Flexible and scalable digital library search

Henk Ernst Blok[1] Menzo Windhouwer[2] Roelof van Zwol[1] Milan Petkovic[1]
Peter M.G. Apers[1] Martin Kersten[2] Willem Jonker[1]
{blok,zwol,milan,apers,jonker}@cs.utwente.nl, {windhouw,mk}@cwi.nl

[1] University of Twente, Enschede, The Netherlands
PO BOX 217, NL 7500 AE, Enschede, The Netherlands
tel. +31 53 489 3690, fax. +31 53 489 2927
[2] CWI, Amsterdam, The Netherlands

1 Introduction

The everlasting search for new methods to explore the Inter- or Intranet is still going on. In this demo we present the combined effort of the AMIS and DMW research projects, each covering significant parts of this problem.

The contribution of this demo is twofold. Firstly, we demonstrate how feature grammars offer a flexible solution for extraction and querying of meta-data from multimedia documents in general. Scalability and efficiency support are illustrated for full text indexing and retrieval. Secondly, we show how for a more limited domain, like an Intranet, conceptual modeling can offer additional and more powerful query facilities. The limited domain case, also allows the extraction and querying of high-level concepts from raw video data.

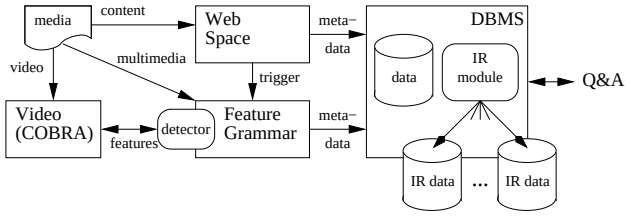

Figure 1: System architecture

Figure 1 shows the integrated architecture of the digital library/web search engine. We refer to the demo website for further information:

http://www.cs.utwente.nl/∼dmw/VLDB2001/

Proceedings of the 27th VLDB Conference, Roma, Italy, 2001

2 Internet search engines

Current search engines provide access to the wealth of textual information on the web by offering keyword based search. However, hidden inside the multimedia content, additional information exists. Feature detectors can be used to extract this information. The retrieval process can be enriched with content-based facilities by storing these features in a meta-index.

To populate the index, the detectors should be executed in the right order, thus a high level description of their dependencies is needed. The *feature grammar*, which forms the core of the Acoi system [WSK99], describes the relationships between meta-data, detectors, and mutual detectors in a set of grammar rules.

Managing the meta-index is done by exploiting the dependencies in the feature grammar. To populate the meta-index the feature grammar is used to generate a parser: the *Feature Detector Engine* (FDE). While proofing that the start symbol of the grammar is valid, the FDE will execute the detectors. These detectors produce the meta-data, and the FDE stores it in the parse tree. When the start rule is valid, the parse tree can be stored in the meta-index.

The flexibility of the feature grammar pays off when either the source data or the feature detector algorithms change, and the meta-index needs to be updated. The *Feature Detector Scheduler* (FDS) analyses the transition graph, deduced from the grammar, to localize the effects of changes, and trigger incremental parses. An incremental parse adds new branches to an existing parse tree or updates the old branches.

Since, the meta-data size will be huge when indexing the entire web, good scalability and efficiency are of crucial importance. Distribution and query optimization are the typical database means to achieve this. For practical reasons we limited the scalability and optimization research to full text information re-

Related publications from the University of Twente and CWI may be found at http://db.cs.utwente.nl/publication.xml *and* http://www.cwi.nl/htbin/ins1/publications *respectively.*

trieval (IR), but we intend to extent the facilities to full fledged multimedia support. IR queries, and multimedia queries in general, typically are top-N queries, *i.e.*, they request a top of best ranked objects. Therefore, we are mostly interested in optimization support for such queries.

By fragmenting [BVBA01] the meta-data tables horizontally, we provide a good means for distribution. Furthermore, the fragmentation facilitates set oriented optimization of the top-N queries. By using a quality prediction model in combination with a typical cost model , the optimizer can trade speed for quality when deciding which fragments to use in the query evaluation. Note that IR is imprecise by nature, so giving up a little quality for a large speed gain can be very usefull. This in particularly holds when the querying is done itteratively when using feedback from the user.

3 Intranet search engines

Finding relevant information on the World-Wide Web (WWW) is often a frustrating task, due to its unstructured nature. Moreover, from a modeling point of view, the topics dealt with on the Internet are too diverse to capture (model) in a database schema. This makes it infeasible to apply database techniques directly to the Internet at a large scale basis.

However, when focusing on smaller portions of the WWW, database techniques can be successfully invoked for the search process. There large collections of documents can be found, containing related information. But, one still has to deal with the *semi-structured* and *multimedia* character of the data involved. The Webspace Method focuses on such domains, like Intranets and large web-sites.

The main contribution of the Webspace Method is that it introduces a whole new category of search engines, by using query formulation techniques, formerly only available within a database environment. More general, it provides an approach for (1) modeling web-data, (2) meta-data extraction and multimedia indexing, and (3) query formulation over a document collection, i.e. a webspace, based on an object-oriented schema.

The object-oriented schema, also called the webspace schema, is formed by concepts. These concepts describe the content of a webspace at a high conceptual level. At the physical level documents are defined, which form a view on (a part of) the webspace schema. The documents are marked up with XML, and form a materialized view over the webspace schema, since they contain both schema and content.

Once a webspace is correctly defined the meta-data extraction phase is started[ZA00]. Conceptual information is extracted from the XML documents, and stored in the meta index. To deal with the multimedia objects involved, the Feature Grammar Engine is integrated in the Webspace Extractor.

Using the concepts defined for a webspace in the query, allows far more specific and complex queries to be formulated over a webspace[ZA00], as compared to standard search engines, which can just deliver a document's URL, based on query formed by keywords. Based on the results of a retrieval performance experiment, we can conclude that searching web-data, as proposed by the Webspace Method results in a significant increase of performance, measured in terms of recall and precision.

One of the multimedia types found on the Australian Open webspace is Video. In order to explore video content and provide automatic extraction of semantic concepts (objects and events) from raw video data, we propose the *COntent-Based RetrievAl* (COBRA) video model [PJ01]. The model is independent of feature/semantic extractors, providing flexibility by using different video processing and pattern recognition techniques. It also includes object and event grammars that formalize the descriptions of these high-level concepts, as well as facilitate their extraction based on features and spatio-temporal reasoning [PJ01].

This rule-based approach results in the automatic mapping from features to high-level concepts. Therefore, a user is able to explore video content specifying very detailed complex queries that include a combination of features, objects, and events, as well as spatio-temporal relations among them.

References

[BVBA01] Henk Ernst Blok, Arjen P. de Vries, Henk M. Blanken, and Peter M.G. Apers, *Experiences with IR TOP N Optimization in a Main Memory DBMS: Applying 'the Database Approach' in New Domains*, Advances in Databases, 18th British National Conference on Databases, BNCOD 18, Lecture Notes in Computer Science, Springer, July 2001, To appear.

[PJ01] M. Petkovic and W. Jonker, *Content-based retrieval of spatio-temporal video events*, Multimedia Computing Track of IRMA Intl. Conf. (Toronto, Canada), May 2001.

[WSK99] M. A. Windhouwer, A. R. Schmidt, and M. L. Kersten, *Acoi: A System for Indexing Multimedia Objects*, International Workshop on Information Integration and Web-based Applications & Services (Yogyakarta, Indonesia), November 1999.

[ZA00] R. van Zwol and P.M.G. Apers, *The webspace method: On the integration of database technology with information retrieval*, In proceedings of CIKM'00 (Washington, DC.), November 2000.