IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITB8932

Chapter XIII

Query-By-Structure Approach for the Web

Michael Johnson
Madonna University, USA

Farshad Fotouhi
Wayne State University, USA

Sorin Draghici
Wayne State University, USA

ABSTRACT

This chapter presents three systems that incorporate document structure information into a search of the Web. These systems extend existing Web searches by allowing the user to request documents containing not only specific search words, but also to specify that documents be of a certain type. In addition to being able to search a local database (DB), all three systems are capable of dynamically querying the Web. Each system applies a query-by-structure *approach that captures and utilizes structure information as well as content during a query of the Web. Two of the systems also employ neural networks (NNs) to organize the information based on relevancy of both the content and structure. These systems utilize a supervised Hamming NN and an unsupervised competitive NN, respectively. Initial testing of these systems has shown promising results when compared to straight keyword searches.*

INTRODUCTION

The vast amount of information available to the users of the World Wide Web is overwhelming. However, what is even more overwhelming for users is trying to find the particular information they are looking for. Search engines have been created to assist

in this process, but a typical keyword search using a search engine can still result in hundreds of thousands of different *relevant* Web documents. Savvy search engine users have learned to combine keywords and phrases with logical Boolean operators to pair down the number of *matched* Web pages. Unfortunately, results from these searches can still yield a significant number of pages that must then be viewed individually to determine whether they contain the content the user is interested in or not.

Search engines support keyword searches by utilizing *spiders* or *webots* to scan the textual content of Web documents, and then indexing and storing the content in a database for future user queries. The stored information typically consists of the document URL, various keywords or phrases, and possibly a brief description of the Web page. However, these search engines maintain very little, if any, information about the context in which the text of the Web page is presented. In other words, a search engine might be able to identify that a particular keyword was used in the title of the Web page or a phrase within the anchor tags. But, it would not distinguish between that same word or phrase being used in a paragraph, heading, or as alternate text for an image. However, the way in which text is *presented* in a Web page plays a significant role in the importance of that text. For example, a Web page designer will usually emphasize particularly important words, phrases, or names. By enabling a search engine to capture how text is presented, and subsequently, allowing the users of the search engine to query based on some presentation criteria, the performance of a search can be greatly enhanced. Since search engines that utilize spiders already scan the entire text of a Web page, it is a simple modification to incorporate a mechanism to identify the context in which the text is presented.

Motivation for this type of modification can best be described by illustration. Consider the following examples:

- A user wants to find Web pages containing images of Princess Diana. Using a typical keyword search, any Web page that mentions Princess Diana will be returned in the results of the request. However, by including presentation tags—namely, the HTML *img* tag—as part of the search, the user can specify that she is only interested in Web pages that contain "Princess Diana" and an image. Granted, this particular search would result in any page mentioning Diana that contains an image, regardless of whether the image was of Diana or not. However, many Web page designers include a brief textual description of an image using the *alt* attribute of the image tag. With this added knowledge, the user could specify that the content "Princess Diana" should appear as part of the *alt* attribute within the image tag.
- There are literally thousands of online publications available today on the Web. Most of these publications use a very rigid style for presenting their articles. For example, article authors are usually distinguished by having their name placed in a particular location within the document or highlighted with a specific font type, style, size, or color. If a user is interested in finding articles that were written by a particular author, a simple keyword search could yield many irrelevant Web pages, particularly if the author has a common name, is famous, or even has a famous namesake. However, a search that specifies that a particular name be presented in a particular way could greatly improve the results.