The 9th ACM International Conference on Web Search and Data Mining (WSDM '16)

BARBARA MADE THE NEWS: MINING THE BEHAVIOR OF CROWDS FOR TIME-AWARE LEARNING TO RANK

Flávio Martins*, João Magalhães* and Jamie Callan†
* NOVA LINCS, Universidade NOVA de Lisboa
† LTI, Carnegie Mellon University

What is happening now prompts users on the Web to produce and interact with new posts about newsworthy events giving rise to trending topics.

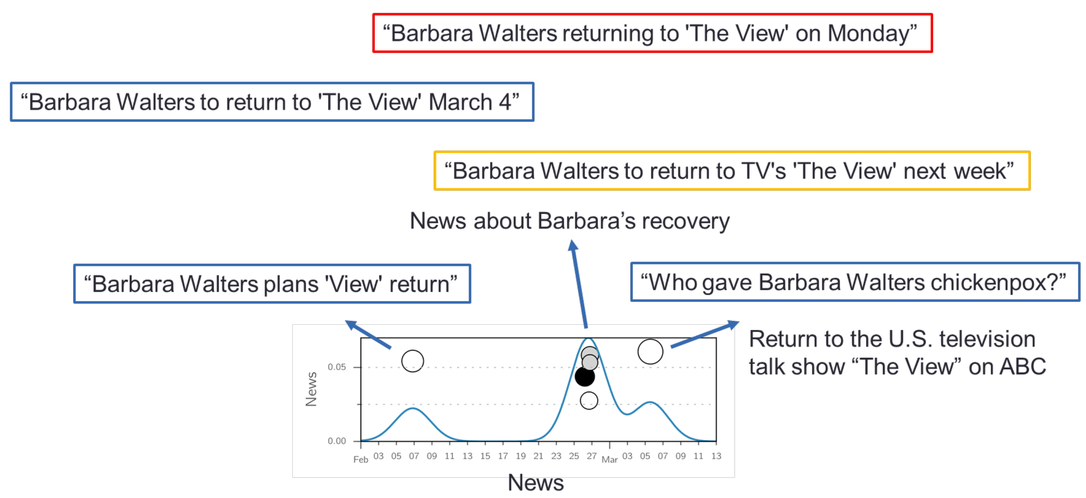
We leverage on the behavioral dynamics of the crowd to estimate a topic's temporal relevance i.e., the most relevant time periods for a search query.

Considering a set $S \in \{S_1, S_2, \ldots, S_k\}$ of information sources that reflect a real-world event, our goal is to discover the relevance of a timestamp for a particular query. In other words, we wish to infer a function:

$$f_{s_k}(q_a, t_b) \in [0,1] \qquad P(r \mid t_{d_b}, q_a, s_k)$$

The probability density function $f_{s_k}(t_{d_b}|q_a)$ is approximated by a kernel density estimation which is advantageous due to the natural smoothness of the resulting function:

$$\widehat{f_{s_k}}(t) = \frac{1}{nh}\sum_{i=0}^{n} w_i^{s_k} K\left(\frac{t - t_i^{s_k}}{h}\right)$$

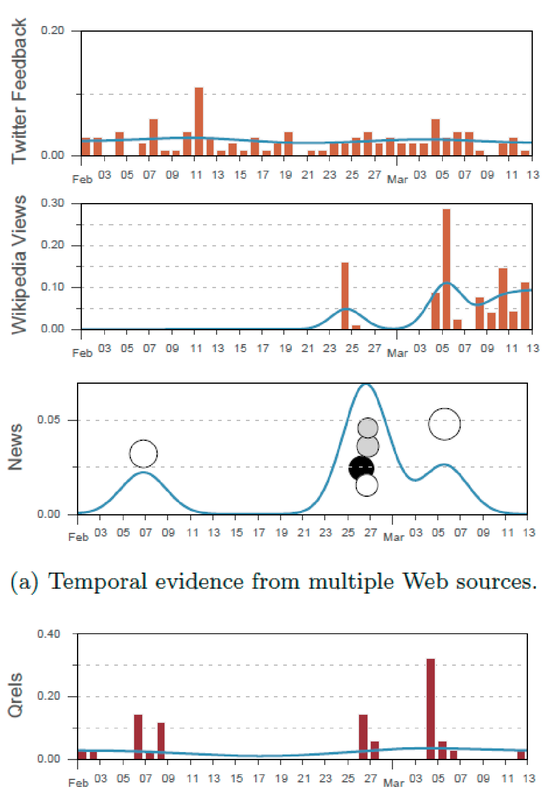


Mining Temporal Crowd Signals from Multiple Sources

For KDE each timestamp is weighted differently depending on the source:

Temporal Feedback	$w_i^{s_t} = QL(q_a \mid d_b)$
Wikipedia Views	$w_i^{s_v} = v_i - \min(v_i, \overline{v})$
Wikipedia Edits	$w_i^{s_e} = \sum_j TF_j(q_a, e)$
News	$w_i^{s_h} = J(q_a, h_i)$

Query: Barbara Walters, chicken pox



(a) Temporal evidence from multiple Web sources.

RMTS: Time-aware Ranking in Twitter with Multiple Sources

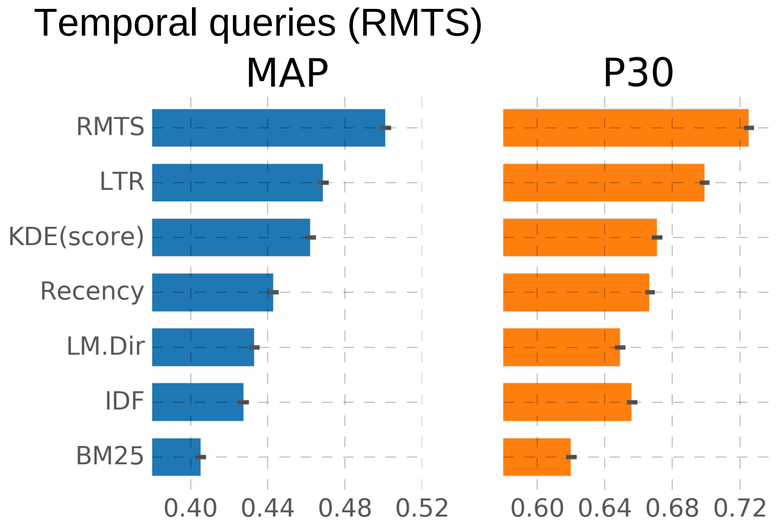
$$RMTS(q_a, t_{d_b}, d_b) = \sum_i \alpha_i f_{l_i}(q_a, d_b) + \sum_j \beta_j f_{c_j}(d_b) + \sum_k \gamma_k f_{s_k}(q_a, t_{d_b})$$

Lexical features · Domain features · Temporal features

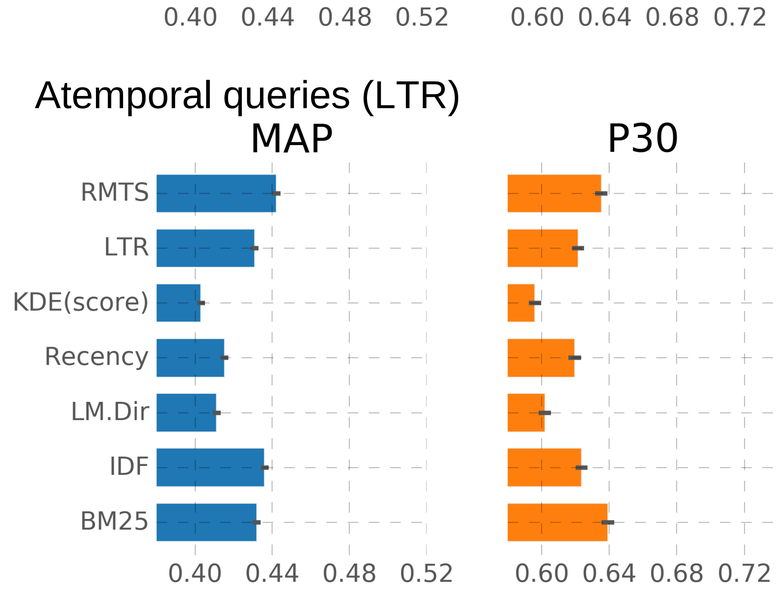
$$\mathrm{LTR}(q_a, d_b) = \sum_i \alpha_i f_{l_i}(q_a, d_b) + \sum_j \beta_j f_{c_j}(d_b)$$

Lexical features · Domain features

Temporal queries (RMTS)

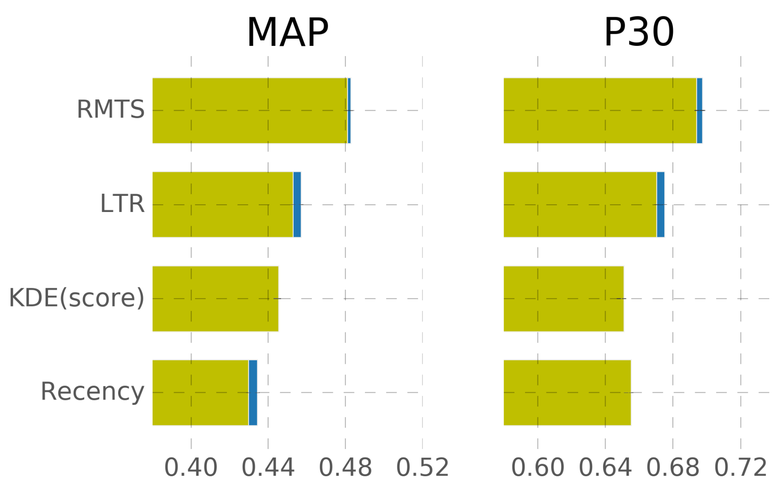


Atemporal queries (LTR)



Results and Conclusions

Global results on TREC 2014 dataset



Retrieval precision: the proposed approach statistically significantly outperforms the BM25 and LM.Dir models by approximately 13.2% and a strong learning to rank model (LTR) by 6.2%.

RMTS is less biased: it explores temporal signals from multiple Web sources to estimate the temporal relevance.

Unified representation of temporal signals: a representation allows predicting temporal relevance from heterogeneous sources.