Search Support in Data Management Systems

Andreas Henrich

Otto-Friedrich-Universität Bamberg, Germany
andreas.henrich@wiai.uni-bamberg.de

Abstract. In consequence of the change in the nature of data management systems the requirements for search support have shifted. In the early days of data management systems, efficient access techniques and optimization strategies for exact match queries had been the main focus. Most of the problems in this field are satisfactorily solved today and new types of applications for data management systems have turned the focus of current research to content-based similarity queries and queries on distributed databases. The present contribution addresses these two aspects. In the first part, algorithms and data structures supporting similarity queries are presented together with considerations about their integration in data management systems, whereas search techniques for distributed data management systems and especially for peer-to-peer networks are discussed in the second part. Here, techniques for exact match queries and for similarity queries are addressed.

1 Motivation

For decades one main focus of data management systems had been the efficient processing of *fact queries*. A typical—yet rather simple—example would be the search for all open invoices for a customer with a given reference number. Current database management systems are optimized to process large numbers of such queries on dynamic data sets and most problems related to queries of that type are already solved more or less satisfactorily (see [12], for example).

On the other hand, the growing amount of digital documents and digital multimedia documents shifts the requirements for data management systems. In situations where text documents or multimedia documents have to be maintained, a strong requirement for content-based queries is induced. Here a typical query is given by an example document or a query text describing the requested documents. Given a set of images, a query can for example be defined by an example image and in this case query processing is concerned with finding *similar* images. Content-based retrieval has been considered for a long time in the area of information retrieval (see, e.g., [1]). However, information retrieval (IR) has mainly addressed flat text documents in the past. Today, structured

T. Härder and W. Lehner (Eds.): Data Management (Wedekind Festschrift), LNCS 3551, pp. 137-157, 2005.

documents (for example, XML documents) and also structured multimedia documents have to be maintained. Many commercial database management systems claim to be the most suitable systems to manage XML data and multimedia data. As a consequence, sophisticated content-based retrieval facilities will be an important distinguishing feature of data management systems.

Another important aspect in data management systems is the decentralized character of cutting edge data management systems. One important trend in this respect are peer-to-peer networks (P2P networks) which are made up of autonomous peers contributing to an administration-free overlay network. The efficient processing of similarity queries in these networks is an important field of research and not yet satisfactorily solved.

In the rest of this contribution, we will first discuss techniques and algorithms for the efficient processing of complex similarity queries in a local scenario. Thereafter, we will discuss approaches towards an efficient processing of content-based similarity queries in P2P networks.

2 Processing Complex Similarity Queries

In recent years, structured multimedia data has become one of the most challenging application areas for data management systems, and search support for structured multimedia data is an important research topic in this respect. To emphasize the key problems in this field, let us assume tree-structured multimedia documents, in which the internal nodes represent intermediate components, such as chapters or sections, and where the leaves represent single media objects such as text, image, video, or audio. In order to support the search for such documents—or document fragments—we need search services which address the following requirements [15]:

- *Dealing with Structured Documents:* The fact that documents are complex-structured objects in our scenario, brings up various interesting research issues.
  First, the search support component of a data management system must allow to search for arbitrary granules ranging from whole documents over intermediate chunks to single media objects. Second, with structured documents many properties of an object are not directly attached to the object itself, but to its components. For example, the text of a chapter will usually be stored in separate text objects associated with the chapter object via links or relationships. Third, additional information about an atomic media object can be found in its *vicinity*. Exploiting the structure of a multimedia document, this concept of vicinity can be addressed navigating one link up and then down to the sibling components.
- *Feature Extraction and Segmentation:* With multimedia data the semantics is usually given implicitly in the media objects. For example, an image might represent a certain mood. Therefore, a search support component should allow to extract features from the media objects potentially describing their semantics. Furthermore, it should