Sifting Micro-blogging Stream for Events of User Interest

Maxim Grinev, Maria Grineva, Alexander Boldakov, Leonid Novak, Andrey Syssoev, Dmitry Lizorkin
Institute for System Programming of the Russian Academy of Sciences
{grinev, rekouts, boldakov, novak, syssoev, lizorkin}@ispras.ru

1. INTRODUCTION

Micro-blogging is a new form of social communication that allows users to post brief text messages to be viewed by either a limited group chosen by a user or by anyone. Micro-blogging encourages users to share information about anything they are seeing or doing, the motivation facilitated by interface simplicity and the ability to use a variety of means for posting messages. *Twitter*, the most popular micro-blogging tool, is exhibiting rapid growth [3]: up to 11% of online Americans are using Twitter by December 2008, compared to 6% in May 2008.

Due to its nature, micro-blogosphere has unique features: (i) It is a source of extremely up-to-date information about what is happening in the world; (ii) It captures the wisdom of millions of people and covers a broad range of domains: from US president inauguration to album release by a little known music band.

These features make micro-blogosphere more than a popular medium of social communication: we believe that it has additionally become a valuable source of extremely up-to-date news on virtually any subject of user interest.

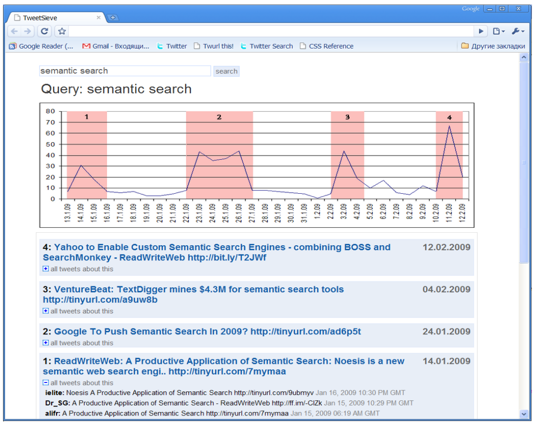
Making use of micro-blogosphere in this new role we meet the following challenges: (A) Since any given subject is generally mentioned in the micro-blogging stream on the continuous basis, a method is needed for locating periods of news on this subject. (B) Additionally, even for such periods, stream filtering is required for removing noise and for extracting messages that best describe the news.

To address these challenges we make and exploit the following observations: (A) For an arbitrary subject, events that catch user interest gain distinguishably more attention than the average mentioning of the subject resulting in message *activity bursts* for it. (B) Most of the messages in an activity burst describe common event in close variations–either rephrased or "retweeted" between the users.

We demonstrate TweetSieve–a system that allows obtaining news on any given subject by sifting the Twitter stream. Our work is related to frequecy-based analysis applied to blogs [1], but higher latency and lower coverage in blogs makes the analysis less effective than in case of micro-blogs.

2. SYSTEM OVERVIEW

In TweetSieve demo, the user is able to express the subject of her interest by an arbitrary search string. The system shows the period of events occuring for the subject and outputs tweets that best describe each of the events. Fig. 1 shows a screenshot of the system for "Semantic search" as a sample subject.



semantic search

Query: semantic search

4: Yahoo to Enable Custom Semantic Search Engines - combining BOSS and SearchMonkey - ReadWriteWeb http://bit.ly/T2JWf 12.02.2009
⊞ all tweets about this

3: VentureBeat: TextDigger mines $4.3M for semantic search tools http://tinyurl.com/a9uw8b 04.02.2009
⊞ all tweets about this

2: Google To Push Semantic Search In 2009? http://tinyurl.com/ad6p5t 24.01.2009
⊞ all tweets about this

1: ReadWriteWeb: A Productive Application of Semantic Search: Noesis is a new semantic web search engi.. http://tinyurl.com/7mymaa 14.01.2009
⊟ all tweets about this
ielite: Noesis A Productive Application of Semantic Search http://tinyurl.com/9ubmyv Jan 16, 2009 10:30 PM GMT
Dr_SG: A Productive Application of Semantic Search - ReadWriteWeb http://ff.im/-CIZk Jan 15, 2009 10:29 PM GMT
alifr: A Productive Application of Semantic Search http://tinyurl.com/7mymaa Jan 15, 2009 06:19 AM GMT

Figure 1: Screenshot of TweetSieve user interface

The underlying process consists of two steps:

Identifying activity bursts. Counting the messages matching the search string in the stream over time, the frequency curve is constructed. Activity bursts in the curve are identified by taking the periods of frequency exceeding the standard deviation from the average.

Selecting messages that best describe news events. For the set of all messages matching the search string in an activity burst, we apply the message-granular variation of our keyphrase extraction algorithm [2] that is specifically suited to efficiently filtering noisy data. The algorithm clusters messages with respect to their similarity to each other and chooses central messages from the most dense clusters. As the similarity measure we use Jaccard coefficient for the "bag of words" representation of messages.

The demonstration illustrates the potential of our approach in bringing news acquisition to a new level of promptness and coverage range.

3. REFERENCES

[1] N. Bansal and N. Koudas. Blogscope: A system for online analysis of high volume text streams. In *VLDB*, pages 1410–1413. ACM, 2007.

[2] M. Grineva, M. Grinev, and D. Lizorkin. Extracting key terms from noisy and multitheme documents. Accepted to the World Wide Web conference (WWW'09).

[3] A. Lenhart and S. Fox. Twitter and status updating. Pew Internet & American Life Project, Feb 2009.

Requirements Description

- Internet connection