What are we tweeting about?

Providing Context for Twitter Analysis

Martin Ebner, Thomas Altmann
Social Learning
Graz University of Technology
Graz, Austria
Email: martin.ebner {at} tugraz.at

Abstract— **Twitter is a medium, which is primarily used for real-time communication. Due to the limitations of retrieving older tweets, archiving them is necessary to enable users to access and analyze old tweets. When analyzing tweet archives, more contexts can lead to better results. This research work aims to determine the value of context for an analysis of tweet archives. First of all the current state of the art of Twitter analysis research is discussed. Afterwards a tool called TweetCollector is introduced, which provides archiving capabilities. Additionally, a further tool for Twitter analysis called TwitterStat is developed. Finally a real-world use case is performed and discussed in depth. The research study points out that providing this context leads to better understanding of the analysis results.**

Keywords- ***Twitter; microblogging; social networks; analyses***

I. INTRODUCTION

Twitter is one of the most popular micro-blogging services in the world [1]. It created a whole new way of communicating. Twitter enables corporations, even countries and other large entities to communicate more directly with individual people or each other, and do so publicly. People can tap into global real-time communication during important events [2]. It is used to voice opinions and to discuss a broad spectrum of topics [3]. Some even gives Twitter credit in facilitating communication of protesters during the Arab Spring revolutions, and some governments now block Twitter as soon as signs of social unrest show themselves [4]. The relevance of this new form of social media is proven [3].

All of this makes Twitter an interesting target for analysis. Many researchers have already done extensive work on this topic [1] [5] [6] [7] [8] [9] [10]. Much of this research abstracts away from the original tweets. This leads to missing context necessary for certain conclusions. Therefore, the research question posed here is the following: What value can the context of a Twitter analysis provide and how can it be done?

To begin analyzing tweets (messages within the platform Twitter), a way to access and retrieve them is necessary. To provide a relevant basis, tweets that are about a similar topic should be taken into account. Usually this is achieved by using so called hashtags (word marked by a prefixed #) to tag tweets as belonging to a particular conversation, topic or event. For example for the conference ED-Media 2013 all participants agreed to use #edmedia13 in their messages. Now any Twitter user can search Twitter for such a keyword or hashtag, but the results are limited and not usable for automated analysis. Therefore we strongly propose that access to an Application Programming Interface (API) is needed. Twitter itself provides powerful APIs for developers to interact with. In general there are two different kinds of APIs: The REST API and the Streaming API.

The REST API enables a developer to make individual requests for sending or retrieving data to and from Twitter. This extends to virtually all interactions possible with Twitter: searching for tweets, following users, sending direct messages, fetching the timeline of a user, posting a tweet and much more [11]. This API is rate limited, so only a certain amount of requests can be made every 15 minutes [12].

The second endpoint Twitter provides is the Streaming API. This API relies on a single persistent connection to the client. Twitter then provides this client with a constant stream of tweets matching the parameters defined when the connection is established [13]. This second model is more complex, but has the benefit of providing real-time access to the stream of tweets.

To achieve analysis on a large scale, access to large amounts of old and current tweets is needed. Due to certain limitations of the Twitter API, this proves difficult when interacting directly with Twitter. The REST API for user timelines is limited to the most recent 3200 tweets of any given user [14] and for search it is limited to the most recent six to nine days of tweets [15]. Additionally, not the full set of tweets for this time period is returned. This leads to an incomplete data set when searching for all tweets containing certain words.

The only way to retrieve all tweets with a certain word or by a certain user is by using the Streaming API, but this necessitates that a client with an active connection to the Streaming API is running when the tweets are written. Therefore, a way to archive tweets is necessary.

II. STATE OF THE ART

Java et al. were among the first researchers to recognize the significance of Twitter. They studied topological and geographical properties of Twitter's social network [1]. In "A

Few Chirps About Twitter", Krishnamurthy conducted similar research [16].

"Social Networks That Matter" examined the relationship between the "declared" network of friends and followers, and a smaller hidden network of real connections that drives the usage of social networks [17]. Zhao and Rossen examined Twitter as a tool for informal communication at work [18]. In "Twitter Power", Jansen et al. examine the role of Twitter as electronic word-of-mouth in relation to brands, and what influence Twitter can have on these brands [3]. Honeycutt and Herring researched how Twitter can be used for collaborative purposes [5]. They did this by looking at the "@" sign as a marker of addressivity and the coherence of exchanges in the noisy environment of Twitter. boyd analyzed the practice of retweeting and how authorship and attribution are handled in this context [7]. Cha tried measuring user influence in Twitter [19]. Using a large dataset of tweets, they compared three different metrics: indegree (number of followers), retweets and mentions. Sentiment analysis and opinion mining on Twitter has been researched by Pak and Paroubek [20]. Kelly et al. write about using TwapperKeeper for Twitter archiving [21]. They discuss the limitations of the Twitter API and the need for an archiving service. In "What is Twitter, a Social Network or a News Media", Kwak studies the topological characteristics and information diffusion of Twitter using quantitative analysis [22]. In "Towards More Systematic Twitter Analysis", Bruns and Stieglitz propose standardized metrics for measuring tweeting activities [23]. Ebner wrote a work detailing the influence of Twitter on the academic environment [10].

Different research was done about possible uses for Twitter in disaster scenarios [24] [25] [26]. Twitter can also be used for making predictions about elections or the stock market [27] [28].

This overview of available literature on the topics of Twitter archiving and analysis shows some similarity between the approaches. To do effective analysis, crawling, retrieval and storage of large amounts of tweets is simply needed. Concerning Twitter analysis, researchers take the approach to separate the individual words of tweets to build ranked lists [29]. This kind of analysis shows interesting results, but most research stops at "most active users" and "most used words/hashtags" [21]. Further lists can be created by refining the analysis. Additionally, when the other forms of analysis like stock market, election and earthquake prediction are considered, one can see that the context of tweets is very important to gain deeper insight. This context is lost when ranked lists are created.

III. Implementation

To improve the current state of Twitter archival and analysis tool, two applications has been developed, called TweetCollector and TwitterStat. TweetCollector serves as a tool that communicates directly with the Twitter API and provides archiving capabilities. TwitterStat uses the tweet archives from TweetCollectors API and performs analysis with the supplied data. It provides an API of its own, enabling further applications. A whole tree structure of applications can be developed this way, with all of them relying on TweetCollector as the root or basis. This is shown in figure 1.

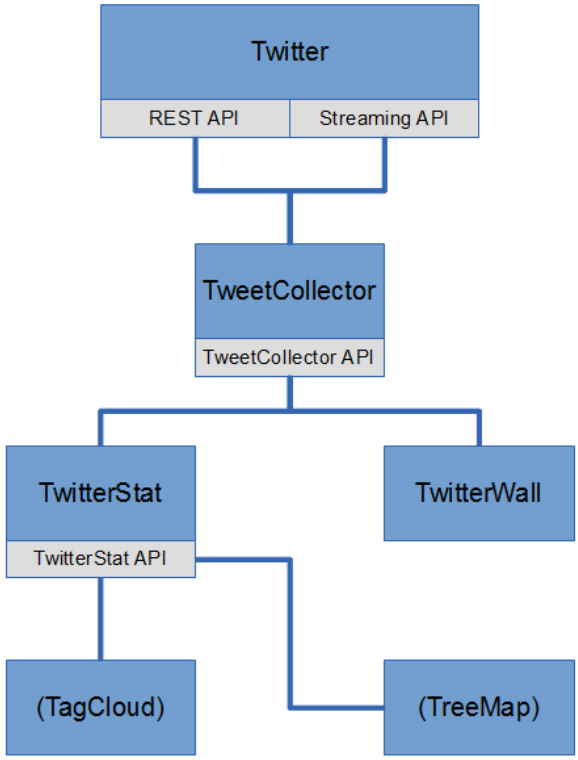

Figure 1. Tree structure of applications using TweetCollector.

A. *TweetCollector*

TweetCollector is the basis of the analysis tools introduced in this paper. It interfaces directly with the Twitter API to collect tweets containing certain words and hashtags or from certain users. These tweets are afterwards stored in tweet archives which can be accessed through a web interface or through a REST API.

Some preconditions need to be met for TweetCollector to work. TweetCollector uses UNIX command line tools to start, stop and manage the archiving processes. Therefore, it requires an operating system that provides access to these tools. TweetCollector has been tested on Debian 7 and Ubuntu 12.04. Running the software on Apple OS X should be possible as well due to the common UNIX heritage. Furthermore a webserver is needed to run TweetCollector. Apache2 was used for development and deployment. In Debian-based operating systems, this is the package "apache2".

TweetCollector uses PHP for server-side processing. It has been tested with PHP versions 5.4 and 5.5. The modules for cURL and PHP command line interface are needed as well. In Debian-based operating systems, the required packages are "php5", "php5-curl" and "php5-cli".

MySQL is used as the database management system. TweetCollector has been tested with MySQL versions 5.5 and 5.6. In Debian-based operating systems, this is the package "mysql-server".

Four processes are responsible for collecting and storing tweets. The first two are called “Crawl Users” and “Crawl Archives”. These two processes are very similar; the only difference is the Twitter API endpoint they retrieve data from. The user process communicates with "statuses/user_timeline", while the keyword/hashtag process interacts with "search/tweets".

This happens in three layered loops.

- Loop over all archives TweetCollector works with.
- Loop over pages of results. The search API provides 100 tweets at a time, while the user API provides 200. If less than the maximum amount of tweets is returned, this means the API is exhausted for this run and the algorithm moves on to the next archive.
- Loop over each individual retrieved tweet. If the tweet fits the parameters and is not yet in the database, it is stored. For user archives, the algorithm also stops looking at older tweets as soon as a tweet already stored in the database is found.

This approach minimizes the computations needed to process the tweets, but it still takes a significant amount of time. Due to rate limiting of the Twitter API, a new request can only be made every 5 seconds. Depending on the number of archives to crawl and the number of search results returned by the Twitter API, this can quickly lead to long pauses until a specific archive is crawled again. Missed tweets can be a result.

To mitigate this fact, the second type of tweet retrieval mechanism in TweetCollector employs the Twitter Streaming API. “Stream Collect” and “Stream Insert” are the processes responsible for this second type of archiving. “Stream Collect” provides an easy communication layer with the Twitter Streaming API. The function "enqueueStatus" is called every time a tweet fitting the specified search terms is received. As this happens often, the execution of this function should take minimal time. Therefore, every tweet is stored in a database table called "rawstream". This table is used by the second streaming process "Stream Insert".

The function "checkFilterPredicates" is called every 30 seconds. This makes it ideal to use "setTrack" and "setFollow" here. These two functions are used to tell the Streaming API which search terms and user names apply to the tweets it should retrieve. The process "Stream Insert" periodically checks the "rawstream" table for new tweets and sorts them into the right tables for each archive.

TweetCollector provides three different APIs:

- "info.php" accepts a "screen_name", "user_id", "keyword" or "id" parameter. Depending on the given parameter, it returns information about a user archive or a keyword/hashtag archive. This information includes the number of tweets in the archive, and whether or not crawling for this archive is active at the moment.
- "list.php" does not accept any parameters. This API simply returns a list of all archives in TweetCollector.
- "tweets.php" accepts "screen_name", "user_id", "keyword" or "id" as a parameter to specify which archive to retrieve tweets from. Additionally, a start and end date can be set. This enables a user to get all tweets from an archive, or just a subset from a specific date range.

B. *TwitterStat*

TwitterStat relies on the archiving function of TweetCollector (as shown in Fig. 1) and provides analysis of these archives.

The core principle of TwitterStat is rather simple: Take the text of each tweet, dissect it into separate words and count how often those words appear in all tweets in the examined archive. This provides the user with a basic understanding of what general topics are discussed in the tweets as well as at the event. This general principle can be applied to more data points in a tweet archive.

TwitterStat requires a webserver and PHP. There is no strict dependence on a specific operating system or type of webserver. PHP should be at least version 5.4. No database software is needed. If it is run on the same server as TweetCollector, all requirements are fulfilled because TweetCollector has more stringent needs than TwitterStat.

TwitterStat provides an API for most of its functionality. Some of the APIs mirror the functionality of the TweetCollector API (list, info), some extend the functionality of TweetCollector (tweets), and some provide data unique to TwitterStat (analyze).

"list.php" provides a list of all archives that are available for analysis, while "info.php" returns information about a single specified archive.

"tweets.php" returns the tweets of a specific archive. These tweets are retrieved from the TweetCollector API, so all the parameters it supports are present as well:

- "archive" defines the archive from which the tweets are to be retrieved.
- "start" defines an optional start date to retrieve only a specific subset of tweets.
- "end" defines an optional end date to retrieve only a specific subset of tweets.

Additionally, "tweets.php" from the TwitterStat API can filter these tweets using various parameters to get a very specific subset. Several more optional parameters are supported for this purpose:

- "from" defines tweets from a specified username.
- "mention1" and "mention2" define tweets where one or two specified usernames are mentioned.
- "word1" and "word2" define tweets where one or two specific words or hashtags are mentioned.

- "rt" denotes tweets that are retweets.
- "links" denotes tweets that contain hyperlinks.
- "safelinks" denotes tweets that contain hyperlinks with encryption (HTTPS).
- "source" defines tweets written with a specified Twitter client.

Only tweets that meet the exact specification are returned. This is used for links on the analysis page which lead back to the analyzed tweets.

"analyze.php" is the centerpiece of TwitterStat. It accepts 4 parameters:

- "archive" defines the tweet archive to be analyzed.
- "parameter" defines an optional parameter to make an analysis more specific.
- "start" defines an optional start date to analyze only a specific subset of tweets.
- "end" defines an optional end date to analyze only a specific subset of tweets.

The software parses these parameters and gets the required tweets from "tweets.php", which in turn retrieves them from the TweetCollector API.

The parameter "archive" is required to define which archive has to be analyzed. Start date and end date are optional because they only limit the scope of tweets that are analyzed. The biggest changes in the result of an analysis are created by the optional second parameter that makes the analysis more specific. Depending on the type of archive and the presence of the second parameter, there are six different kinds of analysis:

- Analysis of a keyword/hashtag archive with no second parameter
- Analysis of a keyword/hashtag archive with keyword/hashtag parameter
- Analysis of a keyword/hashtag archive with user parameter
- Analysis of a user archive with no second parameter
- Analysis of a user archive with keyword/hashtag parameter
- Analysis of a user archive with user parameter

Depending on the type of analysis, different answers are provided.

Some of the results are the same no matter what type of analysis is performed:

- Description of the type of analysis (e.g. This is the analysis of the archive "#test" with the parameter "lecture".)
- Number of tweets in analyzed archive (e.g. There are 469 tweets in this archive.)
- Number and percentage of retweets in the analyzed archive (e.g. There are 163 retweets in this archive (34.75% of all tweets).)
- List of Twitter clients used to write tweets in the analyzed archive. (e.g. what clients are used to write tweets in the archive #test)

The other results depend on the type of analysis performed. Figure 2 shows an overview of the answers provided for each combination of archive type and parameter type.

	Keyword/Hashtag Archive	User Archive
No Parameter	Which persons write about #example? Which keywords are used with #example? Which hashtags are used with #example? Which links are used with #example?	Who does @user talk to? Which keywords are used by @user? Which hashtags are used by @user? Which links are used by @user?
Keyword / Hashtag Parameter	Which persons write #example together with #parameter? Which keywords are used with #example and #parameter? Which hashtags are used with #example and #parameter? Which links are used with #example and #parameter?	Who does @user talk to about #example? Which keywords are used by @user with #example? Which hashtags are used by @user with #example? Which links are used by @user with #example?
User Parameter	Which persons talk to @user about #example? Who does @user talk to about #example? Who else is addressed with @user about #example? Which keywords are used by @user about #example? Which hashtags are used by @user about #example? Which links are used by @user about #example?	Who does @user address together with @person? Which keywords does @user use when talking to @person? Which hashtags does @user use when talking to @person? Which links does @user use when talking to @person?

Figure 2. Overview of provided answers

Each of the parameter-specific results is a list, sorted from the most used word, user or link to the least used. Some of these lists can get very long, especially the keyword list in large archives. Because of this, by default each list only shows the first 100 elements. A link is provided to show the remaining elements as well. An example of a full analysis is shown in figure 3.

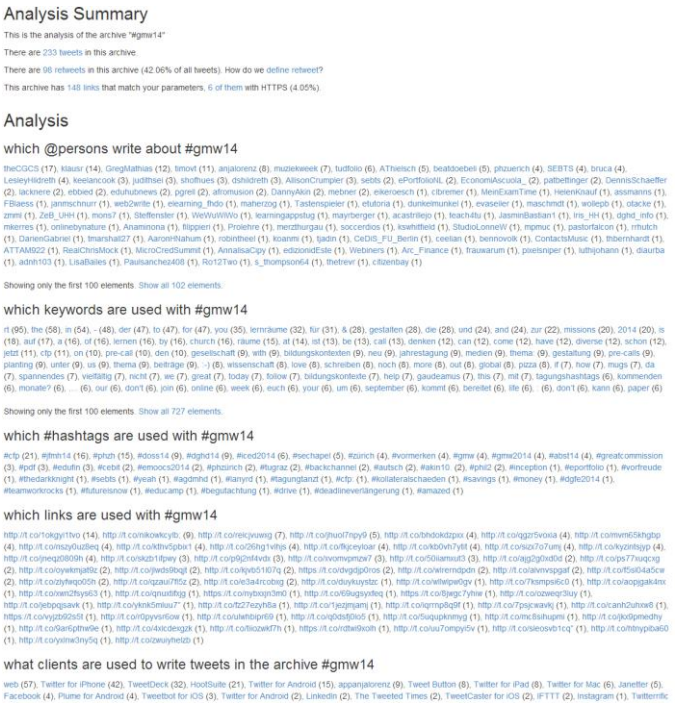

Figure 3. Example analysis of TwitterStat

IV. Use Cases and Discussion

The prime use case of TwitterStat is the analysis of archives from hashtags associated with conferences. Nowadays, most conferences designate a unique hashtag for attendees to use

when tweeting about the conference. Attendees may or may not adhere to this, but because the visibility of tweets is better if they are tagged properly, the incentive to use the hashtag is high.

A. *Analysis of EMOOCS 2014 Conference*

The conference "EMOOCs 2014" was the second European MOOCs Stakeholder Summit. It aims to be a meeting place of European participants in the Massive Open Online Course movement. Due to the nature of the conference, many attendants are interested in technology and are active Twitter users. The official hashtag of the conference is "#emoocs2014".

TweetCollector was able to capture 4359 tweets with this hashtag. The earliest tweet is from February 10th 2014, the latest from March 13th 2014.

For the purpose of this analysis all of these tweets are used. The results shown in this publication are shortened.

At first, a user can start with a general analysis with no second parameter:

analysis.html?archive=%23emoocs2014

The analysis shows that there are 2308 retweets in this archive (52.95% of all tweets). This is a very high percentage. It shows that many users found other tweets very interesting or informative and chose to retweet them to their personal followers or to confirm the tweet's importance.

The analysis also shows that there are 1976 links in the archive. There can be more than one link in a tweet, but if one assumes most tweets with links only contain one link, about 45% of all archived tweets contain links.

A user can view the actual tweets containing links:

tweets.html?archive=%23emoocs2014&links=true

This shows that 1915 tweets contain links (43.93% of all tweets), which proves that most tweets with links contain only one.

The analysis points out several lists:

- Which persons write about #emoocs2014?

  moocf (181), Agora_Sup (137), fuscia_info (130), pabloachard (124), mooc24 (118 , tkoscielniak (100), bobreuter (83), redasadki (79), ziebayves (78), yveszieba (78), crumphelen (75), OpenEduEU (75), DonaldClark (65), paigecuffe (63), anjalorenz (60), PeterMcAllister (59), diando70 (57), stollerschai (57), yprie (49), wfvanvalkenburg (49), celyagd (36), ...

- Which keywords are used with #emoocs2014?

  rt (2369), the (1408), of (1135), to (1113), a (1006), in (857), is (788), and (757), for (683), at (674), moocs (627), mooc (534), on (453), - (370), de (358), are (339), from (324), by (311), not (300), about (296), learning (286), with (275), : (269), la (257), data (226), you (219), be (196), open (196), it (190), des (189), i (183), as (181), les (175), we (169), education (161), le (159), that (157), will (152), an (149), pour (149), have (142), new (138), what (135), simon (134), coursera (133), & (127), track (122), nelson (121), more (118), …

- Which hashtags are used with #emoocs2014?

  #mooc (216), #moocs (201), #futurelearn (55),#vtecl (48), #heie (42), #bigdata (31), #epfl (28), #edtech (27), #itypa (26), #elearning (25), #oldsmooc (22), #edchat (20), #oldsmoop (20), #moocs? (20), #storify (19), #mooc: (18), #emoocs2015 (15), #video (14), #policytrack (14), #coursera (14), #moocs: (14), #emoocs2016 (12), #coer13 (12), #spoc (12), ...

- Which links are used with #emoocs2014?

  http://t.co/rhk4eptgkx (20), http://t.co/7cbp3vbuyv (14), http://t.co/o7yd6dnbq0 (13), http://t.co/qdp84oxukb (13), http://t.co/jv4antkfex (12), ...

- What clients are used to write tweets in the archive #emoocs2014?

  web (1545), Twitter for iPhone (641), TweetDeck (557), Twitter for Android (281), Twitter for iPad (275), HootSuite (192), Mobile Web (M5) (149), Twitter for Mac (123), Tweetbot for iOS (102), Tweetbot for Mac (74), appanjalorenz (58), Tweet Button (54), Twubs (36), iOS (26), Twitter for Windows Phone (24), Scoop.it (24), TweetCaster for Android (21), Buffer (16), ...

This wall of text can be intimidating at first, but a closer look reveals some interesting information.

The first list shows the most active users and provides a further basis for more focused analysis. One can also click on any of the user names to view the tweets this specific user wrote about the conference.

The second list shows the most used words. Because this contains all words that are not hashtags, common words are predominant at the top of the list. Recommendations for filters and other enhancements can be found in the chapter on further works. Nonetheless, some interesting words can be found in the list. "mooc" and "moocs" are present, which is not surprising in a conference dealing with them. Other interesting words are "data", "open", "learning", "education", "simon", "coursera", "track", "business", "european", "social" and others.

This provides a general overview of the topics discussed. If any of the words catches a user's attention, the tweets containing it are just a click away. If a user is interested in which Simon is mentioned, he or she can find the following tweets:

- @BenBrabon: Insightful talks at #emoocs2014 this week. Hear more on #MOOCs from Simon Nelson, Andrew Ng and David Willetts @HumMOOCs conference in May.

- @DonaldClark: #emoocs2014 @brianmmulligan asks great Q: Coursera & Futurelearn not open, but closed and elitist? Simon Nelson eeeh Yes

The user finds out that the Simon mentioned is Simon Nelson, the CEO of Futurelearn (http://www.emoocs2014.eu/speaker/simon-nelson).

The first tweet informs about a different conference about MOOCs. The second tweet describes a Q&A session, where Mr. Nelson seemingly answered a question about the openness of two popular MOOC platforms.

When looking at the rest of the tweets, the talk seems to have been rather controversial:

- @yprie: Not sure that Simon Nelson, as a media guy, is really interested in education, rather in mooc as new form of social media #emoocs2014
- @DonaldClark: #emoocs2014 Simon Nelson talks as if the web was an extension of Radio & TV – it was not, is not and never will be

The list of the most used hashtags shows that Futurelearn and Coursera are mentioned there as well, among other interesting tags. All of these can be explored further.

The list of most used Twitter clients shows a high usage of the Twitter website, as well as Twitter's official mobile clients for Android, iPhone and iPad. TweetDeck is in third place. TweetDeck is Twitter's client for power users, which shows that the people tweeting about this conference prefer more professional solutions for interacting with Twitter.

One can continue this analysis by digging deeper. At first, he/she can look at the tweets written by @yprie. The list of most active users shows that there are 49 tweets by this user. This should be sufficient for analysis.

analysis.html?archive=%23emoocs2014&parameter=@yprie

- Which #hashtags are used by @yprie about #emoocs2014?

  #moocs (2), #annotation (1), #edx (1), #moocs? (1), #vtecl (1), #colorscheme (1), #moocs: (1), #ocwcglobal (1), #mooc (1), #graz (1), #emoocs2015 (1), #emoocs2016 (1), #farfaraway (1), #louvain (1), #heie (1), #bigdata (1)

The most used words and hashtags show an overview of topics the user tweeted about. Universities seem to be an important topic for this user concerning MOOCs, because there are 6 tweets mentioning them. An example:

- @yprie: G.Fischer: identify respective contributions of online learning & core competencies of residential, research-based universities #emoocs2014

The most used hashtags mention "#emoocs2015" and "#emoocs2016", the two following conferences. When clicking through to the tweets, one can see that this is actually the same tweet:

- @yprie: RT @mebner: #Louvain will be hosting #emoocs2015 - afterwards I can invite you all to #Graz for #emoocs2016 #emoocs2014

For the second more detailed analysis, one can add the parameter "#futurelearn". 55 tweets contain "#emoocs2014" together with "#futurelearn".

analysis.html?archive=%23emoocs2014&parameter=%23futurelearn

- Which keywords are used with #emoocs2014 and #futurelearn?

  rt (28), new (24), findings (24), stats (24), & (22), the (18), to (17), of (17), is (13), at (12), moocs (12), simon (12), in (11), by (10), from (10), learning (8), course (8), lot (8), a (8), learn (8), and (6), partners (6), nelson (6), with (6), just (6), on (6), first (6), steps (6), cinema (6), like (6), conclusion (6), this (6), its (6), starts... (6), brilliant (6), storytelling (5), there (5), between (5), social (5), needs (4), learners (4), elearning (4), participation (4), can (4), an (4), analytics. (4), tv (4), complex (4), education (4), data (4), according (4), ...

- Which #hashtags are used with #emoocs2014 and #futurelearn?

  #mooc (19), #fb (7), #mooc: (3), #distancelearning (2), #bbc? (2), #bbc (2), #mlearning (2), #openuniversity (2), #simonnelson (2), #moocs (1), #edtech (1), #moo... (1), #elearning (1), #edchat (1), #unisouthampton (1)

- Which links are used with #emoocs2014 and #futurelearn?

  http://t.co/wle2fju9xn (5), http://t.co/xwi9xxurwq (3), http://t.co/klcqqj30vj (3), http://t.co/t5yrblngdb (2), http://t.co/xbzhaex9az (2), ...

The analysis results in the familiar list of items. The most used words show that there are tweets about storytelling, participation and cinema.

- @yveszieba: #emoocs2014 according to #FutureLearn, Education can learn a lot from complex tv storytelling, and Moocs an learn a lot with data analytics.
- @pbsloep: How open is #Futurelearn to participation of small universities? Not now. In the future? May be! #emoocs2014
- @bobreuter: Brilliant conclusion on MOOCs by Simon from #futurelearn at #emoocs2014 THIS IS JUST THE FIRST STEPS like cinema at its starts...

The most used hashtags lead to tweets with three or more hashtags:

- @LT_tech_HE: #emoocs2014 Simon Nelson announces #BBC collaboration with #futurelearn partners to develop WW1 courses @universityleeds @unileedsonline

- @bobreuter: eLearning needs social learning #futurelearn #fb newline #emoocs2014 to foster rich conversations between learners http://t.co/XWi9xxurwQ

To end this analysis, one can have a look at the most tweeted links. The first link in the list is from a tweet which has been retweeted four times:

- @mhawksey: #eMOOCs2014 #FutureLearn new stats & findings #MOOC http://t.co/wLe2fju9XN

After resolving the Twitter link shortening services, the link leads to a blog post.

http://ignatiawebs.blogspot.co.at/2014/02/emoocs2014-futurelearn-new-stats.html

The post contains a link to a YouTube video of the talk by Simon Nelson at EMOOCS2014.

https://www.youtube.com/watch?v=3NhAjy3Qs6k

After some time, a user can arrive at a video of a talk which our analysis suggested might arguably be one of the most important or controversial talks of the conference. Now he or she can watch the video and form an opinion on the content and see if it fits the conclusions drawn after this analysis.

B. Discussion

When surveying the available literature, it can be pointed out that many researchers have a similar approach to Twitter analysis. The idea to separate tweets into individual words and hashtags to create ranked lists is something that is simple but effective. This leads to the availability of many different tools capable of performing this sort of analysis.

What gets lost in all of these tools is the meaning of the original tweets where the counted words and hashtags are derived from. This context can be valuable to determine what tweets in a certain archive are really about. For example, if the most tweeted hashtag in an archive is "#keynote", this is interesting information. However, the sentiment and context of the tweets containing this hashtag are unknown. Was the keynote good or bad, or are they even talking about a real keynote or the presentation software from Apple?

To achieve this context, TwitterStat offers links in each of the analysis results presented. These links enable the user to follow the results back to the original tweets that led to these results. To continue the example, a user is also able to click on the link and see all tweets in the archive containing the hashtag "#keynote". From these tweets, the original meaning can be determined easily. The tweet list even offers links to view the tweets directly on the Twitter website. If any tweet is part of a larger conversation, the Twitter website can show the whole exchange and provide even more context.

The analysis results provided by TwitterStat can be used in applications that rely on its API. This is possible through the tree structure shown in figure 1. An example of such a use are treemaps that visualize the results in a more accessible way. Figure 4 shows a treemaps of most active users from a conference.

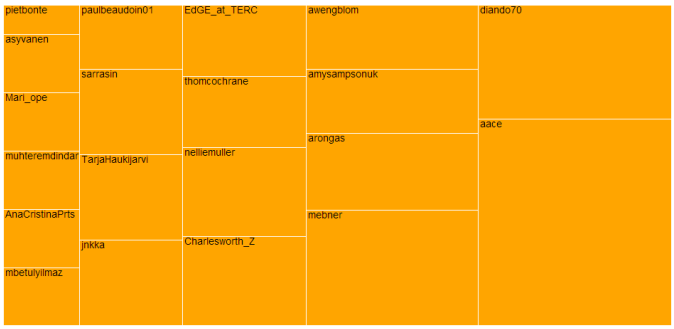

Figure 4. Treemap of most active users

Twitter analysis can provide valuable insight. However, if the abstraction is too far away from the original tweets, context can be lost. By providing a way to get back to the tweets, TwitterStat allows users to dig deep into the details of an archive analysis, but keep track of where the results came from.

To summarize, context can help to:

- Determine the content and sentiment of the original tweets.
- Check if the insights gained from the analysis correspond with the original tweets.
- See tweets as part of a larger conversation.

V. CONCLUSION

The aim of this research study is to show the value of providing context for Twitter analysis. To achieve this, several topics were explored. The state of the art of current academic research on Twitter was surveyed. The research covers a wide variety of topics, from the usage of Twitter during conferences, lectures and academic writing, as well as during disasters such as earthquakes and other crisis events. There are publications on using Twitter to predict elections or the stock market.

A tweet archiving tool called TweetCollector has been developed. TweetCollector creates archives of tweets containing a certain word or hashtag, or from a certain user. The content of these archives is available through an API for other applications to use.

The Twitter analysis tool TwitterStat was introduced. TwitterStat analyzes an archive retrieved from TweetCollector, and shows the most active users and the most used words, hashtags and links in the archive. Depending on further parameters, even more detailed analysis results can be obtained. By clicking on the results, the user can get back to the original tweets.

TwitterStat was used to analyze tweets from a conference. Afterwards, these results were discussed. It was shown that TweetCollector provides value by having an open API that can be used to build application relying on its archives. The "back to tweets" feature of TwitterStat was shown to be valuable for determining context of the original tweets.

This research work points out that the data provided by Twitter itself is not sufficient for many applications. The retrieval and storage of data from Twitter is absolutely necessary to create persistent archives of tweets available for further usage. These generated tweet archives enable a variety of new future applications in the fields of analysis, filtering and visualization. By providing machine readable data through APIs in each stage, a whole tree structure of applications relying on each others data can be constructed. All of this is enabled by the archives.

Twitter is a medium that is becoming more relevant each day. As more and more interactions happen on this medium, analysis of this type of communication is getting increasingly important. The tools introduced in the scope of this paper can be valuable for a variety of users. Finally the defined research questions can be answered, that due to the introduced applications it becomes possible to use Twitter for deeper insights. Whether it is just the beginning, it points out clearly the potential for future studies.

VI. REFERENCES

[1] A. JAVA, X. SONG, T. FININ, B. TSENG. Why we twitter: understanding microblogging usage and communities. Presented at the Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (2007), 56–65.

[2] K. HAEWOON, L. CHANGHYUN, P. HOSUNG, P., S. MOON. What is Twitter: A social network or a news media? Proceedings of the 19th International World Wide Web (WWW) Conference, April 26-30, 2010, Raleigh NC (USA), April 2010.

[3] B. J. JANSEN, M. ZHANG, K. SOBEL, A. CHOWDURY. Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60-11 (2009), 2169–2188.

[4] G. LOTAN, E. GRAEFF, M. ANANNY, D. GAFFNEY, I. PEARCE, D. BOYD. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. International Journal of Communication, 5 (2011), 31.

[5] C. HONEYCUTT, S. C. HERRING. Beyond microblogging: Conversation and collaboration via Twitter. Presented at the Proceedings of the 42nd Hawaii International Conference on System Sciences, (2009), Hawaii.

[6] H. MÜHLBURGER, M. EBNER, B. TARAGHI. @twitter try out #Grabeeter to export, archive and search your tweets. Research 2.0 Approaches to TEL 2010, 76–85.

[7] D. BOYD, S. GOLDER, G. LOTAN. Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter. Presented at the Proceedings of the 43rd Hawaii International Conference on System Sciences, (2010), Hawaii.

[8] L. De VOCHT, S. SOFTIC, M. EBNER, H. MÜHLBURGER. Semantically driven social data aggregation interfaces for research 2.0. In 11th International Conference on Knowledge Management and Knowledge Technologies, 2011, 1-10

[9] P. THONHAUSER, S. SOFTIC, M. EBNER. Thought bubbles - A conceptual prototype for a Twitter based recommender system for research 2.0. iKnow 2012, Proceedings of the 12th Interna- tional Confrerence on Knowledge Management and Knowledge Technologies, 2012.

[10] M. EBNER. The influence of Twitter on the academic environment. Social Media and the New Academic Environment: Pedagogical Challenges. IGI Global (2013), 293–307.

[11] TWITTER, REST API v1.1 Resources, 2014a, retrieved 2014-05-13, https://dev.twitter.com/docs/api/1.1

[12] TWITTER, REST API Rate Limiting in v1.1, 2014b, retrieved 2014-05-13, https://dev.twitter.com/docs/rate-limiting/1.1

[13] TWITTER, The Streaming APIs, 2014c, retrieved 2014-05-13, https://dev.twitter.com/docs/streaming-apis

[14] TWITTER, GET statuses/user_timeline, 2014d, retrieved 2014-05-13, https://dev.twitter.com/docs/api/1.1/get/statuses/user_timeline

[15] TWITTER, Using the Twitter Search API, 2014e, retrieved 2014-05-13, https://dev.twitter.com/docs/using-search

[16] B. KRISHNAMURTHY, P. GILL, M. ARLITT. A few chirps about twitter. In Proceedings of the first workshop on Online social networks, ACM (2008), 19–24.

[17] B. A. HUBERMANN, D. M. ROMERO, F. WU. Social networks that matter: Twitter under the microscope. First Monday, 14/1-5 (2009).

[18] D. ZHAO, M. B. ROSSON. How and why people twitter: the role that micro-blogging plays in informal communication at work. In Proceedings of the ACM 2009 international conference on Supporting group work, ACM (2009), 243–252

[19] M. CHA, H. HADDADI, F. BENEVENUTO, P. K. GUMMADI. Measuring user influence in Twitter: The million follower fallacy. ICWSM, 10 (2010), 10–17.

[20] A. PAK, P. PAROUBEK. Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), 2010, Valletta, Malta.

[21] B. KELLY, M. HAWKSEY, J. O'BRIEN, M. GUY, M. ROWE. Twitter archiving using TwapperKeeper: technical and policy challenges. In 7th International Conference on Preservation of Digital Objects (iPRES 2010). University of Bath, 2010.

[22] H. KWAK, C. LEE, H. PARK, S. MOON. What is twitter, a social network or a news media? In Proceedings of the 19th international conference on World wide web, ACM (2010), 591–600.

[23] A. BRUNS, S. STIEGLITZ. Towards more systematic twitter analysis: Metrics for tweeting activities. International Journal of Social Research Methodology, 16-2 (2013), 91–108.

[24] S. VIEWEG, A. L. HUGHES, K. STARBIRD, L. PALEN. Microblogging during two natural hazards events: What Twitter may contribute to situational awareness. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2010), 1079–1088.

[25] T. TERPSTRA, A. DE VRIES, R. STRONKMAN, G. PARADIES. Towards a realtime twitter analysis during crises for operational crisis management. In ISCRAM'12: Proceedings of the 9th International ISCRAM Conference (2012), Vancouver, Canada.

[26] T. SAKAKI, M. OKAZAKI, Y. MATSUO. Earthquake shakes twitter users: real-time event detection by social sensors. In Proceedings of the 19th international conference on World wide web, ACM (2010), 851–860.

[27] J. BOLLEN, H. MAO, X. ZENG. Twitter mood predicts the stock market. Journal of Computational Science, 2-1 (2011), 1–8.

[28] A. TUMASJAN, T. O. SPRENGER, P. G. SANDNER, I. M. WELPE. Predicting elections with twitter: What 140 characters reveal about political sentiment. ICWSM, 10 (2010), 178–185.

[29] S. SOFTIC, S., M. EBNER, M., L. De VOCHT, E. MANNENS, R. Van De WALLE, A Framework Concept for Profiling Researchers on Twitter using the Web of Data. Proceedings of the 9th International Conference on Web Information Systems and Technologies (WEBIST) 2013, SciTePress 2013, Karl-Heinz Krempels, Alexander Stocker (Eds.), 447-452