Integrated Search Tools for Newspaper Digital Libraries

S.L. Mantzaris, B. Gatos, N.Gouraros and P. Tzavelis
Department of Digital Technologies, Lambrakis Press Archives,
{slm, bgat, nikosg, panost}@dolnet.gr
8, Heyden Str., 104 34 Athens, Greece

Our project aims at the creation of a digital library of newspaper issues dated since 1890. At the moment, all the available source material is property of Lambrakis Press SA, the largest publisher in Greece. The printed material exceeds 1,200,000 pages, half of which are of A2 size. In order to facilitate the access to the digitized material, we have developed a retro-conversion procedure [1,2] according to which articles are traced and catalogued. For the time being, the full text of the articles is only partially available. The reasons for this are that in some cases we have encountered low quality originals, rare old fonts (in our case old Greek fonts) as well as the absence of a suitable dictionary that could correct the OCR outcome and make the recognition reliable. In addition, an integrated set of search tools is provided to the users so that they can easily find the information they are looking for. Finally, when the user decides to see an article, the entire newspaper page image is displayed having the requested article marked appropriately. In this demo, we are focusing in presenting the search tools.

An investigation in our corporate users has shown that a search mechanism should meet the following requirements:

1) To support users of different mentality in searching. Old habits like exact word matching and Boolean queries should be supported since some users do not want to change the way they do things.
2) To support users of different levels of sophistication. Some users are novice searchers while others have significant experience using Internet search tools.
3) To provide simple ways in order to help users to express their information need and to present the results to them.

We believe that a broader spectrum of users would have similar requirements.

In our system a user can search in articles metadata and content, where as content we consider the full text or the titling (headline, top-title, subtitle) in case full text is not available. User weighting of query parts is supported. The search result lists are combined to give a unified result list. This combination is either a fuzzy conjunction in the spirit of [3] or a ranked combination. In the latter case, we additionally take into consideration the number of matching fields between the article and the query.

The FTR engine holds a central position in our system. In order to design it we have taken into account the following points:

- The subjects of the articles vary.
- The articles cover a time period of over 100 years.
- The extensive NLP support of the Greek language. Our previous experience has indicated that simple NLP, for example stemming, improves significantly the search performance [4]. This can be attributed to the fact that for the Greek language a noun has at least four word forms and a verb has at least ten. Furthermore, the difference between modern and older Greek is extended to changes in the spelling of entire words.
- The flexibility of expressing a text query as a set of simple and compound terms, where a simple term consists of a word or a stem or a wildcard expression and a compound term is a set of simple terms which coexist in an article under certain restrictions (word distance, order, paragraph self containment)
- The support of natural language queries and structured queries. To this end we have implemented vector space and extended boolean models. The weighting schemes used are inspired by the work in [5,6].
- The support of concurrent update and search operations.

Regarding the implementation details, metadata are stored in an RDBMS (MS SQL Server 7.0) while images are stored on filesystems. To handle FTR we decided to have an inverted list implementation on top of conventional database tables. The evaluation of the user query is done by SQL statements and middle tier program objects, which are handled by MS Transaction Server.

Currently, we are interested in improving the search effectiveness of small queries.

SIGIR 2000 7/00 Athens, Greece

References

[1] B. Gatos, S.L. Mantzaris, K.V. Chandrinos, A. Tsigris and S.J. Perantonis, "Integrated Algorithms for Newspaper Page Decomposition and Article Tracking," *Proc. of the 5th Intern. Conf. on Document Analysis and Recognition* (ICDAR'99), pp. 559-562, Bangalore, India, September 1999.

[2] B. Gatos, S.L. Mantzaris, S.J. Perantonis and A. Tsigris, "Automatic Page Analysis of a Digital Library from Newspaper Archives", *International Journal of Digital Libraries* (IJODL), to appear in spring 2000 issue.

[3] R. Fagin, "Fuzzy queries in multimedia database systems," *Proc. 17th ACM Symp. on Principles of Database Systems*, Seattle, ed. J. Paredaens, pp. 1-10, 1998.

[4] S.L. Mantzaris, B. Gatos, N. Gouraros and S.J. Perantonis, "Linking Article Parts for the Creation of a Newspaper Digital Library," *Content-Based Multimedia Information Access Intern. Conf.* (RIAO2000), pp. 997-1004.

[5] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," In D.K. Harman ed., TREC-3, NIST, Gaithersburg, MD, 1995.

[6] A. Singhal, "Term weighting revisited," Ph. D. Thesis, Dept Computer Science, Cornell Univ., Ithaca 1997, available also as TR97-1626 technical report.