Research of WWW Link Structure for Search Engines

Ryunosuke Ohshima

School of Information Science,
Japan Advanced Institute of Science and Technology

February 15, 2000

Keywords: Link Structure, World Wide Web, Information Retrieval, Search Engines.

In this paper, I propose some methods for extracting information of link structure on World Wide Web. I also propose some ways to group, classify and order the pages on Web from a viewpoint of extracted link structure's information. To demonstrate some effectiveness of link structure's information, This paper also provides some methods to improve search engines on Web.

It can be said that Web is the most successful field of hypertext and hypermedia. Thanks to their links, We can access its information randomly. This provides not only flexibility but also ambiguousness to users of Web. So, information producers of Web need to guide their users effectively. Although, it is too easy to produce new pages on Web, there are many page producers that are unfamiliar with HTML. Their miscellaneous pages are always added and updated every moment. This fact causes utter chaos. Since Web is not equipped with a function to classify and arrange its information, users are embarrassed with huge and miscellaneous pages on Web. A large gap exists between producers and users to communicate each other.

To cover this gap, Web needs a function to classify and arranges information to navigate users correctly. We need a function that guides users easily. Users need a clear grasp of large quantity pages and need to understand an accurate structure of Web. We need a big help to group, classify and order pages with suitable standard. To bring order to the chaos of Web, We can make use of link structure's information obtained from Web. It can improve pages' reference efficiency. We can use link structure's information to visualize or summarize a group of pages, and can add new information structure to recompose pages on Web. We are also able to make use it to prefetch or cache, to fulfill our unfulfilled desire for a fast Web.

Recently, Web research of page classification and page arrangement has been mainly done on information retrieval. It is a field of natural language processing. On the other

hand, there were not so many researches that treat link structure's information mainly. Although links are an important feature of Web, an expression capability of Web links is still poor. Quality of its links is never also high since there is much miscellaneous noisy information not necessarily to use for a classification and arrangement. However, although there are such difficulties, it is enough to believe that a classification and arrangement of Web by its links can carry out effective use. Moreover, page producers can use a link now more effectively. They can classify and arrange their Web pages efficiently by its links. In this research, I stand on this idea and aim to extract the link structure's information of Web. I propose how to use link structure's information effectively.

I have made a through investigation into Web links. As a result, I have traced links among pages and checked their paths' length. They are indicators of relationship between pages, and I call it "Link Distance". I also pay attention to "Link Influences" between pages. "Link Influences" are derived from pages' number of links. Then, based on "Link Distance" and "Link Influence", I propose some methods to extract link structure's information from the pages on Web. When it comes to finding sentence relationships among pages, they are expressed with some measures: "Degree of Relation", "Degree of Abstraction", and "Degree of Relativity". "Degree of Relation" shows a depth of relations between each page. Similarity of words among pages is a main factor of it. "Degree of Abstraction" is an indicator to distinguish between abstract pages and general pages. "Degree of Relativity" shows whether pages come in front between each page. To detect link structure's information, we use various elements, such as a position where the link exists within HTML, and a relation with other links.

Furthermore, based on these extracted link structure's information, I propose some methods to performe grouping, classifying, and ordering on groups of pages. I divide pages into strongly connected groups (which I call "Content Groups") with "Link Distance". Then, I elect representative pages of "Content Groups" with "Link Influence". "Degree of Relation" fixes relation between each page and I use it as a threshold to divide "Content Groups". Classifying means discovering an informational layered structure using "Degree of Abstraction". In Ordering, I find order among pages with "Degree of Relativity". I show a validity of above proposal methods with some examples. I perform grouping, classifying, and ordering with some example pages on actual Web.

I also propose some improvement methods of information retrieval on Web. We conventionally used full-text-search system called search engines as applications for information retrieval on Web. Full-text-search system uses keywords, but there are many search problems that a keyword cannot become a key to. There exists many pages that are suitable for some users' purpose but are difficult to find by keywords. Many pages accidentally contain some keywords that are unrelated to their contents. So, a suitable page may be buried among many pages that are not suitable, as a result. By adding link structure's information to the full-text-search that uses keywords, suitable pages are taken out and redundant pages are removed. I verify my methods, comparing with some search examples of existing search engines.

To assess our methods, I have implemented a prototype system, which is named "Namazu- Hige". It is based on a full-text search engine "Namazu". I build "Hyperlink Indexer" to make indexes of link structure. It extracts links from pages and compute

pages' "Link Distance" and "Link Influence". I also build "Grouping Engine", which finds "Content Groups". It receives pages' lexical score from the full-text search engine part. Then, it finds "Content Groups" from the indexes of link structure and sorts them with their new score. "Grouping Engine" also discovers representative pages of "Content Groups" and returns them as result. I show some behavior of the "Namazu-Hige" with examples.

As a result of this paper, I throw light on a property of link structure. I use "Link Distance" and "Link Influence" as indicators to measure "Degree of Relationship" between Web pages. Then, I introduce some method to investigate meanings of links. I also propose some methods to extract link structure's information on Web. Finally, I also propose some effective methods to use link structure's information in information retrieval. Our search system cooperates with full-text-search system, and it produces improved results for Web users. I expect that link structure's information can be applied to various fields. They are not only information retrieval but also such as summarization of the information, recognizing of structural information and recomposition of information.