韓國 컴퓨터情報學會 論文誌
第8卷 第2號, 2003. 6.

2003-8-2-1-8

웹 정보의 관리에 있어서 의미적 접근경로의 형성에 관한 연구

이 우 기*

Semantic Access Path Generation in Web Information Management

Wookey Lee

요 약

웹 정보가 폭발적으로 증가하는 정보의 바다 한 가운데에서 이러한 웹 정보를 구조화하는 문제는 매우시의성이 크다. 본 연구는 웹을 노드와 링크로 구성된 그래프로 인식함을 그 출발점으로 한다. 이때 노드는 각 웹 페이지를 나타내고, 링크는 각 페이지를 연결하는 하이퍼텍스트 링크 즉, URL을 나타낸다. 그러면 웹은 웹 페이지와 그들 간의 링크로 연결된 유방향 그래프의 집합이 되는 것이다. 그러나 문제는 이러한 웹 정보가 지수적으로 증가하면서 웹 그래프 역시 지나치게 복잡해짐으로 인해 사용자 즉, 사람 및 검색로봇이 웹 정보를 파악하고 검색함에 있어 복잡성으로 인한 혼란이 야기된다는 것이며, 이를 이름하여 웹 공간에서의 위치혼란(lost in cyber space)라고 부른다. 따라서 이를 적절히 요약 및 추상화하여 방향성(orientation)을 제시하고 전체적으로 웹 공간의 정보를 일목요연하게 표현하는 노력이 필요한 것이다. 이것을 위하여 웹 페이지를 계량적 수치로 나타내야할 필요가 있으며 여기서는 tf-idf를 그 방법론으로 삼았다. tf-idf란 빈도 및 반빈도(term frequency / inverse document frequency)곱을 일컫는 것으로서, 웹 페이지를 용어(keyword)의 벡터로 인식하고, 사용자가 제시하는 용어와의 상관성을 거리공간 벡터값으로 계산하는 과정을 의미한다. 이렇게 웹 정보를 계량화하는 것을 의미적 표현(semantic representation)이라 하고, 그것을 구조화하는 것을 의미적 접근경로라고 하였다. 본 연구의 목표는 궁극적으로는 웹 정보를 의미적 접근경로를 포함하는 계층적 형식(Hierarchical structure)으로 축약하여 사용자로 하여금 웹 정보 검색의 차원을 혁신코자 하는 것이다.

* 현 성결대학교 공과대학 교수

This work was supported by the Korea Science and Engineering Foundation through the Advanced Information Technology Research Center (AITrc), and the author was supported by postdoctoral fellowship program from KOSEF.

I. Introduction

The growing number of web information collections accessible over the global network strongly calls for the assistance of appropriate structuring methods. Web browsers allow users to snoop the World Wide Web with the anchors of hypertext documents. As the browser session progresses, the user is liable to feel lost which is one of the well-known problems: being lost in hyperspace. Using a web browser when the user specifies the URL to jump to a particular site, some information about the Web node's position would be very helpful. If a user wants to compare current pages with past several pages, it would also be up to date a hard work to do. When a user specifies a URL to jump to a particular site, some information about the node's position would be very helpful.

Hasan et al[5]. developed a system that allowed users to get graphical overviews of browsing session in a Web-browsing tool. The views, however, do not suggest the weight of documents corresponding to user's requirements or queries. There have been efforts to visualize the Web information with different points of view. Risse et al[10]. addressed a conceptual approach to generate three-dimensional VRML scenes dynamically from the information stored in a database system. They extended the VRML with a new data type in order to support the server side generation as well as to support the trigger mechanism. Some experimental researches said that graphical representations support better navigation because this type of representation more closely matches a user's mental model of the system[4]. Textual tools, however, suggest further advantages by allowing users to rapidly snoop the extent of Web sites and to search visually in an efficient manner for particular information[5, 6].

Soala [3] is developed with an arc-editing algorithm based on the relative page popularity. It can automatically revise a Web site's page structure to create a more effective hypertext scheme. Unless the mesh structure of a Web site changed to a binary tree, the manipulation with the page popularity would be trifling. There are clustering analyses of effective retrieval in a distributed environment [4] or of a text-database resource discovery [2] that structure the Web pages bag of a retrieval environment around a bag of topics.

Several popular search engines such as AltaVista, Einet Galaxy, Excite, Google, HotBot, Infoseek, Northern Light, Lycos, Yahoo, and Virtual Library attempt to maintain full-text indexes of the World Wide Web. Relying on a single standard search engine, however, has limitations such that they have limited coverage, outdated databases, missing information and are sometimes unavailable[4].

The precision of standard engine results can also vary because they generally focus on handling queries quickly and use relatively simple ranking schemes. Rankings can be further muddled by keyword spamming to increase a page's rank order. Often, the relevance of a particular page is obvious only after loading it and finding the query terms. Metasearch engines, such as MetaCrawler and SavvySearch, attempt to contend with the problem of limited coverage by submitting queries to several standard search engines at once. The primary advantages of meta-search engines are that they combine the results of several search engines and present a consistent user interface. However, most meta-search engines rely on the documents and summaries returned by standard search engines and so inherit their limited precision and vulnerability to keyword spamming[7].

This paper discusses a method to develop the web visualization of the hyperspaces.

Overview diagrams of hyperspaces or part of them have been recognized as important aids to help both orientation and navigation through the web space[5]. The overview shows the web site structure that is consisted of a set of nodes and links. The node is represented as a web document and the link a URL that will be suggested later in detail. It is noted that the web space in this paper is not considered to be a series of individual web documents, but a set of web sites that have individual web documents as components.

This paper is organized as follows. In section 2, we present the data model of Web sites and the Web schema. After a Web site graph preprocessed, a tree structure is derived for the Web visualization in section 3. The Web site graph is viewed with weights and endowed by a semantic measure. The prototype system is explained with its algorithms and motivating examples in section 4. Finally we conclude the paper.

Ⅱ. WEB DATA MODEL

2.1 Web Objects

The Web schema contains the Meta information that represents a bag of Web pages in a Web site. The Web site is a directed graph with Web nodes and arcs, where the Web nodes correspond to HTML files having page contents and the arcs correspond to hypertext links interconnected with the Web pages. The Web node (Wi) is defined as follows :

Wi = [Web_page_identifier, Web_node, Page_link]

Where Web Node = [url, title, meta, format, size, page-depth, modified date, text, weight], and Web Link = [url, parent url, child url, label, url-type]. The Web page contents can be described as the attributes of the Web page such as the Web page identifier, title, Meta, format, size, modified date, text, figures, and multimedia files etc. In this paper for convenience's sake the Web page contents are not described in detail.

2.2 Web Node Processing

A Web site can be viewed as a directed graph that consists of an initial node(called the homepage) and the other nodes inter-connected among them. Complex Web representations do little to help the user orientation within the site and usually tolerate navigation problems themselves. A hierarchical abstraction is useful in organizing information and reducing the number of alternatives that must be considered at any one-time[1, 8]. If the Web site can be represented as a hierarchical structure, those problems such as the multiple paths, the recursive cycle, the multi-path cycle, and multiple parents would be resolved. The problem is treated more specifically in[6, 8].

The Web Node describes the attributes of each web page in the web site, including URL, title, meta, format, size, page-depth, modified date, text, and weight. The page-depth represents the depth of the web page how many steps it is from the home page.

The home page usually indicates the index.html(or default.asp or index.php3) predetermined by the web server. The modified date represents the time-stamp that the web page is created or modified. The weight represents the values resulted from the user's query. It will be discussed later in detail. The Web Link includes the URL of the hypertext, and/or has the parent URL and child URL with label in the web page. The url-type is specified two types: interior link and exterior link. We define the url type is interior link, when the hypertext link points inside of the web site. In reality, the anchor in a web page begins with the directory or with an HTML file, the anchor is represented as an interior link. Even if it begins with a protocol such as 'http:' but has the URL is the same, the anchor is also an interior link. If it begins with a protocol such as 'http:' with different URL, then the url-type is exterior link. In this prototype system, we differentiate the link-type as an exterior link or an interior link.

Example 2-1 : We assume that an example Web site consists of 7 Web pages numbered from 0 to 6 connected with hypertext arcs(Fig. 1). The weight is not defined yet(will be described next section). By the node definition above, the node W0 (home page) has parent URL as Null(for the link of W4 will be discarded in advance), and child URL as W1, W2, and W3. The node W3 has parent URL is W0 and W3(This also makes no sense, for multiple visits to the same node are meaningless), and the child URLs W3, W5, and W5. And the node W6 only has parent URLs W2, W4, and W5.

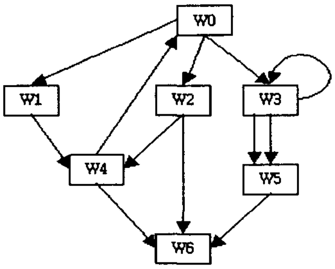


Fig. 1 Example Web site.

In this paper, the URL's are specified two types : interior arcs and exterior arcs. The interior arcs are the URL's that indicate the HTML files somewhere within the Web site, but the exterior arcs out of the Website. We are interested in the interior arcs only, for the structure of a Web site is generated in this paper. But it is needless to differentiate the internal arc in more detail. Because once a Web page is transferred from the Web server, there is no need to access the same Web page physically again.

2.3 Conventional Approaches

Depth first approach(DFA) is easy to adopt to cope with this kind of graphs, for it seems similar with the behaviors of human snoopers. But the DFA seems not applicable in Web environment. Since usual Web pages are complicatedly inter-connected with other Web pages, it may bring about a long series of Web pages. The long series of Web pages may imply long time consumption to access a specific page. On the other hand, there are several strengths in applying the breadth first algorithm (BFA) to Web site graphs. With the BFA an important Web page can easily be accessed. It is done by clicking relatively fewer steps from its homepage rather than by the DFA. It is easy to resolve a graph to a hierarchical tree and to minimize the depths to visit in a Web page. It can also be said that the access time can be minimized.

Ⅲ. WEB REPRESENTATION

3.1 Evaluating the Web

The weight for a Web page indicates how statistically important it is[2, 3, 4]. One common way to compute a Web page(or document) D is the tf-idf that is first to obtain an unnormalized vector $D'{}_1{}'$, ..., $w_{m'}$〉, where each w_i' is the product of a word frequency (*tf*) factor and an inverse document frequency (*idf*) factor. The *tf* factor is equal (or proportional) to the frequency of the i^{th} word within the document. The *idf* factor corresponds to the content discriminating power of the i^{th} word : a word that appears rarely in documents has a high *idf*, while a word that occurs in a large number of documents has a low *idf*. Typically, *idf* is computed by $\log(n/di)$, where n is the total of documents with the i^{th} word. (If a word appears in every document, its discriminating power is 0. If a word appears in a single document, its discriminating power is as large as possible.)

Once D' is computed, the normalized vector D is typically obtained by dividing each w_i' term by $\sqrt{\sum_{i=1}^{m}(w_i')^2}$. The weight can be specified in this paper indicating the importance of the Web page.

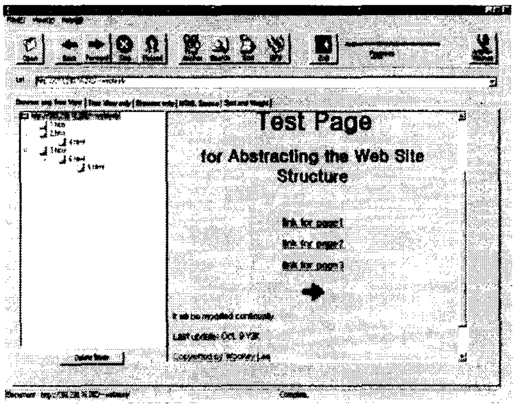
3.2 Semantic Representation

If we can measure the weight of Web nodes corresponding to their significance, then the structure of the nodes can be manipulated by the weight. We introduce the *tf−idf* as the weight measure and it can be used to determine the topological ordering of Web sites. Then simply by comparing the numerical differences of the *tf−idf*, it can be said that a node is closer to a specific node.

In order to investigate the Web site structure, a Web test site [13] is built and implemented for testing the example in Fig. 1. As previously described before, the *tf−idf* measure is applied as a weight of the Web node. The prototype system has been tried to search for the structure of the Web site.

Ⅳ. DESCRIPTIONS ON THE PROTOTYPE SYSTEM

The prototype system is developed to provide users with higher-level summaries and with the structure of Web sites. The system provides a site map, a browser, and a node weight tap. The system has been implemented with VB 6.0 as a client session and with SQL SERVER 7.0 as the server database.



Test Page

for Abstracting the Web Site Structure

link for page1

link for page2

link for page3

Fig. 2 The Prototype system

The system begins with the homepage predetermined by the Web server and uses the interior anchors in the homepage. The anchors are classified with two categories such as interior anchors and exterior anchors. If the homepage begins with a frame without anchors, the system extracts anchors from which the HTML in the frame includes. The result anchors with page contents are stored in the database. The system, for example, finds that the Web test site's homepage includes three interior links(anchor).

The anchors can be expanded in the site map of the system and the example is represented. After getting the url(anchor and frame) extractor module, a web weight vector generator module in the AnchorWoman is represented in Fig. 2. The system can generate the weights of Web pages relevant to the *tf−idf* measure.

V. SUMMARY

The structuring of Web information supports a strong user side viewpoint that a user wants his/her own needs on snooping a specific Web site. Not only the depth first algorithm or the breadth-first algorithm, but also the Web information is abstracted to a hierarchical structure. A prototype system is suggested in order to visualize and to represent a semantic significance. As a motivating example, the Web test site is suggested and analyzed with respect to several keywords. As a future research, the Web site model should be extended to the whole WWW and an accurate assessment function needs to be devised by which several suggested models should be evaluated.

REFERENCES

[1] Chang, C. and Hsu, C. : Enabling Concept-Based Relevance Feedback for Information Retrieval on the WWW, IEEE TKDE, 11(4) : 595-609, 1999.

[2] Garvano, L., Garcia-Molina, H. and Tomasic, A. : GIOSS : Text-Source Discovery over the Internet. ACM TODS, 24(2) : 229-264, 1999.

[3] Garofalakis, J., Kappos, P. and Mourloukos, D. : Web Site Optimization Using Page Popularity, IEEE Internet Computing, 3(4) : 22-29, 1999.

[4] Guillaume, J., Latapy, M. and Viennot, L. : Efficient and Simple Encodings for the Web Graph. Proc. Int'l Conference on WAIM : 328-337, 2002.

[5] Hasan, M. Z., Mendelzon, A. O. and Vista, D. : Applying Database Visualization to the World Wide Web, ACM SIGMOD RECORD, 25(4) : 45-49, 1996.

[6] Huang, M., Eades, P., Wang, J. and Doyle, B. : Dynamic Web Navigation with Information Filtering and Animated Visual Display, Proc. Int'l Conference on APWeb98 : 63-71, 1998.

[7] Lawrence, S., Lee Giles, C. and Bollacker, K. : Digital Libraries and Autonomous Citation Indexing, IEEE Computer, 32(6) : 67-71, 1999.

[8] Lee, W. and Kim, J. : Visualization of Web Site Information with Semantic Weights, Proc. Internet Computing : 253-258, 2000.

[9] Pandurangan, G., Raghavan, P., and Upfal, E. : Using PageRank to Characterize Web Structure. Proc. COCOON : 330-339, 2002.

[10] Risse, T., Leissler, M., Hemmje, M., Aberer, K. and Klement, T. : Supporting dynamic information visualization with VRML and databases, Proc. New paradigms in information visualization and manipulation : 69-72, 1998.

저자소개

이 우 기

서울대학교 학사, 석사, 박사

현 성결대학교 공과대학 교수