Report on the Glasgow IR group (glair4) submission

Mark Sanderson & Ian Ruthven

Department of Computing Science
University of Glasgow
Glasgow G12 8QQ, UK

1 Introduction

This year's submission from the Glasgow IR group (glair4) is to the category B automatic ad hoc section. Due to pressures of time and unexpected complications, our intended application of a technique known as generalised imaging *[Crestani 95]* was not completed in time for the TREC deadline. Therefore, the submission is the output of an IR system running a simplistic retrieval strategy, similar to last year's submission though with some intended improvements. It would appear from comparison with other category B submissions that this strategy is relatively successful.

The following sections of this report contain a description of the retrieval strategy used, a analysis of the results, and finally, a discussion of our intentions for TREC 6.

2 Methodology

The retrieval strategy used was a 'text book' approach. The words of the collection and query documents had their case normalised. Any words appearing in a stop list (the creation of which is described below) were removed. The remaining terms were applied to Porter's stemming algorithm *[Porter 80]*. Document terms were weighted used a *tf•idf* scheme as shown in Equation 1, taken from *[Crestani 95]*. A document was scored in relation to a query by summing the weights of those query terms found within it.

$$w_{ij} = \frac{\log(freq_{ij} + 1)}{\log(length_j)} \bullet \log\left(\frac{N}{n_i}\right) \qquad (1)$$

w_{ij} = tf•idf weight of term i in document j

$freq_{ij}$ = frequency of term i in document j

$length_j$ = number of unique terms in document j

N = number of documents in collection

n_i = number of documents term i occurs in

The stop word list was chosen after a short study of retrieval effectiveness when different lists were used. Those tried were no stop list, the stop list found in Van Rijsbergen *[Van Rijsbergen 79]*, and a list composed of words whose frequency of occurrence, within the document collection being retrieved from, is greater than some level. A number of frequency of occurrence levels were examined, that which was found to produce the highest effectiveness was composed of words that occurred in more that 7.5% of the document collection. As can be seen in

Figure 1, in comparison with the other stop lists, this type of list produced the best effectiveness, therefore it was used in this year's TREC submission.

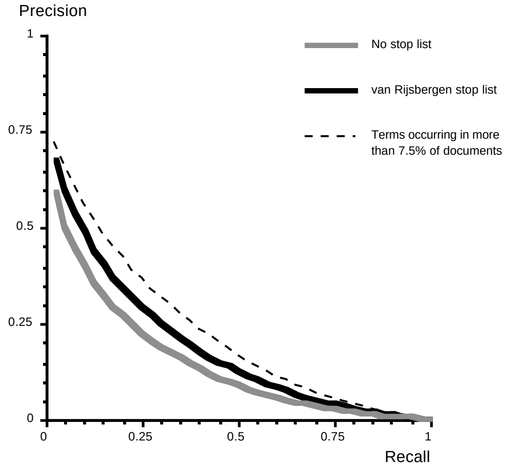

Figure 1. Comparison of retrieval effectiveness using different stop lists.

3 Results

Perhaps surprisingly for such a simplistic retrieval strategy, the glair4 submission appears to have been above average, when compared to the other thirteen category B submissions. Comparing average precision, glair4 was higher than the median average precision in 28 of the 45 topics, equal to the median in 12 topics, and worse in 5.

4 Conclusions and future work

One could say that the relatively good performance of glair4 is a little disheartening. In using a simple retrieval strategy, one might have expected it to perform worse than it did. Nevertheless, as good as this performance appears to be, a comaprison is not being made with the best performing systems in the category A ad hoc section. Therefore, next year it is intended that the simplistic strategy described here will be applied to this larger task.

The work to successfully implement the theoretical approach of generalised imaging on an IR system will continue. Last year, limitations of computational resources was a problem, this has been solved by the purchase of new equipment. It has been found, however, that the initial implementation of this approach requires attention and in the next year an altered implementation will be pursued.

5 References

Crestani 95

F. Crestani, M. Sanderson, I. Ruthven & C.J. van Rijsbergen (1995). The troubles with using a logical model of IR on a large collection of documents, proceedings of the TREC-4 conference.

Frakes 92

W.B. Frakes & R. Baeza-Yeates (1992). Information Retrieval: Data structures & algorithms, in Prentice Hall

Porter 80

M.F. Porter (1980). An algorithm for suffix stripping, in Program - automated library and information systems, 14(3): 130-137.

Van Rijsbergen 79

C.J. van Rijsbergen (1979). Information retrieval (second edition), in London: Butterworths.