Similarity Searching in Text Databases with Multiple Field Types

Kostas Tzeras
GMD-IPSI
Darmstadt, Germany
tzeras@darmstadt.gmd.de

Euripides G.M. Petrakis
Dept. of Electronic and Comp. Engineering
Technical University of Crete
petrakis@ced.tuc.gr

Abstract

Similarity searching in text databases with multiple field types is still an open problem. We experimented with CORDIS and we evaluated the effectiveness of many text retrieval methods in terms of precision, recall and ranking quality.

1. Introduction

On text databases with multiple fields, different field types should be handled with different retrieval methods. An example is CORDIS[1], the research database of the European Union. Currently, CORDIS offers simple search capabilities based on Boolean queries and exact string matching.

2. Methodology and Experiments

CORDIS consists of over 250 fields. In this work we used the fields of Table 1. These are the most commonly used fields and typically occur in text databases with similar structure.

Field name	Num of records	Mean char length
SURNAME	18,971	8.5
TITLE	28,051	86.64
ABSTRACT	18,940	256.37
GEN_INFORM	18,832	1024.21

Table 1. CORDIS fields used in this work.

We used human relevance judgements and the "pooling method" of [2] to compute the effectiveness of each candidate method. The queries were randomly selected from the database and our measurements represent the average over 20 queries on responses with the best 50 answers.

[1] `http://apollo.cordis.lu/cordis/GLOBALsearch.html`

An access method for proper names must be able to retrieve mispelt names. The competing methods are (a) Ukkonen n-grams with $n = 2$ and $n = 3$ and (b) Damerau-Levenstein and (c) Edit distance. Ukkonen digrams perform better than any other method achieving recall 82% and precision 14%.

For longer fields, the queries are short sentences randomly selected from the stored titles. Longer queries are less likely to be formulated by a typical CORDIS user. The candidate methods are (a) SMART [1], (b) LSI [4] and (c) INQUERY [3]. For searching on field TITLE, INQUERY performs slightly better than SMART in both precision (29% and 28% respectively) and recall (77% and 72% respectively) but SMART has better ranking (0.827 and 0.779). On field ABSTRACT, SMART and INQUERY achieve precision 25% but SMART has better recall (58% and 43% respectively) and should be preferred, although INQUERY has slightly better ranking (0.711 and 0.696). On field GEN_INFORMATION, INQUERY is slightly better than SMART in both precision (29% and 28% respectively) and recall (68% and 66%) but SMART has better ranking (0.779 and 0.709). Both SMART and INQUERY are far more effective than LSI. LSI did not exceed recall 30% or precision 12% in all cases.

3. Conclusions

Digrams (for proper names), INQUERY and SMART (for longer fields) provide a good basis for the design of a similarity retrieval mechanism for CORDIS.

References

[1] C. Buckley. *SMART, Version 7*. Computer Science Department, Cornell University.

[2] D. Harman, editor. *The Second Text Retrieval Conference*. National Institute of Standards and Technology Gaithersburg, MD, 1994.

[3] *INQUERY, Release Version 3.0*. University of Massachusetts, Computer Science Dept., June 1995.

[4] *LSI Software (for research purposes only)*. Bellcore Communications Research Inc., 1990.