Web Science Research

Naonori Ueda†

Abstract

What structure does the World Wide Web really have? How will it grow from now on? And how can we use this huge web effectively as a knowledge source? In NTT Communication Science Laboratories, Web science research aims to create intelligent information processing based on new principles and develop core technologies that will provide the answers to these important questions. This special feature explains the latest research results and future plans.

1. Features of the Web

The contents of each Web page are entirely a matter of choice by the page creator and a Web page can use various media, such as sound and images, as well as text. In addition, Web pages are not independent of each other: many are mutually related via hyperlinks. Therefore, the Web can be regarded as an information network consisting of various types of contents. Data on the Web is not collected and managed for some specific overall purpose but is generated through self-organization and is continually changing. For example, when an event occurs in the real world or some new fashion arrives, many Web pages and hyperlinks related to it quickly appear. That is, the Web is an information source in which various information is accumulated in a dynamic manner.

2. Web science research

Although the Web was born only about ten years ago, it has already grown explosively to include billions of pages and this growth is expected to accelerate even more in the future. We can use the Web as a source of reference in place of traditional information sources like dictionaries and we can keep up to date with world news and current affairs by looking at reports on the Web. The Web behaves like a contemporary society in miniature, so it is an interesting research subject that is being studied not only from the engineering viewpoint but also from the sociological viewpoint. It is a subject of research not only in information science, but also in other scientific fields, such as physics and biology. That is why we call this "Web Science".

† NTT Communication Science Laboratories
Soraku-gun, 619-0237 Japan
E-mail: ueda@cslab.kecl.ntt.co.jp

3. Research approach

Scientific research often involves "modeling". Namely, one creates a model to represent some phenomenon and then checks its validity using real data. In Web science research, we have also taken this "modeling" approach. More specifically, our modeling is based on mathematical statistics. This is another reason for using the term Web science research.

4. Research themes

This section describes some of our recent results.

4.1 Text modeling

Although Web pages are written in the html language, there are no restrictions on contents and style as there are with a fixed-form database dedicated to a certain purpose. Therefore, it is much harder to retrieve information from the Web than from a conventional database. For example, the widely used keyword-based retrieval does not work well because it finds many irrelevant pages that contain the keyword but no useful information. For the search to be

more efficient, Web pages should be arranged by topic beforehand.

Many Web pages cover multiple topics, such as "sport" and "music". Quite a few cover three or more topics. By regarding text as a set of words and considering which word is generated by what distribution, we devised a new text model within the framework of mathematical statistics. By applying it to automatic multi-topic detection and similar document retrieval using several tens of thousands of real Web pages, we confirmed that the proposed model is valid. Details are given in the next article: "Multiple Topic Detection by Parametric Mixture Models (PMM)" [1].

4.2 Network visualization

The Web can be regarded as a graph by treating a hyperlink as an edge and a page as a node. If the graph could be visualized, we could intuitively understand the structure of the Web, extract important central pages, and analyze the relationships among communities. However, the Web is a huge network, so such visualization is not an easy task. We have devised an efficient visualization method using only the connections between nodes on a Web network. This method is a useful tool for analyzing not only Web networks, but also many other networks, such as the gene regulatory network between living things, social networks in sociology, and semantic networks between words. Details are given in the third feature article, "Visualizing Large-scale Network Data Structures" [2].

4.3 Growth models of the Web

The Web grows continuously as users create new pages and new hyperlinks based on their personal interests. It is well known that the relationship between the number of pages and the number of links to external sites on them obeys a power law. This is an interesting property that does not occur in randomly organized networks and it is common to other networks that grow through self-organization.

Moreover, on the Web, there are some clusters: each cluster grows independently and this growth produces some directivity in the links. Such a cluster is called a community, which can be regarded as a group of good friends. If we can discover communities, they should help us to analyze macroscopic social phenomena, predict new phenomena, and discover unique phenomena. Considering these characteristics of Web networks, we developed a growth model of the Web. Through experiments using Web data, we studied the performance of Web growth predictions. This is described in the fourth article, "Growth Models of Web Networks" [3].

5. Future developments

In addition to the three themes mentioned above, we have recently been investigating a Web surfing model, anomaly detection, automatic construction of a topic taxonomy, information integration, and intelligent information retrieval using relevance feedback. Due to space limitations, these cannot be covered in detail in this special feature, so they are briefly outlined below.

The Web surfing model psychologically models how users surf the Web. A system using this model can guide users in effective directions according to its predictions of their preferences.

Anomaly detection automatically finds pages that have quite different word distributions from others. This could be useful for qualitatively evaluating information and detecting server crashes by detecting anomalies in message information exchanged between networks.

In automatic construction of topic taxonomy, we have studied some statistical models that extract important words and construct a hierarchy of them.

Information integration is a framework for solving difficult problems that cannot be solved using only one information source by combining multiple information sources. For example, when a user cannot find the required information at a certain portal site, the system can automatically search corresponding topics on other sites, instead of the user having to manually search the other sites. The Semantic Web for metadata is a related technology.

In information retrieval using relevance feedback, we have been studying a method where feedback based on the user's evaluation as a reference result can make the system perform more intelligent searches. For example, when 100 pages have been retrieved by a keyword search, user evaluation of the first ten pages lets the system automatically re-rank the remaining reference pages based on the user's relevance feedback. As a result, the user can get relevant pages efficiently.

One goal for information retrieval is to enable inquiries to be made in natural language. Already, some systems can actually output correct answers, to some extent, when the questions are constrained to fixed forms, such as the name of a place or person. I believe that the day when unrestricted natural lan-

guage-based question-answering systems arrive is not so far in the future.

References

[1] K. Saito, “Multiple Topic Detection by Parametric Mixture Models (PMM) — Automatic Web Page Categorization for Browsing,” NTT Technical Review, Vol. 3, No. 3, pp. 15-18, 2005.
[2] T. Yamada, “Visualizing Large-scale Network Data Structures,” NTT Technical Review, Vol. 3, No. 3, pp. 19-23, 2005.
[3] M. Kimura, “Growth Models of Web Networks,” NTT Technical Review, Vol. 3, No. 3, pp. 24-27, 2005.

Naonori Ueda
Executive manager of the Intelligent Communication Laboratory and Group Leader of the Emergent Learning Research Group, NTT Communication Science Laboratories.
He received the B.S., M.S., and Ph.D. degrees in communication engineering from Osaka University, Suita, Osaka in 1982, 1984, and 1992, respectively. In 1984, he joined the Electrical Communication Laboratories, Nippon Telegraph and Telephone Public Corporation (now NTT), Kanagawa. In 1991, he moved to NTT Communication Science Laboratories, Kyoto. He received the Funai Best Paper Award of FIT2004, Best Paper Awards from the Institute of Electronics, Information and Communication Engineers (IEICE) in 2000 and 2004, Research Awards from Japanese Neural Network Society (JNNS) in 1995 and 2003, and the Telecommunication Advancement Foundation Award in 1997. He is a member of IEICE, JNNS, the Information Processing Society of Japan, and IEEE.