University of Glasgow (UoG_TwTeam) at TREC Microblog 2013

Jesus A. Rodriguez Perez, Andrew J. McMinn, and Joemon M. Jose
{j.rodriguez-perez.1, a.mcminn.1}@research.gla.ac.uk, joemon.jose@glasgow.ac.uk

School of Computing Science
University of Glasgow
Glasgow, UK
G12 8QQ

ABSTRACT

In TREC 2013, we participated in the ad-hoc search task of the Microblog Track. The Microblog track, which is in its third consecutive year, has remained very similar to the last two. This paper describes the approaches we have implemented for Tweet retrieval, which comprehend query expansion, and baseline model selection. The results for all the runs submitted are well above the median achieved for all the automatic runs submitted to TREC. Furthermore, statistically significant improvements in terms of precision at 30, are reported for our automatic query expansion approach with respect to the baseline we chose. The methods here proposed show great potential for enhancing tweet retrieval performance and should therefore be further studied.

1. INTRODUCTION

Microblogs have grown in popularity in recent years, and is gradually transforming the way we communicate with each other, share information, and find out about events. Twitter is a particularly interesting service because it encourages users to discuss events in real-time, and often includes first hand reports of an event as it is developing. This allows for a unique insight into events from the subjective opinions of those directly involved to the discussion between others following the event through social and traditional media.

Ad-hoc retrieval is one of the the most commonly researched tasks in Information Retrieval, where the goal is to return documents that are relevant to an immediate information need, expressed as a query. Tweets are limited to 140 characters in length, and over 340 million tweets are posted every day. Tweets are generally of a low quality as they contain bad grammar or spelling mistakes, and slang is often used to overcome their restricted length.

The very particular characteristics of tweets, makes searching in Twitter a very challenging task, causing traditional ranking models, such as TF-IDF [10], to experience difficulty in retrieving the most relevant documents, as the assumptions in which they were designed may not make sense anymore.

Microblog documents are a few words long, therefore term frequency information becomes limited and sometimes misleading. In addition, the **vocabulary mismatch problem** exacerbates the difficulty of query term matching during retrieval. This problem relates to the vocabulary difference between the query issued and the relevant documents to be retrieved. Due to the limited length, the probability of microblog users utilizing the same terminologies when posting about a particular topic are slim.

In the literature, two main approaches which attempt to bridge the gap between query and document representations have been proposed, namely **automatic query expansion** (AQE) and **document expansion**. Query expansion, refers to the process of enriching an initial query with terms, mined either from an initial set of retrieved documents, or obtained from some external source. Document expansion, on the other hand, attempts to enhance the representation of each of the documents with information from external sources. Common approaches to document expansion include: extracting information from the web pages that microblog documents often point to; finding associated terms in Wikipedia; or mining features from the results list obtained by submitting the query to a commercial search engine.

The contributions of this work include a baseline model selection approach, a novel yet simple approach to term selection in pseudo relevance feedback based AQE, and a clustering based AQE approach. All our approaches have the objective of alleviating the **vocabulary mismatch problem**, by expanding the representation of the queries.

For all our runs, parameters are optimized on the two previous Microblog collections from 2011 and 2012, thus avoiding over-fitting. The results obtained for the 2013 collection show the great potential especially in the case of our automatic query expansion approach.

The rest of the paper is organised as follows: First we briefly cover relevant literature and introduce the concepts utilized throughout this work (Section 2). Then we introduce our approach to selectively using a model given particular conditions, followed by our automatic query expansion methodologies (Section 3). Section 4 sets the evaluation environment in which our experiments are carried out. Finally we present the results and discuss the findings in Section 5 and the document is finalized in section 6 with conclusions and proposed future work.

2. BACKGROUND

In this section we cover the related work, and introduce the background concepts utilized across this work.

2.1 Retrieval Models And Task

Ad-hoc Retrieval tasks are the most commonly studied tasks in IR. The main goal is to retrieve documents from a collection that match a single-aspect information need, expressed as a textual query. The set of retrieved documents are ranked in terms of the decreasing probability of satisfying the information need, as estimated by a retrieval model, and shown to the user. The user model is a simplification under the assumption that a user sequentially evaluates the documents in the result set starting from the top.

Probability Ranking Principle (PRP) is a core concept in Information Retrieval (IR) first introduced by [17]. PRP states that documents should be ranked and presented to the user, based on a document's estimated probability for being relevant given a query. Documents in a search result list are organised in decreasing order of probability.

Inverse Document Frequency (*IDF*) is an estimation of the discriminatory power of a query term. That is, a term q_i is more discriminatory of a particular document d than another term q_j if it is less likely to appear in any documents than term q_j. Therefore the highest IDF scores for a term in any given collection, belong to those terms which appear in a single document.

$$\mathrm{IDF}_{t,D} = \log_2 \left(\frac{|\mathrm{D}|}{|d \in D : t \in d|} \right), \quad (1)$$

where t is the current term, D is the set of all documents in the collection and $|d \in D : t \in d|$ is the number of documents in which term t occurs. *IDF* is commonly used as a component of retrieval models such as TF*IDF [19]. IDF weighting is used as part of the term selections approaches presented in this work.

Other retrieval models. Other modern and representative retrieval models include Divergence From Randomness (DFR) [16], Okapi BM25 (BM25) [18], and Hiemstra's Language Model [9].

Microblog Retrieval Issues. Work by [22] studied the effects that preprocessing had in performance. Their findings showed that the best performance was achieved when applying all preprocessing steps, which include language detection; Emotion removal; Lexical normalization; Mention Removal and Link Removal.

The work by [5, 15] have identified two main problems affecting retrieval models in microblogs, namely **term frequency** and **document length normalization**. Document length normalization [20] has been used by retrieval models to counterbalance the effects of longer documents, which may not necessarily add any new information to a topic, but are more likely to contain higher term frequencies.

Work by [5] studied the effects of query length normalization and term frequency on the BM25 retrieval model whereas [15] looked at these issues from a more generic perspective. Both their findings showed how query length normalization had an undesirable effect on retrieval performance of the systems they were evaluating, since the notion of document length does not have a meaning in microblogs.

2.2 Vocabulary Mismatch Problem

Most retrieval problems in Twitter can be traced back to the **"vocabulary mismatch problem"**, which has been studied as early as in 1987 by [6]. The term mismatch problem, refers to the mismatch between the query representation formulated by a human, and that of relevant documents to be retrieved from the document collection. In the context of microblogs, these two problems are highly accentuated due to the limited information contained within the documents, which often leads to poor matching.

Table 2 shows evidence of this problem for the collections considered. As we can observe, only about half of the documents for a given topic contain any of the original query terms in any of the collections. This shows how deep the repercussions of the **term mismatch problem** in the context of microblogs are, and motivated the work presented in this paper towards alleviating it.

2.3 Query Expansion

Query expansion. Automatic query expansion approaches (AQE) have been the focus of research efforts for many years. Work by [4] does a comprehensive study about these approaches, giving insight on the challenges that these approaches face. Most importantly it introduces critical issues such as parameter setting, efficiency and usability of the approaches which has greatly contributed to the design of our own query expansion approach based on Pseudo Relevance Feedback. In their work they propose a a comprehensive description of the steps involved in any query expansion approach. The following is a description of these four steps in the context of our work:

1. **Preprocessing of Data Source:** In our work this comprehends the tokenization, stop word removal and stemming of those terms found in the initial set of retrieved documents.

2. **Generation and Ranking of Candidate Expansion Features:** This stage refers to estimating the relatedness of terms found in the initial set of retrieved documents, with respect to the initial query. The approaches proposed in this work, specifically target this step.

3. **Selection of Expansion Features:** After the ranking of terms, a number of top ranked terms is selected following a given policy. In our approaches we simply set a maximum number of terms to be utilized.

4. **Query reformulation:** At this stage, terms are added to the initial query following a policy, and normally weights are assigned.

In this work we often refer to "term selection" as the process of ranking and selecting the terms for query expansion, thus involving steps 2 and 3.

Pseudo Relevance Feedback An important concept in query expansion approaches is that of Pseudo Relevance Feedback (PRF) [24]. PRF is a technique by which the top N documents retrieved as a response to a query, are assumed to be relevant. In this work we nominate this set of top N retrieved documents the "Pseudo relevant set". PRF is most often used as part of AQE approaches, as a lightweight and reliable feature source. Algorithm-dependent features

are extracted from the pseudo relevant set (Top N retrieved documents) as to determine the best terms for expansion, in relation with the query.

A known challenge with PRF-based approaches to query expansion is that, they are unstable due to their dependency on top results, as we have no warranties of finding relevant documents. However it has experienced wide use in Microblog retrieval as this approach has been proven to perform effectively in average by previous work such as [23] and [12].

Query Expansion in Microblog Retrieval. Numerous participants including the top performing ones in both 2011 [2, 13, 14] and 2012 [11, 1, 8] TREC Microblog tracks employed QE methods for the ad-hoc task, reporting significant improvements on retrieval effectiveness. However these approaches often fail to filter unrelated terms to the original query, which ultimately can hinder retrieval effectiveness specially for those topics performing badly from the start.

Some approaches utilise external evidence in order to find new terms. The work by [7] successfully used Wikipedia as a source of query expansion terms by finding associations within terms in the articles and those of the original query. A different approach is that proposed by [3], which used a query to search on a commercial search engine such as Google or Bing, and extracted prospective terms from the generated result list.

Document expansion. Document expansion in the context of microblogs is a commonly used technique. Microblogs are very short in length, and often it means that the information contained within is insufficient to make an informed retrieval decision. Document expansion, attempts to add content to the documents from external sources.

The most common approach is to leverage the content pointed by the links contained within the tweets. Work by [11] successfully used the meta-data contained in html pages, such as titles and keywords, to extract prospective terms to enrich each document's representation.

3. APPROACHES

3.1 Model Selection

As it can be clearly observed in Table 1, the two best baseline systems across all collections are DFR and IDF. However there are a number of topics for which IDF performs better than DFR, and vice versa.

Our first approach exploits the differences in pre-retrieval features for each set of topics in which IDF outperforms DFR. These differences are utilized as a selection mechanism which chooses the most appropriate retrieval model given a particular topic.

We experimented with multiple pre-retrieval features, including: query posting sizes, idf values, number entity query terms, query length, etc. The feature that showed most promise is that of the difference in IDF between query terms (IDF_Diff). This is computed as the difference between the highest and lowest IDF scores found in the query.

We found a statistically significant difference, in terms of IDF_Diff in those topics for which IDF out-performed DFR and vice versa. We exploited this difference to select the best performing retrieval model depending the query being issued. To this end IDF is selected when IDF_Diff ranges between 0.09 and 0.22, using DFR otherwise.

Table 1: retrieval results for the Traditional models on each collection

	P@5	***P@10***	***P@15***	***P@20***	***P@30***	***MAP***
	2011					
BM25	0.510	0.467	0.440	0.415	0.370	0.26
DFR	0.600	**0.565**	**0.519**	**0.486**	0.425	0.32
HLM	0.571	0.508	0.466	0.439	0.406	0.32
IDF	**0.620**	0.540	0.503	0.476	**0.444**	**0.36**
	2012					
BM25	0.386	0.372	0.343	0.334	0.305	0.11
DFR	**0.433**	**0.413**	**0.375**	**0.356**	**0.341**	0.14
HLM	0.410	0.378	0.352	0.338	0.327	0.13
IDF	0.410	0.378	0.354	0.338	0.318	**0.14**
	2013					
BM25	0.543	0.481	0.438	0.417	0.377	0.14
DFR	0.630	**0.586**	**0.532**	**0.506**	**0.440**	**0.18**
HLM	0.366	0.325	0.304	0.290	0.263	0.08
IDF	**0.633**	0.573	0.507	0.484	0.425	0.18

3.2 Automatic Query Expansion

Pseudo relevance feedback based query expansion approaches work under the intuition that top retrieved documents are most likely related to the issued query. Thus terms within them have increased probabilities to retrieve documents within that particular topic. Furthermore, within the top N retrieved documents, terms closer to the top of the ranking should be more related than those at the bottom of the top N documents.

We hypothesise that, taking this evidence into consideration should further enhance the term selection process, thus achieving better performance, and more robustness for PRF-based AQE algorithms.

In this subsection we detail an AQE baseline, as well as our approach modelling the decrease in relatedness of top retrieved documents and a clustering approach designed as an enhanced source of related terms.

3.2.1 Base Term Ranking Approach

The following formula provides a common starting point to defining a term scoring approach for PRF-based query expansion. This strategy can be formalized as:

$$QEScore(t, P) = \sum_{r=0}^{|P|} cf(D_r, t), \tag{2}$$

$$cf(D_r, t) = \begin{cases} score(t, r) & \text{if } TF(t, D_r) \geq 1 \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where t is the term being scored, P is the pseudo relevant set of retrieved documents and D_r is the document at rank r. TF is the term frequency of t in the document D_r which conditions the value obtained by the conditional function $cf(D_r, t)$. Finally $score(t, r)$ is a place-holder scoring function which is given by each of the approaches proposed in the remainder of this section.

3.2.2 Baseline: QE_Baseline

The AQE baseline here presented utilizes IDF values to estimate the relevance of terms in the pseudo relevant set. The scoring function if formalised as follows:

$$score(t, r) = QE\_Baseline(t, r) = idf(t), \quad (4)$$

where $idf(t)$ is the IDF score for term t.

Notice that when this scoring function is inserted in Equation 3, the scoring approach is equivalent to computing $TF \times IDF$ over the pseudo relevant set P.

3.2.3 Linearly Discounted: QE_LinearDisc

PRF-based query expansion approaches often rely on term frequency as one of the determining components to discern between related and unrelated terms. In the context of Microblogs, the information contained in documents is limited, and the term frequency assumptions are often not be enough to provide effective term selection, even when treating the whole set of retrieved documents as a unit.

The PRP principle states that the higher the rank, the higher the likelihood of a document to satisfy a user's information need. As a consequence, it is more likely to find better terms for query expansion in the top ranks, than in lower ranks. In order to model this assumption we propose a linear discounting function that reduces the contribution to the total score of terms, as they are found in lower parts of the rank.

In order to model the decreasing value of those terms extracted from tweets located lower in the result set, we employ a linear discounting function dependant only on the document's rank. The score of each candidate term, is the sum of IDF values, whenever it appears in the result set. However the IDF value will be increasingly reduced by the discounting function as we approach the lower ranks of the result set. The QE_LinearDisc approach can be formalized as follows:

$$score(t, r) = QE\_LinearDisc(t, r) = \frac{idf(t)}{r}, \quad (5)$$

The score given for a term at each ranking position r, is linearly reduced by the division by r. The effect of this approach on lower ranks is very strong, thus practically neglecting any information about terms found in the mid-lower range of the result list, in favour of those at the top.

The term selection approach here proposed, builds directly upon the above mentioned QE_baseline as it provides an appropriate and fair point of comparison. However, this discounting term selection policy could be employed as part of any PRF-based AQE approach.

3.3 Entity-based Online Clustering

We believe that entities are a good indication of the subject of a tweet, and use them as a partition so that only tweets which discuss at least one shared entity can be clustered together. This has the effect of significantly reducing the average number of comparisons needed when calculating nearest neighbour, and can be efficiently implemented by maintaining an inverted index for each entity.

As each tweet d arrives, named entities, nouns and verbs are extracted. For each named entity e in the tweet, a list of tweets D is retrieved from the inverted index (the inverted index for e does not contain e) and the maximum TF-IDF weight cosine similarity score is calculated between d and each tweet in D. In order to ensure that our approach is able to run in real-time, we limit the number of tweets which can be retrieved from the inverted index to a fixed number per term, and use only the top 10 TF-IDF weighed terms per tweet. This limited the number of retrieved tweets to $10N$, where N is usually in the range 100-1000 (for our runs, we used 500).

If the maximum score is above a set threshold (in the range $0.3 - 0.7$, using 0.5 for the submitted run), then d is added to the same cluster as its nearest neighbour. If the nearest neighbour does not already belong to a cluster, then a new cluster is created containing both tweets and assigned to entity e. The new tweet is then added to the inverted index for entity e. Algorithm 1 shows the pseudo-code for our entity based cluster approach.

Algorithm 1: Entity-based method of clustering

```
foreach tweet d in corpus do
    foreach entity e in d do
        foreach term t in d do
            foreach tweet d' in index_e that contains t
            do
                update score(d, d');
            end
        end
        score_max(d) = max_d'{score(d, d')};
        if score_max(d) >= threshold then
            add d to cluster(d');
        end
        add d to index_e;
    end
end
```

The clusters formed through this technique are then utilized as a source of related terms for AQE. All clusters were indexed and treated as individual documents. In this way we could then apply the same AQE described above, in particular our discounted IDF. Our expectation is that the enhanced representation in the shape of clusters and the terms within, should provide a stronger evidence in finding related terms with respect to the issued queries.

4. EXPERIMENTAL SETTING

In this section we describe the experimental procedure settings, which includes any pre-processing of data as well as tools utilized in indexing and any other experimental conditions.

Datasets. In this evaluation we have used the three collections from the TREC Microblog track. The 2011 and 2012 collections share the same corpus but have very different relevance assessments. The 2013 corpus is an order of magnitude bigger than previous ones. The relevance assessments for the 2013 corpus are however comparable to those of the 2012 track, in terms of size. Moreover the ratio of relevant documents with respect to non-relevant ones in the relevance judgements, is much more positive for the 2013, than for the 2012 and 2011, which should provide a higher chance for matching relevant documents. Furthermore, there are considerably more relevant documents per

Table 2: Description of 2011, 2012 and 2013 collections.

	2011	2012	2013
Number of topics	50	60	60
Total docs in collection	16M	16M	260M
Total assessed docs	40855	73073	71279
Total assessed non-relevant	38124	66893	62268
Total assessed relevant	2731	6180	9011
Ratio Rel/NonRel	0.07	0.09	0.14
Avg rel. docs per topic	58.45	106.54	150.18
Avg % rel. docs w/query terms	48.33	50.90	52.42

topic in average for the 2012 and 2013 collections than what can be found in the 2011 corpus. In total there are 170 topics to evaluate our approaches on, where queries lengths range between 2 and 3 terms.

Spam and Language Filtering We performed basic filtering to remove spam and non-English Tweets from the corpora. Tweets which contained more than 3 hash-tags, 3 mentions, or 2 URLs were classified as spam and removed. This decision was based upon statistical observations of the corpus, and thorough careful examination of the most common types of spam on Twitter.

Non-English tweets were removed based upon two factors: (1) the author's primary language, and (2) the language of the tweet. For each individual tweets a Java language-detection library[1] was employed. This library uses naive Bayesian filtering and claims over 99% precision for 53 languages.

Indexing. The collections were indexed using the Lucene IR platform [2]. Lucene is an open-source IR platform maintained by Apache, and supported by a very strong community. This year at TREC they decided to offer access to a common Lucene index through a web-service for all participants. This approach eliminated the differences across collections which conditioned the results of previous years. However, this approach was somewhat limiting since the experimenter has no total control over the retrieval process.

Evaluation. We pay attention to Precision at different ranks, with a cut-off point at rank 30, following TREC's Microblog track guidelines. Future evidence is accepted at the collection statistics level, but any document published after the query issuing time is disregarded even if it is relevant. This consideration was carefully studied by TREC organisers to simplify the evaluation procedure and deemed acceptable, having a minimal effect on evaluation results.

5. RESULTS AND DISCUSSION

In this section, we present and discuss the results obtained for each of our experiments in the Microblog 2013 Track.

Figure 1 shows an interpolated precision-recall plot, which gives an overview of the performance achieved by the systems here evaluated. In this figure the differences in performance are not particularly evident, however we do observe a substantial difference between the QE approaches with respect to the DFR and SEL922 baselines.

[1]http://code.google.com/p/language-detection/
[2]http://lucene.apache.org/core/

Figure 1: Interpolated Precision-Recall for all approaches considered

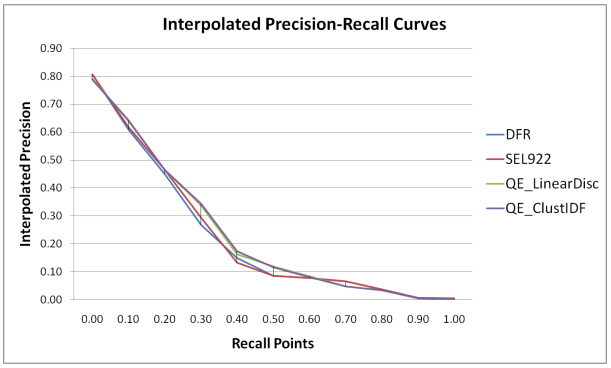

Table 3: Evaluation of systems on the 2013 collection (* denotes significant differences ($p < 0.05$) with respect to the baselines, and (NS) denotes non-submitted runs)

	P_5	P_10	P_15	P_20	P_30
DFR	0.653	0.605	0.558	0.529	0.465
SEL922	0.663	0.615	0.567	0.538	0.474
QE_Baseline(NS)	0.613	0.602	0.580	0.552*	0.501*
QE_LinearDisc	**0.667**	**0.633***	**0.607***	**0.564***	**0.512***
QE_ClustIDF	0.663	0.630*	0.604*	0.563*	0.509*

Table 3 goes more in detail, by showing retrieval metrics for all approaches here described, in terms or precision at different cut-off points.

SEL922 represents our selective model approach which switches between IDF and DFR whenever the difference in IDF scores within the query is between the previously described thresholds. Even thought the differences at each cut-off point are not statistically significant, it can be seen that the selection approach gives a stable boost in performance across all precision metrics, thus making the overall retrieval approach more robust. This approach shows very promising results and we believe that statistical significance could be obtained provided with a larger set of topics, since only on a subset of topics were involved in this boost out of the total 60.

The QE_LinearDisc in the results table (Table 3), represents our discounting IDF policy to term selection within a PRF-based The results for the QE_Baseline, can be observed to improve performance at bottom ranks, particularly at 20 and 30 cut-off points. The results obtained for our linearly discounted approach (QE_LinearDisc) to term selection outperform that of QE_Baseline. This boost in performance can only be explained by the better decision making when selecting terms to be included in the expanded query.

Unlike other AQE approaches which sacrifice precision at higher ranks to produce a higher recall at lower ranks, (Such as QE_Baseline) QE_LinearDisc boosts the performance on the top 10 retrieved documents. This is a highly desired feature since the user model agreed for Twitter ad-hoc searches estimates that users do not read beyond the top 30 retrieved documents. Thus increasing probability of finding relevant documents at the top ranks is essential in this particular context.

Our last experiment concerns the use of clustering to en-

hance the evidence of features, increasing the signal of related terms. The results obtained for the QE_ClustIDF run, are virtually the same as those obtained for QE_LinearDisc. The reason behind these results is that the clusters contained very similar information to that already present in top retrieved tweets, thus the same terms where selected using QE_LinearDisc over the clusters when expanding the original query.

6. CONCLUSION

In this paper, we have presented a selective model approach that will use the IDF retrieval model instead of DFR given the appropriate conditions, producing an overall boost in retrieval performance. Moreover we have addressed the vocabulary mismatch problem by proposing two PRF-AQE techniques. Particularly, we have introduced a term selection approach QE_LinearDisc, which promotes those terms found at the very top tweets of the pseudo relevant set of documents. This approach outperformed the given baselines, bringing statistically significant improvements in terms of precision, which translated in superior performance especially for the very top cut-off points. Finally our assumptions for the clustering approach (QE_ClustIDF) to enhancing the best candidate term signals, resulted in equivalent performance to that already achieved by QE_LinearDisc, thus producing no differences.

In future work, we would like to explore different discounting functions to find which ones produce a better fit to the data provided. Furthermore we would like to test our discounting IDF term selection approach within other existing AQE methodologies.

7. REFERENCES

[1] Y. Aboulnaga, C. L. A. Clarke, and D. R. Cheriton. Frequent itemset mining for query expansion in microblog ad-hoc search.

[2] G. Amati, G. Amodeo, M. Bianchi, G. Marcone, F. U. Bordoni, C. Gaibisso, G. Gambosi, A. Celi, C. Di Nicola, and M. Flammini. Fub, iasi-cnr, univaq at trec 2011 microblog track. In *TREC*, 2011.

[3] A. Bandyopadhyay, K. Ghosh, P. Majumder, and M. Mitra. Query expansion for microblog retrieval. *International Journal of Web Science*, 1(4):368–380, 2012.

[4] C. Carpineto and G. Romano. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1, 2012.

[5] P. Ferguson, N. OâĂŹHare, J. Lanagan, O. Phelan, and K. McCarthy. An investigation of term weighting approaches for microblog retrieval. In *Advances in Information Retrieval*, pages 552–555. Springer, 2012.

[6] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[7] D. F. Gurini and F. Gasparetti. Trec microblog 2012 track: Real-time algorithm for microblog ranking systems. 2012.

[8] Z. Han, X. Li, M. Yang, H. Qi, S. Li, and T. Zhao. Hit at trec 2012 microblog track. *TREC Microblog 2012*, 2012.

[9] D. Hiemstra. *Using language models for information retrieval*. Taaluitgeverij Neslia Paniculata, 2001.

[10] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21, 1972.

[11] Y. Kim, R. Yeniterzi, and J. Callan. Overcoming vocabulary limitations in twitter microblogs. *TREC Microblog 2012*, 2012.

[12] C. Lau, Y. Li, and D. Tjondronegoro. Microblog retrieval using topical features and query expansion. *TREC Microblog Track 2011*, 2011.

[13] Y. Li, Z. Zhang, W. Lv, Q. Xie, Y. Lin, R. Xu, W. Xu, G. Chen, and J. Guo. Pris at trec 2011 microblog track. In *TREC*, 2011.

[14] D. Metzler and C. Cai. Usc/isi at trec 2011: Microblog track. In *Proceedings of the Text REtrieval Conference (TREC 2011)*, 2011.

[15] N. Naveed, T. Gottron, J. Kunegis, and A. C. Alhadi. Searching microblogs: coping with sparsity and document quality. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 183–188. ACM, 2011.

[16] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and D. Johnson. Terrier information retrieval platform. In *Advances in Information Retrieval*, pages 517–519. Springer, 2005.

[17] S. Robertson. The probability ranking principle in ir. *Journal of documentation*, 33(4):294–304, 1977.

[18] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, pages 109–109, 1995.

[19] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513 – 523, 1988.

[20] A. Singhal, C. Buckley, and M. Mitra. Pivoted document length normalization. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 21–29. ACM, 1996.

[21] J. Teevan, D. Ramage, and M. Morris. # twittersearch: a comparison of microblog search and web search. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 35–44. ACM, 2011.

[22] S. K. J. Y. P. Thomas. Searching and filtering tweets: Csiro at the trec 2012 microblog track.

[23] S. Whiting, Y. Moshfeghi, and J. M. Jose. Exploring term temporality for pseudo-relevance feedback. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, SIGIR '11, pages 1245–1246, New York, NY, USA, 2011. ACM.

[24] J. Xu and W. B. Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996.