Similarity Web Pages Retrieval Technologies on the Internet

Rung Ching Chen
Chaoyang University of Technology, Taiwan

Ming Yung Tsai
Chaoyang University of Technology, Taiwan

Chung Hsun Hsieh
Chaoyang University of Technology, Taiwan

INTRODUCTION

In recent years, due to the fast growth of the Internet, the services and information it provides are constantly expanding. Madria and Bhowmick (1999) and Baeza-Yates (2003) indicated that most large search engines need to comply to, on average, at least millions of hits daily in order to satisfy the users' needs for information. Each search engine has its own sorting policy and the keyword format for the query term, but there are some critical problems. The searches may get more or less information. In the former, the user always gets buried in the information. Requiring only a little information, they always select some former items from the large amount of returned information. In the latter, the user always re-queries using another searching keyword to do searching work. The requery operation also leads to retrieving information in a great amount, which leads to having a large amount of useless information. That is a bad cycle of information retrieval. The similarity Web page retrieval can help avoid browsing the useless information. The similarity Web page retrieval indicates a Web page, and then compares the page with the other Web pages from the searching results of search engines. The similarity Web page retrieval will allow users to save time by not browsing unrelated Web pages and reject non-similar Web pages, rank the similarity order of Web pages and cluster the similarity Web pages into the same classification.

In this article, we will introduce the technology of similarity Web page retrieval which includes searching Web pages and classifying similarity Web pages. For searching Web pages, we will specify the types of search engines and the policy of ranking search results first. And then, two algorithms used for finding related Web pages, the Cocitation algorithm and Latent Linkage Information (LLI) algorithm, will be introduced. The classification method can make the Web pages with similarity vector attributed will be clustered in the same class, and the user retrieves the related Web pages more effectively.

BACKGROUND

The Web Pages Searching

With the growing information of the Internet, search information efficiently is more important than ever to do. Search engines are used to find the Web pages with the same keyword on the Internet (Chakrabarti, Joshi, & Tawde, 2001; Wen-Syan & Candan, 2002). Only by keypressing a keyword, search engines can return information quickly and easily. Jenkins classifies the search engines into three categories (Barroso, Dean, & Holzle, 2003; Jenkins, Kackson, Burden, & Wallis, 1998).The categories include the classified directories search engines, automated search engines, and metasearch engines. The classified directories search the engines' classification of Web data by human. The automated search engines collect the Web resources from the WWW automatically. The metasearch engines provide interface that receives the user query and merges the searching results from various search engines. For the three types of searching, engines will be described briefly as follows (Madria & Bhowmick, 1999). The classified directories' search engines collect the Web data from the WWW by the editorial staff and describe the Web site with text mention. The editors must classify the Web pages to suitable category. Therefore, a user can find the useful information by classification directory easily (Baeza-Yates, 2003). The automated search engines employ the retrieval programming (called Robot or Spider) to collect the Web data from WWW voluntarily, construct the URL (Uniform Resource Locator) index and title, and provide that to the user (Wen-Syan & Candan, 2002). A metasearch engine uses a concept of a computer agent, between the users and many search engines. Its responsibility is to make the users search more conveniently for various aspects of the search engines (Tirri, 2003). A metasearch engine receives user's query and passes the query to various searching engines, if reformatting the query is needed, and then

collects and rearranges the results from the search engines (Madri & Bhowmick, 1999; Baeza-Yates, 2003; Wen-Syan & Candan, 2002). Henzinger, Motwani, and Silverstein indicate a number of reasons why the metasearch engines are needed.

The Web Mining

The mining technology is different from general search work. The mining method can find the data in horizontal relation, and the relation between data and data, that general searching work can not do. When mining technologies are applied to finding Web pages, we called it Web mining. The Web mining framework includes three types which are: Web content mining, Web structure mining, and Web usage mining (shown in Figure1) (Flake, Lawrence, Giles, & Coetzee, 2002; Kao, & Lee, 2000; Madria & Bhowmick, 1999; Sundaresan, & Yi, 2000; Zhang, & Dong, 2002).

In traditional search technologies, the user sends a keyword to a search engine, and then the search engine uses the keyword to find related pages, but it spends too much time and returns too many irrelevant pages. Therefore, the hyperlink analysis and anchor text analysis are proposed to find the related pages. Using the hyperlink structure to find the relative pages can find the related pages quickly and easily, but hyperlink structures have no content information. Hence, there are two assumes in hyperlink study. When a page A links to another page B, indicating page B is recommended by author page A. If two pages have the same links, they might be on the same topic (Henzinger, 2000; Qiu, Hemmje, & Neuhold, 2000; Zhengyu, Qingsheng, & Yukun, 2001). Hence, the hyperlink analysis can be used to ranking pages, Web pages community construction, Web searching improvement, Web clustering and visualization, as well as finding the related pages. Tirri thinks the earlier Web search technologies are not good enough by keyword searching, and he used ontology and hyperlink analysis knowledge to do the searching. He considers the next generation of Web search must have the personal search results, easily find the information that users want and make synonym by itself (Hodgson, 2001; Singh, 2002; Tirri, 2003). We described the similarity Web pages searching technology in the following methods.

.

(1) Cocitation Method

Dean and Henzinge proposed the Cocitation algorithm that uses the hyperlink structure to find the related pages (Dean,& Henzinger, 1999; Jingyu, & Yanchun, 2003). The algorithm is based on two definitions: (A) The two pages are co-cited if they have a common parent page; (B) The number of their common parent pages, named degree of cocitation. In this algorithm, a user can use a query term for a search engine to find a require page, and then can use a URL of the require page to construct a vicinity graph by hyperlink, and then can analyze the graph to find the related pages of the required page. Hence, the algorithm employed the hyperlink of Web pages to find the related page. It is simple and computes easily, but when two pages have the same degree of cocitation, the system cannot decide which Web pages are actually related pages. Also, it may have a topic drift problem if the anchor of a Web page constructs malice hyperlink purposely. The similarity measures between two pages are shown in Figure 2. The page “u” and page “a” have the same parent page A. They are cocited, and the degree of the cocitation is 1. In the same way, the in degree of Page B is 2. Hence, the degree of all pages that related to pages u can be computed.

(2) Latent Linkage Information (LLI) Algorithm

Although, the Cocitation algorithm can evaluate the similarity of Web pages, it cannot determine which pages are better when the pages have the same degree. Therefore, Hou and Zhang utilize the linkage matrix to express the relevance of the Web page topology. The matrix can reveal the depth relationships of the pages, and then find the

Figure 1.Web mining classification