How to Stay Up-to-date on Twitter with General Keywords

Mandy Roick, Maximilian Jenders, and Ralf Krestel

Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany

Abstract. Microblogging platforms make it easy for users to share information through the publication of short personal messages. However, users are not only interested in sharing, but even more so in consuming information. As a result, they are confronted with new challenges when it comes to retrieving information on microblogging platforms. In this paper we present a query expansion method based on latent topics to support users interested in topical information. Similar to news aggregator sites, our approach identifies subtopics to a given query and provides the user with a quick overview of discussed topics within the microblogging platform. Using a document collection of microblog posts from Twitter, we compare the quality of search results returned by our algorithm with a baseline approach and a state-of-the-art microblog-specific query expansion method. We introduce a novel, innovative semi-supervised evaluation strategy based on expert Twitter users. In contrast to existing query expansion methods, our approach can be used to aggregate and visualize topical query results based on the calculated topic models, while achieving competitive results for traditional keyword-based search.

1 Searching Microblog Posts

Along with the development of Web 2.0, users have increasingly become content providers. A good example of this trend are microblogging platforms. These platforms allow users to share short text messages, images, or links with interested observers (followers) [5]. Microblogging platforms, such as Facebook, Tumblr, or Twitter, report constantly increasing numbers of users. According to Twitter's website, e.g., the platform has 284 million active users monthly and 500 million shared microblog posts daily, averaging 6,000 tweets per second. However, not all of Twitter's users share content. 44% of the users have never posted anything[1]. These users are only interested in consuming content, thus filtering and searching microblog posts becomes an increasingly important task.

In 2011, Twitter's search engine processed about 1.6 billion search queries daily. An analysis of the search behavior [10] shows that 49% of Twitter users search for timely information, such as trending topics or information related to news, 26% describe an interest in social information about other users, and 36%

[1] Digital Insights http://bit.ly/SMstats2014

report a search for specific topics, such as “astronomy”. Since then, searching microblog posts has become part of the research agenda. The Text REtrieval Conference[2] (TREC) opened a Microblog track in 2011 addressing a real-time search task on microblogging platforms. In 2014, Twitter expanded its search service to allow users to search for all tweets ever posted[3].

In contrast to Web search, searching microblogs displays some characteristic challenges [10]. To cope with the restricted length of tweets, Twitter users not only use abbreviations and emoticons, but also employ use hashtags, which are explicit user-specified topic markers. Another means to artificially condense information to fit in tweets is using a link to another web page with more information on the topic. Hence, many tweets contain URLs. However, these instruments are user-specified and their quality and usability for search depends on how users adopt these instruments. URLs for instance often link to images or videos, which are difficult to interpret for a machine. The given hashtags are very inconsistent through different spellings and different interpretations of users. For example, “#4YearsAgo5StrangersBecame5Brothers”, “#ThankYou1DYouChangedOurLives”, “#4YearsOf1D”, and “#4YearsDownAndForeverToCome” all refer to the four year anniversary of the band One Direction. For a user who does not follow this content on Twitter every day, it is difficult to pose queries that match the language used in tweets. The massive number of tweets every day constitutes an additional challenge to new users who are interested in an overview of the content on Twitter. To overcome the differences in the language used by users who post tweets and users who pose queries, we introduce a new query expansion approach, which allows topic-based searching. This improves the search experience for people searching topical and news-like information on Twitter using rather general keywords such as “politics” or “basketball”.

While many researchers propose query expansion algorithms for microblogging platforms [9], [11], [1], [4], none of them deal with the search for specific topics. Currently, Twitter presents search results in a list view showing the content of tweets and their authors, the time that has passed since the tweets were posted, and, if the tweets link to a news page, a short summary of the news page. The ranking is mainly based on exact query term matching, on recency, and on popularity. While query expansion can help to overcome the problems of exact query term matching, topical queries usually include many subtopics that a user might be interested in. Gaining an overview of these results is difficult using ranked lists. Given the fact that Twitter behaves similar to news media [6], we propose to use our results for query expansion to cluster tweets about similar topics. An application could display a user interface similar to platforms such as Google News[4], where individual news articles are aggregated and categorized.

[2] TREC `http://trec.nist.gov`
[3] Twitter `https://blog.twitter.com/2014/building-a-complete-tweet-index`
[4] Google News `http://news.google.com`

2 Related Work

There are many approaches that use topic models for query expansion in classic information retrieval [13], not so many for microblog posts. Yan et al. [12] present an alternative to LDA specially for short texts: the biterm topic model (BTM). Instead of generating documents, BTM models the generation of biterms (unordered word-pairs that co-occur in short texts) and assumes that each biterm is drawn from one topic. One work similar to ours describes the automatic topic-focused monitor (ATM) [7], which is able to monitor tweets relevant to a given topic. While the strength of ATM lies in the monitoring of tweets over time, our search approach selects keywords firsthand and does not need to know the search query in advance for correct sampling.

Several approaches for query expansion and document expansion have been proposed in the context of the Microblog Track at TREC. For example, Wang et al. [11] use a query expansion by accessing pseudo-relevance feedback and a document expansion through given URLs that some tweets contain. They use this expansion to break ties between tweets that display the same retrieval score, meaning that only tweets with the same retrieval score are considered. In that context, Wang et al. showed that the expansions did not support the ranking but lead to worse results. Bandyopadhyay et al. [1] aim to improve weak queries (e.g., short tweets with different spelling and grammar than a regular search query) and present a query expansion algorithm which is based on pseudo-relevant web documents. The algorithm transfers the original queries to the Google search API and expands the query with the most frequent terms in the resulting titles and snippets which are returned by the search API. Irrespective of the TREC Conference, Massoudi et al. [9] developed a retrieval model for queries that contain trending topics. They extend the model by taking quality indicators, like recency and followers, into account as well as a query expansion through co-occurrence of terms. An approach for document expansion has been described by Efron et al. [4] using a language model which includes a weighted probability for a word given the expanded document. An expansion is achieved by using the document as pseudo-query on the corpus of documents. Liang et al. [8] use pseudo-relevance feedback query expansion based on language models and employ temporal re-ranking to discover recent but relevant information for a query in microblogs. Topic models have been used by Chua et al. [3] to extract representative tweets from a stream for event summarization.

The presented approaches mostly aim to expand the given query to match the language which is used in the short microblog posts [1], [11] or to expand the microblog posts to match the language which is used in a query [11], [4]. In this paper, we concentrate on queries which are intentionally very general and we aim to expand those queries to provide a good overview of the trending subtopics at different levels of granularity.

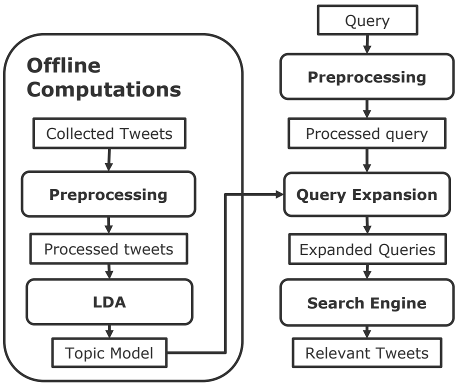

Fig. 1: System architecture

3 Topic-Based Query Expansion

We want to support users searching for general topics, such as “politics” or “Ukraine”. To this end, we propose a query expansion approach based on topic modelling. These models are learned on a daily basis from a data set of crawled and preprocessed tweets and are later used to expand user-specified queries. Figure 1 displays our system’s architecture. The crawling of tweets and topic model construction is handled offline, while the topic model is being used to expand queries in an online fashion at query time. If a new, unkown query term is used which is not present in our offline-computed topic model, we fall back to standard keyword search. However, this essentially does not happen for our targeted general queries. Furthermore, we address recency and popularity in Twitter indirectly via computing new topic models daily so our model reflects trends accordingly.

Topic Model Construction We used latent Dirichlet allocation (LDA) [2] to compute a topic model[5] prior to search, e.g. once a day. The resulting topic model can then be used to infer a topic distribution for a new tweet d, Θ_d. Given a query, the most probable topics can be determined using Φ, the topic-word-distributions. Table 1 shows the 10 most probable topics for a one-day topic model together with the probability of the topic given the query word “politics”.

Using LDA, the number of topics K has to be specified in advance. A larger K leads to splitting of topics, allowing for the separation of ambiguous topics. However, if no ambiguous topics are left, homogeneous ones are split up. For the purpose of query expansion, it is important that different topics can be found for a term, and that the topics found are not ambiguous, as this could lead to topic shifts. We evaluated different values for K on a validation set, which is described in Section 4.

[5] We use Mallet `http://mallet.cs.umass.edu`

Table 1: Top 10 topics from October 20, 2014 for the query *"politics"*

$\hat{p}(i\vert\boldsymbol{q})$	**8 most probable stemmed words**
0.167	obama ebola tcot speech presid reason net ban
0.155	ukip vote ward tori parti peopl nh elect
0.115	bjp part scienc india modi biblic read congress
0.089	vote elect voter earli blue texa gop todai
0.072	gate gamer gamerg peopl women stop game bulli
0.069	presid indonesia jokowi minist presiden japan russia
0.052	isi turkei kurd koban fight kill iran syria
0.044	ari support pakistan ban stand pti khan wesupportari
0.043	energi price compani tax loan pai servic power
0.036	class question teacher answer write english word learn

Query Expansion We are interested in the most probable topics for all words of a query $\boldsymbol{q}$, i.e., we search for topics i where $p(z = i|\boldsymbol{q})$ (in the following $p(i|\boldsymbol{q})$) is maximal. During Gibbs sampling, we sample values for $\boldsymbol{z}$ for each word w in the vocabulary $\boldsymbol{w}$.

We use these samples of $\boldsymbol{z}$ to estimate $\hat{p}(i|\boldsymbol{q})$ with $\frac{n(i,w)}{n(w)}$. In other words, $\hat{p}(i|\boldsymbol{q})$ is estimated by the number of times the query words $\boldsymbol{q}$ were assigned to topic i divided by the total number of occurrences of words $\boldsymbol{q}$ in the corpus. Note that, although our test queries only contain a single term, this formulation also holds for queries with multiple words. For the query expansion, we then use the topics' best representatives, i.e., for a topic i the most probable words based on $p(w|i) = \phi_i^w$.

The quality of the query expansion is heavily influenced by the number of topics the query is expanded with, as well as the number of words chosen from each topic for expansion. We optimized these model parameters on a validation set (see Section 4). Best results were achieved setting K, the number of topics, to 200; the number of terms to use for query expansion to 10, and the threshold to include a topic for an expansion to $\hat{p}(i|\boldsymbol{q}) > 0.05$. For our example in Table 1 the top 7 topics would be used for query expansion, while the rest are disregarded.

4 Experiments

To assess the ability of our algorithm to retrieve topically relevant tweets, we propose a novel, semi-automatic evaluation strategy that produces high-quality labeled data by utilizing expert Twitter users. In addition, we present some example queries together with the expanded queries based on our topic model as anecdotal evidence demonstrating how our algorithm can help users to get a topical overview of subtopics for a given general query.

Data Set Most existing annotated data sets are focused on detailed information needs, such as the Tweets2011 corpus used for the TREC Microblog Track[6]. Gen-

[6] TREC microblog data http://trec.nist.gov/data/tweets

eral topical queries are not included. Therefore, we created our own data set with semi-automatic annotations. We chose 2 general topical queries:“sports”, “politics” and for each general query 2 more specific ones: “baseball”, “basketball”, “Ebola”, “Ukraine”. To find relevant tweets for each of the queries, we hand-picked 10 expert twitter users who primarily tweet on the topic corresponding to the query. Together with the relevance of tweets we used popularity and the number of tweets to select these users. For “politics”, e.g., these users were: @BBCPolitics, @CNNPolitics, @NicRobertsonCNN, @KevinBohnCNN, @TheWhiteHouse, @politico, @thehill, @HuffPostPol, @CBSPolitics, @BarackObama. We then crawled these users’ tweets together with the 1% of general tweets available through the Twitter API. We annotated only our expert users’ tweets as relevant for the respective queries, leading to small values in precision, because some tweets marked as non-relevant are actually relevant. Yet, tweets marked as relevant are in large part actually relevant. Thus, we estimate a method’s tendency for the actual precisions. We constructed two data sets, one for validation and one for testing. Each set includes a training set of one day of twitter data to learn the topic model and the subsequent day to validate or test (Oct. 21st and Dec. 4th 2014, each 1.4m tweets (1% of all tweets)). On average, our expert users published 196 tweets per query per day.

Baseline Approach As baseline approach BL, we search for the given queries without query expansion. Similar to Twitter’s search engine, we search for the query terms in tweets as well as in linked content using BM25. In contrast to Twitter’s search, our ranking is not incorporating recency or popularity.

Next to the baseline approach, we compare our search results with a competing query expansion algorithm that is designed for microblogging platforms and based on word co-occurance. It was proposed by Massoudi et al. [9] and shows improved search results against a standard query expansion with pseudo relevance feedback.

Topic-Based Approach Our topic-based approach results in a set of expanded queries for each initial query according to our topic model. We set α asymmetric and choose the initial value $\alpha_i = K \cdot 0.01$ for all $i \in \{1, 2, \dots, K\}$. In contrast to α, we set β symmetric with initial value $\beta_i = 0.05$. We run Gibbs sampling for 500 iterations.

Each topic i in our model that contains the query term q (i.e. $\hat{p}(i|q) > 0.05$) forms the basis for one query. To compare our search results with other search algorithms and the baseline, we merge the tweets resulting from each expanded query q into one ranking. We calculate a ranking score $sc_q(i, d)$ for each tweet d that was found for a query q. The score depends on the topic (=expanded query) i for which the tweet was found and the tweet d itself. The score combines the probability $\hat{p}(i|q)$ of the query term q belonging to the topic i, the topic’s proportion θ_d^i for tweet d, and the BM25 score for the tweet $BM25(d)$:$sc_q(i, d) = \hat{p}(i|q) \cdot \theta_d^i + BM25(d)$ This score allows to combine the results of all expanded queries for a query term into one ranking, which is needed to compare the precision with other approaches.

Table 2: Average precision for various algorithms for particular queries

	sports	**baseball**	**basketball**	**politics**	**Ukraine**	**Ebola**
BL	0.0035	0.0033	0.0038	0.0000	0.2730	**0.3232**
CB	0.0057	0.2578	0.0106	0.0158	**0.4595**	0.1617
TB	**0.0150**	**0.3175**	**0.0158**	**0.0166**	0.3068	0.2403

Table 3: Example expanded queries for topic-based approach (TB) and co-occurance-based approach (CB) [9] for queries "sports" and "Ebola"

sports				**Ebola**		
CB	**TB**			**CB**	**TB**	
girls	sports	sports	sports	outbreak	ebola	ebola
cespedes	united	hurt	game	americans	dallas	nigeria
boston	goals	head	football	free	health	free
football	game	butt	week	officially	save	big
betting	score	error	win	declared	hospital	plan
sports	mufc	vixx	state	virus	nurse	reason
pretend	west	body	nba	nigeria	patient	declared
smh	liverpool	button	team	health	care	immediate
females	man	touch	play	ebola	disease	someone's
yahoo	manchester	work	season	obama	africa	capricorn

Results The results differ from query to query. Mean average precision (MAP) is 0.101 for the baseline approach (BL), 0.152 for the co-occurance-based approach (CB), and 0.152 for the topic-based approach[7]. The co-occurrence-based query expansion and our topic-based approach improve the results decidedly over the baseline. CB outperforms the topic-based approach only for the query "Ukraine", which results in similar MAP scores, see Table 2. Less general queries, such as Ebola, are less likely to benefit from query expansion since most tweets contain the keyword itself, whereas tweets about baseball are much more likely to contain words such as "MLB" instead of the word "baseball".

The expanded queries give an overview of the topic. The co-occurance-based approach only produces one expanded query, whereas our topic-based approach finds multiple topics for a given keyword and thus can create multiple expanded queries representing subtopics. Table 3 shows how our approach identifies different subtopics related to sports: English soccer, injuries, and American sports, while the co-occurance based approach fails to give a good overview and mixes various sports-related terms. The results are similar for the query Ebola. Here our approach identifies a topic related to Ebola in the U.S. vs. Africa.

Discussion The co-occurance-based expansion is calculated specifically for each query, therefore it benefits from the expansion terms being well suited. Yet, especially for the more general queries, the expanded queries can become ambiguous,

[7] To create comparable MAP scores, each ranking is restricted to 500 tweets

i.e., contain more than one specific topic with considerable topic shifts. In contrast to the co-occurance approach, our topic-based approach discovers more relevant terms for a given query. Thus, the focus of the search can transform to a broader topic than the original one. A strength of our topic-based approach is also the flexibility allowing to expand the query with a variable number of topics and visualize the inherent subtopics.

5 Conclusion

We have analyzed the usage of topic models to support general keyword queries in microblog search. We proposed a query expansion method using latent Dirichlet allocation to find relevant tweets and to group them based on latent topic information. Our experiments have shown that our approach outperforms standard keyword-based search and further demonstrated competitive results compared to a state-of-the-art microblog-specific query expansion algorithm. While standard search algorithms do not by default cluster search results, our approach returns tweets from various subtopics and the topics itself can be inspected to get a quick overview of what is currently discussed in Twitter related to general keywords. Besides a further, large-scale evaluation, for future work we are interested in the development of topics over time. Since Twitter is a highly dynamic platform, we hope to capture trending subtopics for general keywords by substituting LDA with a dynamic topic model.

References

1. Bandyopadhyay, A., Ghosh, K., Majumder, P., Mitra, M.: Query expansion for microblog retrieval. In: TREC. vol. 1, pp. 368–380. NIST (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
3. Chua, F.C.T., Asur, S.: Automatic summarization of events from social media. In: ICWSM. pp. 81–90. AAAI (2013)
4. Efron, M., Organisciak, P., Fenlon, K.: Improving retrieval of short texts through document expansion. In: SIGIR. pp. 911–920. ACM (2012)
5. Kaplan, A.M., Haenlein, M.: The early bird catches the news: Nine things you should know about micro-blogging. Business Horizons 54(2), 105–113 (2011)
6. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: WWW. pp. 591–600. ACM (2010)
7. Li, R., Wang, S., Chang, K.C.C.: Towards social data platform: Automatic topic-focused monitor for twitter stream. In: VLDB. pp. 1966–1977. VLDB Endowment (2013)
8. Liang, F., Qiang, R., Yang, J.: Exploiting real-time information retrieval in the microblogosphere. pp. 267–276. JCDL, ACM (2012)
9. Massoudi, K., Tsagkias, M., de Rijke, M., Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts. In: ECIR, pp. 362–367. Springer (2011)
10. Teevan, J., Ramage, D., Morris, M.R.: #twittersearch: a comparison of microblog search and web search. In: WSDM. pp. 35–44. ACM (2011)

11. Wang, Y., Darko, J., Fang, H.: Tie-breaker: A new perspective of ranking and evaluation for microblog retrieval. In: TREC. NIST (2013)
12. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: WWW. pp. 1445–1456. ACM (2013)
13. Yi, X., Allan, J.: A comparative study of utilizing topic models for information retrieval. In: ECIR. pp. 29–41. Springer (2009)