Chapter X

Flexible Digital Library Search

Menzo Windhouwer and Albrecht Schmidt
CWI Amsterdam, The Netherlands

Roelof van Zwol, Milan Petkovic, and Henk Ernst Blok
University of Twente, The Netherlands

ABSTRACT

In this chapter the development of a specialised search engine for a digital library is described. The proposed system architecture consists of three levels: the conceptual, the logical and the physical level. The conceptual level schema enables by its exposure of a domain specific schema semantically rich conceptual search. The logical level provides a description language to achieve a high degree of flexibility for multimedia retrieval. The physical level takes care of scalable and efficient persistent data storage. The role, played by each level, changes during the various stages of a search engine’s lifecycle: (1) modeling the index, (2) populating and maintaining the index and (3) querying the index. The integration of all this functionality allows the combination of both conceptual and content-based querying in the query stage. A search engine for the Australian Open tennis tournament website is used as a running example, which shows the power of the complete architecture and its various components.
INTRODUCTION

Unfortunately, the Internet is still, in the eyes of many of its users, what appears to be a chaotic organization of information sources. The state-of-the-art means for finding information are full text-based search engines, e.g., Alta Vista and Google, and hierarchical indexes or directories, e.g., Yahoo. They constitute entry points based solely on the textual content of web pages. Still much research is needed to extend the capabilities of search engines to other media such as audio, image or video. The first steps have been made, but the area is still very much in its infancy.

The challenge lies in the presentation orientation of the Internet: site designers usually have more semantic information than is visible on the Internet. For example: the designer of a tennis site knows that “Monica Seles” is the name of a female tennis player. Making this extra domain knowledge explicit would allow a search engine to restrict the range of possible semantic domains and to use more specific multimedia retrieval techniques. This knowledge could thus be used to semantically enrich both indexes and queries. The larger scale and unmoderated nature of the Internet as a whole impedes this approach. And for intranets this approach still incurs high additional costs over deploying a standard search engine restricted to a certain IP-domain.

The integrated database approach, described in this chapter, lowers the costs of developing specialised search engines. For the unlimited, Internet-scale domain, it still uses well-known textual retrieval techniques but, instead of using a special-purpose index structure, now transparently integrated into a DBMS. The proposed architecture consists of three levels: the conceptual, the logical and the physical level. The conceptual level schema enables by its exposure of a domain-specific schema semantically rich conceptual search. The logical level provides a description language to achieve a high degree of flexibility for multimedia retrieval. The physical level takes care of scalable and efficient persistent data storage.

The role played by each level changes during the various stages of a search engine’s lifecycle. This lifecycle consists of the following stages: (1) The initial phase is the (re)creation of the schema which underlies the search engine, describing both the domain specific data and the multimedia meta-data. (2) In the next stage this schema is populated with conceptual data and multimedia meta-data. The schema, the source data and the extraction algorithms may all change, so the stored data has to be maintained to keep its validity. (3) Concurrently to the maintenance stage, the search engine is used for querying.

The integration of all this functionality allows the combination of both conceptual and content-based querying in the query stage. This integration is missing in traditional search engines. It gives users more powerful query primitives. For example, on the Internet scale, the user can ask the following query: “Show me all portraits embedded in pages containing keywords semantically related to the
The Big Data Processing of HF Sky-Wave Radar Sea Echo for Detection of Sea Moving Targets
International Journal of Information Technology and Web Engineering (pp. 56-71).
www.igi-global.com/article/the-big-data-processing-of-hf-sky-wave-radar-sea-echo-for-detection-of-sea-moving-targets/188382?camid=4v1a