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

Title	視覚障害者用音声ブラウザのためのウェブページ構造解析
Author(s)	加藤，邦彦
Citation	
Issue Date	2006-03
Type	Thesis or Dissertation
Text version	author
URL	http://hdl.handle.net/10119/1965
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

Structural Analysis of Web Pages for Voice Browsers

Kunihiko Katou (410030)

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

February 9, 2006

Keywords: voice browser, web page anlaysis, segment, DOM, anchor of link.

By the development of recent WWW, it is possible to get various information by using a computer even if visually impaired user. They browse a web page by information translated into speech through the tools, voice browser or screen reader. " Voice browser " is the specialized tool for translation into the speech of a web page, and " screen reader " is the general-purpose tool that can be translated to the speech for various applications.

But there are problems that need extra efforts till arrive at the place where they would like to get the information, in existing tools. A cause of those problems is the system that uses the method only reading simply in turn from the page top of an HTML source when reading the page.

To resolve these problems, we suggest two following methods to help Web browsing of the visually impaired user more adequately in this study. The first method is that to skip needless information, system performs structure analysis of a web page and detects the segments of the page based on the structure. The second method is that, when a user follows a link, the system estimates the position where the information that a user wants to get from anchor text. After that it starts reading from estimated position. We expect the visually impaired user can perform Web browsing by only sound comfortably, by developing the voice browser which has these functions.

We developed web page segment detection method to realize the system which navigates based on a segment of a web page. Our method detects segments based on a HTML tag used as structure in web pages using DOM. Furthermore, for detected segments, we can improve accuracy of the segment detection by applying segment division method using image and partial tree in the table and merge method of a header part. The division method using image is the technique that assumes the most frequently used image in the segment as border of a segment. The division using partial tree in the table is the technique that divide using a similarity measure of a partial tree of DOM that makes the TD tag or the TR tag a root. If a similar level of those partial trees of DOM is higher than a certain threshold, the system divides a partial tree of DOM as one segment. In addition, the merge method detects a header part for the segment which does not include a header part, and merges into a single segment a header part and the segment.

We experimented on by using real web pages to confirm the effectiveness it. At first, we collected 20 pages and made correct data of segment with hands for these pages. We analyzed those pages, developed a system from the analyzed result. And when having

Copyright © 2006 by Kunihiko Katou

performed the system on these pages, our system was able to detect 49% segments. Furthermore, if the system could detect a start position of a segment precisely, we thought that the system was able to perform an effective skip to some extent. Therefore having computed reproduction rate only for a start position, the result was 64%. In addition, the ratio that a detection segment crossed over plural correct answer segments was under 1.5%. As a result of these, we can regard that proposed method is effective for the segment detection. We confirmed that the performance was improved by applying segment division method and merge method.

As a result of having performed it with the system on new page 20, it was able to detect correct 34% segments. We confirmed the performance of the system was improved by applying the segment division method and the merge method compared with the method based on DOM. In addition, the reproduction rate for only a start position was 58%, and the ratio that a detected segment crossed over plural correct segments was less than 4% at most. From the results of these experiments, we are able to expect that the system can detect many correct segments, by improving these methods in future.

We developed method to identify a reference position of the link. When the user follows a link, the reference position of the link means the position of the text where the information the user wants to know is. Our proposed method performed pattern matching between the anchor text of a link and a text on the page that the link represents, the matched texts output as the reference positions. In that event, the system compare a text length of the text in the page and the anchor text, a shorter text become a pattern, and whether it exists in the long text is examined. By this method, even if extra string was added to the anchor text and either one of a text in a page, the system can detect the text of reference position of the link.

For the experiment of our method, we selected the links of 100 from 20 pages of real web pages, and applied to them the proposed method. As a result, our system was able to detect a reference position of a link definitely in 67 pages. In 67 pages that the system was able to identify, we counted the text read by the voice browser that existed from the top of the page to the reference position. While the result of not using system was 76.1, the result using system was 1.5. By this result, it was proven that our system omitted user's efforts.

We analyzed 33 pages that the system was not able to identify to get the reason that was not able to detect a reference position. As a result, we found that there were various paraphrases between an anchor text and reference position texts of a link. We are going to introduce the comparing method using a word vector for to solve this problem.

Through the experiment results, our system was borne out that help to Web browsing of the visually impaired user, although it was necessary to improve performance of the method by improvement of further method. In addition, we must consider about the interaction with the user, if we applied proposed method to a voice browser. Therefore we must introduce the technique keyword extraction and the technique of the header detection to explain contents of a segment. Therefore we intend to introduce the technique keyword extraction and the technique of the header detection to explain contents of a segment while taking in an opinion of a real user to explain contents of a segment. We expect that we can help Web browsing of the visually impaired user adequately by developing voice browser with these functions.