JAIST Repository
https://dspace.jaist.ac.jp/

Title	参照ページからの情報を利用したWeb探索支援
Author(s)	板橋，英夫
Citation	
Issue Date	2002-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1530
Rights	
Description	Supervisor:白井 清昭，情報科学研究科，修士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Supporting Web Searching Using Information Acquired from Referer Pages

Hideo Itabashi (910008)

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

February 15, 2002

Keywords: WWW, Web browsing, reference relations, extract of reference area.

The Web is a hypertext body of approximately 300million pages that continues to grow at roughly a million pages per day. Page variation is more prodigious than the data 's raw scale: Take as a whole, set of Web page s lacks unifying structure and show far more authoring style and content variation than that seen in traditional text-document collections. This level of complexity makes an " off-the-shelf " database-management and information retrieval solution impossible.

In such situation, this paper aims at supporting web searching. Before the users click the hyperlink and display the linked web page ("target page" hereafter) in their web browsers, browsers show the information about the target page, such as an explanation of the content of the target page, another person's opinion or evaluation about the target page, etc. The method to acquire the information about target pages is the following: first, correcting the web pages the have hyperlinks to the target page ("referer page" hereafter), then extract "reference area" from referer pages. "reference area" is a fragment of a referer page, which contains the information about the target page. By showing reference area to web searchers, they can judge whether the linked page (i.e. target page) is useful for them for not. This paper proposes a method to extract reference areas from ref-

Copyright © 2002 by Hideo Itabashi

erer pages automatically. Furthermore, I analyze what kind of information could be extracted from referer pages.

In order to explore the method to extract reference areas from referer pages, I collected the examples of target pages and their referer pages from real World Wide Web. First, examples of target pages were collected using a search engine with the query "chat" and "Mado no Mori". Then, web pages which have hyperlinks to the target pages were collected as examples of referer pages using the same search engine. Target pages which has less than 10 referer pages were removed from the set of example web pages. As a result, 21 target pages and 582 referer pages were collected. I analyze these pages and explore the method to extract reference area automatically.

I decide to use HTML tag mainly for extracting reference areas. The method to extract reference areas from referer pages are following: I distinguish the hyperlink to the target page ("target anchor" hereafter) which refers the target page only for the in-site navigation. If anchor string is "back", "top" etc., I judge the anchor is hyperlink for the in-site navigation. In such case, no reference areas were extracted. Next, I attempt extracting reference areas. I used the following HTML tag:

- list tag
- ⟨br⟩ tag
- table tag
- other tag

In case of list tag, I extracted the fragment between ⟨li⟩ tag which preceded target anchor and ⟨li⟩ tag which succeeded the target anchor. When the target anchor succeeded ⟨dd⟩ tag, I extract text which succeeded the corresponding ⟨dt⟩ tag as reference area. Furthermore, when the target anchor succeeded ⟨dt⟩ tag, I extract text which succeeded the corresponding ⟨dd⟩ tag as reference area.

In case of ⟨br⟩ tag, I extracted the reference area as following: I search pattern "(anchor) (text) (⟨br⟩)" in the referer page. If such patterns were found, the 'text' which succeeded the target anchor were extracted as reference area. ⟨br⟩ tag can be omitted in the above pattern.

In case of table tag, if anchors were listed vertically using table tags, I extract the text in the right cell of the target anchor. If anchors were listed horizontally using table tags, I extract the text in the below cell of the target anchor. If anchors and text were listed alternatively in the same column of the table, I extract the text in the below cell of the target anchor.

In case of other tags, I extract text near the target anchor. The boundary of the reference area was determined by HTML tags, except for ⟨font⟩ tag, ⟨image⟩ tag, ⟨a⟩ tag and comment, which preceded / succeeded the target anchor.

Next, I analyze what kind of information could be extracted from reference areas. As a result, I found 3 types of information, explanation type, opinion type (web page) and opinion type (contents). The explanation type is an explanation of the target page. Especially, the explanation in language which is different from language used in a target page were sometimes extracted. This cannot be extracted from target page itself. The opinion type (page) is a opinion about the target page itself, such as layout or impression of the web page. The opinion type (contents) is a opinion about a contents of target page. For example, the target page is a catalogue of the handy phone or personal computer, the opinion about handy phone or personal computer could be extracted from referer pages. Web searcher can understand the target page more easily if reference areas are classified into these three types automatically. This is the future work.

Finally, I conducted an experiment to evaluate the proposed algorithm to extract reference areas. We conducted two kind of test, closed test and open test. In closed test, I used examples of web pages mentioned above. In open test, I used 2 target pages. One is "Mayu Kobo", web site which provides pictures for web pages in free. The other is "Libra", web site which compares various commercial products, such as personal computer, car, TV, handy phone, etc. I collected refer pages of these 2 target pages using search engine. Numbers of referer pages of "Mayu Kobo" and "Libra" were 86, 40, respectively. In both open test and closed test, correct reference areas were marked up by hands. In closed test, the recall and precision was 0.57 and 0.49 respectively, when extracted reference areas, which were exactly same as reference areas marked up by

hands, were regarded as correct ones ("exact match standards" hereafter). When extracted reference areas which contains reference areas marked up by hands were also regarded as correct ones ("sub string match standards" hereafter), the recall and precision were 0.85 and 0.82, respectively. In open test, the recall and precision was 0.32 and 0.35 respectively in exact match standards, And 0.62 and 0.69 respectively in sub string match standards. List tag and table tag were applied rarely to extract reference areas, but achieved high precision. When using ⟨br⟩ tag, precision in open test was much poor than that in closed test. As a consequence, using ⟨br⟩ tags is too specialized to collected examples of web pages. Other tags were used often to extract reference pages, and precision is more than 0.5 in sub string standards. However, reference areas extracted using other tags were sometimes quite long text. In such case, only the information about the target page should be selected from extracted reference areas. In addition to the HTML tags, linguistic knowledge will be also required to extract reference areas.