Leveraging Geographical Metadata to Improve Search over Social Media

Alexander Kotov
Emory University
kotov@mathcs.emory.edu

Yu Wang
Emory University
yu.wang@emory.edu

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

ABSTRACT

We propose the methods for document, query and relevance model expansion that leverage geographical metadata provided by social media. In particular, we propose a geographically-aware extension of the LDA topic model and utilize the resulting topics and language models in our expansion methods. The proposed approach has been experimentally evaluated over a large sample of Twitter, demonstrating significant improvements in search accuracy over traditional (geographically-unaware) retrieval models.

Categories and subject descriptors

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

Keywords

Social Media, Microblog Retrieval, Probabilistic Retrieval Models, Language Models, Topic Models

1. INTRODUCTION

Social media provides valuable metadata, such as geographical locations of documents and document authors. Additionally, social media documents (e.g., microblog posts) are becoming increasingly shorter, and thus less likely to contain all the terms as well as topical and geographical aspects of the queries. Therefore, one of the primary goals of microblog retrieval is to bridge the implicit geographical and topical foci of queries and short social media documents. The main idea behind this work is that the missing geo-topical aspects of both the information need and potentially relevant microblog documents can be inferred using *geographically-aware topic models*. While several topic models that take into account geographical lexical variation have been recently proposed [2] [3], there has been no prior work studying application of such models to *microblog retrieval*.

We propose a language modeling-based probabilistic retrieval framework, which leverages geographical metadata to perform geographically-focused document, query and relevance model expansion. While previous studies [5] have shown that topic modeling can improve retrieval over traditional TREC collections, we demonstrate that incorporating geographically-specific topics and language models (LMs) derived from them into retrieval methods can effectively address the issues of geographical query ambiguity and short social media documents.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

2. METHOD

Our geographically-aware retrieval framework is based on the KL-divergence retrieval model with Dirichlet prior smoothing [7] (further referred to as **KL-DIR**) and uses geographically-aware topic model (further referred to as **GLDA**) to leverage geographical metadata. Unlike regular topics, which are multinomial distributions over a vocabulary of terms (i.e. $p(w|t)$), probabilities of terms in geographically-specific topics are also affected by geographical location (i.e. $p(w|t,l)$). Since only 1%∼2% of all microblog posts have explicit geographical coordinates [3], all microblog posts in our experimental collection have been labeled with the location of their author taken from the author's profile. We obtain geographically-specific topics by splitting the original collection into sub-collections of microblog posts labeled with each of the considered geographical locations and running LDA [1] with the same number of topics on each sub-collection.

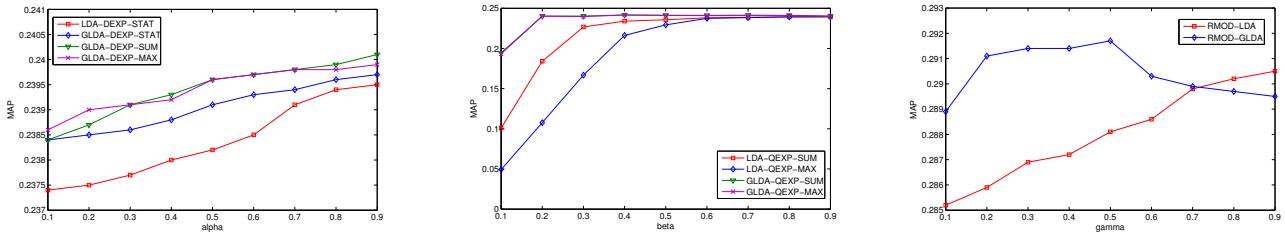
Specifically, we experimentally compare three ways of leveraging geographically-specific topics for microblog retrieval:

1. Document expansion: document expansion LM, $p(w|\hat{\Theta}_D)$, for each document in the collection is interpolated with the original document LM, $p(w|\Theta_D)$, using coefficient α. We propose several different ways of computing $p(w|\hat{\Theta}_D)$:

GLDA-DEXP-STAT: $p(w|\hat{\Theta}_D) = \sum_t p(w|t,l)p(t|D)$, where $p(t|D)$ is the distribution of topics $p(w|t,l)$ for a microblog post D that are specific to l (location of the author of D in her microblog profile), as determined by **GLDA**. The intuition behind this method is that $p(w|\hat{\Theta}_D)$ should include the terms from the most prevalent geographically-specific topics of D. As a baseline, we use the LDA-based document LM expansion method (**LDA-DEXP-STAT**) proposed in [5];

GLDA-DEXP-SUM: $p(w|\hat{\Theta}_D) = \sum_{l'} p(w|l')p(l'|D) = $
$= \sum_{l'} p(w|l') \sum_{w \in D} p(l'|w)p(w|D)$, where $p(w|l') = \sum_t p(w|t,l')$ is a location LM (which can be interpreted as vocabulary preferences of users in a given geographical location) and $p(l'|w) = \frac{p(w|l')p(l')}{p(w)}$ is a specificity of word w to location l'. The intuition behind this method is that instead of using the topics corresponding to l (location of microblog post D taken from its author's profile), we can leverage the terms in LM of D that are characteristic of certain locations to calculate $p(l'|D)$, the distribution of *predicted geographical foci* of D, and use the LMs for all locations in $p(l'|D)$ to form $p(w|\hat{\Theta}_D)$;

GLDA-DEXP-MAX: $p(w|\hat{\Theta}_D) = p(w|l')\,\text{argmax}_{l'}\, p(l'|D) = $
$= \; p(w|l')\,\text{argmax}_{l'} \sum_{w \in D} p(l'|w)p(w|D)$. Unlike **GLDA-DEXP-SUM**, this method uses only the LM of *the most likely predicted geographical focus* of D to form $p(w|\hat{\Theta}_D)$.



(a) **Document expansion** (b) **Query expansion** (c) **Relevance model expansion**

Figure 1: Mean Average Precision (MAP) of (a) document expansion, (b) query expansion and (c) relevance model expansion methods, by varying the interpolation coefficients α, β and γ.

2. Query expansion: query expansion LM, $p(w|\hat{\Theta}_Q)$, is interpolated with the original query LM, $p(w|\Theta_Q)$, using coefficient β. We propose several different ways of computing $p(w|\hat{\Theta}_Q)$:
GLDA-QEXP-SUM: $p(w|\hat{\Theta}_Q) = \sum_{l'} p(w|l')p(l'|Q) =$
$= \sum_{l'} p(w|l') \sum_{w \in Q} p(l'|w)p(w|Q)$, where $p(w|Q)$ is the probability of w in the maximum likelihood LM of the original query Q. This method utilizes the location-specific query terms to calculate $p(l'|Q)$, the distribution of *predicted geographical foci* of Q;
GLDA-QEXP-MAX: $p(w|\hat{\Theta}_Q) = p(w|l')\,\text{argmax}_{l'}\, p(l'|Q) =$
$= p(w|l')\,\text{argmax}_{l'} \sum_{w \in Q} p(l'|w)p(w|Q)$. Similar to **GLDA-DEXP-MAX**, this method uses only the LM of *the most likely predicted geographical focus* of Q to form $p(w|\hat{\Theta}_Q)$;
LDA-QEXP-SUM and **LDA-QEXP-MAX**: these methods are different from **GLDA-QEXP-SUM** and **GLDA-QEXP-MAX** in that they use LDA topics $p(w|t)$, instead of geographically-specific topics $p(w|t,l)$.

3. Pseudo-relevance feedback: relevance model (further referred to as **RMOD**) [4] expansion LM, $p_{RM}(w|\hat{\Theta}_D)$, is formed by interpolating, $p_{RM}(w|\Theta_D)$, the original document relevance model, and $p_{GLDA}(w|D,Q,l)$ with coefficient γ (**RMOD-GLDA**): $p_{RM}(w|\hat{\Theta}_D) =$
$= \sum_{D_i \in R}(\gamma p_{RM}(w|\Theta_{D_i}) + (1-\gamma)p_{GLDA}(w|D_i,Q,l)) \times p(D_i|Q)$, where $p_{GLDA}(w|D_i,Q,l) = \sum_t p(w|t,l)p(t|D_i)p(Q|t,l)$ and $p(Q|t,l) = \prod_{w \in Q} p(w|t,l)$. As a baseline, we use the method to incorporate LDA into relevance model (**RMOD-LDA**), as proposed in [6].

3. EXPERIMENTS

We used the 2011 TREC Microblog track corpus as the base collection for all experiments in this work. For the purpose of topic modeling, we considered a set of 83 geographical locations (union of 50 largest and 50 state capital cities in the United States) and filtered out from the base collection all the tweets, which did not belong to this set, and to compensate for that added new tweets from the considered locations (however, the original TREC collection was still used for all retrieval experiments). We experimentally determined that the optimal performance of **KL-DIR** in terms of MAP is achieved when the Dirichlet prior is set to 200 and that the lowest perplexity for **LDA** is achieved with 800 topics and for **GLDA** with 130 topics per each location.

The main conclusion based on Figure 1 and Table 1 is that document, query and relevance model expansion methods utilizing geographically-specific topics consistently outperform the methods based on standard LDA. Another interesting observation is that leveraging LMs derived from geographically-specific topics is more effective for query expansion than for document expansion. This can attributed to the fact that queries are more likely to have geographically-specific terms than most of their potentially relevant documents and, thus, the predicted geographical focus of queries is likely to be more accurate. In addition to that, using predicted geographical focus of documents is more effective than assuming that all microblog posts have a fixed geographical focus corresponding to the location of their authors.

Method	MAP	GMAP	P@30
KL-DIR	0.2395	0.1642	0.3354
LDA-DEXP	0.2396	0.1640	0.3340
GLDA-DEXP-STAT	0.2397	0.1641	**0.3347**
GLDA-DEXP-SUM	**0.2401**	**0.1643**	**0.3347**
GLDA-DEXP-MAX	0.2399	0.1642	**0.3347**
LDA-QEXP-SUM	0.2392	0.1633	0.3361
LDA-QEXP-MAX	0.2395	0.1635	0.3361
GLDA-QEXP-SUM	**0.2417***†	**0.1663**	**0.3367**
GLDA-QEXP-MAX	0.2415	0.1658	0.3347
RMOD	0.2859	0.1873	**0.3578**
RMOD-LDA	0.2905	0.1895	0.3517
RMOD-GLDA	**0.2916***†	**0.1904**	0.3571

Table 1: Summary of the best results for document, query and relevance model expansion methods (the highest values of performance metrics for each expansion type are highlighted). * and † indicate statistical significance relative to no expansion and LDA-based expansion baselines respectively with $p < 0.05$, according to paired t-test.

Acknowledgments

This work was supported by the DARPA grant D11AP00269.

4. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3.
[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proc. of EMNLP'10*.
[3] L. Hong, A. Ahmed, S. Gurumurthy, A. Smola, and K. Tsioutsiouliklis. Discovering geographical topics in the twitter stream. In *Proc. of WWW'12*.
[4] V. Lavrenko and W. B. Croft. Relevance-based language models. In *Proc. of ACM SIGIR'01*.
[5] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *Proc. of ACM SIGIR'06*.
[6] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. In *Proc. of ECIR'09*.
[7] C. Zhai and J. Lafferty. Document language models, query models, and risk minimization for information retrieval. In *Proc. of SIGIR'01*.