

# CMU DIR Supervised Tracking Report

Yi Zhang, Jamie Callan  
School of Computer Science  
Carnegie Mellon University  
Pittsburgh, PA 15232, USA  
yiz,callan@cs.cmu.edu

## 1 Introduction

This year was the first time the CMUDIR group participated in TDT. However, our research group had extensive prior experience with the adaptive information filtering and the TREC Filtering track [3, 7, 8, 4, 6, 1, 2, 5]. Our goal for our first participation in TDT was to study the effectiveness of an existing adaptive information filtering system for a TDT tracking task.

We submitted 2 runs for the Supervised Tracking task: CMU7 and CMU8. We first ran the TREC-style adaptive filtering system described in [5] and submitted the result as CMU7. Later we discovered an error caused by the difference between TDT input data and TREC data: the TDT Supervised Tracking task does not provide a topic initial description, but our system expected to start with an initial query or topic description provided by the user. We fixed this error and resubmitted the result (CMU8).

The algorithm described in [5] was used for profile learning. This algorithm combines Rocchio and Logistic Regression classifiers. This algorithm uses a simple classifier (Rocchio) to learn user interests when the amount of training data is small, and later switches to a more complex learning algorithm (Logistic Regression) to update its beliefs about user interests as more training data are available. The new algorithm automatically controls its model complexity based on the amount of training data, resulting in a system that can work robustly even with very few or no training data. In other evaluations it has been significantly better than other state-of-the-art information filtering algorithms [5].

The profile-learning algorithm uses both positive and negative training data to learn profiles for each topic. Initially, no documents are labelled as non-relevant. The system randomly sampled 30 documents for each topic at the early stage of filtering and treated each of them as  $\zeta$  off-topic documents to train the Rocchio and logistic regression classifiers, where the  $\zeta$  is the weight of each pseudo-negative document. We set  $\zeta$  to 0.5 in our experiment.

Once a document arrives, the system uses the classifier learned using the LR\_Rocchio algorithm described in [5] to estimate its probability of relevancy (“on topic”) for each topic. A document is delivered to the user if the probability of *on topic* of the document is above 0.09. This threshold is set to optimize the official utility measure

$$U = W_{rel} * R - NR$$

where:

R = number of relevant documents retrieved;

NR = number of nonrelevant documents retrieved; and

WRel = 10 is a constant that determines the relative weighting of relevant vs. non-relevant documents in determining the utility score.

The system updates the classifier each time a document is delivered and relevance feedback is available.

## 2 Experimental Results

The utilities of all submitted supervised adaptation topic tracking results released by NIST are in Table 1. Our run CMU8 is the best in different utility measures.

## 3 Conclusion

Our experimental results demonstrate that a high quality “off the shelf” adaptive information filtering system can be very effective for the TDT Supervised Tracking task. TREC-style adaptive filtering and TDT-style supervised tracking are very similar tasks. Only minor modifications to the TREC system were required to use it for TDT.

RUN	Utility	Utility Normalized	Utility Scaled
CMU1	-4964.58	-70.3956	0.0270
CMU2	-1248.07	-20.1269	0.1091
CMU3	317.37	0.3390	0.6917
CMU4	-1665.99	-11.5721	0.0985
CMU5	-504.75	-21.4876	0.2723
CMU6	161.32	-6.0389	0.4532
CMU7	459.92	0.1037	0.5911
<b>CMU8</b>	449.17	0.5921	<b>0.7281</b>
UMD1	261.81	-1.8287	0.3820
UMD2	233.64	-1.4206	0.3575
UMASS1	340.52	0.4019	0.6104
UMASS2	-614.56	-14.9671	0.1924
UMASS3	391.95	0.4934	0.6672
UMASS4	384.10	0.4148	0.6264

Table 1: The utilities of all submitted supervised adaptation topic tracking results reported by NIST.

We recognize that Supervised Tracking is just one of several tasks of interest to the TDT community, and perhaps not the most important. However, the specialized terminology and metrics used by the TDT community tend to emphasize the (minor) differences between adaptive filtering and supervised tracking, instead of the (more important) similarities. Our low-effort participation and relatively good results remind us that these two research communities have much in common, and suggest that both would benefit from closer cooperation in the future.

## 4 Acknowledgments

We thank Yiming Yang, Nian Li Ma, Jian Zhang, and Shinjae Yoo for providing valuable information about TDT data.

This work was sponsored in part by the Advanced Research and Development Activity in Information Technology (ARDA) under its Statistical Language Modeling for Information Retrieval Research Program. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors, and do not necessarily reflect those of the sponsor.

## References

- [1] J. Allan, J. Callan, W. Croft, L. Ballesteros, D. Byrd, R. Swan, and J. Xu. Inquiry does battle with trec-6. In *Proceeding of the Sixth Text REtrieval Conference (TREC-6)*, 1998.
- [2] J. Callan. Document filtering with inference networks. In *Proceedings of the Nineteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 262–269, 1996.
- [3] J. Callan. Learning while filtering. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1998.
- [4] K. Collins-Thompson, P. Ogilvie, Y. Zhang, and J. Callan. Information filtering, novelty detection, and named-page finding. In *Proceeding of the Eleventh Text REtrieval Conference (TREC-11)*, 2002.
- [5] Y. Zhang. Using bayesian priors to combine classifiers for adaptive. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004.
- [6] Y. Zhang and J. Callan. Maximum likelihood estimation for filtering thresholds. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 294–302, 2001.
- [7] Y. Zhang and J. Callan. Yfilter at TREC-9. In *Proceedings of the Ninth Text REtrieval Conference (TREC-9)*, pages 135–140. National Institute of Standards and Technology, special publication 500-249, 2001.
- [8] Y. Zhang and J. Callan. The bias problem and language models in adaptive filtering. In *The Tenth Text REtrieval Conference (TREC-10)*, pages 78–83. National Institute of Standards and Technology, special publication 500-250, 2002.