ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

An Improved Opinionated Ranking Retrieval System from Blogs

N.Sprushya
M.Tech,
Department of CSE,
AVNIET, JNTUH, Hyderabad.

M.Swapna
Assistant Professor,
Department of CSE,
AVNIET, JNTUH, Hyderabad.

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

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 1million 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 where as 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 judge able 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 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 afinal 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.

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

– We introduce the idea of assigning query independent quality scores (QUIQS, pronounced “quick’s”) 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. 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.

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 . 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 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, where the authors attempt to combine diverse sources of potentially pertinent information. In 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 and Turney and Littman 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 . For instance, Zhang et al. computed the probability of query terms and opinion terms co-occurrence by employing a word window. Similarly, Vechtomova considered a word window around each query term and calculated the distance between the query terms and each word in the window.

In the authors computed the proximity by employing n-grams and experimented with several machine learning classification methods. The authors in proposed a proximity-based opinion propagation model to calculate the opinion density at each point in a document. In addition, Pang et al. employed supervised machine learning techniques to identify positive and negative reviews of movie films, whereas used special words (such as poor and nice) and a machine-learning algorithm to achieve sentiment analysis.

Existing System:

In this work I studied the issue of improving the effectiveness of opinionated blog retrieval. I 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, I 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, I introduced two methods, SBI Rank and Blog Impact Factor (BIF), for the estimation of the value of a blog site.

Proposed System:

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 combination with an opinion lexicon. Another significant related problem is the extraction of objective, unbiased knowledge out of the retrieved opinions.

IMPLEMENTATION:

1.Query Independent Quality Sources (QUIQS):

In this work i 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. 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 Page Rank and HITS. I 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, i 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.

A. 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, 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 Sι,d (which we call ι-score) of a blog post d is determined by the following equation:

Sι,d = w(Ld)wcCd + win|_Dc,d|m=1Sι,dm− wout|_Dr,d|n=1Sι,dn

where w(Ld) is a weight function depending on the length Ld of a post. The symbol wc represents a weight which regulates the contribution of the number of comments Cd, whereas win and wout 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 Page Rank 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 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:

SM,d = γ (Cd + 1)|Dc,d|_θt − td + θ_δ

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.2.Combining opinion and relevance scores with QUIQS

I introduce a new type of scores, SQI(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 SQI are expressed as a linear combination of the entire blog site quality Sb, the blogger's influence score ha, and the overall value of the blog entry Sd. The following equation captures these features:

SQI(d, a, b) = Wb Sb +Waha +WdSd

Where Wb ,Wa, and Wd 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 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 S(d, q) and SQI(d, a, b) are combined to form the final score a candidate post d receives with respect to the query q:

SIR(d, a, b, q) = WS(d, q) + (1 −W)SQI(d, a, b)

Where 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 W = 0.8.

3.Experiment

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

A.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. 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.

B.Blog site rankings:

In this subsection I 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. Furthermore, they confirm that the computation of QUIQS is applicable to large scale data sets.

C.Retrieval effectiveness:

In this subsection I 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. Each query of our test set is accompanied by the corresponding "qrels" file which contains the documents which are relevant, both relevant and opinionated and both relevant and factual.

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.

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.

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. 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.

CONCLUSIONS AND FUTURE WORK:

In this work I studied the issue of improving the effectiveness of opinionated blog retrieval. I 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, I 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, I introduced two methods, SBI Rank and Blog Impact Factor (BIF), for the estimation of the value of a blog site. Our experiments with the TREC blogs08 dataset have shown that the field opinion probabilities (FOS) enhance retrieval effectiveness, whereas the combination of QUIQS with FOS leads to additional gains.

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 combination with an opinion lexicon.

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 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 SIGKDDExplor. 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: Proceedingsof the 2009 IEEE/WIC/ACM International Joint Conferences onWeb 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 acommunity. 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) World WideWeb

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 EmpiricalMethods 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 InternationalACMSIGIR 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 AC