Neurocomputing 152 (2015) 274–286

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Search engine reinforced semi-supervised classification and graph-based summarization of microblogs

Yan Chen [a,*], Xiaoming Zhang [a], Zhoujun Li [a], Jun-Ping Ng [b]

[a] State Key Laboratory of Software Development Environment, Beihang University, China
[b] Bloomberg L.P., USA

ARTICLE INFO

Article history:
Received 29 April 2014
Received in revised form
14 October 2014
Accepted 31 October 2014
Communicated by Y. Chang
Available online 11 November 2014

Keywords:
Microblog
Topic classification
Summarization
Probabilistic graphical model
Semi-supervised
Pagerank

ABSTRACT

There is an abundance of information found on microblog services due to their popularity. However the potential of this trove of information is limited by the lack of effective means for users to browse and interpret the numerous messages found on these services. We tackle this problem using a two-step process, first by slicing up the search results of current retrieval systems along multiple possible genres. Then, a summary is generated from the microblog messages attributed to each genre. We believe that this helps users to better understand the possible interpretations of the retrieved results and aid them in finding the information that they need. Our novel approach makes use of automatically acquired information from external search engines in each of these two steps. We first integrate this information with a semi-supervised probabilistic graphical model, and show that this helps us to achieve significantly better classification performance without the need for much training data. Next we incorporate the extra information into graph-based summarization, and demonstrate that superior summaries (up to 30% improvement in ROUGE-1) are obtained.

1. Introduction

Microblog services have provided a platform for users to convey their thoughts, share their experience and perform virtual social activities. One of the better-known platforms include Twitter,[1] which has more than 140 million active users and 1 billion new microblog messages (or *tweets*) posted every 3 days[2] as of March 2012. Over time, the tremendous number of microblog messages that have been generated makes up a large and informative repository from which users can query to retrieve information.

However the sheer volume of these messages is a double-edged sword. To get the information they want, users have to wade through voluminous search results to locate the information they are interested in. While textual markups like hashtags can help users to zoom-in quickly on messages of interests, the open-ended nature and free styling of these markups affect the effectiveness of such searches. Typical microblog platforms display search results in a ranked list sorted in order of relevance to query keywords or hashtags. Unfortunately, these search terms are very short, potentially ambiguous, or even vague, leading to unsatisfactory search results. For example, searching for the keyword “apple” on Twitter returns a set of messages that is diversified and varied. Fig. 1 shows an example of the results obtained from Twitter. The results span several different genres, ranging from albums (presumably sold on the Apple iTunes store), to the band Beatles, to a reference to the voice recognition feature *Siri* on a phone made by the company named “Apple”. If luck has it, we may probably retrieve tweets on the fruit itself!

We believe that to fully tap on the potential of the large repository of microblog messages, it is important to make it easier for users to retrieve meaningful search results. For example if the search results are presented based on meaningful, loosely structured genres, along with a summary of the information within each genre, users can potentially see at a glance which of these genres are relevant to their information needs. All these are more important if the user is making use of a mobile device with limited screen estate (such as smartphones). In fact recent surveys [1] have confirmed that the use of smartphones and tablets for internet access has increased multi-fold.

A similar problem already exists — news reading, where the huge number of news sources reporting on the same event may overwhelm a user. News aggregators (*e.g.*, Google News[3]) have evolved to tackle this problem by clustering reports on the same events together, while multi-document summarization systems present a short snippet of the highlights in the cluster of reports to

* Corresponding author.
E-mail addresses: chenyan@cse.buaa.edu.cn (Y. Chen), yolixs@buaa.edu.cn (X. Zhang), lizj@buaa.edu.cn (Z. Li), email@junping.ng (J.-P. Ng).

[1] http://twitter.com
[2] http://blog.twitter.com/2012/03/twitter-turns-six.html
[3] http://news.google.com

http://dx.doi.org/10.1016/j.neucom.2014.10.068
0925-2312/

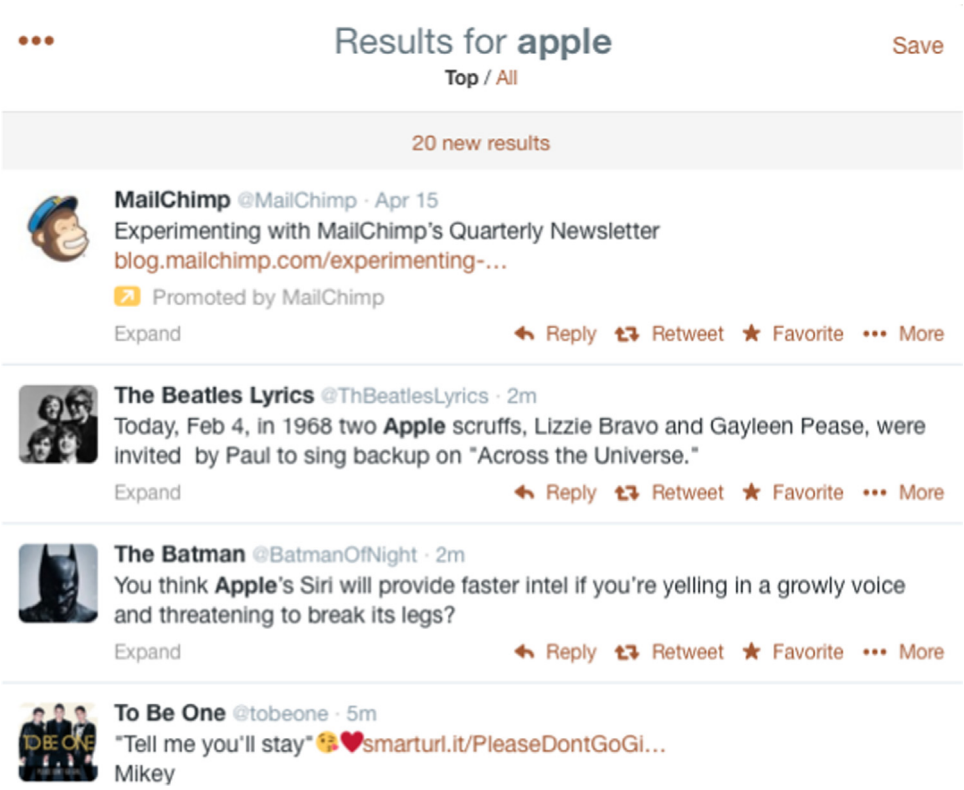

Fig. 1. Extract of search results for "apple" using Twitter's search function.

the user. We are proposing a similar solution here for microblog services. However the challenges involved in dealing with long news articles are different from those faced when working with the typically short text snippets in microblogs.

The main challenge is due to the length limitations imposed on microblog messages. These messages are typically short, consisting of no more than 140 characters. This data sparsity is a problem when trying to classify these messages into different genres. The lack of sufficient contextual information means that traditional similarity measures such as the use of word co-occurrences are ineffective [2]. A secondary problem has to do with the training corpora that are required by popular supervised machine learning-based solutions to achieve good performance. While there is a good number of such corpora for traditional domains like news-wire articles, building up and annotating similar corpora for microblog messages is laborious and time-consuming.

We tackle these two key challenges in this paper by making innovative use of search engines to automatically acquire extra documents and text to enrich the collection of microblog messages we are working with. In doing so, we are able to obtain more relevant text content to overcome the problem of data sparsity. It also allows us to adopt a semi-supervised methodology which requires far less training material than traditional supervised machine learners.

To classify microblog messages into one of the several genres, we propose the use of a semi-supervised probabilistic graphical model which combines textual content from microblog messages and automatically acquired content from search engine results. The model learns a suitable genre distribution with which we can assign individual microblog messages into the most likely genres. To generate a summary for each genre, we evaluate several graph-based summarization algorithms, built on the popular PageRank [3] and HITS [4] algorithms. We further modified these algorithms to take in additional text content from relevant web search results and show that this helps improve summarization performance.

The key contribution of this work is our novel proposal to incorporate the use of external resources to overcome the lack of contextual information inherent with microblog messages. These external resources are obtained automatically and efficiently, and we show that they help to (1) improve the performance of a semi-supervised classification model significantly, thereby reducing our reliance on large annotated datasets, and (2) improve the quality of summaries generated from microblog messages.

2. Related work

Our work overlaps two key areas of research: (1) topic classification and (2) microblog summarization. In this section, we explore related literature for each of them in turn. We also review existing work which adopt a similar "classify-then-summarize" approach to ours and share our aim of easing the information overload that users face today.

2.1. Topic classification

The task of topic classification of microblog messages (or what we refer to subsequently as *genre classification* in the rest of this paper) is to assign messages to one of the several pre-identified class labels. Topic classification is a fundamental task for many applications, including query disambiguation [5], location prediction [6] and hot topic tracking [7].

A common approach to topic classification is with the use of topic models. Of significance here is the work of Hong and Davison [8], where the use of latent Dirichlet allocation (LDA) [9] and author-topic models [10] is explored to automatically detect hidden topic structures within Twitter messages. Several variants of LDA have been proposed [11,12] and have been shown to be competitive for the classification of microblog messages. Although these LDA-based topic models work well generally, they are however hampered by the length limits imposed on most microblog messages.