論文の内容の要旨

論文題目　Entity Information Extraction from the Web Using Search Engine: Methodology and Application
（検索エンジンを利用したウェブからのエンティティ情報抽出手法と応用に関する研究）

氏　名　森　純一郎

The current development of Internet infrastructure such as broadband and wireless network enables users to easily access the Web. Nearly 87 million people in Japan are currently using Internet. Moreover, the current development of Web applications enables users to easily create and disseminate their contents in the Web. For example, using Blogs which are diary-like sites including multimedia contents such as photos and videos, users can easily publish their information. Nearly 8.68 million people in Japan are currently using Blog services.

With the rapidly growing contents on the Web, the recent Web has witnessed the transition from quality to quantity of information. A few years ago, when people tried to find information in the Web, they relied on the several ``authority'' sites that aggregate and disseminate valuable information. The algorithms for ranking the Web sites such as HITS and Pagerank have been developed and applied to such sites. However, the recent information explosion and distribution where users can easily publish their information on the Web has made it difficult to find valuable information only by using such hub and authority-based algorithms. As the contents on the Web are rapidly increasing, the quantity of information is recently becoming more important in the Web.

The importance of quantity of information has been explained with recent ``collective intelligence'' in the Web. Collective intelligence is the capacity of communities to co-operate intellectually in creation, innovation and invention. For example, Wikipedia, an online encyclopedia is based on the notion that every user can add an entry, is a successful site using the idea of the collective intelligence. Folksonomy, a style of collaborative categorization of Web sites using freely chosen keywords (or tags), is

another example of the collective intelligence. As seen in Wikipedia, every single user contributes to creating large quantity of information and then as seen in Folksonomy, the information are organized and guided by user communities.

The collective intelligence is also emerging in huge language resources of the Web documents that contain hundreds of billions of words of text. Therein, search engine plays an important role to access the resources. The simple way to access the language resources in the Web is to leverage hit counts of search engine as word frequencies. For example, when checking the spell, speculater or speculator, Google gives 4,700 for the former and 1,210,000 for the latter. As seen in this example, the collective intelligence of majority decision in the Web can be easily obtained simply by exploiting Google hit counts. With the large quantity of information, the Web has turned to the huge corpus that can be easily accessible source of language material using search engines, which in turn opens new possibility to handle the vast relevant information and mine important structures and knowledge.

In addition to the trend of ``Web as corpus'', another important aspect of the current Web is that our daily life is reflected in the Web. For example, social networking services (SNSs) have recently received considerable attention on the Web. SNSs enable users to maintain an online network of friends or associates for social or business purposes. Therein, the users can create their contents such as profiles and Blogs and communicate with their friends. Information about tens of millions of people and their relationships are published in several SNSs. For example, more than 10 million users are using mixi, the largest SNS in Japan.

As users publish their daily activities and social relationships in Blogs and SNSs, the Web is currently reflecting the information in the real world and the information is constantly updated through the contents that the users create online. Communication and information sharing in the real world are also reflected in the Web. Using several communication tools such as Email, Instant Messenger, and SNSs, users can communicate each other and share information online as they do in the real world. As information and communication in the real world have been reflected in the Web. The Web is becoming another form of our society.

With the current trend of ``Web as Corpus'' and ``Web as Society'', the large amounts of information that are originated from our daily activities in the real world are

available on the Web. In line with these trends in the Web, there is a new tendency of information retrieval that users try to find the ``entity-based'' information rather than documents. Here, entity is defined as the object in the real world such as person, location, and organization. In addition to single entity information, as we can see the recent trend of social networks which are basically representing the structure of relations among entities, relation information among entities from the Web (e.g. relation between two persons or relation between a person and an organization) are also becoming important information to be retrieved by users.

For example, when a user wants to know ``Prof. Mitsuru Ishizuka'', he might put the query ``Mitsuru Ishizuka'' into a search engine and try to find the information about Prof. Ishizuka from the search results. Therein, the final goal of the user is not to find the documents that include descriptions about Prof. Ishizuka but to find the related information of Prof. Ishizuka such as his students, research fields, affiliations, and projects. In other words, what the user wants to know is the information or attributes about Prof. Ishizuka as a person (or more precisely researcher) entity. In order to know about him further, the user might try to find the relation between Prof. Ishizuka and his student, co-author, or colleague. The user might be also interested in the relation between Prof. Ishizuka and his affiliation. As seen in this example, users are currently searching for entity-based information and entity relations on top of existing document-based Web information.

The Semantic Web is one approach to realize the entity-based information retrieval. In the Semantic Web, every resource is annotated with metadata using ontology. For example, ``Prof. Ishizuka'' is explicitly represented as an instance of Person class and related information about him such as affiliations and research fields are described with metadata. Users can easily search for and find the information about Prof. Ishizuka using the annotated metadata. However, because data should be annotated with metadata in advance to fully use the Semantic Web technologies, the annotation of metadata is a major problem to realize the Semantic Web. Therefore, there is still a huge gap between the current Web where most data are unstructured and the Semantic Web.

Aiming at realizing information services based on entity-based information and entity relations toward a next stage of current information retrieval, in this thesis we propose the methods for extracting entity information and entity relations from the

Web. The key features of our approach are to leverage existent search engine and obtain several Web-scale static such as hit counts and snippets in order to assess entity-related information. Applying several text processing technologies such as named entity recognition and clustering to the information obtained from search engine, our methods extract entity information, entity relations and social networks. The extracted information can be applied to several applications that are based on the entity information. We first develop the researcher search system that the information about researchers and relationships are automatically extracted from the Web. We also develop the information sharing system and the expert finding system using the extracted social networks.

Overall, in this thesis we address two major research questions for extracting entity information from the Web: (1) how the search engine can be used to access the Web corpus and extract entity information from the Web and (2) how the extracted entity information can be used to support users in entity-based information services.

For first question, we propose the basic method to use search engine in order to obtain the information about Web-scale static such as hit counts, co-occurrence, and snippets. Using the basic method, we develop the algorithms for extracting entity information, entity relations, and social networks from the Web.

For extracting the entity information, we propose a method of keyword extraction. The proposed method is based on the statistical features of word co-occurrence that are obtained from search engine. The basic idea is a following: if a word co-occurs with an entity in many Web pages, the word might be a relevant keyword about the entity. Importantly, our method extracts relevant keywords depending on the context of the entity. Our evaluation shows better performance to existing keyword extraction.

For extracting the entity relations, we propose a method that automatically extracts descriptive labels of relations among entities automatically such as affiliations, roles, locations, part-whole, social relationships. Fundamentally, the method clusters similar entity pairs according to their collective contexts in Web documents. The descriptive labels for relations are obtained from results of clustering. The proposed method is entirely unsupervised and is easily incorporated with existing social network extraction methods. Our experiments conducted on entities in researcher social networks and political social networks achieved clustering with high precision and

recall. The results showed that our method is able to extract appropriate relation labels to represent relations among entities in the social networks.

For extracting the social networks, we propose a method that leverages a search engine to build a social network by merging the information distributed on the Web. We describe some basic algorithms that extract social networks based on co-occurrence information as well as advanced algorithms that distinguish classes of relations based on a supervised learning. We also address new aspects of social networks: same-name problem, scalability, and keyword extraction.

For second question, we develop three systems that leverage the extracted entity information and social networks: researcher search system, information sharing system, and expert finding system. The systems are aiming at supporting users by using the extracted entity information.

We develop a researcher search system. The system is a Web-based system for an academic community to facilitate communication and mutual understanding based on a social network extracted from the Web. The system provides various types of retrieval on the social network: users can search for researchers by name, affiliation, keyword, and research field; related researchers to a retrieved researcher are listed; and the shortest path between two researchers can be retrieved.

We also develop a real-world-oriented information sharing system that uses social networks. The system automatically obtains users' social relationships by mining various sources in the Web. It also enables users to analyze their social networks to provide awareness of the information dissemination process. Users can determine who has access to particular information based on the social relationships and network analysis.

We finally propose a method that leverages the entity information and social networks of Web communities in order to find experts who have appropriate expertise and are likely to be able to reply to an information request. We develop the system using several data from the actual social network service and provide the service for locating relevant and socially close experts for information seekers.