論文内容の要旨

A Study on the Evolution and Emergence of Web Spam

（ウェブスパムの進化と出現に関する研究）

氏名　鄭　容朱

Since the Web plays an important role in the economy, social activities, and information sharing, search engines have become important tools to access the huge amount of available information on the Web. Considering that half of users look at no more than the top five results in a search result list, it is clear that a higher ranking in the result list brings more traffic and profit to web sites. As a result, many web sites started using unfair methods, so called web spamming, to obtain a higher ranking than they deserve. Detecting web spam is challenging because new spam pages are being continuously created. Once new anti-spamming techniques are proposed, spammers invent new sophisticated techniques to avoid them.

In this thesis, we study the evolution and emergence of web spam using a large-scale of Japanese Web archive.

To understand the evolution of web spam, we analyze temporal changes in structures and contents of web spam. We propose a method for extracting a spam link structure, a link farm, from large-scale of web archives and investigate temporal changes in its distribution. We classify topics of spam pages based on their universal resource locator (URL)s and investigate temporal changes in topic distribution. We show that link farms hardly grow and topic distribution of spam pages hardly changes, which means monitoring link farms is not helpful for newly created spam detection.

To understand the emergence of web spam, we focus on pages that contain link to spam pages. We propose a method for detecting hijacked sites, which are legitimate sites containing links to spam sites, and evaluate its detection precision. We also show that monitoring hijacked sites is helpful in discovering newly created spam sites. On the other hand, we propose a method for identifying hosts that generate links to spam hosts and evaluate its identification accuracy. We find that many links to spam hosts are created by these hosts, and some of them are active for a long period.