

Searching Microblogs: Coping with Sparsity and Document Quality

Nasir Naveed Thomas Gottron Jérôme Kunegis Arifah Che Alhadi
WeST – Institute for Web Science and Technologies
University of Koblenz-Landau, 56070 Koblenz, Germany
{naveed, gottron, kunegis, alhadi}@uni-koblenz.de

ABSTRACT

Two of the main challenges in retrieval on microblogs are the inherent sparsity of the documents and difficulties in assessing their quality. The feature sparsity is immanent to the restriction of the medium to short texts. Quality assessment is necessary as the microblog documents range from spam over trivia and personal chatter to news broadcasts, information dissemination and reports of current hot topics. In this paper we analyze how these challenges can influence standard retrieval models and propose methods to overcome the problems they pose. We consider the sparsity's effect on document length normalization and introduce *interestingness* as static quality measure. Our results show that deliberately ignoring length normalization yields better retrieval results in general and that interestingness improves retrieval for underspecified queries.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Content Analysis and Indexing; H.3.3 [Information Search and Retrieval]: *Information filtering, Relevance feedback, Retrieval models*

General Terms

Algorithms, Experimentation, Performance

Keywords

Microblog, Twitter, Retweet, Interestingness

1. INTRODUCTION

Microblogging documents are – by definition – very short. In the most renown microblogging service Twitter, the length of messages is restricted to 140 characters. While the conciseness of messages has been cited as a major reason for Twitter's success, it is at the same time problematic for text retrieval: Term frequencies, for instance, are typically

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
Copyright 2011 ACM 978-1-4503-0717-8/11/10 ...\$10.00.

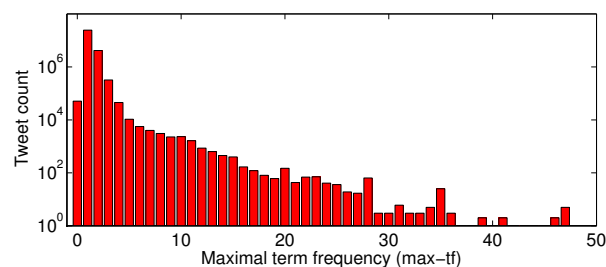


Figure 1: Distribution of maximal term frequencies (*max-tf*) in Twitter messages of the *choudhury-ext* dataset after removing stop words.

used as a parameter in the estimation of term importance within a given document. In short texts, however, this essential feature does not discriminate much between different documents, as it is nearly a binary value. In fact, our analysis across several large datasets shows that about 85% of all Twitter messages contain each term at most once (see Figure 1). In consequence, most retrieval models are effectively reduced to using global term weights, that measure the discriminativeness of terms.

A second challenge for retrieval on microblogs is the nature of the medium. With regard to the purpose Twitter documents are different from documents in the classical sense. Classical documents address a wider, open audience, as in principle everyone has access to the document. Microblog messages, instead, typically have a more restricted and better defined audience provided by the social network of a user. Thus, while the primary purpose of a tweet can be to communicate information (e.g. news, sharing resources, broadcast an alert) it might, alternatively, also serve other primary purposes, such as social interaction, promotion, requesting feedback or expressing emotions [5]. In a general retrieval scenario, the rather private and personal messages in a social interaction context are typically of less interest for a user with a concrete information need. This implies that a retrieval criterion for microblog messages should relate to the *interestingness* of a tweet.

Information retrieval on microblogs needs to address these two challenges that are immanent to the technical and social context of the medium. Classical retrieval models do not consider these aspects; they are directed to longer texts and assume the intention of a document to primarily be the transmission of information. Also, recent approaches transferring the concepts of authority in reference networks, such

as citations or hyperlinks, to the social network of Twitter do not take into consideration that the semantic of the social network is not equivalent to a content motivated reference network.

In this paper we make two contributions:

- We analyze term and length features of microblog messages and provide theoretical and empirical evidence that length normalization introduces an unjustified bias for Twitter.
- We introduce a static quality measure of *interestingness* and show that it improves retrieval results in particular on short, underspecified queries.

2. RELATED WORK

People seek information within microblogs in two ways: by asking questions to their followers or by querying microblogs in order to discover information that has already been posted. Looking at the user’s side, people mainly search microblogs to find in particular timely information (e.g. news, trending topics, events), social information (information related to other users) and topical information (e.g. topic of interest) [15]. Another observed and important difference between search on microblogs and on the Web is the length of queries. While Web queries are on average 3.08 words long, queries on microblogs are far shorter: on average 1.64 words.

Massoudi et al. [8] incorporate quality indicators such as emoticons, post length, shouting, capitalization, hyperlinks, reposts, followers, and recency into a retrieval model for searching microblog posts for a given topic of interest. Previous studies have also measured the influence of users on tweet quality. Cha et al. [4] discovered that a very large number of followers do not necessarily have an impact on a user being retweeted more often.

However, some researchers [7, 10] found that the context of a tweet (number of followers/followees, timestamp) strongly correlates with retweetability along with content features such as the presence of URLs and hashtags. Kwak et al. [10] found that rankings based on the number of followers and PageRank are very similar, while rankings based on the number of retweeted messages differ, concluding that interest does not necessarily correlate with social status. In analogy with PageRank, Weng et al. [16] define the TwitterRank measure to rate users. Nagmoti et al. [11] state that social graph network features (number of followers and followees) can be used as a ranking measure of microblog search.

Hong et al. [7] use retweets as a measure of popularity and apply machine learning techniques to predict how often new messages will be retweeted. The authors analyze the content of messages, temporal information, metadata of messages and users, and the user’s social graph as the features in predicting the messages to be retweeted.

In summary, these recent studies listed here indicate that the likelihood of a tweet to be retweeted is based on context and content of the tweet. Our work focuses on the applicability of a content-based probability of retweet as a static quality measure. Therefore, we put a much stronger emphasis on the content and analyze a wider set of low-level content-based features as well as derived, high-level content-based features (topic and sentiment of a tweet).

3. RETRIEVAL ON MICROBLOGS

In this section we consider the two challenges for retrieval on microblogs we mentioned in the introduction: sparsity and quality. We look at the impact of sparsity on length normalization in retrieval models and motivate to ignore document length in a microblog scenario. Further we introduce a way to measure quality in tweets, which helps to overcome underspecified queries and allows for adding a static quality measure that can compensate for the lack of meaningful term frequencies in microblog messages.

3.1 Term Sparsity and Length Normalization

We already mentioned that microblog messages contain few terms in general and very rarely contain a term more than once (see Figure 1). Intuition dictates that this term sparsity will have an impact in a retrieval setting. The most obvious impact is that a potentially relevant tweet will not be retrieved at all if among its few terms it does not contain one of the query terms. This risk is much higher than with classical documents, as the length restrictions prevent an author from using synonyms or elaborating concepts with additional words. But, a second and more subtle impact lies in the length restriction itself, as we will see in a moment.

Length normalization is an essential ingredient to modern retrieval models. The motivation for length normalization is to counterbalance the potential advantages of longer documents [14] that are commonly explained based on the *verbosity hypothesis* and the *scope hypothesis*.

Verbosity Hypothesis A long document elaborates the same topic longer and repeatedly. Therefore it also contains the same terms repeatedly while not adding further information to the document. This leads to a higher term frequency and in consequence to higher weights for the repeated terms.

Scope Hypothesis A long document addresses several topics. Therefore, it contains more different terms and might seem relevant to wider range of different queries. The general line of thought here is that a user would prefer a short and focused document over a long document relating to several topics.

Given the restriction of microblog messages to contain very few terms, intuition tells us that neither the verbosity nor the scope hypothesis can serve to explain the document length in Twitter messages. To verify our intuition, we analyze a large collection of Twitter messages with respect to two questions: (a) can we observe a tendency of longer tweets to be verbose, i.e. contain terms repeatedly and (b) can we observe a tendency of longer tweets to have a larger scope, i.e. to cover several topics.

To discover verbosity we look at redundancy in a tweet and how it correlates with document length. As measure for document length we employ once the number of characters and once the total number of terms. Given the non-normal distribution of document length (c.f. Figure 2 for the distribution of messages w.r.t. character length) we calculated Spearman’s rank correlation on the two observed values for document length and the amount of redundant words. Between the character length of a tweet and redundancy we observed a correlation of $\rho = 0.381$. This indicates no or at most a very weak correlation. Also for the total number of words and the number of redundant word we found a quite

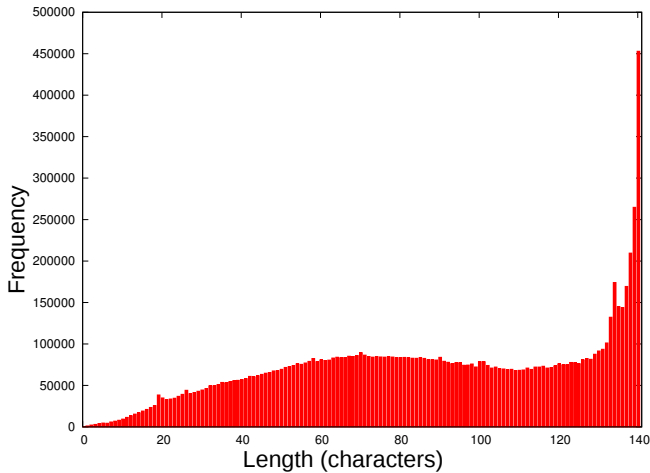


Figure 2: Distribution of document length (in characters) in the *choudhury* dataset.

low value of $\rho = 0.377$, allowing us to conclude that in a microblogging environment document length does not seem to be caused by verbosity.

A larger scope within a document is more difficult to detect. In order to get an idea of how many topics a tweet might cover, we applied latent Dirichlet allocation (LDA) to our dataset. LDA discovers latent topics in text documents and represents them as a distribution over terms. At the same time, documents are seen as a random mixture of topics, i.e. each document is a distribution over topics. The latter allows us to obtain insights on the number of topics a tweet is associated with.

We used LDA to obtain 100 topics and the likelihood of each tweet to belong to this topic. Given this model, we found around 8.5% of the tweets not being strongly related to any topic¹. Among the remaining tweets, we analyzed the tweets for the number of topics that contribute to more than half of the probability of associated topics. To do so we analyzed how many topics we have to consider for each tweet to cover more than 50% of the probability in the topic mixture. We observed that in 77.1% of the tweets this point is already reached with one topic, meaning that the most prominent topic dominates the overall topic of the tweet. And for 99.6% the two top ranking topics contribute half the probability in the topic mixture. Our scope observations are also in line with the work of Zhao et al. [17] that observed that a single tweet usually covers a single topic. As a result, we can say that microblog messages in general are very focused – typically one LDA topic explains very well the composition of the entire tweet.

As neither the verbosity nor the scope hypothesis seem to apply to Twitter, we conjecture that length normalization for Twitter messages is not necessary. On the contrary, it might be counterproductive, as it introduces a bias favoring short over long messages without a justification.

3.2 Interestingness as Static Quality Measure

Microblog documents have different aspects of quality. Obviously, there is the distinction of spam, the trustworthiness or the purpose of a message. Quality is a static

¹No topic had an association of more than 1%

measure for a document that is independent from an actual query. Such a static quality measure is of particular interest when documents are likely to obtain homogeneous relevance values in a retrieval model. This is the case for Twitter, given that term frequency is nearly a binary value on microblogs and queries are typically composed of very few or even single terms.

The motivation of introducing interestingness into a general retrieval setting is, that when searching for messages a user would want to leave his closer social sphere and get results also from other areas in the social graph. We base our notion of interestingness on the *retweet* function. As a retweet is relayed to all the followers of a user, the message being retweeted can therefore be considered of general interest. This notion of a tweet being potentially interesting to other users is then a suitable way to capture the static content quality we have in mind.

Studies of tweets and retweets have revealed that the context of a tweet influences its actual chance to be retweeted. Most prominent context features are the social graph and the time of the original tweet. As our focus is on the message itself, we will deliberately ignore such context information, and rely only on features extracted from the message itself². We use logistic regression on content features to predict the probability of an unseen message to be retweeted and use this probability as a static quality measure for tweets.

We follow our approach from [12] to determine interestingness via the probability of a tweet being retweeted. There, only the following features based on the tweets themselves were considered and a tweet’s author and timestamp was ignored. All features except term odds, sentiment and topic scores are binary.

URLs, usernames and hashtags. Without further differentiation we consider the presence of particular items typical for tweets. These are the presence of a URL, the mention of a username or a hashtag. These constitute three binary features.

Exclamation and question marks. We use the presence of exclamation marks “!” and question marks “?” at the end of tweets as two binary features.

Positive and Negative Terms. We look for positive and negative words from a short predefined list. Positive terms comprise exclamations of happiness, such as *excellent* or *great*, while negative exclamations contain swear words or expressions of disgust (e.g. *eww* or *fail*).

Emoticons. Emoticons or smileys are short character sequences representing emotions. We parse the tweets to find positive emoticons such as :-)) and negative emoticons such as :-(, giving two binary features.

Sentiments. Many tweets are personal and express sentiments. To detect the sentiments expressed by a tweet, we follow previous Twitter research and select a simple dictionary-based approach [9]. We use the Affective Norms of English Words dictionary [3], which provides numerical values for 1,030 English words to capture valence (pleasure vs. displeasure), arousal (excitement vs. calmness) and dominance (weakness vs. strength).

Terms. The most obvious content features in text are the contained terms. We extract terms and normalize them

²Our aim is to capture a relevant and interesting tweet even if it was written by a new user with no followers and at an unsuitable time in the middle of the night.

using case folding and the Porter Stemmer. For each message M we compute the odds of it being a retweet. Modeling a retweet as event RT, assuming independence between the occurrences of terms and employing Bayes’ theorem the odds value can be brought into a form that is easier to handle:

$$O(\text{RT} | M) = O(\text{RT}) \cdot \prod_{t \in M} \frac{P(t | \text{RT})}{P(t | \overline{\text{RT}})}$$

where $O(\text{RT})$ are the a priori odds of a retweet, and the product ranges over the ratios of the probabilities of each contained term to occur in a retweeted or a non-retweeted message. To estimate these probabilities we use maximum likelihood estimation and Laplace smoothing to handle unseen terms.

Topics. The topic of a tweet is a latent feature and can be inferred by analyzing a tweet’s content. Modeling Twitter content requires methods that are suitable for short texts with heterogeneous vocabulary. Recent work shows that one such method is LDA and its extensions [2, 16]. We use LDA to find latent topics that characterize the corpus. In topic modeling the standard model used to describe a topic is as a distribution over terms and documents, generated as a mixture of latent topics.

Based on these features we trained a logistic regression model to be able to obtain for an individual tweet a probability of retweet. We interpret this probability to be the quality of a microblog message. If the probability of retweet is high, the message is seen as interesting for a wider audience and, therefore, of better quality in a general retrieval scenario.

4. EVALUATION

We empirically evaluate our approach in two different ways. We use a relevance-based evaluation following the classical Cranfield paradigm and a subjective evaluation asking users which result sets they prefer for a given query.

To analyze empirically the impact of length normalization and message quality in the sense of interestingness on the retrieval performance we set up three systems making use of two Lucene-based indices over a collection of 10 million tweets. As baseline we use an index-based on an out-of-the-box implementation of Lucene, i.e. a vector space retrieval model (VSM) including length normalization. The second Lucene index was modified not to perform length normalization, but otherwise use the Lucene retrieval function.

In order to incorporate interestingness as static quality measure, we use an approach based on reranking relevant results. For this purpose we take the top-100 entries in a relevance-based result set and rerank them according to descending probability of retweet. This means rank 1 is the tweet that has the highest value for interestingness among the tweets with 100 highest relevance values.

4.1 Methodology

To the best of our knowledge, there is currently no established dataset for evaluating retrieval in a microblog setting. Therefore, we had to develop an own configuration of microblog corpus, topics and relevance decisions. As corpus we employed an existing dataset [6] covering 118,506 users.

Table 1: Twitter datasets used in our experiments.

Dataset	Users	Tweets	Retweets
CHOUDHURY [6]	118,506	9,998,756	7.89%
CHOUDHURY-EXT [6]	277,666	29,000,000	8.64%
PETROVIĆ [13]	4,050,944	21,477,484	8.46%

The ratio of actual retweets in the dataset was 7.89%. The full list of datasets we used is given in Table 1.

To formulate information needs on Twitter data we selected the most prominent terms from 20 distinct LDA topics. To cover different query lengths, we created five queries each of length one to four terms. In this way we could simulate a range of quite general and quite specific information needs, for which the data set should also provide relevant documents.

To assess the objective performance of retrieval on microblogs in a classical Cranfield setting, we needed objective relevance judgements. We applied pooling on the top 5 retrieved tweets for each method and evaluated the messages in the pool for relevance. As relevance is typically judged as *aboutness*, i.e. to which degree a document covers the topic of an information need, we additionally had the judges determine if a tweet was actually interesting. The purpose of this extension is to distinguish between technically relevant (i.e. about the topic, containing the query terms) and actually informative tweets (i.e. about the topic and providing general information on the topic). As we will see later quite often retrieval results on microblog messages are technically relevant, because they contain the query terms, but are practically not informative because they contain no other terms. Other tweets are rather personal messages, that do not satisfy a general information need.

To further measure user satisfaction, we set up a second experiment in which we confronted the judges with two top-10 result lists for a given query, originating from two different retrieval setups anonymized as System A and System B. The users were asked if they preferred the results of System A, of System B, or were indifferent. The intention of this evaluation setup was to capture a subjective preference of a particular system over another one in direct comparison.

For both evaluation tasks we used crowdsourcing to obtain relevance judgements and system preferences respectively. To assert the quality of our crowdsourcing approach we introduced artificial non-relevant results to identify and eliminate spammer results. Further, we collected for each evaluation task the feedback from seven different judgements and derived the final judgement based on the recommendations provided in [1].

4.2 Results

We measured the objective performance using average P@5 and MAP. Figure 3 gives the achieved values on a all queries and a break down for the different query lengths. The plots list the Lucene system in its out-of-the-box configuration (i.e. a VSM with length normalization) as a baseline, the modified Lucene setup which does not perform any length normalization, and the setup using the retweet odds for reranking the top-100 results.

With respect to both evaluation metrics, we observe a global and local trend. Globally we can confirm our theory,

that length normalization on microblog messages is counter-productive. The standard Lucene approach is outperformed on all levels and independent of query length. Looking at the tweets retrieved, length normalization favors shorter messages. In particular for short queries, the tweets quite often consist only of the query terms, thus, not satisfying any information need. Turning off the length normalization bias leads to better results. The gap is narrower for longer queries, but on average, the approach without length normalization still leads to better results.

The second trend is depending on the query length. For short and underspecified queries a simple relevance-based ranking provides relatively poor results. Deactivating length normalization does improve the results, but not to a level that can be observed for longer queries. Here, incorporating interestingness as static quality measure leads to a big improvement. This is of particular interest, as queries on Twitter are typically short: 1.64 words on average [15].

Tables 2 and 3 demonstrate the positive effect of reranking based on interestingness with a very clear and illustrative example of the query *beer*. The top ranking results in Table 2 contain only the query term, repeated many times. Technically these results are relevant as they contain the query term, but they do not convey much information. The results employing reranking based on interestingness in Table 3, instead, are technically relevant and informative.

In general, with an average P@5 value of 1.00 and a MAP value of 0.78, retweet odds reranking on one term queries achieves the overall best results we observed in our evaluation. Looking again at the data we saw two explanations: (a) the top-100 results coming out of the relevance ranking were all more or less related to the topic, thus, reranking did not bring irrelevant documents to the top of the list and (b) the top relevance ranking results consisted mainly of tweets formed by the query term, as already noted above. The longer and more specific the queries are, the less the top relevance ranked documents consists of query terms alone and the more irrelevant tweets come into the top-100 documents used for reranking. Therefore, the advantage of interestingness as static quality measure wears off and might actually put highly interesting, but only marginally relevant tweets to the top of the result lists.

Table 2: Top 5 tweets for the query *beer* using the Lucene-noLen setting.

Tweet
1 Beer beer beer beer beer beer beer beer beer beer beer. Er, guess what I'm looking forward to?
2 BEER ^5. RT @dewbelle: BEER BEER BEER BEER. RT @kulturbrille: BEER BEER BEER. RT @Bluebarrow: BEER BEER. RT @WalterMitty007: BEER
3 http://ping.fm/p/Bnra7 - In!!! BEER, BEER, BEER, BEER, BEER, BEER, BEER, BEER, BEER, BEER, BEER, BEER,
4 Lompoc. beer beer beer beer beer beer beer beer beer beer. http://twitpic.com/168ld
5 Beer, beer beer, beer, beer beer, beer and a little bit more beer.

Our second evaluation looked at subjective performance. Also here we observed the same trends as above. Table 4 shows how many times a system was preferred over an opposed system. Figure 4 further summarizes these results and shows for each query length how many times the result set of an individual system was preferred. We observe the same

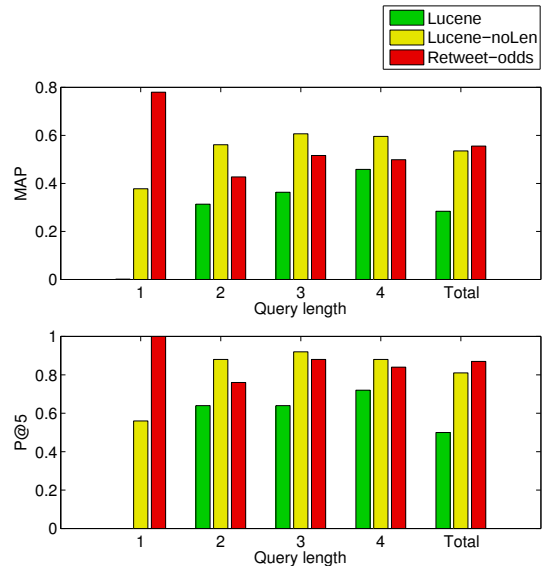


Figure 3: MAP and P@5 performance: in total and resolved by query length.

Table 3: Top 5 tweets for query *beer* using the Retweet-Odds setting.

Tweet
1 UK beer mag declares "the end of beer writing." @StanHieronymus says not so in the US. http://bit.ly/424HRQ #beer
2 beer summit @bspward @jhinderaker no one had billy beer? heehee #narm - beer summit @bspward @jhinde http://tinyurl.com/n29oxj
3 Go green and turn those empty beer bottles into recycled beer glasses! http://bit.ly/2src7F #beer #recycle (via: @td333)
4 Great Divide beer dinner @ Porter Beer Bar on 8/19 - \$45 for 3 courses + beer pairings. http://trunc.it/172wt
5 Interesting Concept-Beer Petitions.com launches&hopes 2help craft beer drinkers enjoy beer they want @their fave pubs. http://bit.ly/11gJQN

tendency, that the classical Lucene implementation is rarely preferred over another setup. Further we note strong preference towards the Retweet-odds based reranking on shorter and underspecified queries, while on longer queries the users prefer the relevance based ranking over the Retweet-odds reranking approach. This corresponds to the observations made for the objective evaluation setting.

5. CONCLUSIONS AND SUMMARY

In this paper we looked into the particularities of information retrieval on microblogs: sparsity and document quality. Sparsity is inherent to microblog documents, as it reflects the technical constraints on the message length. The quality of a document with respect to its ability to satisfy an information need originates from the different purposes and environments in which microblog messages are generated. We motivated from a theoretical and data analytical point of view, that document length normalization introduces an unmotivated bias towards short documents in microblog retrieval. Further, we introduced interestingness as a static quality measure for microblog messages. We empirically showed, that both approaches improve retrieval performance

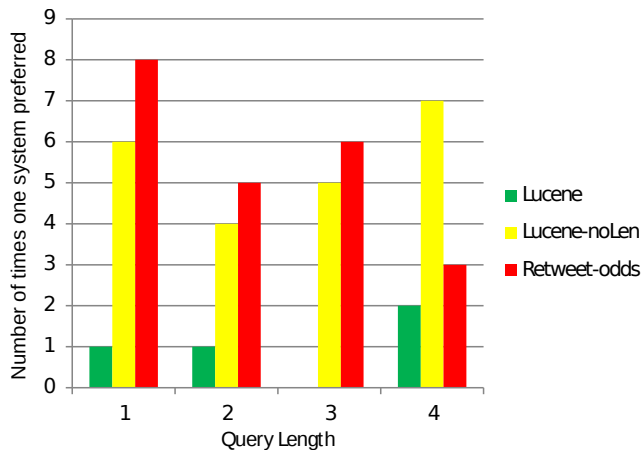


Figure 4: Number of times each system is preferred over the others for different query lengths.

Table 4: User preference of systems.

Opposed	Preferred system		
Q1	Lucene	Lucene-noLen	Retweet-odds
Lucene	–	4	5
Lucene-noLen	1	–	3
Retweet-odds	0	2	–
Q2	Lucene	Lucene-noLen	Retweet-odds
Lucene	–	3	2
Lucene-noLen	0	–	3
Retweet-odds	1	1	–
Q3	Lucene	Lucene-noLen	Retweet-odds
Lucene	–	2	4
Lucene-noLen	0	–	2
Retweet-odds	0	3	–
Q4	Lucene	Lucene-noLen	Retweet-odds
Lucene	–	3	2
Lucene-noLen	0	–	1
Retweet-odds	2	4	–

in the sense of providing more relevant and generally interesting messages in the search results.

As a next step, we plan to incorporate further static quality measures in the process. These will include the social context of a user, the global network structure and the freshness of results.

Acknowledgments

The research leading to these results has received funding from the European Community’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 257859, ROBUST and grant agreement no. 248512, WeGov. This work was supported in part by HEC, Govt. of Pakistan in collaboration with DAAD, Germany. This work was also partly funded by the German Research Foundation (DFG) under the Multipla project (grant 38457858).

6. REFERENCES

- [1] O. Alonso and R. A. Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Proc. European Conf. on Information Retrieval*, pages 153–164, 2011.
- [2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *J. of Machine Learning Research*, 3:993–1022, 2003.
- [3] M. M. Bradley and P. J. Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report, The Center for Research in Psychophysiology, University of Florida, 1999.
- [4] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring user influence in Twitter: the million follower fallacy. In *Proc. Int. Conf. on Weblogs and Social Media*, pages 10–17, 2010.
- [5] A. Che Alhadi, S. Staab, and T. Gottron. Exploring user purpose writing single tweets. In *Proc. Web Science Conf.*, 2011.
- [6] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *Proc. Conf. on Weblogs and Social Media*, pages 34–41, 2010.
- [7] L. Hong, O. Dan, and B. D. Davison. Predicting popular messages in Twitter. In *Proc. Int. World Wide Web Conf.*, pages 57–58, 2011.
- [8] M. Karam, T. Manos, d. R. Marten, and W. Wouter. Incorporating query expansion and quality indicators in searching microblog posts. In *Proc. European Conf. on Information Retrieval*, 2011.
- [9] E. Kim, S. Gilbert, M. J. Edwards, and E. Graeff. Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter. Technical report, Web Ecology Project, Aug 2009.
- [10] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *Proc. Int. World Wide Web Conf.*, pages 591–600, 2010.
- [11] R. Nagmoti, A. Teredesai, and M. De Cock. Ranking approaches for microblog search. In *Proc. Int. Conf. on Web Intelligence and Intelligent Agent Technology*, pages 153–157, 2010.
- [12] N. Naveed, T. Gottron, J. Kunegis, and A. Che Alhadi. Bad news travel fast: A content-based analysis of interestingness on twitter. In *Proc. Web Science Conf.*, 2011.
- [13] S. Petrović, M. Osborne, and V. Lavrenko. The Edinburgh Twitter corpus. In *Proc. Workshop on Computational Linguistics in a World of Social Media*, pages 25–26, 2010.
- [14] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proc. Int. Conf. on Research and Development in Information Retrieval*, pages 21–29, 1996.
- [15] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: a comparison of microblog search and Web search. In *Proc. Int. Conf. on Web Search and Data Mining*, pages 35–44, 2011.
- [16] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential twitterers. In *Proc. Int. Conf. on Web Search and Data Mining*, pages 261–270, 2010.
- [17] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In *Proc. European Conf. on Information Retrieval*, pages 338–349, 2011.