Query Expansion for Microblog Retrieval : 2013

Ayan Bandyopadhyay
ISI Kolkata

1 Introduction

Microblogging sites like `http://twitter.com` have emerged as a popular platform for expressing opinions. Given the increasing amount of information available through such microblogging sites, it would be nice to be able to retrieve useful tweets in response to a given information need. Finding relevant tweets that match a user query is challenging for the following reasons.

- Tweets are short. They contain a maximum of 140 characters.
- Tweets are not always written maintaining formal grammar and proper spelling. Spelling variations increase the likelihood of vocabulary mismatch.

In this preliminary study, we explore standard query expansion approaches as a way to address this problem.

Related work. Our approach is based on a method proposed by Kwok et al. [4] to improve weak ad-hoc queries through "web assistance", in which the Web (accessed via the Google search engine) is used as a source of expansion terms. We took the cue from this paper and used the Google Search API (GSA) to tap the Web as an external corpus for query expansion.

2 Our Approach

2.1 Topic Processing for TREC 2013 Microblog submission

The original queries were submitted to the Google search API[1](GSA). For each query, GSA returned a maximum of 8 pages of results, with a maximum of 8 results per page. Thus, at most $8 \times 8 = 64$ results were returned per query. We collected together all the titles from the list of returned results. We submitted the following runs.

- **R1 (GSAT):** Results were retrieved for each query using the Google Search API. The title words from all the pages returned by Google were sorted in descending order of their frequencies. The most frequent five words were added to the original topic.
- **R2 (GSAS):** This is the same as the above, but uses the snippets returned by GSA instead of the titles.
- **R3 (GSAA):** This is the same as R1, except that we used both the titles and snippets during query expansion.
- **Auto:** This is the query-wise median over 65 automatic submitted runs.
- **All:** This is the query-wise median over all 65 automatic and 6 manual submitted runs.

[1] http://code.google.com/apis/websearch

3 Results

	MAP	Mean of p@30	Mean of R-Prec
GSAT	0.2351	0.4044	0.2920
GSAS	0.2351	0.4044	0.2920
GSAA	0.2412	0.4061	0.2996
Auto	0.2126	0.4217	0.2721
All	0.2212	0.4311	0.2834

Table 1: Results of our three runs

We get a better result when GSAS and GSAT are combined (GSAA).

References

[1] Gianna M. Del Corso, Antonio Gulli, and Francesco Romani. Ranking a stream of news. In *WWW*, 2005.

[2] Anlei Dong, Yi Chang, Zhaohui Zheng, Gilad Mishne, Jing Bai, Ruiqiang Zhang, Karolina Buchner, Ciya Liao, and Fernando Diaz. Towards recency ranking in web search. In *WSDM*, pages 11–20. ACM, 2010.

[3] Miles Efron. Hashtag retrieval in a microblogging environment. *SIGIR*, pages 787–788, 2010.

[4] Kui-Lam Kwok, Laszlo Grunfeld, and Peter Deng. Improving weak ad-hoc retrieval by web assistance and data fusion. In Gary Geunbae Lee, Akio Yamada, Helen Meng, and Sung-Hyon Myaeng, editors, *AIRS*, volume 3689 of *Lecture Notes in Computer Science*, pages 17–30. Springer, 2005.

[5] K. Massoudi, E. Tsagkias, M. de Rijke, and W. Weerkamp. Incorporating query expansion and quality indicators in searching microblog posts. *ECIR 2011*, pages 362–367, 2011.