

Title	リンク構造に基づくWebコミュニティ解析法
Author(s)	松久保, 潤
Citation	
Issue Date	2004-09
Type	Thesis or Dissertation
Text version	none
URL	<a href="http://hdl.handle.net/10119/811">http://hdl.handle.net/10119/811</a>
Rights	
Description	Supervisor: 林 幸雄, 知識科学研究科, 博士

# A method for analysing Web communities based on link structure

Jun Matsukubo

## Abstract

The WWW, the huge communication media consisted of contents and links, is constructed by billions of pages combined mutually, which are created by millions of people. Furthermore, the WWW is not only huge, but is merely continuing rapid growth now. The distributions of page number to in-degree (number of links referenced by others) and to out-degree (number of links referencing to others), i.e. degree-distribution follows power-law. This global character appears on actual networks besides WWW. The networks with degree-distribution following power-law are called scale-free (SF) networks.

On the other hand, on the local link structure of the WWW, it has been clarified that the pages treating similar topics combine each other densely. Recently, the huge amount of information on the WWW attracts the attention of scientists. The scientists have studied the methods based on the contents analysis and the link structure to acquire useful information from billions of pages.

However, there're the following problems in the methods based on contents analysis: (1) Classifying billions of pages by checking the contents is actually difficult due to the huge number of those. (2) Using fixed categories for automatic classification doesn't guarantee that one page treats one topic.

On the other hand, classifying based on link structure discovers a set of the pages with similar topics, using the assumption that a link from a page expresses the concern and interest of the maker to the page referred to by the link. This shows the close relationship between the link structures and the contents of pages. In the background, there are the relationships between page makers with common concerns and interests. Furthermore, the social structure like an acquaintance relationship in the actual world has appeared also on the WWW.

In this study, based on the assumption that Web pages with similar topics constitute link structures combined densely, the link structures are treated as Web communities. The aim of this study is to analyze the structural feature of Web communities for the elucidation of the social structures, which appear on the WWW.

The analysis for the structural feature of Web community is an important subject connected with the following subjects: classification of pages for topics, elucidation for the structure of information space and that for the growing mechanism of the link structure of the WWW, and so on.

This paper consists of 6 Chapters. Chapter 1 describes the introduction of the study. Chapter 2 explains the structural feature of the WWW and the conventional methods for finding a set of pages with similar topics. In Chapter 3, the link structures obtained by the above conventional method are analyzed. Chapter 4 describes the collecting method for high degree pages from the actual WWW, which are cores of communities. In Chapter 5, in order to grasp a global link structure of a network, the link structure is visualized by an automatic drawing method of graph. Chapter 6 describes a conclusion. Hereafter, the outline of each chapter is shown.

In Chapter 1, the relationship between this study and the conventional studies for analyzing the link structure consisted of Web pages and the subjects of this study.

In Chapter 2, we review global and local features of the link structure of the WWW, a conventional method for generating artificially the link structure imitating the structural feature of the WWW, and one for finding pages with similar topics.

In Chapter 3, experimental results show features of a link structure, which consists of a set of nodes obtained by the above method for page with similar topics. In the experiment, an artificial network imitating the WWW structure is used to avoid the influence of sampling from the actual WWW. Consequently, nodes construct the dense link structure near high-degree those on SF network. Therefore, high-degree pages become cores of Web communities.

In Chapter 4, it tries to search efficiently for the high-degree pages considered to be cores of communities. On the one hand, since the link structure of the WWW is always changing quickly, it is difficult to use the whole link structure for search. Thereby, the following method is proposed: preferentially searching for the page with the highest in-degree on a local link structure consisted of only those discovered during search. On the actual Web, features of link structures are compared, which are searched in the

proposed method and breadth-first searching. The results show that search using the proposed method collects high-degree pages efficiently. Moreover, this suggests that link structures searched by the proposed method includes community structures.

In Chapter 5, it shows that using a proposed method for automatically drawing graph enables us to grasp the global link structure. Drawing graph with the proposed method emphasizes the high-degree nodes which are important for the network structure. The example results demonstrate that it is able to find the portal sites and the global user flow on the access structure between the sites about the Oitama area of the Yamagata prefecture.

In Chapter 6, the result obtained by this study is summarized and the validity of the analyzing method based on the link structure of the WWW is shown. In addition, the subject left behind to future is described.