Quality Models for Microblog Retrieval

Jaeho Choi
NHN Corporation
NHN Green Factory, 178-1
Seongnam, Korea
jaehos@nhn.com

W. Bruce Croft
Dept. of Computer Science
Univ. of Massachusetts Amherst
Amherst, MA
croft@cs.umass.edu

Jin Young Kim
Dept. of Computer Science
Univ. of Massachusetts Amherst
Amherst, MA
jykim@cs.umass.edu

ABSTRACT

Microblog services typically contain very short documents (e.g., tweets) containing comments about the latest news and events. Many of these documents are not informative or have very little content due to their personal and ephemeral nature. Providing effective retrieval in a microblog service will require addressing the challenge of distinguishing the high-quality, informative documents from the others. Recent work has focused on finding features that indicate the quality of microblog documents, but the impact these quality features on retrieval is not clear. In this paper, we suggest a low-cost quality model using surrogate judgments based on user behavior (i.e., retweets) that can be collected automatically. We analyze the relationship between document informativeness and relevance judgments for microblog retrieval. Then we demonstrate that our behavior-based quality metric has a high correlation with manual judgments. Also, we perform experiments to study the impact of the quality model on microblog retrieval. The results based on the TREC Microblog track show that the proposed quality model, combined with a variety of retrieval models, can improve retrieval performance and is competitive with a model trained using manual relevance judgments.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation

Keywords

Quality model, microblogs, quality-biased ranking

1. INTRODUCTION

A microblog is medium that is similar to a traditional blog in terms of being used to post personal opinions. However, it is different than a traditional blog in that the users of microblogs tend to post smaller content that is highly related to timely events (e.g., breaking news) and closely connected with offline and online social network relationships. In this paper, we focus on microblog ad-hoc retrieval, which aims to find relevant and recent content for current trend-related queries issued by anonymous users [20]. The realtime ad-hoc task of the TREC 2011 microblog track describes a similar scenario in which the goal is to find "the most recent but relevant" tweets at a specific time. To achieve this goal, participating systems need to detect topically relevant documents and arrange them according to the post time from the most recent to the oldest. In general, most microblog documents are posted by individuals who want to spread breaking news and to express their personal feelings. As a result, we can rarely be assured of the quality of microblog content. In addition, unlike other user-generated contents (e.g., Q&A and Blog), microblog documents often have constraints. For example, Twitter's 140-character limit makes it difficult to distinguish informative content, since the length limitation causes many cases of word abbreviations and poor linguistic usage. In this sense, to improve microblog retrieval, we need to take into account the quality of the document content.

Quality models have been studied in a variety of settings, especially for Web search, since they facilitate distinguishing informative or authoritative content from less useful content. For example, PageRank [3] and HITS [8] are popular models based on link analysis. On the other hand, recent papers show that deterministic quality-biased ranking [2] using content-based features can improve the retrieval performance of Web search. It is notable that simple and easy-to-compute features can distinguish the quality of documents. However, there is a drawback with deterministic quality-biased ranking, in that it depends on supervised learning, which requires training examples to train the model. In our preliminary analysis, we found that retweet behavior (abbreviated to RT), which indicates a user quoting or forwarding other users' content on Twitter, can be used as a surrogate judgment for informative content. For example, if a microblog document is quoted by other users, we can consider that it contains informative or interesting content since users tend to broadcast something worthwhile for their neighbors. While it is unclear what makes people retweet, there are several probable reasons to retweet such as content, network and temporal influence [16]. Among them, we focused on the content of the tweet, which has a large effect on retweeting. We found that retweets influenced by content generally indicate the existence of the informative content, but retweets influenced by a user's network (e.g., most of a celebrity's tweets are retweeted) do not.

The assessment results show that the informativeness and the relevance of the tweet are highly correlated and retweeted messages are more likely to be informative than those not retweeted. We train a quality model based on these surrogate judgments of informative content by using both previously used and novel quality features. We evaluate our RT-based quality model using the TREC 2011 Microblog track corpus. The evaluation results demonstrate that our quality model can improve retrieval performance compared to the baselines and is competitive with quality models trained using manual relevance judgments. Furthermore, we demonstrate that the RT-based quality model tends to demote uninformative content such as

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

spam filtering, rather than promote informative content. Overall, our approach brings some advantages for finding informative content, which have applications beyond the retrieval task.

2. RELATED WORK

A number of quality models have been previously studied, based on both link analysis and content features (e.g., the information-noise ratio). Bendersky et al [2] proposed quality-biased ranking (QBR) for web search, which directly introduces document quality as part of the ranking function. Features based on the content, readability, and spam level of a web page were used. To improve retrieval performance for microblogs, some recent papers have suggested using quality-biased features. For example, Alonso et al. [1] used crowdsourcing to detect "interesting" content from randomly selected tweets and reported the presence of a hyperlink in a tweet was strongly correlated with interestingness. Duan et al. [5] incorporated account authority and tweet-specific features in a learning-to-rank framework. Massoudi et al. [10] adapted quality indicators such as emoticons and capitalization, which came from blog post retrieval. Our work is related to this previous work in that we used quality-biased features to improve microblog retrieval, indeed we adapt many of the features from the previous papers.

However, unlike previous work that has employed human labeling data to train a quality-biased ranking function, our quality model uses surrogate judgments based on user behavior (i.e., retweeting) that can be collected automatically. In fact, some researchers attempted to predict popular messages [6] by focusing on how information spread through retweeting. As a result, topological features (e.g., number of followers) were useful. Here, we focus on capturing a signal of high-quality document that can be incorporated into the retrieval model. Meanwhile, Huang et al. [7] suggested a quality-biased ranking model, which incorporates a regularization factor to overcome the sparse quality judgment data. Our approach has a similar goal. Recently, Naveed et al. introduced the probability of being retweeted [14,15] as a quality measure, and showed that quality improve retrieval performance for shorter queries but deteriorates for longer queries. To the best of our knowledge, their work is the closest to ours. To improve their work, we investigate the relationship between retweets and informativeness based on analysis of our user assessment data. Furthermore, we combine the quality score with state-of-the-art retrieval models. The result of experiments with the TREC Microblog track data shows that our approach improves the performance of a variety of retrieval models.

3. USER STUDY

In this section, we introduce our user assessment study. Our aim is to help understand the difference between two evaluation metrics relevance and informativeness, and how these metrics are related to retweeting.

3.1 Obtaining Data

To obtain assessment data, we selected 59 topics from a topic list (described in Section 4) and randomly sampled 100 tweets per each topic from the tweets containing the topic keywords. Then, we asked the assessors to evaluate a tweet in terms of relevance and informativeness. The voluntary assessors were twelve computer science major students and each assessor evaluated 5 topics, 100 tweets per each topic. An assessor first judged a tweet as to whether it was relevant or not to the given topic on a three-point scale: 0 (non-relevant), 1 (relevant) and 2 (highly-relevant). In addition, we asked them to judge whether the tweet was informative or not: using 1 (informative) or 0. To define informativeness, we asked the assessors to consider the question, "Does it contain specific information that people might search?" which is similar to the "interestingness" used in [1]. This means that even a tweet marked as non-relevant could be marked as informative.

3.2 Result of Analysis

We investigated the relationship between relevance and informativeness. The assessment results are displayed in Table 1.

Table 1: User assessment results

Relevance vs. Informativeness

	R=0	R=1,2	Total
I=0	1799 (64%)	1018 (36%)	2817
I=1	248 (9%)	2573 (91%)	2821

Relevance vs. Retweet

	R=0	R=1,2	Total
RT=0	1600 (39%)	2463 (61%)	4063
RT=1	447 (28%)	1128 (72%)	1575

Informativeness vs. Retweet

	I=0	I=1	Total
RT=0	2217 (55%)	1846 (45%)	4063
RT=1	600 (38%)	975 (62%)	1575

A large portion of documents (approximately 50%) turned out to be uninformative. This result corresponds to previous research [1] on the informativeness of microblog documents. Comparing the conditional probability of relevance given informativeness, an informative document is more likely to be relevant than an uninformative one (c.f., $P(R\geq1|I=1)=0.91$, $P(R\geq1|I=0)=0.36$). We found that tweets containing personal feelings about specific topics were marked as relevant but uninformative documents. Although it is not clear whether these personal tweets can be relevant or not, it is also true that some people want to read other people's thoughts through microblog documents.

The results comparing relevance and retweeting demonstrates that retweeting has a small correlation with the relevance score. ($P(R\geq1|RT=1)=0.72$, $P(R\geq1|RT=0)=0.61$). This means that if we directly use the information of being retweeted as one of the features [5,10], we can only get marginal benefit from it. 72% of relevant tweets were non-retweets, whereas 28% were retweeted. Note that there is only a small number of retweets (e.g., 7%) in real-world data whereas we assess more retweets (e.g., 28%) for the user study. Meanwhile, the assessment results comparing informativeness and retweeting shows that the RTs are more likely to be informative than the non-RTs ($P(I=1|RT=1)=0.62$, $P(I=1|RT=0)=0.45$). In summary, from the user assessment study, we found that the RT-set can be used as surrogate judgments for a quality measure (informativeness), and the estimated quality measure can be used to help identify relevant documents. Cohen's kappa, the inter-annotator agreement of the randomly selected two assessors who evaluated the same topics, was 0.66 for relevance and 0.56 for informativeness.

4. TRAINING A QUALITY MODEL

4.1 Training Set

We extract topic keywords by using a simple outlier detection method directly from the corpus. Topic keywords consist of unigrams, bigrams and trigrams where longer keywords have priority. For example, we select 'Keith Olbermann' rather than 'Olbermann' for topic keywords if both are detected by our

algorithm. In this manner, we extracted 931 topics for 18 days, approximately 50 topics per each day from the training corpus. The extracted topic keywords were used for collecting microblog documents to build a quality model. Meanwhile, previous work on retweets [6,14,15] defined the retweet set by matching *RT* signatures, (i.e., RT @username). However, there are many near-duplicates among the tweets. Since retweets represent positive samples in our quality model approach, creating a high-quality retweet set is crucial for the robustness of the quality model. To this end, we detect near-duplicates among the tweets using the Jaccard similarity measure based on bigram overlapping method. Once we find variations of same content according to the similarity measure, we retain the earliest tweet based on its post time and remove the others from training set. If one tweet in a group contains a RT signature, we consider the earliest tweet of the group as the retweeted tweet, even if the earliest tweet doesn't contain a RT signature. In this manner, we collected 44,734 (3.4%) retweets from the candidate set. The rest of the tweets (96.6%) were denoted as the non-retweet set.

4.2 Quality Model

In general, the quality of document can be seen as a prior probability, denoted *P(D)*, and it is often assumed to be uniform in the language model (LM) [19]. Zhou and Croft suggested the document quality language model [22] for Web search which uses the conditional probability of the document is classified into high quality class, given the quality features. In this work, we introduce a binary random variable *RT* and rank the documents in descending likelihood of *Q* generating the retweeted document *D*. We use the chain rule and the fact that *Q* and *RT* are conditionally independent given *D*. As a result, we formulate the quality model using *P(RT|D)*, that is, the probability of a tweet being retweeted given document as Eq. (1). In this formula, *P*(*Q*|*D*) stands for text matching scores of a variety of language models and *P(D)* is still assumed to be uniform.

$$\begin{aligned} P(D, RT|Q) &\propto P(Q|D, RT)P(D, RT) \\ &\propto P(Q|D, RT)P(RT|D)P(D) \\ &\propto P(Q|D)P(RT|D)P(D) \end{aligned} \quad (1)$$

In [22], the conditional probability was estimated directly from the training data by using a kernel density estimation technique, since they used only two quality features. However, we incorporate more than 20 quality features, and these features are generally correlated with each other. As a result, we use logistic regression to estimate *P*(*RT*|*D*). The advantage of using a logistic regression model is that the output is continuous value, which can be regarded as a probability. In the closest work [14,15], a logistic regression model was also used to estimate the retweet probability.

4.3 Quality Features

We use features from previous work and also introduce novel features: nonRTScore, tweetDocMatch, and tweetUrlMatch. First, we adapt *Web-specific content-based* features that were the most important in previous work [2] for Web search. Those are numVisTerms (number of terms), avgTermLen (average length of terms), entropy (entropy of the content), fracStops (stopword/non-stopword ratio), and stopCover (fraction of terms in stopword list). We also use *microblog-specific content-based* features that used separately in previous work [1,5,6,7,10,13,14,15]. Those are tfIdfScore (sum of tf-idf scores), fracEnglish (fraction of English terms), fracLetter (fraction of alphabet letters), fracCap (fraction of capital letters), fracUnique (fraction of unique terms), maxTermLen (maximal length of terms), minTermLen (minimal length of terms), hasHashTag (presence of a hashtag), hasPerson (presence of person name), hasLocation (presence of location), hasOrgan (presence of organization), hasExcl (presence of an exclamation mark), hasQues (presence of a question mark), isReply (is the tweet a reply tweet), and isBeginTag (is the tweet beginning with a hashtag). In particular, we calculate KL-divergence from non-RT set to RT set to compute a novel feature: nonRTScore (sum of non-RT word scores). This score indicates that the tweet contains terms that are more likely to happen in the non-RT set, so we call it nonRTScore. We show some sample terms sorted by D_{KL}(non-RT||RT) and D_{KL}(RT||non-RT) in Table 2. The left side of the table indicates that there exist many personal postings with some internet slang (i.e., lol) in non-RT set, which were used heuristically to detect newsworthy tweets [4]. In contrast, the right side of the table shows that RT set is biased to certain types of tweets (e.g., celebrities). For this reason, we only use D_{KL}(non-RT||RT) for demoting personal tweets.

Table 2: KL-divergence term density

Term	D_{KL}(non-RT||RT)
i	0.009110
my	0.005642
lol	0.004141
im	0.002979
me	0.002558

Term	D_{KL}(RT||non-RT)
you	0.006728
zodiacfact	0.004811
your	0.004506
justin	0.004242
bieber	0.003646

We also download URLs that are contained in a tweet and use them to calculate *URL-based* features. We hypothesize that if the content of link contains some useful information, they might share the same terms. Therefore, we use hasLink (presence of a hyperlink), *tweetDocMatch* (term overlap between a tweet (T) and the crawled document (D), that is, $|T \cap D| / |T|$), and *tweetUrlMatch* (term overlap between a tweet (T) and the resolved URL (U), that is, $|T \cap U| / |T|$). Lastly, we use numTweets (number of tweets an user posted) as a *User-based* feature.

5. MICROBLOG RANKING

5.1 Retrieval Models

We use the Dirichlet smoothing LM (QL) [21] as a baseline. We found that the smoothing parameters tuned on the training corpus were relatively small (i.e., μ=500~1200) compared to other collections. We hypothesize that is because the microblog document lengths are short and their variances are also small [17]. We also choose the MRF model [11], which uses query term proximity in a document. There are two variants: the sequential dependence model (SDM) and the full dependence model (FDM). In previous work, Metzler and Cai reported that the FDM yielded superior results to the SDM [13]. They hypothesized that was because the FDM promoted tweets that contained more query terms. We evaluate their assumption through the following experiments with uniw=0.8, odw=0.2, and uww=0.0 for model parameters. The last choice of retrieval model is the Relevance model (RM) [9]. By using expanded queries, the RM can potentially address issues related to synonymy and polysemy. Since microblog documents contain many cases of word variations caused by the different language use of microblog users, the RM may help. We use relevance models using QL, SDM, and FDM for the initial retrieval, denoted by RM, SDRM and FDRM respectively. We used fbDocs=10, fbTerms=10 for model parameters and weighted the original and expansion query equally.

Table 3: Performance of the models – averaged over 49 topics (MB050 topic omitted due to the absence of relevant tweet) and † denotes statistically significant difference from the baseline (two sided paired randomization tests[18] : p-value < 0.05)

	QL		SDM		FDM		RM		SDRM		FDRM	
	MAP	P@30	MAP	P@30	MAP	P@30	MAP	P@30	MAP	P@30	MAP	P@30
Base	0.2326	0.4000	0.2304	0.4156	0.2436	0.4286	0.2982	0.4721	0.2874	0.4850	0.3097	0.5027
Base+Q	0.2444	**0.4197**†	0.2407	0.4204	0.2528	**0.4429**†	0.2999	0.4735	0.2917	0.4918	0.3078	0.4966
QBR	0.2511	**0.4347**†	0.2537	**0.4442**†	0.2538	**0.4544**†	0.3010	0.4857	0.2950	0.5034	0.3086	0.5068

5.2 Combining the Quality Model

In Section 4.2, we introduced Eq. (1) which incorporates $P(RT|D)$ as a quality factor. We now define the score of a document as a linear combination of a text matching score and a quality score.

$$SC(Q,D) = \lambda \log P(Q|D) + (1-\lambda) \log P(RT|D) \qquad (2)$$

That is, we take the logarithm and use the parameter λ to combine two probabilities. Here, we need to estimate the λ, which decides how much the quality score affects the retrieval result. In this work, we use a learning-to-rank approach for estimation. Even though we use surrogate judgments to train the quality model, it is very difficult to estimate the ranking parameter without any evidence for relevance. For this reason, we combine the quality model with retrieval models as in QBR. To address the relationship between the quality model and retrieval performance, we also conduct QBR experiments using the same features of the quality model. While our quality model uses surrogate judgments to find optimal weights for quality features, QBR employs human judgments. Indeed, QBR has outperformed other quality models [2,7] because it optimizes quality feature weights according to the resulting performance. As a result, we evaluate QBR performance and compare the model parameters to our quality model. By doing this, we focus on differences between surrogate judgments and human judgments. We used Coordinate-Ascent [12] for learning-to-rank approaches. All experiments are done using 5-folds cross-validation.

6. EVALUATION

6.1 Experimental Setup

The TREC 2011 microblog track released 16M tweets and 50 topics and preferences for the ad-hoc search task. Unlike the other TREC tasks, the microblog track topic contains the timestamp when the query was issued by the user, which means that no documents newer than a given query's timestamp should be retrieved. Furthermore, by the guidelines of the microblog track, the final ranked results must be ordered chronologically. These constraints mean that we retrieve initial results based on the relevance score, and then rank them by posting time. We index all tweets in the microblog copora using the Galago retrieval system, and stem with the Porter2 stemmer. We use the stopword list which was used in [2]. To tune the language model parameters (e.g., μ), we used a training corpus with 59 topics and user assessment results described in Section 3. The training corpus consists of 17M tweets that had been crawled using Twitter API. This data had been crawled between January 8rd and February 8th, 2011. There are overlapping days between the training corpus and the TREC Microblog track data. However, we found that there were no common tweets between the two corpora. We guess that this is caused by the difference of sampling methods used for gathering the data. As a result, the quality model is trained on different data than the evaluation data, that is, the TREC data. We use six variants of language models as our baselines. We denote the performance of our quality model as **Base+Q**, which indicates the quality score combined with the baseline model, and **QBR** for the quality-biased ranking. To evaluate the performance, we used two measures, MAP and precision at 30 (P@30). P@30 was used as the official measurement in the TREC Microblog ad-hoc task.

6.2 Experimental Results

We display the experimental results in Table 3. The results show that our quality model improves the baseline retrieval models in most cases. Significant differences were observed compared to the QL and FDM baselines. We found that these trends were also observed in QBR and our RT-based model is competitive. In the case of the baselines based on relevance models, there are some improvements but they were not significant. Note that this is the case even with the QBR run. This is a very different result than was found with web documents, where pseudo-relevance feedback techniques are not competitive with the quality-biased model. We hypothesize that the retrieval models that use query expansion overcome the low-quality document problem to some degree by matching large number of query terms. That is, the more terms are matched in a document, the more informative the document will be. This effect would be more obvious for microblog documents, which use only small number of words. RM techniques are, however, more expensive in terms of computation than the quality-based models.

To understand the utility of features, we divided them into four groups, and then repeated the experiment removing one group of features each time. We display the results for the quality model and the QBR in Figure 1.

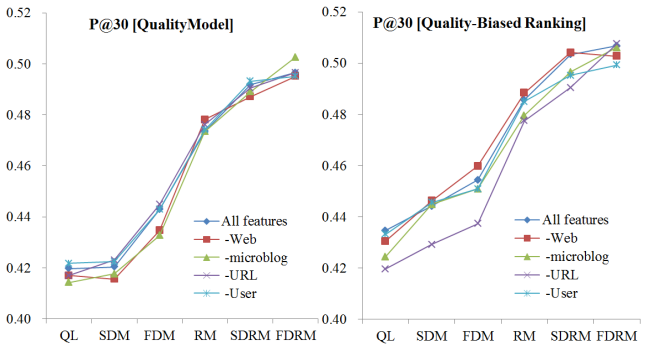

Figure 1: The performance results of feature elimination

We note that there is little difference with regard to the features for RM-based models compared to QL, SDM and FDM. The result of the QBR experiment shows that URL features are the most important. In particular, we found that the overlap ratio between the tweet and the crawled page content (i.e., *tweetDocMatch*) is also of use as much as the existence of hyperlinks. Also, the result shows that microblog-specific content-based features play an important role for improving performance. However, the web-specific content-based features which were useful for Web search have little impact for microblog retrieval. In contrast to the result of the QBR performance, the result of our quality model is more affected by

content-based features, both of Web-specific and microblog-specific, but is less affected by URL features. In Figure 2, we display the feature weight of the QBR and our RT-based quality model in the FDM baseline experiment.

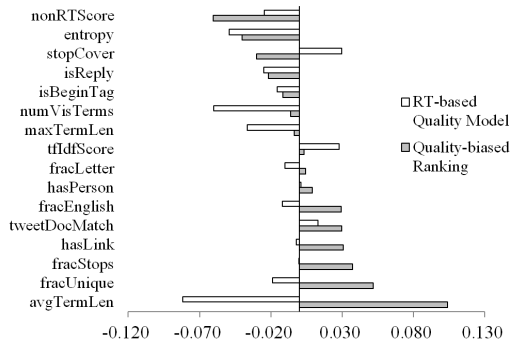

Figure 2: The weight of quality features

Each weight was normalized by the weight of retrieval model score (i.e., output of FDM), so we can compare their relative importance. Interestingly, most of negative weights are highly related, but positive weights of the QBR model are not. This means that our RT-based quality model can demote uninformative documents by using negative features (e.g., nonRTScore, entropy, isReply, etc.), but, does less well at promoting more informative documents. Indeed, the positive features (e.g., fracEnglish, hasLink, fracUnique, etc.) of QBR are highly related to well-formed documents. As a result, we claim that our quality model based on surrogate judgments can be used for filtering out uninformative (or non-relevant) documents for microblog retrieval.

7. CONCLUSION

In this work, we examined document quality, which plays an important role in microblog retrieval. We suggested a quality model using surrogate judgments based on human behavior (i.e., retweets) that can be collected automatically to train the model. Specifically, we described how to train the quality model, which consists of topic extraction, retweet set creation and the quality model parameters estimation. We conducted a user assessment study which demonstrated that retweets can be used to find informative tweets. We evaluated the RT-based quality model on the TREC Microblog track data and showed that it generally improved baseline retrieval models. The results obtained with the RT-based model are competitive with the QBR model, which requires manual judgments for training. The improvements for all quality models were lower compared to baselines that use pseudo-relevance feedback because matching expanded query terms appears to promote quality, which has not been observed in experiments with web pages. Quality-based models have the advantage in terms of being faster to compute. By comparing the model parameters with QBR, which uses human judgments for training, we found that our RT-based quality model demoted low-quality documents effectively. We also used a variety of quality features from the previous work for Web and microblog search, and analyzed which features were useful for the retrieval task.

8. ACKNOWLEDGMENTS

This work was supported in part by NHN Corp. and in part by the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors' and do not necessarily reflect those of the sponsor.

9. REFERENCES

[1] O. Alonso, C. Carson, D. Gerster, X. Ji, and S. U. Nabar. Detecting uninteresting content in text streams. In SIGIR'10 Crowdsourcing for Search Evaluation Workshop, 2010.

[2] M. Bendersky, W. B. Croft, and Y. Diao. Quality-biased ranking of web documents. In WSDM'11, 2011.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. Computer Networks and ISDN Systems, 1998.

[4] C. Castillo, M. Mendoza, and B. Poblete. Information Credibility on Twitter. In WWW'11, 2011.

[5] Y. Duan, L. Jiang, T. Qin, M. Zhou, H. Shum. An empirical study on learning to rank of tweets. In Coling'10, 2010.

[6] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in twitter. In WWW'11, 2011.

[7] M. Huang, Y. Yang, and X. Zhu. Quality-biased ranking of short texts in microblogging services, In IJCNLP'11, 2011.

[8] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. J. ACM, 1999.

[9] V. Lavrenko, W. B. Croft. Relevance-based language models. In SIGIR'01, 2001.

[10] K. Massoudi, E. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In ECIR'11, 2011.

[11] D. Metzler, W. B. Croft. A Markov random field model for term dependencies. In SIGIR'05, 2005.

[12] D. Metzler and W. B. Croft. Linear feature-based models for information retrieval. Information Retrieval, 10(3), 2007.

[13] D. Metzler and C. Cai, USC/ISI at TREC 2011:Microblog Track, In TREC'11, 2012.

[14] N. Naveed, T. Gottron, J. Kunegis, and A. Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In WebSci'11, 2011.

[15] N. Naveed, T. Gottron, J. Kunegis, and A. Che Alhadi. Searching microblogs: Coping with sparsity and document quality. In CIKM'11, 2011.

[16] H.-K. Peng, J. Zhu, D. Piao, R. Yan and J. Y. Zhang. Retweet Modeling Using Conditional Random Fields. ICDM Workshops, 2011.

[17] J. Seo and W. B. Croft. Unsupervised estimation of dirichlet smoothing parameters. In SIGIR'10, 2010.

[18] M. D. Smucker, J. Allan, and B. Carterette, A Comparison of Statistical Significance Tests for Information Retrieval Evaluation, CIKM'07, 2007.

[19] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval, In SIGIR'98, 1998.

[20] J. Teevan, D. Ramage, and M. Morris. #Twittersearch: A comparison of microblog search and web search. In WSDM'11, 2011.

[21] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In SIGIR'01, 2001.

[22] Y. Zhou and W. B. Croft. Document quality models for web ad hoc retrieval. In CIKM'05, 2005.