Information Processing and Management 51 (2015) 89–113

Contents lists available at ScienceDirect

Information Processing and Management

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

Burst-aware data fusion for microblog search

Shangsong Liang *, Maarten de Rijke

University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands

ARTICLE INFO

Article history:
Received 26 March 2014
Received in revised form 31 October 2014
Accepted 31 October 2014
Available online 2 December 2014

Keywords:
Information retrieval
Microblog search
Rank aggregation
Burst detection
Temporal information retrieval

ABSTRACT

We consider the problem of searching posts in microblog environments. We frame this microblog post search problem as a late data fusion problem. Previous work on data fusion has mainly focused on aggregating document lists based on retrieval status values or ranks of documents without fully utilizing temporal features of the set of documents being fused. Additionally, previous work on data fusion has often worked on the assumption that only documents that are highly ranked in many of the lists are likely to be of relevance. We propose BurstFuseX, a fusion model that not only utilizes a microblog post's ranking information but also exploits its publication time. BurstFuseX builds on an existing fusion method and rewards posts that are published in or near a burst of posts that are highly ranked in many of the lists being aggregated. We experimentally verify the effectiveness of the proposed late data fusion algorithm, and demonstrate that in terms of mean average precision it significantly outperforms the standard, state-of-the-art fusion approaches as well as burst or time-sensitive retrieval methods.

1. Introduction

Microblogging platforms, such as Twitter,[1] have become indispensable communication channels through which hundreds of millions of users around the world witness breaking news events. The characteristics of the posts, such as their limited length, along with easy access on many platforms, facilitate regular status updates by large numbers of people (Zhao & Rosson, 2009). Microblogging platforms display fast paced dynamics as reflected by rapidly evolving topics (Yang & Leskovec, 2011). Searching posts in such rapidly changing environments is a challenge (Ounis, Macdonald, Lin, & Soboroff, 2011). To tackle this problem, much previous work has focused on content-based criteria for ranking posts in response to a query, in combination with a broad range of other ranking criteria, including, e.g., existence of hyperlinks, hashtags and retweets.

Fusion is a popular method for generating result lists based on multiple ranking criteria. Previous research has found that data fusion can enhance the retrieval performance in many cases (Dong & Srivastava, 2013; Shaw, Fox, Shaw, & Fox, 1994; Wu, 2012). In this paper, we look at the problem of searching microblog posts as a late data fusion task (Shaw et al., 1994): we fuse ranked lists of posts produced by a diverse set of microblog post rankers into a single final ranked list of posts. In the following, we consider the case where only ranks and publication times are available and no other additional information is provided such as the retrieval status values or the contents of the posts. We focus on a particular microblog search scenario, one that was studied at the Text REtrieval Conference (TREC) 2011 and 2012 Microblog tracks (Ounis et al., 2011; Soboroff,

* Corresponding author.
E-mail addresses: s.liang@uva.nl (S. Liang), derijke@uva.nl (M. de Rijke).

[1] http://www.twitter.com.

http://dx.doi.org/10.1016/j.ipm.2014.10.008
0306-4573/

Ounis, Macdonald, & Lin, 2012). The task uses Twitter data and is defined as follows: given a query with a timestamp, return relevant and interesting tweets.

Fusing multiple document lists that have been retrieved from a corpus in response to a query so as to compile a single result list, has a long history (Kozorovitsky & Kurland, 2011; Shaw et al., 1994; Tsai, Wang, & Chen, 2008), with the CombSUM family (CombMax, CombMin, CombSUM, CombANZ, CombMNX, CombMNZ, etc. Lee, 1995) of fusion methods being the oldest and one of the most successful ones for many IR tasks (He & Wu, 2008; Sheldon, Shokouhi, Szummer, & Craswell, 2011; Tsagkias, de Rijke, & Weerkamp, 2011; Tsai et al., 2008). The lists are often produced by multiple ranking functions, e.g., query representations or document representations (Croft, 2000). Many effective fusion methods are based on the assumption that only documents that are highly ranked in many of the lists are likely to be relevant (Aslam & Montague, 2001; Croft, 2000; Dwork, Kumar, Naor, & Sivakumar, 2001; Kozorovitsky & Kurland, 2011; Lee, 1995; Montague & Aslam, 2002; Shaw et al., 1994; Tsagkias et al., 2011). As a consequence, a relevant document will be ranked low in the final fused list if it appears only in a single list and is ranked low in this list.

The characteristics of microblog environments suggest a different perspective. In such environments news events trigger people to talk about the topics related to an event mostly during specific short time intervals (Chen, Chen, Zhang, Wang, & Bu, 2010; Hoonlor, Szymanski, Zaki, & Chaoji, 2012; Lappas, Arai, Platakis, Kotsakos, & Gunopulos, 2009; Mathioudakis, Bansal, & Koudas, 2010; Peetz, Meij, de Rijke, & Weerkamp, 2012; Vlachos, Meek, & Vagena, 2004). For instance, people talked about the "2014 Eastern Synchronized Skating Sectional Championship" mainly between January 30 and February 1, 2014, which is when the championship was held. Posts created before the beginning or after the ending of the event are less likely to discuss the championship competitions and, hence, are less likely to be relevant. This observation leads to the following intuition about fusing ranked lists of microblog posts. If a post d and (other) relevant posts $d_1, \ldots, d_k$ were published within the same narrow time window, and the relevant posts $d_1, \ldots, d_k$ are ranked highly in many of the lists to be merged, then post d should be "rewarded" by boosting its rank, even if, in the extreme case, it appears in only one list where it is ranked low. Fig. 1 illustrates this intuition; there, post d_2 is ranked low in list L_1 but our intuition suggests that it should be rewarded as it was published in the same narrow time window in which a large number of posts occur that are ranked high in many lists; in contrast, d_8, while ranked high in L_m, receives no such bonus as it was published outside the narrow window.

To tackle the problem of microblog post search, we propose BurstFuseX, a novel probabilistic model that not only utilizes information traditionally used when merging ranked lists, such as ranks, but also exploits temporal information, i.e., the publication timestamps of microblog posts. In our fusion model, we focus on the case where only ranks and publication timestamps are available and no additional information is provided—such as the content of the posts, the post's RSVs (Relevance Status Values), and the resources the posts link to. In fact, accessing the contents of posts may be inefficient and hence inappropriate in dynamic environments such as microblog search. In addition, the content may not be available in all scenarios (Salakhutdinov & Mnih, 2008). Briefly, BurstFuseX first calls a standard document fusion method X to merge a set of ranked lists of microblog posts for a given query. Subsequently, as illustrated in Fig. 1, based on the fused scores produced by method X, we detect windows of timestamps of high-scoring posts. These windows give rise to bursts of posts. We then reward posts that are published in the temporal vicinity of a burst that contains high-scoring posts.

In our experiments aimed at assessing the performance of BurstFuseX, we sample runs that have been submitted to the TREC 2011 and 2012 Microblog tracks and fuse them using BurstFuseX, respectively. For the underlying fusion method X (on top of which BurstFuseX builds), we consider three alternatives: two unsupervised fusion methods, CombSUM (Shaw et al.,

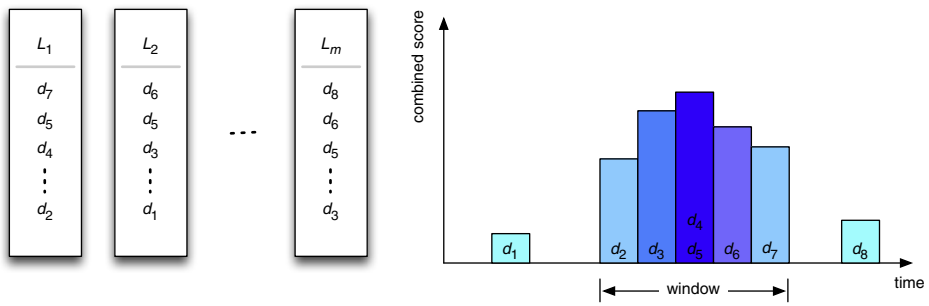

Fig. 1. Rewarding posts that are published in the same narrow time frame as a large number of (supposedly) relevant posts. On the left, we display m ranked lists of posts that were produced in response to a given query; these lists need to be fused. Post d_2 only occurs in list L_1 and it is ranked low in L_2; d_8 also occurs in a single list, L_m, but it is ranked very high. On the right, we show the distribution of the publication timestamps of the documents in the lists to be combined. The vertical axis indicates the combined scores of posts with the same timestamps based on a baseline fusion method, e.g., CombSUM. According to its publication timestamp, d_2 was published in a "good" period for the query: many posts published around the same time as d_2 are highly ranked in many lists; because of this, BurstFuseX will "reward" d_2. In contrast, d_8 does not have a publication time around which many highly ranked posts were published, hence it should not receive a reward.