別紙４－１（課程博士（英文））

Date of Submission（01-11-2018）：

Department	Student ID Number	D145309	Supervisors	Masaki Aono, Kyoji Umemura
Computer Science and Engineering				
Applicant's name	Abu Nowshed Chy			

Abstract（Doctor）

Title of Thesis	Exploiting Temporal and Semantic Information for Microblog Retrieval through Query Expansion and Reranking Approaches (クエリの拡張と再順位付けアプローチによるマイクロブログ検索のための時間的および意味的情報の活用)

Approx. 800 words

Nowadays, microblog websites are not only the places in maintaining social relationships but also act as a valuable information source. Everyday lots of users turn into these sites to fulfill their diverse information needs. Moreover, during a disaster period, microblogs are treated as an important source to serve the situational information needs. Among several microblog services, twitter is now the most popular. Hence, information retrieval in twitter has made a hit with a lot of complaisance.

However, due to the short length characteristics of tweets, people usually use unconventional abbreviations (e.g. use "2day" instead of "today"), poor linguistic phrases (e.g. use "TYT" instead of "take your time"), and URL to express their concise thought. Besides this, some twitter specific syntaxes (e.g. #hashtags, retweets) also very popular among twitter users. Moreover, users usually search the temporally relevant information in twitter, such as breaking news and real-time events. All these characteristics make it challenging for effective information retrieval (IR) over tweets.

In this thesis, we propose two different approaches to tackle the challenges of microblog retrieval. At first, we propose a reranking based approach, where the main focus is to rerank the tweets that are retrieved by using the baseline retrieval model. Whereas, our proposed query expansion based approach augments the original queries with expansion terms that best represent the users' intent. In both approaches, we used the Lucene's implementation of query likelihood as the baseline retrieval model.

In our reranking based approach, we emphasize the alleviation of vocabulary mismatch, and the leverage of the temporal (e.g. recency and burst nature) and contextual characteristics of tweets. One way of alleviating the vocabulary mismatch problem is to reformulate the query via query expansion. In this regard, we propose a three-stage query expansion technique by leveraging the pseudo-relevant tweets at the first stage, made use of Web search results at the second stage, and extracted hashtags relevant to the query at the third stage. For weighting terms, we used the IDF-score of each term.

In the feature extraction stage, we extract several effective features for reranking by leveraging the different tweet characteristics. To extract temporal aspect of tweets, we determine the temporal dimension of the query and if the query is temporally sensitive, we extract our proposed temporal features such as recency score of tweets by utilizing the query time and tweet time. We also hypothesize that any tweet might have burst-time popularity once the tweet has been posted. In order to implement our hypothesis, we propose to introduce a "burst-time" aware temporal feature as well. To extract the content-aware features, we utilize some classical information retrieval models. Along with this direction, we extract the twitter specific features such as URL and retweet count and account related features such as followers count and status count to address the special characteristics of tweets and relations between twitter users. Moreover, we propose some context relevance features based on word embedding, kernel density estimation, and query-tweet sentiment correlation to address the contextual dimension of tweets. We hypothesize that a query is sentimentally sensitive if the largest proportion of the initially retrieved tweets has the similar kind of sentiment polarity. Once our proposed features are extracted, a supervised feature selection method based on regularized regression is applied to select the best feature combination. After estimating the feature importance using the random forest, an ensemble of learning to rank (L2R) framework is applied to estimate the relevance of a query-tweet pair.

However, the naive query expansion technique that we proposed in our reranker framework used the IDF-score to rank each term which might induce irrelevant rare terms from the noisy tweet contexts. Hence, selecting terms by utilizing the unsupervised approach and highly reliant on the top retrieved results without considering temporal relevance may generate the noisy or harmful expansion terms which degrade the retrieval performance.

Considering the above limitations, we present another query expansion approach, where supervised learning is adopted for selecting expansion terms. Upon retrieving tweets by our proposed topic modeling based query expansion (TMQE), we utilize the pseudo relevance feedback (PRF) and a new temporal relatedness (TR) approach to select the candidate tweets. Next, we devise several new features to select the temporally and semantically relevant expansion terms by leveraging the temporal, word embedding, and sentiment association of candidate term and query. Moreover, we also utilize the lexical and twitter specific features to quantify the term relatedness. After supervised feature selection using regularized regression, we estimate the feature importance by applying random forest. Then, we design a linear learning to rank (L2R) model with the aid of feature values and their importance weight to rank the candidate expansion terms.

Experimental result on TREC Microblog 2011 and 2012 test collections over the TREC Tweets2011 corpus demonstrate the effectiveness of our proposed reranking and query expansion approaches over the baseline and state-of-the-art methods.