The Smart Project in Automatic Document Retrieval

Panel Session

Chairman: Gerard Salton, Cornell University

Michael E. Lesk, Bellcore Inc.

- Early Smart developments and applications in future retrieval environments.

Donna Harman, National Institute of Standards and Technology

- Early years at Cornell – Evaluation Issues Revisited.

Robert E. Williamson, Natural Language Access Systems

- How to get the Smart retrieval techniques implemented in the real world.

Edward A. Fox, Virginia Polytechnic Institute and State University

- New developments in information retrieval systems and technology.

Chris Buckley, Cornell University

- Recent Smart implementation and research.

Abstract

The Smart project in automatic text retrieval was started in 1961. It is the oldest, continuously running research project in information retrieval. The panel members are all major contributors to the Smart system work. The discussion covers aspects of the Smart system design and examines the past and future significance of some of the research conducted in the Smart environment.

The Smart Document Retrieval Project

Gerard Salton*

The Smart project was initiated at Harvard University in 1961 with the receipt of a research grant from the Air Force Cambridge Research Laboratory. The first technical report in a series of 22 covering developments in information storage and retrieval was issued at the Harvard Computation Laboratory in November 1961.[1] The National Science Foundation took over sponsorship of the Smart retrieval research in 1964, and NSF has supported the work in one form or another ever since. Much of the research effort was moved from Harvard to Cornell University in 1965, but a research group was maintained at Harvard until 1969.

To support the experimental aspects of the research, computer programs were developed, designed to provide a flexible environment for the automatic analysis, search and retrieval of natural-language texts. Several generations of the Smart system programs were eventually produced, the first one being an implementation in the middle 1960s for an IBM 7094 computer. A second design was prepared at Cornell in the early 1970s for use on IBM 360 and 370 equipment.[2] This was followed in the later 1970s by a partial UNIX-based implementation for Digital Equipment PDP 11-80 and VAX 780 equipment, using the INGRES database system for all file processing operations. The current full, on-line implementation under UNIX operates with various kinds of VAX equipment and most recently with SUN workstations.

From the beginning, the basic premise was that an automatic analysis of query and document texts would form the basis for the retrieval activities. In addition, various kinds of refinements beyond raw matching of query and document words were thought to be essential to obtain reasonable retrieval effectiveness. Accordingly, the basic Smart system design was based on the use of various kinds of stored dictionaries, word suffix lists, phrase tables, and hierarchical term arrangements. Provisions were made early on for the generation of word phrases (in addition to single words) using either statistical word co-occurrence methods or full syntactic analysis methodologies, and for vocabulary expansion by means of phrase dictionaries and preconstructed term hierarchies.[3, 4]

The earliest retrieval experiments were performed with collections of a few hundred documents abstracts, and dictionaries of less than 1,000 entries. The first test results, based on the use of modified recall and precision measurements, indicated that some of the more refined text indexing methods did not provide the expected improvements in retrieval effectiveness. For example, the hierarchical term expansion methods which were to supply narrower and broader term assignments than the ones originally available, were not useful in improving the retrieval output. Similarly, the early tests suggested that contrary to expectation, term phrases obtained by word co-occurrence methods were more powerful in a retrieval setting than phrases generated by syntactic methodologies.[5] Such results were confirmed later by many additional experiments. Almost from the start it became obvious that some form of user-system interaction would be beneficial in the retrieval operations, and the available experimental output suggested that substantial advantages were obtainable with the well-known relevance feedback methodologies.[6] In consequence, relevance feedback and other retrieval strategies introduced in the Smart environment have been utilized over the years in many other retrieval situations.[7-9]

In addition to the system design work, the Smart project work led to advances in many other aspects of automatic text manipulation, including the introduction of new retrieval models (the vector space model

*Department of Computer Science, Cornell University, Ithaca, NY 14853-7501.

and the extended Boolean model), the generation of new automatic indexing methods (term discrimination model, term weighting, term phrase construction), the automatic structuring of text collections (document clustering and cluster searching, dynamic modification of document collections, and text linking methods), and the development of retrieval evaluation methodologies.

Several dozen researchers contributed substantially to the Smart work, among them E.H. Sussenguth, Jr. (tree search methods); E.M. Keen (system evaluation); J.J. Rocchio, Jr. and E.R. Ide (relevance feedback); T.L. Brauen and M.D. Kerchner (document space modification); S.F. Weiss, R.G. Crawford, C.S. Yang and J.L. Fagan (text analysis); D.M. Murray, R.T. Dattola, Y.C.A. Wong, and E.M. Voorhees (collection clustering); C.T. Yu and H. Wu (retrieval modeling), and finally M.E. Lesk, R.E. Williamson, D. Harman, E.A. Fox and C. Buckley (system design and many other areas).

At the present time, much effort is being devoted around the world to the improvement of text analysis and retrieval operations using advanced developments in statistics and probability theory, logic, computational linguistics, and various aspects of artificial intelligence. The Smart project work is thus a precursor of things to come in information retrieval research and the implementation of modern text processing environments.

References

1. Staff of the Computation Laboratory, Information Storage and Retrieval, Scientific Report ISR-1, Harvard University, Cambridge MA, November 1961.

2. G. Salton, ed., The Smart Retrieval System – Experiments in Automatic Document Processing, Prentice Hall Inc., Englewood Cliffs NJ, 1971.

3. G. Salton, A Flexible Automatic System for the Organization, Storage, and Retrieval of Language Data (Smart), Scientific Report ISR-5, Section I, Harvard Computation Laboratory, Cambridge MA, January 1964.

4. M.E. Lesk, The Smart Automatic Text Processing and Document Retrieval System, Scientific Report ISR-8, Section II, Harvard Computation Laboratory, Cambridge MA, December 1964.

5. G. Salton and M.E. Lesk, The Smart Automatic Retrieval System - An Illustration, Communications of the ACM, 8:6, June 1965, 392-398.

6. J.J. Rocchio Jr, Relevance Feedback in Information Retrieval, Scientific Report ISR-9, Section 23, Harvard Computation Laboratory, Cambridge MA, August 1965.

7. V. Vernimb, Automatic Query Adjustment in Document Retrieval, Information Processing and Management, 13:6, 1977, 339-353.

8. T. Noreault, M. Koll and M.J. McGill, Automatic Ranked Output from Boolean Searches in SIRE, Journal of the ASIS, 28:6, November 1977, 333-337.

9. C. Stanfill and B. Kahle, Parallel Free-Text Search on the Connection Machine System, Communication of the ACM, 29:12, December 1986, 1229-1239.