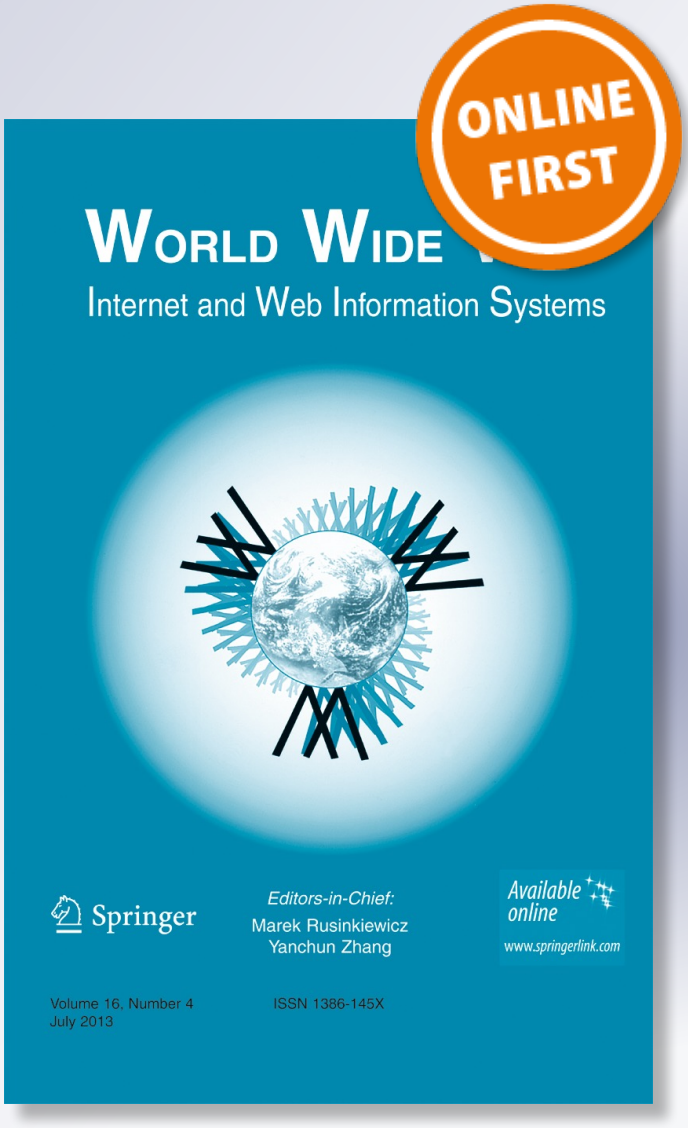
Improving opinionated blog retrieval effectiveness with quality measures and temporal features

Leonidas Akritidis & Panayiotis Bozanis

World Wide Web
Internet and Web Information Systems

ISSN 1386-145X

World Wide Web
DOI 10.1007/s11280-013-0237-1



Springer

World Wide Web
DOI 10.1007/s11280-013-0237-1

Improving opinionated blog retrieval effectiveness with quality measures and temporal features

Leonidas Akritidis · Panayiotis Bozanis

Received: 31 October 2012 / Revised: 25 April 2013 / Accepted: 17 June 2013

Abstract The massive acceptance and usage of the blog communities by a significant portion of the Web users has rendered knowledge extraction from blogs a particularly important research field. One of the most interesting related problems is the issue of the opinionated retrieval, that is, the retrieval of blog entries which contain opinions about a topic. There has been a remarkable amount of work towards the improvement of the effectiveness of the opinion retrieval systems. The primary objective of these systems is to retrieve blog posts which are both relevant to a given query and contain opinions, and generate a ranked list of the retrieved documents according to the relevance and opinion scores. Although a wide variety of effective opinion retrieval methods have been proposed, to the best of our knowledge, none of them takes into consideration the issue of the importance of the retrieved opinions. In this work we introduce a ranking model which combines the existing retrieval strategies with query-independent information to enhance the ranking of the opinionated documents. More specifically, our model accounts for the influence of the blogger who authored an opinion, the reputation of the blog site which published a specific blog post, and the impact of the post itself. Furthermore, we expand the current proximity-based opinion scoring strategies by considering the physical locations of the query and opinion terms within a document. We conduct extensive experiments with the TREC Blogs08 dataset which demonstrate that the application of our methods enhances retrieval precision by a significant margin.

Keywords Information retrieval · Opinionated retrieval · Search · Blog · Post · Blogger · Influence · Impact · Ranking

L. Akritidis (✉) · P. Bozanis
Department of Computer and Communication Engineering,
University of Thessaly, Volos, Greece
e-mail: leoakr@inf.uth.gr

P. Bozanis
e-mail: pbozanis@inf.uth.gr

Published online: 28 June 2013

1 Introduction

The tremendous amount of the information produced and exchanged among the blog users has rendered weblogs a valuable source of knowledge. Blogosphere, the universe which accommodates all blogs, now includes millions of active bloggers and even more readers. In addition, Blogosphere is extremely volatile: it estimated that there are more than 172 million identified blog sites which collectively produce more than 1 million new posts each day.[1] These numbers indicate that the information published and disseminated by the blogs is not only huge, but also, it is accessed by a large number of users.

Within a blog service one or more individuals (the *bloggers*) publish a *post* to express their opinions or experiences about a subject. On the other hand, the readers are allowed to submit their own comments to state their agreement or disagreement to the ideas or opinions contained within the main post. Due to the aforementioned increase in the size of Blogosphere, these opinions are now of crucial importance since they affect a large number of users and their impact is large. For instance, a positive opinion about a product can significantly increase its commercial success whereas in contrast, multiple negative statements about a politician can decrease his/her publicity and affect the success of his/her political career. Similar examples include artists, events, travel locations, service providers, and generally every judgeable aspect of life.

For these reasons, the problem of the opinionated retrieval of blog entries is considered both interesting and challenging and has gained the attention of the research community. In addition, the introduction of the polarity and opinion search task by the Text Retrieval Conference (TREC) in 2006 and 2008 [15, 18, 19] has attracted even more researchers to propose solutions for this problem. In general the suggested opinion retrieval models primarily consist of three basic components: the first one implements a traditional information retrieval (IR) system which identifies topic-relevant documents (i.e. blog posts) from a document set, with respect to a given query. In the sequel, a classification or lexicon-based algorithm is employed to determine whether these posts contain opinions. Finally, a third component assigns opinion scores and combines them with the relevance scores of the IR system to produce a final ranked list of documents.

One of the most challenging issues in opinionated retrieval is to develop an effective method for assigning query-related opinion scores to the documents [10]. The early models did not consider the issue of the opinion relevancy to the query topic; they arbitrarily assumed that each expressed opinion refers to the subject of the query. The most recent approaches addressed this issue by applying either proximity-based strategies or data mining algorithms. However, none of the opinion scores introduced so far embody information which indicates the generic value and impact of the retrieved documents. In this paper we introduce an opinion scoring approach that takes into consideration both query and opinion independent data which indicates the value of the post. Our main motivation is that the opinionated retrieval of blog

[1] http://en.wikipedia.org/wiki/Blogosphere

posts must exploit objective and query independent criteria. Such an improvement would allow an opinion retrieval system to provide rankings which are both relevant and contain *high quality* opinions.

More precisely, the query independent model that we consider in this work is composed by elements which reflect the influence of the blogger who authored each post. The key idea is that an opinion expressed by an influential blogger is apparently more important than the opinion of another blogger who is of lower impact. In this work we also examine the value of the entire blog site which hosts the retrieved opinion. Following a spirit similar to PageRank, we consider that the documents appearing in reputable blog sites are more useful than other posts published in unpopular sites. For the needs of our proposed approach, we examine several metrics which have been proposed for the identification of the influential bloggers and the determination of the value of a blog site. We also introduce two new approaches for the calculation of the value of a blog site, *SBI-Rank* and *Blogs Impact Factor (BIF)*.

Moreover, recent works have demonstrated that the inclusion of proximity information within the ranking component leads to enhanced opinion retrieval effectiveness [10]. Such methods consider that the distance of an opinion term to the query term is a measure of their relatedness; consequently, the opinion terms have stronger connections with the terms which are closer to their position. In this work we extend this idea by taking into consideration the physical location of the document (namely zone, or field) where the query and opinion terms occur. Therefore, our model rewards close term proximities occurring in "important" document fields, such as in the title. Our experiments conducted with the TREC Blogs08 dataset demonstrated that our method outperformed the baseline approach which employs plain IR functions by 39 %, and the term proximity-model of [10] by roughly 7.2 %.

The contributions of this work are summarized in the following list:

- We introduce the idea of assigning query independent quality scores (QUIQS, pronounced "quicks") to the blog posts which are based on the concepts of impact, influence, and time awareness.
- We show how these quality scores can be combined with the existing opinion ranking models to create a new, more effective ranking method.
- Based on the bloggers' productivity and influence metrics, we introduce two methods for the evaluation of a blog site: *Summed Bloggers Influence (SBI-Rank)* and *Blogs Impact Factor (BIF)*.
- We introduce the *Field Opinion Probabilities (FOP)*, an extension which improves the standard opinion probabilities of [10]. This enhanced model takes into consideration not only the proximity of the opinion and query terms, but also, the physical locations (fields or zones) of the document where they occur.
- We measure the performance of our methods by experimenting with the TREC Blogs08 dataset, a repository comprised of approximately 28 million blog posts.

The rest of the paper is organized as follows: In Section 2 we present the state-of-the-art approaches for solving the opinionated retrieval problem. In Section 3 we introduce our query independent scores for the blog posts, bloggers, and blog sites and in Section 4 we show how these scores can be combined effectively with the opinion scores of the literature. Finally, in Section 5 we measure the performance of our proposed model and in Section 6 we end this article by stating our conclusions.

2 Related work

The problem of opinion retrieval and sentiment analysis has attracted the attention of the researchers since 2006, when TREC introduced the polarity and opinion search task [18]. In contrast to the traditional document retrieval, the key problem here is to identify documents which are both relevant to a given query and contain opinion expressions. In [25, 26] the authors adopt a machine learning approach which employs support vector machines to build opinion classifiers. Their proposed system ranks the retrieved documents by computing linearly combined opinion and relevance scores. Support vector machines for sentiment analysis were also used in [16], where the authors attempt to combine diverse sources of potentially pertinent information.

In [17] the authors construct an opinion lexicon with respect to the given query. Their algorithm initially employs a general opinion lexicon which is refined by computing the opinion weights of its words. Moreover, Turney [22] and Turney and Littman [23] study the issue of building lists of subjective words (i.e. *good* against *bad*, or *excellent* against *poor*) with the aim of capturing expressed opinions within a document.

Similarly to some traditional Web ranking models, a number of relevant works takes into account the proximity of the query terms within the retrieved posts to achieve effective ranking [6]. For instance, Zhang et al. [26] computed the probability of query terms and opinion terms co-occurrence by employing a word window. Similarly, Vechtomova [24] considered a word window around each query term and calculated the distance between the query terms and each word in the window. In [7] the authors computed the proximity by employing n-grams and experimented with several machine learning classification methods. The authors in [10] proposed a proximity-based opinion propagation model to calculate the opinion density at each point in a document. In addition, Pang et al. [20] employed supervised machine learning techniques to identify positive and negative reviews of movie films, whereas [22] used special words (such as *poor* and *nice*) and a machine-learning algorithm to achieve sentiment analysis.

Nevertheless, none of the aforementioned approaches take into consideration query-independent information during ranking. In this work we examine how the influence of a post's author and the importance of a blog site can be combined with the aforementioned strategies to enhance ranking. The first work which attempted to identify the influential bloggers in a community is [2], where the authors introduced a post scoring function based on the number of comments and the scores of the incoming and outgoing links. In the sequel, they identified the influential bloggers by the post which received the highest score. Moreover, Akritidis et al. [3, 4] presented numerous methods which take into consideration the temporal aspects of the Blogosphere and the productivity of the bloggers.

In addition, Kritikopoulos et al. [12] introduced BlogRank, a PageRank generalization for ranking weblogs. In this study the authors highlighted the sparseness of the blog graph and detected the inappropriateness of the traditional Web ranking models in Blogosphere. They propose a strategy for enhancing the blog graph with implicit links which are created by considering the participation of some bloggers in multiple social networks. Finally, Tayebi et al. [21] introduced B2Rank, a method for ranking blogs with respect to the behavioral features of the users.

3 Query Independent Quality Scores (QUIQS)

In this work we introduce a scoring model which apart from the established relevance and opinion scores, also takes into account query-independent blog quality information. The key idea is that a blog post not only must be relevant to a given query and contain an opinion, but it also has to be qualitative and highly influential. In other words, a robust opinion retrieval system must consider the issue of the authority of a blog post and rank the important opinions higher.

The authority of a blog entry is reflected by numerous properties: A first important notification is that a blog post inherits the reputation of the author who published it. Therefore, an opinion expressed by an influential blogger is considered more valuable than one published by an individual who is of lower reputation. However, the influence of a blogger is non-static and changes over time [3, 4]. Since the user submits a query at the present time instance, we are mainly interested in measuring the *current* bloggers' influence.

Furthermore, similarly to the original PageRank concept, we consider that the opinions which are published in reputable blog sites are of higher importance than others which are hosted in sites of lower value. One of the objectives of this work is to introduce effective mechanisms for the evaluation of a blog site. Of course, there is a huge amount of research in the traditional Web IR field which proposed eigenvector-based methods for identifying authoritative Web pages, such as PageRank and HITS [13]. However, these methods are not useful to our problem since blog sites in Blogosphere are very sparsely linked [12]. Similarly to the previous occasion, due to the highly dynamic character of Blogosphere, the value of a blog site is not constant. The approaches we present in Section 3.4 adopt the time-aware spirit of the aforementioned blogger influence metrics.

Based on these notifications, we introduce the query-independent quality score (QUIQS) which consists of the following three basic components:

- *The post value:* The importance of a blog post is partially reflected by the impact it has on other bloggers and readers. There are two primary parameters indicating this impact: the number of the Web pages which contain references to the post, and the number of comments submitted by the readers to express their thoughts on the original content. Consequently, since an opinion published in an influential post is accessed by a large number of individuals, we consider it more important than another which was never referenced or commented.
- *The influence of the blogger:* The wide impact of the author who expresses an opinion is a significant factor which affects its importance. Hence, the more readers the writings of a blogger attract, the higher rankings should his/her opinions receive.
- *The impact of the blog site:* The opinions published in a reputable blog site attract a large number of readers and gain more attention. With only a few exceptions, the value of the blog site which hosts a post is a partial indication of the post's value.

In the following subsections we present some state-of-the-art approaches which have been proposed in the literature for identifying influential posts, bloggers and

Table 1 Summary of the used symbols

Symbol	Meaning
A	The set which contains all bloggers
B	The set which contains all blog sites
D	The set which contains all blog posts
a	A blogger $a \in A$
b	A blog site $b \in B$
d	A blog post $d \in D$
D_a	The set which contains all the blog posts of the blogger a
D_b	The set which contains all the blog posts of the site b
C_d	The set of comments to the post d
$D_{c,d}$	The set of posts referring (have a link) to the post d
$D_{r,d}$	The set of posts referenced by the post d
L_d	The length (in words) of the post d
t_d	The time stamp of d
S_d	A score value of the post d
S_b	A score value of the blog b
h_a	A metric for the evaluation of the blogger a

blog sites and we introduce our contributions in the evaluation of a blog site. In Section 4 we examine how these approaches can be combined with the existing opinion retrieval strategies to form a new improved ranking model.

3.1 Preliminaries

Let us begin our analysis by introducing the universe $\mathcal{B}$ which represents Blogosphere. Within Blogosphere we define three sets: The first one is A and contains all bloggers, the second one is B and includes all blog sites, whereas the third one is D and is composed of all the blog entries. From these main sets we identify two important subsets, $D_a \subset D$ which accommodates the blog posts authored by a blogger a, and $D_b \subset D$ which includes the set of posts published by a blog site b.

For each blog entry $d \in D$ we formulate a set of properties which includes: (i) the number of comments C_d submitted by the post readers, (ii) the number of other posts $D_{c,d} \subset D$ and $D_{r,d} \subset D$ referring to and referenced by d, and (iii) t_d which represents the date and time when the post was published expressed as a time stamp.[2] The elapsed time since the creation of a post d (i.e. the age of a post) is symbolized as Δt_d and is expressed in seconds.

Finally, we define three score values h_a, S_d, and S_b: The former reflects the impact of a blogger a, whereas the other two represent the value of a post d and a blog site b respectively. These scores are the foundation upon which we shall build our query independent blog post evaluation mechanisms. All the aforementioned notifications are summarized in Table 1. In the following section we present the current state-of-the-art approaches for the evaluation of these scores.

[2]The time stamp is a 32-bit integer which represents the number of the elapsed seconds since January 1st, 1970.

3.2 Blog post quality scores

The first QUIQS component concerns the value of the blog post which contains the opinion/s of its author. There is a remarkable number of approaches which attempt to evaluate the importance of a blog post by considering several features and properties of the post.

One of the first works which assigned quality scores to a blog post is [2], where the authors introduced a model based on four parameters: recognition (proportional to the incoming links), activity generation (proportional to the number of comments), novelty (inversely proportional to the outgoing links) and eloquence (inversely proportional to the post's length). More specifically, the influence score $S_{\iota,d}$ (which we call ι-score) of a blog post d is determined by the following equation:

$$S_{\iota,d} = w(L_d)\left(w_c C_d + w_{\text{in}} \sum_{m=1}^{|D_{c,d}|} S_{\iota,d_m} - w_{\text{out}} \sum_{n=1}^{|D_{r,d}|} S_{\iota,d_n}\right) \tag{1}$$

where $w(L_d)$ is a weight function depending on the length L_d of a post. The symbol w_c represents a weight which regulates the contribution of the number of comments C_d, whereas w_{in} and w_{out} are weights which adjust the contribution of incoming and outgoing links respectively. The calculation of ι-score is recursive (positive reinforcement from incoming links and negative reinforcement from outgoing links), similar to the PageRank definition.

Nevertheless, ι-score ignores one of the most important factors in Blogosphere: time. According to our earlier notifications, the map in Blogosphere changes rapidly: a high number of new bloggers enter the community each day, whereas millions of posts are being published daily. For this reason, Akritidis et al. [3] introduced two time-aware approaches which are sensible to the temporal aspects of Blogosphere and identify the bloggers who are *presently* influential. The first one, called $MEIBI$, assigns to a post d a time-decaying score determined by the following formula:

$$S_{M,d} = \gamma(C_d + 1)|D_{c,d}|\left(\frac{\theta}{t - t_d + \theta}\right)^{\delta} \tag{2}$$

where t represents the current time stamp. The coefficients γ and δ are constant parameters usually set equal to 4 and 1 respectively. Note that δ does not affect the relative score values, but it is used to determine how quickly the older posts decay (i.e. lose their importance). The MEIBI scores take into consideration popularity statistics (i.e. number of incoming links and number of comments), however, the importance of a post gradually decreases over time.

The second approach, namely MEIBIX, assumes that an old blog post could still be of some importance if it continues to attract references. The key idea is that in case a post is not cited anymore, then it potentially negotiates outdated topics or proposes outdated solutions. On the other hand, if an old post continues to be linked presently, then this is an indication that it contains important and highly influential material. This idea is quantified into the following equation:

$$S_{MX,d} = \gamma(C_d + 1) \sum_{\forall r \in D_{c,d}} \left(\frac{\theta}{t - t_r + \theta}\right)^{\delta} \tag{3}$$

In [4] the authors introduced two additional time-sensitive methods which are linearly combined to identify the bloggers who are both productive and influential. Regarding the first method, called *BP (Blogger's Productivity)*, they adopt a strategy similar to the one of MEIBI; that is, they consider that the value of a blog post gradually degrades as time flows. Equation (4) implements this strategy:

$$S_{BP,d} = \gamma \frac{L_d}{\overline{L}} \left(\frac{\theta}{t - t_d + \theta} \right)^{\delta} \tag{4}$$

where L_d represents the post length in words and $\overline{L}$ is the average post length of the entire collection. Notice that the BP-scores totally ignore the impact of a post (i.e. comments and incoming links). The impact of a post is reflected in the second method, BI. The cornerstone of BI-score is the observation that the influence of a blog post has a dual nature and consists of the proximal impact (denoted by the comments it receives), and the wide impact which is expressed by the number and the age of the incoming links. Equation (5) captures this dual nature by assigning to each blog post a score determined by both the comments and the inlinks:

$$S_{BI,d} = w_l \sum_{\forall r \in D_{c,d}} \left(\frac{\theta}{t - t_r + \theta} \right)^{\delta} + w_c \sum_{\forall c \in C_d} \left(\frac{\theta}{t - t_c + \theta} \right)^{\delta} \tag{5}$$

The parameters w_l and w_c are used for two reasons; (i) they grant the score $S_{BI,d}$ a reasonably large value and (ii) they regulate the balance between the comments and the inlinks. A typical setting is $w_l = 100$ and $w_c = 10$, which means that each incoming link is considered as important as ten user comments. The metrics presented in this section are summarized in the first column of Table 2.

3.3 Bloggers influence scores

Intuitively, the influence of a blogger propagates through the impact of his/her writings. Based on this notification, most of the relevant works quantify the blogger's influence by employing the post quality scores of the previous subsection. More specifically, we primarily encounter two different approaches: the first one was introduced in [2] and suggests that we compute the ι-scores of all posts of a blogger and we select the one which received the highest score $\max(S_{\iota,d})$. This post determines ι-index, which represents the overall influence of a blogger.

Nevertheless, isolating a single post to identify whether a blogger is influential or not, is an over-simplistic approach, and so it would be if they have used gross metrics, like average, median and so on. A blogger may have published only a handful of influential posts and numerous others of low quality, whereas other bloggers may have published several tens of influential blogs with ι-scores lower than the ι-score

Table 2 Metrics for the evaluation of a blogger's influence and productivity

	Metric	Symbol	Post scores
1	ι-index	$h_{\iota,a}$	ι-scores (1)
2	MEIBI	$h_{M,a}$	$MEIBI$-scores (2)
3	MEIBIX	$h_{MX,a}$	$MEIBIX$-scores (3)
4	BP-index	$h_{BP,a}$	BP-scores (4)
5	BI-index	$h_{BI,a}$	BI-scores (5)

of the most influential blog of the former blogger. Therefore, the productivity of bloggers is a significant issue that has been overlooked by this preliminary model.

To overcome this drawback, Akritidis et al. [3] adopted an approach similar to the one applied for calculation of the h-index for scientists [11]. According to this method, we formulate the following definition:

Definition 1 A blogger a is of value h_a, if h_a of his D_a posts get a score $S_d \geq h_a$, and the rest $D_a - h_a$ posts get a score $S_d < h_a$.

This definition initially rewards the productivity of a blogger; it does not isolate a single (the best) blog post, but it takes into consideration the set of posts which received the highest scores. Consequently, a blogger will be influential if s/he has posted several posts of high quality. This definition supports all the post quality scores of the previous subsection. For instance, the definition of the BI-index of a blogger is phrased as follows:

Definition 2 A blogger a is has BI-index $h_{BI,a}$, if $h_{BI,a}$ of his D_a posts get a score $S_{BI,d} \geq h_{BI,a}$, and the rest $D_a - h_{BI,a}$ posts get a score $S_{BI,d} < h_{BI,a}$.

Based on the scores of the previous subsection and the Definition 1, five blogger evaluation metrics can be formed. Table 2 contains all the possible metrics which can be formulated for the identification of the influential bloggers. Notice that in case we use a time-sensitive post evaluation metric (such as MEIBI or MEIBIX), the influence of a blogger will also be time-sensitive. Hence, Definition 2 dictates that a blogger will be presently influential, if his/her posts are presently influential.

3.4 Blog site quality scores

Although the generic problem of evaluating a Web site has been extensively studied in the past and many effective solutions exist now (such as PageRank and its variants, or HITS), the issue of the evaluation of a blog site is still an open research problem. This is due to the inappropriateness of the aforementioned eigenvector-based methods for two reasons [12]: At first, these methods treat blogs as typical interlinked Web pages and do not consider blog-specific information. And second, the blog entries graph is very sparse, hence, the performance of the traditional Web ranking methods is decreased.

In [12] the authors presented BlogRank, a method for ranking weblogs based on the link graph and on several similarity characteristics between weblogs. In this method, the blog graph is initially enhanced with implicit links, that is, virtual links which reveal the participation of the bloggers and the commentators in multiple blog services. In the sequel, they define the BlogRank $S_{BR,b}$ of a blog site which has N incoming links in a PageRank recursive manner:

$$S_{BR,b} = (1 - E) + E \sum_{n=1}^{N} P(n \rightarrow b) S_{BR,n} \tag{6}$$

where E is a damping factor with a typical value equal to 0.85. $P(n \rightarrow b)$ represents the probability that a user who is currently in the blog site n will visit b. In the original PageRank method, we set $P(n \rightarrow b) = 1/M$ where M is the total number of outlinks of n; that is, the user may follow each outlink with equal probability. However, in BlogRank it is assumed that a user has a higher probability to visit a post which shares

the same topic with n. For this reason, the blog graph is further enhanced by adding bidirectional links between the blogs which share the same thematic categories.

Now let us introduce our methods for evaluating a blog site. The first method is based on the idea that *a blog site is as important as its member bloggers are*. In other words, the impact of a blog site is determined by the influence of the blogger/s who publish posts in this specific site. For instance, a blog community which includes numerous influential bloggers is apparently of high importance. Based on this idea, we introduce *SBI-Rank (Summed Bloggers Influence)*, a metric which accumulates the influences of the member bloggers of a blog site into a single quantity $S_{SBI,b}$:

$$S_{SBI,b} = \sum_{\forall a \in A_b} h_a \tag{7}$$

Within Blogosphere we encounter two types of blogs [1]: (a) the community blogs, or multi-authored blogs, where several bloggers may start discussions, and (b) the individual blogs, maintained and updated by one blogger. In the latter case, the SBI-Rank of a blog site is identical to the influence score of its unique author. Furthermore, if the influence metric h_a is time-aware (i.e. MEIBI, MEIBIX, etc), then SBI-Rank is also time-aware since it rewards blog sites which include presently influential authors.

Our second blog evaluation method embodies the spirit of the *journal impact factor (IF)* [9], a metric which ranks scientific journals. The *IF* is based on the average number of citations received by each article of a journal within a given time period. More specifically, if the *IF* of a journal in a year Y is k, then the articles published in the years $Y-1$ and $Y-2$ received on average k citations in the year Y.

Nevertheless, this metric is impractical for evaluating blog sites for two reasons: The first one is that in contrast to the research papers, the vast majority of the blog posts become old very quickly [3]. Consequently, the posts published one or two years ago are probably never read or referenced in the present. The second reason is that *IF* is only based on citations and does not account for the user comments which are also an indication of a post's impact. We can easily overcome the first problem by replacing the years with a smaller time window (i.e. month or week). Regarding the second issue, we remind that the impact of a blog post has a dual nature, reflected by the number of incoming links and the comments. To integrate this dual nature into our analysis, we introduce the *impact units*, a linear combination between the inlinks and the comments.

$$I_d = w_r D_{c,d} + w_c C_d \tag{8}$$

where w_r and w_p are two constants similar to the ones used in the BI-scores which regulate the balance between the inlinks and the comments. Now, we are ready to introduce our second blog quality metric, *Blog Impact Factor (BIF)*, which is formally phrased as follows:

Definition 3 A blog site b has BIF equal to $S_{BIF,b}$ in a time window w, if the posts published in the two previous windows $w-1$ and $w-2$ received on average $S_{BIF,b}$ impact units within w.

For instance, in case a blog site b has $S_{BIF,b} = 5$ in March, then the posts published in the two previous months (January and February) received on average 5 impact units during March.

4 Combining opinion and relevance scores with QUIQS

In this section we examine how the proposed blog site quality scores and the other aforementioned query-independent metrics can be combined with the relevance and opinion scores into a single ranking model. Given a query q and a document set D, our objective is to retrieve a subset of documents $D' \subset D$ which are both relevant to q and contain opinions (*Relevant Opinionative Documents, ROD* [26]).

Initially, the query is processed by a Web IR system which identifies the relevant documents and assigns scores by treating them as typical Web pages. The score $\mathcal{S}(d, q)$ of a blog post d with respect to the query q can be computed by using any of the well-established Web ranking functions such as the inverse document frequency idf, BM25, BM25F, BM25TP [6], etc. The retrieval effectiveness can be enhanced by applying query pre-processing algorithms which (a) attempt to identify concepts within the query, and (b) expand the query with the aim of extending the pool of relevant documents [26]. Query expansion algorithms may include dictionary-based methods which utilize external sites such as Wikipedia, or local context analysis. In this paper we do not study in depth such query pre-processing approaches; we primarily focus on demonstrating the significance of QUIQS in opinionated retrieval.

More specifically, we introduce a new type of scores, $\mathcal{S}_{QI}(d, a, b)$, which indicate the query-independent quality of a blog entry authored by a blogger a and published in a blog site b. Based on our previous discussion, the scores $\mathcal{S}_{QI}$ are expressed as a linear combination of the entire blog site quality S_b, the blogger's influence score h_a, and the overall value of the blog entry S_d. The following equation captures these features:

$$\mathcal{S}_{QI}(d, a, b) = \mathcal{W}_b S_b + \mathcal{W}_a h_a + \mathcal{W}_d S_d \tag{9}$$

where $\mathcal{W}_b$, $\mathcal{W}_a$, and $\mathcal{W}_d$ are three constants used to adjust the importance of the blog site score, the influence score of the blogger, and post quality score respectively. Equation (9) dictates that the overall quality of a blog entry d depends on the influence of its author and the importance of the blog site which published it. Furthermore, the query-independent features of the post in question (i.e. length, number of incoming links and comments) are also considered.

In the sequel, the $\mathcal{S}(d, q)$ and $\mathcal{S}_{QI}(d, a, b)$ are combined to form the final score a candidate post d receives with respect to the query q:

$$\mathcal{S}_{IR}(d, a, b, q) = \mathcal{W}\mathcal{S}(d, q) + (1 - \mathcal{W})\mathcal{S}_{QI}(d, a, b) \tag{10}$$

where $\mathcal{W}$ is a constant parameter which tunes the contribution of the query-dependent and query-independent scores in the overall score of the post. A typical setting which works well in most cases is $\mathcal{W} = 0.8$.

Equation (10) determines the relevance score of a post d with respect to q, however, it is still required that we choose a strategy to compute the opinion score of d. In [10] the authors have shown that the proximity of the query and opinion terms leads to significant gains in retrieval effectiveness. Their model is built upon the idea that an opinion term refers with higher probability to the terms which are located near its position. This method initially considers each document as a

vector $d = (t_1, ..., t_i, ..., t_j, ..., t_{|p|})$ and introduces an opinion probability score at each position i of the document, given by the equation:

$$P(i, d) = \sum_{j=1}^{|d|} p(t_j, d)\, p(j, i, d) \tag{11}$$

where $p(t_j, d)$ is the opinion score of term t_j at the position j of the document. In addition, $p(j, i, d)$ denotes the probability that the term at the position j refers to the the term in the position i, and is calculated as follows:

$$p(j, i, d) = \frac{k(j, i)}{\sum_{j'=1}^{|d|} k(j', i)} \tag{12}$$

where $k(i, j)$ is a non-increasing distance kernel function such as Gaussian or Laplacian which implements the concept that the closer to an opinionated term a query term is, the greater the probability that the opinion refers to this term is. For instance, consider the opinion term *excellent*. In case a term *device* appears right after *excellent*, then the expressed opinion refers to *device* with great probability. In the opposite case where *excellent* and *device* appear in distant locations, then *excellent* may refer to another topic within the post.

Based on the positional opinion scores of (11), the overall probability that the document d expresses and opinion about the query q is calculated as follows:

$$\mathcal{S}_O(d, q) = \frac{1}{|Q|} \sum_{i \in Q} P(i, d) \tag{13}$$

where Q symbolizes a set which contains all the positions of the query terms of q within d. Equation (13) indicates that to estimate the probability that d contains an opinion about q, it is required that we compute the opinion probabilities in the positions where the query terms occur within d. Nevertheless, this model ignores the physical locations of the document where the opinion and query terms may occur. For instance, a blog post which contains an opinion term and the query terms in proximal positions in its title is more important than another which contains them in distant locations within its main body. For this reason, we provide an enhancement of the opinion probabilities of (13) by introducing the *field opinion probabilities (FOP)*:

$$\mathcal{S}'_O(d, q) = \sum_{\forall z \in d} \frac{\mathcal{K}_z}{|Q_z|} \sum_{i \in Q_t} P(i, d) \tag{14}$$

where $\mathcal{K}_z$ is a constant weight parameter which denotes the value of a particular document zone z. Moreover, Q_z symbolizes a set which contains all the positions of the query terms of q within the zone z of d. We experimentally demonstrate later, the FOP extension provides a significant improvement in the retrieval effectiveness over the standard opinion probabilities model; consequently, it is an important contribution.

Finally, the opinion and IR scores of (10) and (14) are factorized in the final scoring function which is given by the following formula:

$$\mathcal{S}(d, a, b, q) = \mathcal{S}_{IR}(d, a, b, q)\mathcal{S}'_O(d, q) \tag{15}$$

5 Experiments

In this section we provide the experimental analysis of our methods. Initially, we provide a brief description of the employed dataset, and we present important implementation details which allow us to apply QUIQS efficiently during query processing. In the sequel, we present measurements which indicate that the inclusion of quality-based query independent scores in opinion retrieval leads to significant performance benefits.

5.1 Dataset characteristics and processing

The dataset we used is the TREC blogs08, a repository comprised of approximately 28.5 million blog posts (documents or permalinks) and 1.3 million blog feeds. The permalinks and the feeds occupy roughly 1,445 GB and 808 GB in uncompressed forms respectively.

Now let us describe the methodology of processing the dataset in order to compute the scores of Sections 3 and 4. Ideally, the most efficient approach dictates that we pre-compute for each blog post the author, blog site, and post QUIQS. In the sequel, it is only required to maintain these scores into an in-memory data structure which will allow us to quickly retrieve these scores during query processing and compute the desired opinion scores. In Table 3 we show a sample record of the aforementioned data structure. In particular, for each blog entry we store:

- An integer document identifier (DocID), which is identical to the one we use to represent the document during inverted index construction.
- An internal identifier assigned by the dataset authors (TREC-ID), which will be used for our own evaluation purposes (i.e. to compare our results with the ones provided by TREC). Of course, TREC-ID can be omitted in real-world implementations.
- The three QUIQS, and
- A pointer which stores the location of the document's full text in the repository. The full text of the post will be used in the second phase of the retrieval by the opinion classifier, to identify whether there are any opinions expressed within the post, or not.

An simple yet very efficient solution is to implement this data structure by employing a standard table indexed and sorted by ascending Document ID; hence,

Table 3 Required metadata for computing QUIQS: For each post we store its integer identifier, the internal TREC identifier, the three QUIQS for the author, the blog site, and the post itself, and a pointer value which stores the location of the full text of the post

	Field
1	DocumentID
2	TREC-ID
3	Author score
4	Blog site score
5	Post score
6	Pointer

in case we need to compute the score for a document d, we merely need to access the data stored in the record $d - 1$.

The construction of the data records of Table 3 requires the preprocessing of the dataset and the extraction of numerous statistics such as its length, the numbers of inlinks, outlinks, comments, etc. From now on, we collectively refer to these statistics as the post's *metadata*. In Table 4 we record the metadata required to compute all QUIQS, whereas in the second column we show the source from where we retrieve the required data. The term "*directly*" indicates that the metadata in question can be directly extracted by accessing and processing the text of the post; "*after processing*" denotes that we obtain the desired information by a special merging procedure which must take place after text processing. Finally, the indication "*feed*" states that the respective metadata can only be retrieved by accessing the corresponding feed file.

The dataset is organized in individual files which contain one thousand of posts each. Initially, our text processor extracts the posts out of each file, and assigns one unique successive integer identifier to each post. In addition, it retrieves the TREC-ID value, the URL, the time stamp, the length (in words) and the number of outlinks of each document. Then, for each of these files, it outputs three structures: a small inverted index for these 1,000 posts, an array which stores the metadata of each post (similar to the one illustrated in Table 4), and a Web graph in the form of *(URL, DocID, number of inlinks, list of [inlinkID, timestamp] pairs)*. Notice that the inlinks list stores not only the identifier of the document referring to a specific URL, but also, its time stamp. This is necessary for the calculation of the MEIBIX and BI-scores because these methods involve computations of the ages of the incoming links.

After the text processing is completed, a merging procedure firstly merges all the small metadata arrays into a single metadata table which also has the form of Table 4. In the sequel, a second procedure merges all the partial Web graphs into the complete Web graph of the dataset. In total, the Web graph of the blogs08 dataset consists of about 571 million vertices (URLs) and 1.05 billion edges (links), leading to an average of 1.84 incoming links for each encountered URL. During the Web graph merging we also store within the metadata structure the number of the incoming links of each document and a pointer value which points to the respective inlinks list.

Finally, a third application merges the partial indexes into the final inverted index structure. For the needs of our experimental evaluation we adopted the block-based

Table 4 Intermediate metadata required to construct the structure of Table 3

	Field	Source
1	DocumentID	Directly
2	TREC-ID	Directly
3	Feed-TREC-ID	Directly
4	Author	Feed
5	Blog site	Directly
6	Time stamp	Directly
7	Number of comments	Feed
8	Length (in words)	Directly
9	Number of outlinks	Directly
10	Number of inlinks	After processing
11	Pointer to the text	Directly
12	Pointer to the inlinks	After processing

index setup introduced in [5] which apart from the positional data, it also stores zone information within each posting. This scheme allows us (i) to apply our proposed field opinion probabilities (FOP) which expand the proximity-based retrieval model of [10], and (ii) to use more sophisticated ranking functions which combine term proximity with zone scoring, such as BM25FTOP. The final merged inverted index structure that we constructed occupied in total roughly 71.8 GB. It consisted of a lexicon with 17,329,126 unique terms, accompanied by an inverted file comprised of 11,693,508,871 postings.

Nevertheless, we are still missing two pieces of information: the author of a blog post, and the number of comments submitted by its readers. Although in most cases this information is present within the text of the document, it is quite impossible to retrieve it without any errors due to the differences in pages formatting, encoding, and languages. For this reason, we choose to access the corresponding XML feed which accompanies the dataset and locate the desired data for each blog post. This strategy ensures both maximum effectiveness and comfort, since we do not have to develop complex and costly data mining algorithms to process the text of each Web page. However, for a percentage of the examined posts the author information was not available. In these occasions, we considered that the blog entries were published by individual sites (i.e. not community blogs) and we set the author name equal to the name of the site.

After the creation of the metadata Table 4 we can easily generate the required metadata of Table 3. Therefore, we successively scan each row by applying the equations of Section 3 and we compute each of desired QUIQS. Notice that some QUIQS may involve a second processing step; for instance, SBI-Rank for blog sites requires that we compute in advance the influence metrics of all bloggers, and then sum up the metric values of all member authors to obtain $S_{SBI,b}$.

Before we proceed to the presentation of our experimental results, it is required that we clarify two additional points which are considered crucial for the computation of QUIQS. The first one concerns the avoidance of synonymies among different bloggers, a problem which could lead to distorted QUIQS evaluation. We resolved this issue by formatting the author names in two parts separated by an "at" sign (@): the prefix reveals the blogger's true name, whereas the suffix holds the name of the blog site on which the author submits his/her writings. This technique is more preferable for distinguishing bloggers than the one adopted in [12] which simply discards common blogger names such as *admin*, *John*, etc.

The second point concerns the selection of the current date. In all the experiments that we conduct here, we consider that the present date is February 15th, 2009. This date is two weeks beyond the last crawl date of the dataset and it was selected instead of the real current date, since in the opposite case our time decaying metrics would assign near-zero values to all the involved scores.

5.2 Blog site rankings

In this subsection we describe the experimental measurements of QUIQS for blog sites proposed in this work, and we present some representative rankings. These rankings demonstrate the differences between BIF and SBI-Rank and they verify the theoretical elements of Section 3. Furthermore, they confirm that the computation of QUIQS is applicable to large scale data sets.

Table 5 Blogs impact rankings according to: BIF (*top*) and SBI-Rank based on MEIBI (*bottom*)

Blog	$S_{BIF,b}$
blogs.adobe.com	23
www.thehindu.com	22
www.planetmysql.org	20
www.mvblogs.org	15
sportsblogs.org	15
planetsun.org	14
planet.haskell.org	14
www.finextra.com	11
www.planetnetbeans.org	11
www.businessweek.com	11

Blog	$S_{SBI,b}$
sportsblogs.org	1546
planetsun.org	1499
www.autosport.com	1335
fashionplanet.worldofSL.com	1009
www.order-order.com	1007
www.libertaddigital.com	878
www.mvblogs.org	780
minagi.akari-house.net	725
www.golem.de	659
www.thehindubusinessline.com	656

In Table 5 we present two rankings with the most qualitative blog sites of the dataset according to the proposed BIF (top table) and SBI-Rank (bottom table). The former rewards the blog sites which contain recently referenced post posts and according to it, the site having the greatest impact is http://www.newyorker.com; the posts published on November and December of 2008 attracted on average 23 incoming links on January 2009. The second and third most influential blog sites on January 2009 were www.thehindu.com and www.planetmysql.org with 22 and 20 incoming links per post respectively. Now regarding SBI-Rank which is based on the influence of the member of bloggers of a community, we used the MEIBI index for our calculations. The corresponding ranking shows that according to this metric, the most influential blog site is sportsblogs.org with $S_{SBI,b} = 1{,}546$, followed by planetsun.org and www.autosport.com ($S_{SBI,b} = 1{,}499$ and $S_{SBI,b} = 1{,}335$ respectively).

5.3 Retrieval effectiveness

In this subsection we present measurements of the retrieval effectiveness of our proposed methods against a set of adversary approaches. For the needs of this experiment we employed a set of 20 opinionated queries used in the blog retrieval task of TREC 2009.[3] Each query of our test set is accompanied by the corresponding "qrels" file which contains the documents which are (1) relevant, (2) both relevant and opinionated and (3) both relevant and factual.

[3]The query set of TREC 2009 is comprised of 50 queries, however, only 21 of them are about opinionated retrieval. For one query out of these 21 queries, TREC does not supply the relevant opinionated documents.

Our experimentation was divided in four phases: During the first phase, we applied only traditional IR functions including BM25, a popular expansion which integrates document zones (BM25F), and a second expansion, BM25FTOP, which combines term-proximity weighting, query term ordering, and zone scoring into a single ranking formula [5]. On the second and third phases, we combined the relevance scores of the first phase with the term-proximity model of [10] (marked as TPM) and our proposed field opinion scores of (14) (FOS) respectively. Finally, on the last stage we attested the usefulness of our query-independent metrics by applying five different QUIQS scenarios.

The results of all these experimental phases are presented in Table 8. Each output ranking was evaluated by calculating three different measures: mean average precision (MAP), R-precision (R-Prec), and precision at the first 10 documents of the ranked list only (p@10). For the first measure we also record the confidence intervals to indicate the importance of the differences achieved using any of our examined methods. The various experimental phases are separated from the others by a horizontal line.

5.3.1 Baseline methods

In this subsection we describe the baseline methods which will be used to evaluate the performance of our proposed methods. The first three rows of Table 8 contain measurements of the retrieval effectiveness of the standard IR functions. These results are the first baseline of our experimentation and our goal is to improve their performance. As expected, BM25F outperformed the traditional BM25 method by a margin of about 4.8 % in terms of MAP. The best-performing scheme was BM25FTOP; its MAP was approximately 6.5 % and 11 % higher than the ones of BM25F and BM25 respectively.

In the sequel, we implemented the term-proximity opinion retrieval model of [10] which is the second adversary approach for our methods. Notice that in order to apply any of the opinion retrieval strategies, it is required that we possess a special opinion lexicon which contains specific sentiment expressing words. In our experiments we utilized a lexicon structure which is based on multiple product reviews and data sheets submitted in the Amazon.com Web service. It was constructed by using senti-WordNet [8], a publicly available resource for opinion mining which assigns to each three sentiment scores: positivity, negativity, and objectivity. The lexicon in question has been proved particularly effective for the requirements of the TREC 2008 blog track [14].

Moreover, since the creators of the term-proximity model report that their approach achieved optimal performance by using the Laplacian kernel function, we replace $k(i, j)$ in (12) by the following quantity:

$$k(i, j) = \frac{1}{2b} exp\left[\frac{-|i - j|}{b}\right] \tag{16}$$

where b is a parameter which we set equal to $6\sqrt{2}$, a value which maximizes the retrieval effectiveness of the Laplacian kernel. The lines 4–6 of Table 8 indicate that the term-proximity model achieved significant improvements in all measurements over the simple IR functions. More specifically, when combined with BM25, it achieves a MAP equal to 0.0698 and performs approximately 14.5 % higher than the best IR function alone (BM25FTOP). However, if we compare the same IR functions

with and without the term proximity model we observe a total performance increase which ranges between 24 % (for BM25F) and 26 % (for BM25).

5.3.2 Field opinion scores

Now let us examine the effectiveness of our proposed Field Opinion Scores (FOS) which expand the aforementioned model. The Eq. (14) dictates that each blog post must be divided into a number of distinct zones; then, the occurrence of opinion and query terms in each document zone is weighted accordingly. In this work we fragmented the blog entries of our dataset into five zones: title, URL, body, anchor text, and text encountered within headings. Each zone was assigned a weight value according to Table 6, a scheme which provides satisfactory IR retrieval as mentioned in [4].

The third part of Table 8 (rows 7–9) records the performance of FOS. In total, the field expansion of the term-proximity model leads to an enhancement of about 1.5–2 %. More specifically, the mean average precision in case FOS is combined with BM25, increases by about 1.8 %, whereas the combination with BM25FTOP leads to an enhancement of 1.9 %. These gains are also noticeable for the other evaluation metrics, i.e., R-Precision and Precision@10.

5.3.3 Retrieval effectiveness with QUIQS

In this subsection we measure the performance of our opinionated blog retrieval system when QUIQS are integrated into its ranking mechanism.

Initially, we notice that the QUIQS can be combined in multiple ways and the number of the possible combinations is large. Nevertheless, we can limit these combinations by taking into consideration the factor of time-sensitivity. For instance, in case we use MEIBI or MEIBIX scores to evaluate the quality of a blog post, we promote the opinions which are *presently* influential. On the other hand, the ι-scores do not provide such abilities. MEIBI, MEIBIX and the BP/BI-index also exhibit this advantage over ι-index, and furthermore, they also reward productivity. For the same reason, we believe that the introduced BIF and SBI-Rank should be preferred over the the time-insensitive BlogRank. In this work we have experimented with a wide variety of such combinations; we choose the five highest performing scenarios with the aim of covering all the introduced QUIQS.

In the first scenario, namely $Q1$, we use the MEIBI-scores $S_{M,d}$ for the query-independent evaluation of a blog post, MEIBI-index $h_{M,a}$ for the bloggers' influence, and BIF for the blog site impact. In the second scheme, we use MEIBIX scores, MEIBIX-index and BIF respectively. The third and fourth settings dictate that we employ the MEIBI-scores and MEIBI-index for the estimation of the posts' value and the blogger's influence; For the computation of the impact of the communities

Table 6 The zone weighting scheme for the field opinion scores

Zone	K_z
Body (normal text)	1
Anchor text	1
Title	6
URL	2
Headings	4

Table 7 Three example QUIQS combinations applied for opinionated retrieval evaluation

	Symbol	Meaning
1	$Q1$	Posts: $S_{M,d}$ - Author: $h_{M,a}$ - Blog Site: $S_{BIF,b}$
2	$Q2$	Posts: $S_{MX,d}$ - Author: $h_{MX,a}$ - Blog Site: $S_{BIF,b}$
3	$Q3$	Posts: $S_{M,d}$ - Author: $h_{M,a}$ - Blog Site: $S_{SBI,b}$
4	$Q4$	Posts: $S_{M,d}$ - Author: $h_{M,a}$ - Blog Site: $S_{BR,b}$
5	$Q5$	Posts: $S_{\iota,d}$ - Author: $h_{\iota,a}$ - Blog Site: $S_{SBI,b}$

we use SBI-Rank ($Q3$) and BlogRank ($Q4$) respectively. Finally, in the last scenario we measure the retrieval effectiveness by using the ι-score for the posts, the ι-index for the bloggers' influence and SBI-Rank for the blog community importance. All these combinations are summarized in Table 7.

Regarding the weights of (9), we also experimented with multiple setups; A representative example which performs reasonably well in all five scenarios is to set $\mathcal{W}_d = 0.3, \mathcal{W}_a = 0.2$, and $\mathcal{W}_b = 0.5$. This setting indicates that among all QUIQS, the one with the most significant contribution in ranking is the objective quality of the blog site which hosts the opinionated post, whereas the objective post quality and the influence of its author is approximately of equal importance.

Now let us discuss the retrieval performance of the five QUIQS scenarios of Table 7. In rows 10–24 of Table 8 we report the values of all three evaluation metrics

Table 8 Evaluation of the retrieval effectiveness using different ranking methods

	Method	MAP (95 %-Conf.)	R-Prec	p@10
IR	BM25	0.0531 (0.0440, 0.0623)	0.0828	0.0650
	BM25F	0.0558 (0.0463, 0.0652)	0.0844	0.0677
	BM25FTOP	**0.0597 (0.0525, 0.0669)**	**0.0903**	**0.0742**
IR + TPM	BM25+TPM	0.0698 (0.0617, 0.0779)	0.1066	0.0857
	BM25F+TPM	0.0724 (0.0644, 0.0804)	0.1099	0.0881
	BM25FTOP+TPM	**0.0808 (0.0739, 0.0877)**	**0.1187**	**0.0974**
IR + FOS	BM25+FOS	0.0711 (0.0636, 0.0786)	0.1078	0.0868
	BM25F+FOS	0.0733 (0.0652, 0.0814)	0.1113	0.0895
	BM25FTOP+FOS	**0.0823 (0.0758, 0.0888)**	**0.1207**	**0.0991**
IR + FOS + QUIQS	BM25+FOS+$Q1$	0.0751 (0.0684, 0.0818)	0.1139	0.0918
	BM25F+FOS+$Q1$	0.0774 (0.0702, 0.0846)	0.1177	0.0947
	BM25FTOP+FOS+$Q1$	0.0868 (0.0801, 0.0935)	0.1275	**0.1048**
	BM25+FOS+$Q2$	0.0748 (0.0680, 0.0816)	0.1139	0.0918
	BM25F+FOS+$Q2$	0.0763 (0.0698, 0.0828)	0.1165	0.0935
	BM25FTOP+FOS+$Q2$	0.0855 (0.0785, 0.0925)	0.1233	0.1022
	BM25+FOS+$Q3$	0.0755 (0.0678, 0.0832)	0.1140	0.0925
	BM25F+FOS+$Q3$	0.0780 (0.0711, 0.0849)	0.1197	0.1002
	BM25FTOP+FOS+Q3	**0.0870 (0.0809,0.0931)**	**0.1277**	**0.1048**
	BM25+FOS+$Q4$	0.0702 (0.0625, 0.0779)	0.1046	0.0844
	BM25F+FOS+$Q4$	0.0717 (0.0670, 0.0764)	0.1055	0.0859
	BM25FTOP+FOS+$Q4$	0.0785 (0.0721, 0.0849)	0.1094	0.0916
	BM25+FOS+$Q5$	0.0544 (0.0464, 0.0624)	0.0603	0.0574
	BM25F+FOS+$Q5$	0.0568 (0.0497, 0.0639)	0.0692	0.0602
	BM25FTOP+FOS+$Q5$	0.0660 (0.0591, 0.0729)	0.0781	0.0720

Bold entries indicate the methods with the highest performance in terms of MAP

for each of the five combinations. Our first notification is that all QUIQS enhance retrieval effectiveness by a margin of 5 to 6 %. The most effective combination is $Q3$, which evaluates the blogs posts by using the MEIBI scores, the authors by employing MEIBI, and the blog sites by using the SBI-Rank metric. In particular, a ranking strategy which utilizes BM25FTOP, FOS and $Q3$ (BM25FTOP + FOS + $Q3$) outperforms the baseline approach with the plain BM25 by enhancing MAP by a percentage touching 39 %. In comparison with the strategy which combines term-proximity model to BM25FTOP, the aforementioned scheme generates rankings with higher MAP by a margin of 7.2 %.

Regarding the other four scenarios, we observe a slightly decreased performance in comparison with $Q3$. However, all of them provide significant improvements over the adversary approaches. More specifically, the MAP for the strategy (BM25FTOP + FOS + $Q2$) was 0.0855, 37.8 % and 5.5 % greater than the MAP of the baseline and term-proximity approaches respectively. In addition, the approach (BM25FTOP + FOS + $Q1$) outperformed the baseline method by 38.8 % and the term-proximity model by 6.9 %. Consequently, the combinations of MEIBI and MEIBIX (for posts and bloggers) with BIF and SBI-Rank (for blog sites), achieve enhanced retrieval effectiveness.

The last two settings also provide improvements over the baseline method; however, the performance gains are limited in comparison to the first three approaches. Therefore, the MAP for strategy (BM25FTOP + FOS + $Q4$) which uses BlogRank for the evaluation of the blog sites, was about 10–11 % lower than the MAP achieved by the first three scenarios. On the other hand, the MAP for the last scenario which utilizes the ι-score and ι-index for blog post and bloggers' influence evaluation, was 0.0660, 24 % lower than the MAP of the first three settings.

6 Conclusions and future work

In this paper we studied the issue of improving the effectiveness of opinionated blog retrieval. We proposed an approach which integrates query-independent and time-sensitive quality metrics (QUIQS) into the current ranking schemes, and combines them with the computed relevance and opinion scores. In particular, we introduced three such metric types for a blog post: The first one takes into consideration the overall value of the post based on its generic impact, the second one depends on the current influence of its creator, whereas the third one evaluates the entire blog site which published it. Regarding the last metric type, we introduced two methods, *SBI-Rank* and *Blog Impact Factor (BIF)*, for the estimation of the value of a blog site. The former is based on the influence scores of the member bloggers of a blog community, whereas the latter takes into consideration the impact of its published posts. In addition, we enhanced the proximity-based opinion retrieval model of [10] by including the physical locations (namely *zones* or *fields*) of the document where the query and opinion terms occur. Our experiments with the TREC blogs08 dataset have shown that the field opinion probabilities (FOS) enhance retrieval effectiveness by 1.5 %–2 %, whereas the combination of QUIQS with FOS leads to additional gains.

Note that the usage of different QUIQS in our proposed model leads to different results. According to our experiments, the best performing scenario dictates that we employ the MEIBI-scores and MEIBI-index for the estimation of the posts' value

and the blogger's influence, and SBI-Rank to measure the overall impact of the blog communities. This QUIQS scheme in combination with FOS further improves effectiveness by a margin of about 6 %. On the other hand, the benefits from the usage of other metrics such as MEIBIX, ι-scores, and BIF were limited.

Our future research is now focused on several interesting and challenging issues. The first one concerns opinion retrieval efficiency and query throughput improvement. This requires extensive examination and study of the query processor, and particularly, the scoring module. We are currently experimenting on performance issues regarding the term-proximity retrieval model of Section 5.3 in combination with an opinion lexicon. Another significant related problem is the extraction of objective, unbiased knowledge out of the retrieved opinions. We plan to classify these opinions according to their contextual polarity (positive, negative, or neutral) either by using opinion lexicons, or by employing text classifiers. Such an application would definitely aid users on multiple manners including decision-making, purchases, traveling, etc.

References

1. Agarwal, N., Liu, H.: Blogosphere: research issues, tools, and applications. ACM SIGKDD Explor. Newslett. **10**(1), 18–31 (2008)
2. Agarwal, N., Liu, H., Tang, L., Yu, P.: Identifying the influential bloggers in a community. In: Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08), pp. 207–218 (2008)
3. Akritidis, L., Katsaros, D., Bozanis, P.: Identifying influential bloggers: time does matter. In: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technologies (WI-IAT'09), vol. 1, pp. 76–83 (2009)
4. Akritidis, L., Katsaros, D., Bozanis, P.: Identifying the productive and influential bloggers in a community. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. **41**(5), 759–764 (2011)
5. Akritidis, L., Katsaros, D., Bozanis, P.: Improved retrieval effectiveness by efficient combination of term proximity and zone scoring: a simulation-based evaluation. Simul. Model. Pract. Theory **22**, 74–91 (2012)
6. Büttcher, S., Clarke, C., Lushman, B.: Term proximity scoring for ad-hoc retrieval on very large text collections. In: Proceedings of the 29th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06), pp. 621–622 (2006)
7. Dave, K., Lawrence, S., Pennock, D.: Mining the peanut gallery: opinion extraction and semantic classification of product reviews. In: Proceedings of the 12th International Conference on World Wide Web (WWW '03), pp. 519–528 (2003)
8. Esuli, A., Sebastiani, F.: Sentiwordnet: a publicly available lexical resource for opinion mining. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06), vol. 6, pp. 417–422 (2006)
9. Garfield, E.: The Application of Citation Indexing to Journals Management. Thomson Reuters (1994)
10. Gerani, S., Carman, M., Crestani, F.: Proximity-based opinion retrieval. In: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10), pp. 403–410 (2010)
11. Hirsch, J.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. U. S. A. **102**(46), 16,569 (2005)
12. Kritikopoulos, A., Sideri, M., Varlamis, I.: Blogrank: ranking weblogs based on connectivity and similarity features. In: Proceedings of the 2nd International Workshop on Advanced Architectures and Algorithms for Internet Delivery and Applications, p. 8 (2006)
13. Langville, A., Meyer, C.: Google Page Rank and Beyond: The Science of Search Engine Rankings. Princeton University Press, Princeton (2006)
14. Lee, Y., Na, S., Kim, J., Nam, S., Jng, H., Lee, J.: Kle at trec 2008 blog track: blog post and feed retrieval. In: Proccedings of TREC 2008 (2008)

15. Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec 2007 blog track. In: Proceedings of TREC 2007 (2007)
16. Mullen, T., Collier, N.: Sentiment analysis using support vector machines with diverse information sources. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP '04), vol. 4, pp. 412–418 (2004)
17. Na, S., Lee, Y., Nam, S., Lee, J.: Improving opinion retrieval based on query-specific sentiment lexicon. LLNCS **5478**, 734–738 (2009)
18. Ounis, I., De Rijke, M., Macdonald, C., Mishne, G.: Overview of the trec 2006 blog track. In: Proceedings of TREC 2006 (2006)
19. Ounis, I., Macdonald, C., Soboroff, I.: Overview of the trec-2008 blog track. In: Proccedings of TREC 2008 (2008)
20. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP '02), pp. 79–86 (2002)
21. Tayebi, M., Hashemi, S., Mohades, A.: B2rank: an algorithm for ranking blogs based on behavioral features. In: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence (WI '07), pp. 104–107 (2007)
22. Turney, P.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02), pp. 417–424 (2002)
23. Turney, P., Littman, M.: Measuring praise and criticism: inference of semantic orientation from association. ACM Trans. Inf. Syst. (TOIS) **21**(4), 315–346 (2003)
24. Vechtomova, O.: Facet-based opinion retrieval from blogs. Inf. Process. Manag. **46**(1), 71–88 (2010)
25. Zhang, M., Ye, X.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: Proceedings of the 31st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08), pp. 411–418 (2008)
26. Zhang, W., Yu, C., Meng, W.: Opinion retrieval from blogs. In: Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM '07), pp. 831–840 (2007)