TweetMogaz: A News Portal of Tweets

Walid Magdy
Qatar Computing Research Institute
Qatar Foundation
Doha, Qatar
wmagdy@qf.org.qa

ABSTRACT

Twitter is currently one of the largest social hubs for users to spread and discuss news. For most of the top news stories happening, there are corresponding discussions on social media. In this demonstration TweetMogaz is presented, which is a platform for microblog search and filtering. It creates a real-time comprehensive report about what people discuss and share around news happening in certain regions. TweetMogaz reports the most popular tweets, jokes, videos, images, and news articles that people share about top news stories. Moreover, it allows users to search for specific topics. A scalable automatic technique for microblog filtering is used to obtain relevant tweets to a certain news category in a region. TweetMogaz.com demonstrates the effectiveness of our filtering technique for reporting public response toward news in different Arabic regions including Egypt and Syria in real-time.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering.

Keywords

Twitter, Microblog Filtering, TweetMogaz, Arabic text.

1. INTRODUCTION

Microblogging sites, such as Twitter, are currently one of the main platforms for exchanging real-time information and discussions. In fact, Twitter and Facebook were instrumental in facilitating the launch of the so-called "Arab Spring". This large amount of information led to the need for an effective platform for information filtering, to allow only relevant information reaches users. The currently implemented and straightforward microblog filtering technique on Twitter is the "follow" feature. This allows users to follow other accounts of entities, persons, or events to be fed with their tweets. This method is personalized according to user's interest. Another method for following specific micrblogs on Twitter is searching for given hashtags (#tags), which is a common way for users to get updates about some topics based on the mention of the hashtag within tweets text. This method is less strict in filtering information, where more tweets are generally presented to user. However, many unneeded tweets would be retrieved because of the misusage of hashtags by some users. Additionally, many relevant tweets to a hashtag topic may not include the hashtag itself, which leads to their absence in the retrieved results.

In this demonstration we present a news portal website, TweetMogaz, which is generated from tweets. TweetMogaz presents the most popular content people share on Twitter regarding the ongoing news in different regions. Visitors of the website can see a comprehensive report of the most popular tweets, jokes, videos, images, and news articles that people share on Twitter related to the top news stories of the day. Standard news sites give their visitors an idea about what is happening in given regions, while TweetMogaz gives visitors an idea about what are the topics in news that people are interested in, and how they react towards them. In addition, it captures additional aspects of news stories shared on social media that may not exist in the news sites.

TweetMogaz applies microblog filtering technique for retrieving tweets. A set of key players in news for a certain region are manually listed, and one or multiple news sites for the same region are set to our system. Potential relevant tweets to news are retrieved; then a relevance classification technique is used to identify the relevant tweets and discard the rest. Classified relevant tweets are then used to generate comprehensive report with most popular content people share as in [3]. Information on website is updated every 15 minutes to keep following fresh news and their response on social media.

Our approach for retrieving and filtering relevant tweets shows high performance. The approach is applied to follow political news in counties in the Arabic region such as the Egypt and Syria. However, it is potentially applicable to different regions and different news categories. The only requirement is preparing a list of key players of a news category in a given region and setting a regional news site.

2. RELATED WORK

Previous work in Microblog retrieval focused on analyzing the search process [4, 8], improving ad-hoc microblog search [6], and developing platforms for improved user experience with search through analyzing results [3, 9]. Additional work studied the role of micrblogs in news reporting and discovery, and how users' profiles can be used for news recommendation [5, 7]. Other work investigated detecting comments about news from Twitter to be presented to readers along with news articles [2].

Recently, microblog filtering grabbed some attention as an important task for allowing users to follow certain topics. Microblog filtering was introduced as a new task in TREC Microblog track 2012, where the aim was to filter a feed of tweets by getting relevant ones to some topics. The best achieved result in the track got a precision of 0.6, which is considerably low for usage in a practical environment [6].

Our system presented in this demonstration applies microblog filtering [6] for getting tweets relevant to regional news [5, 7] in a practical environment with high precision. Search results are presented to user in the form of a comprehensive report [3]. Our system enables user to get a summary about the public response towards ongoing news [2].

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
ACM 978-1-4503-2034-4/13/07.

3. MICROBLOG FILTERING APPROACH

Our filtering approach for retrieving relevant tweets to regional news is shown in Figure 1. The approach is decomposed into the following steps:

1. Retrieving initial set of relevant tweets

Any region has a set of key players who are expected to be a usual actor in news headlines. For example, "Obama" is a key player in US politics. These set of key players are nearly static, where they do not change frequently by time. Therefore, we prepare a list of accurate predefined queries representing key player in a certain region to retrieve an initial set of relevant tweets. Queries include politicians, parties, institutes … etc, and their corresponding Twitter accounts. The list of queries requires updating each few months or years according to changes in the region. These queries are set carefully to achieve high precision to avoid the retrieval of irrelevant tweets. For example, setting a query "Obama" referring to the US president is acceptable, since most of the tweets talking about "Obama" refer to the president himself. While searching for "Clinton" as a query for "Bill Clinton" can lead to the retrieval of a large number of irrelevant tweets for those tweets referring to "Hillary Clinton". Therefore, in the latter case, it is better to have the query as "Bill Clinton" to emphasis on high precision results.

Stream of tweets that match any of the predefined queries are considered relevant. Matching tweets are referred to as Key Players Tweets set ($Tweets_{KP}$).

2. Retrieving set of Potential Relevant Tweets

Tweets about accidental regional news may not be captured with the set of predefined queries. To overcome this problem, news is explored on one or more news sites, and keywords are extracted. Keywords usually exist in news articles as metadata. Collected keywords are then used to retrieve additional tweets. Tweets matching keywords are then assigned to a relevance classifier, since keywords may include general or incorrect terms that can lead to the retrieval of large number of irrelevant tweets. This set of tweets is referred to as Keywords Tweets ($Tweets_{KW}$).

3. Classifying Tweets

An SVM classifier is trained with $Tweets_{KP}$, acting as the positive examples and a set of randomly selected tweets as negative examples ($Tweets_N$). $Tweets_N$ should not match neither the predefined queries nor the extracted keywords from news. This guarantees: $Tweets_N \cap (Tweets_{KW} \cup Tweets_{KP}) = \Phi$.

The number of negative examples is selected to be 10 times the positive examples since the spectrum of irrelevant tweets is expected to be much wider. Positive and negative examples are selected from the past 24 hours to be representing recent data. Of course training examples are not 100% accurate, since they are selected automatically. However, this technique of selection is more scalable and examples are generally correct.

The set of features used to train the SVM classifier are the terms appearing in $Tweets_{KP}$ more than 10 times. In addition, a feature represents the percentage of terms in a tweet that do not match any of these terms is used. Generated model is then used to classify $Tweets_{KW}$. Classified relevant tweets are then added to $Tweets_{KP}$ to form the full set of relevant tweets. Finally a comprehensive report is generated out of these tweets [3].

The process of training the classifier is applied each 15 minutes to keep the user updated with tweets relevant to news in real-time. Typically, the tweets classified as relevant enrich the total number of relevant tweets significantly; especially when accidental news occurs with new entities. Subjectively, the increase of relevant tweets ranges between 50% and 200% according to the type of news at that time, with noticeable high precision.

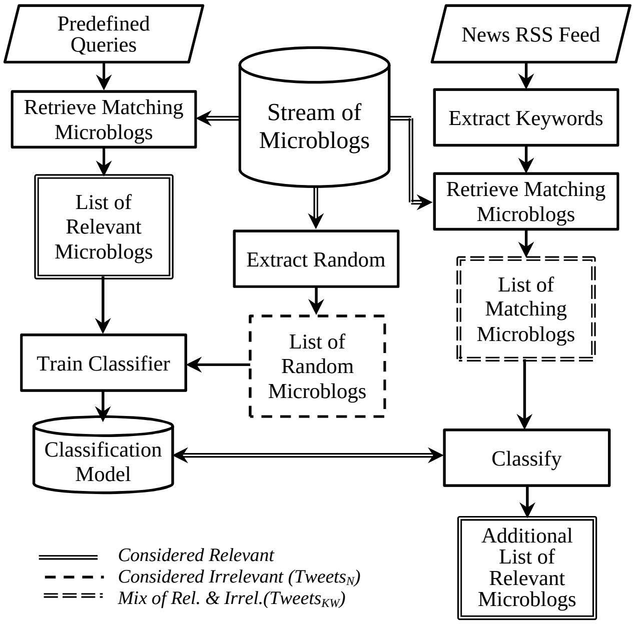
Predefined Queries
Retrieve Matching Microblogs
Stream of Microblogs
News RSS Feed
Extract Keywords
Retrieve Matching Microblogs
List of Relevant Microblogs
Extract Random
List of Matching Microblogs
Train Classifier
List of Random Microblogs
Classification Model
Classify
Additional List of Relevant Microblogs
Considered Relevant
Considered Irrelevant ($Tweets_N$)
Mix of Rel. & Irrel.($Tweets_{KW}$)

Figure 1 Filtering Approach for Relevant Tweets

4. TWEETMOGAZ DEMONSTRATION

TweetMogaz is running over a collection of Arabic tweets. It collects an average of 3-4 million Arabic tweets per day since May 2012. Tweets text is pre-processed using state-of-the-art normalization techniques for social Arabic text [1] and indexed using Solr. Search is enabled by specifying a query and time span, and results are presented in a comprehensive report similar to work in [3]. The homepage of TweetMogaz displays a real-time report about political news in the past 24 hours for countries in the Arabic region including Egypt and Syria. Tens of thousands of tweets are identified daily as relevant to each region using the presented filtering approach. The number of relevant tweets can reach up to 300k tweets on days with hot news, such as the Egyptian presidential elections day, and the days when there are severe battles between the Syrian free army and the regime's army. User can see examples of these days by browsing archived daily reports on the website.

5. REFERENCES

1. K. Darwish, W. Magdy, A. Mourad (2012). Language Processing for Arabic Microblog Retrieval. *CIKM 2012*
2. A. Kuthari, W. Magdy, K. Darwish, A. Mourad, A. Taei. (2013). Detecting Comments on News Articles in Microblogs. *ICWSM 2013*
3. W. Magdy, A. Ali, K. Darwish (2012). A Summarization Tool for Time-Sensitive Social Media. *CIKM 2012*
4. N. Naveed, T. Gottron, J. Kunegis, A. Alhadi. (2011). Searching microblogs: coping with sparsity and document quality. *CIKM 2011*.
5. O. Phelan, K. McCarthy, M. Bennett, and B. Smyth. (2011). Terms of a feather: content-based news recommendation and discovery using twitter. *ECIR 2011*.
6. I. Soboroff, I. Ounis, J. Lin, I. Soboroff. (2012). Overview of the TREC-2012 Microblog Track. *TREC 2012*
7. I. Subasic, B. Berendt. (2011). Peddling or Creating? Investigating the Role of Twitter in News Reporting. *ECIR-2011*
8. J. Teevan, D. Ramage, M. Morris. (2011). #Twittersearch: A comparison of microblog search and web search. *WSDM 2011*.
9. S. R. Yerva, Z. Miklós, F. Grosan, A. Tandrau, K. Aberer. (2012). TweetSpector: Entity-based retrieval of Tweets. *SIGIR 2012*