

A Two-Stage Model for Blog Feed Search

Strategy for identifying a complex information object:

exploratory search
(using salient features)

in-depth examination
of objects

stage 1



stage 2

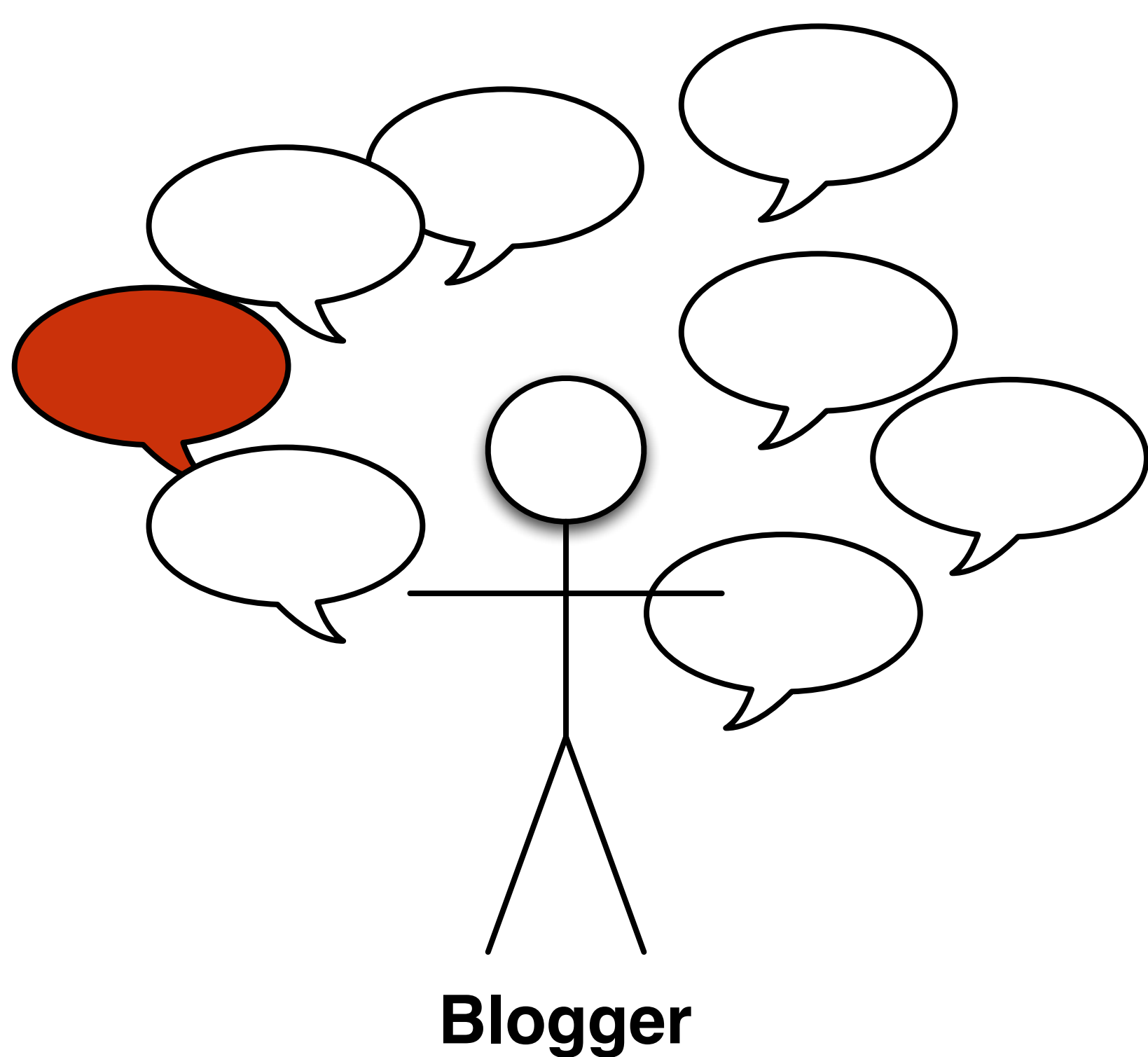


Identify blogs that show a recurring interest in a topic

individual blog posts (salient features) for blog selection

use all posts to determine topical centrality of the blog

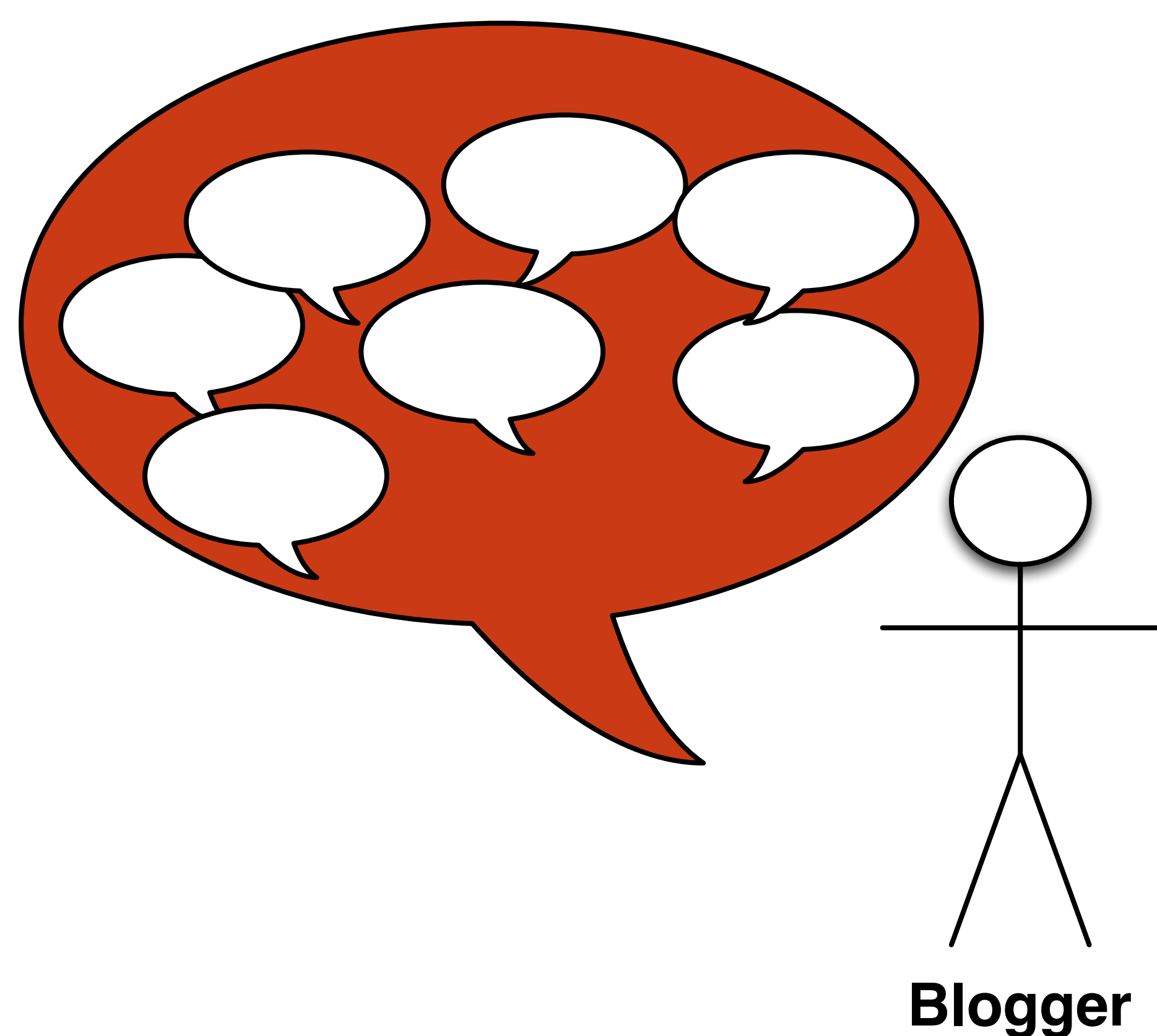
Stage 1



Blogger

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

Stage 2



Blogger

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

Experiments

In stage 1, we can prune the list of blog posts (N), both topic-dependent and independent, and use various document representations.

We test on TREC Blog 2007 and 2008 topics (95 topics).

Conclusions

The two-stage model improves over a blog-based model; topic-dependent pruning and a lean doc. representation improve early precision and efficiency.

Results

2007 topics		MAP	P@5	MRR
<i>Blog-based model</i>		0.3260	0.5422	0.7193
<i>Two-stage model</i>				
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				
<i>Blog-based model</i>		0.2521	0.4880	0.7447
<i>Two-stage model</i>				
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