Hashtag Retrieval in a Microblogging Environment

Miles Efron
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
501 E. Daniel St., Champaign, IL, 61820, USA
mefron@illinois.edu

ABSTRACT

Microblog services let users broadcast brief textual messages to people who "follow" their activity. Often these posts contain terms called hashtags, markers of a post's meaning, audience, etc. This poster treats the following problem: given a user's stated topical interest, retrieve useful hashtags from microblog posts. Our premise is that a user interested in topic x might like to find hashtags that are often applied to posts about x. This poster proposes a language modeling approach to hashtag retrieval. The main contribution is a novel method of relevance feedback based on hashtags. The approach is tested on a corpus of data harvested from `twitter.com`.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Relevance Feedback

General Terms

Experimentation, Performance, Theory

Keywords

microblog, twitter, hashtag, relevance feedback

1. INTRODUCTION

Microblogging services allow users to post brief textual messages that are broadcasted to the user's "followers." Today, the most visible microblogging service is `twitter.com` where users post so-called tweets of no more than 140 characters. While many tweets are inconsequential, others contain information of broad interest, as well as links to external resources (e.g. photos or websites). This poster proposes an approach to one aspect of microblog IR: retrieving hashtags on a topic of interest to a searcher.

Many tweets are marked with so-called hashtags. A hashtag is a character string preceded by a `#` sign. Hashtags often signal aspects of a tweet's meaning such as its topic or its intended audience. Thus `#sigir2010` ostensibly marks tweets related to the 2010 SIGIR conference. A person who is interested in a topic, say *vegetarian recipes*, might want to find hashtags that are often applied to posts about vegetarian recipes, veganism, healthy eating, etc.

SIGIR 2010 Geneva
ACM 978-1-59593-597-7/07/0007.

For a user's topical query q, we wish to find a list of k tags that are relevant to the information need represented by q. The task involves accepting a keyword query and returning a ranked list of hashtags. We approach hashtag retrieval as a type of entity search [1, 2].

Finding useful hashtags offers benefits that are particular to the microblog setting:

- *Tags to follow*: A user may wish to find hashtags that he or she can track on an ongoing basis.
- *Result display*: If users are searching for tweets, people, or posted URLs, returned units of retrieval could be grouped into clusters by their associated hashtags.
- *Query expansion*: Hashtags provide leverage for query expansion during relevance feedback.

The task we address aims to support these actions. This poster explicitly treats only the last item: query expansion. More specifically, we are concerned with query expansion in service to hashtag retrieval.

2. MODEL

Let C be a corpus containing n tweets. Among these n tweets we have m distinct hashtag types. We induce m language models, one per hashtag. To fit a tag t_i's language model we analyze the set of tweets containing t_i, fitting a multinomial over the vocabulary words, with probability vector Θ_i. The maximum likelihood estimator for θ_{wi}, the probability of a word w given Θ_i, is the number of times w occurs in the set of tweets containing t_i divided by this set's total word count. We smooth estimated models $\hat{\Theta}$ by Bayesian updating with Dirichlet priors ($\mu = 2000$). Given a query q generated by the query model Θ_q, we rank hashtags in decreasing order of the negative KL divergence between their models and Θ_q [3]:

$$r(t_i, q) = -D(\Theta_q || \Theta_i) \quad (1)$$

where unless otherwise noted the calculation uses $\hat{\Theta}_q^{ML}$, the maximum likelihood estimator for Θ_q.

2.1 Hashtag Query Expansion

We propose restricting the added query terms to those candidates that are hashtags, stripping candidates of their leading `#`. The topical nature of hashtags motivates this operation.

Let $\mathbf{r}_k$ be the set of the k top-ranked hashtags (by Eq. 1). Here we set $k = 25$. We define Θ_r, a multinomial parameter vector that has non-zero probability over the tags in $\mathbf{r}_k$ and zero probability for all other terms. We derive a feedback

query model: $\hat{\Theta}_{fb} = (1-\lambda)\hat{\Theta}_q^{ML} + \lambda\hat{\Theta}_r$ where λ is a tunable parameter on $(0,1)$ that we fix at 0.2 (a value chosen empirically). As for $\hat{\Theta}_r$, we propose two variants:

- *HFB1*: Non-zero elements of $\hat{\Theta}_r$ are uniformly distributed.
- *HFB2*: each non-zero $\hat{\theta}_{ri}$ is proportional to $\frac{IDF(t_i)}{\max IDF}$

where $IDF(t_i)$ is the inverse document frequency for tag i and $\max IDF$ is the IDF for a tag with document frequency 1. Feedback information enters retrieval by using $\hat{\Theta}_{fb}$ for the query model in Eq. 1.

2.2 Hashtag Association

Given the previous definition of $\mathbf{r}_k$, let $\mathbf{X}$ be the $k \times k$ matrix where for two tags t_i and t_j in $\mathbf{r}_k$, x_{ij} gives the number of times t_i occurs (across the corpus) in a tweet that also contains t_j. Also let $x_{ii} = 1$. We normalize $\mathbf{X}$ so that its columns (rows) are of unit length. Let a tag's association with the retrieved tags $\mathbf{r}_k$ be:

$$a(t_i, \mathbf{r}_k) = \sum_j^k x_{ij}. \quad (2)$$

Eq. 2 is large if a tag co-occurs with many other tags in $\mathbf{r}_k$. Large values for Eq. 2 suggest that a tag has a strong presence in the "neighborhood" of the query. We combine Eq. 1 with Eq. 2, leading to the ranking score, $r_a(t_i, q) = r(t_i, q) + \log a(t_i, \mathbf{r}_k)$. Runs using r_a are designated HFB1a or HFB2a (depending on the HFB used).

3. EVALUATION

We gathered data over a 24-hour period using Twitter's streaming API[1] (cf. Table 1). 29 topical queries were created based on the author's interaction with Twitter. Relevance judgments were obtained using the Amazon Mechanical Turk service[2]. For each query, we created a pool of tags to be judged using runs from three systems: simple KL divergence, KL divergence on a Porter-stemmed corpus, and Rocchio relevance feedback using a TF-IDF model. Each judger was shown a keyword query, a candidate tag, a paragraph-long description of what would make a tag "useful," and a sample of recent tweets using the tag. Judgers ranked each query-tag pair on a four-point scale from 0 (not useful) to 3 (definitely useful). Each query-tag pair was judged by 5 judgers. Usefulness scores were obtained by taking both the mean and median of these 5 scores (we report results based only on means). We take tags with usefulness > 1 to be relevant (graded relevance is used for NDCG).

Table 1: Test collection summary statistics

Number of tweets	3,414,330
Number of hashtag types	50,097
Number of hashtag tokens	571,861
Number of users	874,892
Number of queries	39
Median number relevant tags	28.5

[1] `http://api.twitter.com`
[2] `http://www.mturk.com`

We tested four experimental conditions:

1. Baseline, no feedback: simple KL divergence
2. Baseline, with feedback: KLD retrieval with divergence minimization feedback (divergence minimization gave stronger results than mixture model feedback) [3]
3. Hashtag-based query expansion (HFB1 and HFB2)
4. Hashtag query expansion with association measure (HFB2a).

For all feedback, five terms were added to the initial query, with a weight of 0.2. The baseline feedback model used the top 10 documents. No stemming or stoplists were applied. Runs returned the top 25 tags for each query.

We report three statistics (Table 2): Mean average precision (MAP), normalized discounted cumulative gain (NDCG) at 15, and precision at 10 (P10). *All runs using hashtag-based feedback gave results that were statistically significantly better than the baseline run using standard term-based feedback.*

Table 2: Retrieval effectiveness. All HFB runs show statistically significant improvement over the baseline feedback run on all three measures ($p < 0.05$ on a randomization test). † indicates $p < 0.01$.

Method	MAP	NDCG	P10
Baseline, no FB	0.4268	0.6110	0.7034
Baseline, FB	0.4381	0.6209	0.7138
HFB1	0.4605	0.6431†	0.7483
HFB2	0.4617†	0.6388†	0.7414
HFB2a	0.4684†	0.6488†	0.7483

Table 2 suggests that hashtags provide useful information for relevance feedback. While the baseline (term-based) feedback run was only slightly more effective than the baseline run without feedback, all tag-based feedback runs performed better than the term-based baseline feedback model. HFB2 (using IDF weighting for feedback tags) gives marginal improvement over uniformly weighted expansion, while our association measure gives a bit more of an edge. However, the differences among the three test conditions are slight.

4. CONCLUSIONS

In future work we will identify additional features of hashtags (e.g. from author-based statistics) for use in IR. We will also undertake a more thorough empirical evaluation. The main work, however, lies in defining the relevant problems, applications, and user needs in IR from microblogs.

5. REFERENCES

[1] K. Balog, L. Azzopardi, and M. de Rijke. A language modeling framework for expert finding. *Inf. Process. Manage.*, 45(1):1–19, 2009.

[2] C. Macdonald and I. Ounis. Using relevance feedback in expert search. *Proc. of 29th European Conf. on Information Retrieval, Springer LNCS*, 4425:431–443, 2007.

[3] C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM01: Proc. of the 10th Intl Conference on Information and Knowledge Management*, pages 403–410, 2001.