Saving and accessing the old IR literature

Donna Harman
NIST, Gaithersburg, USA
donna.harman@nist.gov

Djoerd Hiemstra
University of Twente, The Netherlands
hiemstra@cs.utwente.nl

Abstract

This short paper describes the beginnings of a project to digitize some of the older literature in the information retrieval field. So far 14 of the older reports, such as the Cranfield reports and ISR reports have been scanned, along with Karen Sparck Jones's *Information Retrieval Experiment* book. The PDF versions of these are available from the SIGIR web site, which includes a new "museum" of information retrieval that allows searching of this material and provides room for exhibits of historic interest. The paper finishes with some thoughts for future work on making more of our IR literature available for searching.

1 Introduction

As more and more of the world becomes digital, and documents become easily available over the Internet, we are suddenly able to access all kinds of information. The downside of this however is that information that is not digital becomes less accessed, and is liable to be lost to us and to future generations. Whereas there are many scanning projects underway, such as Google books and the Open Library Alliance, these projects are not going to know about, much less find, the specialized scientific literature within various fields.

For this reason a committee of us IR junkies started talking about organizing our own scanning effort back in 2005. We created a list of what would be important to collect, looked at various options for allowing access to this information, and then dreamed a lot about how wonderful it would be to have a digital library for just us IR folks. The ACM digital library had little interest in what they called the "grey literature", probably because there is no end to the amount of stuff that could be collected. Additionally they maintain their library in a proprietary format that would not allow us to insert our own search engine (or multiple engines), which seemed like a good thing for IR people to want to do.

In the end we decided to do our own scanning (professionally, with SIGIR funding) and make the results of that scan available to the whole SIGIR community, both as "raw" text and hopefully within a digital library. We have started with the old literature, but, at least for some of us including Karen Sparck Jones, there was the dream of one day being able to search all of our literature, such as SIGIR proceedings, TREC proceedings, and journal articles. For an example of this, see the wonderful ACL anthology (`http://www.aclweb.org/anthology-index/`), a major resource for the NLP community.

The rest of this article describes the progress so far, what is available and how to get it, and future plans.

The list, as of the end of 2007, contained over 20 titles of "grey literature", including assorted Cranfield reports, the Salton ISR series, many British library reports, plus about 10 miscellaneous ones. There are also 13 books plus 3 proceedings, all out of print. We started with those items that had no copyright issues and that we had access to (many thanks to Doug Oard at the University of Maryland for some "oldies"). This is what is in the initial set of scanned documents that is now finished. The documents were scanned at 600 bpi greyscale and turned into PDF with hidden text via OCR.

Reports

1. Cranfield I report (1960), Cyril Cleverdon, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems* – This is the report on the initial stage of the Cranfield I experiment, setting out the details of the design of the experiment and the execution of the first stage. The experiment, funded by the U.S. National Science Foundation, was to test the effects of four different manual indexing schemes (Universal Decimal Classification, Alphabetic Subject Index, a facetted scheme, and the Uniterm system for co-ordinate indexing), especially in light of future developments in electronic versions of the card catalog. There were 18,000 documents indexed using the four schemes, and this report provides extensive details on the careful design of the project in order to deal with indexer variance, learning effects, etc., and then reports on other issues that were encountered. The report is an excellent example of a well-designed user study!
2. Cranfield I report (1962), Cyril Cleverdon, *Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems* – This report covers the second stage of the Cranfield I experiment, where the indexed documents were then manually searched in order to compare the indexing methods. There were 1200 questions, each submitted by a researcher, and each containing the "source" document for that question. This was defined as the document "which .... would have provided, in the opinion of the compiler of the question, a satisfactory answer to that particular question". Today we would call this a known item search. The report contains the results of the testing, including statistical tests, and the number of successful and failed searches. There are also detailed failure analyses for each index as to why the source document could not be found using this index.
3. Cranfield II report (1966), Cyril Cleverdon, Jack Mills, and Michael Keen, *Factors Determining the Performance of Indexing Systems, Volume 1, Design* – This report starts with an interesting discussion (refutation?) by Cleverdon on the criticisms of Cranfield I, including its use of recall and precision as the measures. This in itself is an interesting commentary on the issues of the times with respect to information retrieval. Chapter 2 lays out the decisions leading to the test design for Cranfield II, known today as the Cranfield paradigm, and resulting in the first re-usable test collection (the Cranfield collection). The goal of this experiment was to move beyond the known item search into true searching experiments and to test specific **types** of indexing methods as opposed to specific indexing schemes as in Cranfield I. These types included the use of single words included in the text, the addition of synonyms, the use of word forms (suffixing), the use of a thesaurus hierarchy, etc. These terms could be manually assigned (at three levels of exhaustivity), or just taken from the natural language of

the abstracts and titles. The report contains incredible detail on how the experiment was run; something to marvel at in these days of running an experiment in a couple of minutes!!

4. Cranfield II report (1966), appendix to Volume 1, contains the input data, etc. for the Cranfield II experiment. This means all of the questions, relevance judgments, the thesaurus, etc. It is **all** here.

5. Cranfield II report (1966), Cyril Cleverdon and Michael Keen, *Factors Determining the Performance of Indexing Systems, Volume 2* – This report contains the results and analysis of the Cranfield II experiment. There are several chapters on the various "display" methods for the results and also various evaluation measures. Of course there are many tables and graphs but maybe the most interesting part of the report, at least to me, is the final chapter presenting the conclusions. Basically it was found that using single terms, as opposed to thesaurus, synonyms, etc. was the best and Cleverdon says "This conclusion is so controversial and so unexpected that it is bound to throw considerable doubt on the methods which have been used .... A complete recheck has failed to reveal any discrepancies .... there is no other course except to attempt to explain the results which seem to offend against every canon on which we were trained as librarians." Indeed, the report did create a furor, however; it also established the basis of the automatic indexing that we do today!!

6. *Measures and Averaging Methods Used in Performance Testing of Indexing Systems* (1966), Michael Keen – This short report contains two sections by Mike Keen, the person mostly responsible for the evaluation measures and statistical testing for the Cranfield II experiments. These sections detail the measures used (and reasons why), and also the averaging methods then in vogue, both at Cranfield and for SMART. It should be required reading for anyone interested in the evolution of our measures during this timeframe.

7. *The Effect of Variations in Relevance Assessments in Comparative Experimental Tests of Index Languages* (1970), Cyril Cleverdon – This report discusses (refutes?) some criticisms of the Cranfield II experiments, particularly the effects of variations in relevance judgments. There was a small (20 question) experiment using 6 different indexing languages, and artificially changing the judgment set by randomly substituting non-relevant documents for relevant ones. There was little change in the order of the results until almost half of the relevant documents have been replaced (similar to what we see today in terms of stability of the test collections).

8. Scientific Report No. ISR-10 (March, 1966), Gerard Salton – This report is Joe Rocchio's thesis, which is often referenced due to the Rocchio feedback algorithm (which it contains). However it also has chapters on automatic indexing, request (question) formulation, matching functions and evaluation.

9. Scientific Report No. ISR-11 (June, 1966), Gerard Salton – This report contains 20 separate papers, including a description of the Harvard version of the SMART system (Lesk), an article by Salton and Lesk on dictionary construction and experiments (stems vs full word vs thesaurus), and three student papers on relevance feedback, request (question) clustering, and document clustering.

10. Scientific Report No. ISR-12 (June, 1967), Gerard Salton – This report contains 12 separate papers, including 7 student papers. These papers are the result of student projects

in such areas as clustering, user interaction, negative relevance feedback, document transformation, the use of bibliographic data, and evaluation of relevance feedback. The report also contains a paper by Salton on optimization of retrieval effectiveness (using a cluster search), a paper by Lesk on statistical testing, and two papers on the current status of the "new" Cornell SMART system.

11. Scientific Report No. ISR-13 (December, 1967), Gerard Salton – This report contains 20 separate papers, including seven by Mike Keen, who was visiting Cornell for a year. These seven papers deal with the test collections and procedures, evaluation measures, and an analysis of the requests (questions). Additionally he reports on four experiments on search matching functions, correlation measures, document length (title vs abstract), suffix dictionaries, and thesaurus, phrase and hierarchy dictionaries. There is a paper by Mike Lesk on word associations and one by Dattola and Murray on automatic thesaurus construction. The papers by Mike Keen, especially the one on evaluation measures, are classics, and are the source of the chapters in the SMART book.

12. Scientific Report No. IRS-15 (1969), Gerard Salton – This report is Eleanor Ide's thesis on relevance feedback. It was the most complete work on relevance feedback up to this point, starting with the Rocchio formula and doing many variations to further test results. It also contains work on evaluation, in particular working with frozen ranked results and some of the other methods used during that time to compensate for the reranking of the relevant documents that were used in the feedback process. It is a good source for information on evaluation of relevance feedback.

13. NIST Monograph 91 (1965, revised 1969), Mary Elizabeth Stevens, *Automatic Indexing: A State of the Art Report* – This monograph contains a survey of the literature on automatic indexing, including machine compiled KWIC indexes, concordances, etc. In addition to surveying the various automatic abstracting techniques, such as Luhn, Edmundson, Wyllys, or Oswald, there is also a chapter on automatic assignment of keywords such as by Swanson, Maron or Williams. The survey includes a chapter on statistical association techniques (Doyle, Giuliano, etc.) and one on the problems of evaluation, which discusses the Cranfield project. There is an extensive bibliography. This survey is one good starting point for those interested in getting some idea of the work done in the late '50s and '60s, particularly from an author **not** associated with the specific research projects.

14. *Evaluation of the MEDLARS DEMAND SEARCH SERVICE* (January 1968), F.W. Lancaster – This is a report on the large evaluation of the MEDLARS service, commissioned by the National Library of Medicine. Some of the goals of this project were to learn the search requirements of the users, to "recognize factors adversely affecting the performance of MEDLARS", and to provide a corpus of documents, requests, and relevance assessments that could be used for further experimentation. Like the Cranfield tests, this test was meticulously designed. One of the problems was finding ways of calculating the recall of the searches and this was resolved by created a "recall base" that consisted of documents that were known before the search began. Note that there was a single recall-precision number that was calculated at the end of a search rather than the continuous numbers we see today (because this was a manual Boolean operation). The report contains exhaustive failure analysis as to why certain searches failed. The collection built for this evaluation was "adapted" by Salton and the Cornell group to form a much smaller test collection known as MEDLARS; this was because the

computers were totally unable to search the size of the MEDLARS collection without using Boolean retrieval at that time.

Books

1. *Information Retrieval Experiment* (1981), Karen Sparck Jones – This book was the classic resource for experimentation in information retrieval; in many ways it still is! The book consists of three parts. The first part deals with testing in general, including chapters by Steve Robertson, Keith van Rijsbergen, Nick Belkin and Jean Tague. The Tague chapter, "The pragmatics of information retrieval experimentation", is a tour de force in how to think about design of an experiment. Part 2 gives specifics about different types of testing, including chapters by F.W. Lancaster, Mike Keen, and Bob Oddy on operational testing and laboratory testing and a fascinating chapter by Bill Cooper on Gedanken experimentation. Part 3 has two blockbuster chapters by Karen (70 pages in all), covering reviews of the retrieval testing from 1958-1978, including 30 pages on the Cranfield tests. There is also a short chapter by Salton on the SMART system.

2 A Museum of IR Research

Two master students of the University of Twente, Tristan Pothoven and Marijn van Vliet developed a method to access the the old IR literature by means of a *Digital Museum of Information Retrieval Research*. The aim of the system is to be more than a traditional digital library by also providing the stories behind the reports and papers. The museum provides two access methods. First of course, the contents can be accessed via a simple *search box*. Second, the contents can be accessed by means of *exhibitions*. The museum is availble from: `http://www.sigir.org/museum/`

Simple, but targeted search

What would a museum of IR be without the possibility to search the information? The museum comes up with a simple search box. Keyword queries allow to search the full text of the collection, that was derived by optical character recognition. A search will find the full reports, or report sections. The system provides relatively long snippets that are displayed in a special panel to support detailed exploration of the collection. Of course, the user may always open the PDF document containing the original report.

Museum exhibitions

Exhibitions in the IR Museum pop up like traditional books with "real" pages that need to be turned by dragging them from right to left. The museum currently has a few short exhibitions, for instance one providing background information for the Cranfield reports, and another providing information about the SMART project. Exhibitions provide a guided tour along the reports and papers that fall under the exhibition's theme.

Technical details and availability

The IR Museum is built using Adobe Flex and PF/Tijah. PF/Tijah is a flexible open source text search system developed at the University of Twente in cooperation with CWI Amsterdam and the University of Munich, and can be downloaded as part of the MonetDB/XQuery database system. Just as the contents of the museum itself, the code for the museum is freely available from the PF/Tijah documentation for others to improve and/or re-use, see: `http://dbappl.cs.utwente.nl/pftijah`.

3 Conclusion

What next? Well, we will be continuing the scanning project. The next step is to tackle the reports from the British Library; we are currently compiling a list and contacting them for permission. Several others have offered their books (copyright free) and at some point we would like to seriously go after the Salton books.

In the meantime, take a look at what we have, think about interesting ways of accessing it, and try out the FIRST access method detailed earlier. Note that this project can provide some interesting challenges, particularly to the digital library community, including how to better access structured documents, how to deal with vocabulary shift over the years, and studies on how a user community would like to access this type of information.

And of course, we hope that this small beginning can one day lead to a digital library for the community that contains more than just "old stuff"!!