Temporal Feedback for Tweet Search with Non-Parametric Density Estimation

Paper Presentation

Axel Magnuson

Department of Computer Science
Boise State University

February 3, 2015

Contributions

1. Investigation of the Temporal Clustering Hypothesis
2. A Method to Characterize Temporal Density

Dataset

TREC 2011/2012

- 1% Sample of Twitter Activity
- 114K Tweets
- 109 Topics
- Classified as “not relevant”, “relevant”, or “highly relevant” by NIST assessors

Temporal Clustering Hypothesis

Temporal Cluster Hypothesis

“In search tasks where time plays an important role[. . .], we hypothesize that relevant documents tend to cluster together in time, and that this property can be exploited to improve search effectiveness.”[1]

Classic Cluster Hypothesis

“Relevant documents tend to share similar content[. . .].” [1]

Temporal Clustering Hypothesis

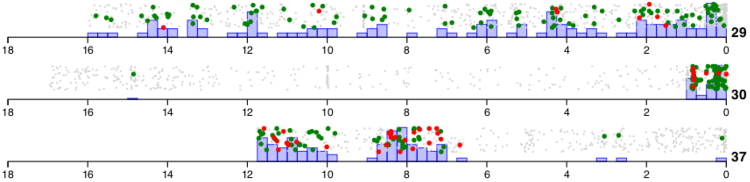


Figure 1: Visualizations illustrating the temporal distribution of retrieved documents and relevant documents for three topics from the TREC 2011 Microblog track: topic 29, “global warming and weather”, topic 30 “Keith Olbermann new job”, and topic 37 “Giffords recovery”. The timeline is measured in days, anchored by the query time on the right edge. Green dots represent relevant documents, red dots represent highly-relevant documents, and gray dots represent non-relevant documents. The bar graphs show bucketed distributions of the relevant and highly-relevant documents.

“These visualizations confirm our intuition that the temporal distribution of relevant tweets is highly non-uniform”[1]

Ranking Model

Idea

We need some way to combine our temporal information with usual retrieval methods.

Probability to the Rescue

There are plenty of IR methods out there that frame document scores in terms of their relevance probability, and sort based on that rank. We can hook into that!

Ranking Model

Query Likelihood

$$P(D|Q) \propto P(Q|D)P(D) \tag{1}$$

Log-Linear Temporal Model

$$\begin{aligned} \log P_\alpha(R|D,Q) = Z_\alpha &+ (1-\alpha)\log P(R|W_D,Q) \\ &+ \alpha \log P(R|T_D,Q) \end{aligned} \tag{2}$$

$$P_\alpha(R|D,Q) \backsim P(R|W_D,Q)^{1-\alpha} \times P(R|T_D,Q)^\alpha \tag{3}$$

- α Expresses the weight of temporal evidence.
- Z_α is a normalization constant.

Ranking Model

Why the Log-Linear Model?

“Lexical and temporal evidence may differ inherently in importance, but this should not be controlled via the temporal model itself.” [1]

$$P_\alpha(R|D,Q) \backsim P(R|W_D,Q)^{1-\alpha} \times P(R|T_D,Q)^\alpha$$

Temporal Feedback

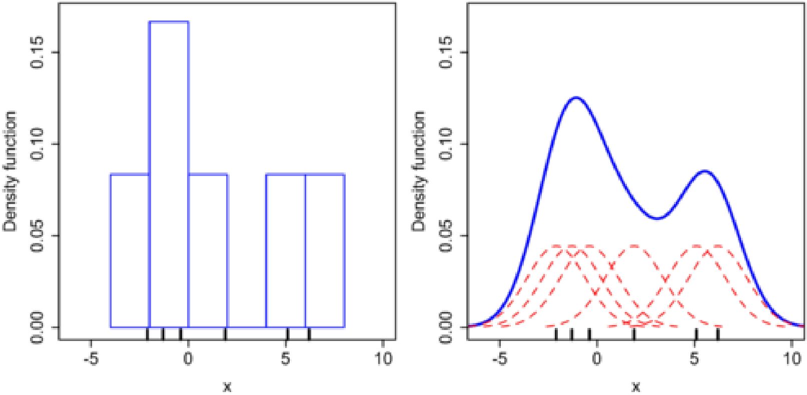
$$P(R|T_D, Q) \tag{4}$$

Distribution of Documents to Q Over Time

- Density f_Q over time
- f_Q is larger when relevant documents are likely to appear

Temporal Feedback

Kernel Density Estimation



Temporal Feedback

Kernel Density Estimation

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=0}^{n} \omega_i K \left(\frac{x - x_i}{h} \right) \tag{5}$$

Kernel Function

$$K \left(\frac{x - y}{h} \right) = \mathcal{N} \left(\frac{x - y}{h}, 0, h \right) \tag{6}$$

In both these cases, *h* is the *bandwidth* variable.

Temporal Feedback

Bandwidth Selection

We want to identify the *optimal* bandwidth h^*. This problem has received a lot of attention in statistical literature, so 3 methods were compared.

1. Silverman's Rule of Thumb (RT)
2. Mean integrated square error (CV)
3. Sheather and Jones method (SJ)

Silverman's Rule of Thumb

$$h^* = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{-\frac{1}{5}} \tag{7}$$

Temporal Feedback

Bandwidth Selection

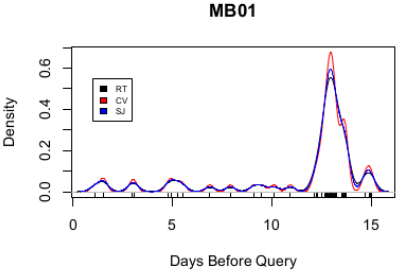


Figure 2: Kernel density estimates for the topic MB01. Each curve corresponds to a different bandwidth selection method.

Temporal Feedback

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=0}^{n} \omega_i K \left(\frac{x - x_i}{h} \right)$$

Weighting Schemes

1. Uniform Weights
2. Score-Based Weights
3. Rank-Based Weights
4. True Feedback Based Weights

Score-Based Weights

$$\omega_i^s = \frac{P(Q|D)}{\sum_{j=1}^{n} P(Q|D_j)} \tag{8}$$

Rank-Based Weights

$$\omega_i^r = \frac{\lambda e^{-\lambda r_i}}{\sum_{j=1}^{n} \lambda e^{-\lambda r_j}} \tag{9}$$

True Feedback Based Weights

$$\bar{\omega}_i^s = \begin{cases} c & \text{if } D_i \text{ is relevant} \\ \omega_i^s & otherwise \end{cases} \tag{10}$$

Experimental Evaluation

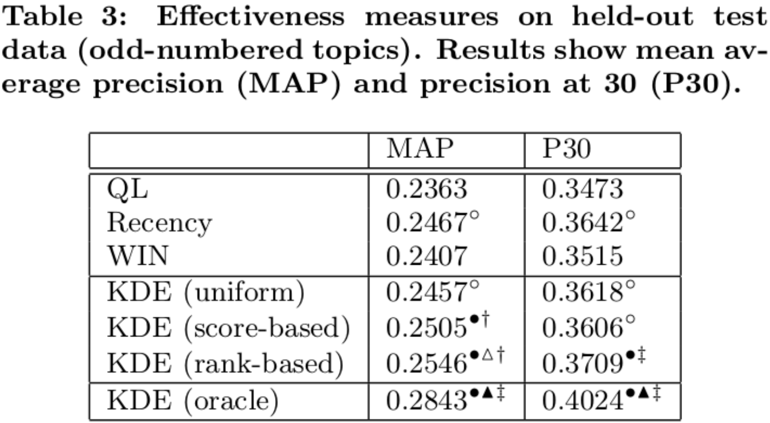


Table 3: Effectiveness measures on held-out test data (odd-numbered topics). Results show mean average precision (MAP) and precision at 30 (P30).

	MAP	P30
QL	0.2363	0.3473
Recency	0.2467°	0.3642°
WIN	0.2407	0.3515
KDE (uniform)	0.2457°	0.3618°
KDE (score-based)	0.2505$^{\bullet\dagger}$	0.3606°
KDE (rank-based)	0.2546$^{\bullet\triangle\dagger}$	0.3709$^{\bullet\ddagger}$
KDE (oracle)	0.2843$^{\bullet\blacktriangle\ddagger}$	0.4024$^{\bullet\blacktriangle\ddagger}$

Experimental Evaluation

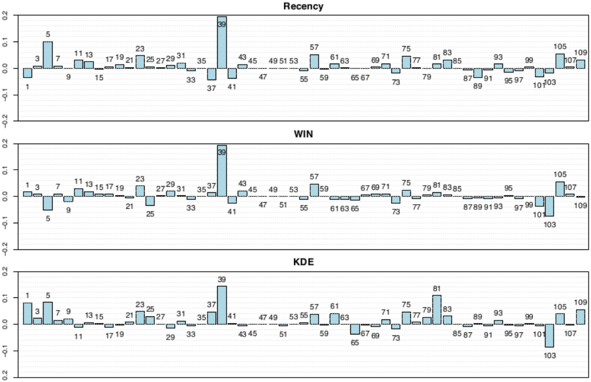


Figure 3: Per-query differences in average precision for each temporal model vs. the query-likelihood baseline.

Experimental Evaluation

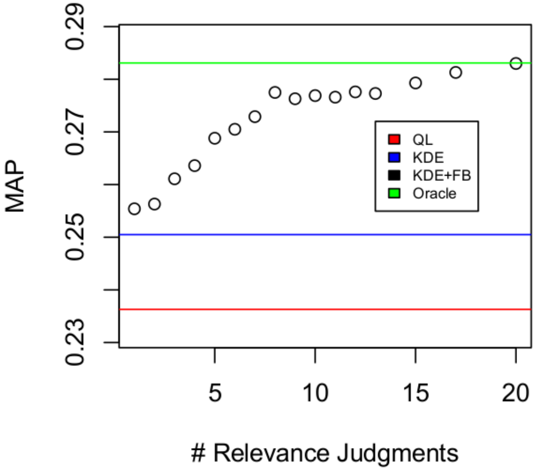


Figure 4: Effectiveness of KDE given an increasing number of user-supplied relevance judgments.

Experimental Evaluation

“Is the improvement that we see due to a signal that is different from information gleaned from document content?” [1]

Lexical Feedback

1. Retrieve initial set of documents
2. Apply the temporal retrieval model (Recency, WIN, or KDE) to rerank results
3. From the top k reranked documents, estimate feedback models: [. . .] [select] the top *k* documents and [assume] that they are relevant

Experimental Evaluation

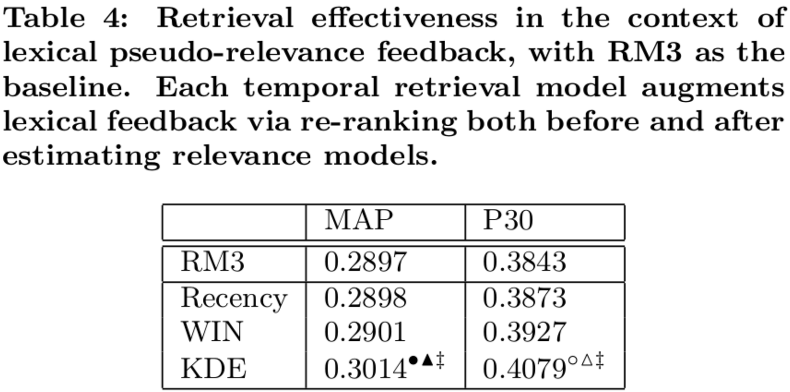


Table 4: Retrieval effectiveness in the context of lexical pseudo-relevance feedback, with RM3 as the baseline. Each temporal retrieval model augments lexical feedback via re-ranking both before and after estimating relevance models.

	MAP	P30
RM3	0.2897	0.3843
Recency	0.2898	0.3873
WIN	0.2901	0.3927
KDE	$0.3014^{\bullet\blacktriangle\ddagger}$	$0.4079^{\circ\vartriangle\ddagger}$

Experimental Evaluation

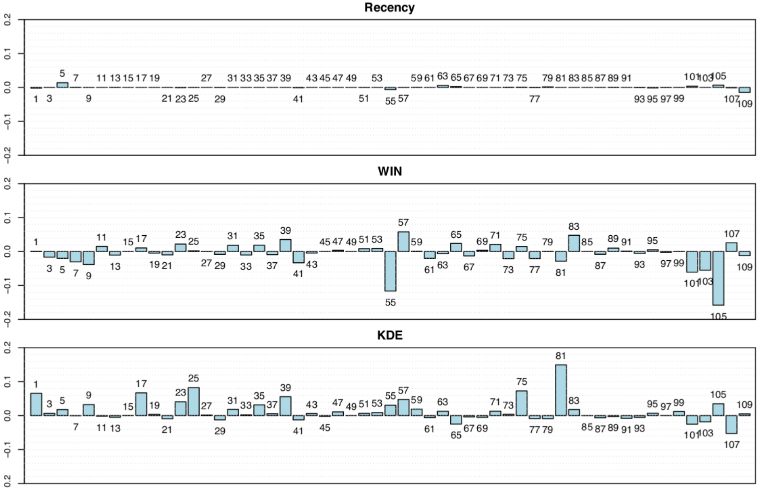


Figure 5: Per-query differences in average precision for each temporal model vs. RM3.

Discussion

Q & A

- Does this confirm the temporal hypothesis?
- What are some computational concerns with this approach?
- How would you try to modify this approach for better results?
- What parts were unclear and need more discussion?

Further Reading I

M. Efron, J. Lin, J. He, and A. de Vries.
Temporal Feedback for Tweet Search with Non-Parametric Density Estimation.
SIGIR 2014 Proceedings, July 6-11, 2014.