Real-Time Data Harvesting Method for Czech Twitter

Pavel Král[1,2], Václav Rajtmajer[1]
[1]*Dept. of Computer Science & Engineering*
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
[2]*NTIS - New Technologies for the Information Society*
Faculty of Applied Sciences
University of West Bohemia
Plzeň, Czech Republic
{*pkral, rajtmajv*}*@kiv.zcu.cz*

Keywords: Czech, Data, Harvesting, Social Media, Twitter

Abstract: This paper deals with automatic analysis of Czech social media. The main goal is to propose an approach to harvest interesting messages from Twitter in Czech language with high download speed. This method uses user lists to discover potentially interesting tweets to download. It is motivated by the fact that only about 20% of Twitter users are posting informative messages, whereas the remaining 80% not and that it is possible to identify the "important" users by the user lists. The experimental results show that the proposed method is very efficient because it harvests about 6 times more data than the other approaches. This approach should be integrated into an experimental system for the Czech News Agency to monitor the current data-flow on Twitter, download messages in real-time, analyze them and extract relevant events.

1 INTRODUCTION

Social media are virtual computer networks that allow individuals, companies, and other organizations to create, share, view and analyze information mainly in the form of short messages. The importance and the size of the today's social media are growing very rapidly which is strictly related to the particular needs of the automatic processing methods.

Twitter is a social net which uses very short messages limited by 140 characters. They are posted online as status updates, so-called *tweets*. The tweets can be accompanied by photos, videos, geolocation, links to other users (words preceded by the sign @) and trending topics (words preceded by the sign #). The posted tweet can be liked, commented by the other tweets, or redistributed by other users by forwarding, so-called *retweet*. Due to its simplicity and easy access, Twitter contains a very wide range of topics from common every day conversations over sport news to news about an ongoing disasters as earthquake, flood or typhoon. Twitter is without doubt a very interesting source of on-line information which can be used for further analysis and data-mining. In this work, we focus on Twitter because of its large size, significant amount of other existing work about this network and particularly because of a number of Twitter users post interesting news from various topics in real-time.

We would like to use Twitter for automatic real-time event detection because it will be very useful for many journals and news agencies in order to discover very quickly new interesting information. Particularly, the Czech News Agency (ČTK[1]) requires a system to automatically harvest data from Czech Twitter and to discover potential events. Several definitions of events exist, however we will use the definition from a Cambridge Dictionary. An event is defined as "anything that happens, especially something important and unusual[2]".

The first main task of this system consists in analyzing of Twitter stream and in harvesting of the appropriate tweets in Czech language in real-time. The second important task is the subsequent analysis of the downloaded data and to discover in such data new events. The main goal of this paper is to propose and implement a novel method to solve the first

[1]http://www.ctk.eu/
[2]http://dictionary.cambridge.org/dictionary/british/event

task described above. Note, that the activity of the Czech Twitter users is significantly lower than of the other ones, which is particularly evident for English or French. Therefore, it is not possible to use common methods provided by Twitter API and a novel method is necessary. The core of the proposed method consists in using user lists to download a sufficient number of Czech tweets in real-time.

The rest of the paper is organized as follows. Section 2 is a short review of Twitter analysis methods. The following section presents an architecture of the whole event detection system. Section 4 describes individual components of this system. The proposed method for tweet harvesting in the Czech language with high speed is presented in Section 4.1.3. Section 5 deals with the results of our experiments. In the last section, we conclude the experimental results and propose some future research directions.

2 SHORT REVIEW OF TWITTER ANALYSIS METHODS

Numerous studies have investigated Twitter, because it offers many possibilities for data processing and analysis. This social net can be used as a data source of sentiment analysis and opinion mining as shown for example in (Pak and Paroubek, 2010). The authors have collected a sentiment analysis corpus from Twitter and they have further built an efficient sentiment classifier on this data. Another work dealing with sentiment analysis from Twitter is proposed in (Kouloumpis et al., 2011). The authors show here the importance of linguistic features for this task.

Twitter data can be further used for sociological surveys as shown for instance in (Yardi and Boyd, 2010). The authors have analyzed a group polarization using the data collected from dynamic debates. Another study analyzes Twitter community (Java et al., 2009) to discover user activities. A taxonomy characterizing the underlying intentions of the users is presented.

Twitter can be also successfully used for event detection as presented for instance in (Sakaki et al., 2010; Earle et al., 2012). These approaches are generally based on the capturing of a presence or an increase of particular key-words. For instance, an increase of the words "earthquake" or "typhoon" is used for disaster detection.

They were also proposed some more sophisticated Twitter analysis approaches as for instance in (Li et al., 2012). The authors propose a system called *Twevent*, which first detects event segments and then, they are clustered considering both their frequency distribution and content similarity to discover events. Wikipedia is used as a knowledge base to derive the most interesting segments to describe the identified events and to discover realistic events. The main advantage of this system from the previous ones is that it is domain independent and therefore, it can identify all event types. The further event detection techniques on Twitter are available in the survey (Atefeh and Khreich, 2015).

Twitter analysis methods are focused particularly on English (sometimes also on French or on Chinese) and relatively few works are oriented to the other languages. Twitter activities of the users in such languages are very high and therefore the common harvesting methods provided by Twitter API are sufficient to get a sufficient amount of the data for a further analysis. We assume, that this fact explains that, to the best of our knowledge, no special Twitter harvesting method exists. Therefore, we will evaluate and compare our proposed harvesting method with the standard ones provided by Twitter.

It is also worth of noting, that no other study about automatic event detection on Czech Twitter exists.

3 SYSTEM DESCRIPTION

In order to show the whole problem, we first describe a general architecture of the event detection system and then, we detail the proposed method for fast harvesting of the Twitter data in Czech language.

The event detection system is composed of three main functional units (*Tweet Stream Analysis*, *Preprocessing* and *Event Detection*) which are further decomposed into six tasks as depicted in Figure 1.

The first task, *Data acquisition*, is beneficial to harvest on-line appropriate data from Twitter in Czech language with high speed. Then, *Spam filtering* is done to remove tweets with useless information (so called "spam"). The third task is *Lemmatization* which is used for word normalization. The next step is *Non-significant word filtering*. While the previous filtering was at the tweet level, this one is at word level and is used to remove non-significant words which could decrease the detection performance. The next step to discover events is *Clustering*. We group together the tweets with similar content using a clustering method. The final decision about an event is based on the thresholding. The last step, *Results representation*, is used to show the detected event to the users in an acceptable form.

All these steps are described below in details with the particular focus on the data acquisition, which is the main contribution of this paper.

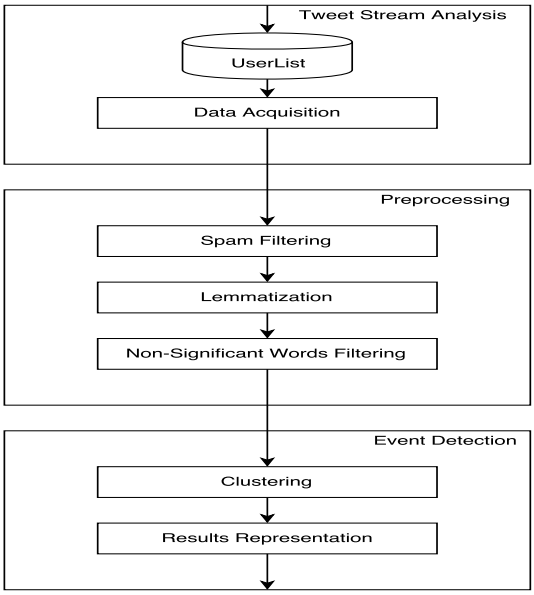

Figure 1: System architecture

4 METHOD DESCRIPTION

4.1 Data Acquisition

We summarize first our requirements to choose an optimal data acquisition method:

- working in real-time;
- downloading of the messages in Czech language;
- harvesting of a "sufficient" number of tweets for a further processing (it means, from our point of view, as much as possible);
- usage for free;
- downloading only informative messages (optional).

We analyze in the following text the different possibilities of Twitter for data harvesting. We show for all the methods the maximum download speed defined by the Twitter constraints. Unfortunately, this speed usually does not correspond to the real one, because the activity of the Twitter users is not sufficient to fill these limits.

4.1.1 Search API

This API is a part of Twitter REST API. It allows queries against the indices of recent or popular tweets and behaves similarly to, but not exactly like the search feature available in web clients. This API searches against a sampling of "recent" tweets published in the past 7 days and its maximum download speed is 72,000 tweets/hour. The query can be restricted by several constraints as for instance by a geolocation or by a target language.

Another important property is that this API is focused on relevance and not on completeness. This means that some tweets and users may be missing from the search results. The first approach, which is further evaluate and compare, uses this API and is hereafter referred as *Search API* method.

4.1.2 Streaming API

This API is intended to monitor (or process) tweets in real-time. Three different streams with three different connection types exist, however only *Public stream* can be suitable for our task. It allows to get public data from different users about different topics, while the other two ones (*User* or *Site*) analyze only the data from specific users.

From the point of view the connection type, we can use only *Filter* connection, because *Sample* provides a sample from all the data and *Firehose* which provides all possible data, is not free of charge. The query can be, as in the case of the *Search API*, restricted by several constraints (e.g. geolocation or target language). The maximum download speed of this method is unfortunately not given. The second evaluated approach uses this API and is hereafter entitled as *Filtered Streaming API* method.

4.1.3 UserList

Design of this method is motivated by the two following facts:

- our preliminary studies have shown that the methods provided by Twitter API are not very suitable for our task;
- about 20% of Twitter users are posting informative tweets, whereas the remaining 80% not (Naaman et al., 2010).

UserList is a Twitter possibility to allow each user to create 20 lists with an option to store up to 5,000 users into one list. These lists can be used to show all tweets that these users have posted and this procedure can be used with Twitter API to get all published data from 100,000 particular users.

The proposed method uses these list for acquisition of the significant amount of tweets in a given language (in Czech in our case, however the method is general enough to handle the other ones). The downloaded messages should contain valuable information for data-mining and further analysis as for instance potentials events.

Our issue is now to select the representative users in order to detect appropriate tweets. Our system is designed for general event detection. Therefore it must cover the all Twitter topics by active authors from all fields. We use a small sample of interesting people provided by Czech News Agency and this sample is automatically extended by our algorithm.

The algorithm to complete the UserList is based on the assumption that:

- We have already a representative group of the users (sample provided by ČTK);
- this set covers a representative part of our domain of interest;
- their followers would be the users with similar interests.

Therefore, we get by the Twitter API detailed information about all the followers of our initial group. Then, we filter out all foreign (no Czech) users and we continue with the first step. Our algorithm is stopped when a requested number of the users is explored.

For every user u, it is then computed a rank R_u which is based on its number of followers Fn and the number of submitted tweets Tn as follows:

$$R_u = w.Fn + (1-w).Tn \quad (1)$$

where w is the importance of both criteria and was set experimentally to 0.5.

Our list is sorted by this rank and the "best" 100,000 users are added to our twitter lists for a further processing.

Twitter "ecosystem" is very dynamic and it evolves very quickly. Therefore, this list must be periodically updated to keep actual information.

Our proposed method then harvests the data from this representative list of the 100,000 users via Twitter API. This method is hereafter referred as *UserList* method. It is also worth of noting that this method is language independent.

4.2 Pre-processing

4.2.1 Spam Filtering

As already stated, this task is realized in order to remove tweets with useless information. These tweets are filtered with a manually defined set of rules (or with a list of entire tweets). Table 1 shows some examples of whole tweets. The rules are based on the predefined patterns.

Of course, this simple method does not filter all useless tweets. However, we assume that they will not be detected as events by our detection algorithm due to their not significant amount. Therefore, it is not necessary for the current system to implement more sophisticated filtering algorithm.

Table 1: Examples of tweets to filter

Tweet	**English translation**
Automatically created messages	
Přidal jsem novou fotku na Facebook.	I have added a new photo on Facebook.
Líbí se mi video @YouTube.	I like @YouTube movie.
Označil(-a) jsem video @YouTube.	I have marked @YouTube movie.
(Everyday) useless tweets created by the users	
Dobré ráno!	Good morning!
Jdu obědvat, dobrou chuť.	I'm going to have lunch, enjoy your meal.

4.2.2 Lemmatization

Lemmatization consists in replacing a particular (inflected) word form by its lemma (base form). It decreases the number of features of the system and is successfully used in many natural language processing tasks. We assume that lemmatization can improve the detection performance of our method. It can be useful particularly in clustering to group together appropriate words.

Following the definition from the Prague Dependency Treebank (PDT) 2.0 (Zeman et al., 2014) project, we use only the first part of the lemma. This is a unique identifier of the lexical item (e.g. infinitive for a verb), possibly followed by a digit to disambiguate different lemmas with the same base forms. For instance, the Czech word "třeba", having the identical lemma, can signify *necessary* or *for example* depending on the context. This is in the PDT notation differentiated by two lemmas: "třeba-1" and "treba-2". The second part containing additional information about the lemma, such as semantic or derivational information, is not taken into account in this work.

4.2.3 Non-Significant Word Filtering

Non-significant words (also sometimes called stop words) are considered words with high frequencies which have in a sentence rather grammatical meaning as for instance prepositions or conjunctions. In this version, the filtering is based on a manually defined list. We plan to implement more sophisticated method based on Part-of-Speech (POS) tags in the further version. However, we assume that this improved removal will play marginal role for event detection.

4.3 Event Detection

4.3.1 Clustering

After getting the data we are facing the problem of extracting events. We use a clustering technique for this purpose. Consider that we get in real-time the filtered and lemmatized tweets which can represent (due to the UserList method) very probably the events. We transform every tweet into a binary representation using a bag of words method, which represents its unique location in n-dimensional space. Then the clustering algorithm is as follows:

1. take an (unprocessed) tweet;
2. calculate the cosine distance between a vector representing this tweet and all the others;
3. choose a closest tweet (or cluster of tweets if any) and group them together (the maximum allowed distance is given by the *threshold* Th);
4. repeat the two previous operations (*go to step 1*) till all tweets are processed.

The clusters created by this algorithm represent the events. Of course, the clustering does not guarantee that the created clusters represent only the events. This should be done by the pre-processing:

- UseList data acquisition method harvests particularly informative tweets which contains mainly the events;
- Spam filtering step removes several useless tweets (no events).

We also define a parameter T which indicates a time period for the clustering. We assume that different events will be produced at different "speed" (different activities of Twitter users). For instance, information about the winner of the football championship can be quicker (more contributions in a short period) than information about a new director of some company.

It is worth of noting, that we have also considered a *gradient* of the frequencies in some event clusters. Unfortunately, this improvement did not work because of the small activity of the users on Czech Twitter.

4.3.2 Results Representation

The results of the clustering are thus the groups of tweets with some common words. This group is represented by the *most significant* tweet. This tweet is defined as a message with the maximum of common words and the minimum of the other words. This representation is used due to the effort to use an answer in natural language, instead of a list of key-words or a phrase.

5 EXPERIMENTAL RESULTS

This section describes the experiments realized to evaluate the proposed tweet harvesting method based on user lists. The global functionality of the proposed event detection system is also evaluated here. This evaluation was done off-line.

5.1 Evaluation of the Data Acquisition

5.1.1 Comparison of the Czech and French Twitter Activity

In the first experiment, we would like confirm our claim that the activity of the Czech Twitter is significantly lower than in the case of the other languages. We have chosen French Twitter and *Search API* method (see Section 4.1.1) for such comparison.

First, we have discovered that, it is not possible to use language constraints to obtain only the Czech tweets. Unfortunately, the Czech constraint is missing and there is available only "sk" field which contains Czech and Slovak tweets together.

Therefore, we have decided to filter tweets according to geolocation. We have chosen a square region, covering most of the territories of the Czech Republic and France, as our area of interest. We have analyzed the download rate in interval from 22 to 29 August 2015. Figure 2 shows the results of this analysis.

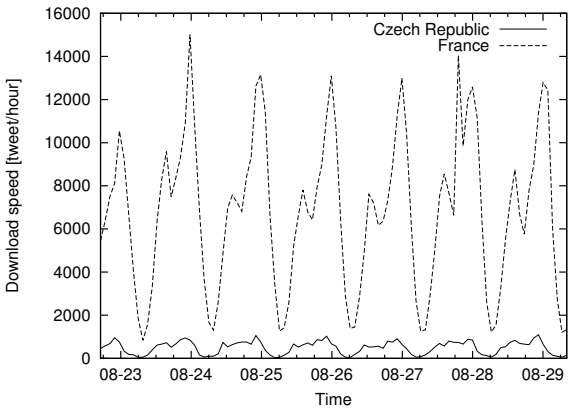

Figure 2: Comparison of the Czech and French Twitter activity

This figure shows that the activity of French Twitter is more than 10 $\times$ higher than the Czech Twitter. The average of the Czech download rate is about 495 tweets/hour. However, after a detailed examination, we have identified that only less than 20% of tweets are written in Czech languages.

Unfortunately, this number is insufficient for a successful further analysis as for instance for event

detection in real-time. Therefore, we must analyze the other approaches for data acquisition.

5.1.2 Comparison of the Different Data Acquisition Methods

In this experiment, we compare the download speed of two standard methods provided by the Twitter API (namely *Search API* and *Filtered Streaming API* methods - see Sections 4.1.1 and 4.1.2, respectively) with the proposed *UserList* approach (see Sec. 4.1.3). We have thus executed all these methods in the same two day period and then we have calculated the average value for one hour.

Table 2: Comparison of the download speed of the different methods on the Czech Twitter

Method	Tweets no. / hour
Search API	43.5
Filtered Streaming API	56.6
UserList *(proposed)*	330.3

The results of this experiment are shown in Table 2. This table shows that the proposed method provides about 6 times more data than the standard methods provided by Twitter API. Based on these results we have chosen the *UserList* approach to integrate into our event detection system.

5.2 Event Detection

We have used 15,856 tweets downloaded by *UserList* approach to evaluate the detection performance of our system. We have executed the event detection algorithm with different values of the acceptance threshold ($Th \in [0;1]$) and analyzed the results. The analysis of the resulting clusters has shown that for results with $Th > 0.5$ the algorithm still detects the majority of events correctly (high *precision*). However, the main interest is to have the *recall* as high as possible. The *precision* is not so important, because of the possibility of manual filtering of incorrectly detected events. Therefore, we set in our system a slightly lower acceptance threshold which causes to detect more events with some false positives.

These preliminary results were shown and discussed with our client who is ready to test this experimental version of the system. It is clear that the current version will already help to the reporters to reduce their work with manual checking of the available data sources.

One sample of the results is depicted in Figure 3. This figure shows that six tweets are saved by our acquisition method (right rectangle). They are then clustered into two groups containing three and two tweets (left "bubbles"). Finally, one representative tweet is chosen from both clusters to be presented to the user (bold text left).

6 CONCLUSIONS AND PERSPECTIVES

The main goal of this paper was to propose an approach to harvest messages from Twitter in Czech language with high download speed. The proposed method uses user lists to discover potentially interesting tweets to harvest. We have experimentally shown that the proposed method is very efficient because it harvests about 6 times more data than the two other approaches provided by the Twitter API. This method will be integrated into our event detection system. We have also experimentally shown that the results of the event detection are promising because the algorithm detects a significant amount of potential events.

The proposed harvesting method is language independent. Therefore, the first perspective consists in evaluation of this method on other (particularly European) languages. Another perspective is a thorough evaluation of the event detection method. We would like also improve this method using more sophisticated semantic similarity functions. Another perspective is an adaptation and evaluation of the whole detection system to the other languages.

ACKNOWLEDGEMENTS

This work has been partly supported by the project LO1506 of the Czech Ministry of Education, Youth and Sports and by Grant No. SGS-2016-018 Data and Software Engineering for Advanced Applications.

REFERENCES

Atefeh, F. and Khreich, W. (2015). A survey of techniques for event detection in Twitter. *Computational Intelligence*, 31(1):132–164.

Earle, P. S., Bowden, D. C., and Guy, M. (2012). Twitter earthquake detection: earthquake monitoring in a social world. *Annals of Geophysics*, 54(6).

Java, A., Song, X., Finin, T., and Tseng, B. (2009). Why we Twitter: An analysis of a microblogging community. In *Advances in Web Mining and Web Usage Analysis*, pages 118–138. Springer.

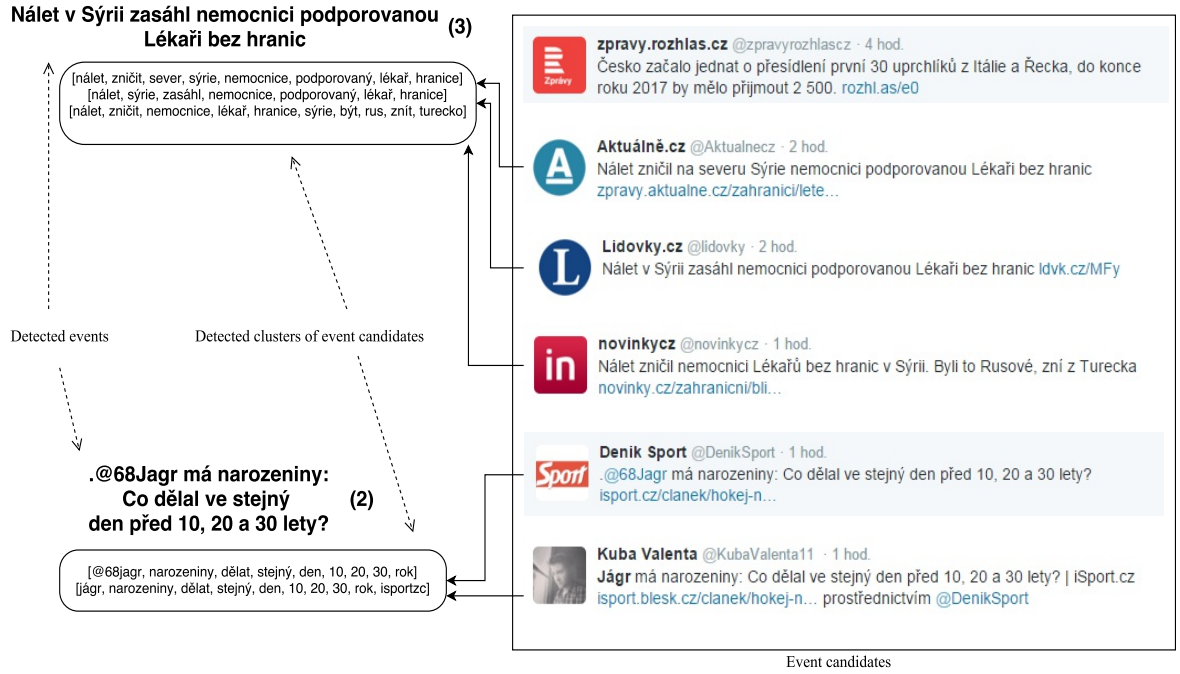

Figure 3: Event detection example (time period $T = 2h$ and acceptance threshold $Th = 0.5$). The rectangle on the right contains six tweets that were saved by our acquisition method. The left "bubbles" show the results of our clustering (two groups containing three and two tweets). The representative tweets are chosen (marked by the bold text on the left side) to be presented to the user.

Kouloumpis, E., Wilson, T., and Moore, J. D. (2011). Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11:538–541.

Li, C., Sun, A., and Datta, A. (2012). Twevent: segment-based event detection from tweets. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 155–164. ACM.

Naaman, M., Boase, J., and Lai, C.-H. (2010). Is it really about me?: message content in social awareness streams. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 189–192. ACM.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326.

Sakaki, T., Okazaki, M., and Matsuo, Y. (2010). Earthquake shakes Twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM.

Yardi, S. and Boyd, D. (2010). Dynamic debates: An analysis of group polarization over time on Twitter. *Bulletin of Science, Technology & Society*, 30(5):316–327.

Zeman, D., Dušek, O., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtskỳ, Z., and Hajič, J. (2014). Hamledt: Harmonized multilanguage dependency treebank. *Language Resources and Evaluation*, 48(4):601–637.