Expert Systems with Applications 41 (2014) 4322–4329

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Determining the titles of Web pages using anchor text and link analysis

Ok-Ran Jeong[a,*], Jehwan Oh[b], Dong-Jin Kim[c], Heetae Lyu[d], Won Kim[a]

[a] Department of Software Design and Management, Gachon University, Republic of Korea
[b] Department of Computer Science, University of Minnesota, United States
[c] NHN Institute for The Next Network, Republic of Korea
[d] Naver Corporation, Republic of Korea

ARTICLE INFO

Keywords:
Anchor text
Link analysis
Title extraction
Web page

ABSTRACT

Determining the titles of Web pages is an important element in characterizing and categorizing the vast number of Web pages. There are a few approaches to automatically determining the titles of Web pages. As an R&D project for *Naver*, the operator of *Naver* (Korea's largest portal site), we developed a new method that makes use of anchor texts and analysis of links among Web pages. In this paper, we describe our method and show experiment results of its performance.

1. Introduction

As the number of Web pages has exploded, it has become important to organize them in order to limit the search space for information. One approach to organize the Web pages is to group them based on key characteristics of their contents. The title of a Web page, or a set of keywords or tags may characterize the contents of the Web page.

Understanding, characterizing or summarizing the contents of a Web page has been the subjects of research. Many methods have been proposed. They may be grouped into at least four classes: Rule-based methods, machine learning methods, content analysis methods, and link analysis methods. Rule-based methods (Giuffrida, Shek, & Yang, 2000, Liddy Mao, Kim, & Thoma, 2002, Mao et al., 2004 & Yilmazel et al., 2004) parse raw HTML documents into DOM trees, and determine the titles of Web pages using predetermined rules. Machine learning methods (Craven, 2003, Crescenzi V. et al., 2001, Evans, Klavans, & McKeown, 2004, Freita, Li, Zaragoza, Herbrich, Shawe-Taylor, & Kandola, 2000 & Li et al., 2002) extract features from Web documents. While most machine learning methods extract textual features, the Vision-based Page Segmentation (VIPS) Algorithm (Cai, Yu, Wen, & Ma, 2003) extracts the location information for different elements on a Web page to determine the title of the Web page. Content analysis methods find the titles of Web pages using almost all features of a Web page. The features used in Hu et al. (2005) include format information, tag information, position information, DOM tree, and linguistic information.

The links that connect Web pages imply relationships between the Web pages. As such, considerable efforts have been made to make use of the links in understanding or organizing the contents of Web pages. Wang, Wang, and Kitsuregawa (2001) proposed the use of links in clustering Web search results. Dumais, Jin, and Jin (2001) proposed augmenting the content information with link information. In particular, it combined content-based similarity measures and link-based scores in order to improve retrieval accuracy. Davison (2000) and Eiron and McCurley (2003) showed how anchor text can be used to enhance search performance. Freitag (2000) proposed finding related documents within a Web site using the words found in anchor texts.

Related research mentioned by the reviewer was added as follows. The research mentioned by the reviewer is a very useful research field in the research field that we proposed, and we believe that it is highly related to our study. However, the below related research and Hu et al. (2005) proposed in this study appears to be different in terms of objective and approach method. Klein and Nelson (2008) is on methodologies for finding pages with different URLs while being related to the internet using LS (lexical signatures) formed in the web page. It is useful research to find the "aboutness" of random Web pages using LS. In Li, Li, Li, Wang, and Qu (2010), a method for extracting title phrases through the document semantic network was proposed. Yan and Yang (2011) proposed a method for extracting titles and contents using the features of news sites on the web. By taking these studies into consideration in future research, we believe that our research can be further developed to become more useful. The updating broken web links: An automatic recommendation system research proposes a method to recommend Web pages with similar contents when a web page cannot be found because there is no link to the web page being searched on the web. We have developed an alternate method. Our method is based on the use of anchor texts and an analysis of links among connected Web pages. Our method is based on the use of anchor texts and an analysis of links among connected Web pages. Our research was conducted in an R&D project for *Naver*, the operator of *Naver*,

* Corresponding author.
E-mail address: orjeong7@gmail.com (O.-R. Jeong).

0957-4174/$ - see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.12.033

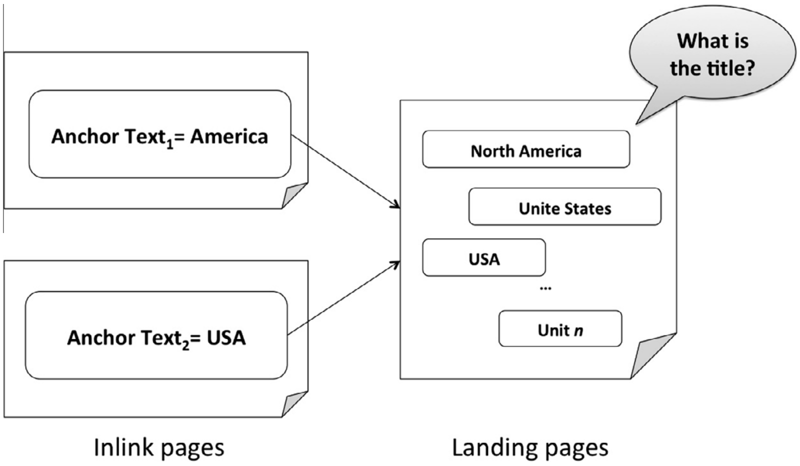

Fig. 1. Motivating example.

Koreas largest portal site.[1] *Naver* needs to continually improve the *Naver* search engine. One of the ways is to quickly categorize Web pages based on their contents. Fig. 1 shows a motivating example for our method. It shows two inlink pages and one landing page. An inlink page, which is an intermediate node on a graph of Web pages, contains an anchor text that links to a separate Web page. The landing page, which is the leaf node on a graph of Web pages, does not contain an anchor text that further links to another Web page. In Fig. 1, each inlink page contains an anchor text: America or USA. The landing page includes 3 page units, each being characterized by the keywords North America, United States, and USA. This study is on automatically extracting titles from landing pages using the anchor text of the inlink page. Here, unit is defined as candidate sentences that can be the title of the landing page. In other words, it is composed of one or more terms. All sentences and phase between HTML tags could be defined as units. A unit generally corresponds to a line of text, which removes HTML tags in landing page. If unit have a URL linked to other pages, the one is to be an anchor text. Our method determines the title of the landing page based on the results of comparison of the keywords that characterize the page units in the landing pages and the anchor texts of the inlink pages. Our method consists of four steps. In the first step, we collect the domain addresses associated with the anchor text on each inlink page. In the second step, we use the hyperlink-induced topic search (HITS) algorithm to analyze the links (inlink and outlinks) on each page. In the third step, we determine, for each Web page, whether it is an inlink page or a landing page. In the final step, we determine the title of each of the landing pages.

To measure the accuracy of our method, we used 28 real Web sites. The total number of landing pages was over 38 million, and the total number of inlink pages was almost 2 billion. We measured the accuracy by comparing the extracted titles with the actual titles of the landing pages. Our method shows up to 79.33% accuracy.

The remainder of this paper is organized as follows. In Section 2, we provide an overview of our method. In Section 3, we describe our method in detail. In Section 4, we describe the performance experiments and analyze the results. In Section 5, we conclude the paper.

2. Overview

A Web site or a Web document consists of many Web pages. A Web page containing an anchor text is linked to another Web page. For example, in the anchor text '<a HREF= "http://www.google.com/"> Google </a>', the text Google is associated with the URL http://www.google.com. When the anchor text is clicked on, the Web page associated with the URL is accessed. In this way, via the anchor texts, the Web pages form a graph structure. (For expository simplicity, we assume for the rest of the paper that the link structure is an acylic graph with a single root node.) Fig. 2 shows the general link structure of Web pages. A Web page that does not contain any anchor text is a landing page. A Web page that contains one or more anchor text is an inlink page. The inlink page is expressed as in Formula (1).

$$\textit{inlink page} = \{(\textit{anchor text}, \textit{URL})_1, (\textit{anchor text}, \textit{URL})_2, \cdots + (\textit{anchor text}, \textit{URL})_n\} \tag{1}$$

If the anchor text is '<a href="http://www.google.com/"> Google </a>', the landing page becomes http://www.google.com. The landing page is expressed as in Formula (2).

$$\textit{landing page} = \{(\textit{unit}_1, \textit{unit}_2, \ldots, \textit{unit}_n), \textit{URL}\} \tag{2}$$

Each Web page may include many units. One unit of the landing page may hold the title of the page. However, it is a challenge for software to automatically determine which of the many units of the landing page holds the title. The objective of our research is to develop a method to automatically determine the title of each of the landing pages on the link structure of an arbitrarily complex Web site or Web document. Now we outline our method, using Figs. 3–6 to illustrate each of the steps involved.

We first identify a set of anchor texts. We select the URL associated with each anchor text included in each Web page. For example, on $page_1$ in Fig. 3, we select the URLs of three anchor texts. The three anchor texts link $page_1$ to $pages_{2,3,and4}$. Next, we use the HITS algorithm to create an adjacency matrix to represent the existence of a link between each inlink page and each of the pages it links to. For example, matrix A in Fig. 4 represents the links between $page_1$ and $pages_{2,3,and4}$. Next, using Levenstein distance (Gusfield, 1997), we distinguish between inlink pages and landing pages. For example, the URLs of the three anchor texts found on $page_1$ in Fig. 3 are shown in Fig. 5. The URLs each consist of 5 components. The first 4 components of the URLs are the same. The URLs differ in only the 5th component, each of which represents the landing page. Next, we determine a set of units for each of the landing pages. We then compare the set of anchor texts (the A-set) for the inlink pages and the set of units (the U-set) for the landing page. The unit with the highest similarity measure is selected as the title of the landing page. For example, in Fig. 6, the U-set includes "U of M Freshmen Celebrate Pride & Sprit night" and the A-set includes "U of M,

[1] http://www.naver.com.