Topic (Query) Selection for IR Evaluation

Jianhan Zhu, Jun Wang, Ingemar Cox
University College London, UK
jianhan.zhu@ucl.ac.uk,
jun_wang@acm.org, ingemar@ieee.org

Vishwa Vinay
Microsoft Research
Cambridge, U.K.
vvinay@microsoft.com

ABSTRACT

The need for evaluating large amounts of topics (queries) makes IR evaluation an uneasy task. In this paper, we study a topic selection problem for IR evaluation. The selection criterion is based on the overall *difficulty* of the chosen set, as well as the *uncertainty* of the final IR metric applied to the systems. Our preliminary experiments demonstrate that our approach helps to identify a set of topics that provides confident estimates of systems' performance while keeping the requirement of the query difficulty.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.3Information Search and Retrieval

General Terms

Experimentation, Measurement, Performance

1 Introduction

Given the tremendous amount of data on the Web and the huge volumes of queries that users issue to search engines, it is a big challenge to evaluate how well a search system can perform well to satisfy users' information needs. In order to tackle the challenge, we propose to select a set of queries from a large query pool so that a system's performance on the selected set can serve as an accurate indicator of the system's performance on the whole pool of queries.

The Cranfield evaluation methodology, which has been adopted in evaluation initiatives such as the Text REtrieval Conference (TREC), uses standard test collections consisting of documents, topics, and relevance judgments. An IR system produces a ranked list of documents for each topic, a metric such as average precision (AP) is calculated on the ranked results and indicates how well this system did on this topic. The mean of such a metric over a number of topics (e.g. mean average precision or MAP), is generally used to measure the system's overall performance.

The robustness and reliability of this methodology has been well studied (e.g. [1, 6, 5]). The sources of error include limited pool depths, incomplete relevance information, topic set size, averaging across topics with different numbers of relevant documents etc. Given such shortcomings, when we use a set of topics and relevance judgements to evaluate a system, there will be some uncertainty in our estimate. Motivated by the Modern Portfolio Theory of finance [2], we attempt to account for, quantify and then minimize the uncertainty involved in running such an evaluation.

In this context, we consider topic difficulty, which potentially convolutes many factors: (i) individual systems find particular topics difficult (ii) topics are likely to be difficult for all systems if they do not have corresponding content in the collection (iii) ambiguous or wide topics can be difficult even if the collection contains relevant content. An estimate of a topic's intrinsic difficulty (i.e., (ii) and (iii) above) can be obtained by looking across systems' performance on that topic.

We first propose our approach in Section 2, present experimental results in Section 3, and conclude in Section 4.

2 Topic Selection

Given a set consisting of N topics, the aim is to select a set of n topics $\{t_k\}, k \in [1, n]$ for system testing. We have the AP (average precision) values of M systems on the N topics, which forms an $M * N$ matrix, where $AP_{i,j}$ denotes the AP of the i-th system on the j-th topic. The mean value of the i-th row is the MAP for the i-th system, MAP_i, and the mean of the j-th column is the expected AP or AAP_j (Average Average Precision) [3] for the j-th topic.

Therefore, for the selected n topics, the expected MAP ($EMAP_n$), i.e., the mean of the expected AP values for the n topics, and the uncertainty associated with the expected MAP (for detailed derivation of Eq. 1 and 2, we refer to [2]) are:

$$EMAP_n = \frac{1}{n}\sum_{k=1}^{n}\hat{AP}_{t_k} \tag{1}$$

$$Unc_n = \frac{1}{n^2}\sum_{k=1}^{n}\sigma_{t_k}^2 + \frac{2}{n^2}\sum_{k=1}^{n}\sum_{l=k+1}^{n}\sigma_{t_k}\sigma_{t_l}\rho_{t_k,t_l} \tag{2}$$

In Equation 2, $\sigma_{t_k}^2$ and $\sigma_{t_l}^2$ are variances/uncertainties due to individual topics t_k and t_l. This quantity can be used to capture factors like uncertainty in the number of relevant documents for the topic, incomplete relevance judgments, etc. The quantity ρ_{t_k,t_l} is the correlation between topic t_k and t_l. It can be seen that if we want to pick the topics with the lowest Unc_n, then one way to do that would be to pick topics with the lowest $\sigma_{t_k}^2$. That is to say, as part of the evaluation set, we pick the topics about which we know most. In this paper, we fix all $\sigma_{t_k}^2 = 0.01$ (or the standard deviation $\sigma_{t_k} = 0.1$), leaving more accurate estimations for future work. Assuming that some level of uncertainty regarding topics is inevitable, attempting to compile a set of n topics with minimum Unc_n boils down to picking topics with low (preferably negative) correlations ρ_{t_k,t_l}.

We estimate ρ_{t_k,t_l} as the Pearson correlation coefficient [4] between the t_k-th and t_l-th columns in the AP matrix. ρ_{t_k,t_l} (which lies between -1 and 1) measures the relationship between two topics. The higher the value of ρ_{t_k,t_l}, the more closely related the two topics are. Note that a high value for the correlation between two topics does not necessarily mean that the pair are related in terms of content, it only means that systems which do well (or badly) on one topic tend to do well (correspondingly, badly) on the other.

To reduce the uncertainty in $EMAP_n$, i.e., Unc_n in Equation 2, we need to have small correlation coefficients (preferably negative) between topics. To understand why this is so, consider having AP values for all systems for a given topic t_j. If we add another topic t_k which is highly positively correlated with t_j, knowing the AP values for the systems on t_k is unlikely to provide any new information regarding the relative goodness of the systems. From an evaluation perspective, it is sufficient to have one of t_j or t_k in the topic set if there is a strong positive correlation between them. If we aim to find the minimal set of topics that provide informative and certain estimates of system effectiveness, diversifying the set of topics will lead to each individual being more informative and the uncertainty associated with $EMAP_n$ will be reduced.

The main goal of topic selection now is to minimize the uncertainty Unc_n when we set a preference for the overall difficulty of selected queries as indicated by $EMAP_n$. We propose a greedy approach for topic selection, which may not produce the globally optimal solution but allows for the trade-off between computational cost and correctness. We start with all the N topics, and remove one topic at each step, until n topics remain. In each step, we remove the topic that can result in the largest reduction in the overall uncertainty Unc_n and at the same time the $EMAP_n$ of the remaining topics matches the criterion of our topic difficulty preference.

3 Experimental Results

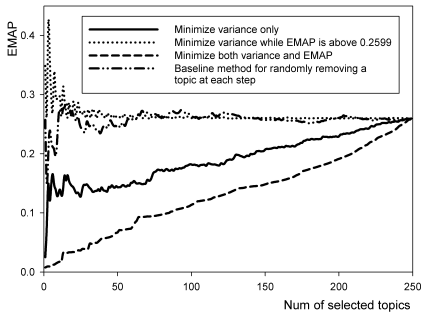
We applied our approach to 249 topics in the TREC 2004 Robust Track collection consisting of topics of different levels of difficulty. There are 110 runs submitted by 14 groups. The range of the MAP is $(0.0756 - 0.3586)$, and range of the expected AP, $\hat{AP}$, is $(0.0077 - 0.7717)$. The EMAP and overall uncertainty of the 249 topics are 0.2599 and 0.0026, respectively. Two strongly correlated topics are topic 392 "Robotics" and 431 "Robotic Technology" with Pearson correlation coefficient as 0.8320.

As indicated before, when going from a topic set of size n to size $n-1$, we have to define a criterion that indicates our preference for the topic to be dropped. We experimented with three policies for removing topics: (1) minimization of uncertainty only, indicating no preference on topic difficulty, (2) minimization of uncertainty while maintaining EMAP above 0.2599 (=the total average), indicating preference for retaining easy topics, and (3) minimization of uncertainty while EMAP decreases monotonically when removing topics, indicating preference for retaining difficult topics.

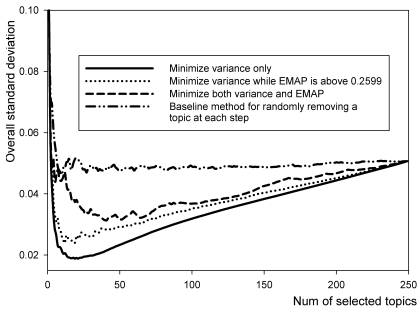
For each set of n topics, we calculate $EMAP_n$ as given in Equation 1. The uncertainty or variance of this value is given by Unc_n, as in Equation 2, we monitor the square-root of Unc_n (the standard deviation) as a function of the size of the topic set. Fig. 1 (a) and (b) plot the EMAP and standard deviation versus n, for $n \in [1, 249]$. For comparison, we also show a baseline policy that randomly drops topics.

All the four curves share the right-most point, when the topic set consists of all the 249 topics. The experiment proceeds by tracing one of the curves (a specified policy for dropping topics) while moving towards the left. The first observation from Fig. 1 (a) is that our three approaches select topics according to the topic difficulty preferences. For the first preference, the EMAP decreases until around 50 topics so that the standard deviation is optimally minimized. For the second preference, EMAP is always above 0.2599 to keep the topics easy overall. For the third preference, EMAP decreases the most quickly in order to find more difficult topics.

From Fig. 1 (b) we see that all our three preference based approaches outperform the baseline in terms of reducing the uncertainty. The overall standard deviation decreases for all three preferences until the number of selected topics is around 50. The standard deviation decreases the most quickly when there are no restrictions on the EMAP. When the number of topics decreases under a small value, the standard deviation starts increasing for all three preferences as well as the baseline, since topics are becoming more independent of each other and there are few negatively correlated topics. This also helps explain that the number of topics for IR evaluation should be above a certain number, such as 50. It is interesting to note that amongst our three policies, preferring difficult topics leads to maximum standard deviation (uncertainty), since systems all perform relatively badly on such topics so that they do not provide us with reliable indicators of system effectiveness.



(a)



(b)

Figure 1: Comparison of three topic selection policies against a random topic selection baseline. We plot the number of topics vs. (a) EMAP and (b) Standard Deviation. All methods start at the same point on the right end and move leftwards while eliminating different topics according to the specified policy.

4 Conclusion

This paper considered the standard IR evaluation task of calculating the relative effectiveness of a group of systems based on a set of topics. The outcome of such a measurement has some inherent uncertainty due to properties of individual queries/topics in the set, as well as interactions between them. Motivated by the Modern Portfolio Theory, we propose a novel mean-variance based topic selection approach for minimizing the uncertainty in the measured effectiveness. Given a restriction on the size of the subset, our approach identifies those topics that reduce the uncertainty of systems' effectiveness, under the condition that the overall difficulty of the set matches predefined preferences.

5 References

[1] C. Buckley and E. M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of SIGIR 2000*.

[2] H. M. Markowitz. *Portfolio Selection: Efficient Diversification of Investments*. John Wiley & Sons, Inc., New York, and Chapman & Hall, Limited, London, 1959.

[3] S. Mizzaro and S. Robertson. Hits hits trec: exploring ir evaluation results with network analysis. In *SIGIR*, pages 479–486, 2007.

[4] J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42:59–66, 1988.

[5] M. Sanderson and J. Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of SIGIR 2005*.

[6] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of SIGIR 1998*.