A Two-Stage Model for Blog Feed Search

Task

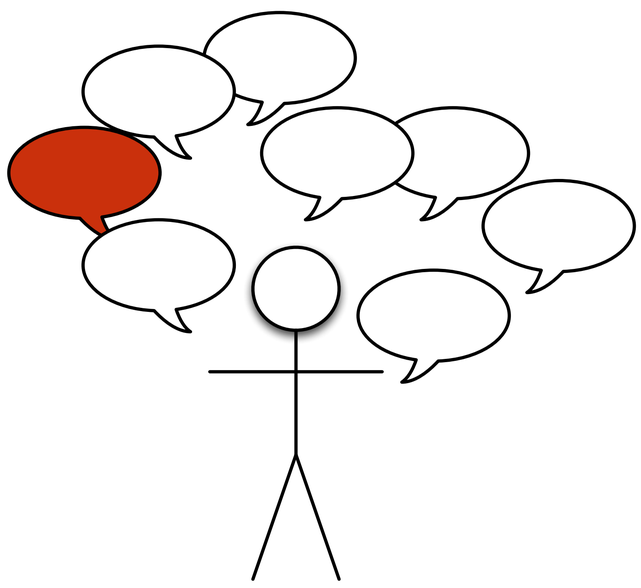
Identify blogs that show a recurring interest in a topic

Approach

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

Stage 1

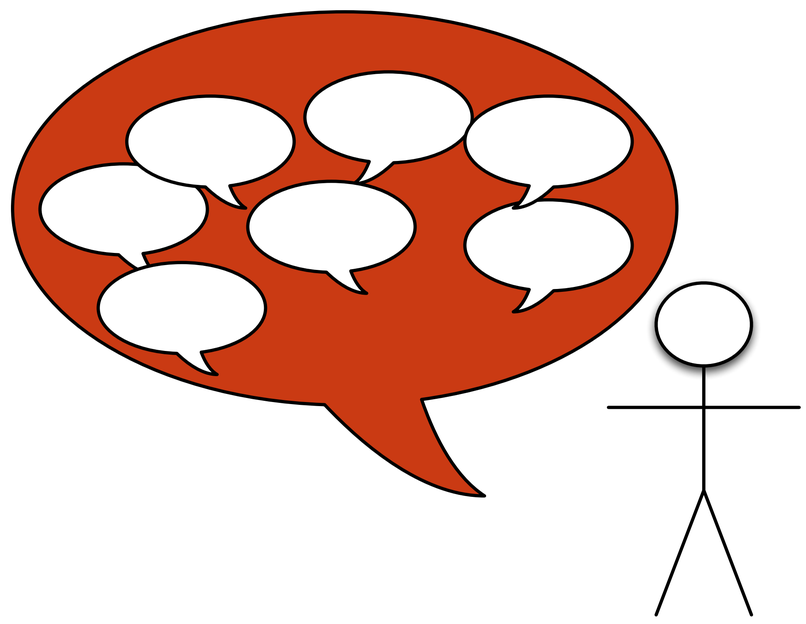
Exploratory search by salient features
➔ select set of blogs $\mathcal{B}$ from interesting posts



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

Stage 2

In-depth examination of objects
➔ use all posts to determine topical centrality of blog, *P(Q|blog)*



$$P(Q|blog) \propto \prod_{t \in Q} P(t|\theta_{blog})^{n(t,Q)}$$

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.

Results

2007 topics		MAP	P@5	MRR
Blog-based model		0.3260	0.5422	0.7193
Two-stage model				
Representation	Pruning			
full content	1,700	0.3348▲	0.5422	0.7213
full content	topic-dep.	0.3611▲	0.5689△	0.7243
title-only	-	0.3549△	0.6444▲	0.8476▲
title-only	7,000	0.3577▲	0.6622▲	0.8587▲
title-only	topic-dep.	**0.3813**▲	**0.6889**▲	**0.8604**▲
2008 topics				
Blog-based model		0.2521	0.4880	0.7447
Two-stage model				
Representation	Pruning			
full content	1,700	0.2551	0.4960	0.7483
full content	topic-dep.	**0.2747**▲	**0.5080**	0.7504
title-only	-	0.2363	0.4880	0.7524
title-only	7,000	0.2368	0.4840	0.7524
title-only	topic-dep.	0.2571	**0.5080**	**0.7591**

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 250430, by the Center for Creation, Content and Technology (CCCT), 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.