Patterns Based Query Expansion for Enhanced Search on Twitter Data

Meryem Bendella[1] and Mohamed Quafafou[2]

[1] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
meryem.bendella@etu.univ-amu.fr
[2] Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
mohamed.quafafou@univ-amu.fr

Abstract. Social microblogging services have an especially significant role in our society. Twitter is one of the most popular microblogging sites used by people to find relevant information (e.g., breaking news, popular trends, information about people of interest, etc). In this context, retrieving information from such data has recently gained growing attention and opened new challenges. However, the size of such data and queries is usually short and may impact the search result. Query Expansion (QE) has a main task in this issue. In fact, words can have different meanings where only one is used for a given context. In this paper, we propose a QE method by considering the meaning of the context. Thus, we use patterns and Word Embeddings to expand users' queries. We experiment and evaluate the proposed method on the TREC 2011 dataset containing approximately 16 million tweets and 49 queries. Results revealed the effectiveness of the proposed approach and show the interest of combining patterns and word embedding for enhanced microblog retrieval.

Keywords: Query expansion, Patterns, Word Embeddings, Microblog retrieval

1 Introduction

Twitter is one of the most popular social media platforms that enable users to post short texts (up to 280 characters for one text) called tweets. Nowadays, users can share ideas, opinions, emotions, suggestions, daily stories and events through this platform. Many of these online services start to be part of daily life of millions of people around the world. However, a huge quantity of information is created in these platforms, hence, finding recent and relevant information is challenging.

There are many users who are interested in collecting recent information from such platforms. This information can be related to a particular event, a specific topic or popular trends. Users express their need through a query to search posts (tweets). According to [25], most people use few search terms and few modified queries in Web searching. However, query formulation becomes difficult for the user in order to express appropriately what he is looking for. Query Expansion (QE) plays a considerable contribution towards fetching relevant results in this case. An expanded query will contain more related terms (called candidate terms) to increase the chances of rendering the maximum number of relevant documents. The objective is to find the microblogs answering to a need for information specified by a user.

In this paper, we propose a new method based on formal concept analysis and word embeddings to expand user queries. In order to achieve this goal, we prepare our dataset collection by preprocessing the tweet text which is a critical step in information retrieval (IR) and Natural Language Processing (NLP). Then, each tweet is represented as a set of words and will be indexed by the Terrier system[3]. Next, we use this system to retrieve tweets according to the TREC2011 query set [19]. We used the BM25 [11] model to retrieve relevant tweets answering original queries. After that, these retrieved tweets are used to extract frequent closed patterns. This can be defined as the frequent closed patterns of words contained in tweets dataset. Furthermore, word embeddings are trained on our textual dataset by using Word2Vec model [16]. Indeed, we expand original queries by combining patterns and word embeddings approaches. This combination consists in enriching the query by finding most closely related words to patterns' words. The proposed model extends the semantics used in the original query and improves the results search of microblogs.

The rest of the paper is organized as follows: We cover related work on query expansion for microblogs retrieval in Section 2. Section 3 describes all steps required in our proposed approach to expanding queries for microblogs retrieval. The proposed approach is given in Section 4. The Section 5 is dedicated to experiments and evaluations. Finally, the conclusion and future work are presented in Section 6.

2 Related Work

Query expansion (QE) has recently gained growing attention in IR domain and there is considerable research addressing the short query problem. Web queries posted by users can be too short and that makes the search results not focused on the topic of interest.

Much effort has been made to improve microblog retrieval. [14] explored use of three different IR query expansion techniques in order to enhance the search results. In [15] authors propose a retrieval model for searching microblog posts for a given topic of interest. This model is based on a combination of quality indicators and the query expansion model.

[3] Terrier is an effective open source search engine (Information Retrieval system), readily deployable on large-scale collections of documents

Query expansion approaches for microblog retrieval can be divided into three groups which are local, global and external [20]:

- **Local**: Local QE techniques select candidate expansion terms from a set of documents retrieved in response to the original (unexpanded) query. This kind of approach is known as Pseudo-Relevance Feedback (PRF). It is widely used for query expansion in research of microblog search [6,15,28]. This approach consists in using terms derived from the top N retrieved documents (relevant documents) to retrieve other similar documents which are also likely to be relevant. In [13], authors propose an algorithm of extracting features from tweets using frequent patterns. Their QE method is based on PRF approach by applying weights for different features.
- **Global**: Global QE approaches select the expansion terms from the entire database of documents. These techniques select candidate terms by mining term-term relationships from the target corpus [20]. This type of techniques has been used in many applications [9,17], and was one of the first techniques to produce consistent effectiveness improvements through automatic expansion [3]. In our work, we use, in a certain way, the global analysis by training word embeddings on the entire dataset in order to extract terms that are most similar to the patterns. In [26], authors propose an QE method using word embeddings and some external ressources.
- **External**: External QE techniques comprise methods that obtain expansion terms from other resources besides the target corpus [20]. Several approaches have been proposed to use external resources such as *Wikipedia*, *WordNet* and *DBpedia* to improve query expansion [12,1,29].

3 Pattern Mining-Based Query Expansion

In this section, we describe all steps required in our proposed approach to expanding queries for microblogs retrieval. As a first step, we have been concentrating on data preprocessing. Then, we describe pattern extraction task and define notations and basic notions necessary for understanding the proposed approach.

3.1 Preprocessing

This process comprise preprocessing of tweet's text i.e., dealing with stop words, emoticons, punctuation, stemming, etc. Microblogs data such as tweets are too short, generally not well written and do not respect the grammar. However, this preliminary step is very crucial to eliminating the noise and cleaning data. We, therefore, prepare the dataset for indexing by filtering tweets as follows: (1) Removing null tweets and short tweets which contain less than two words; (2) Removing Retweets (tweets starting with RT followed by username) as they would be judged as non-relevant; (3) Removing non-English tweets; (4) Eliminating the non-ASCII content found in any of the English tweets; (5)Removing link and mentions from the tweet.

Furthermore, we perform tokenization, tweet normalization, text stemming, and stopwords removal, as part of the preprocessing phase. Tokenization is the process of breaking each tweet up into words or other meaningful elements called tokens. Then, we stem all tokens present in tweets except hashtags. We have used the standard Porter stemmer of Stanford NLP tool[4]. After that, we remove English stopwords[5] that are present in tweets. We also perform normalization of the tweet content, by resolving words containing many repeated letters, such as the word "yes" or "happy", they may appear as "yeeees" or "happyyy" on Twitter. The output of preprocessing task is used for indexing.

3.2 Patterns extraction and Formal concept analysis

Frequent pattern has an important and active role in many data mining tasks [8]. This comprises to find interesting patterns (sets of items) called frequent itemsets from databases. It was initiated by Agrawal et al. [2] and it corresponds to finding the sets of attributes (or items) that appear simultaneously in at least a certain number of objects (or transactions) defined in an extraction context (see definition 1).

Construct Transactional Dataset The transactional dataset is represented by a set of preprocessed tweets. Each tweet is represented as a set of words which are considered as itemsets. We provide an example of transactional dataset in Table 1.

Table 1. Transactional dataset example.

Id	**Items**				
t1	A		C	D	
t2		B	C		E
t3	A	B	C		E
t4		B	C		E
t5	A	B	C		

Pattern Mining We provide the following definitions for patterns extraction:

Definition 1. *(Data mining context). A data mining context is a triple* $\mathcal{D} = (\mathcal{T}, \mathcal{I}, \mathcal{R})$ *composed of a set of transactions* $\mathcal{T}$ *, a set of items* $\mathcal{I}$ *and a binary relation between transactions and items* $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{I}$*. Each couple* $(t, i) \in \mathcal{R}$ *denotes the fact that the transaction* t *is related to the item* i*.*

Definition 2. *(Pattern, cover and support). A pattern* X *is a subset of items* $X \subseteq \mathcal{I}$*. Its cover and its support are defined by:*

$$cover(X) = \{t \in \mathcal{T} | \forall i \in X, (t, i) \in \mathcal{R}\} \qquad support(X) = |cover(X)|$$

Definition 3. *(Frequent patterns). Given a context of data mining* $\mathcal{D}$ *and minsup the minimum support, the set of frequent itemsets in* $\mathcal{D}$ *is:*

$$\mathcal{FI} = \{X \subseteq \mathcal{I} | support(X) \geq minsup\}.$$

[4] https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

[5] https://github.com/ravikiranj/twitter-sentiment-analyzer/blob/master/data/feature_list/stopwords.txt

In this work, we are interested in frequent closed patterns (denoted $\mathcal{FCI}$), which has been proposed by [21]. The pattern X is called closed if none of its supersets have the same support as X. In other words, $\forall Y,\ X \subset Y$, $support(Y) < support(X)$. Here, Y is a superset of X.

Formal Concept Analysis (FCA) Formal concept analysis (FCA) is a theory of data analysis identifying the conceptual structures within data sets. It also presents an interesting unified framework to identify dependencies among data, by understanding and computing them in a formal way [4]. This gives it the advantage to be an effective technique to analyze the different pattern relationships such as in Social Network [24], and Information Retrieval [5]. Given a formal context $\mathcal{D}$, there is a unique ordered set which describes the inherent lattice structure defining natural groupings and relationships among the transactions and their related items. This structure is known as a concept lattice or Galois lattice [7]. Each element of the lattice is a couple (T, I) which consists of a set of transactions (i.e., the *extent*) and a set of items (i.e., the *intent*).

Let X be a closed pattern of items (words), a formal concept is composed of X and of the set of tweets containing this closed pattern.

4 Frequent closed patterns and Word Embeddings to expand the query

In this section, we describe our proposed method to expand queries for the microblog retrieval task. The proposed method is based on frequent concept extraction and Word Embeddings. It is composed of three main steps: (1) Generate frequent closed itemsets and select patterns, (2) Extend patterns using Word Embeddings, and (3) expand the query (see Figure 1).

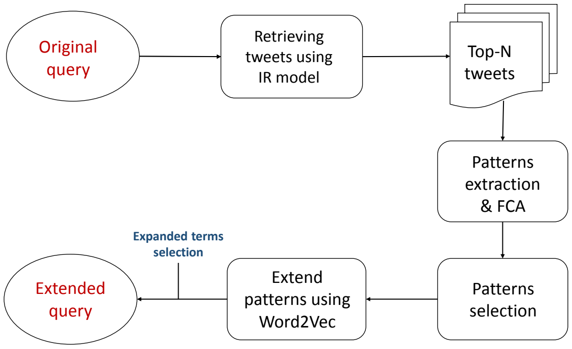
Original query
Retrieving tweets using IR model
Top-N tweets
Patterns extraction & FCA
Patterns selection
Extend patterns using Word2Vec
Expanded terms selection
Extended query

Fig. 1. Overview of the proposed query expansion approach.

4.1 Frequent concept extraction

We perform patterns extraction on top-N tweets returned by Terrier system in the initial search. The topmost relevant tweets are retrieved by using Terrier with original queries. After that, we discover the closed frequent itemsets in this large database of transactions (retrieved tweets). This process is performed by using Charm-L algorithm [27], where the discovery of patterns computes the closed sets of items (i.e., words) that appear together in at least a certain number of transactions (i.e., tweets) recorded in a database. This number is called threshold and it is defined empirically. We compute the frequent concept lattice $\mathcal{L}$ according to the minimal support threshold value minsup. Each node of the formal concept lattice $\mathcal{L}$ represents the correspondence between a pattern and the set of tweets that contain the words of this pattern. The parameters N (for N-top tweets) and *minsup* are chosen according to experiments we conducted (described in section 5.4).

The algorithm for generating the complete set of interesting frequent closed patterns and selection of candidate terms for the query expansion is shown in Algorithm 1, with L_K denoting the set of K closed frequent itemsets, T the set of n transactions which represent tweets, *minsup* the minimal support threshold value, and $\mathcal{Q}_P$ selected patterns for the expanded query. The list of K patterns contains all interesting patterns found in top-N retrieved documents in the initial query. The top-3 patterns with a high support value and with common tweets are selected to represent candidate terms for the expanded query.

Algorithm 1 Patterns extraction

```
Require:
    $\mathcal{T}$: a set of $n$ tweets
    $\mathcal{W}$: Vocabulary of all words contained in tweets dataset
    $minsup$: Minimum support threshold value
    $\mathcal{R}$: a binary relation where $\mathcal{R} \subseteq \mathcal{T} \times \mathcal{W}$
Ensure:
    $\mathcal{L}_K$: List of K patterns
    $\mathcal{Q}_P$: selected patterns for the expanded query

Creation of formal context $\mathcal{D}$ =($\mathcal{T}, \mathcal{W}, \mathcal{R}$);
Computation of frequent concept lattice $\mathcal{L}$ for $\mathcal{D}$ according to $minsup$;
$\mathcal{L}_K = CharmL(\mathcal{T}$, minsup);
$\mathcal{Q}_P$ =Top-3$\{argmax(Support(\mathcal{L}_{K_i})\}$
return $\mathcal{L}_K, \mathcal{Q}_P$;
```

4.2 Word Embeddings: Word2Vec model

Word Embeddings has been used in query expansion task to enhance search [22]. In this paper, we combine them with frequent closed patterns to expand queries for microblogs retrieval. We train word embeddings on the entire dataset in order to extract terms that are most similar to the selected patterns computed in the previous step. For each query, we extend terms of the selected patterns by adding the most similar terms contained in the dataset that are likely to be relevant to the query but do not appear in the patterns. For the training of *embeddings*, we used a project alternative WORD2VEC [6] implemented in programming language JAVA by MEDALLIA team[7] to integrate it into our main program of query expansion implemented in JAVA. The training of the neural network is carried out on the preprocessed corpus TREC 2011 on which search is performed. This model estimates the probability that a term will appear in a position in a text based on terms that appear in a window around this position. Each term in the dataset is represented by a vector embedded in a vector space. Similarities between these vectors were shown to correspond to semantic similarities between terms [16].

For the setting of the neural network, we have set a window with a size of 7 words (the appearance frequency of the words is of a minimum equal to 5, the dimensions of the vectors are 200, the number negative examples is 7 with a use of the hierarchical alternative *softmax*. Specifically, the continuous bag of words model (CBOW) is used.

4.3 Query expansion

Given an original query $q = \{w_{q1}, ..., w_{qn}\}$, the process of expanding q is threefold: (1) retrieving relevant tweets answering q using a retrieval model, (2) selecting a set of candidate terms (CT) for q by extracting *patterns* from the top-N ranked tweets, and using formal concept analysis, (3) selecting the most related terms to the CT set using *Word2Vec* model, to add only terms that are semantically related to q. These terms are then selected to obtain the set of terms which represent the expanded query denoted eq, with $eq = q \bigcup \{w_{eq_1}, ..., w_{eq_m}\}$.

The process of obtaining candidate terms consists in selecting terms from the patterns set as detailed in section 4.1. For extending these terms, we used Word2Vec model in order to select terms that are semantically related to the obtained patterns. The process for selecting these terms computes the cosine similarity between the corresponding pattern-term-vector and each tweet-term-vector in the corpus, and rank the words in decreasing order of the cosine similarity.

5 Experimental Evaluation

In this section, we conduct experiments to evaluate the effectiveness of the proposed query expansion method. To demonstrate the performance of our proposed method, we compare our patterns-based query expansion method with several methods. Our experiments are conducted on the TREC 2011 collection.

[6] https://github.com/medallia/Word2VecJava

[7] http://engineering.medallia.com

5.1 Dataset description

In order to evaluate the proposed approach, we use the Twitter data collection (TREC 2011 Microblog Track Data). This dataset contains approximately 16 million tweets collected over a period of 2 weeks (24th January 2011 until 8th February) [19]. Since the provided dataset contains only tweet ids, on the whole, we have gathered around 12 million tweets with content. The rest of the tweets, either was removed by their editors or we have no access to them. After performing the data filtering and processing task explained in section 2, we obtained a dataset of around 3.5 million tweets on which our experiments were conducted. We have used 49 queries defined by TREC track 2011 [19].

5.2 Retrieval model

We use the well-known Terrier IR system[8] to index our data collection (TREC 2011). Terrier, an open source software, offers a range of document weighting and query expansion models. It has been successfully used for ad-hoc retrieval, cross-language retrieval, Web IR and intranet search [18]. All the tweets (after the preprocessing task) were indexed using Terrier, and the original queries were used to retrieve and rank tweets using the standard BM25 retrieval model [11] of Terrier. BM25 model has been used extensively within the TREC community on a variety of corpora.

5.3 Evaluation metrics

We use Precision, MAP and nDCG metrics, which are widely used in information retrieval [23], to evaluate the proposed method for query expansion. Moreover, we evaluated performance of the proposed method according to its Precision, Recall, F-measure, and R-PREC metrics in order to compare the performance and effectiveness of the proposed approach with other approaches. The MAP (Mean Average Precision) for a set of queries is the mean of the average precision scores for each query.

Normalized discounted cumulative gain (nDCG) is a measure of retrieval quality for ranked documents that, in contrast to precision, makes use of graded relevance assessments [10]. nDCG is computed as follows:

$$nDCG = Z_i \sum_{j=1}^{R} \frac{2^{r(j)} - 1}{log(1+j)'} \tag{1}$$

Here, Z_i is a constant to normalize the result to the value of 1. $r(j)$ is an integer representing the relevance level of the result returned at rank j where R is the last possible ranking position. In our case, the relevance levels are 0 (irrelevant), 1 (relevant), and 2 (high relevant). nDCG@n is a variation of nDCG where only the top-n results are considered.

[8] http://terrier.org/

For our experimental evalutation, we compute nDCG@10, MAP, P@5, P@10, and P@30 by using TREC EVAL [9]. P@5, P@10 and P@30 represent respectively, precision considering only the top 5, top 10, and top 30 results returned by the system.

5.4 Experimental results

We evaluate the effectiveness of our proposed approach by using evaluation metrics detailed in section 5.2. We conducted two different runs in our experiments, the first one is based on patterns, and the second one represents the combination of patterns and Word Embeddings. Table 2 reports these runs performance compared with the baseline run. In this work, we define our baseline as a single run which was generated using Terrier system, selecting the most recent 1000 tweets that contain any of the query terms. In other words, without any query expansion process. We also carried out a test on a method based on Pseudo-Relevance Feedback (PRF) in order to compare our results. This expansion method is applied using the Bo1 term weighting model implemented in terrier.

Table 2. Evaluation results of the proposed approach.

Metrics	P@10	P@30	MAP	nDCG@10	nDCG
Baseline	0.1184	0.0905	0.1025	0.1146	0.2659
PRF	0.2245	0.2116	0.1759	0.2067	0.3876
Run-P	0.2980	0.2415	0.1929	0.2619	0.3938
Run-P-WE	**0.3449**	**0.2878**	**0.2403**	**0.3077**	**0.4476**

Pattern mining utilized for short query expansion was effective for most queries. Due to the shortness of queries and content of tweets, some queries give a poor performance in evaluations. That's why we used Word Embeddings to extend patterns and enrich the query. This additional method added to the query expansion model leads to a significant improvement compared to the pattern and baseline runs as shown in figure 2. Compared to the *baseline*, we have obtained significant improvements over the four measures : +191,30%, +218,01%, +134,43% and +168,5% respectively on P@10, P@30, MAP and nDCG@10.

[9] https://trec.nist.gov/trec_eval/

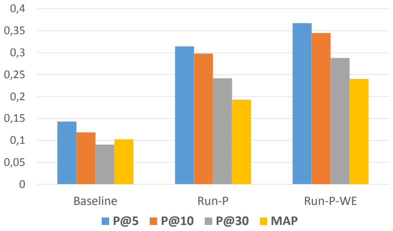

Fig. 2. Comparison of our runs with respect to the baseline run.

Table 3 shows Precision at 30 (P@30), Mean Average Precision (MAP) and Reciprocal Precision (R-Prec) for our proposed approach and for the different expansion methods proposed by authors in [13]. The focus of the comparison is on this approach as since the authors used the concept of co-occurence (frequent patterns) and have evaluated the effectiveness of the proposed approach on the same dataset (TREC 2011).

Table 3. Comparison of our proposed method with the QE method proposed by [13].

Metric/RunID	Run1A	Run2A	Run3A	Run4A	Run-P-WE
P@30	0.1347	0.1694	0.2034	0.1973	**0.2878**
MAP	0.0753	0.0486	0.0673	0.0486	**0.2403**
R-PREC	0.1114	0.0846	0.1191	0.1040	**0.2475**

Table 4 shows MAP, Normalized discounted cumulative gain (NDCG) and F-measure for the proposed approach compared to the query expansion method introduced by authors in [26], where they proposed a new framework for query expansion based on multiple sources of external information. We compared our run with their different runs according to the MAP, NDCG, and F-measure scores. We observe that our proposed model gives a good performance compared to their runs.

Table 4. Comparison of our proposed method with the QE method proposed by [26].

Metric/RunID	NMF+Query	Word2Vec	NMF+W2V	Run-P-WE
MAP	0.036	0.093	0.027	**0.2403**
NDCG	0.226	0.219	0.272	**0.4476**
F-measure	0.101	0.039	0.092	**0.2646**

In overall, the experimental results show that the approach leveraging patterns and Word Embeddings outperforms the baseline and some methods of the literature. The formal concept lattice we used to extend queries selects the related terms which appear together in documents (tweets). Also, the extraction of interesting frequent patterns allows us to select the most important terms related to the initial query by considering the top N retrieved documents.

In our empirical study, we fixed the number of retrieved documents in the initial query to 500 (N). When this number increases to 1000, there is no significant improvement. We also have varied the minsup value for extracting patterns and have chosen 10 (i.e. 2%). All experiments we report on evaluation metrics are performed using TREC Eval.

6 Conclusion

In this paper, we proposed a query expansion method to enhance microblogs search. The shortness of the microblogs and the queries may impact the quality of search. Our proposed method is based on frequent closed patterns and formal concept analysis. The frequent closed patterns are combined with word embedding for finding the words that are most similar to the original query. The results revealed the effectiveness of the proposed approach and show the interest of combining patterns and word embedding to enhance microblog search.

In our future work, it will be interesting to investigate temporal information presented in tweets. We also propose to integrate the location information while searching within the tweets where the query can be composed of region-of-interest (ROI) and text. We will further investigate external resources to compare our proposed method with external QE approaches.

References

1. Aggarwal, N., Buitelaar, P.: Query expansion using Wikipedia and DBpedia. In: CLEF (Online Working Notes/Labs/Workshop) (2012)
2. Agrawal, R., Imieliński, T., Swami, A.: Mining association rules between sets of items in large databases. SIGMOD Rec. **22**(2), 207–216 (Jun 1993). https://doi.org/10.1145/170036.170072
3. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Comput. Surv. **44**(1) (Jan 2012)
4. Codocedo, V., Baixeries, J., Kaytoue, M., Napoli, A.: Contributions to the Formalization of Order-like Dependencies using FCA. In: What can FCA do for Artificial Intelligence? The Hague, Netherlands (Aug 2016)
5. Codocedo, V., Napoli, A.: Formal Concept Analysis and Information Retrieval – A Survey. In: International Conference in Formal Concept Analysis - ICFCA 2015. vol. 9113, pp. 61–77. Springer, Nerja, Spain (Jun 2015)
6. Diaz, F., Mitra, B., Craswell, N.: Query expansion with locally-trained word embeddings. CoRR **abs/1605.07891** (2016)
7. Ganter, B., Wille, R.: Formal concept analysis: mathematical foundations. Springer Science (1999)

8. Han, J., Cheng, H., Xin, D., Yan, X.: Frequent pattern mining: current status and future directions. Data Min. Knowl. Discov. **15**(1), 55–86 (2007)
9. Hu, J., Deng, W., Guo, J.: Improving retrieval performance by global analysis. In: 18th International Conference on Pattern Recognition. vol. 2, pp. 703–706 (2006)
10. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Trans. Inf. Syst. **20**(4), 422–446 (Oct 2002)
11. Jones, K.S., Walker, S., Robertson, S.E.: A probabilistic model of information retrieval: Development and comparative experiments. Inf. Process. Manage. **36**(6) (Nov 2000)
12. Kotov, A., Zhai, C.: Tapping into knowledge base for concept feedback: Leveraging ConceptNet to improve search results for difficult queries. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. pp. 403–412. WSDM '12, ACM, New York, NY, USA (2012)
13. Lau, C., Li, Y., Tjondronegoro, D.: Microblog retrieval using topical features and query expansion. Proceedings of The Twentieth Text REtrieval Conference (November 15-18 2011)
14. Li, W., Jones, G.J.F.: Comparative evaluation of query expansion methods for enhanced search on microblog data: DCU ADAPT @ SMERP 2017 workshop data challenge. In: Proceedings of the First International Workshop on Exploitation of Social Media for Emergency Relief and Preparedness co-located with European Conference on Information Retrieval. pp. 61–72 (2017)
15. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. In: Proceedings of the 33rd European Conference on Advances in Information Retrieval. pp. 362–367. ECIR'11, Springer-Verlag, Berlin, Heidelberg (2011)
16. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. In: Proceedings of the International Conference on Learning Representations. pp. 1–12 (2013)
17. Mittal, N., Nayak, R., Govil, M.C., Jain, K.C.: Dynamic query expansion for efficient information retrieval. In: 2010 International Conference on Web Information Systems and Mining. vol. 1, pp. 211–215 (Oct 2010)
18. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Johnson, D.: Terrier information retrieval platform. In: Proceedings of the 27th European Conference on Advances in Information Retrieval Research. pp. 517–519. Springer-Verlag, Berlin, Heidelberg (2005)
19. Ounis, I., Macdonald, C., Lin, J., Soboroff, I.: Overview of the trec-2011 microblog track. In: In Proceedings of TREC 2011 (2011)
20. Pal, D., Mitra, M., Bhattacharya, S.: Exploring query categorisation for query expansion: A study. CoRR **abs/1509.05567** (2015)
21. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. Inf. Syst. **24**(1), 25–46 (Mar 1999)
22. Roy, D., Paul, D., Mitra, M., Garain, U.: Using word embeddings for automatic query expansion. CoRR **abs/1606.07608** (2016)
23. Sanderson, M.: Test collection based evaluation of information retrieval systems. Foundations and Trends® in Information Retrieval **4**(4), 247–375 (2010)
24. Silva, P.R.C., Dias, S.M., Brandão, W.C., Song, M.A.J., Zárate, L.E.: Formal concept analysis applied to professional social networks analysis. In: Proceedings of the 19th International Conference on Enterprise Information Systems, Volume 1, Porto, Portugal, April, 2017. pp. 123–134

25. Spink, A., Wolfram, D., Jansen, J., Saracevic, T.: Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology **52**, 226 – 234 (02 2001)
26. Yang, Z., Li, C., Fan, K., Huang, J.: Exploiting multi-sources query expansion in microblogging filtering. Neural Network World **27**, 59–76 (01 2017)
27. Zaki, M.J., Hsiao, C.: Efficient algorithms for mining closed itemsets and their lattice structure. IEEE Trans. Knowl. Data Eng. **17**(4), 462–478 (2005)
28. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the Tenth International Conference on Information and Knowledge Management. pp. 403–410. ACM, NY,USA (2001)
29. Zingla, M.A., Chiraz, L., Slimani, Y.: Short query expansion for microblog retrieval. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 20th International Conference KES-2016 **96**(C) (Oct 2016)