The Importance of Being Socially-Savvy: Quantifying the Influence of Social Networks on Microblog Retrieval

Alexander Kotov
Emory University
kotov@mathcs.emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

ABSTRACT

Social media users create virtual connections for various reasons: personal and professional. While significant research efforts have been spent on exploring the dynamics of creation of social network connections, little is known about how those connections influence the content generated by social media users. In this work, we quantitatively evaluate the influence of social networks on social media content providers. Additionally, we propose several document expansion methods, which leverage the content generated by the social networks of the authors of social media documents and compare their effectiveness. Experimental results on a large sample of Twitter data indicate that retrieval models discriminatively leveraging social network content for document expansion outperform both traditional, socially-unaware retrieval models and retrieval models that indiscriminatively utilize all social connections.

Categories and subject descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *Retrieval models*

Keywords

Microblog Retrieval, Social Networks, Document Expansion

1. INTRODUCTION

Retrieval from large collections of short social media documents (e.g. microblogs) is a very challenging task. Microblog posts are often shorter than queries, which exacerbates the problems of vocabulary and semantic mismatch between the queries and documents compared to retrieval from traditional document collections. On the other hand, social media provides different types of metadata, which can be leveraged to improve the accuracy of retrieval results. In this work, we explore how social metadata in the form of the content generated by the social networks of the authors of social media documents can be leveraged for document expansion in the context of Twitter, a popular microblogging platform.

Studying the influence of social networks on content generation from qualitative perspective has recently attracted attention of many social media researchers. Most notably, boyd et al. [2] identified different ways, in which the content generated by Twitter users can be transformed through re-tweets. In this work, however, we are focusing on quantitative and information retrieval aspects of this problem and experimentally verify two assumptions. The first assumption is that the posts generated by social media users (in our case, microbloggers) are to a large extent influenced by the content produced by their social networks or, in the context of Twitter, by the users who the microbloggers follow (i.e. followees). The second assumption is that the documents authored by followees can be effectively used for expansion of the language models (LMs) of microblog posts, ultimately translating into better retrieval results. Introducing additional terms that are semantically related to the LMs of potentially relevant, but short microblog posts, which may not contain some or all of the query terms, can improve their position in the ranked list of retrieval results.

First, we propose to characterize the influence of social networks on social media users (microbloggers) in terms of different similarity measures between all the documents (posts) generated by the microbloggers and their followees. Then we propose and experiment with several microblog post expansion strategies that discriminatively and indiscriminatively utilize textual content from social networks. In particular, we propose two types of discriminative expansion strategies: high-level and low-level. Low-level strategies are based on using some finite number of posts from the social network that are most similar to a given post to form its expansion LM. High-level strategies are based on using some finite number of LMs of the followees that are most similar to the LM of the author of a given post. Indiscriminative strategies utilize the entire social context for social media document expansion. More specifically, after a brief overview of relevant work in Section 2, we provide answers to the following research questions:

– To what extent do social networks influence the content produced by microbloggers? (Section 4.2)

– How to leverage social networks for microblog posts expansion: selectively use LMs of the most similar posts authored by the followees (or LMs of the followees themselves) or all posts authored by the followees? (Section 3)

– What is the optimal number and the best criteria for selecting the LMs of social network posts (followees) for expansion of microblog posts? How to optimally combine the original microblog post LM and expansion LM? (Section 4.3)

CIKM'13, Oct. 27–Nov. 1, 2013, San Francisco, CA, USA.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2263-8/13/10 ...$15.00.
http://dx.doi.org/10.1145/2505515.2507892.

2. RELEVANT WORK

It is generally believed that microblog and traditional information retrieval are fundamentally different [10]. While query [6] and document [9] expansion have both been successfully employed to address the vocabulary mismatch problem in traditional document collections, to the best of our knowledge, there has been no prior work examining the utility of social network content for social media document expansion. Tao et al. [9] demonstrated that semantically relevant documents from the same collection selected using a simple similarity metric can be used for traditional document expansion and that such expansion is particularly effective for shorter documents. In this work, we compliment their study by examining the effectiveness of document expansion at different levels of granularity and by taking into account the inherent social network structure of the collections of microblog posts.

Other types of metadata provided by social media have also previously been shown to improve retrieval. Kotov et al. [5] proposed a method leveraging geographical metadata in combination with topic modeling and Efron et al. [3] proposed a pseudo-feedback method for short document expansion based on temporal metadata. While socially-aware topic modeling has been explored in itself in previous studies (e.g. [8]), in this work, we employ topic modeling as dimensionality reduction technique to more robustly determine semantically similar short documents (followees).

3. METHOD

In this work, we follow the language modeling approach to retrieval and use the KL-divergence based ranking function [11], which involves estimating the query LM, Θ_Q, for a given keyword-based query $Q = \{q_1, q_2, \ldots, q_N\}$ and the document LM, Θ_{D_i}, for each document D_i in the collection $\mathcal{C} = \{D_1, \ldots, D_M\}$. The documents in the collection are scored and ranked according to the Kullback-Leibler (KL) divergence (further referred to as **KL-DIR**):

$$Rank(Q, D_i) \sim \mathrm{KL}(\Theta_Q || \Theta_{D_i}) = \sum_{w \in V} p(w|\Theta_Q) \log \frac{p(w|\Theta_Q)}{p(w|\Theta_{D_i})} \tag{1}$$

where Θ_{D_i} is normally smoothed, for example, using the Dirichlet prior smoothing:

$$p(w|\Theta_{D_i}) = \frac{|d|}{|d| + \mu} p_{ML}(w|\Theta_{D_i}) + \frac{\mu}{|d| + \mu} p(w|\mathcal{C}) \tag{2}$$

where $p_{ML}(w|\Theta_{D_i})$ and $p(w|\mathcal{C})$ are the probabilities of w in the maximum likelihood estimate of document LM and collection LM, respectively, and μ is the Dirichlet prior.

In a language modeling based retrieval framework, document expansion is typically performed by first deriving a document expansion LM, $\hat{\Theta}_{D_i}$, for each document $D_i \in \mathcal{C}$ and then updating the original document LM, Θ_{D_i}, through linear interpolation with coefficient α:

$$p(w|\tilde{\Theta}_{D_i}) = \alpha p(w|\Theta_{D_i}) + (1 - \alpha) p(w|\hat{\Theta}_{D_i}) \tag{3}$$

Our approach is based on deriving $\hat{\Theta}_{D_i}$ from social metadata provided by microblogging services. In particular, the followees of the author of D_i and the posts that the followees have produced. We propose two conceptually different approaches to deriving $\hat{\Theta}_{D_i}$:

– A microblog post D, authored by a user U can be inspired by another post D' (or a set of posts) authored by some user U' in the social network $\mathcal{S} = \{U_1, \ldots, U_{|\mathcal{S}|}\}$ of U. In other words, D can be viewed as a sample from $\Theta_{D'}$, a language model of D'. Following this view, a document expansion LM, $\hat{\Theta}_{D_i}$, can be derived by aggregating one or several LMs of the documents authored by the followees of U;

– Alternatively, a microblog post D, authored by a user U can be viewed as a sample from $\Theta_{U'}$, a language model of some user (or a set of users) $U' \in \mathcal{S}$.

Therefore, both approaches to deriving $\hat{\Theta}_{D_i}$ are based on ranking all documents (users) according to $sim(D, D')$, a measure of similarity between D and D' or $sim(U, U')$, a measure of similarity between U and U'. Since microblog posts are very short, content based document similarity measures (e.g. cosine similarity explored in [9]) may not be a good choice for $sim(D, D')$. Therefore, we utilize topic modeling and determine document similarity based on symmetrized Kullback-Leibler divergence, $\mathrm{SKL}(\Psi_D || \Psi_{D'})$, between topical distributions Ψ_D and $\Psi_{D'}$ for D and D' respectively:

$$\mathrm{SKL}(\Psi_D, \Psi_{D'}) = \frac{1}{2}(\mathrm{KL}(\Psi_D || \Psi_{D'}) + \mathrm{KL}(\Psi_{D'} || \Psi_D)) \tag{4}$$

where:

$$\mathrm{KL}(\Psi_D || \Psi_{D'}) = \sum_{t \in T} p(t|D) \log \frac{p(t|D)}{p(t|D')} \tag{5}$$

is the KL-divergence between topical distributions over T topics for D and D'. This approach is further referred to as **DTOP**.

Since aggregated user documents are larger, there is a broader choice for $sim(U, U')$. In addition to utilizing topical distributions for each user document:

$$sim(U, U') = \mathrm{SKL}(\Psi_{D_U} || \Psi_{D_{U'}}) \tag{6}$$

where D_U is a user document obtained by aggregating all microblog posts authored by U. In addition to calculating symmetrized KL divergence for topical distributions of user documents, we can also apply it directly to user LMs Θ_U and $\Theta_{U'}$:

$$sim(U, U') = \mathrm{SKL}(\Theta_U || \Theta_{U'}) \tag{7}$$

If we only consider W_{D_U} and $W_{D_{U'}}$, sets of unique terms in D_U and $D_{U'}$, then we can calculate similarity between the users as Jaccard coefficient (i.e., normalized term overlap) between W_{D_U} and $W_{D_{U'}}$:

$$sim(U, U') = \frac{|W_{D_U} \cap W_{D_{U'}}|}{|W_{D_U} \cup W_{D_{U'}}|} \tag{8}$$

In summary, in this work we experimented with the following conceptually different approaches for deriving $\hat{\Theta}_{D_i}$:

ALL: create LMs for all followees of the document author by aggregating all their microblog posts and use LMs of all followees for expansion of the original document;

ALL-RT: create a LM for each re-tweeted followee of the document author by aggregating all her microblog posts and use only the LMs of re-tweeted followees weighted by the number of re-tweets for expansion;

RANK-DIV and **RANK-JAC**: create and rank LMs for all followees based on two different similarity functions: symmetrized KL divergence (**RANK-DIV**) and Jaccard coefficient (**RANK-JAC**) and use LMs of the top-ranked followees for expansion;

RANK-UTOP: run a topic model on a collection of user documents and rank all followees according to the similarity of their topical distribution to the topical distribution of the author of the microblog post being expanded;

RANK-DTOP: run a topic model on a collection of microblog posts and rank all the posts authored by followees according to the similarity of their topical distribution to the topical distribution of the post being expanded.

4. EXPERIMENTS

4.1 Datasets

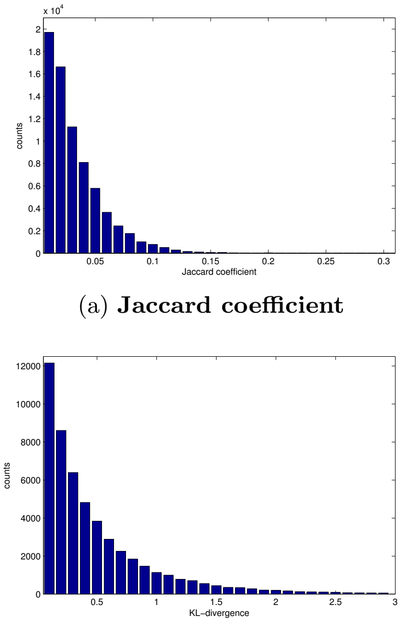
We used 2011 TREC Microblog track [7] corpus as the base dataset for all experiments in this work. Since the original TREC corpus does not include any information about the followees of the authors of the tweets, we randomly sampled 10,000 users from all the authors in TREC corpus and accessed their Twitter profiles to extract their followees and for each followee we crawled 20 most recent microblog posts. After sampling about 5% of all relevant posts in TREC corpus had social context. The extended TREC corpus (**EXT-TREC**) has been post-processed to exclude all user mentions, URLs and remove #-signs from all hashtags. Since collections of microblog posts have been known for having very large vocabularies, which negatively affect the quality of topics, we created two additional corpora with restricted vocabularies based on the post-processed extended TREC corpus. The first corpus (**TMOD-DOC**) was obtained from **EXT-TREC** by excluding all non-English terms (terms that are not in the dictionary of the *aspell* spell-checking program) but retaining all capitalized non-stopwords. This corpus was used to obtain the topical distribution for each microblog post through LDA (Latent Dirichlet Allocation) [1], a popular topic model. The second corpus (**TMOD-USER**) was obtained by aggregating all microblog posts of each user into a single document and used to obtain the topical distribution for each user, also through LDA. Such pre-processing scheme for collections of small documents has been demonstrated [4] to result in topics of better quality, however it was not known whether this would translate into better retrieval results. From both **TMOD-USER** and **TMOD-DOC** we excluded all rare terms (that occurred in less than in 20 documents) and all documents that are too short (had less than 5 terms). Various statistics of experimental datasets used in this work are presented in Table 1.

Corpus	# docs	avg. dlen.	# voc.size
TMOD-USER	3,569,652	54	140,327
TMOD-DOC	19,835,058	7	139,307
TREC	8,790,298	9	2,029,326
EXT-TREC	39,405,811	8	4,544,912

Table 1: Statistics of experimental datasets.

4.2 Quantifying influence

In order to estimate the degree of influence of social networks on social media content generation, for each user in our dataset we first calculated the user LM, Θ_U, by aggregating all the documents authored by U and the social LM, $\Theta_{\mathcal{S}}$, by aggregating all the documents authored by all followees of U. We quantified the degree of influence of social networks using two similarity metrics: KL-divergence and Jaccard coefficient between the user and social LMs and plotted the distribution of values across the entire set of users in Figure 1. As follows from Figure 1, the degree of simple term overlap (without taking into account term probabilities) between the user and social LMs, as measured by Jaccard coefficient, is fairly substantial (5%-10%), while KL-divergence is small. This can be interpreted as the fact that the probabilities of overlapping terms are similar, which indicates that there exist a set of core terms shared by microbloggers and their followees and that there is a significant degree of social network influence on social media content providers.



(a) **Jaccard coefficient**

(b) **KL-divergence**

Figure 1: Distribution of values for Jaccard coefficient and KL-divergence between the user LM and social LM for all users in the dataset.

4.3 Parameter optimization

In the first model parameter optimization experiment we set α to 0.7 and varied the size of the expansion context, measuring P@20 (precision at top 20 retrieved documents averaged over the entire TREC query set) for each context size. P@20 is a more natural performance metric for microblog retrieval tasks, since there is generally very few potentially relevant tweets for each query and users don't want to examine long list of retrieval results. The results of this experiment are shown in Figure 2. Since our goal was to focus on examining the effect of social metadata on the quality of retrieval results, we filtered all the documents for which we did not have a social network context from the retrieved results before measuring the retrieval performance, which explains low values for all the methods. The first interesting observation based on Figure 2 is that user LM based document expansion significantly outperforms document LM based expansion. Secondly, all user based expansion methods follow the same pattern and achieve optimal performance when top 6 followees are used for expansion. Using only a few most similar users or using too many users degrades the performance. This has an intuitive explanation, since each additional user may cover the miss-

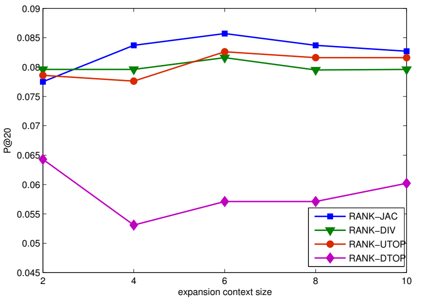


Figure 2: P@20 of different similarity-based document expansion methods by varying the size of the expansion context (number of LMs of top most similar users for RANK-JAC, RANK-DIV, RANK-UTOP and number of top most similar documents for RANK-DTOP)

ing query aspect of a short microblog post, however, the marginal effect of adding more users decreases, and eventually using too many user LMs may cause topic drift. At the same time, content based method for selecting similar users based on normalized set overlap measured by Jaccard coefficient (**RANK-JAC**) is more effective than both topic model based method (**RANK-UTOP**) and content based method using symmetrized KL divergence (**RANK-DIV**). This can be explained by the fact that the vast majority of terms occur only once [3] per microblog post, therefore exact probabilities of terms in LMs are much less important than in retrieval from traditional document collections.

In the next experiment, we used the optimal expansion context size for each method and varied the value of α, the interpolation coefficient between the original and expansion LMs. The results of this experiment are shown in Figure 3. As follows from Figure 3, the best performance is again achieved by **RANK-JAC**. However, when the interpolation coefficient is optimized, the difference between user and document LM based expansion methods decreases. Note that the performance of all expansion methods when $\alpha = 1.0$ corresponds to the baseline performance of **KL-DIR**. As follows from Figure 3, all social-network based expansion methods significantly improve the performance of standard socially-unaware retrieval model.

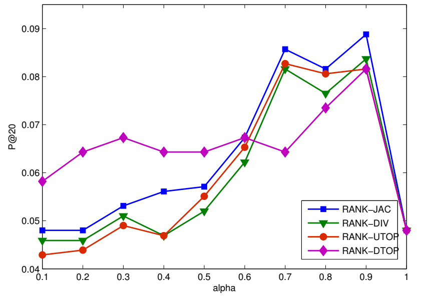


Figure 3: P@20 of different similarity-based document expansion methods by varying the interpolation coefficient α

4.4 Best performance summary

Method	**P@20**	**Δ-BL %**
KL-DIR	0.0480	0.00
ALL	0.0857	+178.54
ALL-RT	0.0847	+176.46
RANK-DTOP	0.0816	+170.00
RANK-JAC	**0.0888**	**+185.00**
RANK-DIV	0.0837	+174.38
RANK-UTOP	0.0816	+170.00

Table 2: Comparison of the best performance of socially-unaware retrieval model (KL-DIR) with indiscriminative (ALL and ALL-RT), discriminative (RANK-JAC, RANK-DIV and RANK-UTOP) user LM based social expansion and discriminative document LM based social expansion (RANK-DTOP).

As follows from Table 2, leveraging social metadata through document expansion translates into substantial improvement in retrieval quality over standard socially-unaware retrieval models.

5. CONCLUSIONS

This paper quantitatively characterizes the influence of social networks on the content generated by microbloggers and experimentally demonstrates that it is fairly strong. It also presents experimental results indicating that leveraging social context for microblog posts expansion translates into significant improvements in retrieval accuracy and that selecting a small number of most similar followees based on simple term overlap-based metric is the most effective strategy for social network based expansion of microblog posts.

Acknowledgments

This work was supported by the DARPA grant D11AP00269.

6. REFERENCES

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

[2] danah boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twiiter. In *Proceedings of IEEE HICSS'10*, 2010.

[3] M. Efron, P. Organisciak, and K. Fenlon. Improving retrieval of short texts through document expansion. In *Proceedings of SIGIR'12*, pages 911–920, 2012.

[4] L. Hong and B. D. Davison. Empirical study of topic modeling in twitter. In *Proceedings of SOMA'10*, pages 80–88, 2010.

[5] A. Kotov, Y. Wang, and E. Agichtein. Leveraging geographical metadata to improve search over social media. In *Proceedings of WWW'13*, pages 151–152, 2013.

[6] A. Kotov and C. Zhai. Tapping into knowledge base for concept feedback: Leveraging conceptnet to improve search results for difficult queries. In *Proceedings of WSDM'12*, pages 403–412, 2012.

[7] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff. Overview of the trec-2011 microblog track. In *Proceedings of TREC 2011*, 2011.

[8] M. Schaal, J. O'Donovan, and B. Smith. An analysis of topical proximity in the twitter social graph. In *Proceedings of SocInfo'12*, pages 232–245, 2012.

[9] T. Tao, X. Wang, Q. Mei, and C. Zhai. Language model information retrieval with document expansion. In *Proceedings of NAACL-HLT'06*, pages 407–414, 2006.

[10] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: A comparison of microblog search and web search. In *Proceedings of ACM WSDM'11*, pages 35–44, 2011.

[11] C. Zhai and J. Lafferty. Document language models, query models, and risk minimization for information retrieval. In *Proceedings of SIGIR'01*, pages 111–119, 2001.