

Three Blog Retrieval Myths

Gilad Mishne

Myth: Blog searchers are looking for highly-linked-to posts.

- Use of link analysis methods does not result in improved performance
 - Average indegree of relevant documents *lower* than non-relevant ones
- Does anchor text help?
 - Propagating anchors results in marginal improvements if any
 - Reasons: >90% of anchors come from inter-blog automated links (e.g., archives); non-internal anchors are mostly the blog title name - good for named-page finding, but not for opinion retrieval

Link usage in the TREC Blog06 corpus

Ave. indegree of relevant documents:	2.4
Ave. indegree of non-relevant documents:	4.6
Posts with anchor text from incoming links:	2.3M
Posts for which anchor text contained new terms:	0.1M

Comparison of RSS/Atom-based and HTML-based indices

Resource	Reduction for RSS/Atom only
Bandwidth/Storage	57%
Index size	67%
Index creation time	91%
Retrieval	34%

Index	MAP	R-Prec	bpref	P@10
RSS/Atom	0.1449	0.2357	0.2393	0.3479
HTML	0.1797	0.2452	0.2564	0.3560

Myth: Indexing RSS content is better than indexing the entire HTML content, since RSS provides clean, structured data.

- Using the syndicated content rather than the HTML has many benefits including lower demands on bandwidth, storage, and computation time
 - Many commercial blog search engines mostly follow RSS/Atom feeds rather than crawl blog HTML
- However, indexing syndicated content results in substantial performance degradation, particularly for recall

Myth: Usage of specialized techniques for opinion retrieval (e.g., NLP approaches) does not improve over robust topical-only ranking, as can be observed from the TREC baselines which are comparable to top-performing systems.

- A good topical ranking formula goes a long way
- However, the following approaches all improve over state-of-the-art topical ranking;
 - Spam filtering through text classification (trained on livejournal vs. blogspot): +6%
 - Lexicon-based sentiment analysis (used the General Inquirer): +11%
 - Recency scoring for “recent” blog post (query date estimated from top ranked docs): +3%
- The following approaches seem promising, but work is preliminary:
 - Taking into account the number of comments on a post
 - Classification-based sentiment analysis

Effectiveness of reranking topical retrieval results using various opinion-retrieval components

TREC submission: RSS	MAP	Post-TREC: HTML	MAP
Baseline (topical retrieval)	0.1449	Baseline (topical retrieval)	0.1797
Baseline+SpamFiltering	0.1523	Baseline+SpamFiltering	0.1864
Baseline+SentimentAnalysis	0.1596	Baseline+SentimentAnalysis	0.1999
Baseline+AllMethods	0.1795	Baseline+AllMethods	0.2241

