Effect of Varying Number of Documents in Blind Feedback

Analysis of the 2003 NRRC RIA Workshop "bf_numdocs" Experiment Suite

Jesse Montgomery[1] Luo Si[2] Jamie Callan[2] David A. Evans[1]

[1]Clairvoyance Corporation
5001 Baum Blvd., Suite 700
Pittsburgh, PA 15213-1854, USA

[2]Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA

{j.montgomery, dae}@clairvoyancecorp.com

{lsi, callan}@cs.cmu.edu

Categories and Subject Descriptors

H.3.3 [**Information Systems]:** Information Search and Retrieval -- *Relevance feedback.*

General Terms

Experimentation, Data Analysis.

Keywords

Information Retrieval, Pseudo-relevance Feedback, Optimal Number of Documents for Feedback, Query Expansion.

1. INTRODUCTION

The "bf_numdocs" experiment was one of several sets of experiments conducted at the 2003 Reliable Information Access (RIA) Workshop, hosted by the Northeast Regional Research Center (NRRC) and MITRE, involving seven research groups. One of the goals for the RIA workshop was to focus on expansion techniques for relevance feedback and pseudo-relevance feedback. With this end in mind, the bf_numdocs suite of experiments was designed to examine the effect of varying the number of documents used to extract expansion terms for pseudo-relevance feedback. We describe the motivation for this set of experiments, the hypotheses tested, an overview of the experimental method, and an analysis of results and conclusions. Given the nature of the workshop, we were able to examine whether there were any system-dependent effects or any topic-dependent effects.

2. HYPOTHESES

Several hypotheses were examined in the bf_numdocs experiment.

1. All the systems would demonstrate a tradeoff in the choice about how many documents should be used for feedback.

2. The optimal number of documents used for feedback would vary from system to system.

3. Query length (as a feature) and the optimal number of documents used for feedback would be negatively correlated—that is, the shorter the query, the more documents would be necessary for optimal performance.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007

4. The number of relevant documents (for a given query) and the optimal number of expansion documents would be positively correlated—that is, the more relevant documents in the corpus, the more documents would be necessary for optimal performance.

5. Different topics would behave similarly (i.e. cluster) in relation to (a) the length of the query, (b) the total number of relevant documents in the database, and (c) the optimal number of documents for expansion.

3. EXPERIMENTAL METHODOLOGY

Nine systems (from seven research groups) participated in the experiment. (For a review of the systems involved in the Workshop, see [1].) Each system, independently, provided results for the same set of experimental configurations with regard to pseudo-relevance feedback. (All other parameters not pertaining to pseudo-relevance feedback were left to the discretion of individual systems.) The required feedback parameters were as follows: in every configuration, each system was allowed to select 20 terms for feedback; however, the number of documents used for expansion was varied from 1 to 20 (for every additional document), and from 25 to 100 (for every five additional documents). This resulted in a total of 36 configurations for every system. Each configuration was denoted as *bf.#docs.20*, where *#docs* stands for the number of documents available for selecting terms for feedback. Not all systems processed documents in the same way; however, the "document" was the unit of additional information for each system, and the amount of additional information for each system increased monotonically in the number of documents.

4. RESULTS AND DISCUSSION

As can be seen in Figure 1, each system had a single peak with regard to the optimal number of documents for feedback. However, the tail for each system behaved differently; some dropped quickly (City, Sabir), while some systems were not sensitive to the addition of more documents for expansion (CMU, UMass). This result serves as the confirmation of our first two hypotheses. The fact that various systems had tradeoffs in the choices of optimal number of documents for feedback shows that usually some amount of documents for feedback can be helpful but too many of them may cause negative influence (Hyp. 1). The fact that each system's performance peaked at a different point shows that the optimal number of feedback documents varied from system to system (Hyp. 2). Those systems whose performance

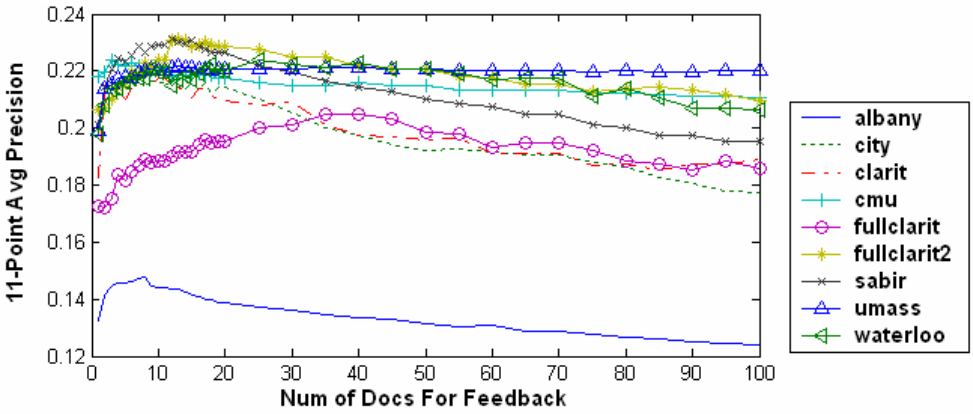

Figure 1. Performance across systems when varying the number of documents used for feedback

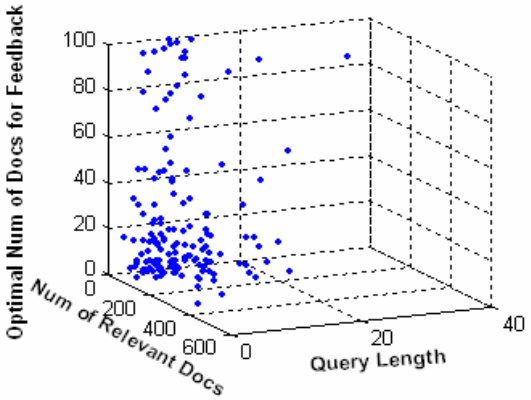

Figure 2. Query feedback behavior distribution for the CMU system, showing the relationship between number of relevant documents or query length with the optimal number of documents for feedback. (Mainly from the point of view of the number of relevant documents.)

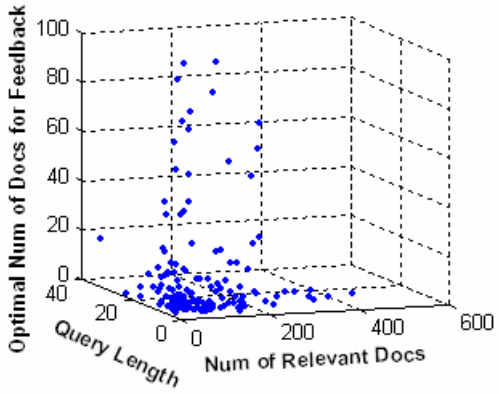

Figure 3. Query feedback behavior distribution for the Fullclarit2 system, showing the relationship between query length or number of relevant documents with the optimal number of documents for feedback. (Mainly from the point of view of query length.)

remains stable in a large range are not sensitive to the choice of the amount of documents for feedback. However, those systems that peak quickly and then drop off would do well to have a conservative feedback strategy, and to focus on optimizing the area proximate to the optimal number of documents.

The remaining hypotheses (Hyp. 3-5) were not supported. There were no correlations between either the query length (Figure 2) or the number of relevant documents (Figure 3) and the optimal number of expansion documents. Additionally, topics did not cluster according to the optimal number of documents for feedback with the above two features. These results can also be seen in Figures 2 and 3. These figures are for only two systems (CMU and Fullclarit2), but they are representative of all the systems, as all behaved similarly.

These results show that short queries or queries with small numbers of relevant documents may need the same number of documents for feedback as long queries or queries with large numbers of relevant documents. Furthermore, no combination of these topic characteristics—query length, total number of relevant documents in the database, and optimal number of documents for feedback—shows any discernable pattern.

5. CONCLUSIONS

The nine systems that participated in the experiment ranged from very traditional vector space models (e.g., the SMART system), to those based on language models (e.g., the CMU Lemur system) and more radical approaches (e.g., the Waterloo system that locates "hot spot" windows within documents). These very different systems chose different features to represent terms, had different weighting schemes, and had a system-dependent choice of the optimal number of documents for feedback. However, all the systems show tradeoffs in the choices of optimal numbers of documents for feedback. Although no explicit relationship has been found between either the query length or the number of relevant documents and the optimal number of documents for feedback, other features such as the score distribution of initial retrieval may be useful in predicting the optimal number of feedback documents. This is the subject of further work.

6. REFERENCES

[1] Harman, D., and Buckley, C. The NRRC Reliable Information Access (RIA) Workshop. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 2004 (SIGIR 2004)*, Sheffield, U.K. ACM Press, 2004.