Implementing Powerful Retrieval Capabilities in a Distributed Environment for Libraries and Archives[1]

Chrisa Tsinaraki, George Anestis, Nektarios Moumoutzis, Stavros Christodoulakis

Laboratory of Distributed Multimedia Information Systems and Applications (MUSIC)
Technical University of Crete
P.O. Box 134, GR, 73100 Chania, Greece
e-mail: {chrisa, ganest, nektar, stavros}@ced.tuc.gr

An on-line distributed environment, which was implemented in the context of the VENIVA project for historical libraries and archives is presented here. Emphasis is given in the presentation of the powerful search capabilities provided to the end-users, which are typical for the end-users (either researchers or ordinary people) of any Digital Library environment.

The information managed resides in a number of different relational databases in one or more institutions (i.e. Libraries and Historical Archives). The end-user of the system uses a WWW client to pose traditional boolean queries, similarity queries or complex queries containing both boolean and similarity terms on the contents of the databases. In the case of similarity queries, the end-user can also select the evaluation formula used to rank the objects that the system returns as the answer to his query. This gives a flexibility to experiment with alternative retrieval models without starting the implementation from scratch. A *Graphical Query Editor* is used in order to construct the queries.

The innovative aspect of this work is that the similarity queries are translated by an appropriate component of the server into a series of traditional SQL queries, so that there is no need to have separate systems to support the various services offered. Only a standard relational DBMS (Database Management System) is used in the core of the system. The software layer that has been implemented on top gives all the additional flexibility. The implementation is based on a sound and flexible mathematical retrieval model.

The framework, which supports queries containing both boolean and similarity terms in the context of VENIVA, is based on a powerful mathematical tool for the description of similarity queries and different evaluation models. It has been implemented on top of a relational DBMS as a distributed system. This framework aims to integrate *Information Retrieval System (IRS)* techniques on top of traditional relational DBMS. In particular, a mathematical model for the description of Information Retrieval Systems based on different fuzzy models has been developed [1].

The framework has been implemented [2] on top of a relational DBMS as a distributed system. The general architecture of this system is shown in Fig. 1. *Schema editor* is a graphical tool, which is used to define the mapping between fuzzy relations and relational tables. It is also responsible to create the relational tables that are used to store the fuzzy relations corresponding to similarity queries. The mappings are then stored in the *Map File*. Based on these mappings, the *Retrieval Engine* is able to translate retrieval requests from any *Application* into SQL queries sent to the *Relational DB* and give back the results

[1] Support for this work was provided by the VENIVA (VENetIan Virtual Archive) ESPRIT project (EP N° 20638)

C. Nikolaou, C. Stephanidis (Eds.): ECDL'98, LNCS 1513, pp. 653-655, 1998.

to the *Application* when requested. The *Retrieval Engine* is also responsible for decomposing a retrieval request in a series of INSERT statements to populate the tables used to store the queries. The current implementation supports the Microsoft Access DBMS and offers queries in conjunctive normal form.

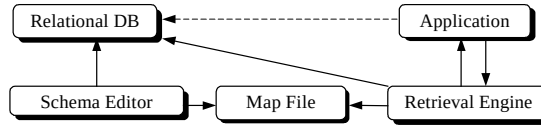

Fig. 1. System Architecture

The procedure we follow in order to use the framework has the following steps:

- Using Schema Editor, on a relational database, the schema of the IRS is defined. Namely there are defined entity sets, relationship sets and fuzzy attributes.
- Retrieval Engine is started, taking as input the schema of the IRS that was created in the previous step using Schema Editor.
- The application that offers the user interface for the formation of the queries and submission is created. The application communicates with Retrieval Engine and sends queries to it. In addition, the application takes care of the presentation of the results to the user. The application implemented in the context of VENIVA is the Graphical Query Editor.

We have presented here the flexible search capabilities provided in the context of the VENIVA project. These services are implemented on a homogenized schema based on popular standards for the description of the contents of Libraries and Historical Archives. The advanced Information Retrieval capabilities supported are built on top of existing relational database technology which has been properly extended to support ranking of objects with different evaluation formulas. The presented approach provides a practical methodology for publishing the contents of historical institutions to the broad public using existing mature technologies such as WWW and Relational DBMSs.

Future research will focus on the integration of relevance feedback techniques so that the end-user can refine his query based on the investigation of the objects that meet his needs. Moreover, support for textual attributes will be integrated with thesauri and special access methods.

References

[1] N. Moumoutzis: “The Design of a System Supporting the Development of Interactive Geographical Applications”, MEng Thesis, Department of Electronic and Computer Engineering, Technical University of Crete, Chania, 1998.

[2] G. Anestis: “Design and Implementation of a Boolean and Similarity Retrieval System on top of Relational Database Management Systems”, Diploma Thesis, Department of Electronic and Computer Engineering, Technical University of Crete, Chania, 1997.

[3] K. Beard, V. Sharma: “Multidimensional ranking for data in digital spatial libraries”, International Journal on Digital Libraries, vol. 1, number 2, pp 153-160, September 1997.

[4] S. DeFacio, A. Daoud, L. A. Smith, J. Srinivasan: “Integrating IR and RDBMS Using Cooperative Indexing” In Proceedings of the 18th ACM SIGIR International Conference on Research and Development in Information Retrieval, 1995, pp 84-92.

[5] S. Dessloch, N. Mattos: "Integrating SQL Databases with Content-specific Search Engines", Proceedings of the 23rd VLDB Conference Athens, Greece, 1997, pp 528-537.
[6] R. Fagin: "Combining Fuzzy Information from Multiple Systems", Proceedings of the 15th Symposium on Principles of Database Systems, Montreal, Canada, June 1996, pp 216-226.
[7] D. A. Grossman, O. Frieder, D. O. Holmes, D. C. Roberts: "Integrating Structured Data and Text: A Relational Approach", Journal of the American Society of Information Science, vol. 48, no. 2, February 1997.
[8] L. Gravano, H. Garcia-Molina: "Merging Ranks from Heterogeneous Internet Sources", Proceedings of the 23rd VLDB Conference Athens, Greece, 1997, pp 196-205.
[9] L. Haas, D. Kossmann, E. Wimmers, J. Yang: "Optimizing Queries across Diverse Data Sources", Proceedings of the 23rd VLDB Conference Athens, Greece, 1997, pp 276-285.
[10] J. Hammer, M. Breunig, H. Garcia-Molina, S. Nestorov, V. Vassalos, R. Yerneni: "Template-based wrappers in the TSIMMIS system", In Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data, pp 532-535.
[11] G. Hjaltason., H. Samet: "Ranking in Spatial Databases", Proceedings of 4th Symposium on Advances in Spatial Databases (SSD '95), Lecture Notes in Computer Science No. 951, Springer-Verlag, 1995, 83-95.
[12] Ad Hoc Commission on Descriptive Standards, "ISAD(G): General International Standard Archival Description"
[13] Working Group on the General International Standard Bibliographic Description set by IFLA, "International Standard Bibliographic Description", 1977
[14] H. V. Jagadish, A. O. Mendelzon, T. Milo: "Similarity-Based Queries", in proceedings of the 14th Symposium on Principles of Database Systems, San Jose CA, USA, 1995, pp 36-45.
[15] Joon Ho Lee, "Properties of Extended Boolean Models in Information Retrieval", In Proceedings of the 17th ACM SIGIR International Conference on Research and Development in Information Retrieval, 1994, 182-190.
[16] R. Kruse, J. Gebhardt, F. Klawonn: "Foundations of Fuzzy Systems", John Wiley & Sons Ltd., ISBN 0-471-94243-X, 1994
[17] C. Lynch, M. Stonebraker: "Extending User-Defined Indexing with Applications to Textual Databases", Proceedings of the 14th VLDB Conference Los Angeles CA, USA, 1988, pp 306-317.
[18] Avram, Henriette D. The MARC pilot project: final report on a project sponsored by the Council on Library Resources, Inc. Washington: Library of Congress, 1968.
[19] M. Persin: "Document Filtering for Fast Ranking", ", In Proceedings of the 17th ACM SIGIR International Conference on Research and Development in Information Retrieval, 1994, pp 339-348.
[20] G. Schmidt, T. Ströhlein: "Relations and Graphs", Springer-Verlag, ISBN 3-540-56254-0, 1993.
[21] G. Shalton: "The use of Extended Boolean Logic in Information Retrieval", In Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data, pp 277-285.
[22] G. Shalton: "Automatic Text Processing", Addison-Wesley, Reading, MA, 1989
[23] V. Vassalos, Y. Papakonstantinou: "Describing and Using Query Capabilities of Heterogeneous Sources", Proceedings of the 23rd VLDB Conference Athens, Greece, 1997, pp 256-265.
[24] W. Moen, M. Tucker, "A Guide to Global Z39.50"