# Effective Pseudo-Relevance for Microblog Retrieval

Khaled Albishre
Science and Engineering Faculty
Queensland University of Technology (QUT)
Brisbane, Australia
Umm Al-Qura University
Makkah, Saudi Arabia
khaled.albishre@hdr.qut.edu.au

Yuefeng Li
Science and Engineering Faculty
Queensland University of Technology (QUT)
Brisbane, Australia
y2.li@qut.edu.au

Yue Xu
Science and Engineering Faculty
Queensland University of Technology (QUT)
Brisbane, Australia
yue.xu@qut.edu.au

## ABSTRACT

Microblog services such as Twitter have become a part of daily life for many users, with thousands of documents published each second. Microblog documents are often too short, overwhelming in their use of informal language and hard to understand due to a lack of contextual clues. Retrieving relevant documents from microblogs is somewhat challenging because of its nature and the massive scale of the data. However, microblog retrieval models suffer from a vocabulary mismatch problem that leads to insufficient performance. In this paper, we address microblog retrieval limitations by proposing a pseudo-relevance feedback model. Our model considers discriminative expansion to meet user interests. Experimental results on TREC 2011 and 2012 microblog datasets show that our model demonstrates significant improvements over the baseline models.

## CCS Concepts

•**Information systems** → **Information retrieval query processing;** *Document topic models; Language models;*

## Keywords

language model; topic model; microblog search; query expansion

## 1. INTRODUCTION

Different contexts are posted daily in social media. These contexts contain images, texts, videos, and hyperlinks. Social media plays a significant role in exchanging text data. The textual data format in social media presents great opportunities for retrieving useful information from these data. Twitter is a famous example of a microblogging service, which has 313 million users per month and features 500 million short messages called 'tweets' that are posted everyday in more than 35 languages[1]. Microblog texts feature some unique aspects, such as time sensitivity, short length, unstructured phrases, and abundant information, which are not found in traditional texts [2]. A significant challenge of social media is its real-time nature.

Many queries are submitted through social media platforms to obtain useful information. These search queries are usually short. Thus, many search results are not often highly relevant to the query keywords. Word mismatch indicates that users regularly use several words to characterize ideas in their queries that a user may use to depict the same ideas in their documents [32, 17]. The essential issues in ambiguity stem from hyponymy and synonymy. Various approaches for handling word mismatch have long been considered [16, 15, 22, 14, 11]. Therefore, taking into account the query words will improve the relevance of retrieved documents.

A well-known methodology for managing vocabulary mismatch is query expansion. The objective of query expansion is to extend a first, unsuccessful query with different words that best catch the user's purpose, or that create a relevant query that will likely retrieve more significant documents [5]. The method is especially viable when the writer's query is ambiguous, short, or needs useful words relating to the expected topic. This procedure can be automatic through pseudo-relevance feedback or it can be manual using explicit relevance feedback.

The relevance feedback technique is a repetitive procedure, in which the user evaluates the importance of documents returned as an initial response to a given query [21, 24]. The user ascertains the preliminary results and marks whether or not the document is relevant. Another query is submitted based on the user feedback from the original query, extended with related words taken from the outcomes. This strategy, called explicit feedback, is repeated until the user is satisfied with the outcomes. Although relevance feedback can be an efficient approach, it puts an enormous weight on the user. Pseudo-relevance feedback (also called implicit feedback, blind feedback, local feedback, or ad hoc) is a method intended to assume what the user may find fascinating without having the user explicitly state the outcomes [30]. Pseudo-relevance feedback assumes that the

[1]http://www.about.twitter.com



*ACSW '17, January 31-February 03, 2017, Geelong, Australia*



DOI: http://dx.doi.org/10.1145/3014812.3014865

top results have a higher accuracy and features those results that are expected to show the user's research topic.

It is hard to apply traditional information retrieval methods to microblogging services such as Twitter without considering the nature of the data. Twitter users try to discover temporarily relevant information such as breaking news and current events or information about other people. Moreover, it seems that users rephrase Twitter queries to observe the related outcomes after several attempts at querying [29].

This paper presents a pseudo-relevance feedback, including relevance feedback and topic-based query expansion for microblog retrieval. The proposed model combines the lexical and topical evidence from pseudo feedback with respect to the original query. The significant benefit of the proposed model is stability and robustness because it does not need any external resources to expand the original query.

The remainder of this article is structured as follows. Section 2 presents a detailed overview of the related works. Section 3 shows the lexical evidence using the Language Model. Section 4 discusses the proposed model. To evaluate the performance of the proposed model, we conduct substantial experiments on the TREC 2011 and 2012 microblog dataset. The empirical results and discussion are reported in Section 6, followed by concluding remarks in the last section.

## 2. RELATED WORK

A critical challenge of microblog retrieval is the vocabulary mismatch problem between query and document. Previous studies addressed the vocabulary mismatch problem from the document and query expansion perspective. Due to the real-time microblog nature, temporal information was extensively presented in previous studies.

Efron proposed a strategy for Twitter document expansion based on pseudo-relevance feedback using lexical and temporal feedback and found that retrieval using query expansion with temporal information resulted in improving the relevance retrieval efficiency [8]. A different technique to expand microblog documents was presented in [10]. El-ganainy and Rafea expanded documents containing hyperlinks with corresponding titles of the hyperlinked documents. In Liang et al. [18], the authors used Twitter-specific features to expand the original query. Their model consists of two steps: first, using the URL for document expansion, and then using temporal features to re-rank the documents. The model showed outperformance results of 26.7% and 9.94% based on P@30 and MAP, respectively. Similarly, Efron [7] used a different strategy for hashtag order to expand the original query. However, the dataset that was used to evaluate the model was not big enough to validate the model results. Miyanishi, Seki, and Uehara proposed relevance feedback model using temporal and lexical feedback by manual tweet selection [22].

Temporal evidence has been widely used in previous microblog retrieval studies, showing that time is important to reflecting the relevance feedback. Efron and Golovchinksy incorporated temporal evidence from the top-ranked document to estimated the rate parameter on query likelihood and use pseudo-relevance feedback to estimate expansion query [9]. Li and Croft incorporated temporal information from pseudo-relevance feedback into the relevance model to enhance query expansion [6]. They selected a period based on user behavior (e.g. retweet) to extract relevant tweets that could be used to expand the original query. Wang and Zhang proposed a microblog retrieval model and query expansion considering lexical and temporal feedback [30].

The major difference between our study and previous studies is how we use pseudo-relevance feedback. The majority of the previous studies utilized lexical expansion from temporal evidence. In our work, we identify the lexical evidence for a given query beside the topical expansion query terms. Moreover, our proposed approach does not involve any external evidence from a knowledge base such as Wikipedia or Freebase.

## 3. LEXICAL EVIDENCE WITH LANGUAGE MODEL

### 3.1 Query Model

The Lexical Evidence with Language Model is a special case of probabilistic retrieval, and the state-of-the-art model for language model is query likelihood. Query likelihood model proposed by Ponte and Croft [25], which assumes the probability of relevance model $P(Q|D)$, can be generated using the probabilities of query's features $Q$ given document $D$. Thus, documents are ranked based on the posterior probability $P(D|Q)$ using Bayes rule:

$$P(D|Q) \propto P(Q|D)P(D) \tag{1}$$

where $P(D)$ is the prior probability that $D$ is relevant to any query and $P(Q|D)$ is the query likelihood on the given document $D$.

The multinomial query likelihood model $P(Q|D)$ is described as follows:

$$P(Q|D) = \prod_{i=0}^{|Q|} P(w_i|D) \tag{2}$$

where $|Q|$ is the the number of query's feature and $P(w|D)$ is the relevance model that computes the probability of word $w$ based on its distribution in document $D$. Different smoothing techniques are proposed in Information Retrieval literature, and an effective technique is presented in Zhai and J. Lafferty [33, 34]. They used Bayesian smoothing for their language model using Dirichlet priors, as follows:

$$P(w|D) = \frac{|D|}{|D|+\mu}\frac{c(w,D)}{|D|} + \frac{\mu}{|D|+\mu}P(w|\mathcal{C}) \tag{3}$$

for $\mu \in [0, +\infty)$. Where $c(w, D)$ is the word frequency in the document, $p(w|\mathcal{C})$ is the collection language model, and $\mu$ is the smoothing parameter. The final form of query likelihood using Dircihlet prior smoothing is described in Zhai and Massung [35]:

$$P(Q,D) = \sum_{w \in Q,D} c(w,Q)\, log\left(1 + \frac{c(w,D)}{\mu \cdot P(w|C)}\right) + |Q| \log \frac{\mu}{\mu + |D|} \tag{4}$$

where $c(w, D)$ is the word count in given document $D$, $c(w, Q)$ is the word count in given query $Q$, $|D|$ and $|Q|$ are the length of document and query and $P(w|C)$ is the probability of the word in the collection that is used to normalized the model.

The biggest challenge for the language model when it works with microblog retrieval tasks is the short document length. The majority of features do not occur more than once. Thus, microblog document like tweet does not have enough statically information.

## 3.2 Pseudo-Relevance Feedback

There are different pseudo-relevance feedback techniques proposed in the literature. All of them seek to enhance the retrieval process and to avoid the vocabulary mismatching problem. One of the robust models is RM3 [13, 1], in which the basic idea is to estimate the relevance feedback using relevance models such as query likelihood, BM25. Then, after the relevance feedback has been estimated, it interpolates with the original query. Lv and Zhai proved that RM3 was the most effective and robust among a number of state-of-art query expansion models [20].

$$P(w|R) = \sum_{D \in \mathcal{D}} P(D)P(w|D) \prod_{i=0}^{n} P(q_i|D) \tag{5}$$

$$P'(w|R) = \gamma\, P(w|R) + (1-\gamma)P(w|Q) \tag{6}$$

Query model $P(w|Q)$ computes using different ways and the simplest way to query feature.

# 4. PROPOSED MODEL

A big challenge that faces microblog retrieval is the vocabulary mismatching problem. This issue happens when the users' interest is not enough to represent the relevant documents. Different techniques can be used to solve this problem, including query expansion. Therefore, selecting discriminative expansions features will improve retrieving relevant documents to meet user interest needs.

Our proposed model takes advantages of the two-stage pseudo-relevance feedback technique to improves the query expansion. In order to overcome the limitation of pseudo-relevance feedback, we propose a pseudo-relevance feedback model employing latent Dirichlet allocation (LDA) [4].

In the first step, we ranked each document $D$ in the collection $C$ using the query likelihood model. The query likelihood $P(Q|D)$ is estimated using Dirichlet smoothing, as shown in equation 4. Intuitively, we make the same assumption for the pseudo-relevance feedback technique, that is, the top $k$ ranked pseudo-documents are relevant. Next, we compute the relevance model probabilities $P(w|Q)$ for the given query $Q = \{q_1, q_2, ..., q_n\}$ and represent the pseudo-relevance feedback $RS$ using RM1 [20], as shown in equation 7.

$$P_{lex}(w|Q) \propto \sum_{D \in RS} P(w|D)P(D) \prod_{i=0}^{n} P(q_i|D) \tag{7}$$

Importantly, we discovered latent topics in pseudo relevance feedback $RS$ using LDA. The result from LDA consists of a set of latent topics, each of which is represented by a multinomial distribution over words, $\Phi_j = (\varphi_{i,1}, \varphi_{i,2}, ..., \varphi_{i,n})$ for topic $j$ where $\varphi_{j,i}$ is the probability of a word $w_i$, and each document in the collection is represented by a multinomial distribution over topics, $\Theta_i = (\vartheta_{i,1}, \vartheta_{i,2}, ..., \vartheta_{i,v})$. $\vartheta_{i,v}$ indicates the proportion of topic $j$ in document $d_i$. $\Phi_j$ is called topic representation for topic $j$ in collection $D$ and $\Theta_i$ is called topic distribution in document $d_i$. In this paper, the topic representation is used as the word distribution $P(w|T)$ for each topic $T$.

Then, after the relevance feedback $P_{lex}(w|Q)$ is estimated as in equation 7, we compute the new relevance feedback model $P_{exp}(w|Q)$ for the combination of relevance model $P_{lex}(w|Q)$ for the pseudo-relevance feedback and the term probability the topic $P(w|T)$ using liner interpolation, as follow:

$$P_{exp}(w|Q) = (1-\lambda)P_{lex}(w|Q) + \lambda P(w|T) \tag{8}$$

where $\lambda \in [0, 1]$ is the something parameter. This step is done to improve the performance of the relevance model estimation.

The final step of our proposed model is a linear computation for the new expansion query words between the original query model $P(w|Q)$ and the relevance feedback model $P_{exp}(w|Q)$.

$$P'(w|Q) = \gamma\, P_{exp}(w|RS) + (1-\gamma)P(w|Q) \tag{9}$$

where $\gamma \in [0, 1]$ is a parameter to balance the using of pseudo-relevance feedback.

We compute the simple form for the original query as follows:

$$P(w|Q) = \frac{c(w, q)}{|q|} \tag{10}$$

Finally, we re-retrieve the results using the new query from the whole collection with respect to the document timestamps.

# 5. EXPERIMENT

In this section, we will demonstrate the experiment environment and describe the performance results and the discussions. Section 5.1 describes our experiment datasets. In section 5.2, we present the preprocessing process. Then, section 5.3 shows the evaluation metrics and parameter settings. Section 5.4 reports the baseline models. Finally, we assess the performance of our model and discussion.

## 5.1 Dataset test collection

The experiments were conducted using the TREC 2011 and 2012 datasets, called Tweets2011 [23, 28]. The size of the dataset is about 16 million tweets over a period of 2 weeks (January 23, 2011 to February 8, 2011), which included significant events including the US Superbowl and the Egyptian revolution events. Different kinds of tweets exist in the dataset, including retweets and and replies, which results in considerable noise. In addition, the tweets were published in various languages. Table 1 shows some statistics about the datasets. Each day in the corpus is split

**Table 1: The Tweets2011 dataset statics.**

| # Tweets | # English | # Retweet | # URL |
|---|---|---|---|
| 11477601 | 2113208 ≈ 18.412 % | 384654 ≈ 3.351 % | 2163585 ≈ 18.854 % |


```
{"text":
"Was at #TheTalk today! Ozzy will be a guest on the show tomorrow! Wish I go be there then!",
"id_str":"34765488198397952",
"id":3476548819839795 2,
"created_at":"Tue Feb 8 24:08:32 +0000 2011",
"retweeted":false,
"retweet_count":0,
"favorited":false,
"user":
{"id_str":"19945462",
"id":19945462,
"screen_name":"HollywoodJunket",
"name":"HOLLYWOOD JUNKET"},
"requested_id":34765488198397952}
```


**Figure 1: An example of Tweet structure.**

into files called blocks, each of which contains approximately 10,000 tweets compressed using gzip. Each tweet is in JSON format, as shown in Figure 1.

The TREC dataset provides 49 topics and relevant judgement for the TREC 2011 microblog and 59 topics for the TREC 2012. Each topic includes the topic number, title, and topic time. The relevant judgement is a multi-point scale that ranks tweets as highly relevant, relevant, and not relevant.

## 5.2 Preprocessing

Preprocessing documents is a critical phase to enhance the retrieval model and improve retrieval performance [3]. We treated the tweets and text queries based on the following steps:

- Tweet filtering: non-English tweets make data noisy, so we discarded these tweets using a language detector called *ldig* [2], in cases where the Tweets2011 dataset did not have a 'lang' attribute. Also, we filtered non-ASCII characters.
- Retweet treatment: we normalized the tweets that start with "RT @".
- Stop word removal and stemming: we applied stop word removal and then stem word using the Porter algorithm [26, 31].

## 5.3 Experiment Metrics and Settings

The next measures applied were some broadly acknowledged and settled assessment measurements. The evaluation metrics which were the precision at 10 and 30 (P@10, P@30 receptively), the Mean Average Precision (MAP), and the Normalized Discounted Cumulative Gain (NDCG) [12]. These measure are extensively used in information retrieval and are the official microblog metric at the TREC conference [23, 28, 19]. Every measurement emphasizes an alternate part of the framework execution.

Precision $p$ takes all retrieved documents into account. Mean Average Precision (MAP) is determined by measuring the precision of each relevant document then averaging the precision over all the subjects. It consolidates precision, relevance ranking, and general review together to measure the nature of the retrieval engines. Moreover, we tested the statistical difference of our evaluation using a two-tailed paired $t$-test with the $p$-value in the test is lower than 0.05.

For the topic model, we used the Java Machine Learning for Language Toolkit (MALLET)[3] Java library in our experimental environment system. We set the LDA parameter as follows, $\alpha = 0.01$ and $\beta = 0.05$.

We completed all experiments taking into account two standard topic sets as a part of the TREC 2011 and 2012 datasets. The primary topics set, which is called *allrel*, reflects minimally and highly relevant document as relevant in more that 49 topics. The second topic set includes highly relevant documents over 33 from 59 topics and is called *highrel*.

## 5.4 Baseline

Our proposed model compares with the query likelihood model that is described in section 3 and Okapi BM25 [27]. We set Dirichlet smoothing as $\mu = 1000$. BM25 is a probabilistic state-of-the-art retrieval model that can compute the similarity between document $d$ and query $q$ containing words $w$.

$$BM25(q,d) = \sum_{w \in q \cap d} \log\left(\frac{N - df(d) + 0.5}{df(d) + 0.5}\right) \times \frac{(k_1+1)c(w,d)}{k_1((1-b) + b\frac{dl}{avdl}) + c(w,d)} \quad (11)$$

where $c(w,d)$ is the document term frequency, $dl$ and $avdl$ are the document length and the average document length, respectively. The $k1$ and $b$ parameters are the experimental parameters (in this paper we set $k1 = 0.9$ and $b = 0.4$, respectively).

In our experiment settings, we set $\lambda = 0.5$ for the relevance model as well as $\gamma = 0.5$ in equation 10. For the relevance model settings, we set $k = 50$ pseudo-relevant documents and selected $n = 20$ feedback terms. The parameter setting set was based on best experimental practices.

## 5.5 Results

Query likelihood and BM25 provided a baseline for our model performance results. Table 2 reports the overall experimental results for the TREC 2011 dataset based on precision at a rank 10 and 30, respectively, Mean Average Precision (MAP), and Normalized Document Gain Cumulative (NDCG), where *%chg* presents the percentage change of TBPRF over QL. With the same metrics in Table 2, Table 3 shows the performance for the TREC 2012 dataset.

It can be clearly seen experimentally that our model (TBPRF) outperforms QL and BM25 for TREC 2011 and 2012. Both Table 2 and Table 3 show that TBPRF achieves excellent performance with a 7.21 per cent improvement on average for TREC 2011 (with a maximum of 13.46 per cent and a minimum of -0.84 per cent). For TREC 2012, TBPRF

[2]http://github.com/shuyo/ldig

[3]http://mallet.cs.umass.edu/

**Table 2: Results comparison of the proposed methods and baselines for allrel documents of TREC 2011.**

| Method | P@10 | P@30 | MAP | NDCG |
|---|---|---|---|---|
| QL | 0.4184 | 0.332 | 0.2705 | 0.5058 |
| BM25 | 0.4000 | 0.3156 | 0.2503 | 0.4896 |
| TBPRF | 0.4149 | 0.3605 | 0.3069 | 0.5444 |
| *%chg* | **-0.84%** | **8.58%** | **13.46%** | **7.63%** |

**Table 3: Results comparison of the proposed methods and baselines for allrel documents of TREC 2012.**

| Method | P@10 | P@30 | MAP | NDCG |
|---|---|---|---|---|
| QL | 0.3729 | 0.3056 | 0.1788 | 0.4324 |
| BM25 | 0.3542 | 0.3073 | 0.1659 | 0.4162 |
| TBPRF | 0.4000 | 0.3232 | 0.2002 | 0.4623 |
| *%chg* | **7.27%** | **5.17%** | **11.97%** | **6.91%** |

reaches outstanding performance, with 7.83 per cent improvement on average for TREC 2012 (with a maximum of 11.97 per cent and a minimum of 5.17 per cent).

The experimental results for highly relevant documents are shown in Table 4. It is obvious that pseudo-relevance feedback using TBPRF has improved the retrieval process when retrieving a highly relevant documents subset for both the TREC 2011 and 2012 datasets. Table 6 shows an example of query analysis for topic number 'MB86' with the topic title 'Michelle Obama's obesity campaign'. It clearly presents the difference between the relevance using the relevance model and our TBPRF relevance model estimation for nominated terms for expansion.

Finally, the statistical significance tests are illustrated in Table 5 to compare the proposed model with the baseline performance models on the TREC 2011 and 2012 microblog datasets. The results show that the proposed TBPRF model is statistically significant, with all $p$ values are less than 0.05.

# 6. CONCLUSIONS

In this paper, we proposed a practical topic-based pseudo-relevance feedback model including query expansion using lexical and topic evidence. We evaluated our model using the TREC 2011 and 2012 microblog datasets. The experimental results illustrate that the proposed model obtained better results than the baseline models. The proposed model

**Table 4: Results comparison of the proposed method and baselines for highrel documents of the TREC 2011 and 2012 datasets.**

| | 2011 | | 2012 | |
|---|---|---|---|---|
| Method | P@30 | MAP | P@30 | MAP |
| QL | 0.0755 | 0.1227 | 0.1644 | 0.1333 |
| BM25 | 0.0605 | 0.1124 | 0.1689 | 0.1229 |
| TBPRF | 0.0782 | 0.1283 | 0.174 | 0.1544 |
| *%chg* | **3.58%** | **4.56%** | **3.02%** | **15.83%** |

**Table 5: p-Values for TBPRF.**

| Method | P@30 | MAP | NDCG |
|---|---|---|---|
| QL | 0.015777 | 0.000162 | 0.000066 |
| BM25 | 0.004086 | 0.000007 | 0.000026 |

**Table 6: Example of expanded terms for a topic numbered MB86: 'Michelle Obama's obesity campaign'.**

| $P_{exp}(w\|Q)$ | | $P'(w\|Q)$ | |
|---|---|---|---|
| term | weight | term | weight |
| obama | 0.2487 | obama | 0.2527 |
| michelle | 0.2284 | michel | 0.2382 |
| campaign | 0.1117 | campaign | 0.1863 |
| lady | 0.0609 | obes | 0.1472 |
| obesity | 0.0457 | ladi | 0.0331 |
| oprah | 0.0355 | atlanta | 0.0157 |
| atlanta | 0.0254 | childhood | 0.0122 |
| obama's | 0.0254 | move | 0.0104 |
| stop | 0.0254 | oprah | 0.0092 |
| childhood | 0.0203 | stop | 0.0091 |
| food | 0.0203 | militari | 0.0083 |
| coming | 0.0203 | role | 0.0080 |
| weight | 0.0203 | plai | 0.0079 |
| role | 0.0203 | utm | 0.0076 |

offered a significant improvement of 7.59 per cent on average improvement over query likelihood using Dirichlet smoothing as well as 12.70 per cent on average improvement over BM25. In this paper, we demonstrated that the proposed model was tested, and the results show that the experimental results were statistically significant. This article presents vital support for the proposed model of stability and robustness. The main challenges that we would like to investigate in future work are how to incorporate temporal evidence with the proposed model.

# 7. REFERENCES


[1] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, and C. Wade. Umass at trec 2004: Novelty and hard. 2004.

[2] C. C. Aggarwal and C. Zhai. *Mining text data*. Springer Science & Business Media, 2012.

[3] K. Albishre, M. Albathan, and Y. Li. Effective 20 newsgroups dataset cleaning. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 3, pages 98–101. IEEE, 2015.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

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

[6] J. Choi and W. B. Croft. Temporal models for microblogs. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 2491–2494. ACM, 2012.

[7] M. Efron. Hashtag retrieval in a microblogging

environment. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788. ACM, 2010.

[8] M. Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.

[9] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 495–504. ACM, 2011.

[10] T. El-Ganainy, W. Magdy, and A. Rafea. Hyperlink-extended pseudo relevance feedback for improved microblog retrieval. In *Proceedings of the first international workshop on Social media retrieval and analysis*, pages 7–12. ACM, 2014.

[11] Y. Gao, Y. Xu, and Y. Li. Topical pattern based document modelling and relevance ranking. In *International Conference on Web Information Systems Engineering*, pages 186–201. Springer, 2014.

[12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002.

[13] V. Lavrenko and W. B. Croft. Relevance based language models. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127. ACM, 2001.

[14] Y. Li, A. Algarni, M. Albathan, Y. Shen, and M. A. Bijaksana. Relevance feature discovery for text mining. *IEEE Transactions on Knowledge and Data Engineering*, 27(6):1656–1669, 2015.

[15] Y. Li, A. Algarni, and N. Zhong. Mining positive and negative patterns for relevance feature discovery. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 753–762. ACM, 2010.

[16] Y. Li and N. Zhong. Mining ontology for automatically acquiring web user information needs. *IEEE transactions on Knowledge and Data Engineering*, 18(4):554–568, 2006.

[17] Y. Li, X. Zhou, P. Bruza, Y. Xu, and R. Y. Lau. A two-stage decision model for information filtering. *Decision Support Systems*, 52(3):706–716, 2012.

[18] F. Liang, R. Qiang, and J. Yang. Exploiting real-time information retrieval in the microblogosphere. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, pages 267–276. ACM, 2012.

[19] J. Lin, M. Efron, Y. Wang, and G. Sherman. Overview of the trec-2014 microblog track. Technical report, DTIC Document, 2014.

[20] Y. Lv and C. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1895–1898. ACM, 2009.

[21] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[22] T. Miyanishi, K. Seki, and K. Uehara. Improving pseudo-relevance feedback via tweet selection. In *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, pages 439–448. ACM, 2013.

[23] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the trec-2011 microblog track. In *Proceeddings of the 20th Text REtrieval Conference (TREC 2011)*, volume 32, 2011.

[24] L. Pipanmaekaporn and Y. Li. Discovering relevant features for effective query formulation. In *Information Retrieval Facility Conference*, pages 137–151. Springer, 2012.

[25] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 275–281. ACM, 1998.

[26] M. F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.

[27] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, M. Gatford, et al. Okapi at trec-3. *NIST SPECIAL PUBLICATION SP*, 109:109, 1995.

[28] I. Soboroff, I. Ounis, C. Macdonald, and J. Lin. Overview of the trec-2012 microblog track. In *TREC*, volume 2012, page 20, 2012.

[29] J. Teevan, D. Ramage, and M. R. 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.

[30] Z. Wang and M. Zhang. Feedback model for microblog retrieval. In *Database Systems for Advanced Applications*, pages 529–544. Springer, 2015.

[31] P. Willett. The porter stemming algorithm: then and now. *Program*, 40(3):219–223, 2006.

[32] 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*, pages 4–11. ACM, 1996.

[33] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 334–342. ACM, 2001.

[34] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214, 2004.

[35] C. Zhai and S. Massung. *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan; Claypool, New York, NY, USA, 2016.