A Two-Stage Model for Blog Feed Search

Task
Identify blogs that show a recurring interest in a topic

Approach
Blogs as complex information objects
Two-stage approach to identifying these objects

Stage 1
Exploratory search by salient features
\( \rightarrow \) select set of blogs \( B \) from interesting posts

\[ B = \{blog | \sum_{post \in N} P(Q | \theta_{post})P(post | blog) > 0 \} \]

Stage 2
In-depth examination of objects
\( \rightarrow \) use all posts to determine topical centrality of blog, \( P(Q | blog) \)

\[ P(Q | blog) \propto \prod_{t \in Q} P(t | \theta_{blog})^{\alpha(t,Q)} \]

Experiments
In stage 1: prune list of blog posts \( (N) \), dependent and independent of topic, and use alternative document representations.

Test on TREC Blog 2007 and 2008 topics.

Conclusions
Two-stage model improves over blog-based model. Topic-dependent pruning and lean document representation improve early precision and efficiency.

Wouter Weerkamp
Krisztian Balog
Maarten de Rijke
ISLA, University of Amsterdam

This research was supported by the European Union’s ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 295430, by the Center for Creation, Content and Technology (CCT), by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802.