A Hybrid Model for Ad-hoc Information Retrieval

Zheng Ye, Jimmy Xiangji Huang
Information Retrieval and Knowledge Management Research Lab
School of Information Technology
York University
Toronto, Canada
{yezheng, jhuang}@yorku.ca

Jun Miao
Information Retrieval and Knowledge Management Research Lab
Department of Computer Science & Engineering
York University
Toronto, Canada
jun@cse.yorku.ca

ABSTRACT

Many information retrieval (IR) techniques have been proposed to improve the performance, and some combinations of these techniques has been demonstrated to be effective. However, how to effectively combine them is largely unexplored. It is possible that a method reduces the positive influence of the other one even if both of them are effective separately. In this paper, we propose a new hybrid model which can simply and flexibly combine components of three different IR techniques under a uniform framework. Extensive experiments on the TREC standard collections indicate that our proposed model can outperform the best TREC systems consistently in the ad-hoc retrieval. It shows that the combination strategy in our proposed model is very effective. Meanwhile, this method is also re-useable for other researchers to test whether their new methods are additive to the current technologies.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval models, Relevance feedback

General Terms

Algorithms, Performance, Experimentation

Keywords

Hybrid Model, Rocchio's Relevance Feedback

1. INTRODUCTION

In the past thirty years, researchers make great progress in the Information Retrieval (IR) area. A plenty of new technologies, e.g., stemming, query expansion and smoothing methods, have been introduced and help to obtain better performance in retrieving relevant documents. Some attempts to combine these technologies show that the strategy of combination is very important because one technology can counteract the affects of others. However, how to make an effective combination is still largely unexplored, especially under a unified framework.

In this paper, we propose a hybrid model to incorporate three different retrieval techniques that have proven to be effective for the ad-hoc retrieval on the TREC collections. We analyze the best TREC systems for ad-hoc retrieval, and extend the Rocchio's feedback method by incorporating three kinds of IR techniques, which are proximity, feedback document quality estimation and query performance prediction techniques, under the pseudo relevance feedback (PRF) framework to boost the overall performance. We mainly focus on how to refine the representation of the query under the PRF framework in order to avoid the drawbacks of traditional PRF methods. Experimental results on various TREC datasets show that our hybrid model consistently obtains better results over the best TREC systems. Because our proposed model is component-based, it is very flexible to import different techniques in the future. Meanwhile, the hybrid model can help researchers to test whether their methods are additive to improve the overall performance of ad-hoc retrieval which was mentioned in [1].

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

2. A HYBRID RETRIEVAL MODEL

Rocchio's algorithm [4] is a classic framework for implementing (pseudo) relevance feedback via improving the query representation. When the negative feedback documents are ignored, the traditional Rocchio's model is as follows:

$$Q_1 = \alpha * Q_0 + \beta * \sum_{r \in R} \frac{r}{|R|} \quad (1)$$

where Q_0 and Q_1 represent the original and first iteration query vectors, R is the set of (pseudo) relevance documents, and r is the expansion term weight vector.

Although the Rocchio's model has been introduced for many years, it is still effective in obtaining relevant documents. According to [7], "BM25 term weighting coupled with Rocchio feedback remains a strong baseline which is at least as competitive as any language modeling approach for many tasks". Meanwhile, it is very flexible to adapt additional components. However, the traditional Rocchio's model can still be reformed to be better. First, the query term proximity information which has proven to be useful is not considered. Second, Rocchio's algorithm views terms from different feedback documents equally. Intuitively, a candidate expansion term in a document with better quality is more likely to be relevant to the query topic. Third, the interpolation parameter α is always fixed across a group of queries. In fact, for a well expressed query, the candidate feedback documents are always more reliable for relevance feedback. In this paper, we use a regression model to predict this interpolation parameter. In order to alleviate influence of these problems, we extend Rocchio's algorithm which re-

Table 1: Direct comparison with the best MAP results in each TREC year. In the Hybrid (Fixed) method, proximity and feedback document quality are utilized. In the Hybrid (Regression) method, all the three techniques are adapted. A "*" indicates a statistically significant improvement when a component technique is added in our algorithm.

Method	TREC1	TREC2	TREC3	TREC6	TREC7	TREC8	TREC2004	TREC2005	TREC2006
BM25	0.2292	0.2058	0.2787	0.2397	0.1819	0.2471	0.2672	0.3403	0.2965
BM25+Prox	0.2461*	0.2111*	0.2929*	0.2507*	0.1936*	0.2582*	0.3148*	0.3661*	0.3459*
Hybrid-Fixed	0.2938*	0.2913*	0.3811*	0.2763*	0.2576*	0.2909*	0.3375*	0.4083*	0.3944*
Hybrid-Reg	0.3012	0.2971	0.3912*	0.2886*	0.2611	0.3103*	0.3431*	0.4193*	0.3921
BEST TREC	0.2062	0.2475	0.3231	0.2876	0.2614	0.3063	0.3052	0.4056	0.3737

fines the query representation as follows.

$$Q_1 = \alpha * (\beta * Q_0 + (1-\beta) * Q_p) + (1-\alpha) * \sum_{r \in R} \frac{r * q(d_r)}{|R|} \quad (2)$$

where β controls how much we rely on the query term proximity information [5], α controls how much we rely on the original query, Q_p is an n-gram of original query terms and $q(d_r)$ is the quality score of document d.

As we can see from Equation 2, our proposed algorithm is very flexible and can evaluate different techniques. In this paper, we adopt the co-occurrence interpretation of term proximity to compute Q_p, where the proximity among query terms is represented by the n-gram frequencies and BM25 is used as the weighting model [2].

Full dependencies of query terms are taken into account. For the document quality factor $q(d_r)$, we simple use the scores from the first-pass retrieval for approximation as in [6]. For the prediction to α, we use the same features as in [3] to train the regression model. The difference is that we do it within Rocchio's framework.

3. EXPERIMENTS AND ANALYSIS

We conduct experiments on three representative test collections: disk1&2, disk4&5, and GOV2, which are used in different TREC years. We present the results for each TREC year such that we can directly compare our results with the best TREC systems. Detailed information about the TREC datasets and the evaluation criteria, please refer to http://trec.nist.gov. For the preprocessing of the collections, we use the Porter Stemmer and a stopword list. In addition, we only use the *title* part of the topics to retrieve.

In our experiments, we first empirically evaluate different combinations of our implemented component techniques, then evaluate how these techniques perform when they are integrated in our hybrid model. We set the parameters to fixed values in a parsimonious way such that each component technique gets considerable improvements on most collections. In other words, there is still room for improvement if the parameters are tuned on a collection-by-collection basis. Particularly, for the basic retrieval model, we use the Okapi BM25 model, and set b in BM25 to 0.3. When only the query term proximity technique is added, denoted as "BM25+Prox", we set β to 0.3. In addition, when query expansion and document quality estimation techniques are added, denoted as "Hybrid-Fixed", we empirically set α, $|R|$ to 0.5 and 30. However, when the query performance prediction technique is used, denoted as "Hybrid-Reg", α is not fixed. But it is obtained from a regression model that is trained as in [3]. When evaluating our hybrid model for a particular TREC year, the queries in the remainder TREC years on the same collection are used as training data.

From Table 1, we can see that our hybrid model with different component techniques can significantly outperforms the basic retrieval model, which reconfirms the effectiveness of these techniques. In addition, when all these component techniques are used in our hybrid model, the retrieval performance can be further improved. It indicates that performance gains from these two component techniques can be added up in our proposed hybrid model. However, when we use a regression component to predict α, the performance gain is not very obvious compared with other components. We conjecture the main reason is as follows: when the feedback document set is more reliable for relevance feedback, the regression component is less useful.

When compared with the best TREC systems, our proposed model obviously outperforms the best TREC systems on most collections . It is of note that the results in our paper are obtained in a uniform setting across all collections while the best TREC results were from different participants independently. We believe the significant improvement is mainly from our successful integration of different IR techniques in the proposed model. In addition, according to Armstrong et al.'s survey, very few published results are better than the best TREC systems, mostly below medium systems. Our proposed model is promising, which provides a good avenue for future IR research, especially for evaluating the overall performance of a system (not a particular component of a system).

4. CONCLUSIONS

In this paper, we propose a hybrid model which can successfully integrate three effective techniques in a uniform model. Extensive experiments show that our approach obviously outperforms the best TREC systems in most cases. In the future, we will investigate more effective techniques in IR and incorporate them into our framework to conduct this research in details.

5. ACKNOWLEDGMENTS

This research is supported by the research grant from the Natural Sciences & Engineering Research Council (NSERC) of Canada and the Early Researcher Award/ Premier's Research Excellence Award, Zhejiang Provincial Natural Science Foundation, Q12F020016.

6. REFERENCES

[1] T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. Improvements that don't add up: ad-hoc retrieval results since 1998. In *CIKM*, pages 601–610, 2009.

[2] B. He, J. X. Huang, and X. Zhou. Modeling term proximity for probabilistic information retrieval models. *Inf. Sci.*, 181(14):3017–3031, 2011.

[3] Y. Lv and C. Zhai. Adaptive relevance feedback in information retrieval. In *CIKM*, pages 255–264, 2009.

[4] J. Rocchio. *Relevance feedback in information retrieval*, pages 313–323. Prentice-Hall Englewood Cliffs, 1971.

[5] C. J. van Rijsbergen. A theoretical basis for the use of co-occurence data in information retrieval. *Journal of Documentation*, 1977.

[6] Z. Ye, B. He, X. Huang, and H. Lin. Revisiting rocchio's relevance feedback algorithm for probabilistic models. In *AAIRS*, pages 151–161, 2010.

[7] C. Zhai. Statistical language models for information retrieval a critical review. *Found. Trends Inf. Retr.*, 2:137–213, 2008.