InTEXT Precision Indexing in TREC4

Mark Burnett, Craig Fisher and Richard Jones

InTEXT Systems,

P.O. Box 310 Deakin West

ACT 2600 Australia
100032.1075@compuserve.com

A. Background

InTEXT Systems is a subsidiary of the CP Software Group, headquartered in Folsom California. The company is a leader in the provision of advanced tools and end-user products for intelligent document management. The InTEXT Research and Development group, based in Canberra Australia, has been in existence for 10 years, and develops leading edge technology in the area of text retrieval, indexing, routing, content analysis and document clustering, using their Heuristic Learning architecture.

The InTEXT Systems R and D group participated in TREC-1, when it formed the Centre for Electronic Document Research (CEDR), in conjunction with CITRI (Collaborative Information Technology Research Institute). [1]

B. Objectives

The principal objective of the experiment INTXT2, which was carried out within the AdHoc (manual) query section of TREC-4, was to establish the effectiveness of the index term selection techniques developed for the InTEXT Precision product.

Unlike most developments in full text retrieval, which concentrate entirely on methods to obtain optimal retrieval from a document collection where all words except a global stop-word list are included as index terms, Precision focuses on improving the process at the indexing stage by being more selective in the words chosen for indexing. In effect Precision selects a stop-word list per document. As the name suggest, a primary objective of the exercise is to improve the precision of the retrieval, it is hoped without significantly reducing the recall.

The objective of the INTXT2 experiment is to establish the effectiveness of the Precision methods for full text retrieval, using a commercially available text retrieval engine (InTEXT

Retrieval Engine) as the retrieval vehicle. In particular, it was hoped that considerably reduced hit lists would be obtained with high precision.

C. The Experiment

The original text files were split into documents, and all fields except the document reference and the full text were stripped. These were run through InTEXT Precision and two derived documents were created from each original. These were:

- a PreciseScope document conforming to the markup for InTEXT Retrieval Engine with stop words inserted to replace the words not to be indexed;
- a Keyword file of keys and associated weights. This was not used in the experiment.

A fragment of an original TREC document together with the two forms of Precision output is shown in Appendix 1. The complete set of PreciseScope documents and weighted keywords are available on request.

Text indexing and retrieval was performed by InTEXT Retrieval Engine, a commercially available full-text retrieval engine which ranks retrieved documents by heuristics that take into account the number and position of search terms within a document and compares these against a system-determined judgment of relevance for these terms.

The queries were created manually, using the TREC test document set with additional research at the National Library of Australia. It was found that a substantial knowledge of USA current affairs was required to understand the implication of some of the queries.

D. The InTEXT Precision Method

The Precision indexing approach uses the techniques employed in the InTEXT Systems Heuristic Learning Architecture, combining elements from:

- identification of document surface structure: headings, paragraphs;
- location of the author's message: cue phrases, repetition;
- Learning from collection characteristics: frequency, discriminating power.

Term Generation: Precision operates by generating a large set of phrases, making use of a large stop word list (some 2,000 words). Grammatical variants are combined using a modified Lovins algorithm [2] and a score assigned to each phrase. This score is computed based upon the frequency and length of each phrase and its position within the document (in heading, first sentence of paragraph etc.) The combined phrases are ordered by their score, and the sum of the scores of the keys defines the Total Information Content (TIC) of the document.

Term Selection: Three classes of document are considered. For documents with a very low TIC, it is recognised that the Precision method does not work effectively and all non-stop words are indexed. A threshold was chosen for this experiment such that very few documents fell into this class. (Typically documents containing <200 words in total). Below the second threshold (about 500 words), all generated terms are used whose combined score adds up to a proportion of the TIC (80% for the experiment). Above this threshold, preference is given to terms that are noun like, particularly proper nouns, again up to a selected proportion of the document TIC.

PreciseScope Document creation: The selected terms including variants are mapped back onto the original document. Noise words are inserted in place of the other words so the positions of the selected words and phrases are correctly preserved. In this way, collocation searches (in same paragraph, within x words) work correctly. The process may be viewed as a dynamic selection of stop words for each document, defined as those which are not important in the document for defining its information content.

The process typically results in 10-20% of the words of the document being indexed. The whole process, including reading and writing all files, processed 30-40 Mbytes of source documents an hour on a 90Mhz Pentium.

The principal benefits of the Precision approach outlined above include:

- much smaller "hit lists" with higher precision, this is key in large document collections. No-one wants to read 1,000 documents in the hope that a few are relevant;
- quicker resolution of information requests;
- smaller index files and quicker generation and update of textbases.

The technique gives the potential for a new generation of full text retrieval, focusing on automatic index term selection, rather than purely on relevance measurement at query time. It is in principle compatible with any existing full text retrieval system.

E. Results of the Experiments

Across all documents some 225K different words were indexed, with a total word instance count of approximately 48M [3].

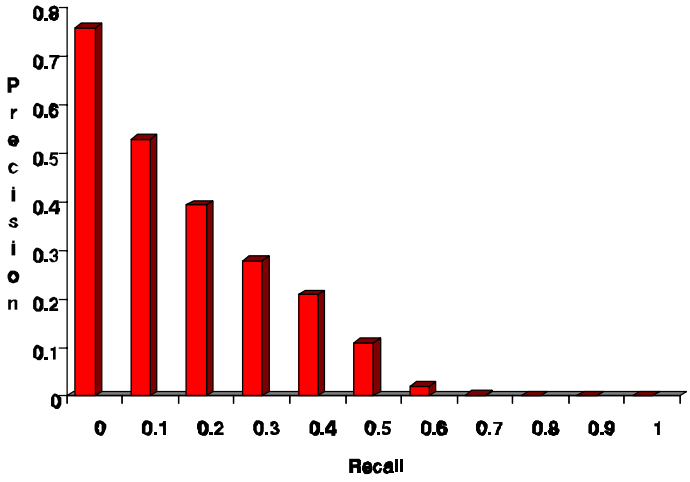
Some of the basic statistics across all queries are as follows:

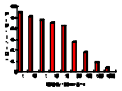
- Total number of documents retrieved over all queries: 8992
- Total number of relevant documents in collection: 6501
- Total number of relevant and retrieved documents: 2161

- The interpolated Recall-Precision averages across all queries are shown in Figure 1.
- Average precision (non-interpolated) over all relevant documents is 0.1814
- The Precision averages across all queries are given in Figure 2.
- R-Precision (precision after R (= num_rel for a query) docs retrieved):
  Exact: 0.2615

Figure 1 - INTXT2 Precision/ Recall all Queries

Figure 2 - INTXT2 Precision/ Relevant Documents all Queries





F. Analysis

When viewed in the light of the experiment objectives, the results are very encouraging. 25% of retrieved documents are relevant and 33% of documents adjudged relevant were retrieved.

It is interesting to note that the precision recall figures are almost identical to those obtained by the InTEXT Routing experiments in TREC-1 [1]. The methods used for determining relevance are very similar to those in InTEXT Retrieval Engine. The difference is of course that in the current experiment, many fewer index terms (one in eight) are being utilised.

At a more detailed level, probably the most favourable statistics were obtained for the number of relevant retrieved at 100. In four queries the result obtained was the highest, in one of these 95 relevant documents were identified. The mean Precision figure for the top 30 retrieved documents across all queries was 0.42 (see Figure 2), a particularly good result. In no statistic was the INTXT2 result the worst!

On detailed analysis of some of the more disappointing queries, it seems that a major reason for poor performance was omission of search terms in the query rather than failure to select important terms for indexing. This is important given the objective of the experiment, and can clearly be addressed by more powerful query term selection methods, such as thesaurus use or automatic term expansion.

G. Conclusion

It is claimed that on the basis of the results of the INTXT2 experiment, the use of PreciseScope documents is a viable way of automatically indexing documents for full text retrieval, particularly where precision is a priority. As mentioned earlier, the InTEXT Precision output from this experiment is available on request

References

[1] R.L. Jones, S.K. Leung, D.L. Pape, "Application of the Automatic Message Router to the Tipster Collection", The First Text REtrieval Conference (TREC-1), National Institute of Standards and Technology, Special Publication 500-207, Gatithersburg, MD. 20899

[2] J. Lovins, "Development of a Stemming Algorithm", Mechanical Translations and Computational Linguistics, **11**, (1968), pp11-32

[3] J. Zobel, RMIT, Melbourne, Private communication

Appendix 1 Example Precision Documents

Original Document

AP901221-0008
<DOC>
.....
<TEXT>
Researchers in California and Utah report, in studies published today, new advances toward isolating the gene linked to a susceptibility to breast cancer in some families.
.....
Skolnick and King said their studies could lead eventually to tests enabling women with known breast cancer risks, based on their family histories, to be treated even before the smallest of tumors have formed.
.....
Science, which published the studies, is the journal of the American Association for the Advancement of Science.
</TEXT>
</DOC>

PreciseScope Document

$$T
AP901221-0008
.....
... a10 STUDIES a01 a02 ... a13 BREAST CANCER a01 a02 FAMILIES
$$P
.....
SKOLNICK a01 KING a01 a02 STUDIES a01 a02 ... a09 BREAST CANCER RISKS a01 a02
$$P
.....
... a04 SCIENCE a01 a02 a03 STUDIES a01 a02
$$P
... a11 SCIENCE
$$N ENDTEXT

Precision Weighted Keys

AP901221-0008
$$N KEYS
#breast#cancer=2050.6
#breast#cancer#risk=926.0
#proliferative#breast#disease=294.5
#breast#cancer#gene=276.5
#inherited#risk#of#breast=211.1
#cancer#risk= 71.2
#King= 64.0
#Skolnick= 36.0
#Science= 25.0
#PBD= 9.0
$$N ENDKEYS