Filtering for Personal Web Information Agents

Gabriel L. Somlo and Adele E. Howe*
Computer Science Department, Colorado State University, Fort Collins, CO 80523, U.S.A.
{somlo, howe}@cs.colostate.edu

Categories and Subject Descriptors: H.3.3 [Information search and retrieval]: Information filtering

General Terms: Algorithms

Keywords: Web IR, adaptive filtering, continuous queries

1. TEXT FILTERING FOR THE WEB

Text filtering algorithms are a main component of a personalized Web information agent. Web information agents, such as WebMate [2] and Syskill & Webert [5], essentially combine text filtering and algorithms that generate the incoming stream of Web documents. However, these algorithms are subject to additional restrictions: negative feedback is considered an inconvenience to the user, the number of training samples is limited, and space is a consideration (thus, storing full text of all examples is discouraged).

We examine two types of adaptive filtering that appear suited to our requirements. The first one is based on TF-IDF and the second, which was used previously for a Web information agent [5], on a naive Bayes classifier (NBC). We present a new variant on NBC that does not require explicit negative feedback. In a study with TREC data, we find that, while all methods are capable of achieving good performance, they each incur a different penalty.

2. COMPARISON ON TREC DATA

The task requirements allowed us to focus on a small set of filtering algorithms. Which is best suited to personalized information filtering for the Web? To address this, we implemented a variant of an existing algorithm that better supports the requirements, evaluated different parameter settings for the methods and compared the performance of the best on pre-judged data, which simulates users.

2.1 Modified NBC

NBC requires both positive and negative training examples. Motivated by our requirements and inspired by Liu's S-EM [4], we implemented a simple algorithm that requires only positive examples:

1. Assume all unlabeled training documents are negative, and build an initial NBC based on this assumption.
2. Classify all unlabeled training documents using the classifier in Step 1, and sort them according to the difference in their conditional probabilities:
$$\Delta = [p(doc|-) \ - \ p(doc|+)]$$
3. Treat the *npdocs* unlabeled documents with the largest Δ, where *npdocs* is the number of *labeled positive* documents, as *pseudo-labeled negatives*.
4. Using the labeled positive and pseudo-labeled negative training samples, build a new NBC. Optionally, use EM on the rest of the unlabeled training samples. Use this newly built classifier on the test documents.

2.2 Experiment Design

We picked three topics each from the FBIS and LATIMES data collections on TREC Disk #5. The FBIS topics were: 189 (584 relevant, 695 non-relevant documents), 301 (339 relevant, 433 non-relevant documents), and 354 (175 relevant, 715 non-relevant documents). The LATIMES topics were: 374 (109 relevant, 315 non-relevant), 422 (98 relevant, 840 non-relevant) and 426 (145 relevant, 626 non-relevant). For each topic and method, a profile was built from the first half of the available relevant documents; we tested with the remaining documents using a simplified version of the Lewis and Gale F-metric [3] that assigns equal importance to both precision and recall, which we call LGF.

2.3 Effects of Parameters

To expedite fairness, we empirically tested to find the best parameter settings for the methods.

2.3.1 N-gram Size:

We tested whether using 2-grams in addition to single word terms would improve performance. For all the methods, paired sample t-tests showed that single words were significant better than 2-grams (e.g., for the TF-IDF method, LGF $\mu = 0.43$ with $sd = 0.19$ for 1-gram and $\mu = 0.26$ with $sd = 0.20$ for 2-gram).

2.3.2 Stop Word Elimination:

Does eliminating stop-words really improve filtering quality? Paired t-tests showed no significant difference in performance due to stop word elimination ($p > 0.05$).

*We thank the anonymous reviewers for their insightful comments. Adele Howe was partially supported by the National Science Foundation under Grant No. IIS-0138690.

SIGIR'04, July 25–29, 2004, Sheffield, UK
ACM 1-58113-881-4/04/0007.

set	topic	adapt.	minmax	NBC	mod-NBC
FBIS	189	0.481	0.607	0.662	0.644
	301	0.495	0.579	0.703	0.628
	354	0.321	0.352	0.509	0.450
LA	374	0.450	0.614	0.523	0.462
	422	0.405	0.406	0.644	0.210
	426	0.477	0.451	0.075	0.351

Table 1: Best LGF values, topics×algorithms

2.3.3 *Dissemination Threshold Computation:*

For the TF-IDF methods, how similar must a document be to the profile before it is disseminated? Several approaches have been proposed to determining the dissemination threshold (e.g., [1, 6]).

We include two methods: an adaptive and a fixed fraction based on the profile size. Callan's adaptive threshold starts low and gradually increases to improve precision [1]. Our more lightweight version of Callan's method incrementally modifies the threshold, based on documents judged relevant and a learning rate [7]. For our second method, we adopt as a threshold a fixed fraction of the size of the profile, which is defined as the distance between the minimum and maximum similarity between known relevant documents and the current profile vector.

We tested a range of learning rates for adaptive and fractions for fixed. We found that both had a significant effect on performance as shown by one-way ANOVAS (for adaptive, $F = 36.1$, $p \leq 0.0001$ and for min-max, $F = 10.9$, $p \leq 0.001$). The best learning rate ($LGF = 0.446$) is 0.09. The best distance fraction ($LGF = 0.509$) is 0.04.

2.3.4 *Inclusion of Unlabeled Documents:*

The corpus includes a large number of unlabeled documents. For the NBC, we tested whether the vocabulary should include terms from the unlabeled documents. We found that including the extra terms significantly degraded performance ($t = 5.596$, $p < 0.0001$ with $\mu = 0.27$ and $sd = 0.32$ for all terms and $\mu = 0.56$ and $sd = 0.16$ for terms from labeled training samples only.) Because the negative class contains many more documents than the positive, the underlying mathematics cause the conditional probabilities of the terms that appear in labeled training samples to be diminished by the terms that never appear in labeled documents, significantly reducing recall.

2.4 Performance of the Methods

Results are presented by topic in Table 1 and Figure 1 for the best parameterizations of the four methods. For the TF-IDF methods, min-max outperforms adapt (higher LGF on all but one topic); a paired sample t-test yields a marginal advantage for min-max ($t = 1.813$, $p = 0.084$).

Min-max demands more storage than adaptive, requiring TF-IDF vectors for all feedback documents. We examined the effect of training and found that while adaptive thresholds fluctuate within topic dependent ranges, min-max thresholds quickly stabilize at about 0.1 for all topics. If hard-coding the dissemination threshold proves viable beyond of our tested topics, it would eliminate the need to store the document vectors.

Generally, the mod-NBC does a little worse than NBC; both perform better on the FBIS topics.

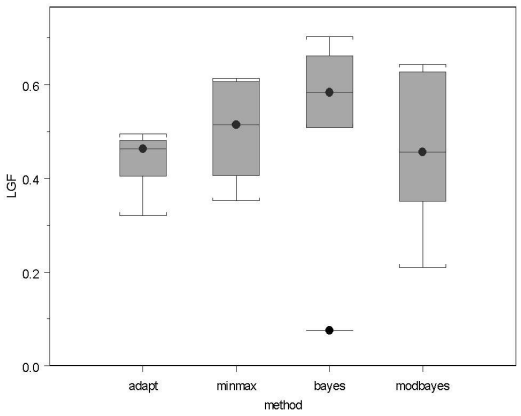

Figure 1: Box plot of best LGF results by algorithm

3. WEB INFORMATION AGENTS

All methods provide acceptable filtering performance, but with different penalties. TF-IDF methods require more training. NBC requires negative feedback, and our version of NBC without explicit negative feedback requires storage of past documents.

In a pilot user study of a system for tracking users' queries on the Web we found that if the user will give negative feedback, it does significantly improve filtering performance. The study also shows that combining filtering decisions by disseminating based on either NBC or TF-IDF leads to improved recall. Thus, we advocate NBC as the best performing method, on both a pre-judged corpus and in a small user study, for applications such as Web information agents which provide restricted feedback and storage.

4. REFERENCES

[1] J. Callan. Learning while filtering documents. In *Proceedings of the 21st International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 1998.

[2] L. Chen and K. Sycara. Webmate: A personal agent for browsing and searching. In *Proceedings of the Second International Conference on Autonomous Agents*, Minneapolis, MN, USA, 1998.

[3] D.D. Lewis and W.A. Gale. A sequential algorithm for training text classifiers. In *Proc. of the 17th Inter. ACM-SIGIR Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994.

[4] X. Liu, B. Li. Learning to classify text using positive and unlabeled data. In *Proc. of the 18th Inter. Joint Conference on Artificial Intelligence*, Acapulco, Mexico, 2003.

[5] M.J. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27:313–331, 1997.

[6] S. Robertson. Threshold setting and performance optimization in adaptive filtering. *Information Retrieval*, 5:239–256, 2002.

[7] G.L. Somlo and A.E. Howe. Adaptive lightweight text filtering. In *Proc. of the 4th Inter. Symposium on Intelligent Data Analysis (IDA 2001)*, Lisbon, Portugal, 2001.