Chapter VI

The Webspace Method

Roelof van Zwol
University of Twente, The Netherlands

Peter M. G. Apers
University of Twente, The Netherlands

ABSTRACT

The main objective of this chapter is to present the Webspace Method for modeling and querying Web-based document collections. When focusing on limited domains of the Internet, like intranets, digital libraries, and large Web sites, document collections can be found that have a high multimedia and semistructured character. Furthermore, the content of such document collections can be assumed to contain related information. Based on conceptual modelling, the Webspace Method introduces a new schema-based approach to enhance the precision, when searching such document collections.

The Webspace Method allows queries to be composed that combine information stored in several documents to satisfy the user’s information need, whereas traditional search engines are only capable of querying a single document at a time. Furthermore, a query over a Webspace allows a user to formulate his information need as the result of a query over a Webspace directly, rather than a collection of URLs pointing to the possible relevant documents.

This chapter appears in the book, Information Modeling for Internet Applications, edited by Patrick van Bommel. Copyright © 2003, Idea Group Inc. Copying or distributing in print or electronic forms without written permission of Idea Group Inc. is prohibited.
INTRODUCTION

Over time, the Internet has grown into an ever more tangled resource of information. The state-of-the-art means for finding information are text-based search engines like Alta-Vista and Google, hierarchical indexes and directories like Yahoo!. These search engines have to base their tools solely on information retrieval techniques, due to the unstructured nature of the Internet. The diversity, irregularity, and incompleteness of the data involved, make it impossible to use database technology at this global level. Besides that, the data has the tendency to change rapidly over time.

However, when focusing on smaller domains of the WWW, database techniques can be invoked successfully to enhance the precision of the search process. On such domains, large collections of documents can be found, containing related information. Although the conditions are better, one still has to deal with the semistructured and multimedia character of the data involved. The Webspace Method (van Zwol & Apers, 1999) focuses on such domains, like intranets, digital libraries, and large Web sites.

The goal of the Webspace Method is to provide sophisticated search facilities for Web-based document collections, using existing database techniques. To achieve this, three stages are identified for the Webspace Method. The first stage deals with conceptual modelling of Web data. During the second stage, conceptual and multimedia meta-data are extracted from the document collection. Finally, in the last stage, the Webspace Method introduces a new approach to query a collection of Web-based documents.

The key behind the Webspace Method is the introduction of a semantical level, which provides a conceptual description of the content of a Webspace. This semantical level consists of a collection of concepts, which are defined in an object-oriented schema. The object-oriented schema is also referred to as the Webspace schema. The second level of the Webspace Method is called the document level. This level is formed by the Web-based document collection. Each document at the document level of a Webspace consists of a view, corresponding to (a part of) the Webspace schema, to assure the relation between the semantical and physical levels of the Webspace Method. The Webspace schema defined in the first stage is also used to extract the relevant meta-data during the second stage and to formulate queries over a Webspace. Both the semantical and the document level form a Webspace.

Revolutionary within the scope of search engines and query formulation over document collections, is that the Webspace Method allows a user to integrate (combine) information stored in several documents in a single query. Nowadays, traditional search engines, e.g., Alta-Vista and Google, are only capable of querying a single document at a time. As a result, the query will return