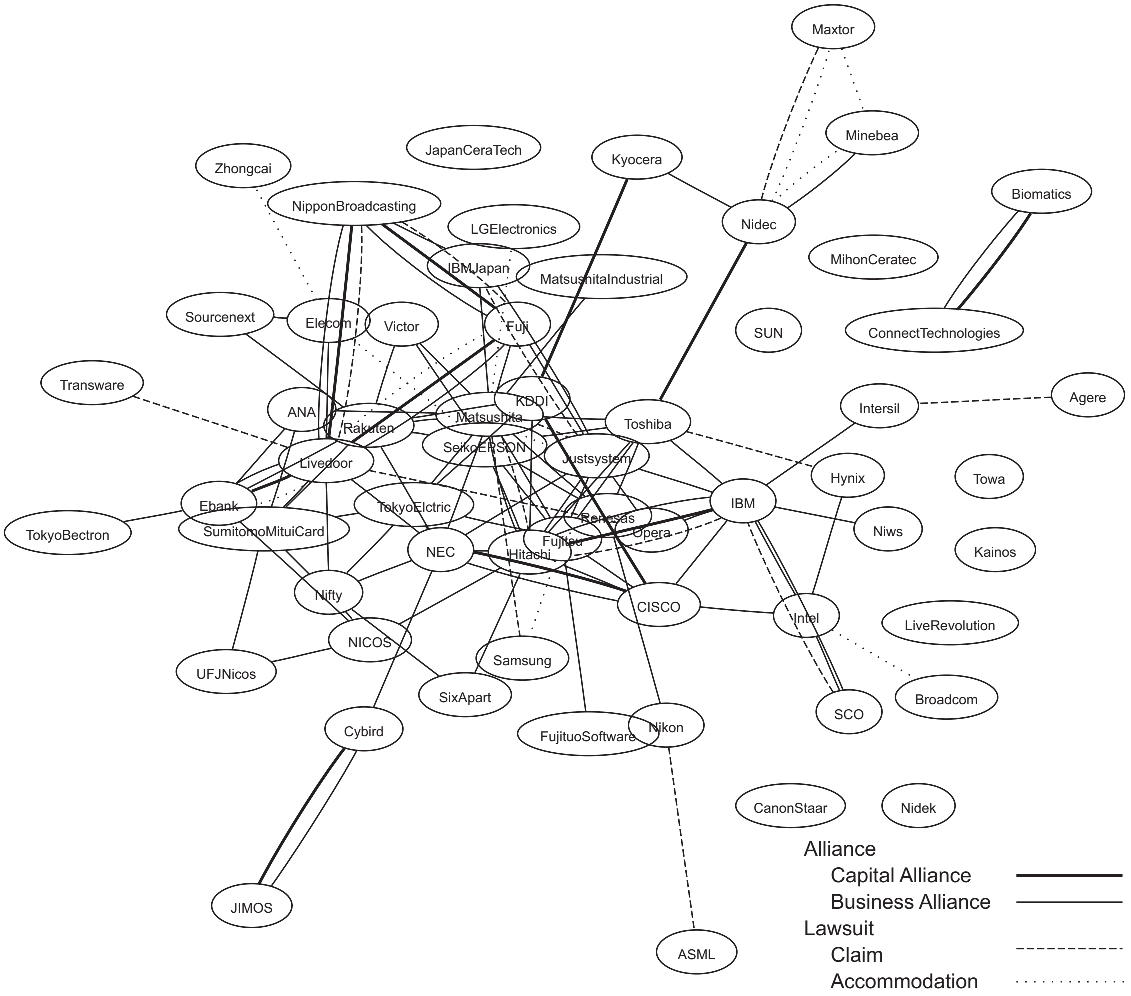


Figure 4.6: Network of 60 companies in Japan.

Table 4.2: Precision and recall of the system.

Target relation	Precision	Recall
Alliance	60.9% (70/115)	62.0% (70/113)
Capital alliance	75.0% (9/12)	42.9% (9/21)
Business alliance	67.4% (60/89)	60.0% (60/100)
Lawsuit	61.5% (16/26)	100% (16/16)
Claim phase	63.6% (14/22)	87.5% (14/16)
Accommodation	72.7% (8/11)	88.9% (8/9)

Table 4.3: Precision and recall in a particular Web site.

Target relation	Precision	Recall
Alliance	100.0% (27/27)	23.8% (27/113)
Capital alliance	100.0% (6/6)	28.6% (6/21)
Business alliance	100.0% (21/21)	21.0% (21/100)
Lawsuit	100.0% (11/11)	68.8% (11/16)
Claim phase	100.0% (11/11)	68.8% (11/16)
Accommodation	100.0% (6/6)	66.7% (6/9)

The simple pattern-based rule can extract relations between companies efficiently. Sometimes, it is unable to address 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 dataset against Nikkei Net and IP News, using the search functionality provided in these sites. We collected all alliance and lawsuit relations from each company's news articles appeared in these sites (Table 4.3), 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 companies 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 readily obtainable from the Web using our algorithm.

4.3.3 Application

The obtained network is useful in several ways. We might find a cluster of companies and characterize a company by its cluster. Business experts often make such inferences based on company relations and company groups. For that reason, the company network might enhance inferential abilities on the business domain. Furthermore, we might use the obtained networks to recommend business partners based on structural advantages. As a related work, F. Gandon et al. described a Semantic Web server that maintains annotations about the industrial organization of Telecom Valley to partnerships and collaboration [37].

We present a prototypical example of applications using a network of companies. We calculate the *centrality*, which is a measure of the structural importance of a node in the network, for each company on the extracted network (on alliance relations). Table 4.4 shows the top 10 companies by eigenvector and betweenness centrality. These companies have remained large and reliable corporations in Japan for decades. It is of particular interest that IBM, Livedoor, and Cisco are on the list. Many of these companies might bridge two or more clusters of companies: IBM and Cisco are United States companies that form alliances with companies 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.

4.4 Related works

Many studies have used search engines to extract social networks automatically from the Web [51, 73, 69, 70]. Co-occurrence of names on the Web is commonly used as evidence of relational strength [51]. 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 [73]. 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 *Polyphonet*, mainly for use

Table 4.4: Centrality of companies in the extracted social network.

Rank	Name	Value
1	Matsushita	0.366
2	Hitachi	0.351
3	NEC	0.289
4	Fujitsu	0.275
5	Toshiba	0.263
6	Rakuten	0.257
7	JustSystem	0.241
8	KDDI	0.208
9	Tokyo Electric	0.207
10	Seiko Epson	0.204

(a) Eigenvector centrality.

Rank	Name	Value
1	Matsushita	168.981
2	IBM	149.192
3	NEC	144.675
4	Hitachi	136.978
5	Toshiba	113.239
6	Rakuten	109.887
7	JustSystem	77.175
8	Livedoor	74.141
9	CISCO	64.558
10	Fujitsu	56.081

(b) Betweenness centrality.

by Japan's AI community [70]. However, 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 might be on the same news pages merely because they made some public statements on the same day. They might be on the pages that list "the most famous people 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 in a search for companies: a company name is similar to a famous person's name; they often co-occur for various reasons, although no formal relations exist among them. When the relation between companies 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 companies.

Web search by query modification and expansion is described in [40]; they extract 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 [83]. They add keywords called "keyword spices" to the user's input query with a Boolean expression for a domain-specific Web search. They sample Web pages using initial keywords and then classify them manually as either relevant or irrelevant, thereby producing a training corpus. Subsequently, they apply a decision-tree learning algorithm to discover keyword spices. Our system sets relation keywords that are added to a query as combinations of one or two terms. Therefore, a Jaccard coefficient is used simply to measure the scores. Other studies such as Flink uses 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 [17].

Battiston et al. extract shareholding relations from stock market information (MIB, NYSE and NASDAQ) and then use those relations to analyze market structure characteristics [10]. Souma et al. extract data published by Tokyo Keizai Inc. to construct Japanese shareholding networks for use in analyzing features of Japanese companies' growth [95]. Our work specifically addresses alliance and lawsuit relations among companies 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 poses an important problem for social network mining. Several reports describe attempts at personal name disambiguation on the Web [12, 17, 59]. However, ambiguity in company (or organization) names is less than that for personal names. We intend to explore ambiguities in company names in our future work.

4.5 Discussion

Various important relations other than alliance relation and lawsuit relation can link companies. For example, a mutual stock-holding relation, capital combination, trade relation, personnel relation (i.e. mutual dispatch of officials), rival and competitive relation. This chapter deals with relations of two types: alliance and lawsuit relations. The alliance relation is further distinguished by two detailed relations: business alliance which includes contacting for new product development, service providing; and capital alliance which includes intergration or transfer of business, merger and acquisitions. The lawsuit relation is distinguished as claim phase relation or accommodation phase relation. These relations are published by news articles or by news releases that might be obtained easily from the Web. Rival and competitive relations can also be found from sites of product comparison, but different extraction methods should be proposed; the approach presented herein does not cover the area. In addition, mutual stockholding and personal relations might be partially published on the Web. Therefore, they are not addressed herein.

In the future, we will extend our algorithm to extract relations of more varieties as well as achieve higher performance. 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 a syntactic tree to improve the precision, or to address tabular data.

4.6 Conclusion

This chapter described a method to extract inter-company networks from the Web. Given a list of names of companies, our system uses a search engine to collect target pages from the Web, and applies text processing to construct a network of companies. 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 particularly examined alliance and lawsuit relations, we plan to extend the proposed method to other types of relations between companies in future studies.

Chapter 5

Social Network Extraction for Inhomogeneous Communities

5.1 Introduction

Social network analyses elicit relations (called *ties* in social sciences) among individuals and groups (called *actors*). They have been studied in social sciences from the 1930s. To analyze social phenomena and social structures [116, 118, 107]. Recently, the relations among individuals have attracted much attention in the information technology field; many studies have been undertaken to connect the information technology field with social network analysis.

It is necessary to acquire relations among individuals to construct a social network for individuals. Originally, these data were collected through interviews, questionnaires, and long-term observations in social science fields. For instance, the network questionnaire done by the General Social Survey (GSS) of America asked respondents to “Please indicate which persons you would regard as your friend.” Using these questions, they can construct and analyze social networks for individuals. However, it is difficult to conduct questionnaires regularly with many people.

On the other hand, many researchers have sought to construct social networks from various sources such as e-mail archives, schedule data, and Web information [1, 105, 75]. The

Web contains 85 billion pages [1], which can be regarded as a reflection of a certain type of human society. Daily news about people and companies, ongoing events, conferences, and exhibitions, personal homepages, blogs: almost all of this information is obtainable from the Web. Therefore, many social scientists have increasingly been devoting attention to the Web [109], and have tried to extract and analyze social networks from the Web.

An early system *Referral Web* used for extracting social networks from the Web was developed by Kautz and Selman in the mid-1990s. The system uses a general search engine to discern a path from a person to a person (e.g., from Henry Kautz to Marvin Minsky) automatically. It finds experts who are connected closely with a person [51]. Another system called *Flink* was developed by Mika during 2006–2007 for extraction, aggregation, and visualization of online social networks for the Semantic Web community [73]. *Polyphonet* also uses a search engine to assess relations among researchers, which system was operated at several AI conferences in Japan (17th, 18th, and 19th Annual Conferences of the Japan Society of Artificial Intelligence) and at The International Conference on Ubiquitous Computing (UbiComp 2005) to promote participants' communications [70]. In addition, more studies have been done for extracting persons related to a given topic from the Web [117], using both e-mail and Web to extract personal relations [12], and using citation and co-citation information to extract networks on the Web [75]. Our study is apparent as one study of this field. Especially, our goal is extracting relations among artists (of contemporary art, performance, and architecture fields), which algorithm considers a new point of view with an approach that has heretofore never been considered.

Some studies of constructing human social networks from the Web (specifically [51, 73, 121, 70, 117]) have used hit numbers from a search engine to measure the relational strength among people. For example, to calculate the relation strength between person x and y, first they put a query "x AND y" to a search engine; then, based on a co-occurrence measure such as the Jaccard coefficient or Overlap coefficient they measure the strength of co-occurrence of names. Then, based on an objective threshold constant for entire community (e.g., connect pairs of names which the number of co-occurrence is higher than 50) they judge the existence of relations from co-occurrence strengths. Finally, they construct a social network. As described herein, we designate this approach as an *objective rule-based approach* that

[1] reported by the Wayback Machine of Internet Archive, which has crawled Web pages since 1996. http://www.archive.org/web/web.php.

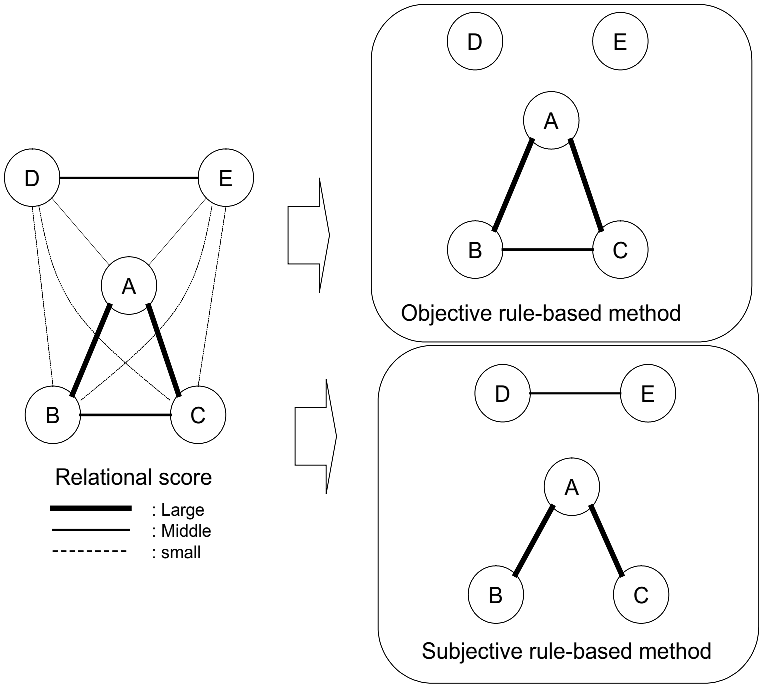


Figure 5.1: Extracted networks based on an objective rule and a subjective rule.

constructs networks according to an objective criteria for an entire community.

However, the objective rule-based approach functions ineffectively when applied to an *inhomogeneous community*. An inhomogeneous community means, in this chapter, a community that includes people of different fields, different nations, or different cultures, where a relation is difficult to obtain using a single criterion (e.g., participants of artists in international exhibition). Although the objective rule-based approach is effective for finding central nodes and main communities in a network, however, the typical problem of this approach is isolating many nodes in the network. (In contrast, to connect isolated nodes, if we were to lower the threshold of objective criteria, the network would become overly heavy.) Therefore, for the purpose of visualization and analyzing inhomogeneous community, it is inappropriate for using the previous approach directly. Originally, the network questionnaire done for social science tasks asked: ”Please indicate which people you would regard as your friend.” There have no objective criteria for an entire community (i.e., no comparison of the importance of relation between person *x* and *y* or with *x* and *w*), but based on the subjective importance of relations for each to construct networks. We designate this approach as the *subjective rule-based approach* in this chapter.

Fig. 5.1 shows a difference of frames between the objective rule-based approach with the subjective rule-based approach for network construction. The edge thickness signifies the relational strength, as calculated by co-occurrence measure. Here, let us consider only the remaining three edges for networks using two approaches. By the objective rule-based approach, edges A–B, A–C, and B–C will be selected based on the decreasing order of co-occurrence strength. However, by the subjective rule-based approach, edges of A–B, A–C, and D–E will be selected based on the highest co-occurrence strength from each perspective of nodes. Therefore, we can obtain different structure of networks using different approaches, even though the co-occurrence strengths were equal. Furthermore, in this case, the network constructed using the subjective rule-based approach appropriately showing high centrality of A as well as two different communities (A–B–C and D–E).

It is not necessarily appropriate to suggest which approach is best for network construction. For example, if the goal is identifying the central persons or main communities in the network, the objective rule-based approach might be more appropriate. However, if the goal is finding out marginal communities, visualizing the entire figure of communities, and supporting the navigation of relations, the subjective rule-based approach is more appropriate. Therefore, it is important to identify and combine these two approaches for different purposes of network construction.

This chapter presents a proposal of an advanced algorithm for network construction, which combines ideas from objective and subjective importance of relations by proper adjustment of parameters. Experimental results underscore the effectiveness of proposed approach. Next, we will discuss properties of parameters further. Our system was operated on the Web site for the International Triennial for Contemporary Arts (Yokohama Triennale 2005 [2]), a famous exhibition of modern art, to navigate users using the extracted social network of artists. As described in this chapter, we use a term *social relation* to present a relation that differs with the cognitive means of relation (e.g., like or dislike) in social sciences, which applies to social actions for people such as collaboration, cooperation, and co-organization.

[2] www.yokohama2005.jp

This chapter is organized as follows. Section 2 describes a basic idea of social network extraction from the Web, and explains some problems in previous approaches. Section 3 gives a detailed explanation for the proposed algorithm and a description of our system flow. Section 4 demonstrates the effectiveness of our approach through experimental results and consideration. Section 5 presents discussion of the position of our work in related works. We conclude this chapter in Section 6.

5.2 Social Network Extraction for Persons

5.2.1 Basic Idea

In human networks, the nodes represent people and edges represent relations among people, which are designated respectively as *actor* and *tie* in social science. Identifying relations among people from the Web means estimating real-world relations among people based on calculating the strengths of social relations on the Web space. Previous studies of social network extraction from the Web have been based on an assumption that the co-occurrence of names on the Web represents the strength of relations among people in the real world. Here, the co-occurrence of names on the Web means names co-occurring on the same Web pages. For example, if two researchers have co-attended many academic communities and conferences, both are members of laboratories, and they are co-authors of papers, then their names might strongly co-occur on the Web. For that reason, we can infer that the social relation between them is strong.

5.2.2 Previous Approach: Objective rule-based Approach

Several co-occurrence indices have been proposed for estimating co-occurrence of names on the Web: the Matching coefficient ($n_{x\cap y}$, Dice coefficient ($2\frac{|x\cap y|}{|x|+|y|}$), Mutual information (log $\frac{N|x\cap y|}{|x||y|}$), and Jaccard coefficient ($\frac{|x\cap y|}{|x\cup y|}$), and Overlap coefficient ($\frac{|x\cap y|}{min(|x|,|y|)}$) [3] [66], where $|x|$,,$|y|$, $|x\cap y|$, and $|x\cup y|$ mean the hit number by putting names, “*x* AND *y*” and“*x* OR *y*”,

[3] No study has ever clarified which co-occurrence indices calculated using a search engine are appropriate for representing relations among people. Therefore, it is noteworthy that social relations we targeted in this chapter are only some of the various relations that can link people.

respectively, as queries.

Referral Web [51] and Flink [73] used the Jaccard coefficient as a co-occurrence measure for identifying relations among academic researchers who attended an international conference. Polyphonet [121, 70] was used to compare several co-occurrence measures, finding that the Overlap coefficient performs well by investigating the probability of co-authorship and researcher community according to their experiments [4]. In addition, NEXAS [117] by Harada uses the *G* score to calculate the relation between given topic words with key persons. Even though all of these studies use different co-occurrence measures, all are based on the same approach to construct social networks. An objective rule-based approach sets a constant threshold for the entire network. Then it adds edges if the relational strength between two people is higher than the objective threshold (Fig. 5.2).

5.2.3 Problem of Previous Approach

A community of academic researchers (of the same research field) usually presents a homogeneous character on the Web. A homogeneous community means, in this chapter, a community that include people of similar attributes, such as similar research field, similar cultures, and similar hobbies. If two researchers frequently attend the same conferences, co-author several papers, and co-organize events, then much information for these social relations might appear on the Web. For that reason, using a single objective criterion to judge the existence of relations based on co-occurrence of names on the Web would enable us to construct a social network of researchers easily.

However, the community for international artists (e.g., participants of artists in international exhibition) presents different characteristics by which the community is apparent as an inhomogeneous community. An inhomogeneous community is, for this discussion, a community that includes people from different fields, different nations, or different cultures, where a relation is difficult to obtain using a single criterion. For example, two Japanese

[4] The Overlap coefficient measures the relational strength between two actors from the perspective of the smaller one, as reflected by the hit number of the actors' names. For example, a student whose name co-occurs almost constantly with that of his supervisor strongly suggests an edge from him to the supervisor. A professor thereby collects edges from her students. Therefore, the relation between a student with his supervisor has different strength for student and for the supervisor, and the Overlap coefficient measures the strength of relation, as seen from the student's perspective.

```
Input: a person name list $L$, and a threshold $T$
Output: a social network $G$

for each $x \in L$
    do set a node in $G$
done
for each $x \in L$ and $y \in L$
    do $r_{x,y} \leftarrow$ GoogleCooc($x$,$y$)
done

/* Invent edges using subjective rule. */
for each $x \in L$ and $y \in L$
    if $r_{x,y} > T$
        do set an edge between $x$ and $y$ in $G$
done
return($G$)
```

* GoogleCooc returns the number of hits retrieved using a given query ("*x* AND *y*") using a search engine (Google).

Figure 5.2: Algorithm of previous method.

artists, "*Taisuke Abe*" (designated as x_1) and "*Jun Oenoki*"(designated as y_1), have no prior relation, but their co-occurrence coefficient is high: $Overlap(x_1, y_1) = \frac{23}{min(113,397)} = 0.2035$, $Jaccard(x_1, y_1) = \frac{23}{960} = 0.024$. Two international artists "*Beat Streuli*" (designated as x_2) from Switzerland and "*Nari Ward*" (designated as y_2) from Jamaica have co-participated in several exhibitions [5], but their co-occurrence coefficient is low: $Overlap(x_2, y_2) = \frac{216}{min(89900,10400)} = 0.0208$, $Jaccard(x_2, y_2) = \frac{216}{175000} = 0.0009$. This happens because the community includes many people from different contexts. For that reason, it is difficult to recognize the relation precisely using a single criterion. This problem also appears when artists are in different

[5] http://www.universes-in-universe.de/car/sharjah/2005/e-artist.htm

fields, or working in different media or genres.

In addition, relations of newly formed artists have few co-occurrence foundations on the Web than those relations of prior formed units. For example, the “ONG Keng Sen” and “Amir Muhammad” first collaborated in a project “Flying Circus” among Yokohama Triennial artists. Only three pages on the Web showed co-occurrence of their names; their co-occurrence strength of the Overlap coefficient and Jaccard coefficient are only 0.005 and 0.0454, respectively. This kind of social relation represents weak relations for an entire community that were overlooked by previous approaches.

Therefore, even though the objective rule-based approach functions well for revealing central nodes and main communities in a network, a salient problem of this approach is that it misses many weak relations in the network, thereby leaving many nodes isolated in the network. Weak relations that connect different nations and different fields might facilitate communications among artists and create new collaborations, which might play an important role for artists. (This kind of relation is also designated as *Weak Ties* in social science.) However based on the previous approach i.e., objective rule-based approach, many weak relations that are lower than a predefined threshold, would be missed by objective criteria. Consequently, for purposes of visualization, navigation, and analyses of an inhomogeneous community, the previous approach is inappropriate.

5.3 Proposed Approach

5.3.1 Subjective rule-based Approach

Inspired by a network questionnaire, we first propose a *subjective rule-based approach* for network construction. By this approach, we collect edges from each viewpoint of nodes even though the co-occurrence value for relations is weak in entire networks. Fig. 5.3 portrays the algorithm. The network was constructed by adding important edges (to M) for each actor. For example, the relation between “*Beat Streuli*” and “*Nari Ward*” is represented as a weak relation in the entire network. Nevertheless, for “*Beat Streuli*”, if no other people show a relation stronger than the relation from “*Nari Ward*”, we might add an edge between them.

```
**Input**: a person name list $L$, and a threshold $M$
**Output**: a social network $G$

for each $x \in L$
    do set a node in $G$
done
for each $x \in L$ and $y \in L$
    do $r_{x,y} \leftarrow$ GoogleCooc($x$,$y$)
done

/* Invent edges using objective rule. */
for each $x \in L$
    do $Y_x \leftarrow$ ConnectedNodes($x$), $\overline{Y}_x \leftarrow L \setminus Y_x$
    while $|Y_x| < M$ and $\overline{Y}_x \neq \phi$
        $y = \operatorname*{argmax}_{y_j \in \overline{Y}_x} r_{x,y_j}$, $\overline{Y}_x \leftarrow \overline{Y}_x \setminus \{y\}$
        do set an edge between $x$ and $y$ in $G$,
            $Y_x \leftarrow Y_x \cup \{y\}$
    done
done
return($G$)
```

* ConnectedNodes returns a node set connected with x; $|X|$ returns the number of elements in a set X.

Figure 5.3: Algorithm of Network Questionnaire.

```
Input: a person name list $L$, and threshold set $< T, M >$
Output: a social network $G$

for each $x \in L$
    do set a node in $G$
done
for each $x \in L$ and $y \in L$
    do $r_{x,y} \leftarrow$ GoogleCooc($x$,$y$)
done

/* First, invent edges using subjective rule.*/ ... (1)
for each $x \in L$ and $y \in L$
    if $r_{x,y} > T$
        do set an edge between $x$ and $y$ in $G$
done

/* Then, invent edges using objective rule.*/ ... (2)
for each $x \in L$
    do $Y_x \leftarrow$ ConnectedNodes($x$), $\overline{Y}_x \leftarrow L \setminus Y_x$
    while $|Y_x| < M$ and $\overline{Y}_x \neq \phi$
        $y = \operatorname*{argmax}_{y_j \in \overline{Y}_y} r_{x,y_j}$, $\overline{Y}_x \leftarrow \overline{Y}_x \setminus \{y\}$
        do set an edge between $x$ and $y$ in $G$,
            $Y_x \leftarrow Y_x \cup \{y\}$
    done
done
return($G$)
```

Figure 5.4: Algorithm of the proposed method.

5.3.2 Objective and Subjective rule-based Approach

However, the subjective rule-based approach is sometimes unreasonable because it treats every actor on the network equally. For example, people (i.e. connector) who hold many connections on the network might not be clarified from the subjective rule-based approach. For a community of academic researchers, it does not fit our intuition that professors and students hold similar connections. The shortcoming of subjective rule-based approach is that it does not incorporate the amount of activity of actors [6]: it treats all nodes as having equal levels of activity.

Herein, we propose a more advanced algorithm that combines ideas from an objective rule-based approach (Fig. 5.2) with a subjective rule-based approach (Fig. 5.3). The proposed algorithm is shown as Fig. 5.4. We employ two criteria corresponding to objective and subjective importance of relations for actors. We first invent edges using objective criteria with a consistent threshold T. Then we invent edges using subjective criteria for actors who have no certain number M of edges. This procedure alleviates the problem of some nodes having too many edges and of some nodes being isolated. The combination of two criteria enables more exhaustive extraction for every node than the previous method, although it sometimes yields low precision. For that reason, we must determine the appropriate parameters so that the target network is extracted as precisely as possible.

5.3.3 Detailed Algorithm in Application

We apply our algorithm to extract social network of artists (of contemporary arts) in Yokohama Triennale 2005. The whole system is illustrated in Fig. 5.5. This system includes online and offline procedures. In the offline procedure, we tune four parameters: T_{ov}, T_{co}, M_1, and M_2. For them, T_{ov} and T_{co} are thresholds to invent edges by the overlap coefficient and matching coefficient; M_1 and M_2 are the minimum quantities of edges for each node [7]. Previous methods have also combined multiple indices to guarantee robustness of mea-

[6] The degree of a node represents a kind of activity of an actor. If an actor holds many connections with others the actor might be an active person, and might maintain height centrality in a network [119].

[7] The matching coefficient directly represents the absolute overlap of two names on the Web in a simple manner. However, a person whose name appears on numerous Web pages will collect many edges. The overlap coefficient is known as the best index for estimating collaboration relations for researchers [70]. However, a

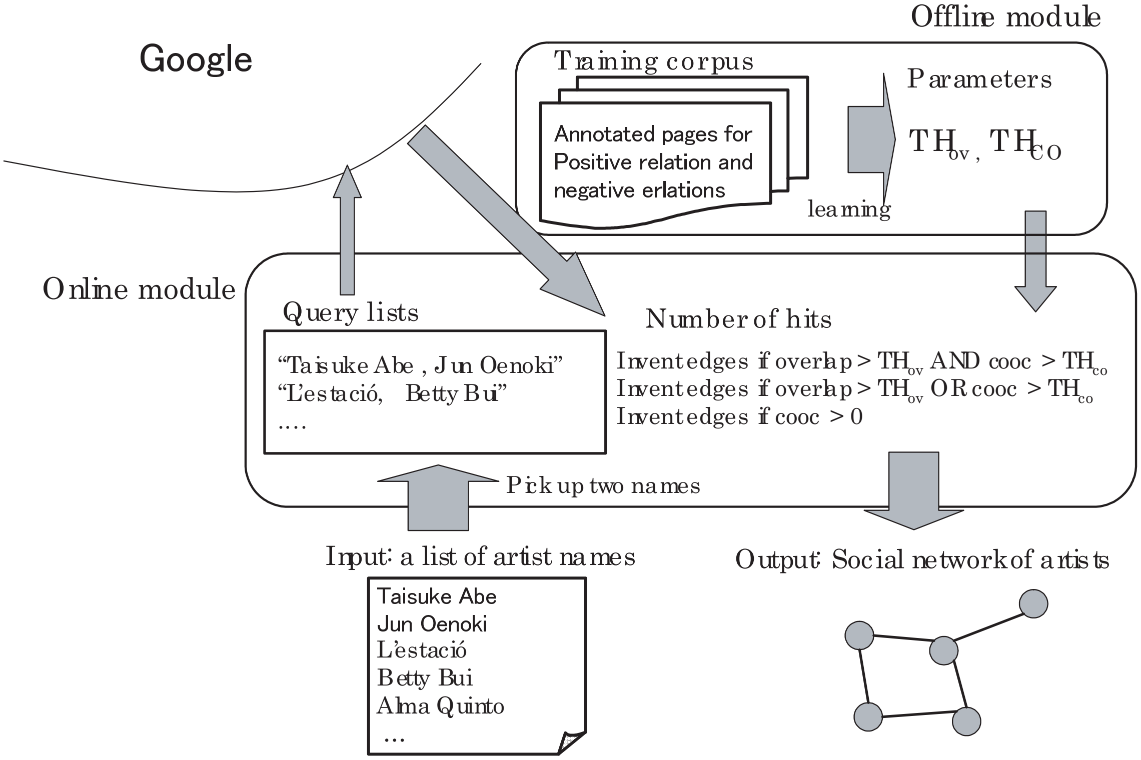


Figure 5.5: System flow to extract an artist network.

surements from a search engine. For example, [121, 70] constrained target researchers as to whose hit-number of names is larger than a threshold; in addition, [73] described only those researchers whose hit-number of names is larger than average.

For the online procedure, a list of names of artists is given as input; the output is a social network of artists. Three steps are used: making queries, Google search, and network construction. First, we make queries for each pair of names. Then we put them into the Google search engine to obtain the hit counts. Finally, we construct a social network after tuning the parameters.

A detailed algorithm to generate a social network is portrayed in Fig. 5.6. Edges are added using an objective criterion (in RULE 1): An edge is added between the nodes if the Overlap coefficient and the Matching coefficient are both over the thresholds. Then subjective criteria are used to add edges (in RULE 2 and RULE 3): We choose nodes that have the strongest relations with node x if node x has less then M_1 edges. Node x is connected to the other nodes until the number of edges reaches M_1 (in RULE 2). After that, if node x has no M_2 edges yet, we add edges in the descending order of overlap coefficient (in RULE 3).

Although the algorithm is highly customized for dealing with Web information, the concept is simple. We use the objective criteria (using T_{ov} and T_{co}) first, and the subjective criteria (using M_1 and M_2) subsequently. It is important to combine multiple criteria to infer the relations among artists correctly from the available Web information. Clearly, if we set $M_1 = 0$, $M_2 = 0$, the algorithm is the same as that used in a previous approach, i.e., an objective rule-based approach. In contrast, if we only consider the subjective criteria M_1 and M_2, the algorithm is identical to a subjective rule-based approach.

5.3.4 System Details

This study only focused on the detection of social relations based on the hit number of the search engine. This procedure was treated as the first step in the Yokohama Triennale system. As a second step, we further identify concrete relation types (labels) from Web pages

person whose name appears on only a few pages will easily create high overlap values for others. Although many other indices for co-occurrence measurement have been reported, in this thesis, we abbreviate the discussion to describe which indices are most appropriate for which relations.

```
Input: a person name list $L$, and threshold set $< T_{ov}, T_{co}, M_1, M_2 >$
Output: a social network $G$

for each $x \in L$
    do set a node in $G$
done
for each $x \in L$ and $y \in L$
    do $r^{ov}_{x,y} \leftarrow$ overlap($x$,$y$), $r^{co}_{x,y} \leftarrow$ cooc($x$,$y$)
done

/* First, invent edges using subjective rule.*/ . . . (1)
for each $x \in L$ and $y \in L$
    if ($r^{ov}_{x,y} > T_{ov}$ AND $r^{co}_{x,y} > T_{co}$) . . . . . . (RULE 1)
        do set an edge between $x$ and $y$ in $G$
done

/* Then, invent edges using objective rule.*/ . . . (2)
for each $x \in L$
    do $Y_x \leftarrow$ ConnectedNodes($x$),
        $\overline{Y}_x \leftarrow L \setminus Y_x$
    while $|Y_x| < M_1$ and $\overline{Y}_x \neq \phi$
        $y \leftarrow \underset{y_j \in \overline{Y}_x}{\text{argmax}}\, r^{ov}_{x,y}$, $\overline{Y}_x \leftarrow \overline{Y}_x \setminus \{y\}$
        if $r^{ov}_{x,y} > T_{ov}$ OR $r^{co}_{x,y} > T_{co}$ . . . . . . (RULE 2)
            do set an edge between $x$ and $y$ in $G$,
                $Y_x \leftarrow Y_x \cup \{y\}$
    done
    $\overline{Y}'_x \leftarrow L \setminus Y_x$
    while $|Y_x| < M_2$ and $\overline{Y}'_x \neq \phi$
        $y \leftarrow \underset{y_k \in \overline{Y}'_x}{\text{argmax}}\, r^{ov}_{x,y}$, $\overline{Y}'_x \leftarrow \overline{Y}'_x \setminus \{y\}$
        if $r^{ov}_{x,y} > 0$ AND $r^{co}_{x,y} > 0$ . . . . . . . . . (RULE 3)
            do set an edge between $x$ and $y$ in $G$,
                $Y_x \leftarrow Y_x \cup \{y\}$
    done
done
return($G$)
```

Figure 5.6: Detailed algorithm used in the Yokohama Triennale 2005.

retrieved by names of artists who connected at the first step; we also filter out noisy edges to improve the system precision. In the system, we considered relations of two types: a *collaboration relation* contains social relations of artists who have collaborated for projects, formed as unit, put together a group, and so on; a *co-attendant relation* includes social relations of artists who have participated in the same exhibitions, same conferences, and same events, and so on. These relations might be used to facilitate the formation of new units, evaluating projects, and holding exhibitions in the future. Details about the second step, i.e., relation type identification, are available from [70]. Here we give only a brief explanation: first, we use several features from Web contents retrieved by two names of artists to create identification rules of relations on the training data. For example, if two names and terms (such as "*form*", "*construct*", "*unit*"), which represent relations, appear in the same sentence, they are judged as a collaboration relation. If two names appearing in the same tables of a Web page and the Web page also contain terms (such as "*exhibition*", "*triennial*", "*participants*") that represent an event, they are judged as co-attendant relations. If no grounds for any rules are satisfied, we treat the pair as having no relation, and delete it from the network.

In our study, we treat the task of relation identification as an external module, and applied it to both the previous approach and the proposed approach. Therefore, in the following evaluation section, we consider only the first step of network construction based on co-occurrence measures.

Additionally, for the Yokohama Triennale network, the nodes can be represented as an artwork of "project". To construct a project network from the artist network, we use the collaboration relations in projects (Fig. 5.7). For example, a member of a project "*Sadaharu Horio*" and a member of project b "*Tomoko Yoneda*" have a co-attendant relation. Therefore, we set an edge between the former project *a* with the latter project *b* as a co-attendant relation.

Fig. 5.8 exhibits the constructed networks (for 132 artists and 71 projects): (a) is the first step of the network; (b) depicts a network that has been improved by identifying relations in the second step. Our system was put into operation on the official support site for Yokohama Triennale 2005 (http://mknet.polypho.net/tricosup/) to provide an overview of the artists (133 artists with 71 projects) along with informational navigation for users. At the exhibitions, it is usual for participants to enjoy and evaluate each work separately. However, our supposition was that if participants knew the background and relations of the artists, they

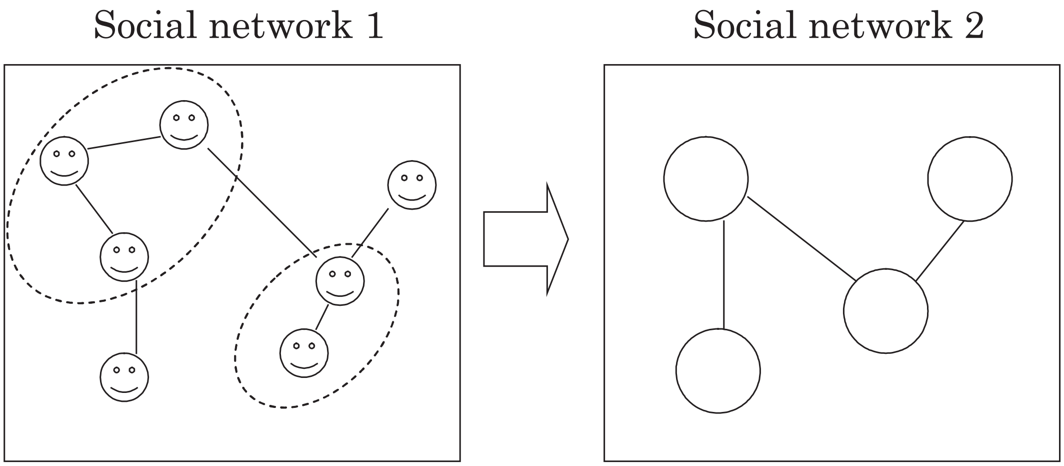


Figure 5.7: Identification of the relations among projects using actors' relations.

might enjoy the event more. For that purpose, the system provided relations of artists and evidential Web pages for users.

The system interface is depicted in Fig. 5.9. It was implemented using Flash display software to facilitate interactive navigation. The system provides a retrieval function. Information about the artist is shown on the left side if a user clicks a node. In addition, the edges from the nodes are highlighted in the right-side network. The user can proceed to view the neighboring artists' information sequentially, and can also jump to the Web pages that show evidence of the relation.

5.4 Evaluation Results

We make two datasets to compare the performance of previous approach with the proposed approach to evaluate the effectiveness of our approach. First we sample 1000 pairs of artists who participated in Yokohama Triennale 2005. Then we compared social networks constructed by previous and proposed approaches. We further investigate characteristics of parameters for networks. Then, we target 50 academic researchers who participated in JSAI 2006 to estimate the applicability and robustness of the proposed approach to a different community.

5.4.1 Evaluation on Social Network of Artists

We sampled 1000 pairs from 132 artists (of ${}_{132}C_2$ pairs) randomly; we quickly checked the relations among these pairs on the Web. We first put each pair as a query to a search engine and collected the top pages as contents of relations. Then we used these contents to judge the existence of relations. For example, if two artists form a group, circus, partner, or unit, then they are judged as having a "*collaboration relation*". If they attended the same exhibition or event, then they were judged as having a "*co-attendant relation*". In addition, because we targeted artists who participated in Yokohama Triennale 2005, the co-attendant relation in this exhibition is a trivial solution. We ignore those relations. To judge the relations from the contents, we used following rules. First, if the Matching coefficient is equal to zero, i.e., if there no contents mention two artists, then they are inferred to have no relation. In addition, cases were judged also as having no relation if they had a Matching coefficient greater than zero, but no information describing any relation between them in the contents. One reason the two names have no relation but co-occurred on the Web pages is the problem of name entity disambiguation. The names such as "*grat*", "*SOI*", and "*Open Circle*" often appeared on the Web, but not only as names of artists. We also checked the collected Web contents by putting each name as a query to consider the recall of relations for each actor. Results show that, among the 1000 pairs or artists, 146 pairs have relations and 854 pairs have no relations. We use these as evaluation data.

For evaluation, we use 132 artists as input, and change four parameters—T_{ov}, T_{co}, M_1, M_2— to construct different networks. We change the values of parameters, classify every pair of artists as positive or negative using the parameters, and find the optimal values at which the F-value is maximized. We change T_{ov} from 0 to 1 by 0.01, T_{co} from 0 to 60 by 5, M_1 from 0 to 10 by 1, and M_2 from 0 to M_1 by 1 [8]. For each network constructed by parameter set $< T_{ov}, T_{co}, M_1, M_2 >$, we evaluate the precision, recall, and F-value among 1000 pairs. The relation strength is calculated using the Overlap coefficient (T_{ov} as threshold), and the constraint of the hit number is based on the Matching coefficient (T_{co} as threshold). We use a

[8] We might use more sophisticated algorithms such as hill-climbing searches. However, we do not examine the optimization method specifically for this chapter. We employed a simple but reliable approach.

general search engine Google [9][10]. Networks constructed using the previous approach can be considered as using only two parameters $< T_{ov}, T_{co} >$. The F-value is defined as follows.

$$Fvalue = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{5.1}$$

Table 5.1 shows the precision, recall, and F-value of the previous approach (i.e., objective rule-based approach) with their parameters T_{ov}, T_{co}. The maximum recall is 100% if we set $T_{ov} = 0$, $T_{co} = 0$. However, the precision in this case is only 14.6%. Conversely, if we only choose highly co-occurred pairs to improve the precision (set $T_{ov} = 0.24$, $T_{co} = 30$), the recall is only 26.7%. The most balanced parameter is $T_{ov} = 0.05$, $T_{co} = 20$, which produces highest F-value as 0.50.

Table 5.2 presents the results obtained using the proposed approach (i.e., objective and subjective rule-based approach). There are four parameters: T_{ov}, T_{co}, M_1, M_2. Even though we set T_{ov} and T_{co} as the same values in cases of table 5.1, if we select appropriate value of M_1 and M_2, we can improve the F-value as 0.55. Numbers in brackets show the number of edges added by RULE 1, RULE 2, and RULE 3, respectively. We can cover edges missed in RULE 1 using the objective criterion M_1. For instance, "*Jun Oenoki*" and "*MIKAN*" collaborated for artwork. Their Overlap and Matching coefficients are, respectively, 0.162 and 162. These relations are missed by RULE 1 because the Overlap coefficient is low, but they can be covered by RULE 2. In addition, some relations missed by RULE 1 and RULE 2 can be covered further by RULE 3. For instance, "*ONG Keng Sen*" and "*Amir Muhammad*" first collaborated in this exhibition. Therefore, only a little information described for this relation is available on the Web: the Overlap and Matching coefficients are, respectively, 0.005 and 3. However, for them no other relations with other participants are strengthened aside from this collaboration; this relation will be covered by RULE 3. As the results showed, we can understand that networks constructed using the previous approach (Fig. 5.10(a)) retain many isolated nodes, while the proposed approach (Fig. 5.10(b)) connects all possible edges for nodes.

Next, we demonstrate the robustness of performance while varying the parameters. Fig. 5.11(a) and Fig. 5.11(b) portray the changes of precision, recall, and F-value in the previ-

[9] www.google.co.jp

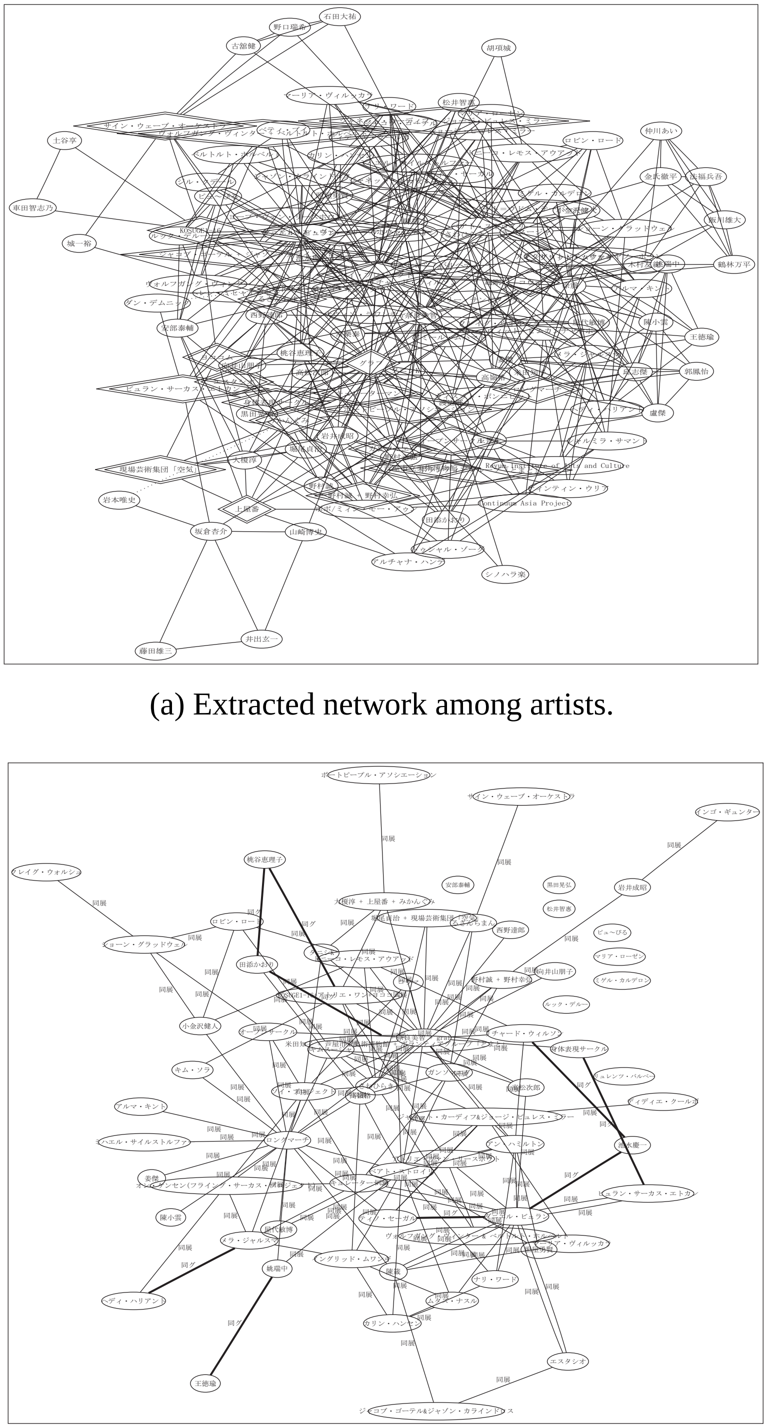
[10] It is noticeable that the search result from search engine is changed dynamically by time. Therefore, the results of our algorithm are slightly affected by the search engine results.

ous approach as well as the proposed approach while varying parameter T_{ov}. Fig. 5.11(a) presents the problem described in Section 2: many weak relations will be overlooked when we improve the threshold T_{ov}; conversely many incorrect edges will be created when we decrease the T_{ov} in the previous approach. On the other hand, by adding parameters $M_1 = 5$, $M_2 = 1$ by the proposed approach, Fig. 5.11(b) depicts stable changes in T_{ov} while retaining high recall.

Fig. 5.12 and Fig. 5.13 portray changes of (a) precision, (b) recall, and (c) F-value with variations of parameters T_{ov} and T_{co} respectively in the previous approach. The performances are changed rapidly by varying parameters in a previous approach (Fig. 5.12) while retaining stable changes in the proposed approach (Fig. 5.13).

Figures 5.14 and 5.15 depict changes of (a) precision, (b) recall, and (c) F-value that occur concomitantly with variations of parameters M_1 and M_2. The recall rises while the precision is falling when increasing M_1 and M_2. From Fig. 5.14, we can understand that the F-value is changed by different parameters, especially by M_1 and T_{ov}. Consequently, using Overlap coefficient to set the objective threshold for the entire community as well as combine subjective criteria for each node to add edges would make the network most appropriate.

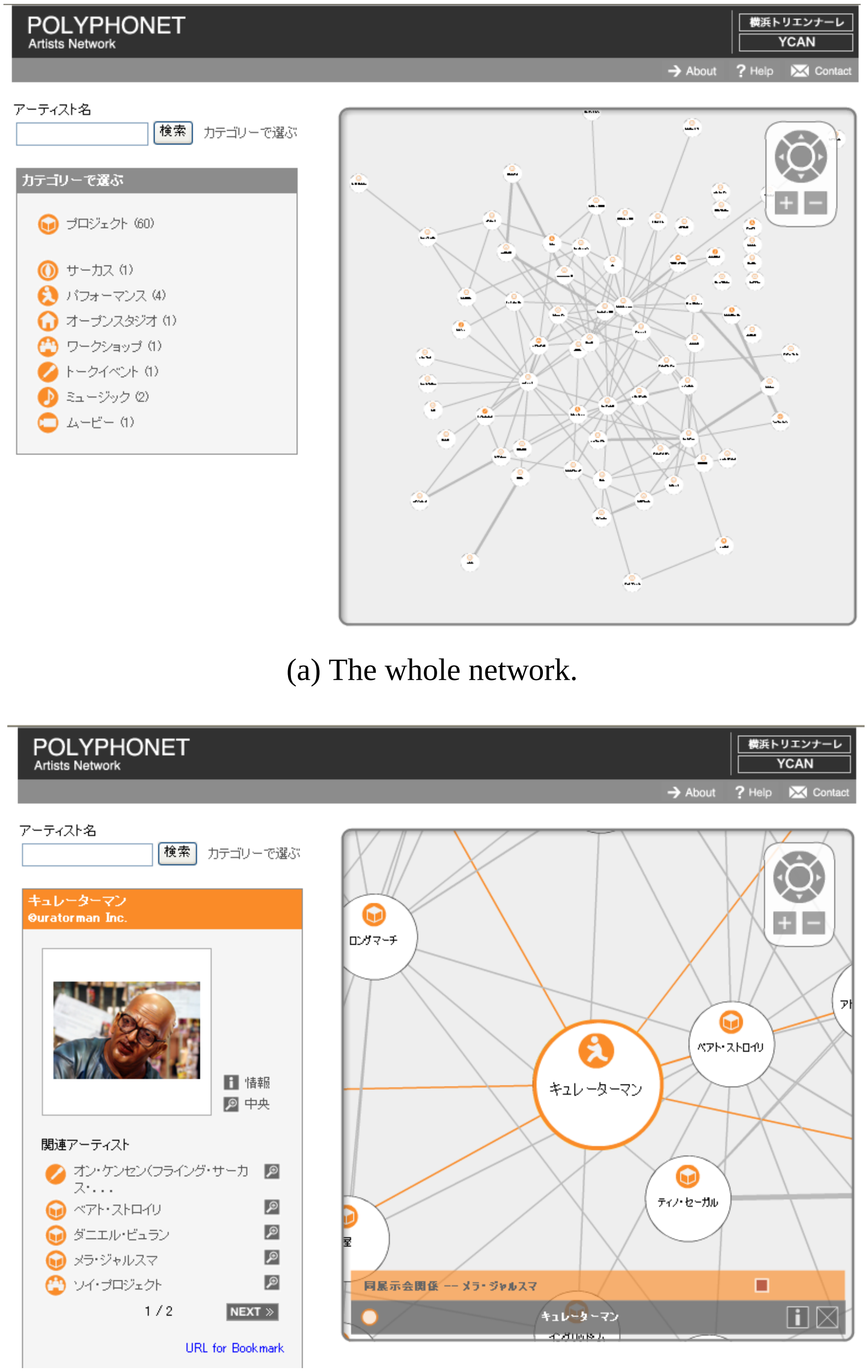
All results described above indicate that the proposed approach is superior to the previous approach. The proposed approach requires more parameters to improve quality, but many studies' objective and subjective criteria of network extraction have unintentionally mixed their extraction processes. Our study treats those procedures as selection parameters and demonstrates the effectiveness of the parameters. These results are expected to inspire various studies of social network extraction from various sources.



(a) Extracted network among artists.

(b) Improved network by identifying relations.

Figure 5.8: Yokohama Triennale 2005 artist network.



(a) The whole network.

(b) Centering artist *Curatorman*.

Figure 5.9: System Interface.

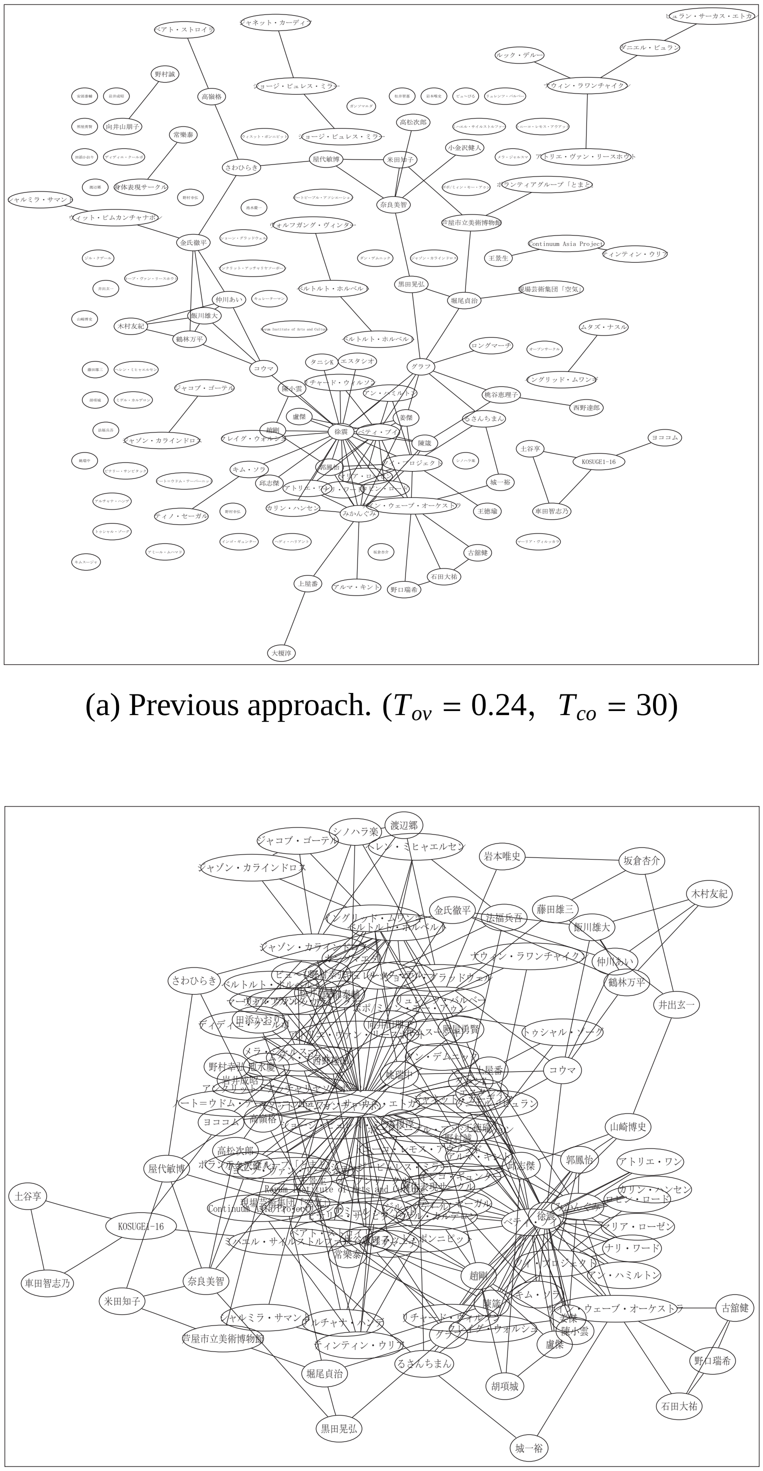
Table 5.1: Precision, recall, and F-value with parameters in the previous approach.

Cases	T_{ov}	T_{co}	P	R	F	#Extracted*	#Correct*
(a): Maximum Precision	0.24	30	92.9%	26.7%	0.41	42 (42,0,0)	39 (39,0,0)
(b): Maximum Recall	0	0	14.6%	100%	0.25	1000 (1000,0,0)	146 (146,0,0)
(c): Maximum F-value	0.05	20	76.4%	37.7%	0.50	72 (72,0,0)	55 (55,0,0)

*: Numbers in brackets are quantities of edges invented in $RULE1$, $RULE2$, and $RULE3$.

Table 5.2: Precision, recall, and F-value with parameters in the proposed approach.

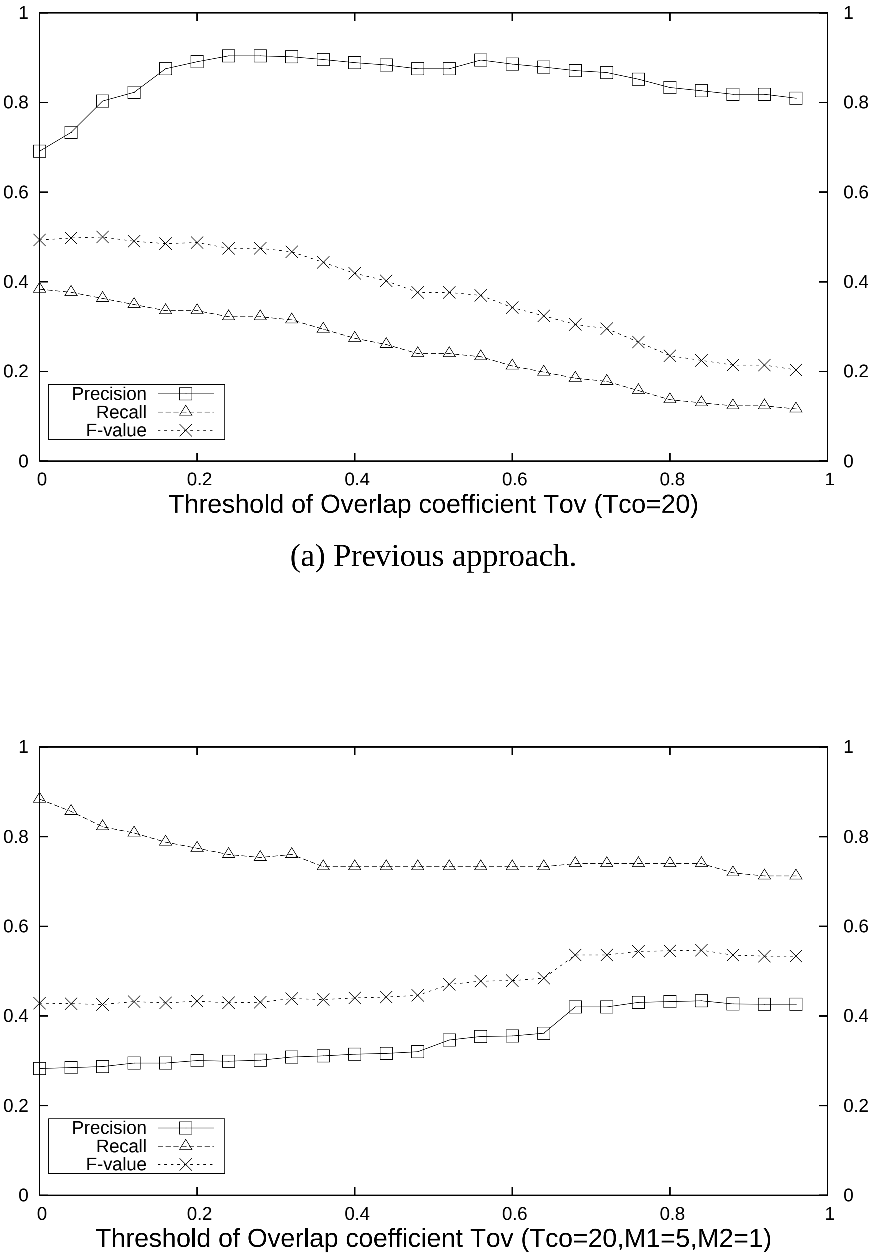
Cases	T_{ov}	T_{co}	M_1	M_2	P	R	F	#Extracted	#Correct
Case (a’)	0.24	30	3	2	34.4%	65.1%	0.45	277 (42,227,8)	95 (39,54,2)
Case (b’)	0	0	0	0	14.6%	100%	0.25	1000 (1000,0,0)	146 (146,0,0)
Case (c’)	0.05	20	1	0	55.4%	49.3%	0.52	130 (72,58,0)	72 (55,17,0)
(d’): Maximum F-value	0.82	20	5	1	43.4%	74.0%	0.55	249 (23,212,14)	108 (19,84,5)



(a) Previous approach. ($T_{ov} = 0.24, \;\; T_{co} = 30$)

(b) Proposed approach. ($T_{ov} = 0.24, \;\; T_{co} = 30, \;\; M_1 = 3, \;\; M_2 = 2$)

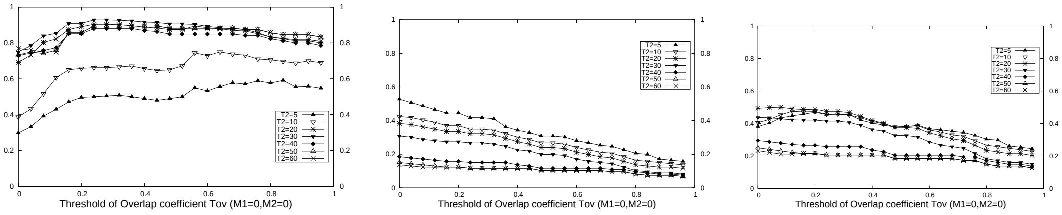
Figure 5.10: Difference of extracted networks.



(a) Previous approach.

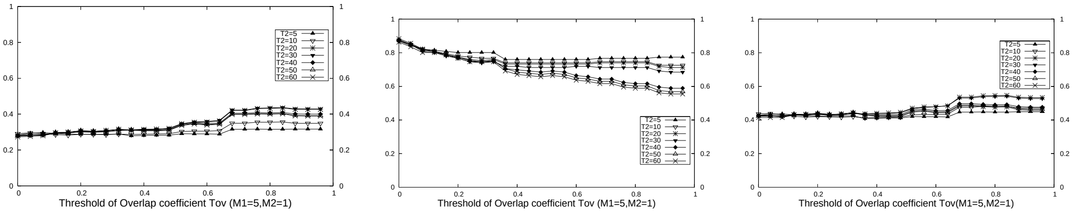
(b) Proposed approach.

Figure 5.11: T_{ov} vs. precision, recall, and F-value.



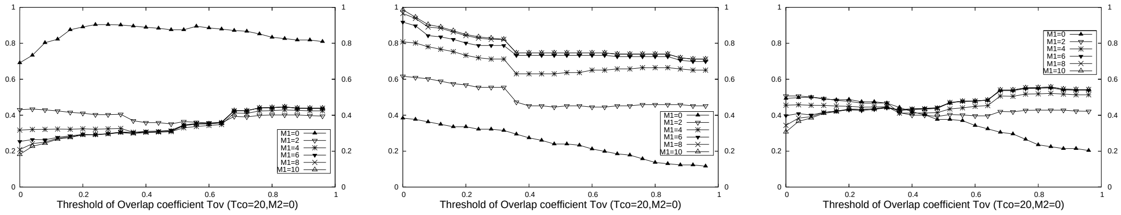
(a) Precision (b) Recall (c) *F*-value

Figure 5.12: T_{ov} and T_{co} vs. performance in the previous approach.



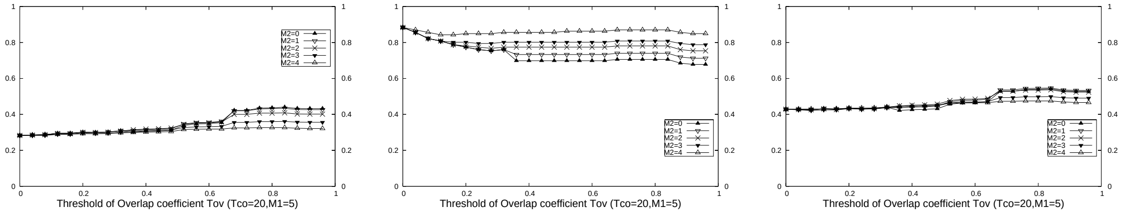
(a) Precision (b) Recall (c) *F*-value

Figure 5.13: T_{ov} and T_{co} vs. performance in the proposed approach.



(a) Precision (b) Recall (c) *F*-value

Figure 5.14: M_1 vs. performance in the proposed approach.



(a) Precision (b) Recall (c) *F*-value

Figure 5.15: M_2 vs. performance in the proposed approach.

Table 5.3: Precision, recall, and F-value with parameters in the previous approach.

Cases	T_{ov}	T_{co}	P	R	F	#Extracted*	#Correct*
(a): Maximum Precision	0.6	5	100%	5.98%	0.11	14 (14,0,0)	14 (14,0,0)
(b): Maximum Recall	0	0	19.1%	100%	0.32	1225 (1225,0,0)	234 (234,0,0)
(c): Maximum F-value	0.2	0	87.4%	47.4%	0.62	127 (127,0,0)	111 (111,0,0)

*: Numbers in brackets are quantities of edges invented in $RULE1$, $RULE2$, and $RULE3$.

Table 5.4: Precision, recall, and F-value with parameters in the proposed approach.

Cases	T_{ov}	T_{co}	M_1	M_2	P	R	F	#Extracted	#Corrct
Case (a’)	0.6	5	3	3	31.7%	59.4%	0.41	438 (14,422,2)	139 (14,124,1)
Case (b’)	0	0	0	0	19.1%	100%	0.32	1225 (1225,0,0)	234 (234,0,0)
Case (c’)	0.2	0	0	0	87.4%	47.4%	0.62	127 (127,0,0)	111 (111,0,0)
(d’): Maximum F-value	0.2	20	7	0	68.0%	71.8%	0.70	247 (30,217,0)	168 (26,142,0)

Table 5.5: Precision, recall, and F-value in the testing data with parameters that produced maximum F-values in the learning data.

(a) previous approach

T_{ov}	T_{co}	F^L_{max}	P	R	F^T
0.18	0	0.66	84.6%	30.6%	0.45
0.8	0	0.66	100%	36.4%	0.53
0.12	0	0.60	60.0%	60.0%	0.60

(b) proposed approach

T_{ov}	T_{co}	M_1	M_2	F^L_{max}	P	R	F^T
0.18	20	5	0	0.75	64.9%	66.7%	0.66
0.2	20	5	0	0.71	72.7%	72.7%	0.73
0.18	20	3	0	0.69	71.1%	67.5%	0.69

5.4.2 Evaluation on Social Network of Academic Researchers

In this section, to examine the generality of our algorithm, we apply it to an academic researcher community—50 researchers (${}_{50}C_2 = 1225$ pairs) who participated in JSAI2006—to construct a social network of researchers as well as tuning parameters for this community. The correct relations among researchers (e.g., co-authored, co-member of laboratory, co-project, co-presentation relations) are gleaned from a questionnaire.

We use these 50 researchers as input and tuning for four parameters. We evaluate the constructed networks of precision, recall, and F-value using different parameters. Table 5.3 shows the maximum F-value with the optimal value of parameters in a previous approach. We can see that an optimal value of parameter $< 0.2, 0 >$ produces the maximum F-value. This means that the Overlap coefficient performs well in analyses of the research community. When we use the same parameter $< T_{ov}, T_{co} >$ as that used in the proposed approach, the result of the F-value is equal to that obtained using the previous approach. However, if we set the parameter as $< 0.2, 20, 7, 0 >$ in our approach, the constructed network outperforms any network constructed using the previous approach.

Table 5.6: Precision, recall, and F-value in the subjective rule.

M	P	R	F	#Extracted	#Correct
1	65.1%	12.0%	0.20	43	28
2	59.0%	19.7%	0.29	78	46
3	57.8%	28.6%	0.38	116	67
4	56.8%	37.6%	0.45	155	88
5	56.1%	47.4%	0.51	198	111
6	**53.8%**	**54.7%**	**0.54**	**238**	**128**
7	49.1%	58.1%	0.53	277	136
8	45.7%	61.5%	0.52	315	144
9	43.1%	64.1%	0.52	348	150
10	41.3%	67.5%	0.51	383	158

For indicating the robustness of function of parameters, we use 40 researcher names as a training dataset to tune parameters; we then use the obtained parameters to evaluate the remaining 10 researchers. We iterate this process three times and take the mean. Table 5.5 presents the precision, recall, and F-value in testing data (designated as F^T) with optimal parameters which produce the maximum F-value in training data (F^L_{max}): (a) shows results of a previous approach that uses only objective criteria; and (b) shows results of proposed approach that uses both objective and subjective criteria. It is apparent that parameters from training data serve stable functions in testing data. In addition, in every case, the proposed approach outperforms the previous approach.

We further compare the proposed approaches: subjective rule-based approaches (Fig. 5.3) with objective and subjective rule-based approaches. Table 5.6 presents results obtained using the subjective rule-based approach. We add M edges (sorted by the Overlap coefficient) for each node. Then we evaluate the precision, recall, and F-value of each constructed network. When creating M, the most appropriate network produces an F-value as 0.54 when M equals 6. No case in the subjective rule-based approach outperforms the objective and subjective rule-based approach because the subjective rule-based approach ignores key persons who actually hold many connections (much greater than M).

5.5 Discussion

The proper values of parameters used by network construction vary according to the target community characteristics. To set the most appropriate values of parameters for constructing networks, we need some training data. No reported general approach describes how to prepare training data. At the Yokohama Triennale system, we initially used artists involved in the same projects published on the homepage as training data to define the parameters. We can further prepare an interface that enables users' input for relations; thereby, we can modify the parameters automatically.

Previous studies of social network extraction from the Web have been based on the assumption that the target community is homogeneous. Based on this assumption, they only addressed strong relations from the entire community. It becomes possible to find core members as well as a connector with many ties in a community. This kind of relation is called *Strong Ties* in social sciences fields. However, actors connected with strong ties are mutually similar; the coverage of information is narrow. In contrast, *Weak Ties*, as they are known in social sciences fields, connect different fields and different communities, serving as a bridge for information transmission and social integration [44]. For supporting navigation and communication for participants of artists in international exhibition, it is important to discern relations between artists in different fields and different nations. Our proposed approach is inspired by network questionnaire in social science, which is not only considered subjective perspective but also considered objective perspective. This algorithm is effective for extracting various social networks for humans with different domains.

Although our algorithm appears to be simple, it can be considered from the perspective of the most important question: "*what is a network*". Social network analysis in sociology explains phenomena from relations among actors, not from attributes of actors. Relations have two different meanings: relational events such as e-mail changes and telephone conversations, and the importance from each actor according to who is most important to whom. Social networks do not exist but instead have the appearance of a social phenomenon. Our research considers these essential problems of social networks.

5.6 Conclusion

As described in this chapter, we proposed an advanced algorithm for network extraction that can extract weak relations among contemporary artists. Experimental results demonstrate the effectiveness of our approach. We will further discuss characteristics of parameters in future reports. The obtained network for artists was operated on the Web site for the Yokohama Triennale 2005. Future studies will discern appropriate parameters for different networks. Additionally, we will continue to consider the relation identification module to improve the precision of the entire system.

Chapter 6

General Model of Social Network Extraction

6.1 General Model of Social Network Extraction

Based on the two case studies described in preceding chapters, this chapter presents and explains an architecture to support general social network extraction from the Web using a search engine. The types of social networks depend on their purpose [91]. A "good" social network is expected to represent a target domain most appropriately.

We consider that the model of social network extraction (Fig. 6.1) is generally written as

$$f(\mathbb{S}_r(X, Y), \Theta) \to \{0, 1\} \tag{6.1}$$

where $\mathbb{S}_r(X, Y)$ is an m-dimensional vector space $(S_r^{(1)}(X, Y), S_r^{(2)}(X, Y), \ldots, S_r^{(m)}(X, Y))$ to represent various measures for X and Y in relation r. For example, $S_r^{(i)}(X, Y)$ can be either a $n_{X \cap Y}$ (matching coefficient), a $n_{X \cap Y}/n_{X \cup Y}$ (Jaccard coefficient), or a $n_{X \cap Y}/min(n_X, n_Y)$ (overlap coefficient). It can be a score function based on sentences mentioning both mentions of X and Y (similarly to the algorithm presented in chapter 4). The parameter Θ is an n-dimensional vector space $(\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)})$. For example, Θ can be characterized as a combination of T_{ov}, T_{co}, M_1, and M_2, as presented for the algorithm in chapter 5. The function f determines whether an edge should be invented or not based on multiple measures and parameters.

A social network is expected to represent the particular relations of entities depending on purposes. Therefore, function f need not always be the same. A method to infer an

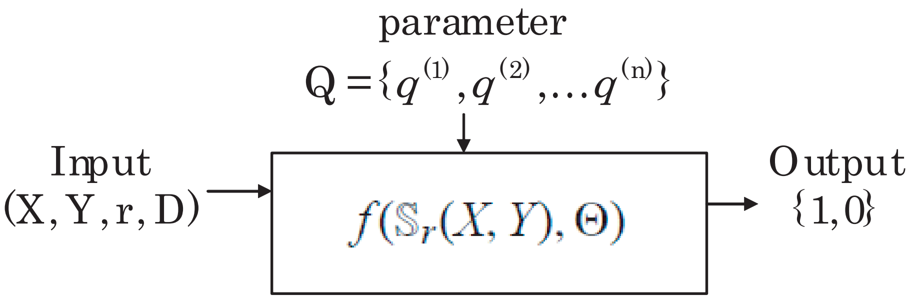

parameter
$Q=\{q^{(1)},q^{(2)},\ldots q^{(n)}\}$
Input
(X, Y, r, D)
$f(\mathbb{S}_r(X,Y),\Theta)$
Output
{1,0}

Figure 6.1: General model of social network extraction.

appropriate function f is necessary; thereby the algorithm invariably consists of an offline module and an online module. Function f is learned from the training examples and provides good classification to other examples.

6.2 General Procedures

In the online phase, it is important to extract a social network from the Web in an efficient manner. We must consider how to use a search engine better and how to process Web documents efficiently and correctly. Generally, the procedure consists of three steps:

Making queries Two entities are used to generate a query. Basically, we put a query X AND Y to a search engine. As described in this paper, we add relation keywords to extract a particular type of relation efficiently. A combination of multiple queries might improve the result, as explained in chapter 4. Entity disambiguation is another important issue that has already been addressed in several studies [8, 12].

Google search We put the queries into a search engine. Sometimes the counts are used to infer relational strength. In other cases, we download some documents (or snippets) and investigate the mentions of X and Y. A good combination of Google counts and text analysis would make the search more efficient and scalable, as discussed in [70].

Network construction We use Google counts and downloaded text as evidence to construct a social network. The value of function f is calculated and the existence of an edge is determined. Usually, the obtained social network is visualized and reviewed. Sometimes we must change the settings of the algorithm (or increase the training data) and repeat the entire process to improve the quality.

Previous studies have emphasized the study of how to calculate the strength of two names on the Web in the **Google search** step, simply using X AND Y as a query and constructing networks based on objective criteria. The method presented herein, which includes *relation identification* and *threshold tuning* is proposed for **Making queries** and **Network construction** steps, respectively, for complex and inhomogeneous communities. All of these methods are combined into our architecture of general extraction of social networks for various entities.

The obtained network is useful for Semantic Web studies in several ways. For example (inspired by [7]), we can use a social network of artists for detecting COI among artists when they make evaluations and comments related to others' work. We might find a cluster of companies and characterize a company by its cluster. Business experts often make such inferences based on company relations and company groups. Consequently, the company network might enhance inferential abilities in the business domain. As a related work, Gandon et al. built a Semantic Web server that maintains annotations about the industrial organization of Telecom Valley to partnerships and collaboration [37]. We presented a prototypical example of applications using a social network of companies in chapter 4. We calculated the *centrality*, which is a measure of the structural importance of a node in the network, for each company on the extracted social network (on alliance relations). Such information can only be inferred after extracting a social network. There seem to be many potential applications that can make use of social networks in the Semantic Web.

6.3 Conclusion

This chapter presented a description of methods of extracting various social networks from the Web. To date, numerous studies have addressed the researcher domain to estimate extraction methods. It is an important test-bed. Nevertheless, the next step must be taken to depart from the domain of researchers. The proposed architecture toward general extraction of social networks, which bundles these different extraction methods, will enable us to extract various social networks from available information related to the Web. In addition to some direct applications of social networks, we believe that a network perspective is important for knowledge integration and articulation and for (lightweight) ontology emergence. The combination of social networks and ontology emergence might prepare a fertile ground for Semantic Web research.

Part II

Application of Social Networks

Abstract

Many rankings existing for popularity, recommendation, evaluation, election, etc. can be found in the real world as well as on the Web. Many efforts are undertaken by people and companies to improve their popularity, growth, and power, the outcomes of which are all expressed as rankings (designated as *target rankings*). Are these rankings merely the results of its elements' own attributes? In the theory of social network analysis (SNA), the performance and power (i.e. ranking) of actors are usually interpreted as relations and the relational structures they embedded. For example, if we seek to rank companies by market value, we can extract the social network of the company from the Web and discern, and then subsequently learn, a ranking model based on the social network. Consequently, we can predict the ranking of a new company by mining its relations to other companies. We can learn from existing rankings to expect other rankings. Furthermore, we can understand the kinds of relations which are important for the target rankings, we can determine the type of structural extension of companies that can improve the target rankings. Part 2 specifically examines the application of a social network that provides an example of advanced utilization of social networks mined from the Web. In chapter 7, we present ranking learning approaches using a social network that is mined from the Web based on the approaches described in Part 1. The proposed model combines social network mining and ranking learning, which further uses multiple relations on the Web to explain arbitrary rankings in the real world.

Chapter 7

Ranking Entities Based on the Social Networks

7.1 Introduction

People prefer to use rankings for comparing companies, discussing elections, and evaluating goods. For example, job seekers like to apply for employment at popular and high-paying companies; investors seek to invest funds in fast-growing and stable companies; consumers tend to buy highly popular products. Therefore, many efforts are undertaken by companies to improve their popularity, growth, and power, the outcomes of which are all expressed as rankings. Conventionally, these rankings are evaluated and ranked by values from statistical data and attributes of actors such as income, education, personality, and social status. The following are noteworthy examples. The Fortune-1000 lists the 1000 largest American companies ranked by revenues alone. Popular companies are ranked by the number of applications from job-seekers. Goods are ranked by their unit sales.

In the theory of social network analysis (SNA), social networks are used to analyze the performance and valuation of social actors [107, 106, 99, 93]. Network researchers have argued that relational and structural embeddedness influence individual's behavior and performance, and that a successful company must therefore emphasize relation management. For studies in company networks as example, Bernstein et al. [14] construct company networks from business news stories and presented an interesting result that more than 50% of the top 30 “most central ”technology companies are Fortune-1000. Especially for analyzing companies in terms of relational construction, various relations are targeted: Rowley et al. [99] use strategic alliance networks to analyze such embeddedness of companies; Bengtsson et al. [13] analyze cooperation and competition in relations among companies in business networks; Souma et al. and Battiston et al. analyze structural features of shareholding networks. They then use those results to explain features of companies’ growth [95] and market structure characteristics [10], respectively. Multiple relations clearly exist in the world with different impacts; the companies might be tied together closely in one relational network, but can differ greatly from one to another in a different relational network. The question arises: *relations of what kind are important for actors?* Unfortunately, the answers of important relations used by analysis have been decided according to the judgment of researchers themselves.

To identify the prominence or importance of an individual actor embedded in a network (i.e., ranking network entities), centrality measures have been used in social sciences: degree centrality, betweenness centrality, and closeness centrality. These measures often engender distinct results with different perspectives of “actor location ” i.e., local (e.g. degree) and global (e.g. eigenvector) locations, in a social network [107]. On the other hand, a ranking network entities is an important topic in link mining [38]. Given a network among entities, the goal is to find a good ranking function to calculate the ranking of each entity using the relational structure. PageRank [84] and HITS [54] algorithms can be considered as famous examples for ranking in the context of information retrieval, i.e., to rank Web pages based on the link structure. Another question arises: *what kind of centrality indices are most appropriate for ranking actors?* That question can be extended as *what kind of structural embeddedness of actors makes them more powerful?* Although quite a few studies of learning-to-rank fields (particularly targeted on information retrieval) have investigated many attribute-based ranking functions learned from given preference orders [24, 30, 5, 86], only a few studies have concluded that such an impact arises from relations and structures [86, 5].

This chapter presents a description of an attempt to learn the ranking of named entities from a social network that has been mined from the Web. It enables us to have a model to rank entities for various purposes: one might wish to rank entities for search and recommendation, or might want to have the ranking model for prediction. Given a list of entities, we first extract different types of relations from the Web based on our previous work [70, 47].

Subsequently, we rank the entities on these networks using different network indices. In this thesis, we propose three approaches: The first approach is based on an intuitive idea: based on the correlation between rankings from different networks (designated as *network rankings*) with target ranking, simply choose the most predictive type of relation along with centrality indices. For the second approach, we combine multiple relations into one network (designated as the *combined-relational network*) to learn a ranking model. The third approach is more systematic: we integrate features generated from networks for each and then use these features to learn and predict rankings. We designate features generated from network as the *network-based features*. The important characteristic of our model is *target-dependent*, which suggests that the important relations and advantages of structural embeddedness on a network differ according to target rankings. We conducted two experiments: related to social networks among 312 companies of the electronics industry in Japan to discern the target rankings of market capitalization, average income, and excellent ranking; related to social networks among researchers to learn and predict the ranking of researchers' productivity.

Several findings including social networks vary according to different relational indices or types even though they contain the same list of entities. Relations and networks of different types differently impact on target of ranking. Multiple networks have more information than single networks for explaining target ranking. Well-chosen attribute-based features have good performance for explaining the target ranking. However, by combining proposed network-based features, the prediction results are further improved.

The contributions of this study can be summarized as follows. We provide an example of advanced utilization of a social network mined from the Web. The results illustrate the usefulness of our approach, by which we can understand the important relations as well as the important structural embeddedness to predict features of entities. Multi-relational networks are extracted from the Web and are then used. They are more realistic than single-relational networks. The proposed ranking learning model combines various network features. The model can be combined with a conventional attribute-based approach. Results of this study will provide a bridge between relation extraction and ranking learning for advanced knowledge acquisition for Web intelligence.

The following section presents a description of an overview of the ranking learning model. Section 7.3 introduces our previous work for extracting social networks from the

Web. Section 7.4 describes ranking learning models based on extracted social networks. Section 7.5 describes experimental results on a case study of learning to rank companies. Section 7.6 describes another experimental results on a case study of learning to rank researchers. Section 7.7 presents some related works before the chapter concludes.

7.2 System Overview

Our study explores the integration of mining relations (and structures) among entities and the learning ranking of entities. For that reason, we first extract relations and then determine a model based on those relations. Our reasoning is that important relations can be recognized only when we define some tasks. These tasks include ranking or scores for entities, i.e., *target ranking* such as ranking of companies for job-seekers, CD sales, popular blogs, and sales of products.

Our study is motivated by our desire to infer various relations among entities from the Web. However, what we are often interested in is not the relation itself, but a combination of relations (e.g., finding a path), or the aggregated impact of the relations on each entity (e.g. network structure of the entity) [106, 112, 23]. If we can identify a type of relation or a typed network that is influential to some attributes of each entity, we can understand that the types of relation as well as the type of structural embeddedness are important, and that it would be possible to execute an analysis using the extracted network. For example, two companies might have shareholding relations, alliance relations, lawsuit relations, neighboring offices, the same field of business, and so on. Although many relations exist, why do many methods described in the literature use shareholding relations [95] or alliance relations [99] to assess a company's influence? The readily available answer is that such relations contribute to an analytical task: this intuition implicitly or explicitly exists in our lives. In short, our approach consists of two steps;

Step 1: Constructing Social Networks Given a list of entities with a target ranking, extract a set of social networks among these entities from the Web based on approaches introduced into Part 1.

Step 2: Ranking learning Learn a ranking model based on the relations and structural features generated from the network.

Once we obtain a ranking model, we use it for prediction for unknown entities. Additionally, we can obtain the weights for each relation type as well as relation structure, which can be considered as important for target rankings. The social network can be visualized by specifically examining its relations if the important relations are identified. Alternatively, social network analysis can be executed based on the relations.

7.3 Constructing Social Networks from the Web

In this step, our task is, given a list of entities $V = \{v_1, \ldots, v_n\}$, construct a set of social networks $G_i(V, E_i)$, $i \in \{1, \ldots, m\}$, where m signifies the number of relations, and $E_i = \{e_i(v_x, v_y)|v_x \in V, v_y \in V, v_x \neq v_y\}$ denotes a set of edges with respect to the *i-th* relation, and where $e_i(v_x, v_y)$ is equal to 1 if entities v_x and v_x have relation i; it is 0 otherwise. In this paper, we are interested only in undirected networks.

A social network is obtainable through various approaches; one is to use Semantic Web data. With developments in the Semantic Web, the Web includes growth of machine-readable descriptions of people: FOAF documents. The FOAF provides an RDF/XML vocabulary to describe personal information, including name, mailbox, homepage URI, interest, friends, and so on. Using FOAF documents, we can construct social networks among people. Given a list of persons V, we first use *foaf:Person* to mapping each name with FOAF instances, then connect persons with several meaning of relational properties such as *foaf:knows*, *foal:interest*, *foaf:location*, *foaf: publications*, and *foaf: currentProject* properties. Consequently, we can construct social networks G_i of different kinds. When a person is described in more than one FOAF document, we must fuse information from multiple sources using identical properties such as *foaf:mbox*, *foaf:homepage* and *foaf:Weblog* and generate aggregated information about the person [35]. Furthermore, by combining FOAF documents to DBLP data, we can construct more kinds of social networks such as *authorship* network, *citation* network [7, 115].

Another is to extract social networks using Web mining. Several studies have particularly addressed the use of search engines as well as text mining for social network extraction. Through this study, we detail the co-occurrence approach and relation-identification approach used by Matsuo et al. [70] and Jin et al. [47], respectively, as a basis of our study.

We are interested only in undirected networks.

7.3.1 Co-occurrence-based approach

The social network of the first kind is extracted using a co-occurrence-based approach. This approach was used originally by Kautz et al. [51], and was recently applied and modeled by Mika [73] and Matsuo et al. [70] to extract researcher networks automatically from the Web. The fundamental idea underlying the co-occurrence approach is that *the strength of a relation between two entities can be estimated by co-occurrence of their names on the Web*. The strength of relevance of two persons, x and y, is estimated by putting a query x AND y to a search engine: If x and y share a strong relation, we can usually find various evidence on the Web such as links found on home pages, lists of co-authors of technical papers, organization charts, and so on. An edge will be invented when the relation strength by the co-occurrence measure is higher than a predefined threshold.

Subsequently, we use the Overlap coefficient $n_{x \wedge y} / \min(n_x, n_y)$ (used by [70]) as well as the Matching coefficient as relational indices and thereby construct co-occurrence-based networks of two kinds: an overlap network ($G_{overlap}$) and a cooc network (G_{cooc}). Many advanced algorithms are described in [70].

7.3.2 Classification-based approach

The classification-based approach was proposed by Matsuo et al. [70], and also applied by our Yokohama Triennale system described in chapter 5. This approach identifies concrete relation types (labels) from Web pages retrieved by names of actors who connected at the Web co-occurrence-based networks; it also filters out noisy edges to improve the system precision. First, it uses several features from Web contents retrieved by two names of actors to create identification rules of relations on the training data. For example, if two names and terms (such as "*department*", "*graduate*", "lecture"), which represent relations, appear in the same Web page, they are judged as co-affiliation relations. If two names appearing in the same tables of a Web page and the Web page also contain terms (such as "*project*", "*committee*", "*member*") that represent an project or an event, they are judged as co-project relations.

Subsequently, in our experiment, based on overlap network among researchers, we classify the edges into two kinds of relational networks: an co-affiliation network ($G_{affiliation}$) and a co-project network ($G_{project}$).

7.3.3 Relation-identification approach

We proposed the *relation-identification* approach to extract target relational social networks in chapter 4 This approach emphasizes real-world relations such as a mutual stock holding relation, capital combination, trade relation, personal relation (i.e., mutual dispatch of officials), rivalry, and a competitive relation. These relations are published in news articles or by news releases that might be obtained easily from the Web.

Given a list of companies and target relations as input, the method extracts a social network of entities. To collect target relational information from the tops of Web pages, it makes elaborate queries to emphasize a specific relation, and applies text processing to those pages to form an inference of whether or not the relation actually exists. First, queries are produced by adding *relation keywords* (such as "*alliance* AND *corporate*") to each pair of companies. Relation keywords are in advance for each target relation by measuring the Jaccard relevance from given seed words. Then, to extract target relations from Web documents, a simple pattern-based heuristic is useful: First pick all sentences that include the two company names (x and y), and assign each sentence the sum of relation keyword scores in the sentence. The score of companies x and y is the maximum of the sentence scores. An edge is invented between the two companies if that score is greater than a certain threshold. Subsequently, we extract two kinds of relational networks: a business-alliance network ($G_{business}$) and a capital-alliance network ($G_{capital}$).

Extracted networks for 312 companies related to the electrical products industry from Japan and for 253 researchers of The University of Tokyo are portrayed in Fig. 7.1 and Fig. 7.4, respectively. It is apparent that the social networks vary with different relational indices or types even though they contain the same list of entities.

7.4 Ranking Learning Model

For the list of nodes $V = \{v_1, \ldots, v_n\}$, given a set of networks $G_i(V, E_i)$, $i \in \{1, \ldots, m\}$ (constructed by section 7.3) with a target ranking $\mathbf{r}^*$ ($\in R^t$) (where $t \leq n$, and r^*_k denotes k-th element of the vector $\mathbf{r}^*$ and means the target ranking score of entity v_k), the goal is to learn a ranking model based on these networks.

First, as a baseline approach, we follow the intuitive idea of simply using approach from SNAs (i.e. centrality) to learn ranking. As the second approach, multiple relations are combined into one to consider a combination model for ranking. Finally, to learn ranking, we propose a more useful algorithm that generates various network features for individuals from social networks.

7.4.1 Baseline Model

In this section, based on the intuitive approach, we first overview commonly used indices in social network analysis and complex network studies. Given a set of social networks, we rank entities on these networks using different network centrality indices. We designate these rankings as *network rankings* because they are calculated directly from relational networks. We use $\mathbf{r}_i$ ($\in R^n$) to denote network ranking that is directly attributable to the i-th relational network G_i. Our task is to find a ranking model based on network rankings that maximally explain the target ranking.

Choosing the most predictive type of relation

To address the question of what kind of relation is most important for companies, we intuitively compare rankings caused by relations of various types. Although simple, it can be considered as an implicit step of social network analysis given a set of relational networks. We merely choose the type of relation that maximally explains the given ranking. We rank each type of relational network; then we compare the *network ranking* with the *target ranking*. Intuitively, if the correlation to the network ranking $\mathbf{r}_{\hat{i}}$ is high, then the relation $\hat{i}$ represents the important influences among entities for the given target ranking. Therefore, this model is designed to find an optimal relation $\hat{i}$ from a set of relations:

$$\hat{i} = \underset{i \in \{1,\ldots,m\}}{\operatorname{argmax}} Cor(\mathbf{r}_i, \mathbf{r}^*) \tag{7.1}$$

We define a ranking function $h(G)$ that returns a vector of network ranking ($\in R^n$) for given network $G(V, E)$. Therefore, the i-th network ranking $\mathbf{r}_i$ is obtained from $h(G_i)$. Here are the other questions for what kind of ranking indices are most appropriate to explain the target ranking. In the next section, we treat several *centrality* measures from SNAs as our different network ranking function $h(G)$.

Choosing the most predictive type of centrality indices

Different meanings of prominence and importance can be generated from a network, such as "having a powerful position", and having "more opportunities" and ”fewer constraints”. Several *centrality* measures are useful to rank network entities with these different meanings: degree centrality, betweenness centrality, and closeness centrality and other centralities. Bellow, we introduce these different meanings of centrality.

- *Degree centrality* is an assessment of the number of relations that any given actor is engaged in. Actors with more ties to other actors might be in advantaged positions, which can be defined as

  $$C_d(v_l) = \frac{d(v_l)}{(n-1)},$$

  Therein, $d(v_l)$is the degree of node v_l, and n is the number of nodes.

- *Betweenness centrality* measures an actor as central if it lies between other actors on their geodesics. More actors depend on one actor v_l to make connections with other actors (geodesics passing through).

  $$C_b(v_l) = \frac{\sum_{(v_p,v_q)\in(V\times V),v_p\in V,v_q\in V} g_{v_p,v_q}(v_l)/g_{v_p,v_q}}{(n-1)(n-2)}$$

  where q_{v_p,v_q} is the number of shortest geodesic paths from node v_p to v_q, and $g_{v_p,v_q}(v_l)$ is the number of shortest paths from v_p to v_q that pass through node v_l

- *Closeness centrality* is a sophisticated measure that is defined as the mean shortest path between an actor i and all other actors that are reachable from that actor. Closeness can be regarded as a measure of how long it will take information to spread from a given actor v_l to other reachable actors in the network.

$$C_c(v_l) = \frac{\sum_{v_p \in V, v_p \neq v_l} g_G(v_l, v_p)}{(n-1)}$$

In that equation, $g_G(v_l, v_p)$ is the shortest geodesic paths from v_l to reachable node v_p.

These measures characterize some aspects of the local (i.e. degree) or global (i.e., closeness, betweenness) network structure, as indicated by a given actor's embeddedness in the network [107]. Intuitively, given a target ranking, the most predictive type of centrality measure is finding optimal centrality measure $h_{\hat{j}}$ for target ranking $\mathbf{r}^*$ from a set of ranking functions.

$$\hat{j} = \underset{h_j \in \{h_1, \ldots, h_s\}}{\text{argmax}} \; Cor(\mathbf{r}_{,j}, \mathbf{r}^*) \tag{7.2}$$

For different relational networks, the network ranking from i-th network with j-th ranking can be presented as $\mathbf{r}_{i,j}$ $(\in R^n)$, which is obtainable from $h_j(G_i)$, where $h_j \in \{h_1, \ldots, h_s\}$, $i \in \{1, \ldots, m\}$. Therefore, the first method can be extended simply to find a pair of optimal parameters $< \hat{i}, \hat{j} >$ (i.e., i-th network by j-th ranking indices) that maximizes the coefficient between network rankings with a target ranking.

$$< \hat{i}, \hat{j} >= \underset{i \in \{1, \ldots, m\} \; h_j \in \{h_1, \ldots, h_s\}}{\text{argmax}} \; Cor(\mathbf{r}_{i,j}, \mathbf{r}^*) \tag{7.3}$$

7.4.2 Network Combination Model

Many centrality approaches related to ranking network entities specifically examine graphs with a single link type. However, multiple social networks exist in the real world, each representing a particular relation type, and each of which might be integrated to play a distinct role in a particular task. We combine several extracted multiple social networks into one network and designate such a social network as a *combined-relational network* (denoted as $G_c(V, E_c)$). Our target is using combined-relational network, which is integrated with multiple networks extracted from the Web, to learn and predict the ranking. The important questions that must be resolved here is *how to combine relations to describe the given ranking best.*

For $G_c(V, E_c)$, the set of edges is $E_c = \{e_c(v_x, v_y)|v_x \in V, v_y \in V, v_x \neq v_y\}$. Using a linear combination, each edge $e_c(v_x, v_y)$ can be generated from $\sum_{i\in\{1,\ldots,m\}} w_i e_i(v_x, v_y)$, where w_i is i-th element of $\mathbf{w}$ (i.e., $\mathbf{w} = [w_1, \ldots, w_m]^T$). Therefore, the purpose is to learn optimal combination weights $\hat{\mathbf{w}}$ to combine relations as well as optimal ranking method h_j on G_c:

$$< \hat{\mathbf{w}}, \hat{j} >= \underset{\mathbf{w}, h_j \in \{h_1,\ldots,h_s\}}{\operatorname{argmax}} Cor(\mathbf{r}_{c,j}, \mathbf{r}^*). \tag{7.4}$$

Cai et al. [21] regard a similar idea with this approach: They attempt to identify the best combination of relations (i.e., relations as features) which makes the relation between the intra-community examples as tight as possible. Simultaneously, the relation between the inter-community examples is as loose as possible when a user provides multiple community examples (e.g. two groups of researchers). However, our purpose is learn a ranking model (e.g. ranking of companies) based on social networks, which has a different optimization task. Moreover, we propose innovative features for entities based on combination or integration of structural importance generated from social networks.

In this study, we simply use Boolean type ($w_i \in \{1, 0\}$) to combine relations. Using relations of m types to combine a network, we can create $2^m - 1$ types of combination-relational networks (in which at least one type of relation exists in the G_c). We obtain network rankings in these combined networks to learn and predict the target rankings. Future work on how to choose parameter values will be helpful to practitioners.

7.4.3 Network-based Feature Integration Model

The most advanced method in our research is to integrate multiple indices that are obtained from multiple social networks. A feature by itself (e.g. a centrality value) may have little correlation with the target ranking, but when it is combined with some other features, they may be strongly correlated with the target rankings [114]. The idea in this model is the integration of all network features for individuals from networks as a context of the actors to learn the target ranking. Those features are expected to be useful to interpret a given target ranking accurately.

We integrate multiple indices from social networks, thereby combining several perspectives of importance for individuals from different relational structures. Simply, we can integrate various centrality values (described in the Baseline model) for each actor, thereby combining different meanings of importance to learn the ranking model. Furthermore, we can generate more relational and structural features from a network for each, such as how many nodes are reachable, how many connections one's friends have, and the connection status in one's friends. We might understand some about the behavior and power about the individual as well as we predict their ranking if we could know the structural position of individuals. Herein, we designate these features generated from relations and networks as *network-based features*. The interesting question is *how to generate network-based features from networks for each*, and *how to integrate these features to learn and predict rankings*. Below we will describe the approach of generating and integrating network-based features.

Generating Network-based Features for nodes

For each x, we first define node sets with relations that might effect x. Then we apply some operators to the set of nodes to produce a list of values. Subsequently, the values are aggregated into a single feature value. Therefore, we can generate several structural features for each nodes. For example, when calculating the closeness centrality (i.e., average distance from node x to all others) of node x, we discern its value fundamentally in three steps: we first select reachable nodes from x; secondly, we calculate the distance between node x and each node; finally, we take the average of these distances. Additionally, we can discern the value of the closeness centrality of node x. For that reason, we can construct indices used in SNAs through these steps. Below, we explain each step in detail.

- Step 1: Defining a node set First, we define a node set. Most straightforwardly, we can choose the nodes that are adjacent to node x. The nodes are those of distance one from node x. The nodes with distances of two, three, and so on are definable as well. We define a set of nodes $C_x^{(k)}$ as a set of nodes within distance k from x. For example, we can denote the node set adjacent to node x as C_x^1. In addition, we use $C_y^{(k)}$ to express a set of nodes within distance k from y (where $y \neq x$).

- Step 2: Operation on a Node Set Given a node set, we can conduct several calculations for the node set. Below, we define operators with respect to two nodes; then we expand it to a node set with an arbitrary number of nodes.

  The simple operation for two nodes is to check whether the two nodes are adjacent or not. We denote these operators as $s^{(1)}(x, y)$, which returns 1 if nodes x and y are mutually connected, and 0 otherwise. We also define operator $t(x, y) = argmin_k\{s^{(k)}(x, y) = 1\}$ to measure the geodesic distance between the two nodes on the graph. These two operations are applied to each pair of nodes in N if given a set of more than two nodes (denoted as N). This calculation can be defined as follows.

$$Operator \circ N = \{Operator(x, y) | x \in N, y \in N, x \neq y\}$$

For example, if we are given a node set $\{ n_1, n_2, n_3\}$, we can calculate $s^{(1)}(n_1, n_2)$, $s^{(1)}(n_1, n_3)$, and $s^{(1)}(n_2, n_2)$ and return a list of three values, e.g., $(1, 0, 1)$. We denote this operation as $s^{(1)} \circ N$.

In addition, to s and t operations, we define two other operations. One operation is to measure the distance from node x to each node, denoted as t_x. Instead of measuring the distance between two nodes, $t_x \circ N$ measures the distance of each node in N from node x. Another operation is to check the shortest path between two nodes. Operator $u_x(y, z)$ returns 1 if the shortest path between y and z includes node x. Consequently, $u_x \circ N$ returns a set of values for each pair of $y \in N$ and $z \in N$. The other is to calculate the structural equivalence between node x and y. This is denoted as $e_x(y)$.

- Step 3: Aggregation of Values

  Once we obtain a list of values, several standard operations can be added to the list. Given a list of values, we can take the summation (Sum), average (Avg), maximum (Max), and minimum (Min). For example, if we apply Sum aggregation to a value list $(1, 0, 1)$, we obtain a value of 2. We can write the aggregation as e.g., $Sum \circ s^{(1)} \circ N$. Although other operations can be performed, such as taking the variance or taking the mean, we limit the operations to the four described above. The value obtained here results in the network-based feature for node x.

  Additionally, we can take the difference or the ratio of two obtained values. For example, if we obtain 2 by $Sum \circ s^{(1)} \circ C_x^{(1)}$ and 1 by $Sum \circ s^{(1)} \circ C_x^{(k)}$, the ratio is $2/1 = 2.0$.

We can thereby generate a feature by subsequently defining a nodeset, applying an operator, and aggregating the values. The number of possible combinations is enormous. Therefore, we apply some constraints on the combinations. First, when defining a nodeset, k is an arbitrary integer theoretically; however, we limit k to be 1 for a nodeset of neighbors, k to be 3 for a nodeset of reachable nodes simplicity. Operator $s^{(k)}$ is used only as $s^{(1)}$. We also limit taking the ratio only to those two values with neighbor nodeset $C_x^{(1)}$ and reachable nodeset $C_x^{(\infty)}$. The nodesets, operators, and aggregations are presented in Table 7.1. We have $2(nodesets) \times 5(operators) \times 4(aggregations) = 40$ combinations. There are ratios for $C_x^{(1)}$ to $C_x^{(k)}$ if we consider the ratio. In all, there are 4×5 more combinations: there are 60 in all. Each combination corresponds to a feature of node x. Some combinations produce the same value. One example is that $Sum \circ t_x \circ C_x^{(1)}$ is the same as $Sum \circ s \circ C_x^{\infty}$, representing the degree of node x.

The resultant value sometimes corresponds to a well-known index, as we intended in the design of the operators. For example, the network density can be denoted as $Avg \circ s^{(1)} \circ N$. It represents the average of edge existence among all nodes; it therefore corresponds to the network density. These features represent some possible combinations. Some lesser-known features might actually be effective.

Table 7.1: Operator list

Notation	Input	Output	Description
$C_x^{(1)}$	node x	a nodeset	adjacent nodes to x
$C_x^{(k)}$	node x	a nodeset	nodes within distance k from x
$s^{(1)}$	a nodeset	a list of values	1 if connected, 0 otherwise
t	a nodeset	a list of values	distance between a pair of nodes
t_x	a nodeset	a list of values	distance between node x and other nodes
γ	a nodeset	a list of values	number of links in each node
u_x	a nodeset	a list of values	1 if the shortest path includes node x, 0 otherwise
Avg	a list of values	a value	average of values
Sum	a list of values	a value	summation of values
Min	a list of values	a value	minimum of values
Max	a list of values	a value	maximum of values
$Ratio$	two values	value	ratio of value on neighbor nodeset $C_x^{(1)}$ by reachable nodeset $C_x^{(\infty)}$

Network-based features with SNAs indices

It is readily apparent that centralities described in baseline approach are also a particular case of this model because our network-base feature include those centrality measures and other SNAs indices for each node. Below, we describe other examples that are used in the social network analysis literature.

- diameter of the network: $Min \circ t \circ N$
- characteristic path length: $Avg \circ t \circ N$
- degree centrality: $Sum \circ s_x^{(1)} \circ C_x^{(1)}$
- node clustering: $Avg \circ s^{(1)} \circ C_x^{(1)}$
- closeness centrality: $Avg \circ t_x \circ C_x^{(\infty)}$
- betweenness centrality: $Sum \circ u_x \circ C_x^{(\infty)}$,
- structural holes: $Avg \circ t \circ C_x^{(1)}$

When we set the element $Sum \circ s_x^{(1)} \circ N_x^{(1)}$ in a feature vector equal to 1, and all others to 0, we can elucidate the effect of degree centrality for predicting target ranking.

Network-based feature Integration Model

Next, generated network-based features to learn rankings are used for entities. The goal of learning is to integrate all features from networks into a single ranking of individuals. Combined, they are expected to be useful to interpret a given target ranking most accurately.

After we generate various network-based features for individual nodes, we integrate them to learn ranking. This integration is accomplished through regression of features. We introduce an f-dimensional feature vector F, in which each element represents a network-based feature for each node. We identify the f-dimensional combination vector $\mathbf{u} = [u_1, \ldots, u_f]^T$ to combine network-based features for each node. The inter-product $\mathbf{u}^T\dot{\mathbf{F}}$ for each node produces n-dimensional ranking. For relational networks of m kinds, the feature vector can be expanded to $m \times 56$-dimensions. In this case, the purpose is finding out whether optimal combination weight $\hat{\mathbf{u}}$ to $\mathbf{u}^T\dot{\mathbf{F}}$ maximally explains the target ranking:

$$\hat{\mathbf{u}} = \underset{\mathbf{u}}{\operatorname{argmax}} \, Cor(\mathbf{u}^T \bullet F, \quad \mathbf{r}^*) \tag{7.5}$$

This model can be extended easily to add attributes (or profiles) of entities as features such as Sales, Assets, or number of employees of a company. We can use any technique, such as SVM, boosting and neural network, to implement the optimization problem. In this study, we consider using the Ranking SVM technique. Ranking SVM utilizes instance pairs and their preference labels in training. The optimization formulation of Ranking SVM is the following:

$$\min \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + \mathbf{C}\ \boldsymbol{\Sigma}_{\mathbf{i,j,q}}\ \boldsymbol{\zeta}_{\mathbf{i,j,q}}.$$

$$s.t.\forall(d_i, d_j) \in r_q^* : w\phi(q, d_i) \geq w\phi(q, d_j) + 1 - \zeta_{i,j,q} \tag{7.6}$$

where $\mathbf{w}$ is a weight vector that is adjusted by learning to minimize the upper bound $\Sigma\ \zeta_{i,j,q}$. In addition, C is a parameter that enables trading-off of the margin size against training error. The result is a ranking function that has few discordant pairs with respect to the observed of the target ranking.

For multi-relational networks, we can generate features for each single-relational network. Subsequently, we can compare the performance among them to understand which relational network produces more reasonable features. Thereby, we can see which relation(s) is important for the target ranking.

In the following sections, we describe results and thereby clarify the effectiveness of ranking learning on extracted social networks in two different fields: company and researcher. For the first trial, we use 312 electrical-product-related companies listed on the Tokyo Stock Exchange [1] to predict rankings of companies. For the second trial, we use 253 researchers from The University of Tokyo to predict a ranking of researchers.

[1] http://profile.yahoo.co.jp/industry/electrical/electrical1.html

7.5 Case Study 1: Ranking Companies using Social Networks

7.5.1 Datasets

We extract social networks for companies from 312 electrical product-related industry companies that are listed on the Tokyo Stock Exchange. All financial information about these companies is published in Yahoo! Finance[2]. For these companies, we extract social networks of seven kinds (Fig. 7.1) from the Web using a search engine Yahoo! Search Boss [3] and information from Toyo Keizai Inc.[4]: the cooc network (G_{cooc}) and overlap network (G_{over}) network are extracted using the co-occurrence-based approach described in Section 7.3.1; the business-alliance network ($G_{business}$) and capital-alliance network ($G_{capital}$) are extracted using the relation-identification approach described in Section 7.3.3 (details in chapter 4); same-market network (G_{market}) includes links that connect two companies listed on the same stock market; shareholding network ($G_{shareholder}$) connects shareholding relations among companies; similar-age network (G_{age}) connects two companies if their average age is similar (age-gap is less than two years); Each extraction method and corresponding figure of networks is listed in Table 7.2.

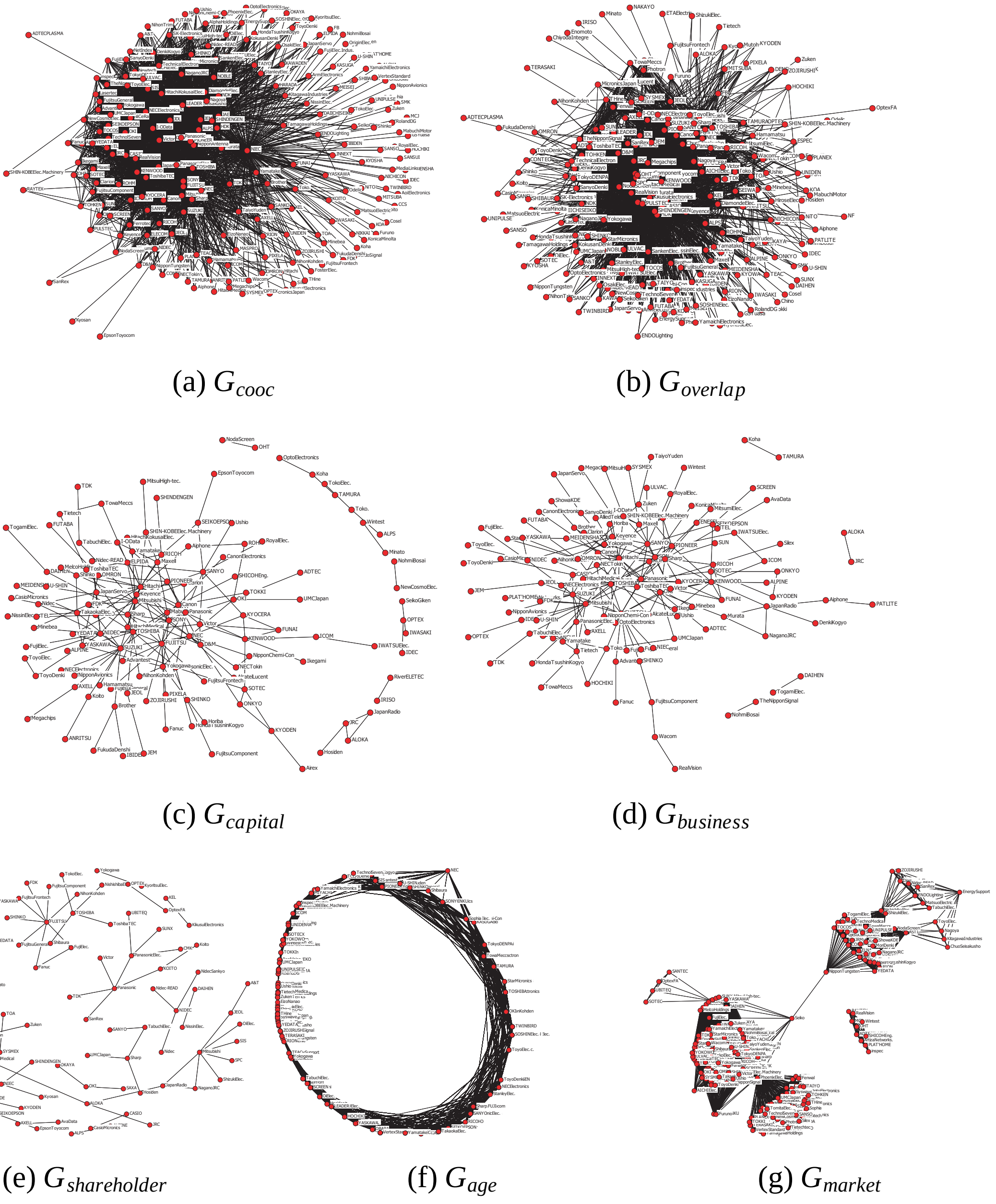
For our experiments, we set the target ranking of the companies by market capitalization (designated as Market-Cap), ranking of average annual income (designated as Avg-In), and the ranking of excellent accounts (designated as Excellent). The target ranking of Avg-In is collected from quarterly corporate reports from Toyo Keizai Inc. Market-Cap represents the market's valuation of all the equity in a corporation. From Yahoo! Finance we can obtain all Market-Cap information for listed companies in Japan. The ranking of Excellent is published by Nihon Keizai Shimbun Inc.[5] every year in March. They rank companies based on evaluating factors of flexibility & sociality, earning & growth ability, development & research, age of employees, etc. The top 300 excellent companies include 22 electrical industry companies used in our experiments. Table 7.3 shows the top 25 companies ranked

[2] http://profile.yahoo.co.jp/industry/electrical/electrical1.html

[3] http://developer.yahoo.com/search/boss/

[4] Toyo Keizai Inc. (http://www.toyokeizai.co.jp/): a Japanese book and magazine publisher.

[5] http://www.nikkei.co.jp/



(a) G_{cooc} (b) $G_{overlap}$

(c) $G_{capital}$ (d) $G_{business}$

(e) $G_{shareholder}$ (f) G_{age} (g) G_{market}

Figure 7.1: Social networks for companies in electrical industrial with different relational indices or types.

Table 7.2: Constructed networks of electrical industry companies.

G_i	Network name	Extraction Method	Figs.
G_{cooc}	cooc network	Section 7.3.1	Fig. 7.1(a)
$G_{overlap}$	overlap network	Section 7.3.1	Fig. 7.1(b)
$G_{business}$	business-alliance network	Section 7.3.3	Fig. 7.1(c)
$G_{capital}$	capital-alliance network	Section 7.3.3	Fig. 7.1(d)
G_{market}	same-market network	connect companies listed on the same stock market	Fig. 7.1(e)
$G_{shareholder}$	shareholding network	connect shareholding relations	Fig. 7.1(f)
G_{age}	similar-age network	connect similar average-age companies	Fig. 7.1(g)

by Avg-In, Market-Cap, and Excellent in the electrical industry.

In our experiments, we conducted three-fold cross-validation. In each trial, two folds of actors are used for training, and one fold for prediction. The results we report in this section are those averaged over three trials. We use Spearman's rank correlation coefficient (ρ) [96] to measure the pairwise ranking correlation.

$$\rho = 1 - \frac{6\Sigma_i^2}{n(n^2 - 1)} \tag{7.7}$$

In that equation, d_i is the difference between the ranks of corresponding values X_i and Y_i.

7.5.2 Ranking Results

First, we rank companies on different networks according to their network rankings. Table 7.4 and Table 7.5 show the top 20 companies ranked by degree centrality and betweenness centrality, respectively, on different types of networks in the electrical industry field. Results show that *Hitachi*, *NIEC*, and *Fujitsu* have good degree centrality in different networks. In addition, *Hitachi* has good betweenness centrality in the networks: we can implicitly understand that *Hitachi* has good network embeddedness in the electrical industry. Additionally, these results reflect that companies have different centrality rankings even if they are in the same type of relational network. For instance, *Phoenix Elec.* and *SanRex* have good degree

Table 7.3: Top 25 companies ranked by target rankings i.e. Avg-In, Market-Cap, and Excellent in an electrical industry field.

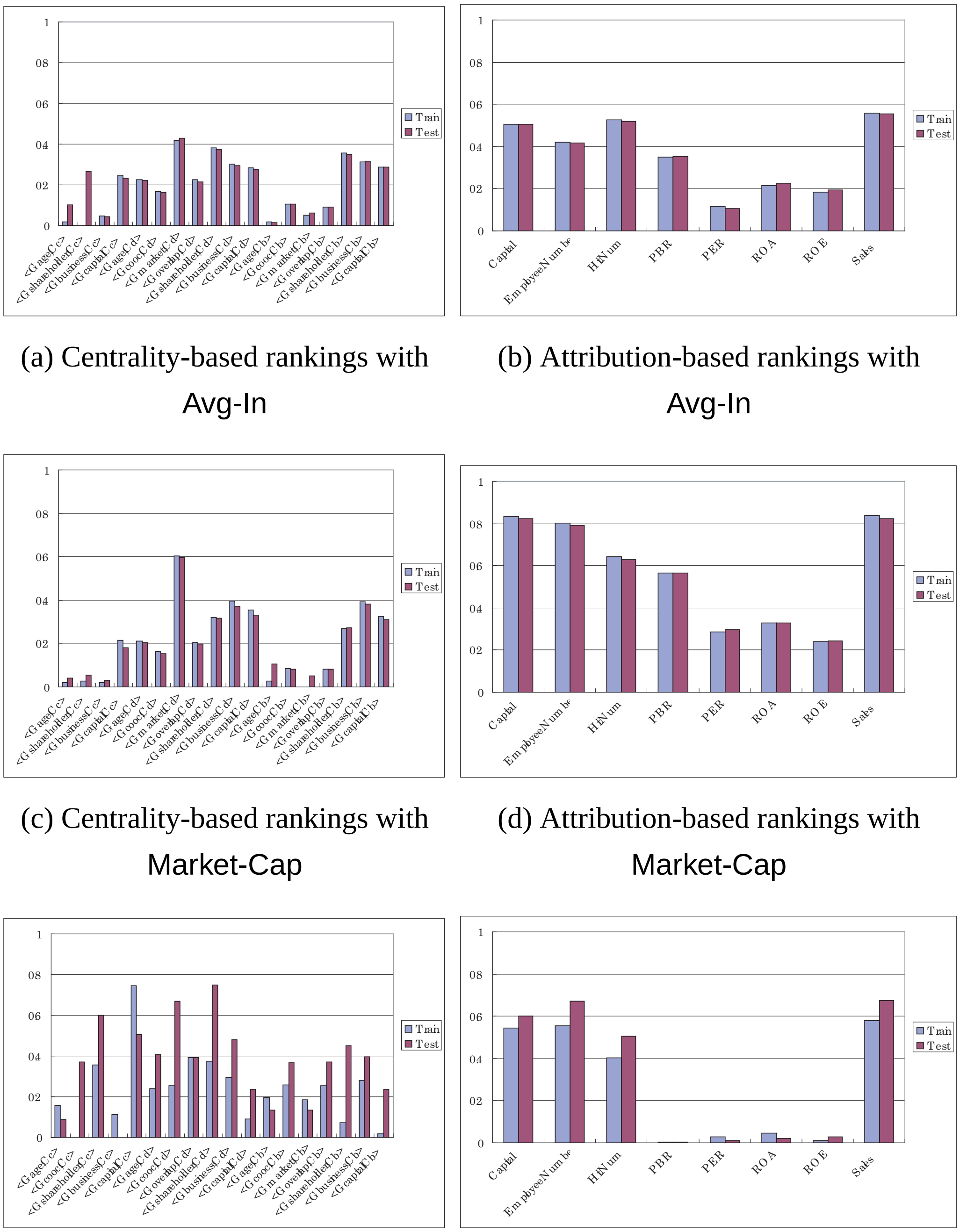
r*	Avg-In	Market-Cap	Excellent
1:	Keyence	Canon	Canon
2:	Advantest	Sony	Fanuc Ltd.
3:	AXELL	Panasonic	TDK
4:	Lasertec	Toshiba	Omron
5:	Fanuc Ltd.	Hitachi	Kyocera
6:	TEL	Mitsubishi	Sysmex
7:	Sony	Fanuc Ltd.	Ricoh
8:	Screen	Sharp	Toshiba
9:	Yokogawa	Kyocera	Ibiden
10:	Elpida	Fujitsu	Rohm
11:	Canon	Ricoh	Sharp
12:	Nihon Kohden	Murata	Sony
13:	Panasonic	Keyence	Eizo Nanao
14:	Megachips	Ibiden	Fujitsu
15:	Ricoh	TEL	Optex
16:	Nippon Signal	Nidec	Cosel
17:	Ulvac	Rohm	Daihen
18:	Hirose Elec.	Konica Minolta	SMK
19:	SK Elec.	TDK	Yamatake
20:	Panasonic Elec.	NEC	Ulvac
21:	Fujitsu	Panasonic Elec.	Hioki E.E.
22:	Omron	Omron	Nihon Kohden
23:	Toshiba	Advantest	
24:	Casio	Elpida	
25:	Yaskawa	Hirose Elec.	

rankings in G_{market} and G_{age} networks respectively, but do not have good betweenness rankings in those networks. We also use seven carefully chosen fundamental indices as attributes of companies for comparison of our proposed network indices: Capital, Emplyee Number, Sales, return on equity (ROE), return on assets (ROA), the price earnings ratio (PER), and the price to book value ratio (PBR). Each of them has been used traditionally for company valuation. Additionally, we use the number of hits of names (HitNum) on the Web as another attribute (i.e. popularity on the Web) of a company. Table 7.6 shows the top 20 companies ranked by each attribute in the electrical industry field.

As a baseline model, we use three centrality indices (i.e., degree centrality C_d, closeness centrality C_c, and betweenness centrality C_b) on different networks (G_{cooc}, $G_{overlap}$, $G_{capital}$, $G_{business}$, $G_{shareholder}$, G_{age}, G_{market}) as network rankings, and calculate the correlation between network rankings with each target ranking: Avg-In, Excellent, and Market-Cap. For comparison, we also rank companies according to previously described attributes (i.e., seven fundamental indices and hit number of names on the Web), and calculate the correlation with target rankings. Fig. 7.2 presents correlations (mean of three tries) of each network ranking as well as each attribute-based ranking with different target rankings on training and testing data in the electrical industry. These results demonstrate that rankings of betweenness centrality in same-market network ($\mathbf{r}_{G_{market},C_b}$) and in shareholding relational network ($\mathbf{r}_{G_{shareholder},C_b}$) have good correlation with the target ranking of Avg-In. Betweenness centralities in the cooc network ($\mathbf{r}_{G_{cooc},C_b}$), betweenness centralities and degree centralities in the business-alliance network as well as the capital-alliance network ($\mathbf{r}_{G_{business},C_b}$, $\mathbf{r}_{G_{capital},C_b}$, $\mathbf{r}_{G_{business},C_d}$, $\mathbf{r}_{G_{capital},C_d}$) all show good correlation with the target ranking of Market-Cap. Betweenness centralities in the capital-alliance network and shareholding relational network correlate well with the target ranking of Excellent.

In the combination model, we simply use Boolean type ($w_i \in \{1, 0\}$) to combine relations. Using relations of seven types to combine a network $G_{overlap-business-capital-market-shareholder-age-cooc}$, we can create $2^7 - 1$ (=127) types of combination-relational networks (in which at least one type of relation exists). We obtain network rankings in these combined networks to learn and predict the target rankings. The top 50 correlations between network rankings in combined-relational network and target rankings are presented in Fig. 7.3. Results demonstrate that degree centralities on combined-relational network produce good correlation with target rankings. For the target ranking of Avg-In, a network $G_{1-0-0-1-1-0-1}$ comprising overlap relations, same-market relations, shareholding relations, and cooc relations shows good correlation. They outperform the baseline approach. For the target ranking of Market-Cap, the combined-relational networks which combined by overlap relation, capital-alliance relation, same-market relation, and shareholding relation $G_{1-1-1-1-1-0-0}$, $G_{1-0-1-1-1-0-0}$ show good correlation. For the target ranking of Excellent, closeness centralities in the capital-alliance network outperform other combinations. Future work on how to choose parameter values will yield results that will be especially helpful to practitioners.

We execute our feature integration ranking model (with several varies) to single and multi-relational social networks to train and predict three different targets rankings: Avg-In, Excellent, and Market-Cap. We use Ranking SVM to learn the ranking model which minimize pairwise training error in the training data; then we apply the model to predict rankings on training data (again) and on testing data. Comparable results for several varieties of model are presented in Table 7.7. Below we will explain a trial of each and interpret the results. First, we integrate the attributes of companies (i.e., several fundamental indices plus hit number of names on the Web) as features, and treat it as a baseline of feature-integration models to learn and predict the rankings. We can obtain 0.389 correlation for Avg-In, 0.571 correlation for Excellent, and 0.718 correlation for Market-Cap using these attribute-based features. This means that fundamental indices are quite good features for explaining target rankings, and are especially good for Market-Cap. Then, we integrate proposed network-based features obtained from each type of single network as well as multi-relational networks to train and predict the rankings. These results show that integrating the features in the network of G_{market}, G_{age}, $G_{capital}$ yields good performance for explaining the ranking of Avg-In, features in the G_{cooc}, $G_{shareholder}$ explain ranking of Excellent, and features in the G_{market}, $G_{business}$, and $G_{capital}$ have good performance for explaining the ranking of Market-Cap. These results reflect that relations and networks of different types produce different impacts on different target of rankings. Some examples are the following. Listing on the same stock market and connection with similar average-age companies are related to higher average incomes of companies. Co-occurence with many other companies on the Web, shareholding relations with big companies are associated with a company being more well-known; consequently, the company has an excellent ranking. Active collaboration with other companies through business and capital alliances are associated with higher market value company. Using the features from multi-relational networks G_{ALL}, the prediction results are higher than those of any other single-relational network. This conforms to the intuition that multi-relational networks have more information than single networks to explain real-world phenomena. Furthermore, we combine network-based features with attribute-based features to train the model. The prediction results for any target ranking outperform each of the use of attribute-based features alone or network-based features alone. The correlation with target ranking of Market-Cap improved little from 0.718 (attribute only), 0.645 (network only) to 0.756 (both); the correlation with Avg-In shows remarkable changes from 0.389 (attribute only), to 0.584 (network only) and 0.601 (both), which means that market values are explained more by fundamental attributes than relations among companies, although average incomes for companies are more understandable according to relations among companies than fundamental indices. The overall results demonstrate that, even thought the attribute-based features have good performance for explaining Market-Cap than network-based features, by combining network-based features with attribute-based features, the prediction results are improved. The target rankings of Avg-In and Excellent are more explainable by integrating network-based features than attribute-based features. Demonstrably combining both network and attribute-based features yields further improved prediction results.



(a) Centrality-based rankings with Avg-In

(b) Attribution-based rankings with Avg-In

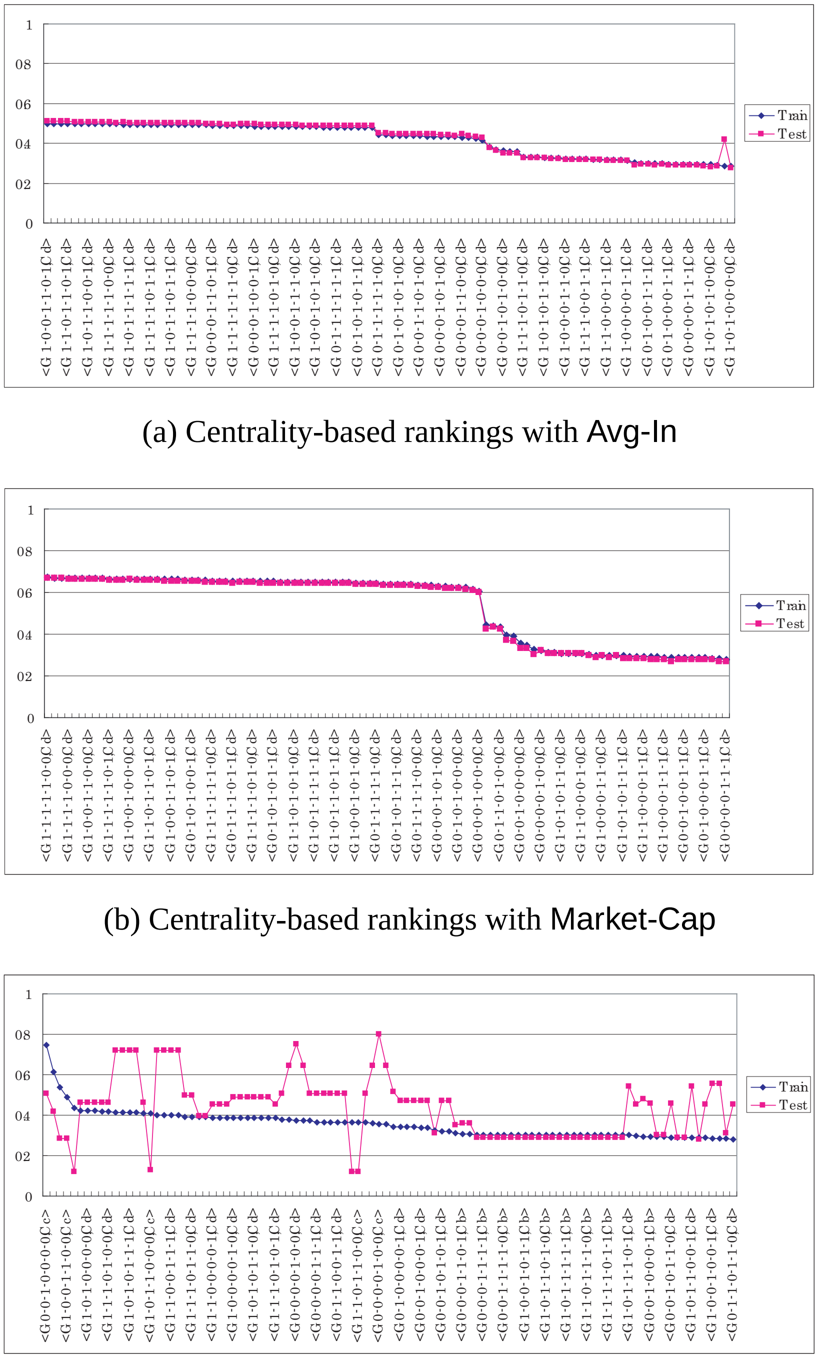
(c) Centrality-based rankings with Market-Cap

(d) Attribution-based rankings with Market-Cap

(e) Centrality-based rankings with Excellent

(f) Attribution-based rankings with Excellent

Figure 7.2: Evaluation for each centrality-based ranking, along with a attribute-based ranking with different target rankings in the electrical industry.



(a) Centrality-based rankings with Avg-In

(b) Centrality-based rankings with Market-Cap

(e) Centrality-based rankings with Excellent

Figure 7.3: Evaluation for network rankings in a combined-relational network with different target rankings in the electrical industry.

Table 7.4: Top 20 companies ranked by degree centrality on different social networks in the electrical industry.

$\mathbf{r}_{i,Cd}$	$\mathbf{r}_{cooc,Cd}$	$\mathbf{r}_{overlap,Cd}$	$\mathbf{r}_{business,Cd}$	$\mathbf{r}_{capital,Cd}$	$\mathbf{r}_{market,Cd}$	$\mathbf{r}_{shareholder,Cd}$	$\mathbf{r}_{age,Cd}$
1:	NIEC	Keyence	Hitachi	Hitachi	Phoenix Elec.	Hitachi	SanRex
2:	JEM	Shindengen	Fujitsu	Suzuki	NIEC	Fujitsu	ALOKA
3:	Toshiba	HDK	Suzuki	Fujitsu	Shibaura	Mitsubishi	Koito
4:	JAE	Casio	Panasonic	Toshiba	Hamamatsu	Panasonic	TOA
5:	Pioneer	JAE	NEC	Mitsubishi	Nihon Kohden	Toshiba	Hitachi Medical
6:	JDL	Murata	Mitsubishi	Panasonic	Kenwood	Panasonic Elec.	Maxell
7:	Sony	Pulstec	Sharp	Nidec	Pixela	Sharp	Ichikoh
8:	Nippon Antenna	Clarion	Toshiba	Sony	ALOKA	ALOKA	Noble
9:	Chuo Seisakusho	Real Vision	Kenwood	NEC	Iwasaki	Nidec	Lasertec
10:	Panasonic	Kenwood	Oki	Sharp	JRC	Japan Radio	Soshin Elec.
11:	Shindengen	Hitachi Medical	Pioneer	Canon	JAE	Hitachi Medical	HitachiKokusaiElec.
12:	Leader	Kikusui Elec.	Sony	Sanyo	Mutoh	NIEC	Minebea
13:	Fujitsu	Ikegami	Sanyo	Kenwood	Ikegami	Oki	Twinbird
14:	Canon Elec.	Toshiba	Omron	Yokogawa	Shinko	Fuji Elec.	Daido Signal
15:	Nagoya	Fujitsu Component	Canon	Takaoka Elec.	Optex	Fanuc Ltd.	Omron
16:	ADTEC	Sony	Nidec	Victor	ENPLAS	Elpida	Toyo Denki
17:	Murata	Noda Screen	Victor	Ricoh	D&M	Canon	Shindengen
18:	HDK	JRC	Tietech	Oki	Casio	Koito	SPC
19:	NEC	SPC	Kyocera	D&M	Shindengen	JEOL	Meidensha
20:	Victor	Epson Toyocom	Casio	Keyence	FB	Clarion	ETA Elec.

Table 7.5: Top 20 companies ranked by betweenness centrality on different social networks in the electrical industry.

$\mathbf{r}_{i,Cb}$	$\mathbf{r}_{cooc,Cb}$	$\mathbf{r}_{overlap,Cb}$	$\mathbf{r}_{business,Cb}$	$\mathbf{r}_{capital,Cb}$	$\mathbf{r}_{market,Cb}$	$\mathbf{r}_{shareholder,Cb}$	$\mathbf{r}_{age,Cb}$
1:	NIEC	Shindengen	Hitachi	Hitachi	Shinko	Hitachi	Konica Minolta
2:	JEOL	Keyence	Fujitsu	Suzuki	Eneserve	Fujitsu	Brother
3:	Toshiba	Casio	Mitsubishi	Mitsubishi	Konica Minolta	Mitsubishi	NetIndex
4:	Sony	HDK	Omron	Fujitsu	Hitachi	Panasonic	Sanyo Denki
5:	Fujitsu	JAE	Sharp	Nidec	Ibiden	Panasonic Elec.	Minebea
6:	ADTEC	JEOL	Panasonic	Toshiba	Nishishiba Elec.	Elpida	Hitachi
7:	Mitsubishi	Murata	Suzuki	NEC	Brother	Clarion	Mitsubishi
8:	Chuo Seisakusho	Toshiba	Oki	Takaoka Elec.	Noda Screen	ALOKA	Daiichi Seiko
9:	JEM	Sony	Nidec	Canon	Energy Support	Japan Radio	Shinko
10:	NEC	Mitsubishi	NEC	Sanyo	Showa KDE	JEOL	Ibiden
11:	Panasonic	Pulstec	Sony	Panasonic Elec.	Toyo Elec.	Toshiba	Morio Denki
12:	Pioneer	Kenwood	Toshiba	Kenwood	Tabuchi Elec.	Sharp	Nidec
13:	Panasonic Elec.	Real Vision	Sanyo	Oki	Sophia	Koito	Showa KDE
14:	JDL	NEC	Kenwood	Sharp	NIEC	Hitachi Medical	Syswave
15:	Nagoya	Hitachi Medical	Pioneer	Yokogawa	Ferrotec	Nidec	MCJ
16:	Sharp	Fujitsu	Keyence	Ricoh	Shibaura	Tabuchi Elec.	Origin Elec.
17:	Japan Radio	Suzuki	Panasonic Elec.	Panasonic	Santec	Canon	Sophia
18:	Canon Elec.	SEIWA	Toko.	Casio	Hamamatsu	NIEC	NIEC
19:	I-O Data	ADTEC	Maxell	Brother	Mimaki	Oki	Ferrotec
20:	Canon	JEM	Japan Radio	Omron	Enomoto	TDK	Shibaura

Table 7.6: Top 20 companies ranked by attributes of companies in the electrical industry.

$\mathbf{r}_A$	Capital	Employee Number	Sales	PER	PBR	ROA	ROE	HitNum
1:	Sony	Hitachi	Panasonic	ENPLAS	Sanyo	AXELL	Nagano JRC	NEC
2:	NEC	Panasonic	Sony	Santec	Meisei	Keyence	KKDI-Nikko Eng.	Sony
3:	Fujitsu	Toshiba	Toshiba	FDK	Tokki	OptexFA	TEAC	Suzuki
4:	Sanyo	Sony	Fujitsu	SK Elec.	TEAC	Canon Elec.	AXELL	Toshiba
5:	Hitachi	Fujitsu	Canon	NEC	Wacom	TEL	Yaskawa	Sharp
6:	Toshiba	NEC	Sharp	Sanko	Ibiden	Lasertec	Tabuchi Elec.	Fujitsu
7:	Panasonic	Canon	Hitachi	Fujitsu General	SPC	Roland DG	Fujitsu Component	Pioneer
8:	Sharp	Mitsubishi	Mitsubishi	Seiko Giken	Japan Servo	Hioki E.E.	Canon Elec.	Canon
9:	Mitsubishi	Sanyo	NEC	ALPS	AXELL	Wacom	Shinko	Mitsubishi
10:	Canon	Nidec	Sanyo	Maxell	Roland DG	Cosel	Konica Minolta	Ricoh
11:	Elpida	Seiko Epson	Ricoh	Sony	Yaskawa	Ibiden	TEL	Hitachi
12:	Panasonic Elec.	Ricoh	Panasonic Elec.	Anritsu	Nidec	Nidec-READ	Epson Toyocom	Panasonic
13:	Ricoh	Kyocera	TEL	Hosiden	Nagano JRC	Fanuc Ltd.	NEC	Kyocera
14:	Kyocera	TDK	Seiko Epson	UNIPULSE	Tabuchi Elec.	Optex	Lasertec	Sanyo
15:	Rohm	Panasonic Elec.	NEC Elec.	Raytex	Sysmex	Syswave	Ibiden	Panasonic Elec.
16:	NEC Elec.	Minebea	Pioneer	Iwasaki	Mimaki	JEM	Wacom	Omron
17:	Oki	Sharp	Kyocera	Miyakoshi	Keyence	Canon	Elpida	AXELL
18:	Murata	Mabuchi Motor	Murata	Wacom	Foster Elec.	CCS	Nidec-READ	Yamatake
19:	Fanuc Ltd.	Mitsumi Elec.	Casio	ETA Elec.	Micronics Japan	Noda Screen	Terasaki	TDK
20:	Minebea	Pioneer	Elpida	Nishishiba Elec.	Hamamatsu	Techno Medica	Mimaki	KEL

Table 7.7: Results of feature integration in the electrical industry.

Electrical	Feature	Avg-In		Excellent		Market-Cap	
		Train	Test	Train	Test	Train	Test
Network	G_{age}	0.357	0.341	0.443	-0.107	0.361	0.233
	G_{cooc}	0.247	0.120	0.364	0.619	0.346	0.197
	G_{market}	0.535	0.475	0.425	0.357	0.761	0.651
	$G_{overlap}$	0.409	0.284	0.423	0.381	0.519	0.295
	$G_{shareholder}$	0.397	0.190	0.771	0.400	0.514	0.117
	$G_{business}$	0.501	0.182	0.699	-0.500	0.590	0.421
	$G_{capital}$	0.641	0.329	0.818	0.300	0.643	0.350
	G_{ALL}	0.758	**0.584**	0.912	**0.574**	0.685	**0.645**
Attributes	ALL	0.559	**0.389**	0.811	**0.571**	0.735	**0.718**
Network + Attributes	G_{age}+A	0.681	0.573	0.762	0.429	0.791	0.710
	G_{cooc}+A	0.572	0.396	0.804	0.429	0.725	0.700
	G_{market}+A	0.643	0.555	0.746	0.595	0.808	0.754
	$G_{overlap}$+A	0.604	0.418	0.631	0.452	0.745	0.655
	$G_{shareholder}$+A	0.580	0.438	0.739	0.456	0.764	0.625
	$G_{business}$+A	0.596	0.396	0.873	0.619	0.747	0.692
	$G_{capital}$+A	0.592	0.470	0.811	0.524	0.752	0.705
	G_{ALL}+A	0.812	**0.601**	0.947	**0.580**	0.811	**0.756**

Table 7.8: Effective features in various networks for Avg-In, Market=Cap, and Excellent, respectively, among companies.

Top	Features for Avg-In	Features for Market-Cap	Features for Excellent
1	$Max \circ \gamma \circ C_x^{(1)} \circ G_{overlap}$	$Ratio \circ (Sum \circ u_x \circ C_x^{(1)}, Sum \circ u_x \circ C_x^{(\infty)}) \circ G_{overlap}$	$Min \circ s^{(1)} \circ C_x^{(\infty)} \circ G_{business}$
2	$Min \circ s^{(1)} \circ C_x^{(\infty)} \circ G_{business}$	$Min \circ s^{(1)} \circ C_x^{(\infty)} \circ G_{shareholder}$	$Max \circ s^{(1)} \circ C_x^{(1)} \circ G_{business}$
3	$Avg \circ u_x \circ C_x^{(1)} \circ G_{capital}$	$Avg \circ \gamma \circ C_x^{(\infty)} \circ G_{business}$	$Ratio \circ (Max \circ s^{(1)} \circ C_x^{(1)}, Max \circ s^{(1)} \circ C_x^{(\infty)}) \circ G_{business}$
4	$Max \circ \gamma \circ C_x^{(\infty)} \circ G_{age}$	$Max \circ \gamma \circ C_x^{(\infty)} \circ G_{business}$	$Avg \circ t_x \circ C_x^{(1)} \circ G_{capital}$
5	$Avg \circ \gamma \circ C_x^{(\infty)} \circ G_{capital}$	$Min \circ \gamma \circ C_x^{(\infty)} \circ G_{business}$	$Max \circ t_x \circ C_x^{(1)} \circ G_{capital}$
6	$Ratio \circ (Sum \circ u_x \circ C_x^{(1)}, Sum \circ u_x \circ C_x^{(\infty)}) \circ G_{age}$	$Ave \circ s^{(1)} \circ C_x^{(1)} \circ G_{cooc}$	$Min \circ t_x \circ C_x^{(1)} \circ G_{capital}$
7	$Max \circ \gamma \circ C_x^{(\infty)} \circ G_{market}$	$Max \circ t \circ C_x^{(1)} \circ G_{market}$	$Min \circ t_x \circ C_x^{(\infty)} \circ G_{capital}$
8	$Ave \circ s^{(1)} \circ C_x^{(1)} \circ G_{business}$	$Max \circ \gamma \circ C_x^{(\infty)} \circ G_{market}$	$Ratio \circ (Min \circ t_x \circ C_x^{(1)}, Min \circ t_x \circ C_x^{(\infty)}) \circ G_{capital}$
9	$Avg \circ u_x \circ C_x^{(\infty)} \circ G_{capital}$	$Avg \circ \gamma \circ C_x^{(\infty)} \circ G_{shareholder}$	$Max \circ t_x \circ C_x^{(\infty)} \circ G_{capital}$
10	$Avg \circ u_x \circ C_x^{(\infty)} \circ G_{age}$	$Sum \circ t \circ C_x^{(\infty)} \circ G_{shareholder}$	$Max \circ s^{(1)} \circ C_x^{(\infty)} \circ G_{capital}$
11	$Min \circ \gamma \circ C_x^{(1)} \circ G_{cooc}$	$Avg \circ \gamma \circ C_x^{(\infty)} \circ G_{capital}$	$Avg \circ t_x \circ C_x^{(1)} \circ G_{cooc}$
12	$Sum \circ \gamma \circ C_x^{(1)} \circ G_{cooc}$	$Max \circ t \circ C_x^{(\infty)} \circ G_{capital}$	$Max \circ t_x \circ C_x^{(1)} \circ G_{cooc}$
13	$Min \circ \gamma \circ C_x^{(\infty)} \circ G_{business}$	$Avg \circ u_x \circ C_x^{(\infty)} \circ G_{age}$	$Max \circ s^{(1)} \circ C_x^{(\infty)} \circ G_{cooc}$
14	$Ratio \circ (Avg \circ s^{(1)} \circ C_x^{(1)}, Avg \circ s^{(1)} \circ C_x^{(\infty)}) \circ G_{business}$	$Min \circ s^{(1)} \circ C_x^{(1)} \circ G_{age}$	$Max \circ \gamma \circ C_x^{(1)} \circ G_{cooc}$
15	$Max \circ t \circ C_x^{(1)} \circ G_{age}$	$Min \circ \gamma \circ C_x^{(\infty)} \circ G_{age}$	$Max \circ \gamma \circ C_x^{(\infty)} \circ G_{cooc}$

7.5.3 Detailed Analysis of Useful Features

We use network-based features separately to train and expect the target rankings to clarify their usefulness. Leaving out one feature, the others are used to train and predict the rankings to evaluate network-based features. In fact, k-th feature is a useful feature for explaining the target ranking if the result worsens much when leaving out the feature k from the feature set. Table 7.8 presents the effective features for the different target rankings of Market-Cap, and Excellent, respectively, in company networks. For example, the maximum number of links in the neighbor nodeset of x from overlap network $Max \circ \gamma \circ C_x^{(1)} \circ G_{overlap}$ is effective for the target ranking of Avg-In, which means that if a famous company is reachable from a company, the company's income can be more high. The ratio of the sum of paths through x among neighbors to the sum of paths through x among reachable nodes from overlap network $Ratio \circ (Sum \circ u_x \circ C_x^{(1)}, Sum \circ u_x \circ C_x^{(\infty)}) \circ G_{overlap}$ is effective for the target ranking of Market-Value, which means that maintaining high betweenness among neighbors from all of reachable nodes in the Web makes the company' market value higher. The minimum number of edges among reachable companies from the business-alliance network $Min \circ s^{(1)} \circ C_x^{(\infty)} \circ G_{business}$ is effective for the target ranking of Excellent, which means that x will be more excellent when the reachable companies have little business-alliance among them.

We understand that various features have been shown to be important for real-world rankings (i.e. target ranking). Some of them correspond to well-known indices in social network analysis. Some indices seem new, but their meanings resemble those of the existing indices. The results support the usefulness of the indices that are commonly used in the social network literature, and underscore the potential for additional composition of useful features.

Summary:

Several conclusions are suggested by the experimental results presented above: Social networks vary according to different relational indices or types even though they contain the same list of companies; Companies have different centrality rankings even though they are in the same type of relational networks: Relations and networks of different types differently impact on different targets of rankings: Multi-relational networks have more information than single networks to explain target rankings. Well-chosen attribute-based features have good performance for explaining target rankings. However, by combining proposed network-based features, the prediction results are further improved: various network-based features have been shown to be important for real-world rankings (i.e., target ranking), some of which correspond to well-known indices in social network analysis such as degree centrality, closeness centrality, and betweenness centrality. Some indices seem new, but their meanings resemble those of the existing indices.

7.6 Case Study 2: Ranking Researchers using Social Networks

7.6.1 Datasets

We extract social networks for researchers (253 professors of The University of Tokyo) to learn and predict the ranking of researchers. We use the ranking by the number of publications (designated as Paper) as a target ranking, as presented in Table 7.9. Academic papers are often the product of several researchers' collaboration. Therefore, a good position in a social network is derived through good performance. Is there any relation that is important to predict productivity?

We construct social networks among researchers from the Web using a general search engine. We detail the co-occurrence-based approach (Section 6.3.1) to extract co-occurrence-based networks of two kinds in English-language Web sites and Japanese Web sites respectively: cooc network (G_{Ecooc}, G_{Jcooc}) and overlap network ($G_{Eoverlap}$, $G_{Joverlap}$). Actually, we used English/romanized names of researchers as a query to obtain co-occurrence information for G_{Ecooc} and $G_{Eoverlap}$, and used Japanese names of researchers as a query to obtain co-occurrence information for G_{Jcooc} and $G_{Joverlap}$. Then, based on Web co-occurrence networks (in Japanese Web sites), we use the context of Web pages retrieved by two names of persons to classify the relations using C4.5 as a classifier (details presented in [70]). We use Jaccard network constructed by above approach, then classify the edges into relational networks of two kinds: a co-affiliation network ($G_{affiliation}$) and a co-project network ($G_{project}$). Extracted networks for 253 researchers are portrayed in Fig. 7.4.

For this experiment, we also use two types of researchers attributes: the number of hits on Japanese Web sites JhitNum (using Japanese names as a query) and the number of hits on the English-language Web sites EhitNum) (using English/romanized names as a query).

In our experiments, we conducted three-fold cross-validation. In each trial, two folds of actors are used for training, and one fold for prediction. The results we report in this section are those averaged over three trials. We use Spearman's rank correlation coefficient (ρ) [96] to measure the pairwise ranking correlation.

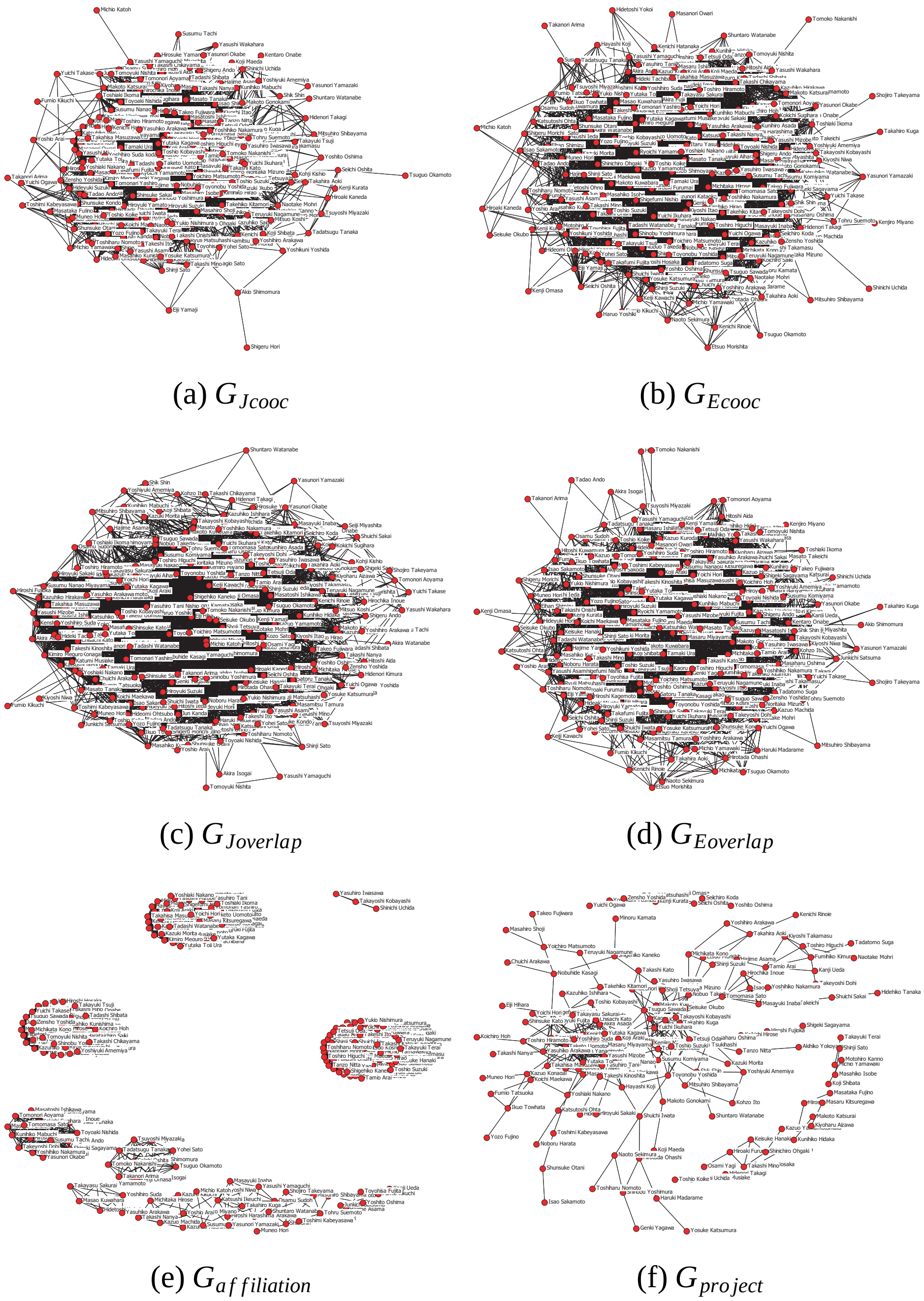
$$\rho = 1 - \frac{6\Sigma_i^2}{n(n^2 - 1)} \tag{7.8}$$

In that equation, d_i is the difference between the ranks of corresponding values X_i and Y_i.

7.6.2 Ranking Results

First, we rank researchers on different network rankings. Table 7.10 presents the degree centrality rankings of different types of networks in researcher networks. Results show that *Yutaka Kagawa* has good degree centrality on a cooc network of Japanese Web sites G_{Jcooc} and that a co-affiliation network $G_{affiliation}$, and *Masaru Kitsuregawa* has stable centralities on several networks.

For the baseline model, three centrality indices (degree centrality C_d, closeness centrality C_c, and betweenness centrality C_b) are used on different networks (G_{Ecooc}, $G_{Eoverlap}$, G_{Jcooc}, $G_{Joverlap}$, $G_{affiliation}$, and $G_{project}$) as network rankings. We calculate the correlation between network rankings with each target ranking of Paper. For comparison, we also rank companies according to previously described attributes (i.e., JhitNum and EhitNum), and take correlation with target ranking. Fig. 7.5 portrays correlations (mean of three tries) of each network rankings as well as each attribute-based rankings with different target rankings on training and testing data among researchers. Results show that the hit number of names on Japanese Web sites is a good attribute of researchers for predicting the creditability of publications. Furthermore, degree centralities in overlap network as well as in cooc network on English-language Web sites ($\mathbf{r}_{G_{Eoverlap},C_d}$ and $\mathbf{r}_{G_{Ecooc},Cd}$) exhibit a good correlation with target ranking. We can say that researchers who are famous on Japanese Web sites and who frequently co-occur with other researchers on English-language Web sites are the more creative researchers.



(a) G_{Jcooc}

(b) G_{Ecooc}

(c) $G_{Joverlap}$

(d) $G_{Eoverlap}$

(e) $G_{affiliation}$

(f) $G_{project}$

Figure 7.4: Web-based social networks for researchers with different relational indices or types.

Table 7.9: Ranking of the number of pages for the top 50 researchers of The University of Tokyo.

r*	Name	**r***	Name
1:	Yasuhiko Arakawa	26:	Kazuhiko Saigo
2:	Kazunori Kataoka	27:	Tadatomo Suga
3:	Kohji Kishio	28:	Tamio Arai
4:	Yuichi Ikuhara	29:	Akira Isogai
5:	Kazuhiko Ishihara	30:	Ryoichi Yamamoto
6:	Yasuhiro Iwasawa	31:	Takayasu Sakurai
7:	Genki Yagawa	32:	Michio Yamawaki
8:	Kazuhito Hashimoto	33:	Hiroshi Harashima
9:	Hiroyuki Sakaki	34:	Takayoshi Kobayashi
10:	Hideki Imai	35:	Fumio Tatsuoka
11:	Masaharu Oshima	36:	Takehiko Kitamori
12:	Kazuyuki Aihara	37:	Teruyuki Nagamune
13:	Kazuro Kikuchi	38:	Masahiko Isobe
14:	Yoshiaki Nakano	39:	Motohiro Kanno
15:	Shinichi Uchida	40:	Kazuo Hotate
16:	Hidenori Takagi	41:	Mitsuhiro Shibayama
17:	Hiroyuki Fujita	42:	Hajime Asama
18:	Katsushi Ikeuchi	43:	Satoru Tanaka
19:	Yutaka Kagawa	44:	Isao Shimoyama
20:	Nobuo Takeda	45:	Yozo Fujino
21:	Masaru Miyayama	46:	Takayuki Terai
22:	Toshiro Higuchi	47:	Yoichiro Matsumoto
23:	Tsuguo Sawada	48:	Nobuhide Kasagi
24:	Kiyoharu Aizawa	49:	Yoshiyuki Amemiya
25:	Kimihiko Hirao	50:	Kunihiro Asada

Table 7.10: Top 20 researchers ranked by degree centrality on different social networks.

$\mathbf{r}_{i,Cd}$	$\mathbf{r}_{Ecooc,Cd}$	$\mathbf{r}_{Eoverlap,Cd}$	$\mathbf{r}_{Jcooc,Cd}$	$\mathbf{r}_{Joverlap,Cd}$	$\mathbf{r}_{affiliation,Cd}$	$\mathbf{r}_{project,Cd}$
1:	Masatoshi Ishikawa	Yoshiaki Nakano	Yutaka Kagawa	Shoji Tetsuya	Koichi Maekawa	Yutaka Kagawa
2:	Yasuhiko Arakawa	Hiroyuki Fujita	Masatoshi Ishikawa	Haruo Yoshiki	Michio Yamawaki	Kazuhiko Hirakawa
3:	Masaru Kitsuregawa	Susumu Tachi	Masaru Kitsuregawa	Yasuhiro Tani	Keiji Kawachi	Tsuguo Sawada
4:	Genki Yagawa	Akira Watanabe	Yasuhiko Arakawa	Shigefumi Nishio	Ikuo Towhata	Masanori Owari
5:	Yutaka Kagawa	Masataka Fujino	Tsuguo Sawada	Michikata Kono	Genki Yagawa	Masao Kuwahara
6:	Yasuhiro Iwasawa	Koji Maeda	Yasuhiro Iwasawa	Seisuke Okubo	Hitoshi Kuwamura	Yasuhiko Arakawa
7:	Masao Kuwahara	Kunihiko Mabuchi	Keiji Kawachi	Michio Katoh	Yoshihiro Arakawa	Makoto Kuwabara
8:	Kiyoharu Aizawa	Yasushi Mizobe	Makoto Kuwabara	Shigehiko Kaneko	Shuichi Iwata	Takahisa Masuzawa
9:	Takahisa Masuzawa	Isao Shimoyama	Genki Yagawa	Akio Shimomura	Makoto Kuwabara	Koji Araki
10:	Toshimi Kabeyasawa	Kazuro Kikuchi	Masao Kuwahara	Koji Araki	Takeo Fujiwara	Hidetoshi Yokoi
11:	Koichi Maekawa	Taketo Uomoto	Kazuhiko Hirakawa	Minoru Kamata	Takahisa Masuzawa	Shuichi Iwata
12:	Takeo Fujiwara	Takeshi Kinoshita	Takahisa Masuzawa	Hideaki Miyata	Kazuhiko Hirakawa	Jun Yanagimoto
13:	Yuichi Ogawa	Yoichi Hori	Masanori Owari	Tomoko Nakanishi	Yutaka Kagawa	Yasushi Mizobe
14:	Shuichi Iwata	Tamaki Ura	Takeo Fujiwara	Hiroshi Hosaka	Masaru Kitsuregawa	Ikuo Towhata
15:	Makoto Kuwabara	Kazuyuki Aihara	Kiyoharu Aizawa	Hitoshi Kuwamura	Kiyoharu Aizawa	Taketo Uomoto
16:	Tsuguo Sawada	Chisachi Kato	Chuichi Arakawa	Eiji Hihara	Tsuguo Sawada	Koichi Maekawa
17:	Kazuhiko Hirakawa	Kenshiro Takagi	Shuichi Iwata	Yutaka Toi	Masatoshi Ishikawa	Kenichi Hatanaka
18:	Ikuo Towhata	Kohji Kishio	Koichi Maekawa	Yutaka Kagawa	Takayuki Terai	Susumu Nanao
19:	Chuichi Arakawa	Takayasu Sakurai	Ikuo Towhata	Tomonari Yashiro	Shigeru Morichi	Yasuhiro Iwasawa
20:	Yoshihiro Arakawa	Hideyuki Suzuki	Hitoshi Kuwamura	Kenichi Hatanaka	Noritaka Mizuno	Yoshihiro Arakawa

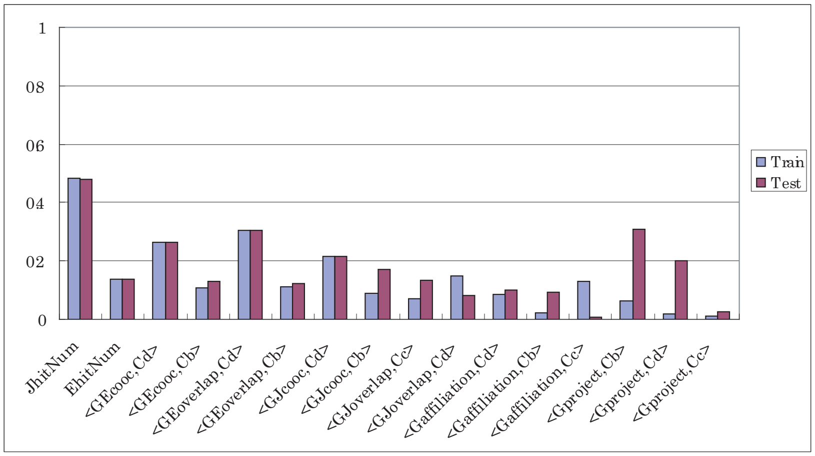


Figure 7.5: Evaluation for each attribute-based ranking as well as centrality-based ranking with target ranking among researchers.

In the combination model, we also use Boolean type ($w_i \in \{1,0\}$) to combine the relations. Using relations of six types to combine a network $G_{affiliation-Ecooc-Eoverlap-Jcooc-Joverlap-project}$, we can create $2^6 - 1$ (=63) types of combination-relational networks (in which at least one type of relation exists). We obtain network rankings in this combined network to learn and predict the target rankings. The top 50 correlations between network rankings in combined-relational network and target rankings are portrayed in Fig. 7.6. Results show that degree centralities on combined-relational network produce good correlation with target rankings. For instance, combining cooc relations on English-language Web sites with co-project relations ($G_{0-1-0-0-0-1}$), or combining cooc relation and overlap relations on English-language Web sites with cooc relation on Japanese Web sites ($G_{0-1-1-1-0-0}$) makes the networks more reasonable for predicting a target ranking.

We execute our feature integration ranking model (with several varies) to single and multi-relational social networks to train and predict rankings of researchers' Paper. We use Ranking SVM to learn the ranking model which minimize pairwise training error in the training data. Then we apply the model to predict the rankings on training data (again) and on testing data. Comparable results on several varies of model are presented in Table 7.11. First, we integrate attribute-indices (i.e., hit number of names on the Japanese Web sites and on the English-language Web sites) of researchers as features as a baseline of this model

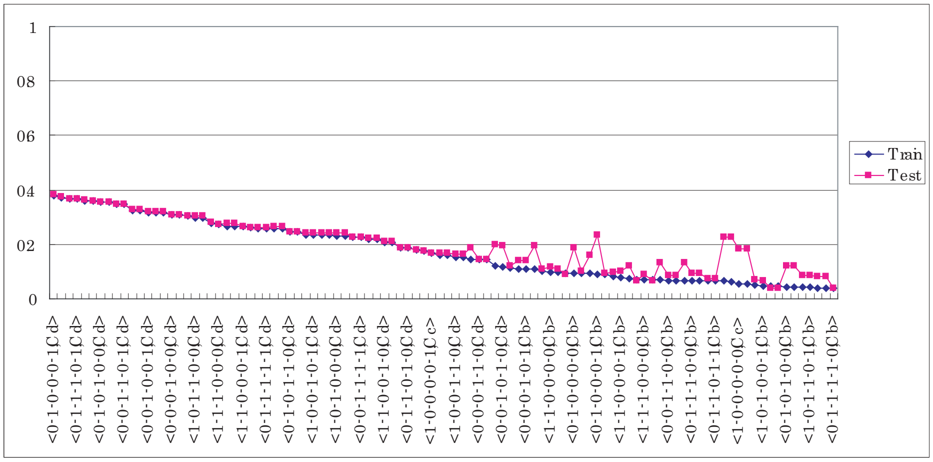


Figure 7.6: Evaluation for network rankings in combined-relational network with Paper among researchers.

to learn and predict the rankings. We can obtain a 0.448 correlation coefficient between predicted rankings and target rankings, which is explainable: famous researchers are also famous on the Web sites. Then, we integrate proposed network-based features obtained from each type of single network as well as multi-relational networks among researchers to train and predict the rankings. The co-occurrence-based networks G_{Ecooc}, $G_{Eoverlap}$, $G_{Joverlap}$ (especially on English-language Web sites) appear to be a better explanation of target ranking of Paper than the co-affiliation network $G_{affiliation}$ or the co-project network $G_{projext}$. Using features from multi-relational networks G_{ALL}, the prediction results are better than for any other single-relational network. Furthermore, when we combine network-based features with attribute-based features to learn the model, the results outperform each using attribute-based features only and network-based features only.

7.6.3 Detailed Analysis of Useful Features

We use network-based features separately to train and expect the target rankings to clarify their usefulness. Leaving out one feature, the others are used to train and predict the rankings to evaluate network-based features. In fact, k-th feature is a useful feature for explaining the target ranking if the result worsens much when leaving out the feature k. Table 7.12 presents

Table 7.11: Results of feature integration among researchers.

Professor	Feature	PaperNum	
		Train	Test
Network	G_{Ecooc}	0.470	0.413
	$G_{Eoverlap}$	0.508	0.411
	G_{Jcooc}	0.443	0.261
	$G_{Joverlap}$	0.585	0.325
	$G_{affiliation}$	0.178	-0.011
	$G_{project}$	0.540	0.043
	G_{ALL}	0.821	**0.417**
Attributes	ALL	0.491	**0.448**
Network + Attributes	G_{Ecooc}+A	0.514	0.429
	$G_{Eoverlap}$+A	0.544	0.404
	G_{Jcooc}+A	0.481	0.284
	$G_{Joverlap}$+A	0.519	0.420
	$G_{affiliation}$+A	0.497	0.159
	$G_{project}$+A	0.548	0.304
	G_{ALL}+A	0.811	**0.456**

the effective features for the target ranking of Paper in the researcher field. For example, the maximum number of links in the reachable nodeset of x from cooc network from English-language Web sites $Max \circ \gamma \circ C_x^{(\infty)} \circ G_{Ecooc}$ is effective for the target ranking, which means that if a famous researcher is reachable from a person, that person can be more productive. The minimum number of links in the neighbor nodeset of x from the cooc network from Japanese Web sites $Min \circ \gamma \circ C_x^{(1)} \circ G_{Jcooc}$ is also effective, which means that if a direct neighbor is productive, then x will be more productive. The ratio of the number of edges among neighbors to the number of edges among reachable nodes from co-project network $Ratio \circ (Sum \circ s^{(1)} \circ C_x^{(1)}, Sum \circ s^{(1)} \circ C_x^{(\infty)}) \circ G_{project}$ means that binding neighbors from all of reachable nodes in projects makes the researcher more productive.

We understand that various features have been shown to be important for real-world

Table 7.12: Effective features in various networks for Paper among researchers.

Top	Effective Features for Paper
1	$Max \circ \gamma \circ C_x^{(\infty)} \circ G_{Ecooc}$
2	$Min \circ \gamma \circ C_x^{(1)} \circ G_{Jcooc}$
3	$Avg \circ \gamma \circ C_x^{(\infty)} \circ G_{Eoverlap}$
4	$Max \circ t \circ C_x^{(\infty)} \circ G_{Joverlap}$
5	$Avg \circ u_x \circ C_x^{(1)} \circ G_{Eoverlap}$
6	$Min \circ \gamma \circ C_x^{(1)} \circ G_{Eoverlap}$
7	$Min \circ \gamma \circ C_x^{(\infty)} \circ G_{Jcooc}$
8	$Ratio \circ (Sum \circ s^{(1)} \circ C_x^{(1)}, Sum \circ s^{(1)} \circ C_x^{(\infty)}) \circ G_{project}$
9	$Avg \circ \gamma \circ C_x^{(1)} \circ G_{Joverlap}$
10	$Min \circ \gamma \circ C_x^{(1)} \circ G_{Ecooc}$
11	$Ratio \circ (Sum \circ s^{(1)} \circ C_x^{(1)}, Sum \circ s^{(1)} \circ C_x^{(\infty)}) \circ G_{Ecooc}$
12	$Ratio \circ (Sum \circ u_x \circ C_x^{(1)}, Sum \circ u_x \circ C_x^{(\infty)}) \circ G_{Ecooc}$
13	$Min \circ u_x \circ C_x^{(1)} \circ G_{Jcooc}$
14	$Ratio \circ (Avg \circ u_x \circ C_x^{(1)}, Avg \circ u_x \circ C_x^{(\infty)}) \circ G_{Jcooc}$
15	$Min \circ \gamma \circ C_x^{(\infty)} \circ G_{Joverlap}$

rankings (i.e. target ranking). Some of them correspond to well-known indices in social network analysis: degree centrality, closeness centrality, and betweenness centrality. Some indices seem new, but their meanings resemble those of the existing indices. The results support the usefulness of the indices that are commonly used in the social network literature, and underscore the potential for additional composition of useful features.

Summary:

Social networks vary according to different relational indices or types even though they contain the same list of researchers; Researchers have different centrality rankings even though they are in the same type of relational networks: Multi-relational networks have more information than single networks to explain target rankings. Well-chosen attribute-based features have good performance for explaining target rankings. However, by combining proposed network-based features, the prediction results are further improved: various network-based features have been shown to be important for real-world rankings (i.e., target ranking), some of which correspond to well-known indices in social network analysis such as degree centrality, closeness centrality, and betweenness centrality. Some indices seem new, but their meanings resemble those of the existing indices.

7.7 Related Works

Recently, many studies deal with social networks among various online resources such as social network services (SNSs) [115], online Instance Messengers (IM) [93], as well as Friend-of-a-Friend (FOAF) instances [31, 35]. Unfortunately, these resources are not specifically applicable to relations among companies or other organization structures. However, many relations among companies are published on the Web in news articles or news releases. Our work emphasizes the investigation of such published relations on the Web. A news site might deal little with information related to small companies and foreign corporations. Therefore, we use a search engine to extract more coverable relations among any given set of companies.

The location of actors in multi-relational networks and the structure of networks composed of multiple relations are interesting areas of SNAs. Recent efforts to address this problem adopt consideration of multi-modal networks—a network composed of a set of different kind of nodes—and mainly consider the relations among these nodes [115, 81, 89]. They usually use papers, authors, and conferences (or journals) as different types of nodes, and considering the relational impact from different models (or layers) paper-paper, paper-author, as well as paper-conference (or journal) relations to calculate document similarity for document recommendation as well as support the scholarly communication process. This paper presents different views of multi-relational networks comprising multiple different kinds of relations (ties) among the same set of social actors (nodes) to elucidate what kinds of relations are important, as well as what kinds of relational combinations are important.

In the context of information retrieval, PageRank [84] and HITS [54] algorithms can be considered as well known examples for ranking Web pages based on the link structure. Recently, more advanced algorithms have been proposed for ranking entities. Several studies have examined learning certain relational weights as conductance of Markovian walks on a network, given preference orders over nodes using gradient descent [24], error backpropagation [30], and approximate Newton method [23]. Our networks are social networks with connections among nodes according to relations. Therefore, we neither give assumptions that the network must be a Markov network nor that the weight is positive only (because negative relations such as a lawsuit relation might damage the company ranking). Furthermore, our model is target-dependent: the important features of relations and structural embeddedness vary among different tasks.

Relations and structural embeddedness influence behavior of individuals and growth and change of the group [93]. Several researchers use network-based features for analyses. L. Backstrom et al. [9] describe analyses of community evolution, and show some structural features characterizing individuals ’positions in the network. D. Liben-Nowell et al. [60] elucidate features using network structures for link prediction in the link prediction problem. We specifically examine relations and structural features for individuals and deal with various structural features from multi-relational networks systematically. Our generated features include those described in works from Backstrom and Liben-Nowell.

Our approaches are similar to text classification given the document features and correct categories. Features are designed beforehand. Similarly, the relation is defined beforehand. The classifier learns the model to predict the given categories. Similarly, the ranking is given and is used for learning. Specifically regarding feature weights, we can understand which features are important for categorization, thereby yielding a better classification model. Furthermore, examining the weights of each relation, we can understand which relations are important for ranking. Cai et al. [21] regarded a similar idea with this approach: They try to identify the best combination of relations (i.e., relations as features) which makes the relation between the intra-community examples as tight as possible. Simultaneously, the relation between the inter-community examples is as loose as possible when a user provides multiple community examples (e.g., two groups of researchers). However, our purpose is learning of a ranking function (e.g.,, ranking of companies) based on social networks, which has a different optimization task. Moreover, we propose innovative features for entities based on the combination or integration of structural importance generated from social networks. However, our purpose is learn the ranking function (e.g. ranking of companies) based on social networks, which has different optimization task. Moreover, we propose innovative features for entities based on combination or integration of structural importance generated from social networks.

7.8 Conclusion

This chapter described methods of learning the ranking of entities from multiple social networks mined from the Web. Various relations and relational embeddedness pertain to our lives: their combinations and their aggregate impacts are influential to predict features of entities. Based on that intuition, we constructed our ranking learning model from social networks to predict the ranking of other actors. We first extracted social networks of different kinds from the Web. Subsequently, we used these networks and a given target ranking to learn the model. We proposed three approaches to obtaining the ranking model. Results of experiments using two domains (i.e., companies in the electrical industry in Japan and researchers in The University of Tokyo) reveal that effectiveness of our models for explaining target rankings of actors using multiple social networks mined from the Web. Our models provide an example of advanced utilization of a social network mined from the Web. The results underscore the usefulness of our approach, by which we can understand the important relations as well as important structural embeddedness to predict the rankings. We use multiple social networks extracted from the Web, which are more realistic than a single relational network. In addition, the model can be combined with a conventional attribute-based approach. Our model provides an example of advanced utilization of a social network mined from the Web. More kinds of networks and attributes for various target rankings in different domains can be designated for improving the usefulness of our models in the future.

Chapter 8

Conclusion and Future Work

Because of the wealth of information available on the Web, many studies have cast attention on the extraction and application of useful data from the Web. This thesis described novel methods for extracting social networks from the Web using a general search engine. Furthermore, this thesis presented a ranking learning model using social networks mined from the Web. The key features of our approach are using simple algorithms to process huge amounts of information . The extracted social networks are applicable to several applications. We proposed an advanced model for ranking learning from the networks.

Overall, in this thesis, we addressed two research topics for social networks on the Web: social network extraction from the Web and application of those extracted social networks.

Regarding the first topic, we initially defined problems in social network extraction and examined two assumptions and shortcomings in previous studies. To assess the first assumption, we proposed the *relation identification* approach, which enables us to address complex communities. We used companies as instances to extract inter-company networks from the Web using the proposed approach. Given a list of names of companies, our system uses a search engine to collect target pages from the Web. The system then applies text processing to construct a network of companies. 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 particularly addressed alliance and lawsuit relations, in future work, extension of the proposed method to other types of relations among companies will be undertaken. To assess assumption two, we proposed the *threshold tuning* approach, which enables us to address inhomogeneous communities. We used artists (of contemporary art) as instances to extract weak relations among them to construct a social network. The experimental results described herein demonstrate the effectiveness of our approach. Additional characteristics of parameters will be discussed in future reports. The obtained network for artists was operated on the Web site for the Yokohama Triennale 2005. Future studies will support discernment of appropriate parameters for different networks.

For the second topic, we specifically examined the application of a social network that provides an example of advanced utilization of social networks mined from the Web. We described methods of learning the ranking of companies from a social network mined from the Web. Various relations and relational embeddedness apply to our lives: their combinations and their aggregate impacts are influential to predict features of entities. Based on that inference, we constructed our ranking learning model from social networks to predict the ranking of other entities. We first extracted social networks of different kinds from the Web. Then, we used these networks and a given target ranking to discern important relations of ranking indices. We then proposed three approaches to obtain the ranking model. Results of experiments on the field of companies and researchers demonstrated that the important relation depends on the purpose of the target analysis. Our model provides an example of advanced utilization of a social network mined from the Web. The results underscore the usefulness of our approach, by which we can understand the important relations as well as the important structural embeddedness to predict the rankings. We use multi-relational networks extracted from the Web, which are more realistic than single-relational networks. The proposed ranking learning model combines various network features. Moreover, the model can be combined with a conventional attribute-based approach. Our approach suggests an interesting and important direction for advanced Web mining.

Our approach is that of using the Web as a huge resource and using search engines as an interface to obtain information. This thesis expands social network mining from the Web so that is applicable to various domains. Two major improvements are proposed and described: *relation identification* and *threshold tuning*. We presented examples and evaluations for companies' and artists' networks. We provided an example of advanced utilization of a social network mined from the Web. The results emphasize the usefulness of our approach, by which we can understand the important relations as well as the important structural embeddedness to predict features of entities. Furthermore, we used multi-relational networks extracted from the Web, which are more realistic than single-relational networks.

The proposed ranking learning model combines various network features; the model can be combined with a conventional attribute-based approach. Results of this study will provide a bridge between relation extraction and ranking learning for advanced knowledge acquisition for Web Intelligence.

Several tasks can be undertaken in future examinations of this topic.

- We can 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, and addressing tabular data.

- We can extract relations and networks of greater variety from the Web to explain the real-world ranking (i.e. target rankings) more appropriately. For example, this method constructs social networks not only using a search engine, but also from structured or semi-structured relational data such as DBLPs, wikis, and SNSs.

- The extracted networks from the Web can be applied further in various applications such as identifying political cliques and hidden competitive relations etc.

- We can construct a methodology for system construction by applying advanced Web mining techniques. For example, if we input a list of names of companies with target rankings, the system can construct various social networks, which explains the input ranking.

Bibliography

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

[2] L. A. Adamic. The small world web. In *Proc. ECDL'99*, pages 443–452, 1999.

[3] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: Everyone knows something. In *Proc. WWW2008*, 2008.

[4] Eytan Adar, Li Zhang, Lada A. Adamic, and Rajan M. Lukose. Implicit structure and the dynamics of blogspace. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*, 2004.

[5] A. Agarwal, S. Chakrabarti, and S. Aggarwal. Learning to rank networked entities. In *Proc. KDD'06*, 2006.

[6] R.D. Alba. A graph-theoretic definition of a sociometric clique. *Journal of Mathematical Sociology*, pages 113–126, 1973.

[7] B. Aleman-Meza, M. Nagarajan, C. Ramakrishnan, A. Sheth, I. Arpinar, L. Ding, P. Kolari, A. Joshi, and Tim Finin. Semantic analytics on social networks: Experiences in addressing the problem of conflict of interest detection. In *Proc. WWW2006*, 2006.

[8] N. Aswani, K. Bontcheva, and H. Cunningham. Mining information for instance unification. In *Proc. ISWC2006*, 2006.

[9] L. Backstrom, D. Huttenlocher, X. Lan, and J. Kleinberg. Group formation in large social networks: Membership, growth, and evolution. In *Proc. SIGKDD'06*, 2006.

[10] S. Battiston. Inner structure of capital control networks. *Physica A*, 338:107–112, 2004.

[11] M. A. Beauchamp. An improved index of centrality. *Behavioral Sceince*, pages 161–173, 1965.

[12] R. Bekkerman and A. McCallum. Disambiguating web appearances of people in a social network. In *Proc.WWW2005*, Chiba, Japan, 2005.

[13] M. Bengtsson and S. Kock. Cooperation and competition in relationships between competitors in business networks. *Journal of Business & Industrial Marketing*, 14:178–194, 1999.

[14] A. Bernstein, S. Clearwater, S. Hill, C. Perlich, and F. Provost. Discovering knowledge from relational data extracted from business news. In *SIGKDD-2002 Workshop on Multi-Relational Data Mining*, 2002.

[15] J. Boissevain. *Friends of friends: Networks, manipulators and coalitions*. Oxford: Basil Blackwell, 1974.

[16] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring semantic similarity between words using web search engines. In *Proc. WWW2007*, 2007.

[17] D. Bollegara, Y. Matsuo, and M. Ishizuka. Disambiguating personal names on the web using automatically extracted key phrases. In *Proc. ECAI 2006*, 2006.

[18] Phillip Bonacichi. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, pages 113–120, 1972.

[19] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proc. 7th WWW Conf.*, 1998.

[20] M. Cafarella and O. Etzioni. A search engine for natural language applications. In *Proc. WWW2005*, 2005.

[21] D. Cai, Z. Shao, X. He, X. Yan, and J. Han. Mining hidden community in heterogeneous social networks. In *LinkKDD ’05*, pages 58–65, 2005.

[22] C. Cardie. Empirical methods in information extraction. *AI Magazine*, 18:99–101, 1997.

[23] S. Chakrabarti and A. Agarwal. Learning parameters in entity relationship graphs from ranking preferences. In *Proc. ECML/PKDD*, volume 4213, pages 91–102, 2006.

[24] H. Chang, D. Cohn, and A. McCallum. Creating customized authority lists. In *Proc. ICML2000*, 2000.

[25] C. Chen. Visualising semantic spaces and author co-citation networks in digital libraries. *Inf.Process.Manage.*, 35(3):401–420, 1999.

[26] H. Chen, M. Lin, and Y. Wei. Novel association measures using web search with double checking. In *Proc. COLING-ACL2006*, pages 1009–1016, 2006.

[27] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In *Proc. WWW2004*, pages 462–471, 2004.

[28] F. Crimmins, A. Smeaton, T. Dkaki, and J. Mothe Tëtrafusion. Information discovery on the internet. *IEEE Inteligent Systems*, 14:55–62, 1999.

[29] A. Culotta, R. Bekkerman, and A. McCallum. Extracting social networks and contact information from email and the web. In *CEAS-1*, 2004.

[30] M. Diligenti, M. Gori, and M. Maggini. Learning web page scores by error back-propagation. In *Proc. IJCAI2005*, 2005.

[31] L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of foaf. *Proceeding of the 38th International Conference On System Sciences, Hawaii*, 2005.

[32] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in knowitall. In *Proc. WWW2004*, 2004.

[33] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. From data mining to knowledge discovery: An overview. *Advances in Knowledge Discovery and Data Mining*, pages 1–34, 1996. AAAI Press.

[34] R. Ferrer and R. V. Sole. The small world of human language. In *Proceedings of The Royal Society of London. Series B, Biological Sciences*, volume 268, pages 2261–2265, 2001.

[35] T. Finin, L. Ding, L. Zhou, and A. Joshi. Social networking on the semantic web. *The Learning Organization*, 12(5):418–435, 2005.

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

[37] F. Gandon, O. Corby, A. Giboin, N. Gronnier, and C. Guigard. Graph-based inferences in a semantic web server for the cartography of competencies in a telecom valley. In *Proc. ISWC2005*, 2005.

[38] L. Getoor and C. P. Diehl. Link mining: A survey. *SIGKDD Explorations*, 2(7), 2005.

[39] S. Ghita, W. Nejdl, and R. Paiu. Semantically rich recommendations in social networks for sharing, exchanging and ranking semantic context. In *Proc. ISWC2005*, 2005.

[40] E. Glover, G. Flake, S. Lawrence, W. P. Birmingham, A. Kruger, C. L. Giles, and D. Pennock. Improving category specific web search by learning query modifications. In *Symposium on Applications and the Internet, SAINT*, pages 23–31, 2001.

[41] J. Golbeck and J. Hendler. Inferring trust relationships in web-based social networks. *ACM Transactions on Internet Technology*, 7(1), 2005.

[42] J. Golbeck and B. Parsia. Trust network-based filtering of aggregated claims. *International Journal of Metadata, Semantics and Ontologies*, 2006.

[43] W.H. Goodenough. *Cognitive Anthropology*, chapter Rethinking “status” and “role”: Toward a general model of the cultural organization of social relationships, pages 311–330. New York: Holt, Rinehart, and Winston, 1969.

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

[45] Info-plosion. http://www.infoplosion.nii.ac.jp/info-plosion/.

[46] Angelo Di Iorio, Fabio Vitali, and Stefano Zacchiroli. Wiki content templating. In *Proc. WWW2008*, 2008.

[47] Y. Jin, Y. Matsuo, and M. Ishizuka. Extracting social networks among various entities on the web. In *ESWC2007*, 2007.

[48] Y. Jing and S. Baluja. Pagerank for product image search. In *Proc. WWW2008*, 2008.

[49] J. Kaiser. It's a small Web after all. *Science*, 285:1815, 1999.

[50] N. Kambhatla. Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. In *Proc. ACL04*, pages 178–181, 2004.

[51] H. Kautz, B. Selman, and M. Shah. The hidden Web. *AI magazine*, 18(2):27–35, 1997.

[52] F. Keller and M. Lapata. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29:459–484, 2003.

[53] L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *Proc. WWW2008*, 2008.

[54] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *Proc. ACM-SIAM Symposium on Discrete Algorithms*, pages 668–677, 1998.

[55] P. Knees, E. Pampalk, and G. Widmer. Artist classification with web-based data. In *Proc. 5th International Conf. on Music Information Retrieval (ISMIR)*, 2004.

[56] M. Kochen. *The Small World*. Ablex Publishing Corporation, New Jersey, 1989.

[57] R. Kosala and H. Blockeel. Web mining research: A survey. *ACM SIGKDD Explorations*, 1(2):1–15, 2000.

[58] X. Li, L. Guo, and Y. E. Zhao. Tag-based social interest discovery. In *Proc. WWW2008*, 2008.

[59] X. Li, P. Morie, and D. Roth. Semantic integration in text: From ambiguous names to identifiable entities. *AI Magazine Spring*, pages 45–68, 2005.

[60] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *Proc. CIKM*, pages 556–559, 2003.

[61] Y. Lin, Y. Chi, S. Zhu, H. Sundaram, and B. L. Tseng. Facetnet: A framework for analyzing communities and their evolutions in dynamic networks. In *Proc. WWW2008*, 2008.

[62] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*. Hardcover, 2007.

[63] Y. Lu and C. Zhai. Opinion integration through semi-supervised topic modeling. In *Proc. WWW2008*, 2008.

[64] R.D. Luce and A.D. Perry. A method of matrix analysis of group structure. *Psychometrika*, pages 94–116, 1949.

[65] P. Maes. Agents that reduce work and information overload. *Communications of the ACM*, (37):30–40, 1994.

[66] C.D. Manning and H. Schütze. *Foundations of statistical natural language processing*. The MIT Press, London, 2002.

[67] P. Marsden. Measuring tie strength. *Social Forces*, 63:482–501, 1984.

[68] P. Massa and P. Avesani. Controversial users demand local trust metrics: an experimental study on epinions.com community. In *Proc. AAAI-05*, 2005.

[69] Y. Matsuo, M. Hamasaki, H. Takeda, J. Mori, D. Bollegala, Y. Nakamura, T. Nishimura, K. Hasida, and M. Ishizuka. Spinning multiple social networks for semantic web. In *Proc. AAAI-06*, 2006.

[70] Y. Matsuo, J. Mori, M. Hamasaki, K. Ishida, T. Nishimura, H. Takeda, K. Hasida, and M. Ishizuka. POLYPHONET: an advanced social network extraction system. In *Proc. WWW2006*, 2006.

[71] Y. Matsuo, H. Tomobe, K. Hasida, and M. Ishizuka. Finding social network for trust calculation. In *Proc. ECAI2004*, pages 510–514, 2004.

[72] J. McCarthy, D. McDonald, S. Soroczak, D. Nguyen, and A. Rashid. Augmenting the social space of an academic conference. In *Proc. CSCW2004*, 2004.

[73] P. Mika. Flink: semantic web technology for the extraction and analysis of social networks. *Journal of Web Semantics*, 3(2):211–223, 2005.

[74] P. Mika. Ontologies are us: A unified model of social networks and semantics. In *Proc. ISWC2005*, 2005.

[75] T. Miki, S. Nomura, and T. Ishida. Semantic web link analysis to discover social relationship in academic communities. *IEEE/IPSJ Symposium on Applications and the Internet (SAINT-05)*, 00(0), 2005.

[76] R.J. Mokken. Cliques, clubs and clans. *Quality and Quantity*, pages 161–173, 1979.

[77] J.L. Moreno. *Who Shall Survive?: Foundations of Sociometry, Group Psychotherapy, and Sociodrama*. Washington, 1934.

[78] J. Mori, M. Ishizuka, T. Sugiyama, and Y. Matsuo. Real-world oriented information sharing using social networks. In *Proc. ACM GROUP2005*, 2005.

[79] I. Muslea. Extraction patterns for information extraction tasks: A survey. In *Proc. AAAI-99 Worlkshop on Machine Learning for Information Extraction*, 1999.

[80] X. Ni, G.-R. Xue, X. Ling, Y. Yu, and Q. Yang. Exploring in the weblog space by detecting informative and affective articles. In *Proc. WWW2007*, 2007.

[81] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: Bringing order to web objects. In *Proc. WWW2005*, 2005.

[82] J. Nieminen. On centrality in a graph. *Scandinavian Journal of Psychology*, pages 332–336, 1974.

[83] S. Oyama, T. Kokubo, and T. Ishida. Domain-specific web search with keyword spices. *IEEE Transactions on Knowledge and Data Engineering*, 16(1):17–27, 2004.

[84] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[85] N. V. Papakyriazis and M. A. Boudourides. Electronic weak ties in network organisations. In *4th GOR Conference*, 2001.

[86] T. Qin, T. Liu, X. Zhang, D. Wang, W. Xiong, and H. Li. Learning to rank relational objects and its application to web search. In *Proc. WWW2008*, 2008.

[87] G. Ramakrishnan, S. Chakrabarti, D. Paranjpe, and P. Bhattacharyya. Is question answering an acquired skill? In *Proc. WWW2004*, 2004.

[88] M. Rege, M. Dong, and J. Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In *Proc. WWW2008*, 2008.

[89] M.A Rodriguez. A multi-relational network to support the scholarly communication process. *International Journal of Public Information Systems*, 2007:ISSN: 1653–4360, 2007.

[90] L.D. Sailer. Structural equivalence: Meaning and definition, computation and application. *Social Networks*, pages 73–90, 1978.

[91] J. Scott. *Social Network Analysis: A Handbook (2nd ed.)*. SAGE publications, 2000.

[92] S.B. Seidman and B.L. Foster. A graph-theoretic generalization of the clique concept. *Journal of Mathematical Sociology*, pages 139–154, 1978.

[93] P. Singla and M. Richardson. Yes, there is a correlation - from social networks to personal behavior on the web. In *Proc. WWW2008*, 2008.

[94] S. Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 34:233–272, 1996.

[95] W. Souma, Y. Fujiwara, and H. Aoyama. Shareholding networks in japan. *Science of Complex Networks: From Biology to the Internet and WWW; CNET 2004, AIP Conference Proc.*, 776:298–307, 2005.

[96] C. Spearman. The proof and measurement of association between two things. *Amer. J. Psychol.*, pages 72–101, 1904.

[97] S. Staab, P. Domingos, P. Mika, J. Golbeck, L. Ding, T. Finin, A. Joshi, A. Nowak, and R. Vallacher. Social networks applied. *IEEE Intelligent Systems*, 20(1):80–93, 2005.

[98] K. Stephenson and M. Zelen. Rethinking centrality: Methods and applications. *Social Networks*, pages 1–37, 1989.

[99] D. Krackhardt T. Rowley, B. Dean. Redundant governance structures: an analysis of structural and relational embeddedness in the steel and semiconductor industries. *Strategic Management Journal*, 21:369–386, 2000.

[100] B. Tan and F. Peng. Unsupervised query segmentation using generative language models and wikipedia. In *Proc. WWW2008*, 2008.

[101] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *Proc. WWW2008*, 2008.

[102] P. Turney. Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Proc. ECML2001*, pages 491–502, 2001.

[103] P. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. In *Proc. ACL02*, pages 417–424, 2002.

[104] P. D. Turney. Measuring semantic similarity by latent relational analysis. In *IJCAI-05*, 2005.

[105] J. Tyler, D. Wikinson, and B. Huberman. *Email as spectroscopy: automated discovery of community structure within organizations*, pages 81–96. Kluwer, B.V., 2003.

[106] B. Uzzi. Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly*, 42:35–67, 1997.

[107] S. Wasserman and K. Faust. *Social network analysis. methods and applications*. Cambridge University Press, Cambridge, 1994.

[108] D. J. Watts. *Six Degrees: The Science of a Connected Age*. Norton, New York, 2003.

[109] B. Wellman. *The Global Village: Internet and Community*, volume 1(1) of *Idea&s - The Arts & Science Review*. University of Toronto, 2004.

[110] F. Wu and D. S. Weld. Automatically refining the wikipedia infobox ontology. In *Proc. WWW2008*, 2008.

[111] J. Yang, L. A. Adamic, and M. S. Ackerman. Crowdsourcing and knowledge sharing: Strategic user behavior on taskcn. In *Proc. EC2008*, 2008.

[112] H. Yeung. The firm as social networks: An organisational perspective. *Growth and Change*, 36:307–328, 2005.

[113] S. Zhao and R. Grishman. Extracting relations with integrated information using kernel methods. In *Proc. ACL 2005*, 2005.

[114] Z. Zhao and H. Liu. Searching for interacting features. In *Proc. IJCAI2007*, 2007.

[115] D. Zhou, S. Zhu, K. Yu, X. Song, B.L. Tseng, H. Zha, and C.L. Giles. Learning multiple graphs for document recommendations. In *Proc. WWW2008*, 2008.

[116] 金光 淳. **社会ネットワーク分析の基礎 –社会的関係資本論にむけて**–. 勁草書店, 2003.

[117] 原田 昌紀, 佐藤 進也, and 風間 一洋. Web 上のキーパーソンの発見と関係の可視化. In **情報処理学会研究報告**, volume DBS-130/FI-71, 2003.

[118] 安田 雪. **社会ネットワーク分析 –何が行為を決定するか**–. 新曜社, 1997.

[119] 安田 雪. **実践ネットワーク分析**. 新曜社, 2001.

[120] 古川 忠延, 松澤 智史, 松尾 豊, 内山 幸樹, and 武田 正之. Weblog におけるユーザの繋がりと閲覧行動の分析. **電子情報通信学会論文誌**, J88-B(7):1258–1266, 2005.

[121] 松尾 豊, 友部 博教, 橋田 浩一, and 石塚 満. Web 上の情報からの人間関係ネットワークの抽出. **人工知能学会論文誌**, 20(1), 2005.

Publications

Journal Papers & Chapters in Books

1. Y. Jin, Y. Matsuo, M. Ishizuka: “Ranking Companies on the Web using Social Network Mining”, Book chapter in I.-H. Ting (Ed.), Web Mining Applications in E-commerce and E-services, Springer–Verlag, Heidelberg Germany, ISBN: 9783540880806, pp. 137–152, 2009

2. Y. Jin, Y. Matsuo, M. Ishizuka: “Extracting Inter-Firm Networks from World Wide Web Using General-Purpose Search Engine”, Journal of Online Information Review, ISSN: 1468-4527, Vol. 32, No. 22, 2008

3. 金英子，松尾豊，石塚満: “Web上の情報を用いた弱い社会的関係のネットワーク抽出手法”，電子情報通信学会論文誌，Vol. J91-D, No. 3 pp. 709–722, 2008

4. 金英子，松尾豊，石塚満: “Web上の情報を用いた企業間関係の抽出”，人工知能学会論文誌，Vol. 22，No. 1，pp. 48–57，2007

Conference & Workshop Publications

1. Y. Jin, Y. Matsuo, M. Ishizuka: “Learning to Rank Entities Based on Multiple Social Networks Mined from the Web”, Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09) (submitted), 2009

2. Y. Jin, Y. Matsuo, M. Ishizuka: “Ranking Entities on the Web using Social Network Mining and Ranking Learning”, WWW2008 Workshop on Social Web Search and Mining (SWSM2008), Beijing, China, 2008

3. Y. Jin, Y. Matsuo, M. Ishizuka: “Extracting Social Networks among Various Entities on the Web”, Proc. 4th European Semantic Web Conf. (ESWC ’07), Innsbruck, Austria, 2007

4. Y. Jin, Y. Matsuo, M. Ishizuka: “Extracting a Social Network among Entities by Web mining”, Web Content Mining with Human Language Technologies, Proc. 5th International Semantic Web Conf. (ISWC’06) Workshop, Athens, GA, USA, 2006

5. Y. Jin, Y. Matsuo, M. Ishizuka: “Extracting Inter-Firm Networks from World Wide Web”, The 9th IEEE International Conference on E-Commerce Technology and The 4th IEEE International Conference on Enterprise Computing, E-Commerce and E-Services (CEC-EEE 2007) pp. 635-642, Tokyo, Japan, 2007

6. H. Prendinger, C. L. Ma, Y. Jin, A. Nakasone, M. Ishizuka: “Understanding the effect of life-like interface agents through users’ eye movements”, 7th International Conference on Multimodal Interfaces (ICMI-05), ACM Press, Trento, Italy, 2005

7. H. Prendinger, C. L. Ma, Y. Jin, K. Kushida, M. Ishizuka: “Evaluating the interaction with synthetic agents using attention and affect tracking”, 4th International Conference on Autonomous Agents and Multi Agent Systems, Utrecht, The Netherlands, 2005

8. 金英子，松尾豊，石塚満：“Webから抽出した社会ネットワークに基づくエンティティランキングの学習”，人工知能学会第22回全国大会，2008

9. 金英子，松尾豊，石塚満：“Webから抽出した企業間ネットワークに基づく企業価値の予測，情報処理学会全国大会”，2008

10. 金英子，松尾豊，石塚満：“Web上からさまざまなエンティティ同士のソーシャルネットワーク抽出”，人工知能学会第21回全国大会，2007

11. 金英子，松尾豊，石塚満：“Web上からソーシャルネットワークを抽出する汎用モデル，情報処理学会全国大会”，2007

12. 金英子，松尾豊，石塚満：“Web上の情報を用いたアーティスト間の弱い社会的関係の抽出，人工知能学会第20回全国大会”，2006

13. 金英子，松尾豊，石塚満：“Web 上の情報を用いた弱い社会的関係のネットワーク抽出”，第５回 Web インテリジェンスとインタラクション研究会，WI2-2006-34, pp129-134，2006

14. Y. Jin，Y．Matsuo，M．Ishizuka: “Extracting inter-business relationships from World Wide Web”, 社会情報学フェア 2005 -Web が生み出す関係構造と社会ネットワーク分析ワークショップ，pp29-36，2005

15. 金英子，松尾豊，石塚満： “Web 上の情報を用いた企業間関係の抽出”，FIT2005 第 4 回情報科学技術フォーラム』(CD-ROM)，D_001，pp.1-2，2005

16. 金英子，松尾豊，石塚満： “Web 上の情報を用いた企業間関係の抽出”，第 70 回人工知能学会知識ベースシステム研究会，SIG-KBS-A501-05，pp25-30，2005