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）

金　英　子

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 analyze social phenomena. Interaction patterns reveal relations among actors (such as persons, groups, companies), which can be merged to produce valuable information as a network structure. The power of social network analyses has become apparent through its use for orienting ideas and specific bodies of methods. Many applications are relevant to social networks: trust realization, ontology construction, community discovering, knowledge sharing, etc. While shaping our own knowledge, social networks are a strong influence shaping our lives.

Originally in social sciences, social networks are extracted through observations and interviews. The typical approach of questionnaire surveys is often used to obtain social networks, e.g., asking “please indicate which persons you would regard as your friend.” Recent technologies have enabled us to obtain social network data from e-mail archives, electronic libraries, schedule data, and Web citation information. By virtue of the current trends of “Web 2.0”, “Semantic Web”, and “Web 3.0”, a huge amount of information has become available on the Web. Current Web applications such as Wikis and blogs enable users to create and publicize their contents on the Web easily. Moreover, social networking services (SNSs) provide a foundation for users to communicate with their friends on the Web space. As an entrance to the Web corpus, a search engine is useful to access various information that is available on the Web. Extracting and analyzing social networks from the Web is expected to yield timely, comprehensive information of the real world on the Web.

Overall, in this thesis, we address two research topics for social networks on the Web:
(1) How to extract various social networks from the Web, and
(2) How to use and apply social networks mined from the Web.

Regarding the first topic, several methods exist to extract social networks including people (particularly researchers) from the Web using a search engine. In this thesis, we describe expansion of existing techniques to obtain social networks among various entities on the Web so that they become applicable to various domains. We propose two improvements—*relation identification* and *threshold tuning*—which enable us to address complex and inhomogeneous communities. For the first improvement of *relation identification* for the complex communities, we extract social networks among companies as examples. Extraction of relations between a pair of companies is realized using a search engine and text processing. Because names of companies co-appear coincidentally on the Web, we

propose an advanced algorithm, which is characterized by the addition of keywords (*relation keywords*) to a query. The relation keywords are obtained from the Web using a Jaccard coefficient. This method is a first attempt to extract companies' social networks from the Web using a search engine. The approach is also applicable to other actors, such as famous persons, organizations or other multiple relational entities. For the second improvement of *threshold tuning* for inhomogeneous communities, we extract social networks among artists (contemporary), who participated in the International Triennale of Contemporary Art (Yokohama Triennale 2005) to facilitate the navigation of artists' information. The algorithm can identify even tenuous relations among the artists. We first describe the basic idea of extracting social networks from the Web; then indicate that *objective rule-based methods* do not function well when applied to inhomogeneous communities. We propose a *subjective rule-based method* that is inspired by network questionnaires in social science. Furthermore, we propose a more advanced algorithm, an *objective and subjective rule-based method*, which enables more exhaustive extraction than that available under the previous methods.

For the second topic, we provide an example of advanced utilization of social networks mined from the Web. We seek to learn a ranking model of entities from a social network that has been mined from the Web. We assume that the performance and power (i.e. ranking) of social actors are usually interpreted as relations and structural features embedded in the network. Therefore, the ranking of entities can be learned and predicted from the social network. For example, if we seek to rank companies by market value, we can extract the social network of the company from the Web, then discern and 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. Consequently, we can understand the kinds of relations that are important for what sorts of rankings; additionally, knowledge about the structural embeddedness (i.e. network features of entities) can improve the rankings. We can address various questions from the Web using our algorithm: Why are some companies successful and others are not? Why are some researchers more productive than others? Similarly to methods described in many existing studies of social network analysis, our algorithm is useful to uncover answers from a structural perspective rather than from an attributional perspective. 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. Moreover, we specifically examine a new approach to using Web information for advanced analysis by integrating multiple relations and network features of entities for interpreting rankings among them.

Our study will expand social network mining from the Web so that is applicable to various domains. Furthermore, results of our study will provide a bridge between relation extraction and ranking learning for advanced knowledge acquisition for use in Web Intelligence.