The LIVE-project

Retrieval Experiments Based on Evaluation Viewpoints

P. Bollmann, F. Jochum, U. Reiner, V. Weissmann, H. Zuse
Technische Universitaet Berlin

Extended Abstract

The LIVE-project (Leistungsbewertung von Information Retrieval Verfahren) at the Technische Universitaet Berlin, West Germany is concerned with the evaluation of information retrieval systems. Two fields are mainly under investigation.

One area is about the investigation of methodological foundations of retrieval experiments. There are many authors /1/ who state, that there are still many problems to be solved. A summary on these problems can be found in /2/. Results of the LIVE-project in this area can be seen on three different areas: on the one hand measurement-theoretical criteria for the application of similarity and evaluation measures in retrieval experiments have been considered and developed. Some of the results can be found in /3/, /4/, /5/. Further work has been done in the application of statistical principles of experimental design in information retrieval. Especially control structures of factors in retrieval tests have been investigated and some aspects of statistical models for experimentation in information retrieval. Some of these results can be found in /6/, /7/, /8/, /9/.

The other topic in the LIVE-project — which is the main content of this paper — is the conduction of a retrieval experiment in co-operation with FIZ4 (FachInformationsZentrum 4) which is an information service center for mathematics, physics and energy in Karlsruhe, West Germany. Amongst the many databases which FIZ4 is offering the LIVE-project uses the database about physics in their retrieval experiment. FIZ4 is using the information retrieval system GRIPS (General Relation based Information Processing System) which was developed by DIMDI (Deutsches Institut fuer Medizinische Dokumentation und Information). The query language of GRIPS is an extended Boolean language /10/. Besides the operators 'and', 'or' and 'not' the GRIPS retrieval language contains thesaurus — operators to extend the query and truncation — and context-operators for freetext and Boolean searching.

To a given query GRIPS partitions the document collection into two sets: the retrieved and the not retrieved documents. To the user the retrieved documents are presented in the reversed order of their registration into the database. Under the assumption that this temporal order is not correlated with relevance the set of retrieved documents is considered as an unordered set.

The research hypotheses for our retrieval experiment in this context is: 'the retrieval output of the GRIPS-system can be improved by the application of a similarity measure'.

The evaluation was done by comparing the unordered outputs of the GRIPS-system with ordered outputs, which were obtained by applying a similarity measure to the retrieved documents. For the evaluation problems arise such as: which objects are measured, what does 'better' mean, what is an appropriate measure, which averaging methods should be used?

In the LIVE-project we did not use the usual evaluation measures such as recall-precision-graph because the measurement values are difficult to interpret. Instead in co-operation with FIZ4 evaluation viewpoints were defined. An example of an evaluation viewpoint is: 'a user wants exactly five relevant documents and wants to get as few nonrelevant documents as possible. A retrieval result R1 is better than another one R2 if the user gets fewer nonrelevant documents with R1 than with R2'. For this evaluation viewpoint the 'expected search length' of Cooper /11/ is an appropriate evaluation measure. As in the viewpoint defined above only the order of preference of retrieval results is defined, the measure may only be used as an ordinal scale. To use it as an interval scale the assumption is made, that every additional nonrelevant document which the user retrieves causes the same additional amount of unproductive labour.

In a similar way several other viewpoints and evaluation measures were defined and applied in the retrieval experiment. Under the assumption that the evaluation measure is an interval scale averaging is done by calculating the arithmetic mean.

As levels of the experimental factor /8/ the following similarity measures were used: inner product measure, cosine measure, overlap measure, coefficient of Jaccard and Euclidean distance. As situative factors /8/ the number of documents retrieved by GRIPS, the number of descriptiors of the queries, generality and topic of documents were used.

The retrieval experiment is not yet finished completely but several results have already been obtained. For example in the average (over 81 queries) for the above defined viewpoint the ranking with the inner product measure does not indicate a significant improvement compared with the GRIPS-output. In the case of the Euclidean distance measure it seems that in the average the user has to inspect less nonrelevant documents. This means an improvement compared with the unordered retrieved set from the GRIPS-output.

For more details of the so called 'two-level retrieval process' and further experimental results we refer to the long version of this paper.

References

/1/ Sparck Jones, K. (ed.):
Information Retrieval Experiment.
Butterworths, London, 1981

/2/ Reiner, U.:
Experimente im Gebiet des Information Retrieval — Ueberblick und Stand der Forschung.
TU Berlin, Fachbereich Informatik, Institut fuer Angewandte Informatik, LIVE-Bericht Nr. 6/83

/3/ Bollmann, P.:
The Normalized Recall and Related Measures.
Proceedings of the Sixth Annual International ACM Conference 'Research and Development in Information Retrieval', June 6 - 8, 1983, Bethesda, Maryland

/4/ Bollmann, P.; Weissmann, V.:
Probleme der Mittelwertbildung.
Deutscher Dokumentartag, Goettingen, 1983, Saur-Verlag, Muenchen

/5/ Bollmann, P.:
Two Axioms for Evaluation Measures in Information Retrieval.
Proceedings of the Third Joint BCS and ACM Symposium, King's College, Cambridge, 1984

/6/ Jockum, F.; Reiner, U.:
Probleme bei der Planung von Information Retrieval Experimenten.
Deutscher Dokumentartag, Goettingen, 1983, Saur-Verlag, Muenchen

/7/ Jockum, F.; Weissman, V.:
Struktur und Elemente des Information Retrieval Experiments.
Studien zur Klassifikation, Band 14, Indeksverlag, Frankfurt am Main, 1985

/8/ Jochum, F.:
On Retrieval tests with an Inhomogeneous Query Collection.
Proceedings of the Eighth Annual International ACM Conference 'Research and Development in Information Retrieval', 1985, Montreal, Canada

/9/ Weissman, V.E.:
Zur Rolle der Statistik bei experimentellen Untersuchungen.
TU Berlin, Fachbereich Informatik, Institut fuer Angewandte Informatik, LIVE-Bericht NR. 8/83

/10/ Konrad, E.; Reiner, U:
Eine semantische Analyse der Suchkomponente des Information-Retrieval-Systems GRIPS.
in preparation (LIVE-Bericht Nr. 2/86)

/11/ Cooper, W.S.:
Expected Search Length: A Single Measure of Retrieval Effectiveness Based on the Weak Ordering Action of Retrieval Systems.
American Documentation, Vol. 19, pp 30-41, 1968