UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Credibility improves topical blog post retrieval

Weerkamp, W.; de Rijke, M.

Publication date
2008

Published in
ACL-08: HLT: 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Proceedings of the conference

Link to publication

Citation for published version (APA):
Weerkamp, W., & de Rijke, M. (2008). Credibility improves topical blog post retrieval. In *ACL-08: HLT: 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Proceedings of the conference* (pp. 923-931). Association for Computational Linguistics (ACL). http://www.aclweb.org/anthology/P/P08/P08-1105

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:13 Oct 2026

Credibility Improves Topical Blog Post Retrieval

Wouter Weerkamp
ISLA, University of Amsterdam
weerkamp@science.uva.nl

Maarten de Rijke
ISLA, University of Amsterdam
mdr@science.uva.nl

Abstract

Topical blog post retrieval is the task of ranking blog posts with respect to their relevance for a given topic. To improve topical blog post retrieval we incorporate textual credibility indicators in the retrieval process. We consider two groups of indicators: *post level* (determined using information about individual blog posts only) and *blog level* (determined using information from the underlying blogs). We describe how to estimate these indicators and how to integrate them into a retrieval approach based on language models. Experiments on the TREC Blog track test set show that both groups of credibility indicators significantly improve retrieval effectiveness; the best performance is achieved when combining them.

1 Introduction

The growing amount of user generated content available online creates new challenges for the information retrieval (IR) community, in terms of search and analysis tasks for this type of content. The introduction of a blog retrieval track at TREC (Ounis et al., 2007) has created a platform where we can begin to address these challenges. During the 2006 edition of the track, two types of blog post retrieval were considered: *topical* (retrieve posts about a topic) and *opinionated* (retrieve opinionated posts about a topic). Here, we consider the former task.

Blogs and blog posts offer unique features that may be exploited for retrieval purposes. E.g., Mishne (2007b) incorporates time in a blog post retrieval model to account for the fact that many blog queries and posts are a response to a news event (Mishne and de Rijke, 2006). Data quality is an issue with blogs—the quality of posts ranges from low to edited news article-like. Some approaches to post retrieval use indirect quality measures (e.g., elaborate spam filtering (Java et al., 2007) or counting inlinks (Mishne, 2007a)).

Few systems turn the *credibility* (Metzger, 2007) of blog posts into an aspect that can benefit the retrieval process. Our hypothesis is that more credible blog posts are preferred by searchers. The idea of using credibility in the blogosphere is not new: Rubin and Liddy (2006) define a framework for assessing blog credibility, consisting of four main categories: blogger's expertise and offline identity disclosure; blogger's trustworthiness and value system; information quality; and appeals and triggers of a personal nature. Under these four categories the authors list a large number of indicators, some of which can be determined from textual sources (e.g., literary appeal), and some of which typically need non-textual evidence (e.g., curiosity trigger); see Section 2.

We give concrete form to Rubin and Liddy (2006)'s indicators and test their impact on blog post retrieval effectiveness. We do not consider all indicators: we only consider indicators that are textual in nature, and to ensure reproducibility of our results, we only consider indicators that can be derived from the TRECBlog06 corpus (and that do not need additional resources such as bloggers' profiles that may be hard to obtain for technical or legal reasons).

We detail and implement two groups of credibility indicators: *post level* (these use information about individual posts) and *blog level* (these use information from the underlying blogs). Within the post level group, we distinguish between topic dependent and independent indicators. To make matters concrete, consider Figure 1: both posts are relevant to the query "tennis," but based on obvious surface level features of the posts we quickly determine *Post 2* to be more *credible* than *Post 1*. The most obvious features are spelling errors, the lack of leading capitals, and the large number of exclamation marks and

Proceedings of ACL-08: HLT, pages 923–931,
Columbus, Ohio, USA, June 2008. ©2008 Association for Computational Linguistics

Post 1
as for today (monday) we had no school! yaay labor day. but we had tennis from 9-11 at the highschool. after that me suzi melis & ashley had a picnic at cecil park and then played tennis. i just got home right now. it was a very very very fun afternoon. (...) we will have a short week. mine will be even shorter b/c i wont be there all day on friday cuz we have the Big 7 Tournament at like keystone oaks or sumthin. so i will miss school the whole day.

Post 2
Wimbledon champion Venus Williams has pulled out of next week's Kremlin Cup with a knee injury, tournament organisers said on Friday. The American has not played since pulling out injured of last month's China Open. The former world number one has been troubled by various injuries (...) Williams's withdrawal is the latest blow for organisers after Australian Open champion and home favorite Marat Safin withdrew (...).

Figure 1: Two blog posts relevant to the query "tennis."

personal pronouns—i.e., topic independent ones—and the fact that the language usage in the second post is more easily associated with credible information about tennis than the language usage in the first post—i.e., a topic dependent feature.

Our main finding is that topical blog post retrieval can benefit from using credibility indicators in the retrieval process. Both post and blog level indicator groups each show a significant improvement over the baseline. When we combine all features we obtain the best retrieval performance, and this performance is comparable to the best performing TREC 2006 and 2007 Blog track participants. The improvement over the baseline is stable across most topics, although topic shift occurs in a few cases.

The rest of the paper is organized as follows. In Section 2 we provide information on determining credibility; we also relate previous work to the credibility indicators that we consider. Section 3 specifies our retrieval model, a method for incorporating credibility indicators in our retrieval model, and estimations of credibility indicators. Section 4 gives the results of our experiments aimed at assessing the contribution of credibility towards blog post retrieval effectiveness. We conclude in Section 5.

2 Credibility Indicators

In our choice of credibility indicators we use (Rubin and Liddy, 2006)'s work as a reference point. We recall the main points of their framework and relate our indicators to it. We briefly discuss other credibility-related indicators found in the literature.

2.1 Rubin and Liddy (2006)'s work

Rubin and Liddy (2006) proposed a four factor analytical framework for blog-readers' credibility assessment of blog sites, based in part on evidentiality theory (Chafe, 1986), website credibility assessment surveys (Stanford et al., 2002), and Van House (2004)'s observations on blog credibility. The four factors—plus indicators for each of them—are:

1. *blogger's expertise and offline identity disclosure* (*a*: name and geographic location; *b*: credentials; *c*: affiliations; *d*: hyperlinks to others; *e*: stated competencies; *f*: mode of knowing);
2. *blogger's trustworthiness and value system* (*a*: biases; *b*: beliefs; *c*: opinions; *d*: honesty; *e*: preferences; *f*: habits; *g*: slogans)
3. *information quality* (*a*: completeness; *b*: accuracy; *c*: appropriateness; *d*: timeliness; *e*: organization (by categories or chronology); *f*: match to prior expectations; *g*: match to information need); and
4. *appeals and triggers of a personal nature* (*a*: aesthetic appeal; *b*: literary appeal (i.e., writing style); *c*: curiosity trigger; *d*: memory trigger; *e*: personal connection).

2.2 Our credibility indicators

We only consider credibility indicators that avoid making use of the searcher's or blogger's identity (i.e., excluding 1a, 1c, 1e, 1f, 2e from Rubin and Liddy's list), that can be estimated automatically from available test collections only so as to facilitate repeatability of our experiments (ruling out 3e, 4a, 4c, 4d, 4e), that are textual in nature (ruling out 2d), and that can be reliably estimated with state-of-the-art language technology (ruling out 2a, 2b, 2c, 2g). For reasons that we explain below, we also ignore the "hyperlinks to others" indicator (1d).

The indicators that we do consider—1b, 2f, 3a, 3b, 3c, 3d, 3f, 3g, 4b—are organized in two groups,

depending on the information source that we use to estimate them, *post level* and *blog level*, and the former is further subdivided into *topic independent* and *topic dependent*. Table 1 lists the indicators we consider, together with the corresponding Rubin and Liddy indicator(s).

Let us quickly explain our indicators. First, we consider the use of *capitalization* to be an indicator of good writing style, which in turn contributes to a sense of credibility. Second, we identify Western style *emoticons* (e.g., `:-)` and `:-D`) in blog posts, and assume that excessive use indicates a less credible blog post. Third, words written in all caps are considered *shouting* in a web environment; we consider shouting to be indicative for non-credible posts. Fourth, a credible author should be able to write without (a lot of) *spelling* errors; the more spelling errors occur in a blog post, the less credible we consider it to be. Fifth, we assume that credible texts have a reasonable *length*; the text should supply enough information to convince the reader of the author's credibility. Sixth, assuming that much of what goes on in the blogosphere is inspired by events in the news (Mishne and de Rijke, 2006), we believe that, for news related topics, a blog post is more credible if it is published around the time of the triggering news event (*timeliness*). Seventh, our *semantic* indicator also exploits the news-related nature of many blog posts, and "prefers" posts whose language usage is similar to news stories on the topic. Eighth, blogs are a popular place for spammers; *spam* blogs are not considered credible and we want to demote them in the search results. Ninth, *comments* are a notable blog feature: readers of a blog post often have the possibility of leaving a comment for other readers or the author. When people comment on a blog post they apparently find the post worth putting effort in, which can be seen as an indicator of credibility (Mishne and Glance, 2006). Tenth, blogs consist of multiple posts in (reverse) chronological order. The temporal aspect of blogs may indicate credibility: we assume that bloggers with an irregular posting behavior are less credible than bloggers who post *regularly*. And, finally, we consider the topical fluctuation of a blogger's posts. When looking for credible information we would like to retrieve posts from bloggers that have a certain level of (topical) *consistency*: not the fluctuating behavior of a (personal) blogger, but a solid interest.

indicator	**topic dependent?**	**post level/ blog level**	**related Rubin & Liddy indicator**
capitalization	no	post	4b
emoticons	no	post	4b
shouting	no	post	4b
spelling	no	post	4b
post length	no	post	3a
timeliness	yes	post	3d
semantic	yes	post	3b, 3c
spam	no	blog	3b, 3c, 3f, 3g
comments	no	blog	1b
regularity	no	blog	2f
consistency	no	blog	2f

Table 1: Credibility indicators

2.3 Other work

In a web setting, credibility is often couched in terms of authoritativeness and estimated by exploiting the hyperlink structure. Two well-known examples are the PageRank and HITS algorithms (Liu, 2007), that use the link structure in a topic independent or topic dependent way, respectively. Zhou and Croft (2005) propose collection-document distance and signal-to-noise ratio as priors for the indication of quality in web ad hoc retrieval. The idea of using link structure for improving blog post retrieval has been researched, but results do not show improvements. E.g., Mishne (2007a) finds that retrieval performance decreased. This confirms lessons from the TREC web tracks, where participants found no conclusive benefit from the use of link information for ad hoc retrieval tasks (Hawking and Craswell, 2002). Hence, we restrict ourselves to the use of content-based features for blog post retrieval, thus ignoring indicator 1d (hyperlinks to others).

Related to credibility in blogs is the automatic assessment of forum post quality discussed by Weimer et al. (2007). The authors use surface, lexical, syntactic and forum-specific features to classify forum posts as bad posts or good posts. The use of forum-specific features (such as whether or not the post contains HTML, and the fraction of characters that are inside quotes of other posts), gives the highest benefits to the classification. Working in the community question/answering domain, Agichtein et al. (2008) use a content features, as well non-content information available, such as links between items and

explicit quality ratings from members of the community to identify high-quality content.

As we argued above, spam identification may be part of estimating a blog (or blog post's) credibility. Spam identification has been successfully applied in the blogosphere to improve retrieval effectiveness; see, e.g., (Mishne, 2007b; Java et al., 2007).

3 Modeling

In this section we detail the retrieval model that we use, incorporating ranking by relevance and by credibility. We also describe how we estimate the credibility indicators listed in Section 2.

3.1 Baseline retrieval model

We address the baseline retrieval task using a language modeling approach (Croft and Lafferty, 2003), where we rank documents given a query: $p(d|q) = p(d)p(q|d)p(q)^{-1}$. Using Bayes' Theorem we rewrite this, ignoring expressions that do not influence the ranking, obtaining

$$p(d|q) \propto p(d)p(q|d), \tag{1}$$

and, assuming that query terms are independent,

$$p(d|q) \propto p(d) \prod_{t \in q} p(t|\theta_d)^{n(t,q)}, \tag{2}$$

where θ_d is the blog post model, and $n(t, q)$ denotes the number of times term t occurs in query q. To prevent numerical underflows, we perform this computation in the log domain:

$$\log p(d|q) \propto \log p(d) + \sum_{t \in q} n(t, q) \log p(t|\theta_d) \tag{3}$$

In our final formula for ranking posts based on relevance only we substitute $n(t, q)$ by the probability of the term given the query. This allows us to assign different weights to query terms and yields:

$$\log p(d|q) \propto \log p(d) + \sum_{t \in q} p(t|q) \log p(t|\theta_d). \tag{4}$$

For our baseline experiments we assume that all query terms are equally important and set $p(t|q)$ set to be $n(t, q) \cdot |q|^{-1}$. The component $p(d)$ is the topic independent ("prior") probability that the document is relevant; in the baseline model, priors are ignored.

3.2 Incorporating credibility

Next, we extend Eq. 4 by incorporating estimations of the credibility indicators listed in Table 1. Recall that our credibility indicators come in two kinds—post level and blog level—and that the post level indicators can be topic indepedent or topic dependent, while all blog level indicators are topic independent. Now, modeling topic independent indicators is easy—they can simply be incorporated in Eq. 4 as a weighted sum of two priors:

$$p(d) = \lambda \cdot p_{pl}(d) + (1 - \lambda) \cdot p_{bl}(d), \tag{5}$$

where $p_{pl}(d)$ and $p_{bl}(d)$ are the post level and blog level prior probability of d, respectively. The priors p_{pl} and p_{bl} are defined as equally weighted sums:

$$\begin{aligned} p_{pl}(d) &= \textstyle\sum_i \frac{1}{5} \cdot p_i(d) \\ p_{bl}(d) &= \textstyle\sum_j \frac{1}{4} \cdot p_j(d), \end{aligned}$$

where $i \in \{$*capitalization*, *emoticons*, *shouting*, *spelling*, *post_length*$\}$ and $j \in \{$*spam*, *comments*, *regularity*, *consistency*$\}$. Estimations of the priors p_i and p_j are given below; the weighting parameter λ is determined experimentally.

Modeling topic dependent indicators is slighty more involved. Given a query q, we create a query model θ_q that is a mixture of a temporal query model $\theta_{temporal}$ and a semantic query model $\theta_{semantic}$:

$$\begin{aligned} &p(t|\theta_q) = \\ &\quad \mu \cdot p(t|\theta_{temporal}) + (1 - \mu) \cdot p(t|\theta_{semantic}). \end{aligned} \tag{6}$$

The component models $\theta_{temporal}$ and $\theta_{semantic}$ will be estimated below; the parameter μ will be estimated experimentally.

Our final ranking formula, then, is obtained by plugging in Eq. 5 and 6 in Eq. 4:

$$\begin{aligned} \log p(d|q) \propto &\log p(d) \\ &+ \beta \left(\textstyle\sum_t p(t|q) \cdot \log p(t|\theta_d) \right) \\ &+ (1 - \beta) \left(\textstyle\sum_t p(t|\theta_q) \cdot \log p(t|\theta_d) \right). \end{aligned} \tag{7}$$

3.3 Estimating credibility indicators

Next, we specify how each of the credibility indicators is estimated; we do so in two groups: post level and blog level.

3.3.1 Post level credibility indicators

Capitalization We estimate the capitalization prior as follows:

$$p_{capitalization}(d) = n(c, s) \cdot |s|^{-1}, \quad (8)$$

where $n(c, s)$ is the number of sentences starting with a capital and $|s|$ is the number of sentences; we only consider sentences with five or more words.

Emoticons The emoticons prior is estimated as

$$p_{emoticons}(d) = 1 - n(e, d) \cdot |d|^{-1}, \quad (9)$$

where $n(e, d)$ is the number of emoticons in the post and $|d|$ is the length of the post in words.

Shouting We use the following equation to estimate the shouting prior:

$$p_{shouting}(d) = 1 - n(a, d) \cdot |d|^{-1}, \quad (10)$$

where $n(a, d)$ is the number of all caps words in blog post d and $|d|$ is the post length in words.

Spelling The spelling prior is estimated as

$$p_{spelling}(d) = 1 - n(m, d) \cdot |d|^{-1}, \quad (11)$$

where $n(m, d)$ is the number of misspelled (or unknown) words and $|d|$ is the post length in words.

Post length The post length prior is estimated using $|d|$, the post length in words:

$$p_{length}(d) = \log(|d|). \quad (12)$$

Timeliness We estimate timeliness using the time-based language models $\theta_{temporal}$ proposed in (Li and Croft, 2003; Mishne, 2007b). I.e., we use a news corpus from the same period as the blog corpus that we use for evaluation purposes (see Section 4.2). We assign a timeliness score per post based on:

$$p(d|\theta_{temporal}) = k^{-1} \cdot (n(date(d), k) + 1), \quad (13)$$

where k is the number of top results from the initial result list, $date(d)$ is the date associated with document d, and $n(date(d), k)$ is the number of documents in k with the same date as d. For our initial result list we perform retrieval on both the blog and the news corpus and take $k = 50$ for both corpora.

Semantic A semantic query model $\theta_{semantic}$ is obtained using ideas due to Diaz and Metzler (2006). Again, we use a news corpus from the same period as the evaluation blog corpus and estimate $\theta_{semantic}$. We issue the query to the external news corpus, retrieve the top 10 documents and extract the top 10 distinctive terms from these documents. These terms are added to the original query terms to capture the language usage around the topic.

3.3.2 Blog level credibility indicators

Spam filtering To estimate the spaminess of a blog, we take a simple approach. We train an SVM classifier on a labeled splog blog dataset (Kolari et al., 2006) using the top 1500 words for both spam and non-spam blogs as features. For each classified blog d we have a confidence value $s(d)$. If the classifier cannot make a decision ($s(d) = 0$) we set $p_{spam}(d)$ to 0, otherwise we use the following to transform $s(d)$ into a spam prior $p_{spam}(d)$:

$$p_{spam}(d) = \frac{s(d)}{2|s(d)|} + \frac{-1 \cdot s(d)}{2s(d)^2 + 2|s(d)|} + \frac{1}{2}. \quad (14)$$

Comments We estimate the comment prior as

$$p_{comment}(d) = \log(n(r, d)), \quad (15)$$

where $n(r, d)$ is the number of comments on post d.

Regularity To estimate the regularity prior we use

$$p_{regularity}(d) = \log(\sigma_{interval}), \quad (16)$$

where $\sigma_{interval}$ expresses the standard deviation of the temporal intervals between two successive posts.

Topical consistency Here we use an approach similar to query clarity (Cronen-Townsend and Croft, 2002): based on the list of posts from the same blog we compare the topic distribution of blog B to the topic distribution in the collection C and assign a 'clarity' value to B; a score further away from zero indicates a higher topical consistency. We estimate the topical consistency prior as

$$p_{topic}(d) = \log(clarity(d)), \quad (17)$$

where $clarity(d)$ is estimated by

$$clarity(d) = \frac{\sum_w p(w|B) \cdot \log\left(\frac{p(w|B)}{p(w)}\right)}{\sum_w p(w|B)} \quad (18)$$

with $p(w) = \frac{count(w,C)}{|C|}$ and $p(w|B) = \frac{count(w,B)}{|B|}$.

3.3.3 Efficiency

All estimators discussed above can be implemented efficiently: most are document priors and can therefore be calculated offline. The only topic dependent estimators are *timeliness* and *language usage*; both can be implemented efficiently as specific forms of query expansion.

4 Evaluation

In this section we describe the experiments we conducted to answer our research questions about the impact of credibility on blog post retrieval.

4.1 Research questions

Our research revolves around the contribution of credibility to the effectiveness of topical blog post retrieval: what is the contribution of individual indicators, of the post level indicators (topic dependent or independent), of the blog level indicators, and of all indicators combined? And do different topics benefit from different indicators? To answer our research question we compared the performance of the baseline retrieval system (as detailed in Section 3.1) with extensions of the baseline system with a single indicator, a set of indicators, or all indicators.

4.2 Setup

We apply our models to the TREC Blog06 corpus (Macdonald and Ounis, 2006). This corpus has been constructed by monitoring around 100,000 blog feeds for a period of 11 weeks in early 2006, downloading all posts created in this period. For each permalink (HTML page containing one blog post) the feed id is registered. We can use this id to aggregate post level features to the blog level. In our experiments we use only the HTML documents, 3.2M permalinks, which add up to around 88 GB.

The TREC 2006 and 2007 Blog tracks each offer 50 topics and assessments (Ounis et al., 2007; Macdonald et al., 2007). For topical relevancy, assessment was done using a standard two-level scale: the content of the post was judged to be topically relevant or not. The evaluation metrics that we use are standard ones: mean average precision (MAP) and precision@10 (p@10) (Baeza-Yates and Ribeiro-Neto, 1999). For all our retrieval tasks we use the title field (T) of the topic statement as query.

To estimate the timeliness and semantic credibility indicators, we use AQUAINT-2, a set of newswire articles (2.5 GB, about 907K documents) that are roughly contemporaneous with the TREC Blog06 collection (AQUAINT-2, 2007). Articles are in English and come from a variety of sources.

Statistical significance is tested using a two-tailed paired t-test. Significant improvements over the baseline are marked with $^{\triangle}$ ($\alpha = 0.05$) or $^{\blacktriangle}$ ($\alpha = 0.01$). We use $^{\triangledown}$ and $^{\blacktriangledown}$ for a drop in performance (for $\alpha = 0.05$ and $\alpha = 0.01$, respectively).

4.3 Parameter estimation

The models proposed in Section 3.2 contain parameters β, λ and μ. These parameters need to be estimated and, hence, require a training and test set. We use a two-fold parameter estimation process: in the first cycle we estimate the parameters on the TREC 2006 Blog topic set and test these settings on the topics of the TREC 2007 Blog track. The second cycle goes the other way around and trains on the 2007 set, while testing on the 2006 set.

Figure 2 shows the optimum values for λ, β, and μ on the 2006 and the 2007 topic sets for both MAP (bottom lines) and p@10 (top lines). When looking at the MAP scores, the optimal setting for λ is almost identical for the two topic sets: 0.4 for the 2006 set and 0.3 for the 2007 set, and also the optimal setting for β is very similar for both sets: 0.4 for the 2006 set and 0.5 for the 2007 set. As to μ, it is clear that timeliness does not improve the performance over using the semantic feature alone and the optimal setting for μ is therefore 0.0. Both μ and β show similar behavior on p@10 as on MAP, but for λ we see a different trend. If early precision is required, the value of λ should be increased, giving more weight to the topic-independent post level features compared to the blog level features.

4.4 Retrieval performance

Table 2 lists the retrieval results for the baseline, for each of the credibility indicators (on top of the baseline), for four subsets of indicators, and for all indicators combined. The baseline performs similar to the median scores at the TREC 2006 Blog track (MAP: 0.2203; p@10: 0.564) and somewhat below the median MAP score at 2007 Blog track (MAP: 0.3340) but above the median p@10 score: 0.3805.

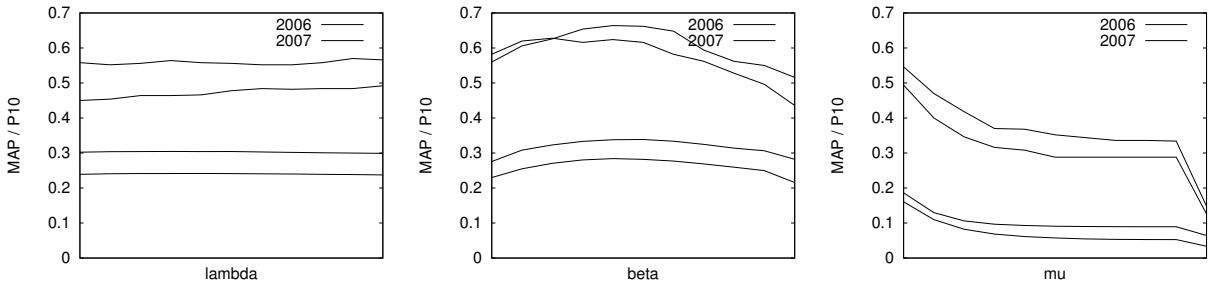

Figure 2: Parameter estimation on the TREC 2006 and 2007 Blog topics. (Left): λ. (Center): β. (Right): μ.

	2006		2007	
	map	p@10	map	p@10
baseline	0.2156	0.4360	0.2820	0.5160
capitalization	0.2155	0.4500	0.2824	0.5160
emoticons	0.2156	0.4360	0.2820	0.5200
shouting	0.2159	0.4320	0.2833	0.5100
spelling	0.2179$^{\triangle}$	0.4480$^{\triangle}$	0.2839$^{\blacktriangle}$	0.5220
post length	0.2502$^{\blacktriangle}$	0.4960$^{\blacktriangle}$	0.3112$^{\blacktriangle}$	0.5700$^{\blacktriangle}$
timeliness	0.1865$^{\blacktriangledown}$	0.4520	0.2660	0.4860
semantic	0.2840$^{\blacktriangle}$	0.6240$^{\blacktriangle}$	0.3379$^{\blacktriangle}$	0.6640$^{\blacktriangle}$
spam filtering	0.2093	0.4700	0.2814	0.5760$^{\blacktriangle}$
comments	0.2497$^{\blacktriangle}$	0.5000$^{\blacktriangle}$	0.3099$^{\blacktriangle}$	0.5600$^{\blacktriangle}$
regularity	0.1658$^{\blacktriangledown}$	0.4940$^{\triangle}$	0.2353$^{\blacktriangledown}$	0.5640$^{\triangle}$
consistency	0.2141$^{\blacktriangledown}$	0.4220	0.2785$^{\triangledown}$	0.5040
post level (topic indep.)	0.2374$^{\blacktriangle}$	0.4920$^{\blacktriangle}$	0.2990$^{\blacktriangle}$	0.5660$^{\blacktriangle}$
post level (topic dep.)	0.2840$^{\blacktriangle}$	0.6240$^{\blacktriangle}$	0.3379$^{\blacktriangle}$	0.6640$^{\blacktriangle}$
post level (all)	0.2911$^{\blacktriangle}$	0.6380$^{\blacktriangle}$	0.3369$^{\blacktriangle}$	0.6620$^{\blacktriangle}$
blog level	0.2391$^{\blacktriangle}$	0.4500	0.3023$^{\blacktriangle}$	0.5580$^{\blacktriangle}$
all	0.3051$^{\blacktriangle}$	0.6880$^{\blacktriangle}$	0.3530$^{\blacktriangle}$	0.6900$^{\blacktriangle}$

Table 2: Retrieval performance on 2006 and 2007 topics, using $\lambda = 0.3$, $\beta = 0.4$, and $\mu = 0.0$.

Some (topic independent) post level indicators hurt the MAP score, while others help (for both years, and both measures). Combined, the topic independent post level indicators perform less well than the use of one of them (*post length*). As to the topic dependent post level indicators, *timeliness* hurts performance on MAP for both years, while the semantic indicator provides significant improvements across the board (resulting in a top 2 score in terms of MAP and a top 5 score in terms of p@10, when compared to the TREC 2006 Blog track participants that only used the T field).

Some of the blog level features hurt more than they help (*regularity*, *consistency*), while the *comments* feature helps, on all measures, and for both years. Combined, the blog level features help less than the use of one of them (*comments*).

As a group, the combined post level features help more than either of the two post level sub groups alone. The blog level features show similar results to the topic-independent post level features, obtaining a significant increase on both MAP and p@10, but lower than the topic-dependent post level features.

The grand combination of all credibility indicators leads to a significant improvement over any of the single indicators and over any of the four subsets considered in Table 2. The MAP score of this run is higher than the best performing run in the TREC 2006 Blog track and has a top 3 performance on p@10; its 2007 performance is just within the top half on both MAP and p@10.

4.5 Analysis

Next we examine the differences in average precision (per topic) between the baseline and subsets of indicators (post and blog level) and the grand combination. We limit ourselves to an analysis of the MAP scores. Figure 3 displays the per topic average precision scores, where topics are sorted by absolute gain of the grand combination over the baseline.

In 2006, 7 (out of 50) topics were negatively affected by the use of credibility indicators; in 2007, 15 (out of 50) were negatively affected. Table 3 lists the topics that displayed extreme behavior (in terms of relative performance gain or drop in AP score). While the extreme drops for both years are in the same range, the gains for 2006 are more extreme than for 2007.

The topic that is hurt most (in absolute terms) by the credibility indicators is the 2007 topic 910: *aperto network* (AP -0.2781). The semantic indicator is to blame for this decrease is: the terms included in the expanded query shift the topic from a wireless broadband provider to television networks.

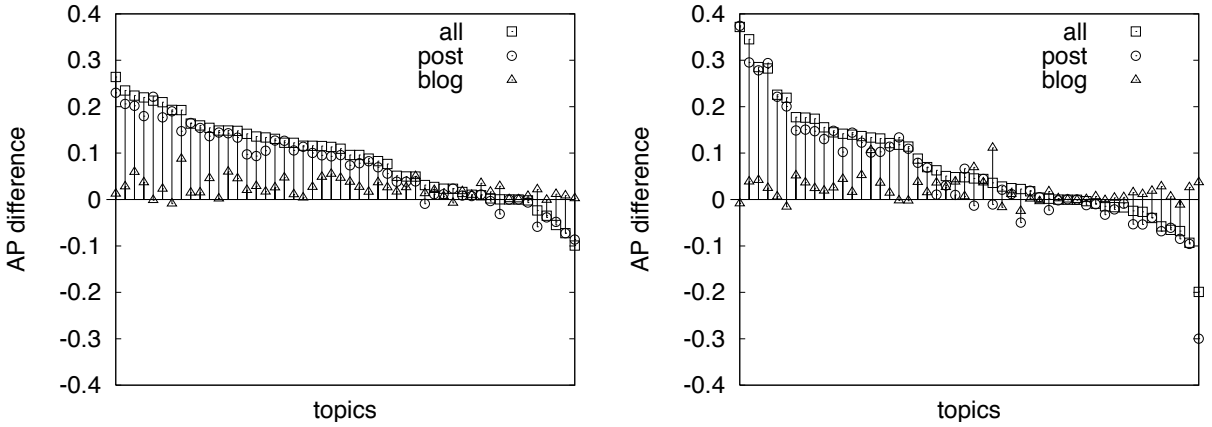

Figure 3: Per-topic AP differences between baseline run and runs with blog level features (triangles), post level features (circles) and all feature (squares) on the 2006 (left) en 2007 (right) topics.

Table 3: Extreme performance gains/drops of the grand combination over the baseline (MAP).

2006		
id	topic	% gain/loss
900	mcdonalds	+525.9%
866	foods	+446.2%
865	basque	+308.6%
862	blackberry	-21.5%
870	barry bonds	-35.2%
898	business intelligence resources	-78.8%
2007		
id	topic	% gain/loss
923	challenger	+162.1%
926	hawthorne heights	+160.7%
945	bolivia	+125.5%
943	censure	-49.4%
928	big love	-80.0%
904	alterman	-84.2%

Topics that gain most (in absolute terms) are 947 (*sasha cohen*; AP +0.3809) and 923 (*challenger*; AP +0.3622) from the 2007 topic set.

Finally, the combination of all credibility indicators hurts 7 (2006) plus 15 (2007) equals 22 topics; for the post level indicators get a performance drop in AP for 28 topics (10 plus 18, respectively) and for the blog level indicators we get a drop for 15 topics (4 plus 11, respectively). Hence, the combination of all indicators strikes a good balance between overall performance gain and per topic risk.

5 Conclusions

We provided efficient estimations for 11 credibility indicators and assessed their impact on topical blog post retrieval, on top of a content-based retrieval baseline. We compared the contribution of these indicators, both individually and in groups, and found that (combined) they have a significant positive impact on topical blog post retrieval effectiveness. Certain single indicators, like *post length* and *comments*, make good credibility indicators on their own; the best performing credibility indicator group consists of topic dependent post level ones. Other future work concerns indicator selection: instead of taking all indicators on board, consider selected indicators only, in a topic dependent fashion.

Our choice of credibility indicators was based on a framework proposed by Rubin and Liddy (2006): the estimators we used are natural implementations of the selected indicators, but by no means the only possible ones. In future work we intend to extend the set of indicators considered so as to include, e.g., stated competencies (1e), by harvesting and analyzing bloggers' profiles, and to extend the set of estimators for indicators that we already consider such as reading level measures (e.g., Flesch-Kincaid) for the literary appeal indicator (4b).

Acknowledgments

We would like to thank our reviewers for their feedback. Both authors were supported by the E.U. IST programme of the 6th FP for RTD under project MultiMATCH contract IST-033104. De Rijke was also supported by NWO under project numbers 017.001.190, 220-80-001, 264-70-050, 354-20-005, 600.065.120, 612-13-001, 612.000.106, 612.066.-302, 612.069.006, 640.001.501, and 640.002.501.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. (2008). Finding high-quality content in social media. In *WSDM '08*.

AQUAINT-2 (2007). URL: `http://trec.nist.gov/data/qa/2007_qadata/qa.07.guidelines.html#documents`.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Chafe, W. (1986). Evidentiality in English conversion and academic writing. In Chaf, W. and Nichols, J., editors, *Evidentiality: The Linguistic Coding of Epistemology*, volume 20, pages 261–273. Ablex Publishing Corporation.

Croft, W. B. and Lafferty, J., editors (2003). *Language Modeling for Information Retrieval*. Kluwer.

Cronen-Townsend, S. and Croft, W. (2002). Quantifying query ambiguity. In *Proceedings of Human Language Technology 2002*, pages 94–98.

Diaz, F. and Metzler, D. (2006). Improving the estimation of relevance models using large external corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York. ACM Press.

Hawking, D. and Craswell, N. (2002). Overview of the TREC-2001 web track. In *The Tenth Text Retrieval Conferences (TREC-2001)*, pages 25–31.

Java, A., Kolari, P., Finin, T., Joshi, A., and Martineau, J. (2007). The blogvox opinion retrieval system. In *The Fifteenth Text REtrieval Conference (TREC 2006)*.

Kolari, P., Finin, T., Java, A., and Joshi, A. (2006). Splog blog dataset. URL: `http://ebiquity.umbc.edu/resource/html/id/212/Splog-Blog-Dataset`.

Li, X. and Croft, W. (2003). Time-based language models. In *Proceedings of the 12th International Conference on Information and Knowledge Managment (CIKM)*, pages 469–475.

Liu, B. (2007). *Web Data Mining*. Springer-Verlag, Heidelberg.

Macdonald, C. and Ounis, I. (2006). The trec blogs06 collection: Creating and analyzing a blog test collection. Technical Report TR-2006-224, Department of Computer Science, University of Glasgow.

Macdonald, C., Ounis, I., and Soboroff, I. (2007). Overview of the trec 2007 blog track. In *TREC 2007 Working Notes*, pages 31–43.

Metzger, M. (2007). Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journl of the American Society for Information Science and Technology*, 58(13):2078–2091.

Mishne, G. (2007a). *Applied Text Analytics for Blogs*. PhD thesis, University of Amsterdam, Amsterdam.

Mishne, G. (2007b). Using blog properties to improve retrieval. In *Proceedings of ICWSM 2007*.

Mishne, G. and de Rijke, M. (2006). A study of blog search. In Lalmas, M., MacFarlane, A., Rüger, S., Tombros, A., Tsikrika, T., and Yavlinsky, A., editors, *Advances in Information Retrieval: Proceedings 28th European Conference on IR Research (ECIR 2006)*, volume 3936 of *LNCS*, pages 289–301. Springer.

Mishne, G. and Glance, N. (2006). Leave a reply: An analysis of weblog comments. In *Proceedings of WWW 2006*.

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., and Soboroff, I. (2007). Overview of the trec-2006 blog track. In *The Fifteenth Text REtrieval Conference (TREC 2006) Proceedings*.

Rubin, V. and Liddy, E. (2006). Assessing credibility of weblogs. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs (CAAW)*.

Stanford, J., Tauber, E., Fogg, B., and Marable, L. (2002). Experts vs online consumers: A comparative credibility study of health and finance web sites. URL: `http://www.consumerwebwatch.org/news/report3_credibilityresearch/slicedbread.pdf`.

Van House, N. (2004). Weblogs: Credibility and collaboration in an online world. URL: `people.ischool.berkeley.edu/~vanhouse/Van\%20House\%20trust\%20workshop.pdf`.

Weimer, M., Gurevych, I., and Mehlhauser, M. (2007). Automatically assessing the post quality in online discussions on software. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 125–128.

Zhou, Y. and Croft, W. B. (2005). Document quality models for web ad hoc retrieval. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 331–332.