ABSTRACT

On various microblogging platforms like Twitter, the users post short text messages ranging from news and information to thoughts and daily chatter. These messages often contain keywords called Hashtags, which are semantico-syntactic constructs that enable topical classification of the microblog posts. In this poster, we propose and evaluate a novel method of semantic enrichment of microblogs for a particular type of entity search – retrieving a ranked list of the top-k hashtags relevant to a user’s query $Q$. Such a list can help the users track posts of their general interest. We show that our techniques significantly improved microblog retrieval as well. We tested our approach on the publicly available Stanford sentiment analysis tweet corpus. We observed an improvement of more than 10% in NDCG for microblog retrieval task, and around 11% in mean average precision for hashtag retrieval task.

1. INTRODUCTION

Microblogging services enable communication at a massive scale. Twitter, one of the most popular microblogging services reportedly has 284 monthly active users, who post around 500 million short text posts (limited to 140 characters) called “tweets” everyday. Often users include keywords prefixed with ‘#’, called hashtags to indicate and organize the contextual meanings of their tweets. Hence, to retrieve information related to a user’s interest, for instance, “Rock concerts”, it’d be very helpful to the user if they can be suggested a list of hashtags which are commonly used in relation to “Rock concerts”. By tracking these hashtags, a user can gain information about rock concerts via the posted tweets. However, it’s not possible for the user to manually figure out all the hashtags that are used across Twitter, relevant to their interest. In this poster, we address this problem and present a system (publicly accessible at http://bit.ly/SemanticHashtagRetrieval) that takes a query $Q$ from the user and returns a ranked list of hashtags most relevant to $Q$.

In the past, microblog retrieval has been studied in various contexts: Sakaki et al. [5] exploit the real-time nature of tweets to discover events. TweetMotif [4] presents an exploratory search interface to deal with microblog retrieval, trying to summarize topics by analyzing co-occurrence patterns. More recently, Efron et al. [3] have studied hashtag retrieval from a query expansion point of view. However, almost all these retrieval approaches are strictly term-based, which are sensitive to polysemy, and term-use variation. In this poster, we propose “Semantically Enriched Microblog Document (SEMD)” structure, which enables semantic retrieval of hashtags and microblog posts, trying to overcome these limitations to a greater extent. Table 1 shows the top two hashtags retrieved for a small subset of queries.

2. SEMANTIC ENRICHMENT

Traditionally most of the microblog IR has either ignored hashtags for analysis, or has treated them as single words. However, that might not be true for most of the hashtags - #WWW2015Firenze refers to “WWW 2015 Firenze”. Recent work by Bansal et al. [1] presents a machine learning based approach to segment the hashtags and link the entities in hashtags to Wikipedia. This allows to extract latent semantic information about hashtags. It’s worth noting here that their proposed approach also performs entity linking on the rest of the tweet text. We follow a similar approach with a few modifications to reduce latency and ensure high throughput, which is critical for a real time search engine such as Twitter. Notably, we make the following modifications:

1. Unlike Bansal et al. [1], we replace Microsoft Web N-
We perform our evaluation for two tasks - NDCG 0.591 0.713 0.815 and MAP 0.475 0.794. Our Approach 0.759 0.499 0.55.

NDCG 0.451 0.698 0.687 and MRR 0.569 0.670 0.886 0.761 0.845 0.686. MRR

We used TF-IDF weights for pseudo relevance feedback. The quality of hashtags retrieved, while boosting the recall. For any given query Q, BM25F associates weight to each document field inversely proportional to its length. Hence, shorter fields like f_{LT}, f_H and f_{SH} have more weight in retrieval process. Our premise is that these fields have greater importance than f_{LTH} and f_{LSSH}, since hashtags tend to be the gist of the tweet.

Semantically Enriched Microblog Document: We propose a virtual document structure that is enriched with semantic information obtained as described in the above section. The proposed document has 5 fields as can be seen in Table 2. We use Whoosh library’s BM25F implementation to retrieve microblogs most relevant to a given query Q. BM25F associates weight to each document field inversely proportional to its length. Hence, shorter fields like f_{LT}, f_H and f_{SH} have more weight in retrieval process. Our premise is that these fields have greater importance than f_{LTH} and f_{LSSH}, since hashtags tend to be the gist of the tweet.

3. RETRIEVAL PROCEDURE

For a given user query Q, we obtain a list of top 500 microblogs ranked by their relevance according to SEMD structure. In order to retrieve most relevant hashtags, we propose and experiment with a few hashtag ranking approaches -

1. GlobalRank (GR), where the hashtags in top 500 retrieved microblogs were ranked on the basis of their frequency in the overall corpus.
2. RetrievedHashtagRank (RHR), where the hashtags in top 500 retrieved microblogs were ranked on the basis of their frequency in top 500 microblogs.
3. TF-IDFReRank (TFIDFR), where we iteratively refine the quality of hashtags retrieved, while boosting the recall. We used TF-IDF weights for pseudo relevance feedback.
4. KLDivergenceRank (KLD), where we use KLD divergence for blind feedback.

We observed that KLD performed significantly better than the other proposed approaches. A detailed comparative analysis is present in Table 3.

4. EXPERIMENTS AND RESULTS

Dataset: We used Stanford Sentiment Analysis tweet corpus [1] for testing our approach. The dataset consists of 1.6 million tweets collected between April 6, 2009 and June 25, 2009.

Experiments: We perform our evaluation for two tasks - 1. Hashtag Retrieval, and 2. Microblog Retrieval. 5 experienced Twitter users were explained the problem statements. Subsequently, they were asked to suggest 10 queries each, relevant to the given tasks, hence obtaining a list of total 50 queries. We define the baseline as BM25F search conducted on original microblog posts. The users were asked to search for the 50 queries on our search system, ranking each result on a 5 point Likert scale. Our evaluation interface presented the results from baseline, as well as from 

http://webscope.sandbox.yahoo.com
http://weblm.research.microsoft.com
https://pypi.python.org/pypi/whoosh
http://cs.stanford.edu/people/slecago/trainingandtestdata.zip

Table 3: Comparative results for hashtag retrieval

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Baseline</th>
<th>Our Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>MAP</td>
<td>MRR</td>
</tr>
<tr>
<td>GR</td>
<td>0.392</td>
<td>0.557</td>
</tr>
<tr>
<td>RHR</td>
<td>0.435</td>
<td>0.577</td>
</tr>
<tr>
<td>TFIDFR</td>
<td>0.475</td>
<td>0.687</td>
</tr>
<tr>
<td>KLD</td>
<td>0.567</td>
<td>0.689</td>
</tr>
</tbody>
</table>

Table 4: Comparative results for microblog retrieval

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Baseline</th>
<th>Our Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>Rel. Recall</td>
<td>MAP</td>
</tr>
<tr>
<td>GR</td>
<td>NA</td>
<td>0.761</td>
</tr>
<tr>
<td>RHR</td>
<td>0.845</td>
<td>0.815</td>
</tr>
</tbody>
</table>

5. CONCLUSIONS AND FUTURE WORK

We have presented and evaluated a semantic search system in context of hashtag and microblog retrieval. We demonstrate how our approach is better than the existing approaches by detailed experiments. In the future, we’d experiment by enriching microblog posts with additional semantic information. This would include data mined from external links present in the microblog posts, author information and location et cetera.

6. REFERENCES