Discovering Document Semantics QBYS: A System for Querying the WWW by Semantics

MICHAEL JOHNSON  mjohnson@madonna.edu
College of Science and Mathematics, Madonna University, Livonia, MI 48150, USA

FARSHAD FOTOUHI  fotouhi@cs.wayne.edu
SORIN DRĂGHICI  sod@cs.wayne.edu
MING DONG  mdong@cs.wayne.edu
DUO XU  xuduo18@cs.wayne.edu
Department of Computer Science, Wayne State University, Detroit, MI 48202, USA

Abstract. This paper describes our research into a query-by-semantics approach to searching the World Wide Web. This research extends existing work, which had focused on a query-by-structure approach for the Web. We present a system that allows users to request documents containing not only specific content information, but also to specify that documents be of a certain type. The system captures and utilizes structure information as well as content during a distributed query of the Web. The system also allows the user the option of creating their own document types by providing the system with example documents. In addition, although the system still gives users the option of dynamically querying the web, the incorporation of a document database has improved the response time involved in the search process. Based on extensive testing and validation presented herein, it is clear that a system that incorporates structure and document semantic information into the query process can significantly improve search results over the standard keyword search.

Keywords: query-by-semantics, document features, document type, neural network

1. Introduction

The task of accessing and processing information in today’s business world is a necessity. However, as more and more information is recorded, being able to sift through all this information to capture what is relevant has become increasingly more difficult. Traditional database systems have, for the most part, been able to keep up with the increased amount of data because hardware technology such as processor speed and memory size has increased as well. However, in recent years a significant paradigm shift towards relying on distributed information repositories has occurred, largely due to the phenomenal growth of the World Wide Web. Once again, database systems were able to keep abreast thanks in part to new networking technology and research into distributed database systems. However, the new-style distributed database systems still rely heavily on the old-style structure of tables, records, and fields that are inherent in traditional database systems. As a result, these systems cannot easily adapt to handling unstructured and semi-structured data.

From a closed-world perspective, not being able to adapt to unstructured or semi-structured data would not be an issue. In fact, it has always been the responsibility of
the data to adapt to the structure of the database system. However, for many database traditionalists who originally perceived the Web as the “goose that laid the golden egg”, it has become a realization that in actuality, the Web is much more like “Pandora’s Box”, and unfortunately, it has already been opened. In the time span of just a few short years, the number of web pages has grown exponentially. There are billions of pages on the Web (according to their website, Google [9] alone has indexed over 3 billion web pages). Although some web sites use traditional-style databases to generate page content, most web pages are written in Hypertext Markup Language (HTML). When viewed in a web browser, these documents are essentially unstructured. However, the source code of an HTML document contains specific structure information relating to how the document should be rendered by the browser. Hence an HTML web page is actually a semi-structured document. Unfortunately, although these types of documents are visually appealing, to the traditional database query system, about the only useful information is the text within the page.

Early web search systems (many of which are still in use today) performed queries of the web by capturing and storing page information in a database. Users then search the database by providing key search words. The results of the search are simply a list of links to the pages that contain the specified words. However, these types of search engines are much less effective today because of the vast size of the Web. This is the case not because the query system cannot handle searching a large amount of data but rather because typical searches can result in thousands of purportedly relevant web documents. Hence, a significant amount of research has gone into improving web searches. However, most of the research into improving searches has involved analyzing the content of a document. Our approach, on the other hand, has been to take advantage of the fact that a web page is a semi-structured document, and that web designers utilize structure to emphasize particularly important aspects of their documents. Our prior research, documented in “A Neural Network Net Query-by-Structure Approach” [15] has shown that capturing and using structure information during the query process can significantly enhance performance.

2. Query-by-semantics system

Our research began with a query by-structure approach using a system we designed called CLaP system [7]. CLaP was designed in an attempt to enhance Web searches by allowing the user to search for web pages containing words that were structured and presented in a specific way within a web page. However, due to the extensive amount of information required of the user during the query, it became apparent that another approach was necessary. Hence, two new systems that employed neural networks to organize the information based on relevancy of both the content and the structure of the document were created [15]. The primary goal of these systems, called AN3 systems, was to test the feasibility of using neural networks to improve query performance while eliminating the need for the complex user queries. The initial results from the testing were promising enough to warrant further research. However, it was clear that significant changes were required. In particular, the system used prototype vectors that were handcrafted to represent the designer’s interpretation of a specific document type.