Filtering for Personal Web Information Agents

Gabriel L. Somlo
and
Adele E. Howe
Computer Science Dept.
Colorado State University

Objectives

- Incorporate text filtering into personal Web information agents
- Desired properties of agent-embedded filtering:
  - Avoid negative feedback
  - Learn quickly, with limited training
  - Incremental learning (avoid storing training instances)

Filtering Algorithms and Parameters

- TF-IDF representation + cosine similarity
  - 1- and 2-grams
  - stop-word pruning (y/n)
  - adaptive vs. min-max-ratio dissemination threshold
- Naïve Bayes Classifier
  - use terms from unlabeled documents?
  - how to avoid using labeled negatives?

Evaluation

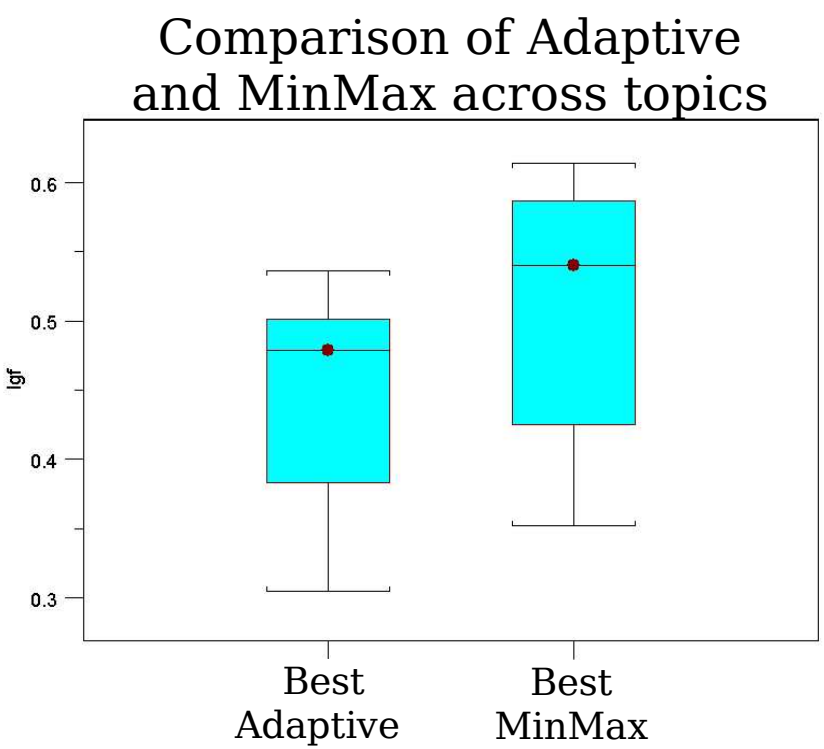
- Data Set: TREC Disk #5
  - FBIS: 130,471 documents
  - LATimes: 127,742 documents
- Six topics:

Topic	Corpus	# Relevant	# Non-Rel.	Med. Rel. Pos.
189	FBIS	584	695	62,595
301	FBIS	339	433	50,695
354	FBIS	175	715	64,424
374	LATimes	109	315	58,943
422	LATimes	98	840	70,875
426	LATimes	145	626	67,988

- Metric: harmonic mean $$\mathrm{HM} = \frac{2 \cdot \mathrm{recall} \cdot \mathrm{precision}}{\mathrm{recall} + \mathrm{precision}}$$

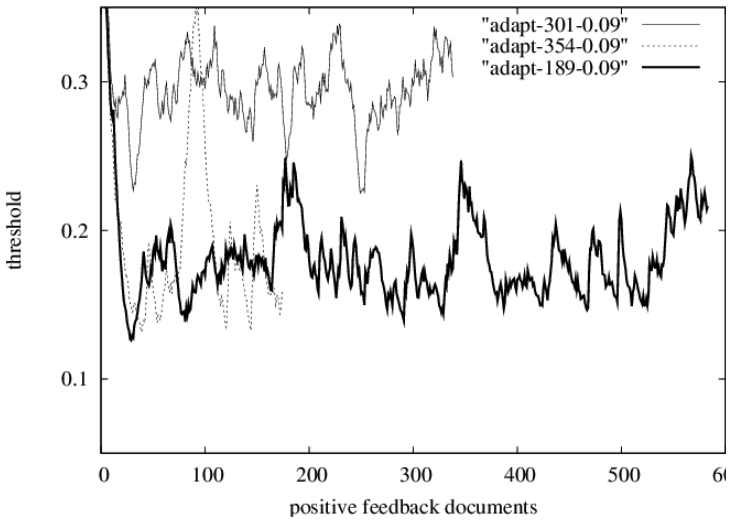
TF-IDF Parameter Analysis

- 2-grams perform worse than single terms
- Stop-word removal does not improve HM
- Threshold learning: min-max outperforms adaptive learning

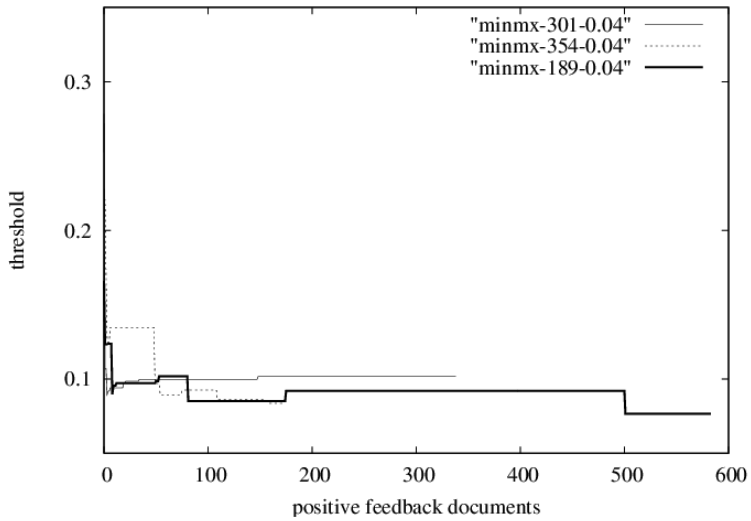


TF-IDF: Analysis of Learning the Dissemination Threshold

adaptive



min-max



- Static threshold = 0.1 comparable to best learned threshold – learning may be unnecessary!

Naïve Bayes Parameter Analysis

- Because corpus is biased toward non-relevant documents...
- Using terms from unlabeled documents is a terrible idea:

$$p(t|C) = \frac{1}{npos(C) + nterms}$$

$$npos(C) \equiv \# \text{ of term positions in class } C$$

$$nterms \equiv \text{size of vocabulary}$$

$$npos(+) \ll npos(-)$$

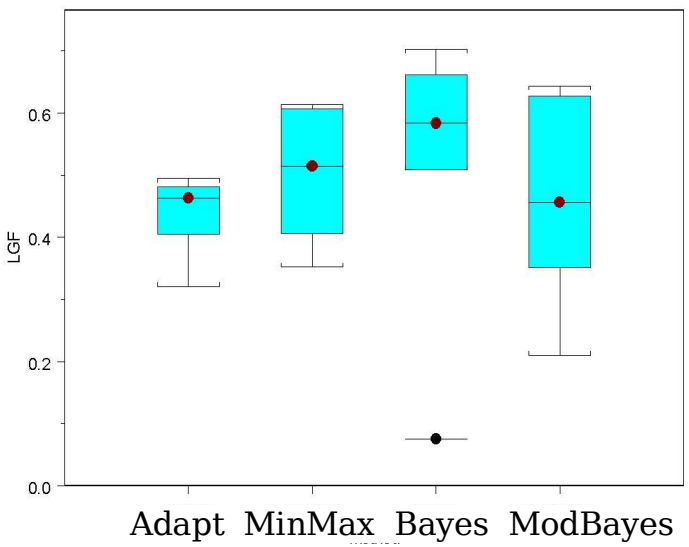
- Recall goes down as “positive” terms are discounted!

Modified Naïve Bayes without Negative Feedback

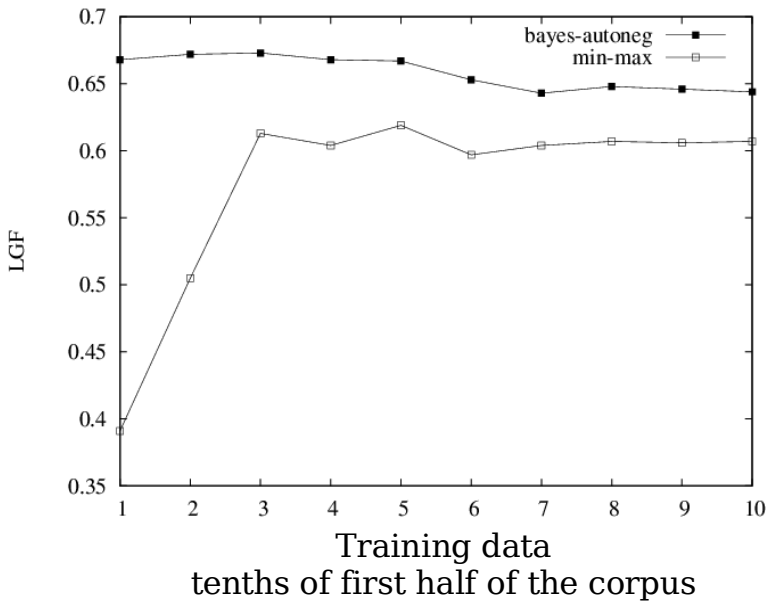
- Avoiding explicit negative feedback:
  1. Build initial classifier assuming all unlabeled docs are non-relevant
  2. Classify all unlabeled docs using initial classifier, and sort by $\triangle = [\, p(doc|-) - p(doc|+) \,]$
  3. From unlabeled docs with largest $\triangle$, pick a number equal to that of the labeled relevant docs
  4. Build a new classifier from labeled relevant and picked non-relevant docs
- Equal numbers of relevant and non-relevant docs avoids problem shown on previous slide

Four Algorithms Compared

Harmonic Mean



learning speed



- Bayes performs better and learns faster !

Conclusions

Bayes: best performance, but requires negative feedback

ModBayes and MinMax not incremental

We may be able to bypass TF-IDF threshold learning and hard-code to 0.1

Bayes wins if we can convince users to supply negative feedback!