

Assessing the Term Independence Assumption in Blind Relevance Feedback

Jimmy Lin G. Craig Murray
College of Information Studies
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742, USA
{jimmylin,gcraigm}@umd.edu

ABSTRACT

When applying blind relevance feedback for *ad hoc* document retrieval, is it possible to identify, *a priori*, the set of query terms that will most improve retrieval performance? Can this complex problem be reduced into the simpler one of making independent decisions about the performance effects of each query term? Our experiments suggest that, for the selection of terms for blind relevance feedback, the term independence assumption may be empirically justified.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Relevance feedback*

General Terms: Experimentation

Keywords: pseudorelevance feedback, automatic query expansion

1. INTRODUCTION

Blind relevance feedback (e.g., [3, 4, 2]) is a common information retrieval technique that has been shown, on average, to increase retrieval performance. Because this strategy hurts performance in some cases, the ability to predict whether or not a set of feedback terms is beneficial remains an open research question that has important implications (cf. recent NRRC RIA workshop). If we could differentiate “good” from “bad” terms, the overall gains from blind relevance feedback could be further boosted. Our study explores the term independence assumption in blind relevance feedback. Can we individually predict the performance effect of each feedback term, or are we hampered by interactions between multiple terms? The answer appears to be *yes* and *no*, respectively. The quality of each feedback term can be decided independently, suggesting that the selection of “good” terms for blind relevance feedback can be recast as a binary classification problem on each individual term.

2. TERM INDEPENDENCE

Our experiments involved 150 topics from TREC-6 through TREC-8, and were performed with Indri¹ [1]. Candidate feedback terms were gathered by using the description field

¹<http://www.lemurproject.org/>

Query	MAP
description	0.1591
description + all terms	0.1745 (+9.68%)
description + “good” terms	0.2113 (+32.8%)

Table 1: Performance of blind relevance feedback

of the topics as queries and collecting the twenty highest scoring terms from the top twenty Indri hits using a simple *tf.idf* measure; no stemming was employed in these experiments. The baseline results from adding all the candidate feedback terms are shown in Table 1.² For each topic, we then ran twenty separate queries, one with each of the candidate feedback terms and the original query. The difference between the mean average precision of original query and the one-off query was calculated; we called this difference the marginal MAP gain (MMG), since it quantified the effect of adding that additional term. If a term has positive MMG, we considered it “good”; otherwise, it was a “bad” term.

Since the goal of this study was to explore the term independence assumption, we first established an upper bound on performance. If an oracle told the system the MMG of each candidate feedback term, and it only added terms that were “good”, how well would our system perform? The results of this experiment are shown in Table 1. As we can see, if a system were able to correctly choose “good” terms, the improvement in MAP would be more than three times that of simply adding all feedback terms.

Although these numbers are encouraging, they tell us nothing about the potential interactions between multiple query terms. For the simplest case of two term interactions, this can be quantified by comparing the MMGs of all possible two-off queries against the sum of individual term MMGs. This scatter plot is shown in Figure 1. The distribution can be well-fit by a straight line with a slope of 0.78, which means that for pairs of terms, marginal MAP gains are additive with a constant discount factor. Of the 28,500 possible queries, there were only 315 cases where the MMG sums incorrectly predicted increased performance (those points in the fourth quadrant).

To truly assess the validity of the term independence assumption, we needed to compare the performance achieved

²By simple parameter tuning, we established 0.4 as the optimal weight for feedback terms.

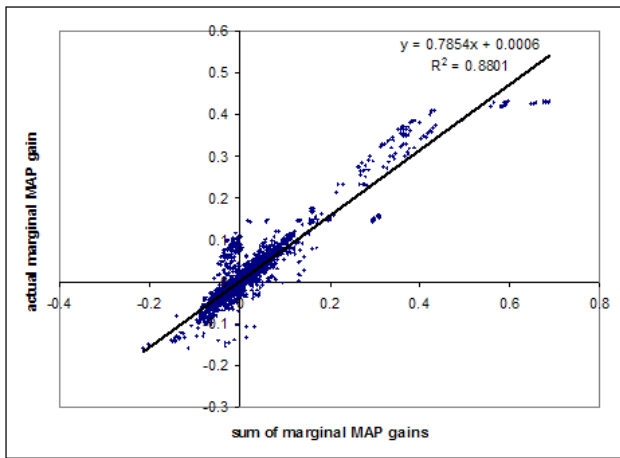


Figure 1: Scatter plot shown two-term interactions.

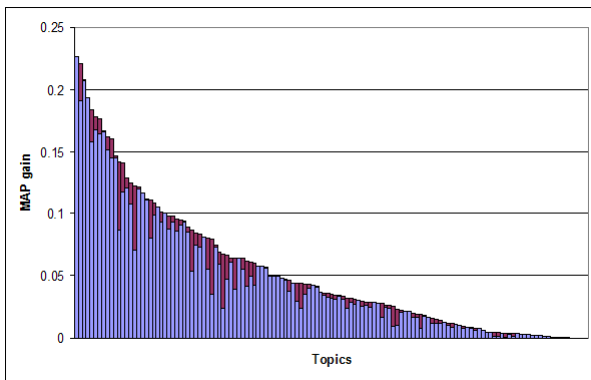


Figure 2: The validity of the term independence assumption.

by the optimal set of query terms and the performance obtained by independently making decisions about each query term. Naturally, the optimal set of feedback terms can only be determined by considering the power set of all candidates, which, for twenty terms, is unrealistic. However, since the previous experiment suggested that adding two “good” terms is unlikely to decrease performance, we might approximate the optimal set by exhaustively searching the power set of only “good” terms. Due to the exponential nature of power sets, even this was impractical. Thus, we considered topics that had 16 or fewer “good” terms, which yielded results for 138 topics. In total, 826,759 queries were executed by Indri in this experiment. Overall, we discovered that adding all the “good” terms yielded a performance improvement in MAP that was, on average, 87% that of the optimal power set approximation. For 43 of the topics, adding all the “good” terms *was* the best combination. Figure 2 shows these results graphically. Each bar represents a topic, with the shorter bar indicating MAP gain under the term independence assumption, and the longer bar the power set results. These numbers appear to suggest that term–term interactions are not a significant cause for concern in selecting terms for blind relevance feedback.

To further study the term independence assumption in the context of blind relevance feedback, we explored the effect of

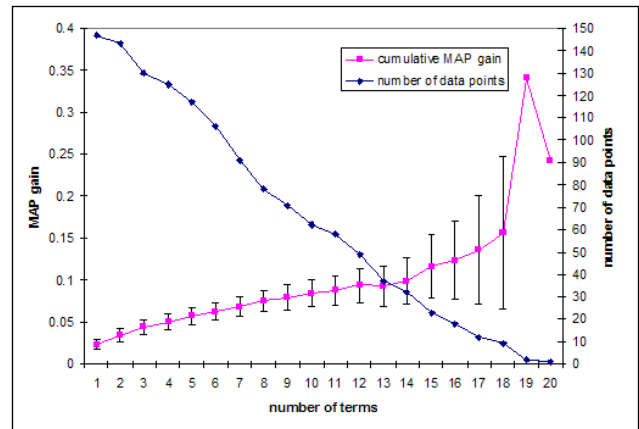


Figure 3: Cumulative MAP gain by number of “good” terms added.

adding different numbers of “good” terms. Do most of the improvements in performance come from only a few terms? Is there a point of diminishing returns in adding more terms? To answer these questions, we first sorted the “good” feedback terms for each topic by their MMGs, in descending order. We iteratively added each individual term and plotted the change in mean average precision—essentially, a cumulative distribution. These results are shown in Figure 3. Obviously, as the number of feedback terms increases, the number of data points decreases; for example, only 62 out of 150 topics have 10 or more “good” terms. The error bars denote the 95% confidence intervals³ and reflect the decreasing sample size as the number of terms increases. It is interesting to note that the MMG for adding all 20 feedback terms (0.0154) is worse than adding only the best term (MMG=0.0236, 95% confidence interval= ± 0.0057).

3. CONCLUSION

Our study suggests that the problem of choosing the optimal set of feedback terms can be approximated by individual binary decisions about each term independently—that the term independence assumption in choosing feedback terms may be empirically justified. This reformulation of the problem immediately suggests a classification approach based on machine learning techniques.

4. REFERENCES

- [1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, D. Metzler, M. D. Smucker, T. Strohan, H. Turtle, and C. Wade. UMass at TREC 2004: Notebook. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, 2004.
- [2] C. Carpineto, R. de Mori, G. Romano, and B. Bigi. An information-theoretic approach to automatic query expansion. *ACM Transactions on Information Systems*, 19(1):1–27, 2001.
- [3] S. Gauch, J. Wang, and S. M. Rachakonda. A corpus analysis approach to automatic query expansion and its extension to multiple databases. *ACM Transactions on Information Systems*, 17(3):250–269, 1999.
- [4] J. Xu and W. B. Croft. Improving the effectiveness of information retrieval with local context analysis. *ACM Transactions on Information Systems*, 18(1):79–112, 2000.

³Omitted for 19 and 20 terms because the data points are too few to give meaningful intervals.