On Evaluation of Adaptive Topic Tracking Systems

Tamer Elsayed
Department of Computer Science and UMIACS
University of Maryland, College Park, MD 20742
telsayed@cs.umd.edu

Douglas W. Oard
College of Information Studies and UMIACS
University of Maryland, College Park, MD 20742
oard@glue.umd.edu

ABSTRACT

Summative evaluation methods for supervised adaptive topic tracking systems convolve the effect of system decisions on present utility with the effect on future utility. This paper describes a new formative evaluation approach that focuses on future utility for use in the design stage of adaptive systems. Topic model quality is assessed at a predefined set of points using a fixed document set to enhance comparability. Experiments using a vector-space topic tracking system illustrate the utility of this approach to formative evaluation.

Categories and Subject Descriptors: H.3.4 [Information Storage and Retrieval]: Systems and Software – Performance Evaluation.

General Terms: Design, Measurement, Performance.

Keywords: Adaptive filtering, topic tracking, DET curve, TDT, formative evaluation.

1. INTRODUCTION

The supervised adaptive topic tracking task, recently introduced in Topic Detection and Tracking (TDT) evaluations, is similar to the TREC adaptive filtering task. Both resemble the traditional "batch" filtering task, adding a simulation of user interaction by the feedback provided when the system delivers a putatatively relevant document. Systems make an immediate decision whether to display each incoming document, usually by assigning a confidence score then applying a threshold. Displaying a document has two main effects: the immediate effect on user satisfaction ("present utility"), and the effect on the quality of the topic model that the system will rely as a basis for future decisions ("expected future utility," which may increase if the model is adapted based on feedback from the user) [2]. This view suggests that an evaluation approach that separates the two effects could provide greater insight into the consequences of system design decisions. TDT and TREC adopted somewhat different evaluation strategies, but both convolve the two effects.

In TDT, a Detection Error Tradeoff (DET) curve is used to plot the probability of missing an on-topic story against the probability of presenting an off-topic story. The points on the curve are obtained by sweeping a threshold value across all topics simultaneously. A cost function is computed for each point, and the system is then evaluated by both the actual cost (for whatever threshold was actually chosen) and by the minimum cost over the whole curve. This implicitly assumes that scores for different topics are comparable. When all scores are normalized consistently and the topic model does not vary over time, this is a reasonable assumption. In adaptive tracking, however, the topic model changes as additional relevance judgments become available. Reducing the threshold can improve future utility at the expense of present utility, but the DET curve focuses solely on present utility; failing to reflect the effect on future utility. Changing the threshold and re-running the system may therefore not actually yield the miss and false alarm rates that were depicted on the DET curve.

An alternative approach was adopted in TREC, where a utility measure (conceptually, the inverse of cost) was plotted against time. That curve has the virtue of illustrating how the net effect accumulates over time, but utility measures at different points in time convolve the effects of present utility, future utility, and different document sets. In a sense, this is the converse of the DET curve's limitations; DET curves are insensitive to changes in future utility, while the TREC plots mask the future utility effect in a larger range of factors. In this paper, we propose a new approach that offers useful insights into the effect of model design on future utility to support formative evaluation of adaptive topic tracking systems.

2. EVALUATION DESIGN

The main goal of our proposed approach is to systematically characterize the effect of topic model adaptation on future utility in a way that is directly comparable over time. In order to trace model quality, we sample a set of topic models that the system incrementally builds during the adaptation process. Since system behavior might change any time feedback becomes available, we would ideally like to sample a topic model immediately after each feedback instance. However, we must balance the allocation of relevant documents between model adaptation and evaluation; once a document has been used for model adaptation, it would make little sense to measure how well the new model does on that same document. We therefore make only the first r relevant documents available for model adaptation (in our experiments, r=10). At each of these r potential adaptation points, we freeze the topic model and perform "batch"

SIGIR'05, August 15–19, 2005, Salvador, Brazil.
ACM 1-59593-034-5/05/0008.

(i.e., non-adaptive) topic tracking over the entire collection (including previously seen documents, but excluding all r relevant documents that were reserved for potential use in model adaptation). The process is repeated for each topic for which at least $r+n$ relevant documents are known (in our experiments, n=10).

We use the mean across the topics of the uninterpolated average precision (MAP) as a measure of expected future utility at each point in time. The MAP measure has the desirable characteristic that it is insensitive to any deficiencies in cross-topic score normalization. By plotting the MAP value for each of the r potential update points, we obtain a curve that shows how expected future utility evolves over time in a manner that is insensitive to coincident effects on present utility. The topic model could, of course, also be adapted any time a non-relevant document is selected by the system; the effect of such adaptations on expected future utility are accumulated until the next relevant document. This decision limits the resolution of the depiction somewhat, but at the cost of significant savings in computational complexity (for 15 topics with r=10, 180 system runs are required).

To summarize, an adaptive system S_{ad} is evaluated for tracking a topic T that is represented by a set of training examples t in the following 3 steps:

1. Run S_{ad} once, given the initial topic model M_0^T built using t, to generate additional r topic models M_i^T.
2. r+1 different sets of scores $O_{nad}(M_i^T)$ are obtained by running S_{nad} (a non-adaptive system) with each topic model. For each set, MAP is computed to measure the expected future utility of that topic model.
3. Plot MAP for each point.

3. DEMONSTRATING THE TECHNIQUE

We implemented a variant of the TDT-2004 "UMASS-2" adaptive vector space topic tracking system using a fixed threshold [1]. We evaluated this system using topics from the TDT-5 collection for which there are at least 20 known relevant documents in the evaluation epoch. We excluded a few topics for which assessment was terminated due to time constraints before adequate exhaustiveness was achieved (as determined by the Linguistic Data Consortium). That results in 15 English topics with average of 51 relevant documents (out of 254,000 documents) in the evaluation epoch (including the 10 reserved update points). We illustrate the effect of three static thresholds on expected future utility (0.075, 0.15, and 0.25). Figure 1(a) shows the results; for contrast, Figure 1(b) shows a similar plot for normalized detection cost (which conflates present and expected future utility in a manner similar to the TREC utility plot, but using a traditional TDT measure).

The curves indicate that the expected future utility of each system gradually improves as additional feedback becomes available, reflecting improvements in model quality over time from a relatively good initial model. We also notice that the system can gain a good future utility at the expense of the present utility. The presence of a sharp rise in the TDT detection cost for the lowest threshold value can clearly be seen to result from present cost rather than model quality, since that rise is not reflected in MAP. Indeed, the stable MAP at that point suggests that (when averaged over topics) the system is being overly aggressive in selecting documents early on (when the model is weakest). This suggests that starting with a relatively strict threshold and relaxing that threshold somewhat as relevant documents are discovered might be a productive strategy when using models of this design. The curves also show that the system achieves its best performance at the seventh positive feedback, after which it stabilizes, suggesting that 7 relevant documents could be good enough to initialize such system.

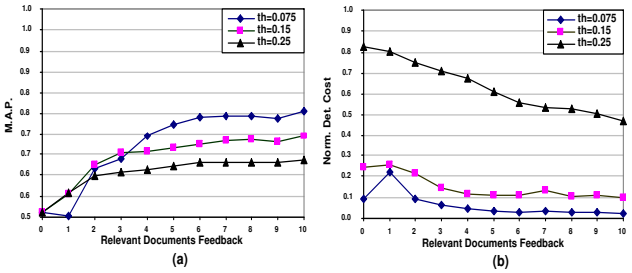

Figure 1: Adaptation effectiveness: (a) Model quality [high=good] (b) TDT detection cost [low=good]

4. CONCLUSION AND FUTURE WORK

In this paper, we have proposed a new approach to formative evaluation of adaptive topic tracking systems that is based on decomposing one adaptive system run into a few (7–10) non-adaptive runs that can be compared over time using a simple measure of topic model quality. The approach does, however, have two limitations. The first is that only topics with a substantial number of known relevant documents in the evaluation epoch can be used; that in turn limits the number of suitable topics. That limitation may well be acceptable for formative evaluation in which the goal is system tuning, not definitive comparisons. The second limitation is that the number of required system runs is multiplied by r+2; for slow systems, that may limit the number of variants that it would be practical to compare.

Some variants of our approach are also possible. Time rather than relevant documents could be plotted on the x-axis if elapsed time is particularly important in the application scenario. When two-sided models that learn from non-relevant documents are being compared, non-relevant documents selected for display by any system may also need to be excluded from the stable evaluation sets. As with any approach to formative evaluation, the design of the evaluation must naturally track closely with the insights that the developers seek to obtain.

5. ACKNOWLEDGMENTS

We wish to thank Gary Kuhn for asking some of the questions that inspired this work and for his valuable suggestions throughout. This research has been supported in part by DoD cooperative agreement N660010028910.

6. REFERENCES

[1] M. Connel, A. Feng, G. Kumaran, H. Raghavan, C. Shah, and J. Allan. UMass at TDT 2004. *In Working Notes of the TDT-2004 Evaluation*, 2004.

[2] Y. Zhang, W. Xu, and J. Callan. Exploration and exploitation in adaptive filtering based on bayesian active learning. *ICML*, 2003.