Query Expansion Based on a Feedback Concept Model for Microblog Retrieval

Yashen Wang
Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China
yswang@bit.edu.cn

Heyan Huang*
Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China
hhy63@bit.edu.cn

Chong Feng
Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China
fengchong@bit.edu.cn

ABSTRACT

We tackle the problem of improving microblog retrieval algorithms by proposing a Feedback Concept Model for query expansion. In particular, we expand the query using knowledge information derived from Probase so that the expanded one could better reflect users' search intent, which allows for microblog retrieval at a concept-level, rather than term-level. In the proposed feedback concept model: (i) we mine the concept information implicit in short-texts based on the external knowledge bases; (ii) with the relevant concepts associated with short-texts, a mixture model is generated to estimate a concept language model; (iii) finally, we utilize the concept language model for query expansion. Moreover, we incorporate temporal prior into the proposed query expansion method to satisfy real-time information need. Finally, we test the generalization power of the feedback concept model on the TREC Microblog corpora. The experimental results demonstrate that the proposed model outperforms the previous methods for microblog retrieval significantly.

Keywords

Microblog Retrieval; Pseudo-Relevance Feedback; Short-Text Conceptualization; Query Expansion

1. INTRODUCTION

Collections of microblog documents pose difficult challenges and offer unique opportunities to retrieval systems at the same time. Microblog retrieval systems need to overcome severe vocabulary mismatch problem (i.e., how to retrieve very short documents, which might be conceptually relevant, but do not explicitly contain some or all of the query terms), while having to deal only with scarce relevance signals that could be derived from the text of the tweets alone. Moreover, for short-text, neither parsing nor topic modeling works well because there are simply not enough signals in the input [34, 35]. To solve the problem, we must: (i) derive more signals from the input by combining it with external knowledge bases, and (ii) devise a framework that enables the signals to fully interplay, so that we have more power to disambiguate and understand a short-text. In the rest of the paper, we address the two challenges above: (i) for the first aspect, we mine the concept information implicit short-text; and (ii) for the second aspect, we integrate the correlative concept into pseudo-relevance feedback framework to improve the semantic relevance of microblog retrieval.

Query Expansion (QE) methods based on Pseudo-Relevance Feedback (PRF) [17, 20, 37] are widely used in microblog search to mitigate the problems mentioned above. However, these methods rely much on the assumption that the top ranked documents in the initial search are relevant and contain good words for query expansion. Nevertheless, in real world, this assumption does not always hold in microblogosphere [5, 24], considering the example that the query contains proper nouns difficult to understand. What's more, even if the top ranked documents are highly relevant to the topic, it is still very likely that they contain numerous topic-unrelated words due to the informality of the tweet content [24].

To overcome the limitations of existing methods, we utilize Probase [36] as the knowledge source to infer more concept-related context information for each query and each tweet. In this paper, we propose a feedback concept model for query expansion. In this model, the ultimate aim is to capture optimal concept language model to expand the original query language model. We firstly preprocess the coming query and the entire tweet collection via short-text conceptualization [31, 35], and obtain the relevant concepts respect to query and tweets in the collection. With the identified

*The Corresponding Author.

WWW 2017, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4913-0/17/04.
http://dx.doi.org/10.1145/3038912.3052710

concepts, we obtain latest concept-relevant feedback tweets, which is used to estimate concept language model upon a maximum likelihood estimation strategy. More specifically, during the estimation, we assume that the concept-relevant tweets are generated by a mixture model, which consists of the concept language model, the collection language model and the contextual-concept model. In summary, the general idea of the paper is that, properly identifying feedback concepts in queries and tweets could potentially allow for retrieval at a concept-level, rather than term-level. In order to satisfy real-time information need in microblog retrieval, we follow the work of [15] and incorporate a temporal prior distribution regarding to the recency of documents into the proposed feedback concept model. We assign each top ranked pseudo-relevance document with a time prior so that the words appearing more in recent documents are associated with higher probability.

The main contributions of this paper include: (i) we leverage a novel approach for generating knowledge information from the probabilistic knowledge base (e.g., Probase) to expand the original query in concept-level, which leads to better understanding of usersą́r information need; (ii) To make signals fully interplay, the feedback concept model is estimated with the latest concept-relevant tweets by a mixture model based upon a maximum likelihood criterion; (iii) the temporal evidence is incorporated into our QE method to trade off between relevance and recency; (iv) We examine the proposed framework on two real-world tweet collections published by TREC by comparing it to previous representative methods, and the experimental results show that proposed framework achieves higher accuracy and efficiency in microblog retrieval task. To the best of our knowledge, although there exists researches incorporating concept information into traditional search [7, 23, 25], the proposed research is the first to incorporate short-text conceptualization into pseudo-relevance feedback framework for microblog retrieval.

The outline of the paper is as follows. Section 2 surveys the related researches. Section 3 formally describes the proposed query expansion framework based on feedback concept model. Corresponding experimental results are shown in Section 4. Finally, Section 5 concludes the paper.

2. RELATED WORK

2.1 Query Expansion based on PRF

It is widely used to leverage Pseudo-Relevance Feedback (PRF) for Query Expansion (QE) in research of microblog search [13, 22, 9, 37], which assumes that most of the frequent terms in the pseudo-relevance documents are useful. [24] proposed a query expansion method based on two-stage relevance feedback that models search interests by manual tweet selection and integration of lexical evidence into its relevance model. Observing that users tend to use entities in their queries to express the certain information need, [20, 12] proposed a two-stage feedback entity model based on a mixture strategy, which consists of the entity model, the domain-specific language model and the collection language model. [28] studied the utility of syntactic patterns for microblog retrieval, by proposing an efficient way to encode tweets into linguistic structures and using kernels for automatic feature engineering. However, these methods somehow are easily affected by the errors propagating from traditional NLP pipelines (e.g., entity recognition or parsing) in preprocess and the entire inference is nearly based on the word-level analysis. Moreover, [12] only generates discrete entity model for each entity in query, which is powerless to release the global semantic representation for the entire query. Just like the reported researches, we also attempt to improve the effectiveness of pseudo-relevance feedback based query expansion. However we seek help from another semantic signal from short-text, i.e., concept, which have been demonstrated effective in knowledge representation [12, 24, 34].

2.2 Query Expansion based on Concept

Although many researcher aim at using semantics to enhance microblog search [1, 38], most of them failed to enable the signals to fully interplay and rare researches pay attention to concept-based microblog search. Many recent studies in IR consider query terms relevant with each other [2]. Other models have been proposed to capture the exact dependencies between more than two terms. As mentioned in [23], most previous term dependence models have failed to show robust, significant improvements over baseline bag-of-words models, with only a few exceptions such as the dependence language models and the Markov random fields (MRF) model [23]. As a concept-based query expansion technique, the Latent Concept Expansion (LCE) derived from MRF shows significant improvements over relevance models on MAP across different data sets, but only small, insignificant improvements on precision at 5, 10, and 20. However, one problem of LCE is that it adds the words of new concepts to the given query based on word statistics regardless of the semantics and syntax of the expanded concepts. Unfortunately, microblog texts usually donąŕt contain enough signals for statistical inference. Moreover, for concept-based IR, some empirical studies have shown that retrieval models using only concepts have inconsistent results, since not all documents and queries could be effectively represented using concepts. Different methods are investigated to combine relevance scores from both word and concept retrievals in [6, 18] . Hence, external knowledge base is introduced as priors in our study, to enhance the signals which contribute to inference.

2.3 Query Expansion based on Temporary

Previous works have demonstrated that temporal prior has a strong effect on information retrieval [10, 8, 20]. To explore the relationship between time and relevance, [15] proposed a time-based language model by incorporating time into both query-likelihood models and relevance models. [11] introduced a temporal factor into language model smoothing and query expansion using pseudo-relevance feedback, and showed their effectiveness for recency queries. For the temporal re-ranking in information retrieval , [17] suggested several methods to evaluate the temporal aspects of documents. to deal with real-time filtering in the Microblogging platform, [3] modified the user profile to balance between the importance of the short-term interests for a given topic. [24] utilized a two-stage PRF based on similarity of temporal profiles of the query and top retrieved documents. In our study, the temporal prior is leveraged for the expansion method to enhance the importance of the words those are often used to describe the concept recently.

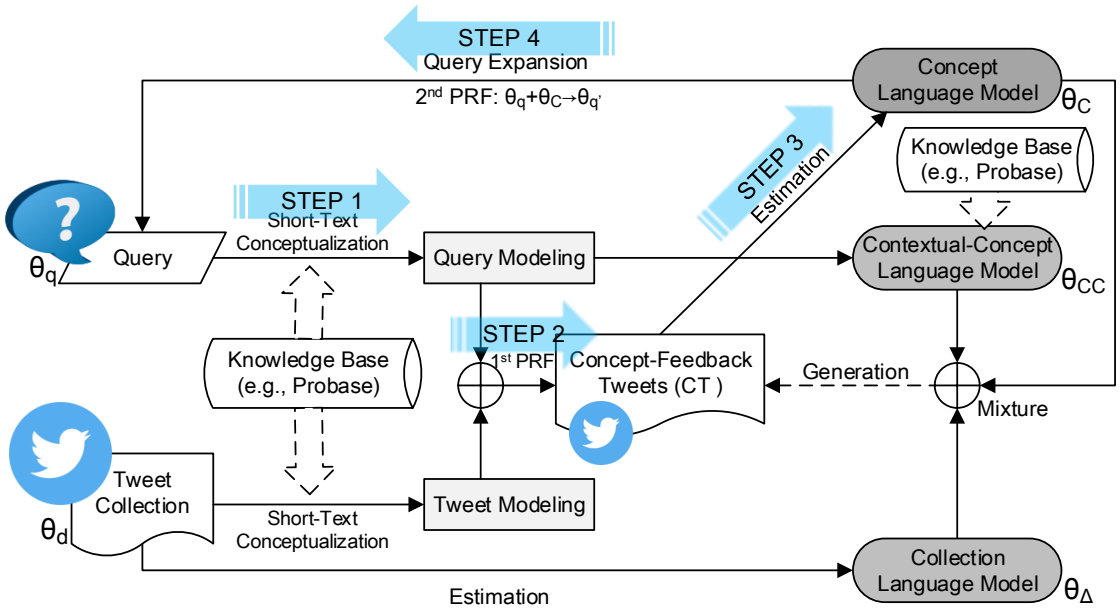

Figure 1: The scheme of the proposed Feedback Concept Model for Query Expansion (QE) in microblog retrieval task.

2.4 Short-Text Conceptualization

Short-text understanding is a challenging task. Since short-texts are usually lack of context, mapping short texts to concepts can help better make sense of text data, extend the texts with categorical or topical information, and facilitate many applications. For example, it has been verified very useful for short-text categorization [33], text classification [30, 34], search relevance measurement [29], and search logo mining [13]. An increasing number of studies have paid much attention to the research attempted to acquire knowledge from external sources to obtain additional query terms. During query conceptualization, [31] groups instances by their conceptual similarity, and uses simple Bayesian analysis to conceptualize each group. [16] is a Wikipedia-based query expansion algorithm, which utilizes Wikipedia as external corpus to understand query for improving ad-hoc retrieval performance. Observing that available context (i.e., the verbs related to the instances, the adjectives and attributes of the instances) do provide valuable clues to understand instances, [35] utilize a knowledge base that maps instances to their concepts, and build a knowledge base that maps non-instance words, including verbs and adjectives, to concepts. However, queries usually do not observe the syntax of a written language, nor do they contain enough signals for statistical inference.

3. FEEDBACK CONCEPT MODEL FOR QUERY EXPANSION

In this section, we describe the details of the proposed query expansion framework based on the feedback concept model. More specially, this feedback concept model is estimated by a mixture model with the latest concept-relevant tweets. As discussed in [12, 21], the mixture model [37] is reported with a relatively better retrieval performance among the state-of-the-art PRF-based query expansion methods.

3.1 The Framework of Feedback Concept Model

Language modeling [14] presumes that if it is highly probable that the document generates the query, then the content of the document is more likely to be relevant to the information need underlying the user's request. Therefore, it usually utilizes the probability that a language model that the document would generate the query as the ranking function [37]. Hence, we assume that a query q is generated by the query language model θ_q and a document (e.g., tweet) d is generated by the document language model θ_d. After estimating θ_d and θ_q according to [14], the relevance score of d with respect to q could be computed by the following KL-divergence function:

$$S(q|d) = -KL(\theta_q \| \theta_d) \propto \sum_{w \in V} P(w|\theta_q) * \log P(w|\theta_d) \quad (1)$$

Wherein V is the vocabulary dictionary, and w is the word in V. In this study, we mainly focus on how to estimate θ_q by leveraging concept information. To this end, we propose a query expansion framework based on feedback concept model. The conceptual scheme of this framework is illustrated in Figure 1. Moreover, the proposed framework aims at solving the problem with minimum supervision. Although manually labeled data are insufficient, prior knowledge on the structured semantics or lexical information of language are massively available because many knowledge bases have emerged in recent years, including Probase, Freebase, DBpedia and so on. Here, we incorporate Probase [36] into the proposed framework.

The proposed query expansion framework is based on two-stage Pseudo-Relevance Feedback (PRF) strategy, and it consists of two parts, an offline part and an online part:

Offline Part: In offline part, we construct semantic network representing relationships between instance terms and corresponding concepts by utilizing knowledge base Probase

(described in Section 3.2), and preprocess the indexed tweet collection via short-text conceptualization algorithm [35] to capture the relevant concepts respect to tweets in the collection.

Online Part: In the online part, we firstly preprocess the coming query via short-text conceptualization algorithm [35], and obtain its relevant concepts. With the identified concepts, we obtain latest concept-relevant feedback tweets (CT) while an original concept-based retrieval (described in Section 3.3), which is the 1st step in PRF. Then, the obtained feedback tweets is used to estimate Concept Language Model: Specifically, we assume that the concept-relevant feedback tweets are generated by a mixture model, which consists of three language models: (i) the concept language model, (ii) the contextual-concept language model (described in section 3.5) and (iii) the collection language model. Afterwards, the concept language model is estimated with the concept-relevant feedback tweets based upon a maximum likelihood estimation (MLE) strategy (described in section 3.4). Finally, the estimated concept language model is used to expand the original query (2nd step in PRF). Besides, in order to further satisfy users' real-time information need, we incorporate temporal evidences into the proposed expansion method (described in section 3.6).

3.2 Why We Choose Probase

As discussed in [31, 35], it is essential to utilize lexical knowledge bases to understand query, rather than encyclopedic knowledge bases (e.g., Wikipedia, DBpedia, Freebase, Yago etc.). That is, the knowledge of the language should be used. Because such encyclopedic knowledge bases contains facts such as Barack Obama's birthday and birthplace, which are useful to answer questions. Inversely, lexical knowledge bases could definitely indicate that birthplace and birthday are properties of a person.

We use Probase here as the knowledge base to demonstrate the generate feedback terms. Probase is a lexical knowledge base, which is widely used in research about short-text understanding [31, 32] and text representation [34]. However, our techniques can be applied to other knowledge bases such as Yago. Probase uses an automatic and iterative procedure to extract concept knowledge from 1.68 billion Web pages. It contains 2.36 millions of open domain terms. Each term is a concept, an instance, or both . Meanwhile, it provides around 14 millions relationships with two kinds of important knowledge related to concepts: concept-attribute co-occurrence ($isAttrbuteOf$) and concept-instance co-occurrence (isA) .

The semantic network is constructed, which will make contribution to short-text conceptualization. This kind of semantic network consists of terms, concepts, and their relationships. And a sub-graph of this semantic network is shown in Figure 2. In particular, each term could map to several concepts. For a short-text s, the words in evoke a subgraph in the semantic network. For any word w in s, we want to maximize the likelihood of the concept c given short-text, which can be formulated as the following optimization problem: $\arg\max_c P(c|w,s)$. Figure 2 shows a subgraph of the semantic network taking query "microsoft unveils office for apple's ipad" as an example. As we can see, there are two types of vertices: one represents concept (shown as ellipse), and the other represents an instance or attribute (shown as rectangle). Among these vertices, there are t-

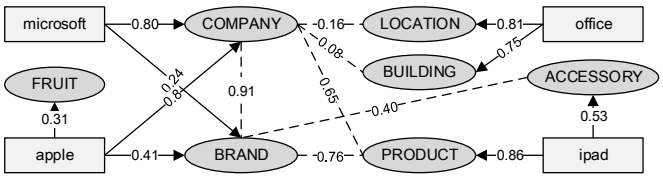

Figure 2: Subgraph of the semantic network.

wo kinds of edges denoting different relationships: isA relationship between instances (attributes) and concepts (i.e., term-concept, shown as directed links), *correlation* relationship between two concepts (i.e., concept-concept, shown as dashed lines). Wherein, correlation relationship measures how strongly two concepts are related (e.g., `Company` and `Brand` are strongly related). Intuitively, the semantic network is a heterogeneous network.

Since the edge could represents isA or *correlation* relationship, we discuss each case to quantify the weights of edges, separately. In case of isA, which represents an edge from a non-concept word w (instance or attribute) to a concept c, the weight of term-concept edge is denoted as $P(c \mid w)$, We compute it as follows: $P(c \mid w) = n(w,c)/\sum_{c_i} n(w,c_i)$. Wherein, $n(w,c)$ is the frequency observed as isA relationship (ie., w isA c) in a corpus. In case of *correlation*, we denote the relatedness between two concepts (c_1 and c_2) as $P(c_2 \mid c_1)$. And following the strategy discussed in [31], the probability is derived by aggregating the co-occurrences between all instances of the two concepts, as follows: $P(c_2 \mid c_1) = \sum_{t_i \in c_1, t_j \in c_2} n(t_i,t_j)/\sum_c \sum_{t_i,t_j \in c} n(t_i,t_j)$.

3.3 1st Step in Feedback Concept Model: Concept Feedback Tweet Construction

The proposed query expansion algorithm with Feedback Concept Model is based on pseudo-relevance feedback (PRF) framework. Generally, the PRF framework consist of two stage: (i) in the first stage, the Pseudo-Relevance Documents (PRD) are obtained; and (ii) in the second stage, original query is expanded based on some query expansion strategy.

As the 1st stage in the pseudo-relevance feedback strategy, we should construct the Pseudo-Relevance Documents, which is called the Concept-Relevant Feedback Tweets (denoted as CT) in this paper. To address this problem, we simultaneously investigate concept-relevance and temporal information: Firstly, we identifies relevant concepts of the query and the indexed tweets in collection. To achieve this goal, we introduce an algorithm for short-text conceptualization [35]. Then we could obtain the feedback tweets (CT) by collecting all the tweets which are also relevant with these concepts containing in query. Considering that temporal-anchored ad-hoc search strategy is essential in microblog retrieval [26], we maintain the temporal characters of the feedback concept model by selecting the most recent M tweets before query issue time T_q. Overall, this method is a proper balance of relevance and recency.

The employed algorithm for short-text conceptualization [35] consists of three components: (i) Firstly, we segment the query into a set of candidate terms; (ii) Secondly, we create a graph out of the terms and their relationships; (iii) Finally, a random walk based iterative process is used to find the most likely mapping from terms to concepts. In short-text

segmentation, we ignore terms such as prepositions (e.g., example in Figure 2 ignores preposition for). Although these terms provide useful linguistic clues to understand term dependency, the proposed model do not consider about them. Because we focus on annotating terms in a query with their concepts, while there is no proper concepts with respect to prepositions. Moreover, each tweet was stemmed using the Porter algorithm and stopwords were removed using the In-Query stopword-list. We segment a query into a set of words. Then we take advantage of Probase as lexicon to identify all occurrences of terms in the query. In particular, we only consider maximum terms, which are not completely contained by other terms.

3.4 2nd Step in Feedback Concept Model: Estimation of Concept Language Model

Overall, we would like to estimate a concept language model (θ_C) for the identified concepts in query q which is used to expand the original query language model (θ_q), as shown in Figure 1. Specially, this concept language model is a probabilistic distribution of words $\{P(w|\theta_C)\}_{w\in V}$ and represents the common semantic categories of a given concept. Following [36], we define a 'concept' as a set or class of 'entities' or 'things' within a domain, such that words belonging to similar classes get similar representations. For instance, $Jeep$ and $Honda$ could be represented by concept `Car`. Clearly, we have $\sum_{w\in V} P(w|\theta_C) = 1$. Note that we introduce Probase as knowledge base to estimate this concept language model. and we will describe details about how to estimate this model later.

With the concept language model θ_C we could expand the original query language model (θ_q), i.e., $\theta_q + \theta_C \rightarrow \theta_{q'}$, which is our ultimate goal. To address this query expansion, we take advantage of the linear interpolation for combing the original query language model (θ_q) and the concept language model (θ_C) as follows:

$$P(w|\theta_{q'}) = (1-\alpha) * P(w|\theta_q) + \alpha * P(w|\theta_C) \qquad (2)$$

Wherein $\alpha \in [0,1]$ is a parameter to control the weight of the concept language model, determining the influence of feedback concept information to original query: when $\alpha = 0$, we only use the original query model (i.e., θ_q), and while $\alpha = 1$, our framework emphasizes the feedback concept model most.

Therefore, the next intractable challenge we are faced with is: how to accurately estimate the concept language model (θ_C) for original query, based on concept-relevant feedback tweets (denoted as CT). Intuitively, it is a natural way to solve this problem by assuming that the feedback tweets (CT) are generated by a probabilistic language model just as $P(CT|\Lambda) = \Pi_{d_i}\Pi_{w\in V} P(w|\Lambda)^{n(w,d_i)}$, wherein Λ denotes the entire parameter set. Actually, such assumption seems to be somehow idealistic. However, the content in these feedback tweets is apparently of high diversity and redundancy [12], as they generally may contain rich background noise and irrelevant concept information. On the other hand, since pseudo-relevance feedback strategy extremely relies on feedback documents [21], the quality of the feedback tweets should be guaranteed.

Therefore, we should filter the noise feedbacks and purify the feedback tweets in advance, to make use of feedback tweets more effectively. To address this problem, we propose a mixture model to estimate the concept language model (θ_C) based on a maximum likelihood estimation (MLE) criterion. And we assume that the observed feedback tweets (CT) is generated by this mixture model. Specifically, the proposed mixture model incorporates not only the concept language model (θ_C), but also the collection language model (θ_Δ) and the contextual-concept language model (θ_{CC}):

(i) Concept Language Model (θ_C): $P(w|\theta_C)_{w\in V}$, is the objective of our estimation and will be ultimately used to expand original query (θ_C) rather than [12] generates discrete entity model for each entity in query, θ_C denotes concept model for the whole query in our study.

(ii) Collection Language Model (θ_Δ): $P(w|\theta_\Delta)_{w\in V}$, is a probabilistic distribution of words, which is estimated based on the entire twitter collection, similar to [12]. It helps to filter global background noise, which is widely used as filtration method in microblog retrieval.

(iii) Contextual-Concept Language Model (θ_{CC}): $P(w|\theta_{CC})_{w\in V}$, models prior knowledge of specific concepts including in the given query. That is, it plays a role of leveraging concept-prior information from knowledge base. Assuming that there exist k concepts $(c_1, c_2, \ldots, c_k)$ in original query q after short-text conceptualization, we have $\theta_{CC} = \{\theta_{c_1}, \theta_{c_2}, \ldots, \theta_{c_k}\}$. Each of θ_{c_k} is also a probabilistic distribution of words, which is estimated based on the information of specific concept c in knowledge base. θ_{CC} is used to filter local background noise of special concepts in given query. Meanwhile, we have $\sum_{w\in V} P(w|\theta_{CC}) = 1$, and $\sum_{w\in V} P(w|\theta_\Delta) = 1$.

By utilizing this mixture model to estimate the concept language model (θ_C), the log-likelihood for the concept-feedback tweets (CT) is reformulated as follows.

$$\begin{aligned}\log P(CT|\Lambda) &= \textstyle\sum_{d_i}\sum_{w\in V} n(w,d_i) \\ * \log\{(1-\lambda_C) * [(1-\lambda_\Delta) * P(w|\theta_C) &+ \lambda_\Delta * P(w|\theta_\Delta)] \\ +\lambda_C * \textstyle\sum_j \mu_j * P(w|\theta_{c_j})\}& \end{aligned} \qquad (3)$$

wherein, d_i denotes the tweet in the feedback tweets (CT), k is the number of the corresponding concepts included in given query, and $n(w,d_i)$ is the frequency of word w occurs in tweet d_i. With the effects above, parameter set Λ is extended as follows: (i) the concept language model (θ_C); (ii) the collection language model θ_Δ; (iii) the contextual-concept language models $\theta_{CC} = \theta_{c_1}, \theta_{c_2}, \ldots, \theta_{c_k}$, and (iv) the corresponding weights for contextual-concept language models $\mu_1, \ldots, \mu_k$. Parameters λ_Δ and λ_C control the weight of global background noise and local background noise of specific concepts, respectively. Note that, θ_Δ could be estimated from the whole tweet collection easily, while the estimation of $\theta_{CC} = \theta_{c_1}, \theta_{c_2}, \ldots, \theta_{c_k}$ should introduce concept-priors from knowledge base (details in Section 3.5). Finally, Expectation Maximization (EM) algorithm could be applied to compute this maximum likelihood estimation.

3.5 Concept-Priors Incorporation for Contextual-Concept Language Model

Let's discuss how to construct the contextual-concept language model (θ_{CC}), which incorporate concept-priors from knowledge base (e.g., Probase) into the proposed feedback concept model for query expansion. Such prior knowledge contributes greatly to estimating the contextual-concept language model $\theta_{CC} = \theta_{c_1}, \theta_{c_2}, ..., \theta_{c_k}$, as shown in Figure 1. More specifically, we want to build a language model θ_{c_k},

i.e., $\{P(w|c_k)\}_{w \in V}$, for each pre-defined concept c based on information in Probase. For concept c_k we generate its concept-document (D_{c_k}) based on the information with respect to c_k in Probase: we group all the texts corresponding to c_k, including sub-concepts and instances (and their attributes). Then we generate the language model for each concept-document (D_{c_k}) as follows: $P(w|D_{c_k}) = n(w, D_{c_k}) / \sum_{w' \in V} n(w', D_{c_k})$. Wherein, $n(w, D_{c_k})$ denotes the frequency that word w occurs in the concept-document D_{c_k}. Additionally, a Dirichlet prior could be defined on each unigram language model following [12]. Assume that σ_{c_k} is the confidence parameter for the Dirichlet prior, and $\sigma_{c_k} = 0$ if we do not have prior knowledge for concept c_k. Ultimately, the contextual-concept language model (θ_{CC}) with concept-prior, could be formulated as follows:

$$P(w|\theta_{CC}) \propto \prod_{w \in V} \prod_{j} P(w|\theta_{c_j})^{\sigma_{c_j} * P(w|D_{c_j})} \quad (4)$$

3.6 Temporal-Prior Incorporation for Feedback Concept Model

Moreover, it has been reported that a good expansion term in microblog retrieval should satisfy the following criteria [20]: (i) The term should be semantically relevant with the concept from the original query (q); (ii) The term extracted from Probase should also be widely adopted in the Twitter corpus while talking about the concept; (iii) As the user's intent may change and events related to the given topic will develop over the time, the ranking function should favor the short-term words that are mostly used in recent tweets.

Obviously, with efforts above, the first criterion has been achieved. To meet the second criterion, we follow the work of [20] and re-score the candidate terms based on the top ranked M concept-relevant feedback tweets (CT), as follows:

$$Score(w) = \sum_{d_i \in CT} P(d_i) * P(w|d_i) * \prod_{q_j \in q} P(q_j|d_i) \quad (5)$$

Wherein $P(d_i)$ is the document prior which is usually assumed to be uniform, and $\prod_{q_j \in q} P(q_j|d_i)$ is the query likelihood given the document model. To meet the third criterion, we follow the work of [15] and incorporate the temporal evidence into the document prior ($P(d_i)$) above, as follows:

$$P(d_i|T_{d_i}) = r * e^{-r*(T_q - T_{d_i})} \quad (6)$$

Wherein r is the parameter that controls the temporal evidence, T_q is the query issue time and T_{d_i} is the tweet post time. Note that T_{d_i} is constantly less than T_q as we cannot use the future evidence.

4. EXPERIMENTS AND RESULTS

We show experiments on two microblog datasets to evaluate our algorithm in several aspects. Firstly, we investigate the influence of concept-based query expansion by comparing with other kinds of query expansion algorithms. Secondly, we conduct corresponding analysis on parameters of the proposed framework.

4.1 Datasets

Our dataset are the official tweet collections used in the TREC Microblog track, **Tweet11** (TMB2011 and TMB2012) and **Tweet13** (TMB 2013) [26, 19]. Using the official API [26], we crawled a set of local copies of the corpus. Our local **Tweet11** collection has a sample of about 16 million tweets, and a set of 49 (TMB2011) and 60 (TMB2012) timestamped topics. While **Tweet13** collection contains about 259 million tweets, and a set of 60 (TMB2013) timestamped topics. In our experiments, we make use of all of the topics. The tweets in the collections should be preprocessed in many aspects. We firstly discarded those non-English tweets, removed all the retweets, normalize elongations, and normalize URLs and author IDs. Moreover, each tweet was stemmed, and the stop-words were removed. For a temporally-anchored ad-hoc search task, each query q in the topic set is assigned with a timestamp. That is only tweets posted prior to T_q were assessed for relevance [26], and no tweets newer than a given timestamp should be retrieved. Hence, based on the preprocessed collections above, we built the final collections consisting of tweets whose timestamps are prior to T_q. We index all tweets in these collections using Lemur toolkit .

As external corpus, Wikipedia articles are used for operating baselines. We preprocess the Wikipedia articles as follows: First, we remove the articles less than 100 words and remove the articles less than 10 links; then we remove all the category pages and disambiguation pages. Moreover, we move the content to the right redirection pages. Finally we obtain about 3.74 millions Wikipedia articles indexed by Lemur toolkit.

4.2 Alternative Algorithms

We compare the proposed models with the following comparative algorithms, including state-of-the-art concept-based algorithm, entity-based algorithm, and topic-based algorithm. Moreover, we also compare with the best submitted results of the previous TREC Microblog Track. The comparative algorithms are listed as follows:

LM: The simple language model [27] is employed as a basic baseline. KL-divergence is used to estimate both θ_q and θ_d with empirical word distribution.

WikiQE: The Wikipedia-based query expansion algorithm utilizes Wikipedia as external corpus to understand query for improving ad-hoc retrieval performance [16].

RTR: Real-time ranking model proposed in [17], is the state-of-the-art real-time ranking model. This approach utilized a two-stage pseudo-relevance feedback query expansion to estimate the query language model. Besides, it adopts a temporal re-ranking component to evaluate the temporal aspects of tweets

EntityQE: As the state-of-art entity-based algorithm, [12] proposed a feedback entity model and integrated it into a two-stage adaptive language modeling framework. Since only its first stage is comparative to ours, we omit the second stage from the entire model as **EntityQE** here.

LCE: Based on latent concept expansion model [23], [25] proposed a microblog-variant by utilizing a temporal relevance model that uses the temporal variation of concepts (e.g., terms and phrases) on microblogs. It is reported as the state-of-art concept-based algorithm.

ConceptQE: Our proposed query expansion framework based on a feedback concept model, with concept information incorporation via short-text conceptualization [35].

TopicQE: It is a variant of the proposed **ConceptQE** by replacing concept information with topic information: It utilizes LDA [4] to generate topic-distribution to represent short-text rather than concept-distribution used in **Con-**

Table 1: Experimental results on TMB2012 and TMB2013.

Query Set	TMB2012		TMB2013	
Metric	MAP	P@30	MAP	P@30
LM	0.271	0.407	0.307	0.493
WikiQE	0.291	0.424	0.317	0.508
TopicQE	0.314	0.442	0.308	0.504
RTR	0.324	0.446	0.351	0.520
EntityQE	0.318^{δ}	0.459^{δ}	0.322^{δ}	0.506^{δ}
LCE	$0.365^{\delta\varphi}$	0.454^{δ}	$0.344^{\delta\varphi}$	$0.511^{\delta\varphi}$
ConceptQE	$\mathbf{0.407}^{\delta\rho\varphi}$	$\mathbf{0.498}^{\delta\rho}$	$\mathbf{0.410}^{\delta\rho\varphi}$	$\mathbf{0.557}^{\delta\rho\varphi}$

ceptQE. Hence, **TopicQE** is a topic-based feedback algorithm.

4.3 Experiment Settings

With the limitation of space, we briefly describe the experimental settings here, mainly including the parameter settings, and the elaborate details are presented in the URL mentioned above. For **WikiQE**, Wikipedia articles are ranked by using KL-divergence, and we select 2/3/5/7 terms according to descending TF-IDF scores from top-5/10/20/100 articles. Then the selected terms are interpolated into the original query. For the proposed **ConceptQE**, the model parameters λ_C and λ_Δ are set as 0.8 and 0.5, respectively (see Section 4.5 for details). The interpolation coefficient α is set as 0.6 in Eq. 2. The exponential parameter r for temporal prior is set as 0.1 in Eq. 6. Wikipedia articles are used for providing expansion terms for **WikiQE** and training **TopicQE**.

Although it is critical to accurately estimate θ_q and θ_d respectively (as mentioned in Section 3), we mainly focus on the estimation of θ_q. Therefore, similar to [24], all the alternative algorithms estimate θ_d as **LM** uniformly. During the constructing pseudo-relevance feedback tweets for all the alternative algorithms, we only collect the most recent 500 tweets as for each query q. In TREC Microblog Track, tweets were judged on the basis of the defined information using a three-point scale [26]: (i) irrelevant (labeled as 0), (ii) minimally relevant (labeled as 1), and (iii) highly relevant (labeled as 2). The main official evaluation metric is Mean Average Precision (MAP) for top 1,000 documents and Precision at N (P@N), which are widely used in traditional IR. Therefore, MAP and P@30 with respect to allrel (i.e. tweet set judged as highly or minimally relevant) are used in this paper. Moreover, statistical t-test are employed here. To decide whether the improvement by method `A` over method `B` is significant, the t-test calculates a value p based on the performance of `A` and `B`. The smaller p is, the more significant is the improvement. If the p is small enough ($p < 0.05$), we conclude that the improvement is statistically significant.

4.4 Testing Performance Summary

We display the experimental results with statistical significance test results in Table 1. The results show the proposed Feedback Concept Model, **ConceptQE**, improves the baseline retrieval models in most cases. The superscript δ, ρ and φ respectively denote statistically significant improvements over **LM**, **LCE** and **EntityQE** ($p < 0.05$). Moreover, **ConceptQE** outperforms the best results of previous TREC Microblog tasks significantly. More specifically, for

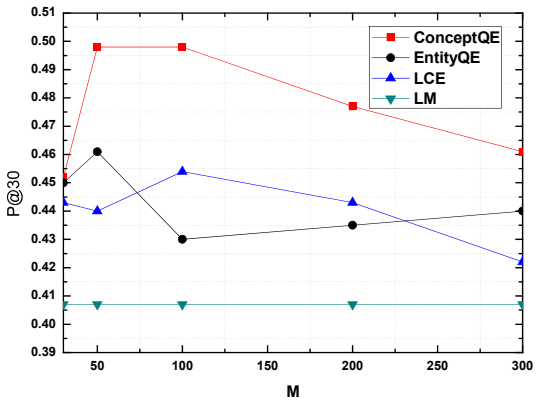

Figure 3: Effect of the increasing number of feedback documents (M) on TMB2012.

TREC 2012 topics (TMB2012), our method **ConceptQE** improves MAP and P@30 over the best submitted result of the task by 50.7% and 68.5%, respectively; while for TREC 2013 topics (TMB2013), our method improves the MAP over the best submitted result of the task by 16.4%.

Note that all the methods estimate the document model (θ_d) as **LM**. As we can see, all of the query expansion methods have significant MAP and P@30 improvements compared with the **LM** method, which indicates the effectiveness of query expansion in microblog retrieval.

The Wikipedia-based method **WikiQE** performs better than **LM**, which indicates the importance of external expansion strategy. Moreover, **ConceptQE** is better than **WikiQE** on both MAP and P@30. This shows the superiority of our Probase-based query expansion method and demonstrate the effectiveness of the structured data. Compared with concept-based (**ConceptQE**) and entity-based algorithms (**EntityQE**), the topic-based **TopicQE** performs worse, demonstrating short-texts are more challenging for topic model. Moreover, our method also beats the state-of-the-art baseline **RTR**, which uses a two-stage query expansion method.

Specifically, for **Tweet13**, the **ConceptQE** improves the MAP over that of **EntityQE** by 19.18%; while the corresponding increment for **Tweet11** topics is relatively faint, which is 11.54%. **EntityQE** generates entity model for each identified entity from the given query, and then weights average of all entity models as the unified entity model by using inverse document frequency (IDF). However, we argue that the entire entity model for a query should not be averaged by all the each entity model simple, which may loss much information. Futhermore, **EntityQE** relies extremely on named entity recognition toolkits which faces problem of vocabulary mismatching, and its time consumption is huge because it has to generate entity model for each entity. On the other hand, IDF does not work well on short-texts, because suffers extremely from the sparsity and noise. **LCE** is reported as the state-of-the-art concept-based algorithm, which is most similar to us in methodology. Table 1 shows that the performance of **ConceptQE** is better than that of **LCE**, indicating the importance of incorporating concept-

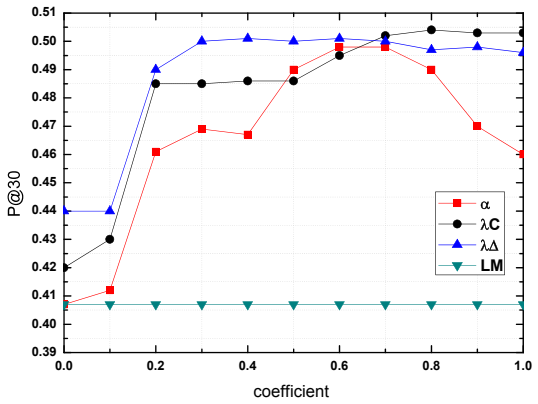

Figure 4: Effect of the increasing coefficients (λ_Δ, λ_C, and α) on TMB2012.

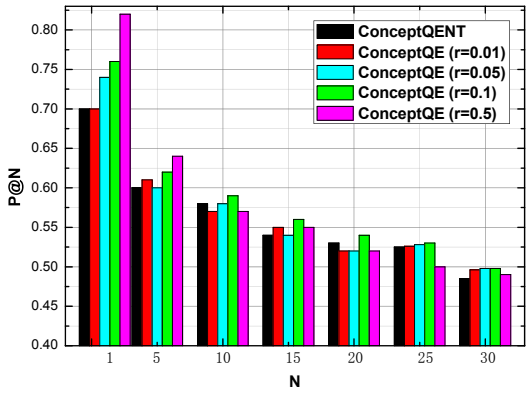

Figure 5: Effect of the exponential rate parameter r for ConceptQE model on TMB2012.

prior information from knowledge base. Moreover, **ConceptQE** needs less feedback tweets compared with **LCE**, which promotes the efficiency of system greatly in practical application.

4.5 Optimization of Parameters

In this section, we analyze the robustness of the parameter setting in the proposed framework, which could affect the overall performance: (i) number of concept-relevant feedback tweets M, (ii) global background noise coefficient λ_Δ, (iii) coefficient of background noise of specific concepts λ_C, (iv) interpolation coefficient α, and (v) temporal rate parameter r. All these experiments in this section are run on TREC 2012 topics (TMB2012). As mentioned above, setting the value for number of feedback tweets (M) is difficult as larger values may introduce a degenerative behavior in the model, as more effort is spent predicting concept language models that are conditioned on unrelated terms, while smaller values of M may lead to cases where the feedback information is not sufficient enough include terms that are semantically related. Figure 3 demonstrates performance change of **LM**, **LCE**, **EntityQE** and **ConceptQE** across different number of feedback documents, which indicates our PRF methods require few feedback document.

Parameters λ_Δ and λ_C control the weight of *global* background noise and *local* background noise of specific concepts, respectively. When evaluating the sensitivity of them, we set interpolation coefficient α as 0.6. Figure 4 shows the performance changes of the **ConceptQE** across different λ_Δ while setting λ_C as 0.8, and across different λ_C while setting λ_Δ as 0.5. If the λ_C is larger, the proposed feedback concept model could filter more specific-concept noise, indicating the importance of incorporating concept-prior information from knowledge base. Conversely, performance change with different λ_Δ is slight, so we set it as 0.5 in experiments. Besides, from Figure 4 we could observe that, when setting interpolation coefficient α around 0.7, **ConceptQE** could reach their optimal P@30 scores, while the performance is descend when the value of α become larger. as it leads to much information loss about the original query.

Previous works [11, 20] have demonstrated that setting temporal rate parameter r (in Eq. 6) for the exponential distribution has a strong effect on retrieval, and affects the expansion terms selected from knowledge base. Apparently, a large r favors the terms that are used recently in the pseudo-relevance documents (CT). For comparison, we generate a variant of the proposed **ConceptQE**, denoted as **ConceptQENT**, which totally ignores temporal evidence. P@N scores of **ConceptQE** with different values of r and **ConceptQENT** are shown in Figure 5 ($N = 1/5/10/15/20/25/30$). From the experimental results, we could observe that an appropriate r can improve the retrieval performance compared with **ConceptQENT**. As has been assumed, retrieval performance benefits from larger value of r when N is small, however small value of r show superiority when N is large.

5. CONCLUSIONS

Microblog retrieval is a challenging task. To derive more semantic signals for inference, we seek help from concept information by leveraging short-text conceptualization. With the knowledge derived from the Probase, the queries in microblogosphere can be more comprehensible and thus more relevant documents can be retrieved. Meanwhile, a mixture model are designed to enable all signals to fully interplay. With the above efforts, we propose a Feedback Concept Model to finish concept-level query expansion for microblog retrieval task. Moreover, we incorporate the temporal evidence into query representation. Thus the proposed method favors recent tweets which satisfies the real-time information need in microblog retrieval. The experimental results demonstrate that the proposed model performs the best and shows significant improvement over the previous methods for microblog retrieval.

6. ACKNOWLEDGMENTS

The work was supported by the National Basic Research Program of China (973 Program, Grant No. 2013CB329303), and the State Key Program of Joint Funds of the National Natural Science Foundation of China (Grant No. U1636203).

7. REFERENCES

[1] F. Abel, I. Celik, G. J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on twitter. In *International Semantic Web Conference*, pages 1–17, 2011.

[2] F. Ahmed and A. Rnberger. Evaluation of n-gram conflation approaches for arabic text retrieval. *Journal of the Association for Information Science and Technology*, 60(7):1448–1465, 2009.

[3] M. D. Albakour, C. Macdonald, and I. Ounis. On sparsity and drift for effective real-time filtering in microblogs. In *ACM International Conference on Conference on Information and Knowledge Management*, pages 419–428, 2013.

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

[5] G. Cao, J. Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July*, pages 243–250, 2008.

[6] P. Castells, M. Fernĺćndez, and D. Vallet. An adaptation of the vector-space model for ontology-based information retrieval. *Knowledge and Data Engineering IEEE Transactions on*, 19(2):261–272, 2007.

[7] S. Choi, J. Choi, S. Yoo, H. Kim, and Y. Lee. Semantic concept-enriched dependence model for medical information retrieval. *Journal of Biomedical Informatics*, 47(2):18–27, 2014.

[8] W. Dakka, L. Gravano, and P. G. Ipeirotis. Answering general time sensitive queries. In *ACM Conference on Information and Knowledge Management, CIKM 2008, Napa Valley, California, Usa, October*, pages 1437–1438, 2008.

[9] F. Diaz, B. Mitra, and N. Craswell. Query expansion with locally-trained word embeddings. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 367–377, 2016.

[10] A. Dong, R. Zhang, P. Kolari, J. Bai, F. Diaz, Y. Chang, Z. Zheng, and H. Zha. Time is of the essence: improving recency ranking using twitter data. In *19th International Conference on World Wide Web*, pages 331–340, 2010.

[11] M. Efron and G. Golovchinsky. Estimation methods for ranking recent information. In *Proceeding of the International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2011, Beijing, China, July*, pages 495–504, 2011.

[12] F. Fan, R. Qiang, C. Lv, and J. Yang. Improving microblog retrieval with feedback entity model. In *The ACM International*, pages 573–582, 2015.

[13] W. Hua, Y. Song, H. Wang, and X. Zhou. Identifying users' topical tasks in web search. In *ACM International Conference on Web Search and Data Mining*, pages 93–102, 2013.

[14] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119, 2002.

[15] X. Li and W. B. Croft. Time-based language models. In *Twelfth International Conference on Information and Knowledge Management*, pages 469–475, 2003.

[16] Y. Li, W. P. R. Luk, K. S. E. Ho, and F. L. K. Chung. Improving weak ad-hoc queries using wikipedia asexternal corpus. In *SIGIR 2007: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, July*, pages 797–798, 2007.

[17] 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, 2012.

[18] N. Limsopatham, C. Macdonald, and I. Ounis. Learning to combine representations for medical records search. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 833–836, 2013.

[19] J. Lin, M. Efron, Y. Wang, and G. Sherman. Overview of the trec-2014 microblog track. 2015.

[20] C. Lv, R. Qiang, F. Fan, and J. Yang. Knowledge-based query expansion in real-time microblog search. In *Asia Information Retrieval Symposium*, pages 43–55, 2015.

[21] Y. Lv and C. X. Zhai. A comparative study of methods for estimating query language models with pseudo feedback. In *ACM Conference on Information and Knowledge Management*, pages 1895–1898, 2009.

[22] K. Massoudi, M. Tsagkias, M. D. Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. In *Advances in Information Retrieval - European Conference on Ir Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*, pages 362–367, 2011.

[23] D. A. Metzler. Automatic feature selection in the markov random field model for information retrieval. In *Sixteenth ACM Conference on Information and Knowledge Management, CIKM 2007, Lisbon, Portugal, November*, pages 253–262, 2007.

[24] T. Miyanishi, K. Seki, and K. Uehara. Improving pseudo-relevance feedback via tweet selection. In *ACM International Conference on Information and Knowledge Management*, pages 439–448, 2013.

[25] T. Miyanishi, K. Seki, and K. Uehara. Time-aware latent concept expansion for microblog search. In *8th International AAAI Conference on Weblogs and Social Media*, pages 366–375, 2014.

[26] I. Ounis, C. Macdonald, and J. Lin. Overview of the trec-2011 microblog track. 2011.

[27] J. M. Ponte. A language modeling approach to information retrieval. In *21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 275–281, 1998.

[28] A. Severyn, A. Moschitti, M. Tsagkias, R. Berendsen, and M. D. Rijke. A syntax-aware re-ranker for microblog retrieval. In *37th International ACM SIGIR*

Conference on Research and Development in Information Retrieval, pages 1067–1070, 2014.

[29] Y. Song and D. Roth. On dataless hierarchical text classification. In *28th AAAI Conference on Artificial Intelligence*, pages 1579–1585, 2014.

[30] Y. Song, H. Wang, W. Chen, and S. Wang. Transfer understanding from head queries to tail queries. In *23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1299–1308, 2014.

[31] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen. Short text conceptualization using a probabilistic knowledgebase. In *IJCAI 2011, Proceedings of the International Joint Conference on Artificial Intelligence*, pages 2330–2336, 2011.

[32] Y. Song, S. Wang, and H. Wang. Open domain short text conceptualization: a generative + descriptive modeling approach. In *International Conference on Artificial Intelligence*, pages 3820–3826, 2015.

[33] F. Wang, Z. Wang, Z. Li, and J. R. Wen. Concept-based short text classification and ranking. In *The ACM International Conference*, pages 1069–1078, 2014.

[34] Y. Wang, H. Huang, C. Feng, Q. Zhou, J. Gu, and X. Gao. Cse: Conceptual sentence embeddings based on attention model. In *54th Annual Meeting of the Association for Computational Linguistics*, pages 505–515, 2016.

[35] Z. Wang, K. Zhao, H. Wang, X. Meng, and J. R. Wen. Query understanding through knowledge-based conceptualization. In *International Conference on Artificial Intelligence*, pages 3264–3270, 2015.

[36] W. Wu, H. Li, H. Wang, and K. Q. Zhu. Probase:a probabilistic taxonomy for text understanding. In *ACM SIGMOD International Conference on Management of Data*, pages 481–492, 2012.

[37] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *10th International Conference on Information and Knowledge Management*, pages 403–410, 2001.

[38] Z. Zhang and L. Man. Estimating semantic similarity between expanded query and tweet content for microblog retrieval.