Credibility-inspired Ranking for Blog Post Retrieval

Weerkamp, W.; de Rijke, M.

DOI
10.1007/s10791-011-9182-8
Publication date
2012
Document Version
Final published version
Published in
Information Retrieval

Inf Retrieval (2012) 15:243–277
DOI 10.1007/s10791-011-9182-8

INFORMATION RETRIEVAL FOR SOCIAL MEDIA

Credibility-inspired ranking for blog post retrieval

Wouter Weerkamp · Maarten de Rijke

Received: 9 April 2011 / Accepted: 23 December 2011 / Published online: 11 February 2012

Abstract Credibility of information refers to its believability or the believability of its sources. We explore the impact of credibility-inspired indicators on the task of blog post retrieval, following the intuition that more credible blog posts are preferred by searchers. Based on a previously introduced credibility framework for blogs, we define several credibility indicators, and divide them into post-level (e.g., spelling, timeliness, document length) and blog-level (e.g., regularity, expertise, comments) indicators. The retrieval task at hand is precision-oriented, and we hypothesize that the use of credibility-inspired indicators will positively impact precision. We propose to use ideas from the credibility framework in a reranking approach to the blog post retrieval problem: We introduce two simple ways of reranking the top n of an initial run. The first approach, Credibility-inspired reranking, simply reranks the top n of a baseline based on the credibility-inspired score. The second approach, Combined reranking, multiplies the credibility-inspired score of the top n results by their retrieval score, and reranks based on this score. Results show that Credibility-inspired reranking leads to larger improvements over the baseline than Combined reranking, but both approaches are capable of improving over an already strong baseline. For Credibility-inspired reranking the best performance is achieved using a combination of all post-level indicators. Combined reranking works best using the post-level indicators combined with comments and pronouns. The blog-level indicators expertise, regularity, and coherence do not contribute positively to the performance, although analysis shows that they can be useful for certain topics. Additional analysis shows that a relative small value of n (15–25) leads to the best results, and that posts that

Initial ideas in this paper have been published at ACL 2008 (Weerkamp and de Rijke 2008). This paper builds on these ideas, but introduces new indicators and new ways of incorporating credibility-inspired indicators. Since the baselines used in this paper are also different, results are not comparable between the two papers.

W. Weerkamp (✉) · M. de Rijke
ISLA, University of Amsterdam, Amsterdam, The Netherlands
e-mail: w.weerkamp@uva.nl

M. de Rijke
e-mail: derijke@uva.nl

move up the ranking due to the integration of reranking based on credibility-inspired indicators do indeed appear to be more credible than the ones that go down.

Keywords Credibility · Blog post retrieval · Reranking

1 Introduction

In recent years there has been an ever growing usage of social media, web-based platforms that allow the easy creation and exchange of user generated content. Social media can be centered around video (e.g., Youtube, Vimeo), audio (e.g., Last.fm, MySpace), pictures (e.g., Flickr, Picassa), other media types (bookmarks, books, etc.), and people (e.g., Facebook, Friendster). However, one of the most popular media types is still text. Textual social media come in various forms, each with its own characteristics and users: examples are (micro)blogs, forums, mailing lists, reviews, and comments. In this paper we look at one particular type of social media, blogs. Blogging platforms allow people (bloggers) to write diary entries (blog posts) about topics of their choice.

The growing amount of social media content available online creates new challenges for the information retrieval (IR) community, in terms of search and analysis tasks for this type of content (Weerkamp 2011). One of the main challenges lies in the fact that creators of social media content, the bloggers, are given a large degree of freedom: operating without top-down editorial rules and editors, they produce blog posts of hugely varying quality. Some of the posts are edited, news article-like, whereas others are of very low quality. The quality of a blog post may have an impact on its suitability of being returned in response to a query.

Although some approaches to blog post retrieval (finding blog posts that are relevant to a given topic) use indirect quality measures like elaborate spam filtering (Java et al. 2007) or counting inlinks (Mishne 2007b), few systems turn the *credibility* (Metzger 2007) of blog posts into an aspect that can benefit the retrieval process. Our hypothesis is that we can use credibility-inspired indicators to improve topical blog post retrieval. In this paper we explore the impact of these credibility-inspired indicators on the task of blog post retrieval.

To make matters concrete, consider Fig. 1: both (blog) 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, the large number of exclamation marks and personal pronouns, 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.

Another case in which credibility plays an important role is so-called online reputation management (Klewes and Wreschniok 2009): companies monitor online activities, for example on blogs and social networking sites, to find mentions of themselves or of their products and services. The goal here is to identify potentially harmful messages, and try to respond fast and adequately to these. While monitoring a company's reputation, one can come across posts like the ones in Fig. 2: The first post is an extensive and well-written description of someone's encounter with company X's help desk. The second is a short, apparently angry shout by a frustrated customer. Company X might decide to act fast after spotting the first post, given that this post sounds reliable, and other people reading it might believe it. The second post is useful for overall statistics on reputation, but is not as important as an individual post.

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 (...).

Fig. 1 Two blog posts relevant to the query "tennis"

Post 3
Yesterday I tried to contact company X to ask a question regarding their service Y. After waiting for at least 30 minutes, the woman "helping" me didn't know what I was talking about. (...) I guess I won't be trying to contact them ever again, I should probably switch to company Z instead.

Post 4
Aarrggghhh, u got 2be joking... I HATE X!!!

Fig. 2 Two blog posts about "Company X"

Similarly, when looking for information on company X, searchers might be more interested in reading the first post than the second. The first will give them insight in what particular service of this company is not as it should be; the second post does not contain much information besides conveying an opinion.

The idea of considering 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). We discuss the indicators in Sect. 3.

Although the Rubin and Liddy (2006) framework is not the only available credibility framework, it is the only framework specifically designed for the blogosphere. Other credibility assessments in social media, like Weimer et al. (2007)'s assessment of forum posts and Agichtein et al. (2008)'s quality detection in cQA, have the advantage that they already identified measurable indicators and have tested the performance of these indicators, but these "frameworks" are specifically designed for other social media platforms. This results in a large group of indicators that do not necessarily apply to our (blog) setting, like content ratings ("thumbs up"), user ratings, and inclusion of HTML code, signatures, and quotes in posts. The indicators proposed by Rubin and Liddy (2006) are not (yet) instantiated and give us the freedom to find appropriate ways of measuring these indicators.

In this paper, we instantiate Rubin and Liddy (2006)'s indicators in a concrete manner and test their impact on blog post retrieval effectiveness. Specifically, 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 collection at hand (see Sect. 5) and that do not need additional resources such as bloggers' profiles, that may be hard to obtain for technical or legal reasons. We identify two groups of indicators: (1) blog-level, and (2) post-level indicators. The former group refers to the blog as a whole, that is, to the blogger, and the latter group deals only with characteristics of the post at hand. Blog post retrieval is a precision-oriented task, similar to web search (Manning et al. 2008, Chapter 19). Taking credibility-inspired indicators into account in the retrieval process aims at enhancing precision; there is no obvious reason why these indicators should or could improve recall.

Note that we do not try to explicitly measure the credibility of posts. Although this would be a very interesting and challenging task, we currently have no ways of evaluating the performance on such a task. Rather, we take ideas from the credibility framework and propose a set of credibility-inspired indicators that we put to use on the task of blog post retrieval.

We ask the following research questions:

1. Given the credibility framework developed by Rubin and Liddy (2006), which indicators can we measure from the text of blog posts?
2. Can we incorporate credibility-inspired indicators in the retrieval process, keeping in mind the precision-oriented nature of the task? We try two methods: (i) "Credibility-inspired reranking" based on credibility-inspired scores and (ii) "Combined reranking" based both on credibility-inspired scores and retrieval scores.
3. Can individual credibility-inspired indicators improve precision over a strong baseline?
4. Can we improve performance (further) by combining indicators in blog and post-level groups? And by combining them all?

In our extensive analysis we discuss five issues that were raised during the experiments:

1. What is the performance of our (simple) spam classification system?
2. Given the reranking approaches we take, how do these actually change the rankings of blog posts?
3. Which specific posts are helped or hurt by the credibility-inspired indicators?
4. What is the impact on performance of the number of results we use in reranking?
5. Do we observe differences between topics with regard to the performance of credibility-inspired indicators?
6. Which of the credibility-inspired indicators have most influence on retrieval performance and why is this?

Our main findings are that reranking the top results based on credibility-inspired scores is beneficial for precision. Especially indicators on the post level contribute to a great extent to this improvement. We can choose for a more radical reranking approach, leading to high gains and losses, or a smoothed version, leading to more stable results.

In Sect. 2 we discuss related work. We follow in Sect. 3 with the introduction of the credibility framework. We define our credibility-inspired indicators in Sect. 4, and describe the experimental setup for testing their impact on retrieval effectiveness in Sect. 5. We discuss the results of our two methods for incorporating credibility-inspired indicators in Sect. 6 and analyze them in detail in Sect. 7. Finally, we draw conclusions in Sect. 8.

2 Related work

Related work comes in two kinds. First, we briefly introduce work related to credibility assessment in web settings. Next, we zoom in on social media and credibility. Then, the next section introduces the credibility framework by Rubin and Liddy (2006) that we use as basis for our work.

2.1 Credibility on the web

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. The idea behind these algorithms is that more pages linking to a certain document is an indication of this page being more authoritative. In calculating the authoritativeness for a page, the authoritativeness of pages linking to it is taken into account.

The idea of using link structure for improving blog post retrieval has been studied, but results do not show improvements, e.g., Mishne (2007b) finds that retrieval performance decreased, probably because linking in blogs indicates a social relation rather than a vote of authoritativeness. 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). And although some work suggests that social links can be useful in quality prediction (Lu et al. 2010), this mostly works in (dense) social networks. The blog data at hand contains too little social linkage to show this.

Mandl (2006) tries to determine the quality of web pages using a machine learning approach and uses this automatic assessment in a web search engine; features are mainly extracted from the HTML code and DOM tree.

2.2 Credibility in social media

Credibility-related work in social media comes in various forms, and is applied to different platforms. Weimer et al. (2007) discuss the automatic assessment of forum post quality; they 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 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. In the same domain, Su et al. (2010) try to detect text trustworthiness by incorporating evidentiality (e.g., "I'm *certain* of this") in their feature set.

To allow for better presentation of online reviews to users, O'Mahony and Smyth (2009) try to determine the helpfulness of reviews. Their features are divided in reputation features, content features, social features, and sentiment features. Follow-up work also includes readability features (O'Mahony and Smyth 2010).

For blogs, most work related to credibility is aimed at trying to identify blogs worth following. Sriphaew et al. (2008) try to identify "cool blogs," i.e., blogs that are worth exploring. Their approach follows a combination of credibility-like features with topic consistency, as used in blog feed search (Macdonald et al. 2008b). Similar work is done by

Chen and Ohta (2010), who try to filter blog posts using topic concentration and topic variety. One of our indicators (see Sect. 4) is post length, which was further explored by Hearst and Dumais (2009). They found that there is a correlation between the length of posts in a blog and the popularity of that blog. Mishne and de Rijke (2006)'s observation that bloggers often report on news events is the basis for the credibility assessment in Juffinger et al. (2009). The authors compare blog posts to news articles about the same topic, and assign a credibility level based on the similarity between the two. We use a similar technique, but acknowledge that not all blog posts are about news events, hence the need for other indicators. Spam identification may be part of estimating credibility, not only for blogs (or blog posts), but also for other (web) documents. Spam identification has been successfully applied in the blogosphere to improve retrieval effectiveness, for example by Java et al. (2007) and Mishne (2007a).

Recently, credibility-inspired indicators have been successfully applied to post finding in a specific type of blog environment: microblogs (Massoudi et al. 2011). Besides translating indicators to the new environment, the authors also introduced platform-specific indicators like followers, retweets, and recency. For the task of exploring trending topics on Twitter, Castillo et al. (2011) use a similar set of indicators to assess credibility of tweets, and use human assessments to test their approach.

Research into credibility of content is not restricted to textual content. Tsagkias et al. (2010) try to establish the credibility of a particular type of audio: podcasts. They show that, besides podcast-wide metadata (e.g., podcast logo, description length), episode data also plays an important role in determining credibility. We use a similar notion by combining blog level and post level indicators in our work. Finally, Diakopoulos and Essa (2010) explore credibility in video, mainly through the use of smart interfaces and knowledge sharing.

3 Credibility framework

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. 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 listed below.

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

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

In our decision which indicators to include in our experiments, we followed the following steps. For each, we indicate which of the credibility indicators from Rubin and Liddy (2006)'s framework are excluded.

A. We do not use credibility indicators that make use of the searcher's or blogger's identity (excluding 1a, 1c, 1e, 2e);
B. We include indicators that can be estimated automatically from available test collections only so as to facilitate repeatability of our experiments (excluding 3e, 4a, 4c, 4d, 4e);
C. We only select indicators that can be reliably estimated with state-of-the-art language technology (excluding 2b, 2c, 2d, 2g).
D. Finally, given the findings by Mishne (2007b), we ignore the "hyperlinks to others" indicator (1d).

Of the 11 indicators that we do consider—1b, 1f, 2a, 2f, 3a, 3b, 3c, 3d, 3f, 3g, 4b—one is part of the baseline retrieval system (3f), and does not require an indicator. The others are organized in two groups, depending on the information source that we use to estimate them: *post level* and *blog(ger) level*. The former depends solely on information contained in an individual blog post, and ignores the blog it belongs to. The latter aggregates or averages information from posts to the blog level; these indicator values are therefore equal for all posts in the same blog.

In the next section we explore the 10 selected indicators from Rubin and Liddy (2006)'s credibility framework and introduce ways of estimating these indicators so that they can be applied to the task at hand: blog post retrieval.

4 Credibility-inspired indicators

In this section we introduce our credibility-inspired indicators, explain how they are related to the work by Rubin and Liddy (2006) that was described in the previous section, and

Table 1 Our credibility-inspired indicators and their origins in Rubin and Liddy (2006)

Blog-level indicator	Rubin and Liddy (2006)	Post-level indicator	Rubin and Liddy (2006)
Comments	Credentials	Post length	Completeness
Expertise	Mode of knowing	Semantics	Accuracy/appropriateness
Regularity	Habits	Timeliness	Timeliness
Consistency	Habits	Capitalization	Literary appeal
Spamminess	Information quality	Emoticons	Literary appeal
Pronouns	Biases	Shouting	Literary appeal
		Spelling	Literary appeal
		Punctuation	Literary appeal

offer ways of estimating the indicators. Table 1 summarizes this section, and lists our credibility-inspired indicators and their originating counterpart.

Next, we specify how each of the credibility-inspired indicators is estimated, and briefly discuss why and how these indicators address the issue of credibility. We start with the eight post-level indicators (Sect. 4.1) and conclude with the six blog-level indicators (Sect. 4.2).

4.1 Post-level indicators

As mentioned previously, post-level indicators make use of information contained within individual posts. We go through the indicators capitalization, emoticons, shouting, spelling, punctuation, post length, timeliness, and semantics.

4.1.1 Capitalization

We estimate the capitalization score as follows:

$$S_{capitalization}(post) = \frac{n(caps, s_{post})}{|s_{post}|}, \tag{1}$$

where $n(caps, s_{post})$ is the number of sentences in post *post*starting with a capital and $|s_{post}|$ is the number of sentences in the post; we only consider sentences with five or more words. We consider the use of capitalization to be an indicator of good writing style, which in turn contributes to a sense of credibility.

4.1.2 Emoticons

The emoticons score is estimated as

$$S_{emoticons}(post) = 1 - \left(\frac{n(emo, post)}{|post|}\right), \tag{2}$$

where $n(emo, post)$ is the number of emoticons in the post and $|post|$ is the length of the post in words. We identify Western style emoticons (e.g., :-) and :-D) in blog posts, and assume that excessive use indicates a less credible blog post.

4.1.3 Shouting

We use the following equation to estimate the shouting score:

$$S_{shouting}(post) = 1 - \left(\frac{n(shout, post)}{|post|}\right), \tag{3}$$

where $n(shout, post)$ is the number of all caps words in blog post *post* and $|post|$ is the post length in words. Words written in all caps are considered shouting in a web environment; we consider shouting to be indicative for non-credible posts. Note that nowadays the use of repeated characters could also be considered shouting, but that we did not try to detect this notion of shouting.

4.1.4 Spelling

The spelling score is estimated as

$$S_{spelling}(post) = 1 - \left(\frac{n(error, post)}{|post|}\right), \tag{4}$$

where $n(error, post)$ is the number of misspelled or unknown words (with more than 4 characters) in post *post* and $|post|$ is the post length in words. 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.

4.1.5 Punctuation

The punctuation score is calculated as follows:

$$S_{punctuation}(post) = 1 - \left(\frac{n(punc, post)}{|post|}\right), \tag{5}$$

where $n(punc, post)$ is the number of repetitive occurrences of dots, question marks, or exclamation marks (e.g., "look at this!!!", "wel…", or "can you believe it??") and $|post|$ is the post length in words. If $n(punc, post) \cdot |post|^{-1}$ is larger than 1, we set $S_{punctuation}(post) = 0$. We assume that excessive use of repeated punctuation marks is an indication of non-credible posts.

4.1.6 Post length

The post length score is estimated using $|post|$, the post length in words:

$$S_{length}(post) = \log(|post|). \tag{6}$$

We assume that credible texts have a reasonable length; the text should supply enough information to convince the reader of the author's credibility, and it is an indication of "completeness."

4.1.7 Timeliness

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

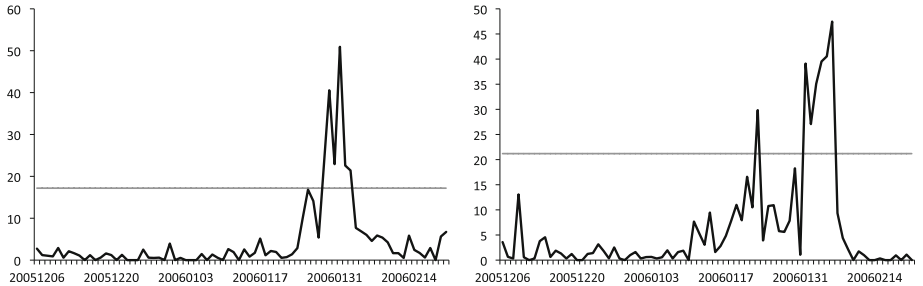

Fig. 3 Peaks in news articles for (*Left*) topic 853, *State of the Union*, which was held on January 31, 2006. (*Right*) topic 882, *Seahawks*, an American football team that won the NFC on January 22, 2006, and played the Super Bowl on February 5, 2006

credible if it is published around the time of the triggering news event: it is timely. Bloggers that take (much) longer to respond to news events are considered less timely. To estimate timeliness, we first identify peaks for a topic in a collection of news articles, by summing over the retrieval scores for each date in the the top 500 results, and taking dates with a value higher than twice the standard deviation to be "peak dates". Two example topics and their peaks are given in Fig. 3.

Having identified peaks for certain topics, we take the timeliness to be the difference in days between the peak date and the day of the post. More formally:

$$S_{timeliness}(post, Q) = \begin{cases} e^{-(|\tau_{post} - \tau_{peak_Q}|)} & \text{if } \tau_{post} - \tau_{peak_Q} > -2 \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

Here, τ_{peak_Q} is the date of the peak (in case the peak spans several days, it is the date closest to the post date), and τ_{post} is the post date. The difference between the dates is calculated in days.

4.1.8 Semantics

For news-related topics, we are looking for posts that "mimic" the semantics of credible sources, like actual news articles. For this, we use a query expansion approach, based on previous work (Diaz and Metzler 2006; Weerkamp et al. 2009). We query the same news collection as before for the topics, and select the top 10 retrieved articles. From these articles we select the 10 most important terms, using Lavrenko and Croft (2001)'s relevance model 2. The selected terms, $\theta_{semantic,Q}$, represent credible semantics for the given topic, and we use these terms as query to score blog posts on the semantics indicator. Table 2 shows the extracted credible terms for three example topics.

4.1.9 Text quality

To limit the number of experiments to run, we combine the following indicators into one *text quality* indicator: spelling, emoticons, capitalization, shouting, and punctuation. To combine these indicators, we first normalize each individual indicator using min-max normalization (Lee 1995). Then, we take the average value over all these indicators to be the text quality indicator.

Table 2 Terms indicating credible semantics for three topics: *Macbook Pro* deals with laptops by Apple; *Cheney hunting* discusses a hunting accident involving vice-president Cheney and his friend Whittington; *David Irving* is an Austrian historian on trial for denying the Holocaust

Topic 856: *Macbook Pro*	Topic 867: *Cheney hunting*	Topic 1042: *David Irving*
Macbook	Cheney	Irving
Intel	Whittington	David
Apple	President	Holocaust
Computer	Accidentally	Court
Start	Hunting	British
Pro	Shot	Austian
Chip	Attack	Monday
Shipping	Heart	Urgent
Notebook	Doctors	Historian
Laptop	Minor	Prison

4.2 Blog-level indicators

Blog-level indicators say something about the blog as a whole, or about the blogger who wrote the posts. Most indicators aggregate information from individual posts to the blog level, and they all lead to posts from the same blog having equal scores. Here, we go through the indicators spamminess, comments, regularity, consistency, pronouns, and expertise.

4.2.1 Spamminess

To estimate the spamminess of a blog, we take a simple approach. First, we observe that blogs are either completely spam ("splogs") or not (i.e., there are no blogs with half of the posts spam and half of them non-spam), and this is why we consider this indicator on the blog level. We train an SVM classifier on a labeled splog blog dataset (Kolari et al. 2006) using the top 1,500 words for both spam and non-spam blogs as features. We then apply the trained classifier to our set of blog posts, and assign a spam or no-spam label to each post. We calculate the ratio of spam posts in each blog, and use this ratio as indication of spamminess for the full blog.

$$S_{spam}(post) = \frac{n(post_{spam}, blog)}{|blog|}, \tag{8}$$

where $n(post_{spam}, blog)$ is the number of spam posts in the blog, and $|blog|$ is the size of the blog in number of posts. Splogs are not considered credible and we want to demote them in or filter them from the search results. Although the list of splogs for our test collection (see Sect. 5.1) is available, we do not use it in any way in this paper, ensuring our results are still comparable to previously published results. Future work could look at the performance of our spam classification.

4.2.2 Comments

We estimate the comment score as

$$S_{comment}(post) = \log\left(\frac{\sum_{post \in blog} n(comment, post)}{|blog|} + 1\right), \quad (9)$$

where $n(comment, post)$ is the number of comments on post *post*, and $|blog|$ is the size of the blog in number of posts. 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).

4.2.3 Regularity

To estimate the regularity score we use

$$S_{regularity}(post) = \log(\sigma_{interval,blog}), \quad (10)$$

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

4.2.4 Topical consistency

We take into consideration 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. The coherence score indicator (He et al. 2009) is a relatively cheap, topic-independent way of estimating this. Given a set of blog posts from a blog, $blog = \{post_i\}_{i=1}^{M}$, which is drawn from a background collection C, i.e., $blog \subseteq C$ (i.e., the blogosphere), the coherence score is defined as the proportion of "coherent" pairs of blog posts with respect to the total number of post pairs within *blog*. The criterion of being a "coherent" pair is that the similarity between the two posts in the pair should meet or exceed a given threshold. Formally, the coherence (Co) of a blog *blog* is defined as

$$Co(blog) = \frac{\sum_{i \neq j \in \{1,\ldots,M\}} \delta(post_i, post_j)}{\frac{1}{2}M(M-1)}, \quad (11)$$

where $\delta(post_i, post_j)$ is 1 if posts are similar and 0 otherwise. We use cosine similarity to determine the similarity between two blog posts. More details on the coherence score can be found in (He et al. 2008, He et al. 2009).

4.2.5 Pronouns

We estimate the pronouns score as follows

$$S_{pronouns}(post) = 1 - \left(\frac{\sum_{post \in blog} \frac{n(pron,post)}{|post|}}{|blog|}\right), \quad (12)$$

where $n(pron, post)$ is the number of first person pronouns (I, me, mine, we, us, ...) in post *post*, $|post|$ is the size of the post in words, and $|blog|$ is the size of the blog in number of posts. First person pronouns express a bias towards ones own interpretation, and we feel

this could harm the credibility of a blog (post). Note that we use simple string matching for this indicator and that this might lead to an overestimation for some pronouns (e.g., "mine" can be used as noun and verb as well). We believe, however, that this is only a marginal issue and should not influence the results of this indicator.

4.2.6 Expertise

To estimate a blogger's expertise for a given topic, we use the approach described in Weerkamp et al. (2011). We look at the posts written by a blogger, and try to estimate to what extent the given topic is central to the blog. Blogs that are most likely to be relevant to this query are retrieved, and we assign posts in those blogs a higher score on the expertise indicator. As an example, consider topic 856, *macbook pro*: the top retrieved blogs are (1) *MacBook Garage*, (2) *Enterprise Mac*, and (3) *tech ronin*. The first two are very Apple/Mac oriented, and the third result is more general technology-oriented, but with an interest in Macs. We consider posts from these blogs, blogs with a *recurring interest* in the topic, to be more credible than posts from blogs mentioning the topic only occasionally.

$$S_{expertise}(post, Q) = P(blog|Q), \tag{13}$$

where $P(blog|Q)$ is the retrieval score for blog *blog* on query Q as given by the Blogger model from Weerkamp et al. (2011).

On top of the individual credibility indicators, we report on the performance of combinations of indicators. We combine indicators into our two levels (post and blog level) and into a full combination, using these steps: (1) normalize indicator scores using min-max normalization (Lee 1995) and (2) average over the indicators belonging to the combination at hand (post level, blog level, or all).

We already introduced the difference between post-level and blog-level indicators, but there is one more dimension on which we can seperate indicators: whether or not the indicator depends on the topic. Most of the indicators get their score independent of the topic (e.g., spelling errors, capitalization), however, three indicators do depend on the topic: semantics, timeliness, and expertise. To summarize this section, Table 3 shows all our indicators and their characteristics.

Table 3 Our credibility-inspired indicators and their characteristics

	Topic independent	Topic dependent
Post level	Post length	
	Spelling	
	Shouting	Semantics
	Emoticons	Timeliness
	Capitalization	
	Punctuation	
Blog level	Regularity	
	Comments	
	Coherence	Expertise
	Spamminess	
	Pronouns	

Table 4 Collection statistics before and after preprocessing

Period	12/06/2005–02/21/2006
Original data	
Number of blog posts	3,215,171
Number of blogs	100,649
- of which splogs	17,969
After preprocessing	
Number of blog posts	2,574,356
Number of blogs	76,358
Average post length	506

5 Experimental setup

This section describes the task and collection we use to test our credibility-inspired indicators (Sect. 5.1), the general retrieval framework (Sect. 5.2), and the evaluation metrics (Sect. 5.3).

5.1 Task and collection

We apply our credibility-inspired indicators to the task of retrieving topically relevant blog posts. This task ran at the Text REtrieval Conference (TREC), as part of the blog track, in 2006–2008 (Macdonald et al. 2008a; Ounis et al. 2006, 2009). Given a set of blog posts and a query, we are asked to return relevant blog posts for that query. We apply our model and indicators 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, and ignore syndicated (RSS) data. We perform two preprocessing steps: (1) keep long sentences (Hofmann and Weerkamp 2008), and (2) apply language identification using TextCat,[1] to select English posts. The collection statistics are displayed in Table 4.

The TREC 2006, 2007, and 2008 Blog tracks each offer 50 topics and assessments, offering us 150 topics in total. 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. For all our retrieval tasks we only use the title field (T) of the topic statement as query; this boils down to the use of keyword queries. Table 5 lists several statistics of the queries in our test collection. We see that for 2006 more posts were assessed than for 2007 and 2008, which leads to more relevant posts per query. As to the number of terms per query, we see that 2008 queries are, on average, quite a bit longer than the 2006 and 2007 queries.

To estimate the semantics and timeliness credibility indicators, we need a collection of news papers. Here, we use AQUAINT-2, a set of about 907,000 newswire articles (AQUAINT-2. Guidelines 2007) from six different news sources. Of these articles, 135,763 are contemporaneous with the TREC Blog06 collection, and we use only this subset in our experiments. All news articles are written in English.

[1] http://odur.let.rug.nl/~vannoord/TextCat/.

Table 5 Query statistics for 2006, 2007, and 2008

	2006	2007	2008
Queries	50	50	50
Assessed posts	67,382	54,621	53,815
Relevant posts	19,891	12,187	11,735
Rel. posts/query	397	244	235
Query terms	99	85	128
Terms/query	2.0	1.7	2.6

5.2 Retrieval framework

Our retrieval framework uses a language modeling for IR approach (Croft and Lafferty 2003), where we estimate the probability of a document generating the query. We select this framework because it is theoretically sound and has shown good and robust performance on a broad range of retrieval tasks. We use the implementation as provided by Indri.[2]

5.3 Evaluation and significance

As explained before, we consider blog post retrieval to be a precision-oriented task, and focus mainly on precision metrics. The evaluation metrics on which we focus are standard precision-oriented IR metrics: mean reciprocal rank (MRR), and precision at ranks 5 and 10 (P5 and P10) (Baeza-Yates and Ribeiro-Neto 2010). For the sake of completeness we also report on the commonly used mean average precision (MAP) metric. In each table the best performing run per metric is bold-faced.

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

6 Results

We present our results in three sections. First, we show the performance of our baseline, see how it compares to previous approaches at TREC, and we show what the influence of spam filtering is (Sect. 6.1). We continue by applying our credibility-inspired indicators on top of our (spam filtered) baseline. Since we aim at improving precision using credibility, we mainly aim at reranking originally retrieved results, assuming that the baseline has a sufficiently strong recall. We start by reranking the top n of the initial run based solely on the credibility-inspired scores (Credibility-inspired reranking) in Sect. 6.2. We then take a step back, and combine retrieval scores and credibility-inspired scores in our Combined reranking approach in Sect. 6.3, and explore reranking the top n results using this combined score.

Both our reranking approaches are applied on the top n of the baseline ranking after spam filtering. We need to decide on a value for n to use, and to make results from the two approaches comparable, we choose the same n for both of them. For the result section we

[2] We used Lemur version 4.10, http://www.lemurproject.co.

take $n = 20$, as this value allows measuring changes in early precision (at ranks 5 and 10), without ignoring the initial ranking too much. In Sect. 7.3 we come back to this issue, and explore the influence of n on the performance of our approaches.

6.1 Baseline and spam filtering

We start by establishing our baseline: Table 6 shows the results on the three topic sets. Note that the baseline is strong: Its performance is better than or close to the best performing runs at TREC for all 3 years (our runs would have been at rank 1/15, 4/20, and 8/20). This is impressive knowing that the participating systems incorporate additional techniques like (external) query expansion, especially in 2007 and 2008.

We detailed our spam classification approach in Sect. 4.2, where we assigned a score to each blog based on the ratio of spam posts in that blog. To turn this score into a filter, we need a threshold for this ratio: every blog that has a higher ratio than this threshold is considered a splog and is removed from the results. Given the orientation towards precision we consider blogs that have >25% of their posts classified as spam posts to be splogs. This threshold leads to the removal of 6,412 splogs (198,065 posts).

Table 7 shows the results after filtering out spam. Results show similar performance on the precision metrics and a slight, though significant, drop in terms of MAP. We revisit the results of our spam classifier in Sect. 7.1.

In the remainder of the paper we have two notions of a "baseline." First, when it comes to comparing performance of our approaches, we do so against the baseline (row one in Table 7). Second, the ranking that is produced after filtering splogs (spam-filtered baseline; row two in Table 7) serves as the starting on top of which we apply our two reranking approaches: Credibility-inspired reranking and Combined reranking; put differently, in our discussions below reranking always includes spam filtering.

6.2 Credibility-inspired reranking

The first method of reranking we explore is Credibility-inspired reranking. As the name indicates, this approach takes only the credibility-inspired scores into account when

Table 6 Preliminary baseline scores for all three topic sets and their combination (150 topics)

Year	MRR	P5	P10	MAP
2006	0.7339	0.6880	0.6720	0.3365
2007	0.8200	0.7200	0.7240	0.4514
2008	0.7629	0.6760	0.6920	0.3800
All	0.7722	0.6947	0.6960	0.3893

Table 7 Results before and after filtering spam. Significance tested against the baseline

Run	MRR	P5	P10	MAP
Baseline	0.7722	0.6947	0.6960	0.3893
Spam-filtered baseline	**0.7894**	**0.7107**	**0.7087**	0.3774^{∇}

Table 8 Results for Credibility-inspired reranking on the top 20 results based on each of the credibility-inspired indicator scores for all 150 topics. Significance tested against the baseline

Run	MRR	P5	P10	MAP
Baseline	0.7722	0.6947	0.6960	**0.3893**
Upperbound	0.9806	0.9507	0.8787	0.3976
Post-level indicators				
Quality	0.8200	0.7040	0.6980	$0.3749^{\blacktriangledown}$
Document length	0.7702	0.6907	0.6840	$0.3731^{\blacktriangledown}$
Timeliness	0.8138^{Δ}	0.7213	0.7127	$0.3782^{\triangledown}$
Semantics	0.8144	0.7200	0.7167	$0.3751^{\blacktriangledown}$
Blog-level indicators				
Comments	0.8252^{Δ}	0.7187	0.7120	$0.3743^{\blacktriangledown}$
Pronouns	0.7270	$0.6173^{\blacktriangledown}$	$0.6620^{\triangledown}$	$0.3716^{\blacktriangledown}$
Coherence	0.7648	0.6720	0.6707	$0.3730^{\blacktriangledown}$
Regularity	$0.7080^{\triangledown}$	$0.6493^{\triangledown}$	$0.6640^{\triangledown}$	$0.3705^{\blacktriangledown}$
Expertise	0.7595	0.6653	0.6793	$0.3766^{\triangledown}$
Combinations				
Post level	**0.8289**	**0.7347**	**0.7193**	$0.3748^{\blacktriangledown}$
Blog level	0.7659	0.6560	$0.6673^{\triangledown}$	$0.3741^{\blacktriangledown}$
All	0.8163	0.7067	0.6920	$0.3755^{\triangledown}$

reranking the top 20 results of our baseline ranking. That is, we take the ranking produced after filtering spam, ignore retrieval scores for the top 20 results, and assign to each of the top 20 posts the score as assigned by each credibility-inspired indicator (viz. Sect. 4), and construct the new ranking based on these scores. The posts ranked lower than position 20 keep their original retrieval score/ranking.

We present the results of Credibility-inspired reranking in Table 8. The results are divided into four groups: (1) the baseline and the manual upper bound (which reranks the posts based on their relevance assessments), (2) the individual post-level indicators, (3) the individual blog-level indicators, and (iv) the combined indicators on post level, blog level, and both. We first focus on the individual indicators.

The individual indicators show a wide range in performance. All indicators show a drop in MAP compared to the baseline, but this was expected. We focus on the precision metrics and here we observe that almost all post-level indicators seem to improve over the baseline, although only the improvement on MRR by timeliness is significant. Looking at the blog-level indicators, we find that only the comments indicator improves over the baseline, with MRR showing a significant increase. The other blog-level indicators perform worse than or similar to the baseline. The highest scores on the precision metrics, when looking at the individual indicators, are achieved by three different indicators: comments on MRR, timeliness on P5, and semantics on P10.

Next, we shift our attention to combinations of indicators (the bottom part of Table 8). From these results we observe two things. First, the combined blog-level indicators do not improve over the baseline run on any metric, which is disappointing, but expected given the scores of individual indicators on this level. Second, the combined post-level indicators have the highest scores on the precision metrics, but improvements are not significant.

As an aside, given the strong performance of the comments indicator, it is natural to wonder what would happen if this blog level indicator were included with the post level indicators. That is, we take all post-level indicators and combine these with the comments indicator only. Using this combination we achieve the following scores: MRR 0.8280; P5 0.7280; P10 0.7167; and MAP 0.3744. Here, we find that performance on all metrics is still slightly below post-level indicators only.

Summarizing, we see that the Credibility-inspired reranking approach works well for post-level indicators, although it is hard to obtain significant improvements. The blog-level indicators, with the exception of comments, perform rather disappointing. Given the fact that we completely ignore the retrieval score once we start the reranking process, the results obtained by post-level indicators are quite remarkable and show the possibilities of taking ideas from the credibility framework on board as precision enhancement.

6.3 Combined reranking

Completely ignoring the initial retrieval score sounds like a "bad" idea: there is a reason why certain posts get assigned a higher retrieval score than others and we probably should be using these differences in scores. In this section we take another approach to incorporating ideas from the credibility framework in ranking blog posts: we combine the original retrieval score and the credibility-inspired score of posts to rerank the baseline ranking. We, again, look only at the top 20 results of the original ranking and multiply the retrieval score of each document by the (normalized) score on each credibility-inspired indicator. We present the results similar to the previous section: (1) the baseline and

Table 9 Results for Combined reranking using a combination of retrieval and credibility scores, and reranking the top 20 results based on this score for all 150 topics. Significance tested against the baseline

Run	MRR	P5	P10	MAP
Baseline	0.7722	0.6947	0.6960	**0.3893**
Upperbound	0.9806	0.9507	0.8787	0.3976
Post-level indicators				
Quality	0.7986^{Δ}	0.7120	0.7020	0.3768^{∇}
Document length	0.8009	0.7107	0.7013	0.3768^{∇}
Timeliness	0.8151^{Δ}	0.7253^{Δ}	**0.7147**	0.3781^{∇}
Semantics	0.8210^{Δ}	$\mathbf{0.7347}^{\Delta}$	0.7173	0.3779^{∇}
Blog-level indicators				
Comments	$\mathbf{0.8311}^{\Delta}$	0.7200	0.7093	0.3754^{∇}
Pronouns	0.7796	0.7093	0.7027	0.3772^{∇}
Coherence	0.7531	0.6760	0.6707^{∇}	0.3757^{∇}
Regularity	0.7624	0.6787	0.6787	$0.3743^{\blacktriangledown}$
Expertise	0.7608	0.6827	0.6827	0.3782^{∇}
Combinations				
Post level	0.8098^{Δ}	0.7227^{Δ}	0.7113	0.3771^{∇}
Blog level	0.7622	0.6827	0.6747	0.3766^{∇}
All	0.7895	0.7160	0.7027	0.3769^{∇}

upperbound, (2) the individual post-level indicators, (3) the individual blog-level indicators, and (4) the combinations of indicators. The results are listed in Table 9.

Results show that most post-level indicators are able to improve over the baseline on precision metrics. Especially scores on MRR improve significantly and both the timeliness and semantics indicators show large improvements on MRR and P5 compared to the baseline. Compared to the Credibility-inspired reranking approach in the previous section, we observe better performance on the precision at 5 and 10 metrics, as well as more significant (stable) improvements. Looking at the individual blog-level indicators we see a similar pattern as before: the comments indicator works well on MRR, but coherence, regularity, and expertise cannot improve over the baseline on any metric. An interesting difference with the previous approach is that both the pronouns and regularity indicators, which dropped significantly in performance compared to the baseline in Sect. 6.2 are now comparable to the baseline.

When combining the credibility-inspired indicators on our two levels we notice that scores for the post-level combination are, in absolute sense, slightly below the results of Credibility-inspired reranking, but they do show significant improvements over the baseline on precision metrics, indicating a more stable improvement.

Given the below-baseline performance of some of the blog-level indicators, we experiment by excluding them from the final (all) combination. Table 10 shows the results of using only comments and using both comments and pronouns in this final combination. Results here show that we can indeed improve over the combined post-level indicators when adding comments and pronouns to the combination. The final two runs show a (strong) significant improvement over the baseline on MRR and precision at 5.

Summarizing, we find that Combined reranking resembles a "smoothed" version of Credibility-inspired reranking: It takes away the outliers, leading to slightly lower absolute scores than for Credibility-inspired reranking, but the improvements over the baseline are more often significant. Again, post-level indicators are the better performing ones, although this time we find that combining these with two blog-level indicators (comments and pronouns) leads to even better performance. Combined reranking is a powerful way of incorporating ideas from the credibility framework, resulting in stable improvements.

In the analysis section (Sect. 7), we often look at the two best performing runs from both approaches. For Credibility-inspired reranking this is the post-level combination run, and for Combined reranking it is the post-level + comments + pronouns run.

7 Analysis and discussion

We presented the overall results of our two credibility-inspired reranking approaches in the previous section. These results, however, hide a lot of detail, which could be important to

Table 10 Results for combining post-level indicators and one or two blog-level indicators. Significance tested against the baseline

Run	MRR	P5	P10	MAP
Baseline	0.7722	0.6947	0.6960	**0.3893**
Post level	0.8098^{Δ}	0.7227^{Δ}	**0.7113**	0.3771^{∇}
Post level + comments	$0.8107^{\blacktriangle}$	$\mathbf{0.7253}^{\Delta}$	0.7100	0.3770^{∇}
Post level + comments + pronouns	$\mathbf{0.8113}^{\blacktriangle}$	0.7240^{Δ}	0.7107	0.3770^{∇}

understanding what exactly is happening. In this section we perform extensive analyses on our results from four perspectives. First, in Sect. 7.1, we look at the performance of our spam classifier. In Sect. 7.2 we acknowledge the fact that we are looking at reranking strategies and give more details on how our approaches really affect ranking by looking at swaps, the positions of relevant posts, and specific (relevant) posts that move significantly up or down the ranking. Sect. 7.3 deals with per-topic analyses of our indicators and reranking approaches and compares various runs on a per-topic basis and explores which specific topics show improvement or drops in performance. We discuss the setting of n, the number of results we rerank, in Sect. 7.4, and finally, we explore the interplay between credibility-inspired ranking and relevance in Sect. 7.5.

7.1 Spam classification

The official collection was purposefully injected with spam by gathering blog posts from known splogs. In total, 17,958 splogs were followed during the 11 week period of crawling. As mentioned before, we use a relatively simple approach to splog detection based on a rather small training set and a limited set of features (unigrams). From the 6,412 blogs classified as splogs, 4,148 are really splogs (precision 65%). The recall for our classifier is rather low, with 4,148 out of 17,958 splogs identified (recall 23%).

7.2 Changes in ranking

Our two approaches for incorporating credibility-inspired indicators are based on reranking an initial ranking of posts. Besides looking at scores produced by each of the (re)rankings, we can also look at the rankings themselves and explore how they differ between runs. First, we look at the number of swaps in the top 20 after reranking. The higher this number, the more changes in positions between the baseline and the reranked result lists. We compare the various indicators and also the two reranking approaches, in Table 11. Note that for most analyses in this section the numbers for the timeliness indicators might seem

Table 11 Average number of swaps (changes in ranking) per topic between each run and the (spam-filtered) baseline

	Swaps	
Indicator	Reranking	Combining
Quality	19.0	15.1
Document length	18.9	15.6
Timeliness	6.2	6.1
Semantics	17.2	16.5
Comments	19.0	18.4
Pronouns	19.0	7.2
Coherence	19.0	18.2
Regularity	19.0	17.7
Expertise	18.7	17.9
Post level	18.8	14.9
Blog level	18.7	16.5
All	18.8	14.8

out of the ordinary, but this is because this indicator only affects 50 of the 150 topics, which influences the averages quite a bit.

We observe that in the Credibility-inspired reranking approach more swaps are generated than in the Combined reranking approach, although in some cases (e.g., timeliness) the difference is only marginal. The reason for the difference between the two approaches is that in the Combined reranking approach the initial retrieval score acts as a kind of "smoothing," making the changes less radical. In general we see that most of the results in the top 20 get a different position after applying our reranking techniques.

To examine how successful the swaps are, we combine the swaps with relevance information; Tables 12 (Credibility-inspired reranking) and 13 (Combined reranking) show the average number of *relevant* posts per topic that go up or down in the ranking after reranking has been applied and the average number of positions each of these posts gains or loses. We should note that relevant posts going down in the ranking is not necessarily a problem, as long as the posts crossing them are relevant too.

Comparing the two approaches on these numbers, we observe that all the numbers (except the ratios) are higher for Credibility-inspired reranking than for Combined reranking: more relevant posts go up, more relevant posts go down and for both the average number of positions is higher. The only numbers that are consistently higher for Combined reranking are the ratios of number of relevant posts going up vs. relevant posts going down. Here, we see that for most indicators this ratio is above 1 for Combined reranking, whereas it is above 1 for only two indicators for Credibility-inspired reranking.

Looking at the individual indicators for Combined reranking, we notice some interesting differences. The quality indicator has by far the highest ratio of relevant posts up vs. down, but the average number of positions is almost the lowest over all indicators. The comments indicator on the other hand has a mediocre up vs. down ratio, but the average number of positions relevant posts move (either up or down) is much higher than most other indicators.

Table 12 Credibility-inspired reranking: average number of *relevant* posts per topic that go up or down the ranking after reranking, and the average number of positions these posts go up or down. Also: the ratio of rising vs. dropping relevant posts per indicator

Indicator	Up		Down		Ratio up/down
	Posts	Positions	Posts	Positions	
Quality	6.43	7.03	6.63	6.75	0.97
Document length	6.24	6.31	6.71	6.26	0.93
Timeliness	1.91	2.47	2.48	1.84	0.77
Semantics	6.19	5.36	5.63	5.68	1.10
Comments	6.55	6.61	6.51	6.43	1.01
Pronouns	6.18	6.33	6.89	6.75	0.90
Coherence	6.23	6.43	6.84	6.63	0.91
Regularity	6.41	6.53	6.69	6.96	0.96
Expertise	6.09	6.09	6.79	6.38	0.90
Post level	6.65	6.56	6.29	6.45	1.06
Blog level	6.06	6.04	6.81	6.41	0.89
All	6.37	6.33	6.55	6.56	0.97

Table 13 Combined reranking: average number of *relevant* posts per topic that go up or down the ranking after reranking, and the average number of positions these posts go up or down. Also: the ratio of rising vs. dropping relevant posts per indicator

Indicator	Up		Down		Ratio up/down
	Posts	Positions	Posts	Positions	
Quality	7.02	2.38	3.37	5.11	2.08
Document length	5.40	2.91	5.32	3.17	1.02
Timeliness	2.13	2.21	2.24	1.88	0.95
Semantics	5.75	4.41	5.46	4.29	1.05
Comments	6.68	5.65	5.95	6.08	1.12
Pronouns	2.39	1.12	2.45	1.31	0.98
Coherence	6.39	5.25	6.13	6.13	1.11
Regularity	6.35	4.25	5.78	5.18	1.10
Expertise	5.89	5.18	6.41	5.31	0.92
Post level	6.37	2.54	3.96	3.85	1.61
Blog level	5.56	3.50	5.82	4.09	0.96
All	5.63	2.54	4.67	3.47	1.21

7.2.1 Per-post analysis

Next, we drill down to the level of individual posts and look at example posts that show "interesting" behavior. First we look at posts that move up or go down most when comparing our approaches to the baseline. Table 14 shows the average of these maxima per topic for two selected indicators and the best performing run per approach. We observe that Credibility-inspired reranking leads to posts going up and also going down a lot, whereas Combined reranking is more modest in both cases.

We zoom in and look at the posts themselves. Table 15 shows four examples of posts that are relevant to a topic and that show the largest "bump" for that topic after using Combined reranking (with post-level + comments + pronouns). For each example post we give the topic to which it is relevant, the change in positions, the ID, a part of the post's text, and the reasons why this post went up in the ranking.

The example posts show that we are able to push more credible posts up the ranking. As to the indicators that matter most in these examples, we observe that most have a high (text) quality (few spelling mistakes, correct use of punctuation and capitalization), have

Table 14 Average maximum number of positions per topic a relevant post goes up or down the top 20 of the ranking for two individual indicators and the best run per approach

Approach	Indicator	Avg. max. up	Avg. max. down
Credibility-inspired	Quality	14.6	15.0
	Comments	14.6	14.1
	Post level	14.0	14.1
Combined	Quality	4.5	9.2
	Comments	12.7	13.5
	Post level + comments + pronouns	5.2	7.4

Table 15 Examples of relevant posts helped by credibility after reranking using Combined reranking (post-level + comments + pronouns)

Topic	Ann Coulter (854)
Change in positions	3 (5 to 2)
Post ID	BLOG06-20060131-018-0031501574
Conservative commentator Ann Coulter has come under media fire yet again, this time for joking that U.S. Supreme Court Justice John Paul Stevens should be poisoned so that conservatives can gain a majority on the high court. Coulter is an articulate conservative and an outspoken Christian, but it is becoming increasingly clear that her "bomb throwing" style does more harm than good to these cause.	
Why?	Many comments
	High quality
	Few pronouns
Topic	Cheney hunting (867)
Change in positions	10 (20–10)
Post ID	BLOG06-20060213-013-0027595552
Today the AP reported: WASHINGTON - Vice President Dick Cheney accidentally shot and wounded a companion during a weekend quail hunting trip in Texas, spraying the fellow hunter in the face and chest with shotgun pellets. Vice President Cheney explained the shooting this way: "I was tracking a covey of quail with my gun barrel. Suddenly Whittington just popped up from the grass, directly in the way, so I shot him. I know my critics on the left will point out that Whittington is not a bird, but he was between the quail and my gun.	
Why?	Very timely
	Many comments
	High quality
	Few pronouns
Topic	Seahawks (882)
Change in positions	6 (12–6)
Post ID	BLOG06-20060207-025-0012517965
DETROIT – Shoulders slumped. Eyes drooped, some red with the hint of earlier tears. Heads sagged. The Seahawks' locker room was a sad and somber place. In many of their minds, the Seahawks were the better team in Super Bowl XL. The scoreboard at Ford Field said differently, however, and that was all that mattered. The greatest Seahawks season ended in bitter disappointment Sunday, a 21-10 loss to the Pittsburgh Steelers. The way the Seahawks lost—with mistake after mistake—left them disconsolate.	
Why?	Very timely
	High quality
	Proper semantics
Topic	Qualcomm (884)
Change in positions	4 (6–2)
Post ID	BLOG06-20060212-028-0007415694
A federal district court in California permanently barred chip maker Broadcom from prosecuting several of its patent infringement claims against Qualcomm before the International Trade Commission, ruling that the dispute must be resolved under the court's own jurisdiction in San Diego. Judge Rudi M. Brewster said in his ruling the week of Feb. 6 that Broadcom cannot pursue two individual claims from its patent case with the ITC in Washington, or in another California District Court, based on the details of a licensing agreement signed by the companies related to the legal dispute.	
Why?	Proper semantics
	High quality
	Few pronouns

many comments, are timely (i.e., published on the day of the related event), and share semantics with related news articles.

We perform a similar analysis for relevant posts that drop in the ranking after using Combined reranking. Table 16 shows four of these posts, again with a snippet from the post and the reasons why the system believes these posts should drop.

Table 16 Examples of relevant posts hurt by credibility after reranking using Combined reranking (post-level + comments + pronouns)

Topic	Hybrid car (879)
Change in positions	−15 (1–16)
Post ID	BLOG06-20051219-075-0006828953
If your goal is to find out whether a hybrid car is right for you or your biggest desire is reducing your impact on the environment buy using a hybrid car, then take advantage of the advantages of hybrid car material that we have pulled together. Browse the site for additional Hybrid Cars information.	
Why?	Few comments
	Short
	Improper semantics
Topic	Qualcomm (884)
Change in positions	−4 (2 to 6)
Post ID	BLOG06-20051211-081-0015735208
I have been analyzing wireless communications for 26 years. I am president of Wireless Internet & Mobile Computing, a pioneering consulting firm that helps create new and enhance existing wireless data businesses in the United States and abroad. Previously, I created the world's first wireless data newsletter, wireless data conference, cellular conference and FM radio subcarrier newsletter. I was instrumental in creating and developing the world's first cellular magazine. I also helped create and run the first association in the U.S. for the paging and mobile telephone industries.	
Why?	Few comments
	Improper semantics
	Many pronouns
Topic	Oprah (895)
Change in positions	−14 (6–20)
Post ID	BLOG06-20060211-010-0023506187
George: I appreciate that. Fighting evil, it's hard work. I, um ... my SUV, um ... Oprah: George, you just go ahead, cry if you want to. I'm not ashamed to tell you that when I watched your speech, I cried. George: I really appreciate that, Oprah. Oprah: But George, I have to be straight with you now. I ... I have to say it is difficult for me to talk to you because I also feel really duped.	
Why?	Not timely
	Low quality
	Improper semantics
Topic	Lance Armstrong (940)
Change in positions	−2 (3 to 5)
Post ID	BLOG06-20051209-083-0015483759
When is enough, well... enough? Lance Armstrong "was" possibly the most tested athlete of all time never being tested positive once for using performance enhancing drugs yet the European press simply will not let it go. Again the press have attacked Lance Armstrong for using a drug called "EPO" which increases performance in athetes. Maybe we can let Armstrong retire a champ instead continuing down this road... ??	
Why?	Few comments
	Low quality
	Short

Looking at these posts, we feel that, although relevant, they are less credible than the posts in Table 15. The first post is a collection of links to other sources and contains in itself not much information, which is reflected by its short length and lack of comments. The second post sounds more credible, but is quite biased (i.e., a high number of pronouns) and has again only few comments. The third post is a fake "conversation" between Oprah and George Bush and is considered less credible because improper semantics and low text quality. Finally, the fourth post is characterized by punctuation "abuse" (. . . , ??), short length, and very few comments.

In general we see that Credibility-inspired reranking is a more radical reranking approach, leading to many changes in the ranking and many (relevant) posts moving up and down. This is risky; it can lead to high gains, but also to large drops in performance. Combined reranking is a more careful, "smoothed" approach, which shows (slightly) fewer changes and moves in the ranking, but is more stable in its improvements (i.e., the ratio of posts going up and down), leading to significant improvements.

Looking at examples of relevant posts that are helped or hurt by credibility-inspired indicators, we find that posts that are pushed up the ranking are indeed more credible, whereas the posts that are pushed down seem to be less credible (although still relevant). There is not one indicator that leads to these changes, but it is always a combination of indicators (like comments, timeliness, semantics, and quality). We revisit the influence of individual indicators and the interplay between credibility-inspired ranking and relevance in Sect. 7.5.

7.3 Per topic analysis

Performance numbers averaged over 150 topics hide a lot of details. In this section we analyze the performance of our approaches on a per-topic basis and see how their behavior differs for various topics. We start by looking at the results of our best performing Credibility-inspired reranking and Combined reranking runs as compared to the baseline. The plots in Fig. 4 show the increase or decrease on precision metrics for each topic when comparing thetwo approaches to the baseline.

The plots show some interesting differences between the two reranking approaches. First, both approaches have topics on which they improve over the baseline, as well as topics for which the baseline performs better. In general, we observe that Credibility-inspired reranking has more topics that improve over the baseline than Combined reranking, but also more topics that drop in performance. Both gains and losses are higher for Credibility-inspired reranking compared to Combined reranking. The actual number of topics going up or down for both approaches compared to the baseline are listed in Table 17.

We move on to the analysis of a selection of individual indicators. Figure 5 shows similar plots as before for four individual indicators; We only show precision at 5, to keep the number of plots limited.

The quality indicator shows similar behavior as the combinations of indicators: numbers for Credibility-inspired reranking are higher across the board. This pattern is, however, not so strong for timeliness and comments, where both approaches show similar behavior (i.e., equal number of topics increasing and decreasing compared to the baseline). We included the expertise indicator to show that, although overall performance of this indicator was below the baseline, we can improve over the baseline for a number of topics (32 topics for Credibility-inspired reranking and 30 for Combined reranking).

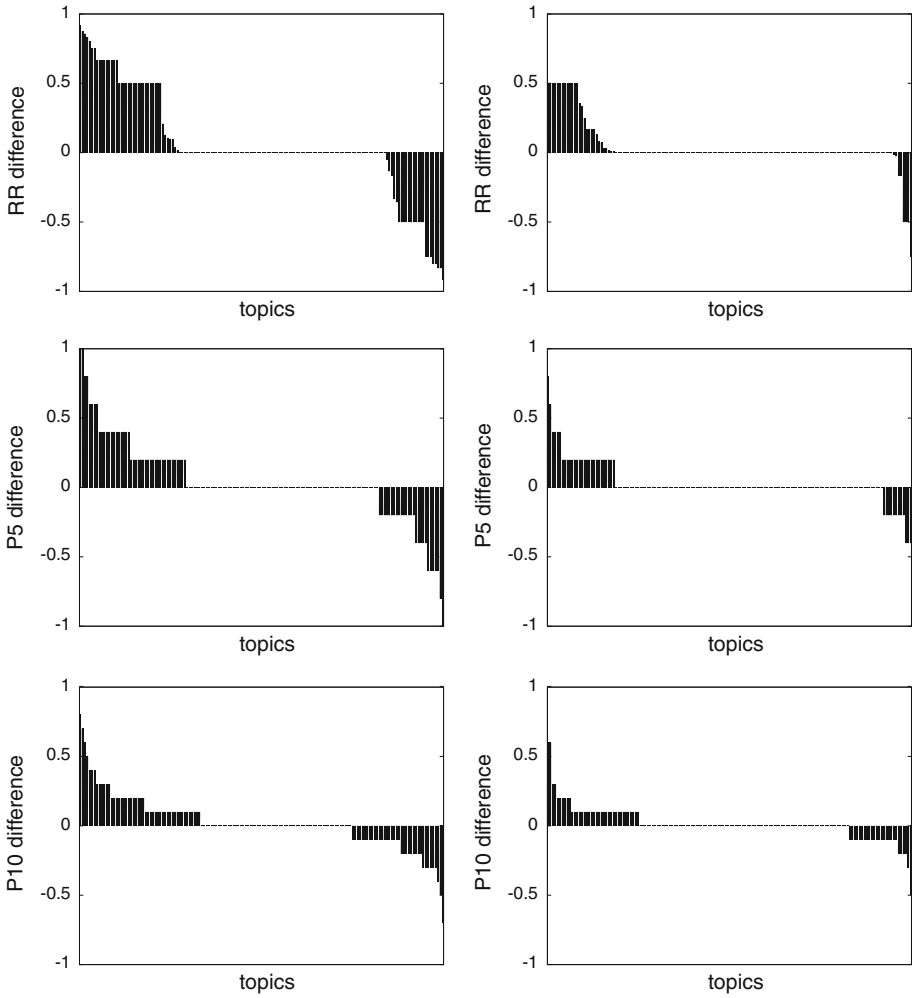

Fig. 4 Comparing the baseline against (*Left*) Credibility-inspired reranking (post-level indicators) and (*Right*) Combined reranking (post-level + comments + pronouns). A *positive bar* indicates the topic improves over the baseline, a *negative bar* indicates a drop compared to the baseline

Table 17 Number of topics that increase or decrease as compared to the baseline for both approaches on precision metrics

Approach	RR		P5		P10	
	Up	Down	Up	Down	Up	Down
Credibility-inspired reranking	42	24	44	27	50	38
Combined reranking	29	9	28	12	38	26

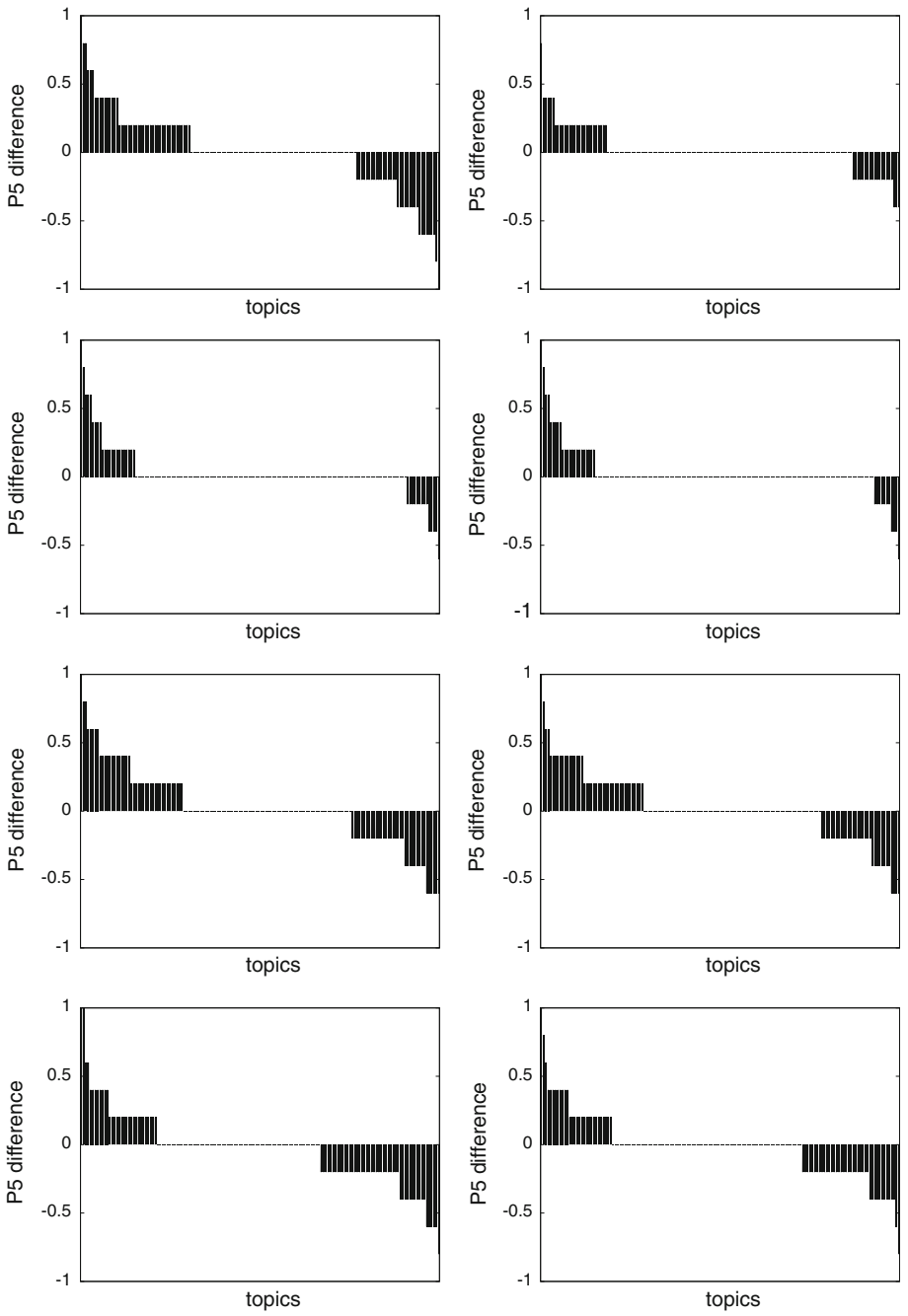

Fig. 5 Comparing the baseline against (*Left*) Credibility-inspired reranking (*post-level* indicators) and (*Right*) Combined reranking (post-level + comments + pronouns) on precision at 5 for four individual indicators: (1) quality, (2) timeliness, (3) comments, and (4) expertise. A *positive bar* indicates the topic improves over the baseline, a *negative bar* indicates a drop compared to the baseline

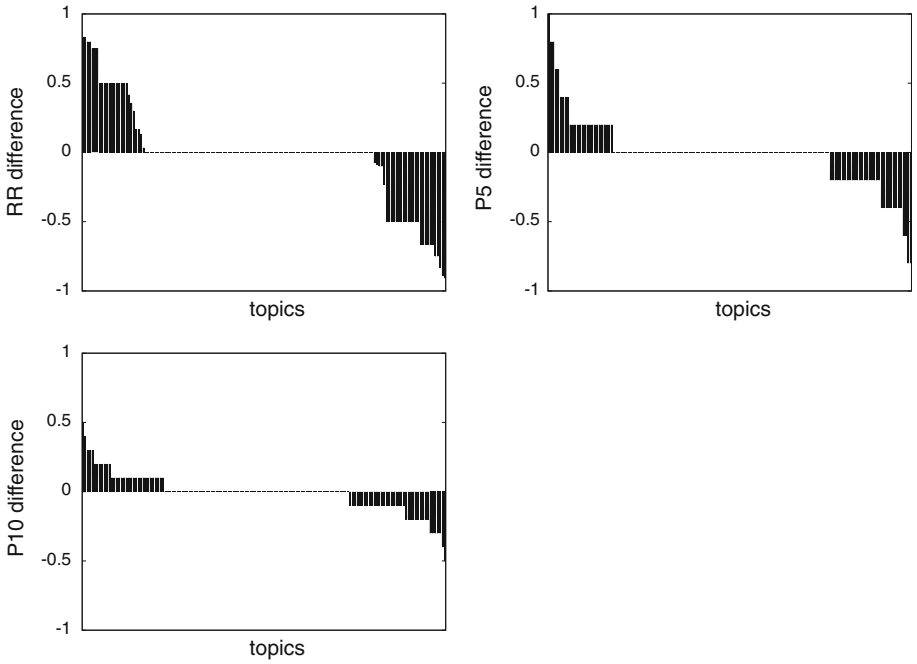

Fig. 6 Comparing Credibility-inspired reranking (*post-level* indicators), as baseline, to Combined reranking (post-level + comments + pronouns) on (*Top left*) RR, (*Top right*) P5, and (*Bottom*) P10. A *positive bar* indicates that Combined reranking makes the topic improve over Credibility-inspired reranking, a *negative bar* indicates the opposite

Finally, we compare the two reranking approaches in the same way: per topic. Figure 6 shows the number of topics that prefer either Credibility-inspired reranking ("negative" bars) or Combined reranking ("positive" bars) on the precision metrics.

The plots show that both reranking approaches have topics on which they clearly outperform the other, although in general the Credibility-inspired reranking is preferred for slightly more topics. To be precise, Credibility-inspired reranking is preferred for 30 (RR), 34 (P5), and 40 (P10) topics, whereas Combined reranking is preferred for 26 (RR), 27 (P5), and 34 (P10) topics.

7.3.1 Very early precision

We shift focus to MRR, the ability to rank the first relevant post as high as possible. We see that our Combined reranking approach is capable of moving the first relevant post from position 2 to position 1 for 13 topics, while another 16 topics show an increase in RR as well. On the other hand, only 9 topics show a decrease in RR. Table 18 shows on the left hand side the topics that improve the most after reranking and on the right the topics that drop the most.

We perform the same comparison between Credibility-inspired reranking using post-level indicators and the baseline. Table 19 shows the topics that show the largest difference on RR between the two runs. In total, 42 topics go up in RR, and 24 go down.

Table 18 Topics that increase or decrease most on RR using Combined reranking (post-level indicators + comments + pronouns), compared to the baseline

Increase			Decrease		
#	Topic	ΔRR	#	Topic	ΔRR
942	Lawful access	0.5000	929	Brand manager	−0.7500
1,018	Mythbusters	0.5000	921	Christianity today	−0.5000
1,011	Chipotle restaurant	0.5000	943	Censure	−0.5000
1,023	Yojimbo	0.5000	869	Muhammad cartoon	−0.5000
903	Steve jobs	0.5000	870	Barry bonds	−0.1667
885	Shimano	0.5000	893	Zyrtec	−0.1666
913	Sag awards	0.5000	1,038	Israeli government	−0.0250
895	Oprah	0.5000	1,012	Ed norton	−0.0139
873	Bruce bartlett	0.5000	881	Fox news report	−0.0047
947	Sasha cohen	0.5000			
879	Hybrid car	0.5000			
878	Jihad	0.5000			
1,042	David irving	0.5000			

Table 19 Topics that increase or decrease most on RR using Credibility-inspired reranking (post-level indicators) compared to the baseline

Increase			Decrease		
#	Topic	ΔRR	#	Topic	ΔRR
1,034	Ruth rendell	0.9167	921	Christianity today	−0.9167
1,012	Ed norton	0.8750	1,014	Tax break for hybrid automobiles	−0.8333
940	Lance armstrong	0.8571	937	Lexisnexis	−0.8333
923	Challenger	0.8333	950	Hitachi data systems	−0.8000
1,035	Mayo clinic	0.8000	1,039	The geek squad	−0.8000
887	World trade organization	0.7500	1,022	Subway sandwiches	−0.7500
941	Teri hatcher	0.7500	1,025	Nancy grace	−0.7500
1,007	Women in saudi arabia	0.6667	1,019	China one child law	−0.7500
1,013	Iceland european union	0.6667	915	Allianz	−0.5000
933	Winter olympics	0.6667	855	Abramoff bush	−0.5000
880	Natalie portman	0.6667	943	Censure	−0.5000
890	Olympics	0.6667	918	Varanasi	−0.5000
1,008	Un commission on human rights	0.6667	938	Plug awards	−0.5000
1,047	Trader joe's	0.6667	867	Cheney hunting	−0.5000
893	Zyrtec	0.6667	866	Whole foods	−0.5000
900	Mcdonalds	0.6667	925	Mashup camp	−0.5000

Some interesting observations can be made from the tables with topics. E.g., we notice that for topic 921 ("christianity today") it is hard to maintain a relevant post at the first position for both approaches and the same goes for topic 943 ("censure"). Credibility-inspired reranking is capable of pushing the first relevant result quite a bit up for topics 893

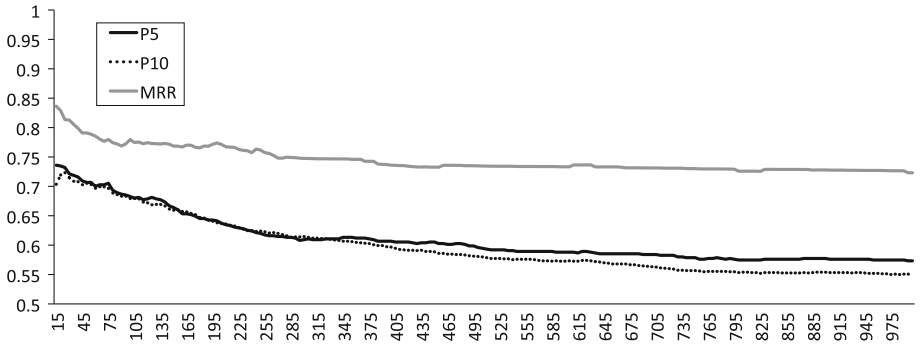

Fig. 7 Influence of reranking top n (x-axis) on precision at 5 (P5) and 10 (P10) and MRR for Credibility-inspired reranking using post-level indicators

("zyrtec") and 1012 ("ed norton"), whereas these drop for Combined reranking. All other topics that either increase or decrease are different between both approaches, which again supports the notion that certain topics are helped by Credibility-inspired reranking and others by Combined reranking.

7.4 Impact of parameters on precision

So far, we have looked at the results of reranking only the top 20 of the initial ranking. What happens if we change the value of n and rerank not 20, but the first 15 or 500 results of the ranking? We first explore the impact of different values of n on Credibility-inspired reranking on precision metrics, and then look at Combined reranking.

The plot in Fig. 7 shows the change in performance for Credibility-inspired reranking on precision when using increasing values of n. We start at $n = 15$, so that we can measure a difference in P10 after reranking. On all metrics performance drops quite rapidly with n going up and it keeps dropping all the way up to $n = 1{,}000$.

The best performance for Credibility-inspired reranking is achieved using either $n = 15$ (for P5 and MRR) or $n = 25$ (for P10). Results of these two runs and the baseline are reported in Table 20. The results for MRR using $n = 15$ are higher than before and show a significant increase over the baseline. For P5 and P10 the results are slightly higher, but are still not significantly better.

Looking at Combined reranking we find a very stable performance on all metrics over all n's. Smoothing the credibility scores with the initial retrieval score leads to improvements, but the ranking does not change anymore going further down the ranking than position 15–20. The best performance is already achieved using $n = 20$ and there is no need to present further results here.

7.5 Credibility-inspired ranking vs. relevance ranking

We have seen that the effects of using credibility-inspired indicators on blog post retrieval are positive, but why this is the case? One issue that we should raise is the fact that assessors in the blog post retrieval task are asked to judge whether a blog post is *topically relevant* for a given topic. This relevance is assessed regardless of other factors that could otherwise influence judgements (e.g., readability, opinionatedness, quality). If we would follow this line of reasoning, we might wonder why credibility-inspired indicators have an

Table 20 Results for the best values of n (15 and 25), our baseline, and the run presented before ($n = 20$) for Credibility-inspired reranking (using post-level indicators). Significance tested against the baseline

Run	MRR	P5	P10	MAP
Baseline	0.7722	0.6947	0.6960	**0.3893**
$n = 15$	**0.8364**$^{\Delta}$	**0.7360**	0.7033	0.3754$^{\nabla}$
$n = 20$	*0.8289*	*0.7347*	*0.7193*	*0.3748*$^{\blacktriangledown}$
$n = 25$	0.8134	0.7320	**0.7233**	0.3723$^{\blacktriangledown}$

effect on the performance at all. In order to gain a better understanding of this matter, we explore the topics that show the biggest increase or decrease in terms of precision at 10 and identify reasons for the change in performance. Below we list the factors that are most influential in performance changes.

Spam filtering: We already discussed the issue of spam classification in Sect. 7.1. In this analysis we find that spam filtering is one of the main contributors to both improvements and drops in performance. By removing spam blogs, proper blog posts are promoted to higher ranks, leading to better results. Similarly, when spam classification fails and non-spam blogs are filtered out, non-relevant blog posts might take their place in the ranking, leading to a drop in performance.

Timeliness: For topics that are time sensitive, the timeliness indicator is very influential. It often leads to relevant blog posts being pushed up in the ranking, while non-relevant blog posts are pushed down. Since this indicator is topic-dependent it does not influence all topics.

Semantics: Another topic-dependent indicator, semantics, shows a large degree of influence on performance. As with the other indicators, semantics can make relevant posts move up the ranking and non-relevant posts down, but also the other way around.

Comments: We observe that the number of comments a post receives is among the more influential indicators. One of the reasons why this indicator has so much influence could be that the text of the comments is considered to be part of the blog post and thus is being considered when determining relevance. A larger number of comments leads to extra text associated with the post and possibly to a better match between blog post and topic.

Post length: The influence of the length of a document has attracted a lot of interest over the years (see e.g., Losada and Azzopardi 2008; Singhal et al. 1996), and its influence on retrieval performance is well-studied. In this chapter we also find that post length is one of the indicators with most influence on performance.

We observe that the credibility-inspired indicators each have their own reasons for improving (topical) blog post retrieval performance. However, the credibility framework offers us a principled way of combining these indicators and leaves space to include other indicators as well. Moreover, although we do not have the test collections to prove it, anecdotes suggest that the credibility-inspired indicators do indeed push more credible posts up the ranking.

8 Conclusions

In this paper we explore the use of ideas from a credibility framework in blog post retrieval. Based on a previously introduced credibility framework for blogs, we define

several credibility-inspired indicators. These indicators are divided into post-level and blog-level indicators. Post-level indicators include spelling mistakes, correct capitalization, use of emoticons, punctuation abuse, document length, timeliness (when related to a news event), and how its semantics matches formal (news) text. On the blog level we introduce the following indicators: average number of comments, average number of pronouns, regularity of posting, coherence of the blog, and the expertise of the blogger.

Since the task at hand is precision-oriented and we expect credibility to help on precision, we propose to use inspiration from the credibility framework in a reranking approach and we introduce two ways of incorporating the credibility-inspired indicators in our blog post retrieval process. The first approach, Credibility-inspired reranking, simply reranks the top n of a baseline based on the credibility-inspired score. The second approach, Combined reranking, multiplies the credibility-inspired score of the top n results by their retrieval score and reranks based on this score.

Results show that Credibility-inspired reranking leads to larger improvements over the baseline than Combined reranking, but both approaches are capable of improving over an already strong baseline. For Credibility-inspired reranking the best performance is achieved using a combination of all post-level indicators. Combined reranking works best using the post-level indicators combined with comments and pronouns. The blog-level indicators expertise, regularity, and coherence do not contribute positively to the performance, although analysis shows that they can be useful for certain topics.

Analyses revealed that reranking on credibility-inspired scores alone (Credibility-inspired reranking) leads to higher gains and higher drops: its absolute scores are higher than for Combined reranking, but less stable. Combined reranking managed to improve significantly over the baseline on MRR and P5 and Credibility-inspired reranking can only do that after optimizing n to 15. Examples of posts that are affected by the reranking approaches indicated that we get the desired effect of moving credible posts up the ranking, but this is not always reflected in retrieval performance, as our test collection does not allow for direct measurement of credibility. We identified the most influential indicators and explained why these indicators lead to improvements in retrieval performance.

Concluding, in this paper we have shown that we can translate certain credibility indicators to measurable indicators from blog posts and their blogs. Applying two reranking approaches shows that the (early) precision of blog post retrieval can benefit from incorporating credibility-inspired indicators. Interestingly, ignoring the original retrieval score when reranking leads to the highest scores, although combining the two scores leads to more significant improvements in precision. The credibility framework offers us a principled way of adding indicators to a retrieval model, although the real effect on credibility ranking needs to be examined when an appropriate collection is available. Future work focuses around the blog-level indicators, that have proven to be harder to estimate than post-level indicators. We believe that blog-level indicators are important, but that we need other ways of estimating coherence, regularity, an expertise. An important future direction is the direct measurement of credibility using our indicators; for this, we need new collections or assessments.

Acknowledgments We are grateful to our reviewers and the editors of the journal for providing valuable comments and feedback. 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 (GALATEAS), the 7th Framework Program of the European Commission, grant agreements no. 258191 (PROMISE) and no. 288024 (LiMoSINe), the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.-061.-814, 612.-061.-

815, 640.-004.-802, 380-70-011, the Center for Creation, Content and Technology (CCCT), the Hyperlocal Service Platform project funded by the Service Innovation & ICT program, the WAHSP project funded by the CLARIN-nl program, under COMMIT project Infiniti, and by the ESF Research Network Program ELIAS.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G. (2008). Finding high-quality content in social media. In *Proceedings of the 1st ACM international conference on web search and data mining (WSDM 2008)*, (pp. 183–194), New York, NY: ACM.

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

Baeza-Yates, R., & Ribeiro-Neto, B. (2010). *Modern information retrieval: The concepts and technology behind search*. Boston: Addison Wesley.

Castillo, C., Mendoza, M., & Poblete, B. (2011) Information credibility on twitter. In *Proceedings of the 20th international conference on World Wide Web (WWW 2011)* (pp. 675–684), New York, NY: ACM.

Chafe, W. (1986). Evidentiality in English conversion and academic writing. In W. Chaf & J. Nichols (Eds.), *Evidentiality: The linguistic coding of epistemology* (Vol. 20, pp. 261–273). New York: Ablex Publishing Corporation.

Chen, M., & Ohta, T. (2010). Using blog content depth and breadth to access and classify blogs. *International Journal of Business and Information, 5*(1), 26–45.

Croft, W. B., & Lafferty, J. (Eds.) (2003). *Language modeling for information retrieval*. Berlin: Kluwer.

Diakopoulos, N., & Essa, I. (2010). Modulating video credibility via visualization of quality evaluations. In *Proceedings of the 4th workshop on information credibility on the Web (WICOW 2010)* (pp. 75–82). New York, NY: ACM.

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

Hawking, D., & Craswell, N. (2002). Overview of the TREC-2001 Web track. In *The 10th text REtrieval conference proceedings (TREC 2001)*. Gaithersburg, USA: NIST.

He, J., Larson, M., & de Rijke, M. (2008). Using coherence-based measures to predict query difficulty. In *Advances in information retrieval—30th European conference on IR research (ECIR 2008), volume 4956 of Lecture Notes in Computer Science* (pp. 689–694). Berlin: Springer.

He, J., Weerkamp, W., Larson, M., & de Rijke, M. (2009). An effective coherence measure to determine topical consistency in user generated content. *International Journal on Document Analysis and Recognition, 12*(3), 185–203.

Hearst, M. A., & Dumais, S. T. (2009). Blogging together: An examination of group blogs. In *Proceedings of the 3rd international conference on weblogs and social media (ICWSM 2009)*. Menlo Park: AAAI Press.

Hofmann, K., & Weerkamp, W. (2008). Content extraction for information retrieval in blogs and intranets. Technical report, University of Amsterdam, 2008. URL http://ilps.science.uva.nl/biblio/content-extraction-information-retrieval-blogs-and-intranet.

Java, A., Kolari, P., Finin, T., Joshi, A., & Martineau, J. (2007). The blogVox opinion retrieval system. In *Proceedings of the 15th text REtrieval conference (TREC 2006)*. Gaithersburg, USA: NIST.

Juffinger, A., Granitzer, M., & Lex, E. (2009). Blog credibility ranking by exploiting verified content. In *Proceedings of the 3rd workshop on information credibility on the Web (WICOW 2009)* (pp. 51–58). New York, NY: ACM.

Klewes, J., & Wreschniok, R. (2009). *Reputation capital*. New York: Springer.

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

Lavrenko, V., & Croft, W. B. (2001). Relevance-based language models. In *Proceedings of the 14th international ACM SIGIR conference on research and development in information retrieval (SIGIR 2001)* (pp. 120–127). New York, NY: ACM.

Lee J. H. (1995). Combining multiple evidence from different properties of weighting schemes. In *Proceedings of the 18th international ACM SIGIR conference on research and development in information retrieval (SIGIR 1995)* (pp. 180–188). New York, NY: ACM.

Liu, B. (2007). *Web data mining*. Heidelberg: Springer.

Losada, D. E., & Azzopardi, L. (2008). An analysis on document length retrieval trends in language modeling smoothing. *Information Retrieval Journal, 11*(2), 109–138.

Lu, Y., Tsaparas, P., Ntoulas, A., & Polanyi, L. (2010). Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World Wide Web (WWW 2010)* (pp. 691–700), New York, NY: ACM.

Macdonald, C., & 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., & Soboroff, I. (2008a) Overview of the TREC 2007 blog track. In *Proceedings of the 16th text REtrieval conference (TREC 2007)*. Gaithersburg, USA: NIST.

Macdonald, C., Ounis, I., & Soboroff, I. (2008b) Overview of the TREC 2007 blog track. In *Proceedings of the 16th text REtrieval conference (TREC 2007)*. Gaithersburg, USA: NIST.

Mandl, T. (2006). Implementation and evaluation of a quality-based search engine. In: *Proceedings of the 17th conference on hypertext and hypermedia* (pp. 73–84). New York, NY: ACM.

Manning, C., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.

Massoudi, K., Tsagkias, M., de Rijke, M., & Weerkamp, W. (2011). Incorporating query expansion and quality indicators in searching microblog posts. In: *Advances in information retrieval—33rd European conference on IR research (ECIR 2011), Vol. 6611 of Lecture Notes in Computer Science* (pp. 362–367). Berlin/Heidelberg: Springer.

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

Mishne, G. (2007a) Using blog properties to improve retrieval. In *Proceedings of the 1st international conference on weblogs and social media (ICWSM 2007)*. Menlo Park: AAAI Press.

Mishne, G. (2007b) *Applied text analytics for blogs*. PhD thesis, Amsterdam:University of Amsterdam.

Mishne, G., & de Rijke, M. (2006). A study of blog search. In *Advances in information retrieval—28th European conference on IR research (ECIR 2006), Vol. 3936 of Lecture Notes in Computer Science* (pp. 289–301), Berlin/Heidelberg: Springer.

Mishne, G., & Glance, N. (2006) Leave a reply: An analysis of weblog comments. In *Proceedings of the WWW 2006 workshop on weblogging ecosystem: Aggregation, analysis and dynamics*.

O'Mahony, M. P., & Smyth, B. (2009) Learning to recommend helpful hotel reviews. In *Proceedings of the 3rd ACM conference on recommender systems (RecSys 2009)* (pp. 305–308). New York, NY: ACM.

O'Mahony, M. P., & Smyth, B. (2010). Using readability tests to predict helpful product reviews. In *Proceedings of the 9th international conference on adaptivity, personalization and fusion of heterogeneous information (RIAO 2010)* (pp. 164–167).

Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., & Soboroff, I. (2006). Overview of the TREC-2006 blog track. In *Proceedings of the 15th text REtrieval conference (TREC 2006)*. Gaithersburg, USA: NIST.

Ounis, I., Macdonald, C., & Soboroff, I. (2009). Overview of the TREC-2008 blog track. In *Proceedings of the 7th text REtrieval conference (TREC 2008)*. Gaithersburg, USA: NIST.

Rubin, V., & Liddy, E. (2006). Assessing credibility of weblogs. In *Proceedings of the AAAI spring symposium: Computational approaches to analyzing weblogs (CAAW)*.

Singhal, A., Buckley, C., & Mitra, M. (1996). Pivoted document length normalization. In *Proceedings of the 19th international ACM SIGIR conference on research and development in information retrieval (SIGIR 1996)* (pp. 21–29). New York, NY: ACM.

Sriphaew, K., Takamura, H., & Okumura, M. (2008). Cool blog identification using topic-based models. In *Proceedings of the 2008 IEEE/WIC/ACM international conference on Web intelligence and intelligent agent technology (WIIAT 2008)* (pp. 402–406). Washington, DC: IEEE Computer Society.

Stanford, J., Tauber, E., Fogg, B., & Marable, L. (2002). Experts vs online consumers: A comparative credibility study of health and finance web sites, http://www.consumerwebwatch.org/dynamic/web-credibility-reports-experts-vs-online-abstract.cf.

Su, Q., Huang, C.-R., & Yun Chen, K. (2010). Evidentiality for text trustworthiness detection. In *Proceedings of the 2010 workshop on NLP and linguistics: Finding the common ground (NLPLING 2010)*, (pp. 10–17), Stroudsburg, PA: Association for Computational Linguistics.

Tsagkias, M., Larson, M., & de Rijke M. (2010). Predicting podcast preference: An analysis framework and its application. *Journal of the American Society for Information Science and Technology, 61*(2), 374–391.

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

Weerkamp, W. (2011). *Finding people and their utterances in social media*. PhD thesis, University of Amsterdam.

Weerkamp, W., & de Rijke, M. (2008). Credibility improves topical blog post retrieval. In *Proceedings of the 46th annual meeting of the association for computational linguistics (ACL 2008)* (pp. 923–931), Stroudsburg, PA: Association for Computational Linguistics.

Weerkamp, W., Balog, K., & de Rijke M. (2009). A generative blog post retrieval model that uses query expansion based on external collections. In *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language. Processing of the AFNLP (ACL-IJCNLP 2009)* (pp. 1057–1065). Stroudsburg, PA: Association for Computational Linguistics.

Weerkamp, W., Balog, K., & de Rijke, M. (2011). Blog feed search with a post index. *Information Retrieval Journal, 14*(5), 515–545.

Weimer, M., Gurevych, I., & Mehlhauser, M. (2007). Automatically assessing the post quality in online discussions on software. In *Proceedings of the ACL 2007 demo and poster sessions*, pp. 125–128.