Thomas Altmann

Potential of Twitter Archives

Master's Thesis

Graz University of Technology

Institute of Information Systems and Computer Media
Head: Prof. PhD Frank Kappe

Supervisor: Assoc. Prof. PhD Martin Ebner

Graz, April 2014

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, ________________________ ____________________________

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am ________________________ ____________________________

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Acknowledgements

I want to thank my family for supporting me over the whole course of my university studies.

I also want to thank my girlfriend for believing in me during the creation of this thesis.

Thank you to my supervisor Martin Ebner, for making this thesis possible and giving me a job during my master's program.

Thanks to my colleagues in the department of social learning for all the help and interesting discussions.

Last but not least, I want to thank all of my friends at university and everywhere else, who made life as a student just that much better.

Thank you!

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. Through these archives, users are able to access and analyze old tweets. In the course of this thesis, the value created through the archiving of tweets is to be determined.

When analyzing tweet archives, more context can lead to better results. Therefore, this work also tries to determine the advantage of context for an analysis of tweet archives.

In the course of this thesis, an exploration of the current state of the art of Twitter archival and analysis tools is conducted. Furthermore, current publications and research about these topics are discussed. Then a tool called TweetCollector is introduced, which builds on the foundation of the aforementioned tools and provides improved archiving capabilities. Additionally, two other tools for Twitter analysis and filtering are introduced: TwitterStat and TwitterWall.

To show the application of the aforementioned tools, several real-world use cases are performed and discussed. Concerning value of Twitter archives, it can be seen that archiving tweets is vital for any kind of further usage of tweet data. Regarding value of context for Twitter archive analysis, the research shows that providing this context leads to better understanding of the analysis results.

In addition to the analysis of existing tools and the introduction of TweetCollector, TwitterStat and TwitterWall, suggestions for the improvements of the developed tools are given. Now as well as in the future, the analysis of tweets provides an important starting point for the efficient usage of Twitter.

Kurzfassung

Twitter ist ein Medium, welches vorwiegend für Echtzeitkommunikation genutzt wird. Da der Zugang zu älteren Tweets eingeschränkt ist, ist eine Archivierung notwendig, um es Nutzern zu ermöglichen, alte Tweets aufzufinden und zu analysieren. In dieser Arbeit wird versucht, die durch die Archivierung von Tweets geschaffene Wertschöpfung zu bestimmen.

Bei der Analyse von Tweet-Archiven kann mehr Kontext oft zu besserem Verständnis von Analyseergebnissen führen. Aus diesem Grund wird des Weiteren der daraus entstehende Mehrwert für den User untersucht.

In dieser Arbeit wird eine Untersuchung des aktuellen Standes der Technik von Twitter Archivierungs- und Analysetools durchgeführt. Weiters werden gegenwärtige Arbeiten zum Thema besprochen. Danach wird das Tool TweetCollector vorgestellt, welches auf die bestehende Forschung als Grundlage aufbaut und verbesserte Archivierungsfunktionen zur Verfügung stellt. Im Zuge der Arbeit werden noch zwei weitere Tools für die Analyse und Filterung von Tweets vorgestellt: TwitterStat und TwitterWall.

Um die Anwendung der zuvor genannten Tools zeigen zu können, werden diese anhand von echten Anwendungsfällen behandelt und diskutiert. Zum Thema Wertschöpfung durch Twitter-Archive zeigt sich, dass eine Archivierung der Daten notwendig ist, um jegliche Art weiterer Verwendung von Tweet-Daten zu ermöglichen. Der Mehrwert durch den Kontext bei Twitter-Analysen entsteht durch ein besseres Verständnis der Analyseergebnisse.

Neben der Analyse von bestehenden Tools und der Einführung in die Tools TweetCollector, TwitterStat und TwitterWall werden zusätzlich auch Vorschläge für die Verbesserung der vorgestellten Tools gegeben. Auch in Zukunft wird die Analyse von Tweets einen wichtigen Ansatzpunkt für die effiziente Nutzung von Twitter darstellen.

Contents

List of Figures

Acronyms

AJAX Asynchronous JavaScript and XML

API Application Programming Interface

GUI Graphical User Interface

HTML HyperText Markup Language

HTTP Hypertext Transfer Protocol

HTTPS Hypertext Transfer Protocol over Secure Socket Layer

IT Information Technology

JSON JavaScript Object Notation

MOOC Massive Open Online Course

REST Representational State Transfer

RSS Rich Site Summary or Really Simple Syndication

SQL Structured Query Language

URL Uniform Resource Locator

XML Extensible Markup Language

1. Introduction

Twitter is one of the most popular micro-blogging services in the world [Java et al., 2007]. It created a whole new way of communicating. Twitter enables corporations, countries and 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. It is used to voice opinions and to discuss a broad spectrum of topics [Jansen et al., 2009]. Some even give 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 [Lotan et al., 2011]. The relevance of this new form of social media is proven [Jansen et al., 2009].

All of this makes Twitter an interesting target for analysis. Many researchers have already done extensive work on this topic [boyd et al., 2010, Ebner, 2013, Honeycutt and Herring, 2009, Java et al., 2007]. To achieve analysis on a large scale, access to large amounts of old and current tweets is needed. Due to certain limitations described in chapter 2, this proves difficult when interacting directly with Twitter. Therefore, a way to archive tweets is necessary.

This thesis introduces a tool to retrieve and store data from Twitter. The initial use of this data is for analysis. The availability of those archives enables various other use cases like filtering or visualization.

1.1. Research Questions

This thesis deals with the potential of Twitter archives. Twitter is mainly a real-time communication network, but for some types of analysis and usage, persistent archives of old tweets are necessary. This leads to the first question:

- What value can tweet archives provide?

Due to the interesting nature of Twitter, a large amount of analysis has already been conducted on the topic of Twitter. Much of this research abstracts away from the original tweets. This leads to missing context necessary for certain conclusions. The current thesis tries to provide a solution for that problem. Therefore, the second question is:

- What value can the context of an analysis provide?

1.2. Structure

Chapter 2 introduces the terms and definitions of microblogging, Twitter and tweets, as well as Twitter conventions like hashtags, mentions and retweets. The Twitter Application Programming Interface is introduced, and the problems concerning its limitations are described.

In Chapter 3, the state of the art for both scientific research on the topic of Twitter, as well as existing tools for Twitter archiving and analysis are detailed.

Chapter 4 describes TweetCollector, the tweet archiving tool created for this thesis.

Chapters 5 and 6 deal with TwitterStat and TwitterWall, respectively. These two tools build on TweetCollector to provide analysis and filtering of tweet archives.

Chapter 7 shows use cases this collection of tools can be applied to. These applications and their significance for the research questions are discussed in chapter 8.

In Chapter 9, the outlook and future works are reviewed. This includes possible improvements to the tools discussed in the previous chapters.

Chapter 10 contains the concluding remarks of this thesis.

2. Terms and Definitions

In this chapter, various terms and definitions used throughout this thesis, as well as the problem that needs to be solved to answer the research questions are explained.

2.1. Microblogging

Microblogging is a form of blogging that differentiates itself from regular blogging mainly through the type of content that is contained in a post. Microblogging ″allows users to exchange small elements of content such as short sentences, individual images, or video links″ [Kaplan and Haenlein, 2011].

Posts in microblogs are typically shorter than posts in regular weblogs, which leads to the term ″micropost″ to describe them. The shorter posts can occur voluntarily due to conventions of a certain blogging service, or due to an artificial limit on the length of posts (like a maximum character limit of 140). This leads to a blurry line between microblogs and regular blogs.

2.2. Twitter

Twitter is a microblogging service and social networking site launched in July 2006. Since then, it has become one of the most popular microblogging platforms worldwide. Twitter has 241 million monthly active users, and 500 million tweets are sent per day.[1]

[1] https://about.twitter.com/company, 2014-04-21

Twitter is also an asymmetrical social network. By default, tweets are public. If a user has not protected his or her tweets, other users can ″follow″ this user and subscribe to the tweets of the user without intervention by the followed user. This makes one user the ″follower″ while the user being followed is called the ″followee″.

A user with protected tweets has to explicitly grant others the right to see and subscribe to his or her tweets.[2]

2.3. Tweet

Posts on Twitter are called tweets. Tweets are limited to 140 characters. In addition, metadata about the tweet like author, creation date, language, location and client software is stored by Twitter. When a tweet contains a link to certain media sources, they are displayed.

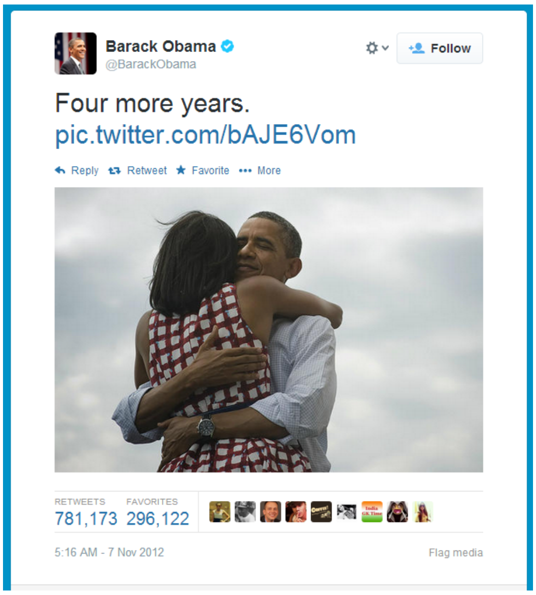
Tweets are shown to Twitter users in a reverse chronological timeline. Figure 2.1 shows a single tweet.

2.4. Retweet

A retweet is a syndication of a tweet from one user to the followers of a second user. If the second user wants to share the content of a tweet with his followers, he or she can retweet it.

This can be done by clicking the retweet button, which is the officially supported way and copies the original tweet into the timeline of another user. Prior to the existence of this function, retweeting was done by copying the text of the original tweet and mentioning the original author together with the letters ″RT″ (short for retweet). Different styles evolved, therefore making it difficult to precisely distinguish between tweets and retweets [boyd et al., 2010].

[2] https://support.twitter.com/articles/14016-about-public-and-protected-tweets, 2014-04-21



Barack Obama
@BarackObama

Four more years.
pic.twitter.com/bAJE6Vom

RETWEETS 781,173
FAVORITES 296,122

5:16 AM - 7 Nov 2012

Figure 2.1.: A tweet. Source: https://twitter.com/BarackObama/status/266031293945503744, 2014-04-21

2.5. Mention

Mentions are usernames of Twitter users in the text of a tweet, prepended with the symbol ″@″. Twitter automatically detects these mentions and provides links to the relevant profile pages. Mentions can serve various purposes, from notifying a user that you are talking about them to replying to other tweets and having a conversation [Honeycutt and Herring, 2009].

2.6. Hashtag

Hashtags are words prefixed with the symbol ″#″. They are used to tag tweets as belonging to a certain topic. A tweet can contain multiple hashtags. Similar to mentions, Twitter automatically detects hashtags and links them to searches for the clicked tag.[3]

2.7. Twitter Application Programming Interface

Twitter provides a powerful API[4] for developers to interact with. There are two different kinds of APIs: The REST[5] 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.[6]

This API is rate limited, so only a certain amount of requests can be made every 15 minutes.[7]

[3] https://support.twitter.com/articles/49309-using-hashtags-on-twitter, 2014-04-21

[4] Application Programming Interface

[5] Representational State Transfer: all requests are treated independently from each other

[6] https://dev.twitter.com/docs/api/1.1, 2014-04-21

[7] https://dev.twitter.com/docs/rate-limiting/1.1, 2014-04-21

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.[8]

This second model is more complex, but has the benefit of providing real-time access to the stream of tweets.

2.8. Problem Description

The mission statement of Twitter as a company is ”to give everyone the power to create and share ideas and information instantly, without barriers.”[9] This reflects in the fact that Twitter is an inherently transient medium. The most important tweets are the ones written right now. Nonetheless, every tweet is kept forever, unless someone deletes it. The problem is finding these old tweets.

If the URL[10] of a tweet is known, it is trivial to find it again. The following tweet by Barack Obama, shown in figure 2.1, serves as a good example:

`https://twitter.com/BarackObama/status/266031293945503744`

If the user who wrote a certain tweet is known, it is possible to visit the profile page of this user and scroll down the timeline until the desired tweet is found. This is a tedious process and can only be done by hand. The corresponding API is limited to the most recent 3200 tweets of any given user.[11] This prevents finding and retrieving tweets older than the most recent 3200 by a certain user automatically.

An exception to this occurs if access to the user account is available. In December 2012, Twitter enabled a feature where users can download their own tweets as an archive containing a local webpage and tweets in computer readable form.[12] This enables retrieval, storage and analysis of old tweets,

[8] https://dev.twitter.com/docs/streaming-apis, 2014-04-21

[9] https://about.twitter.com/company, 2014-04-21

[10] Uniform Resource Locator

[11] https://dev.twitter.com/docs/api/1.1/get/statuses/user, 2014-04-21

[12] https://blog.twitter.com/2012/your-twitter-archive, 2014-04-21

but only for user accounts with known access credentials. Retrieving all tweets from other users is still not possible.

Recently, Twitter introduced ”data grants” for a limited amount of research institutions.[13] This enables selected partners to have access to large datasets of tweets. The drawback is that a research institution needs to be accepted to get access to this program.

If only the tweet's content or part of it are known, it can be impossible to find it again. The tweet shown in figure 2.1 is one of the most popular tweets of all time, yet it does not show up in the search results when searching for ”Four More Years” on the Twitter website:

`https://twitter.com/search?q=four%20more%20years`

The Search API is limited to the most recent six to nine days of tweets.[14] Additionally, not the full set of tweets for this time period is returned. This leads to incomplete data 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. This necessitates that a client with an active connection to the Streaming API is running when the tweets are written.

To maximize the chances of archiving all tweets, a combination of the Search/User API and the Streaming API needs to be used.

These limitations show that retrieval and analysis of older tweets is only feasible if archiving of tweets is done at time of creation or shortly thereafter.

[13] https://blog.twitter.com/2014/introducing-twitter-data-grants, 2014-04-21

[14] https://dev.twitter.com/docs/using-search, 2014-04-21

3. State of the Art

Twitter introduced a new type of communication, which makes it a very interesting target for analysis. This chapter deals with some of the academic research done on the topic of Twitter, as well as some tools available to conduct research and analysis.

3.1. Scientific Research

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 [Java et al., 2007]. This included the growth and properties of the network, and the geographical distribution of the users.

In their analysis, they found different kinds of intentions each user has for using Twitter. They distinguished between four broad categories:

- Daily chatter
- Conversations
- Sharing information
- Reporting news

While most of the tweets they found were daily chatter, the other categories are more interesting. Conversations are tweets with mentioned user names after the @ symbol, while sharing information means tweets containing links. The tweets reporting the news show indicators of people using Twitter as a different type of personalized RSS[1] aggregator.

They also found three distinct types of users:

[1] Rich Site Summary, used to subscribe to frequently updated content

- Information source
- Friends
- Information seeker

Information sources have many followers and post frequent or valuable information, while information seekers post rarely and follow many users. Friends classifies people who use Twitter more like Facebook and follow their immediate offline social contacts.

In ″A Few Chirps About Twitter″, Krishnamurthy et al. conducted similar research [Krishnamurthy et al., 2008]. They also characterized Twitter users and proposed different classes:

- Broadcasters
- Acquaintances
- Miscreants and Evangelists

Broadcasters contain media organisations that publish their headlines via Twitter. The class of acquaintances is similar to the ″friends″ found by Java et al [Java et al., 2007]. Miscreants and evangelists share similar characteristics. They are users with few followers and many followees. With a negative intention, this can be seen as typical for spammers or stalkers, while the positive intention might be reaching as many people as possible, hoping to be followed back.

″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 [Huberman et al., 2008]. This is demonstrated on the example of Twitter.

Huberman et al. created the definition of a friend as a person that a user has directed at least two posts to using mentions. Even when the number of followees rises, the number of friends eventually saturates.

The resulting social network of friends as opposed to the declared network of followees is much more sparse, but also more relevant. The implication is that ″attention is the scarce resource in the age of the web″, and valuable insights can be gained by finding the real social networks users devote their attention to.

The work of Huberman et al. does not distinguish between any classes of Twitter users, but looks at the average. It shows that most Twitter users have a small core of friends that they interact with regularly, and a larger group of users that they follow because they are interested in their status updates. This is possible because Twitter is a more interest-based social network as opposed to Facebook, where symmetrical friendship connections are the norm.

Zhao and Rossen examined Twitter as a tool for informal communication at work [Zhao and Rosson, 2009]. They listed various benefits of informal communication, both relational (person perception, common ground, connectedness) and personal benefits in the form of valuable information for personal goals.

The method used was phone interviews with employees of a large IT[2] firm. They observed that people use Twitter for its content and technology features. Content features include ″frequent brief updates about personal life activities″, ″real-time information″ and ″people-based RSS feed″. Technology features are ″brevity″, ″mobility and pervasive access″ and the ″broadcast nature″ of Twitter.

The research showed that people valued Twitter for the positive effects on relational benefits and for ″work-relevant information sharing and expertise seeking″ concerning personal benefits. The study also showed issues with security and integration within the structures of a company.

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 [Jansen et al., 2009]. They examine various aspects of this: the trends, characteristics and patterns of brand microblogging.

A scale was developed to classify tweets about brands into 5 different sentiments: wretched, bad, so-so, swell and great. In addition, there is a category for tweets without sentiment. The researchers then used a tool called ″Summize″ to analyze the sentiment of tweets from their dataset. Summize has since been acquired by Twitter.[3]

[2] Information Technology

[3] http://techcrunch.com/2008/07/15/confirmed-twitter-acquires-summize-search-engine, 2014-04-21

Their findings indicate that people use microblogging to express and form opinions, and their recommendation for brands is to be present on such services to influence the discussion.

boyd et al. analyzed the practice of retweeting and how authorship and attribution are handled in this context [boyd et al., 2010]. The paper gives an overview of Twitter background and conventions, such as mentioning users with the "@" sign, assigning tweets to topics with hashtags, and retweeting tweets.

Due to the focus on retweeting, the paper discusses the various ways a retweet can be constructed. Twitter now has a dedicated retweet button for every tweet, but before this feature became available, a variety of ways have been developed by users to syndicate tweet content. The most popular one is prepending "RT @user:" to the content of a tweet. Other ways include mentioning the original user with the word "via", or adding additional comments to the content of the original tweet.

Twitter itself only recognizes a tweet as a retweet if the official way to retweet is used. All of this makes it difficult to determine exactly what constitutes a retweet.

The study researches how, why and what people retweet. This was done using a random sample of tweets captured with the Twitter API, as well as questions asked to the Twitter followers of one of the authors.

They found different practices of people who try to preserve the original tweet as much as possible and people who shorten or adapt the original tweet to have room to comment on it within the 140 character limit. There are users who retweet for others and users who retweet for social action like donations. The different reasons why people retweet are numerous.

The researchers also found out that people use retweets for conversations. In this usage and others, issues with authorship, attribution, missing context and missing content can emerge. It takes just one user who doesn't credit the original source to make finding it difficult. Similarly, shortened tweets or tweets stripped of some context due to character limitations may be misleading.

The study concludes that although retweeting has issues and different approaches, users embrace it.

Cha et al. tried measuring user influence in Twitter [Cha et al., 2010]. Using a large dataset of tweets, they compared 3 different metrics: indegree (number of followers), retweets and mentions. Their reasoning is that "indegree represents popularity of a user; retweets represent the content value of one's tweets; and mentions represent the name value of a user".

They found little overlap in the top users of each measure of influence. The most followed users were public figures and news outlets, while the most mentioned users were celebrities. The most retweets were achieved by tweets from content aggregation services, businessmen and news sites.

The research showed that popular users are not necessarily influential and that gaining influence requires a concerted effort. Becoming influential on Twitter requires dedication and commitment. This may make it possible to predict emerging influential users.

Kelly et al. write about using TwapperKeeper for Twitter archiving [Kelly et al., 2010]. This is the same service that has been used in earlier versions of the tools described in this thesis.

They discuss the limitations of the Twitter API and the need for an archiving service. After exploring the available options, they decided to fund the development of TwapperKeeper. The paper explains the technical, policy and sustainability issues concerning this project.

TwapperKeeper was used for archiving conference tweets from the International World Wide Web Conference 2010 in Raleigh, North Carolina, using the hashtag "#www2010".[4] Afterwards, the data from this TwapperKeeper archive was used with the service "Summarizr" to analyze it for data like most active users and to create tag clouds of most used words.

In "Towards More Systematic Twitter Analysis", Bruns and Stieglitz propose standardized metrics for measuring tweeting activities [Bruns and Stieglitz, 2013]. These include user metrics, temporal metrics and combined user/temporal metrics.

[4] http://www2010.org/www, 2014-04-21

Examples for user metrics are ″replies sent″ or ″mentions received″, while ″tweets per period of time″ is an example for temporal metrics. Combined metrics include constructs like ″currently active users from the most active one percent for each time period″. They show the application of this approach on tweets captured using yourTwapperKeeper, the open source[5] version of TwapperKeeper.

The paper claims that these standard metrics for analyzing hashtag archives provide better comparability between different datasets. They show this by comparing tweets tagged ″#tsunami″ with tweets tagged ″#royalwedding″, where the former has a higher percentage of retweets and tweets containing URLs. Comparing this to other hashtag archives results shows clustering of certain archives. This means that certain topics behave similar to some and distinct to others.

Sentiment analysis and opinion mining on Twitter has been researched by Pak and Paroubek [Pak and Paroubek, 2010]. They performed linguistic analysis and classified the sentiment of individual tweets.

To achieve this, they collected tweets with positive, negative and no emotions. Tweets containing the happy smiley face :-) were classified as containing positive sentiment, while tweets with the sad smiley face :-(were sorted into the opposite category. To get neutral tweets, they collected headlines posted among others by the New York Times and Washington Post Twitter accounts.

These tweets were used as training data for a sentiment classifier. Using machine learning algorithms, they were able to determine the sentiment of a tweet with high accuracy.

In ″What is Twitter, a Social Network or a News Media″, Kwak et al. study the topological characteristics and information diffusion of Twitter using quantitative analysis [Kwak et al., 2010].

Twitter users were ranked by number of followers, by the PageRank algorithm [Page et al., 1999] and by retweets. A very high correlation between number of followers and PageRank was discovered. The top list of retweets

[5] A computer program where the source code is available to everyone for use and modification

looked very different, suggesting that other factors than popularity play a role in motivating people to retweet.

Kwak et al. also looked at trending topics on Twitter, specifically the retweeting of trends, the participation in trends and the active period of trends. They found out that the majority of trending topics are persistent news stories.

To research the information diffusion on Twitter, retweets were analyzed concerning their audience and when they happen. They found that any retweet has an average audience of 1000 users, signifying very fast information diffusion.

Honeycutt and Herring researched how Twitter can be used for collaborative purposes [Honeycutt and Herring, 2009]. They did this by looking at the ″@″ sign as a marker of addressivity and the coherence of exchanges in the noisy environment of Twitter.

They found evidence that 90 percent of tweets with the ″@″ sign directly addressed other users. Tweets containing no ″@″ mostly fell into the category of posts answering the question asked by Twitter: ″What are you doing?″

Regarding coherence, they found that a surprisingly high amount of over 30 percent of tweets addressed to someone else received a public response within half an hour. In their sample, most conversations spanned two persons and three to five messages over a period of 15 to 30 minutes. Most of these messages used the ″@″ sign.

This shows that by using proper addressing, coherence of longer conversations can be achieved even in the noisy environment of public messages that is Twitter.

Twitter also has possible uses in disaster scenarios. In ″Earthquake Shakes Twitter Users″, Sakati et al. developed a system that can detect where earthquakes are happening based on the tweets of affected Twitter users, and warn people [Sakaki et al., 2010]. They developed this system in Japan, which is uniquely suited to this because of a high number of Twitter users as well as occurring earthquakes.

They asked if they can detect earthquakes by real-time monitoring of tweets. Each Twitter user is assumed to be a sensor, while each tweet represents sensory information. These ”social sensors” are very varied: some are very active, others are not. A sensor may be inactive if the user is sleeping or busy.

The researchers determine that their social sensors are very noisy. To mitigate this, they use semantic analysis with machine learning. Only tweets with a location (either from the tweet itself, or from the profile information of the user) are assumed to be relevant.

In 2009, this prototype was employed during a real earthquake and typhoon and got very accurate results. This led to the development of an earthquake reporting system called ”torreter”, which in most cases of earthquakes notifies users even before the Japan Meteorological Agency.

Terpstra et al. conducted similar research on the example of a storm incident in the Belgian town Kiewit [Terpstra et al., 2012]. They postulate that ”utilizing Twitter's potential for operational crisis management [...] requires information extraction tools that digest the information content in realtime, and in a reliable fashion.”

In 2011, the Pukkelpop pop festival in Belgium was hit by a storm. After the event, the researchers analyzed and visualized tweets about the disaster with a tool called ”Twitcident”.[6]

They could identify warnings before the storm. During the disaster, the volume of tweets per minute increased significantly. The topic of the tweets was mostly damage and casualty reports. After the storm, they found tweets for citizen initiatives to initiate disaster relief.

In their conclusion, the researchers recommended visualization of tweet volume and location to facilitate crisis management.

A third important work of research concerning Twitter and crisis situations was done by Vieweg et al by collecting tweets during two natural disasters: The Oklahoma grass fires of April 2009 (5 days worth of tweets) and the Red River floods in March/April 2009 (51 days of tweets). With this data, they attempted to identify information that contributes to enhancing situational awareness.[Vieweg et al., 2010]

[6] http://twitcident.com, 2014-04-21

At first they looked at the geolocation information, and found that 78 percent of Twitter users from the Oklahoma dataset and 86 percent from the Red River dataset wrote at least one tweet with location information. This suggests that users find this information useful in such situations.

The content of the tweets mostly consists of situational updates. The researchers developed a framework that suggests the design of a system for information extraction from such tweets.

There has also been research about Twitter as a tool for making predictions. Tumasjan et al. looked at Twitter in the context of the 2009 parliament elections in Germany [Tumasjan et al., 2010]. Their approach was split into three parts.

First they asked if Twitter is a suitable vehicle for online political deliberation, and found that Twitter is used as a forum for such discussions, but dominated by a small number of very active users. Their second question was whether Twitter messages reflect the current offline political sentiment, and they found that to be true.

The third question was if Twitter chatter before the election can be used to predict the outcome of the election and the coalitions formed afterwards. In this case this was true, since the percentages of Twitter mentions of the parties were very close to the actual votes received and the coalition formed. This was true despite the fact that Twitter users are not a representative sample of all German voters.

Bollen et al. tried to use Twitter to predict the stock market [Bollen et al., 2011]. They defined seven mood dimensions and monitored tweets to sort them into these categories. Concurrently, the Dow Jones Industrial Index was monitored.

Their research showed that the most influential mood dimension was "calmness". Changes on this dimension correlated with changes in the stock index three to four days later. They concluded that the calmness of the public is more predictive than positive or negative sentiment. It is acknowledged that this shows correlation but no causation.

Ebner et al. conducted a variety of research on the role of microblogging in the academic environment. One of the first works of Ebner and Schiefner introduces microblogging as a form of mobile learning [Ebner and Schiefner, 2008]. The researchers created a group dedicated to ″elearning″ on the microblogging platform Jaiku. They found that the most interesting contributions to this group were microposts from conferences. This went as far as using microblogging as a back-channel to pose questions to keynote speakers, where questions from posts were answered after the presentation.

Further research in this direction was undertaken in ″Introducing Live Microblogging″ [Ebner, 2009]. The question posed in this work was if microblogging can enhance a live event. This was tested during the ED-MEDIA 2008 conference. Participants were invited to participate using the hashtag ″#edmedia08″. During a keynote, the tool ″Twemes″ was used on screen next to the slides of the presentation, to help everyone follow the Twitter conversation. The study found four distinct types of tweets during this presentation: concerning the presentation, discussion, links and comments.

Ebner and Maurer applied microblogging to a lecture at Graz University of Technology [Ebner and Maurer, 2009]. In the course ″Social Aspects of Information Technology″, students were split into four groups. Each group had a different task in order to get a grade for the lecture: writing a scientific paper, reviewing a scientific paper, writing blog posts or writing microblog postings. The evaluation of this approach showed that students in the blogger and microblogger groups experienced positive effects: They wrote about their topics for a longer period of time and in more detail. The discussion with the microblogging group led to more personal opinions and reflection on the topics discussed.

A similar experiment was undertaken at a University of Applied Sciences in Upper Austria [Ebner et al., 2010]. The microblogging platform ″identi.ca″ was used in tandem with MediaWiki to give students a platform for posting. The study found that microblogging can have advantages for informal learning as well as process-oriented learning.

The paper concludes that ″microblogging can help users to be partially and virtually present and to be part of a murmuring community, that is working on a specific problem without any restrictions of time and place.″

Two different papers analyzed the Twitter community of the ED-MEDIA 2009 conference [Reinhardt et al., 2009, Ebner and Reinhardt, 2009]. The approaches were different, with one study conducting an online survey of participants to get qualitative answers, while the other used the tools "twitterVisBT" and "Yahoo Term Extraction Web Service" to achieve quantitative results. They survey provided reasons why people use Twitter during conferences: exchange of resources and social activities, documentation, announcements, feedback, comments and discussion. The quantitative tools visualized the most active users and most used keywords and hashtags of the conference.

Mühlburger et al. developed a tool called "Grabeeter" to archive tweets from Twitter users [Mühlburger et al., 2010]. Grabeeter was a combination of a web app that managed the creating and archiving of tweets, and a desktop client that could download these archives to the local storage of a computer.

The predecessor of the tools described in this thesis was called "STAT" and was able to archive and analyze hashtag and keyword archives, as well as person archives like Grabeeter. Softic et al. used Grabeeter and STAT to conduct a semantic analysis of Twitter archives [Softic et al., 2010]. In "Twitter Analysis of #edmedia10", Ebner et al. provide a more comprehensive overview of how STAT is used to enable analysis of tweets from scientific conferences [Ebner et al., 2011].

In 2013, Ebner wrote a work detailing the influence of Twitter on the academic environment [Ebner, 2013]. The paper references much of the work described in the last few paragraphs and gives an overview of the different ways Twitter can be applied to learning, universities and scientific conferences. Preconditions to achieve a microblogging community are listed: "mobility", since many people use Twitter from mobile devices; "communication", since microblogging is a short and efficient way to stay in contact; and "collection", since using hashtags enables storing tweets. The paper names semantic analysis of social networks as a further direction of research.

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 needed. This was achieved in various ways by tracking person archives with Grabeeter and hashtag archives with yourTwapperKeeper [Kelly et al., 2010, Mühlburger et al., 2010]. However, none of these tools provided both options.

When looking at the research on Twitter analysis, many papers take the approach to separate the individual words of tweets to build ranked lists. This kind of analysis shows good results, but most research stops at ”most active users” and ”most used words/hashtags”. 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.

3.2. Existing Tools

Due to the increased interest in Twitter, many tools and websites that can analyze and filter various aspects of Twitter have emerged. This section enumerates some of them and details their abilities.

3.2.1. TwapperKeeper and yourTwapperKeeper

The first version of the tools described in this work was primarily concerned with the analysis of tweets. This led to the discovery of the limitations of the Twitter API concerning older tweets. A website called TwapperKeeper offered a service where archives of tweets from a certain user or containing a certain word or hashtag could be created.[7] TwapperKeeper had to shut down in March 2011.[8] The ability to export tweets in this form is a violation of the Twitter API terms of service.[9]

[7] http://twapperkeeper.com/index.html, 2014-04-21

[8] http://chronicle.com/blogs/profhacker/the-end-of-twapperkeeper-and-what-to-do-about-it, 2014-04-21

[9] https://dev.twitter.com/terms/api-terms, 2014-04-21

This led to the release of the archiving tool as open source in the form of yourTwapperKeeper. [10] Hosting and using this code was still in violation of the API terms of service, but the small scale and distribution across more users made enforcement of those rules unnecessary and difficult. yourTwapperKeeper removed the ability to archive tweets from certain users, allowing only keyword and hashtag archives. Figure 3.1 shows yourTwapperKeeper archiving the hashtag "#twitter".

The developer of TwapperKeeper eventually joined HootSuite, which develops a social media management suite by the same name.[11] A feature of this tool called HootSuite Archives provides similar archiving capabilities.[12]

3.2.2. Tweet Archivist

Tweet Archivist is a Twitter archival and analysis service. Users can create tweet archives of a certain word or hashtag in advance of events. The service can analyze the archive and provide lists of the top users, words and links. Figure 3.2 shows a screenshot of Tweet Archivist.[13]

One disadvantage of Tweet Archivist is that it is a paid service. Other shortfalls are the lack of user archives and the lack of real-time updates. Because this is a consumer-facing product, there are no APIs available to use this dataset or extend functionality.

The provided analysis is less comprehensive than those of the tools described in this work, but Tweet Archivist also has some advantages. It provides visualization of top tweeted images and better analysis of top tweeted links.

[10] https://github.com/540co/yourTwapperKeeper, 2014-04-21

[11] https://hootsuite.com, 2014-04-21

[12] https://help.hootsuite.com/entries/21840213-Creating-Tweet-Archives, 2014-04-21

[13] https://www.tweetarchivist.com, 2014-04-21

3.2.3. twXplorer

twXplorer is a tool developed by the Northwestern University Knight Lab.[14] It is shown in figure 3.3

The tool provides analysis similar to Tweet Archivist: Most used words, hashtags and links. The difference is that there is no archiving service. A user can specify a search term, and the service just analyzes the last 500 tweets retrieved when searching Twitter for this term. A snapshot of this analysis can be stored for later viewing.

The lack of any archiving keeps the analysis of tweets very limited. A larger amount than 500 tweets would be necessary to gain deeper insight.

3.2.4. TWUBS

TWUBS is a tweet archiving service for hashtags.[15] After registering a hashtag, a visually rich page for this hashtag is created. It shows most recent tweets as well as the most recently tweeted pictures. There is no analysis function or API to retrieve raw data. TWUBS is shown in figure 3.4.

3.2.5. TweetDeck

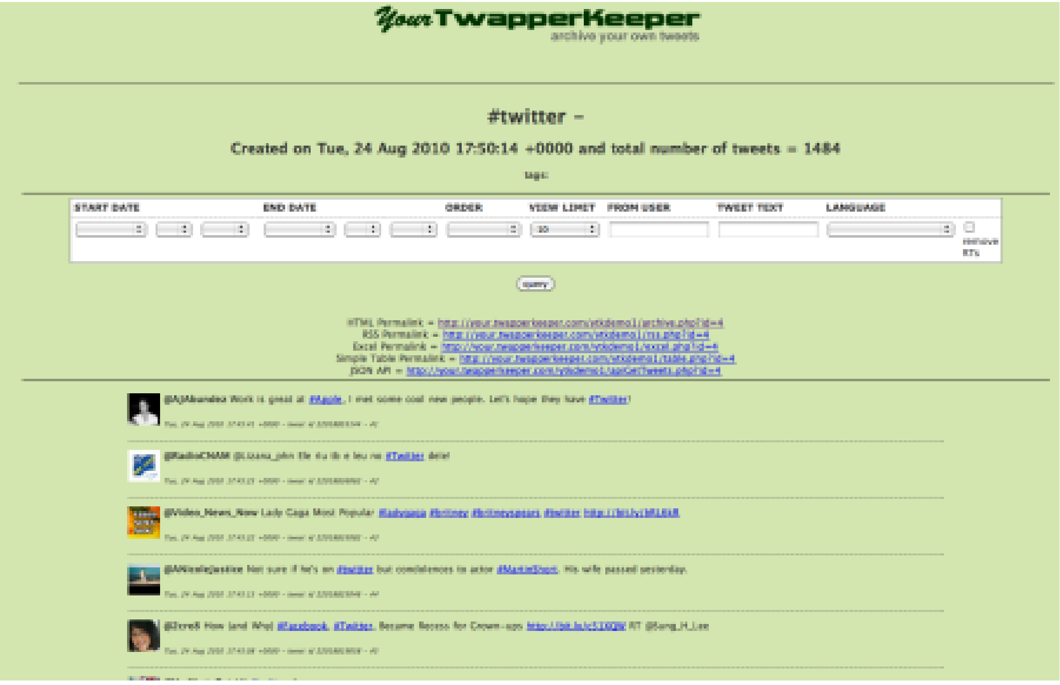
TweetDeck is a tool by Twitter for more professional real-time tracking, organizing and engagement.[16] A user can enter the credentials for multiple Twitter accounts and monitor the activities for all of them on a single page. It is also possible to create columns for search results, thereby tracking activity for certain hashtags. The product description page of TweetDeck is shown in figure 3.5.

TweetDeck shares many similarities with the tool "TwitterWall" described in this thesis. A more detailed overview of the similarities and differences is available in chapter 6.

[14] http://twxplorer.knightlab.com, 2014-04-21

[15] http://twubs.com, 2014-04-21

[16] https://about.twitter.com/products/tweetdeck, 2014-04-21



#twitter –

Created on Tue, 24 Aug 2010 17:50:14 +0000 and total number of tweets = 1484

tags:

START DATE			END DATE			ORDER	VIEW LIMIT	FROM USER	TWEET TEXT	LANGUAGE	
							10				□ remove RTs

HTML Permalink = http://your.twapperkeeper.com/ytkdemo1/archive.php?id=4
RSS Permalink = http://your.twapperkeeper.com/ytkdemo1/rss.php?id=4
Excel Permalink = http://your.twapperkeeper.com/ytkdemo1/excel.php?id=4
Simple Table Permalink = http://your.twapperkeeper.com/ytkdemo1/table.php?id=4
JSON API = http://your.twapperkeeper.com/ytkdemo1/apiGetTweets.php?id=4

@AJAbundez Work is great at #Apple. I met some cool new people. Let's hope they have #Twitter!
[illegible]

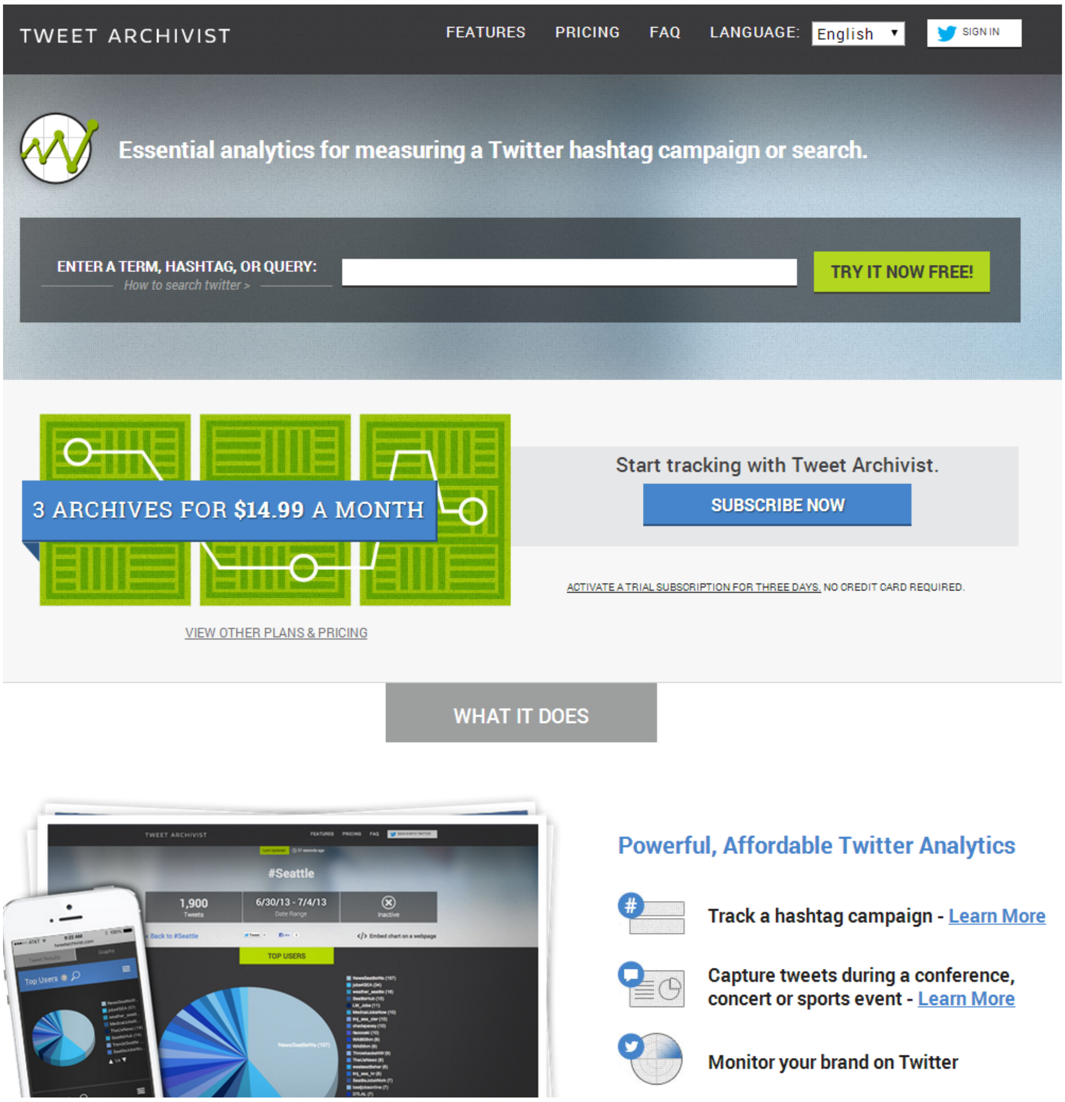
@RadioCHAM @Lizana_john Ele tiu ib e leu ne #Twitter dele!
[illegible]

@Video_News_Now Lady Gaga Most Popular #ladygaga #britney #britneyspears #twitter http://bit.ly/cbRL8kR
[illegible]

@ANicoleJustice Not sure if he's on #twitter but condolences to actor #MartinShort. His wife passed yesterday.
[illegible]

@Zero8 How (and Who) #Facebook, #Twitter, Became Recess for Grown-ups http://bit.ly/c51XQW RT @Sang_H_Lee
[illegible]

Figure 3.1.: yourTwapperKeeper. Source: http://twapperkeeper.wordpress.com/2010/08/25/announcing-yourtwapperkeeper-archive-your-own-tweets-on-your-own-server, 2014-04-21



TWEET ARCHIVIST

FEATURES PRICING FAQ LANGUAGE: English ▾ SIGN IN

Essential analytics for measuring a Twitter hashtag campaign or search.

ENTER A TERM, HASHTAG, OR QUERY:

How to search twitter >

TRY IT NOW FREE!

3 ARCHIVES FOR $14.99 A MONTH

Start tracking with Tweet Archivist.

SUBSCRIBE NOW

ACTIVATE A TRIAL SUBSCRIPTION FOR THREE DAYS. NO CREDIT CARD REQUIRED.

VIEW OTHER PLANS & PRICING

WHAT IT DOES

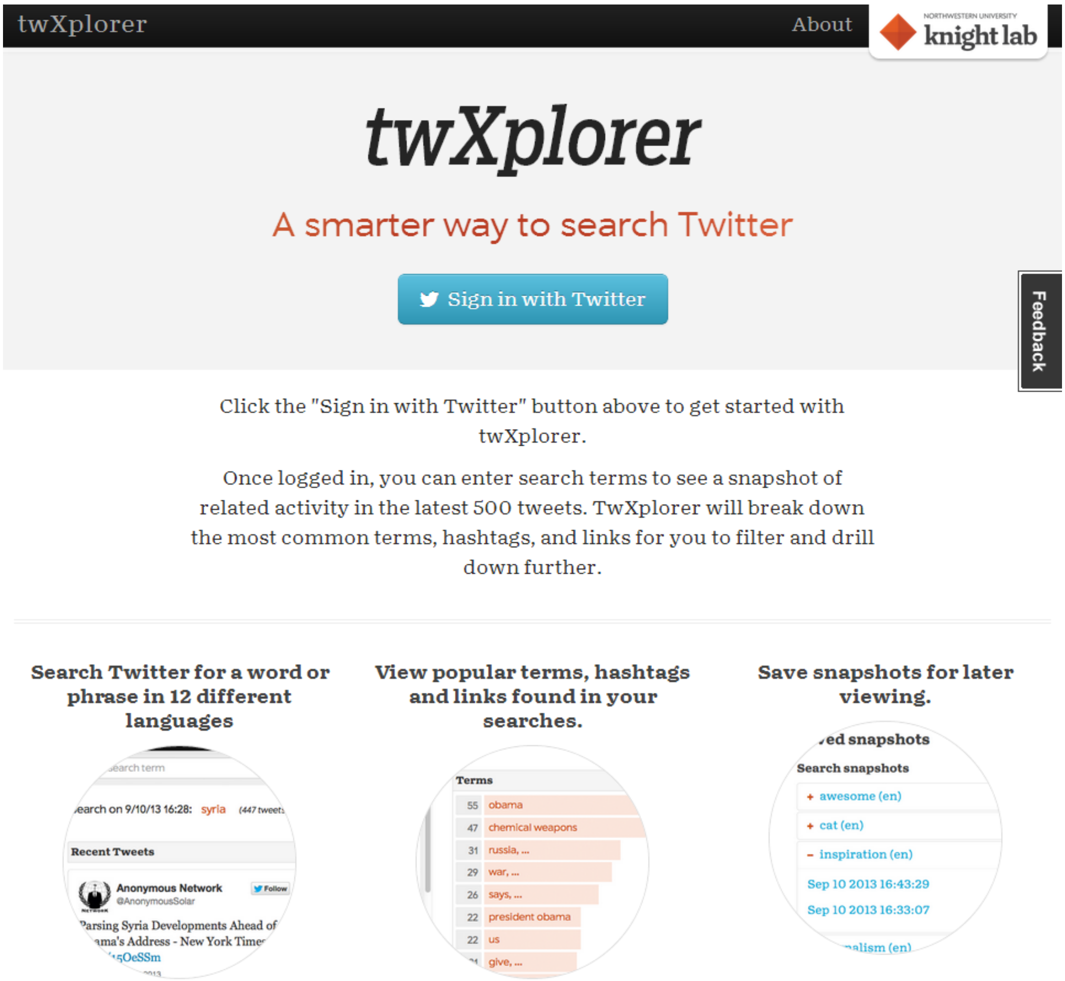
Powerful, Affordable Twitter Analytics

Track a hashtag campaign - Learn More

Capture tweets during a conference, concert or sports event - Learn More

Monitor your brand on Twitter

Figure 3.2.: Tweet Archivist. Source: https://www.tweetarchivist.com, 2014-04-21



twXplorer

A smarter way to search Twitter

Click the "Sign in with Twitter" button above to get started with twXplorer.

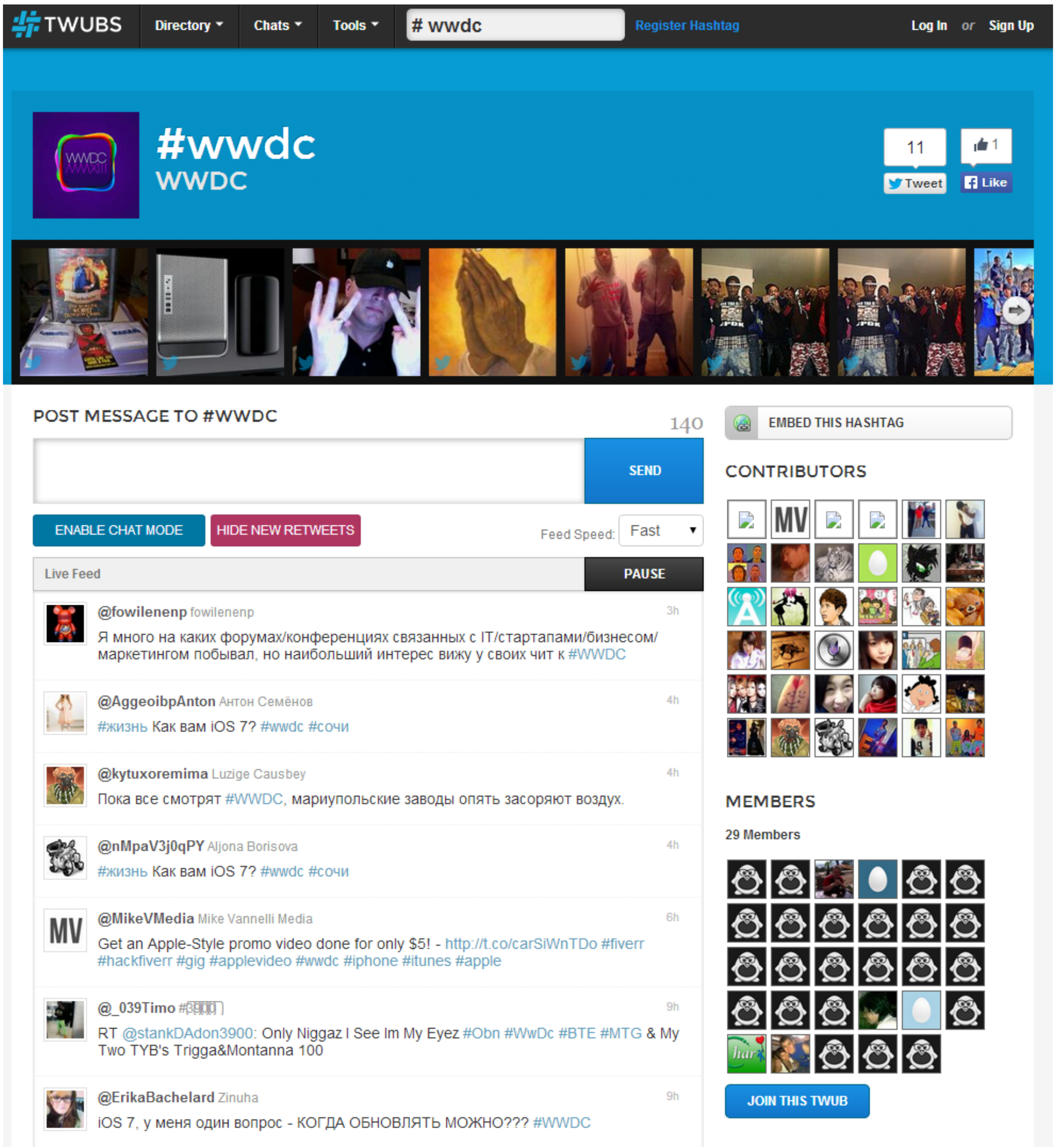
Once logged in, you can enter search terms to see a snapshot of related activity in the latest 500 tweets. TwXplorer will break down the most common terms, hashtags, and links for you to filter and drill down further.

Search Twitter for a word or phrase in 12 different languages

View popular terms, hashtags and links found in your searches.

Save snapshots for later viewing.

Figure 3.3.: twXplorer. Source: http://twxplorer.knightlab.com, 2014-04-21



@fowilenenp fowilenenp 3h
Я много на каких форумах/конференциях связанных с IT/стартапами/бизнесом/маркетингом побывал, но наибольший интерес вижу у своих чит к #WWDC

@AggeoibpAnton Антон Семёнов 4h
#жизнь Как вам iOS 7? #wwdc #сочи

@kytuxoremima Luzige Causbey 4h
Пока все смотрят #WWDC, мариупольские заводы опять засоряют воздух.

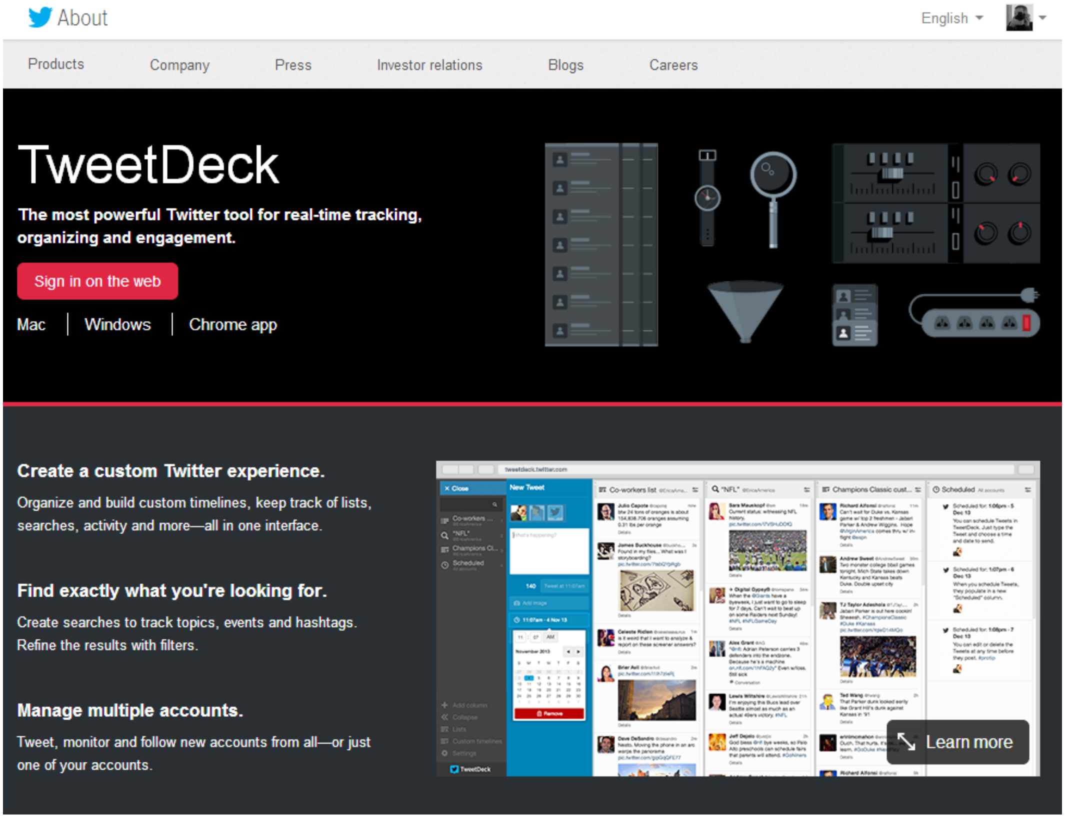
@nMpaV3j0qPY Aljona Borisova 4h
#жизнь Как вам iOS 7? #wwdc #сочи

@MikeVMedia Mike Vannelli Media 6h
Get an Apple-Style promo video done for only $5! - http://t.co/carSiWnTDo #fiverr #hackfiverr #gig #applevideo #wwdc #iphone #itunes #apple

@_039Timo #[illegible] 9h
RT @stankDAdon3900: Only Niggaz I See Im My Eyez #Obn #WwDc #BTE #MTG & My Two TYB's Trigga&Montanna 100

@ErikaBachelard Zinuha 9h
iOS 7, у меня один вопрос - КОГДА ОБНОВЛЯТЬ МОЖНО??? #WWDC

Figure 3.4.: TWUBS. Source: http://twubs.com, 2014-04-21



About

English

Products Company Press Investor relations Blogs Careers

TweetDeck

The most powerful Twitter tool for real-time tracking, organizing and engagement.

Sign in on the web

Mac | Windows | Chrome app

Create a custom Twitter experience.

Organize and build custom timelines, keep track of lists, searches, activity and more—all in one interface.

Find exactly what you're looking for.

Create searches to track topics, events and hashtags. Refine the results with filters.

Manage multiple accounts.

Tweet, monitor and follow new accounts from all—or just one of your accounts.

Figure 3.5.: TweetDeck homepage. Source: https://about.twitter.com/products/tweetdeck, 2014-04-21

4. TweetCollector

TweetCollector is the foundation of the whole stack of tools developed as a part of this thesis that are introduced in the following chapters. It interfaces with the Twitter API to collect tweets containing certain words and hashtags or from certain users. These tweets are stored in tweet archives which can be accessed through a web interface or through a REST API.

The following chapter details how TweetCollector works, which technologies were used to create it, and what kind of interfaces it provides for tools relying on it.

4.1. Predecessor

TweetCollector is based on yourTwapperKeeper, specifically on version 0.5.6. Certain structures in the source code can still be traced back to the original program, but a number of changes have been made to the source code to adapt it to different needs.

These changes include:

- Compatibility with the Twitter API version 1.1 and OAuth 1.0A
- New database structure to store more information about tweets
- Added support for user archives
- Reworked web interface to support adaptive web design and mobile devices

4.2. Implementation Details

This section provides in-depth details of the implementation of TweetCollector.

4.2.1. Dependencies

Some preconditions need to be met for TweetCollector to work. These requirements are detailed here.

Operating System

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.

Webserver

A webserver is needed to run TweetCollector. Apache2 was used for development and deployment. In Debian-based operating systems, this is the package ″apache2″. The tool has not been tested with any other servers.

PHP

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

MySQL is used as a 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”.

4.2.2. Libraries

TweetCollector uses several libraries. All of them are built into the source code, so there are no external dependencies.

TwitterOAuth

The first library is TwitterOAuth.[1] This is used for authentication of registered users of TweetCollector, and for authenticating TweetCollector in requests to the Twitter REST API 1.1.

Phirehose

The second library is called Phirehose.[2] Phirehose is used to interface with the Twitter Streaming API.

Bootstrap

Bootstrap is a front-end framework for developing responsive, mobile-ready websites created by Twitter.[3] It is used for the layout of the webpages. The version used is 3.0.2.

[1] https://github.com/abraham/twitteroauth, 2014-04-21

[2] https://github.com/fennb/phirehose, 2014-04-21

[3] http://getbootstrap.com, 2014-04-21

jQuery

jQuery is a JavaScript library designed for versatility and extensibility.[4] Version 2.0.3 of this library is used.

4.2.3. Installation and Configuration

TweetCollector requires a preconfigured MySQL database. The structure of this database can be found in the file TC.SQL. This SQL[5] dump can be used to create a database via MySQL command line tool or phpMyAdmin.

TC.SQL creates the tables ″archives″ and ″users″, which contain a list of keyword/hashtag archives and user archives run by TweetCollector. The table ″processes″ is created and filled with the names of the 4 processes designed to retrieve tweets.

The file config.php is used to define parameters for the operation of the program and set access credentials for the Twitter API and the MySQL database.

4.2.4. Collector Processes

In the subdirectory ″collectors″, four PHP files are responsible for collecting and storing tweets.

- tweetcollector_crawl_users.php
- tweetcollector_crawl_archives.php
- tweetcollector_stream_collect.php
- tweetcollector_stream_insert.php

These four files run as concurrent processes.

[4] http://jquery.com, 2014-04-21

[5] Structured Query Language

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.

1. Loop over all archives TweetCollector works with.
2. 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.
3. 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

This process uses the aforementioned Phirehose library. Phirehose provides an easy communication layer with the Twitter Streaming API.

The function "enqueueStatus" is called everytime Phirehose receives a tweet fitting the specified search terms. 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 Phirehose which search terms and user names apply to the tweets it should retrieve.

Stream Insert

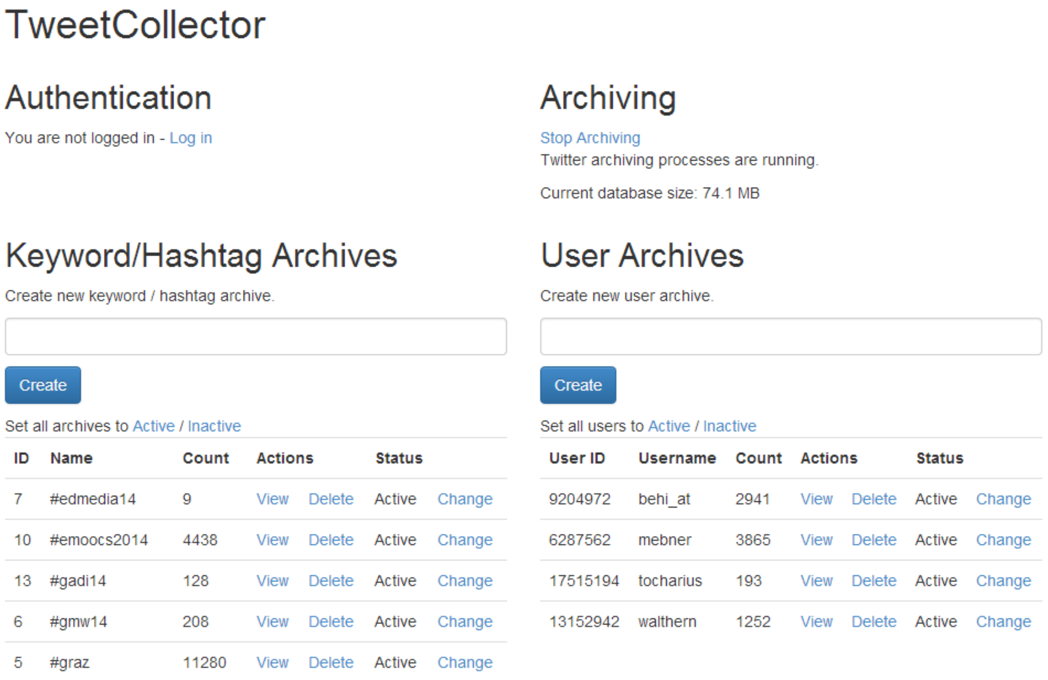
The process ″stream insert″ periodically checks the ″rawstream″ table for new tweets and sorts them into the right tables for each archive. The Phirehose library has a slightly different definition of the search parameters than required. An example is that tweets mentioning a user name are also provided when a user is set in ″setFollow″. TweetCollector archives only store tweets that were written or retweeted by a certain user.

Therefore, each tweet in the rawstream needs to be checked if it complies with one or more of the archives. For example, if there is a user archive called X and a keyword archive called Y, and the rawstream contains a tweet by user X with the word Y, it is added to both archives. If neither is true because the tweet only mentions user X (as in the example above), the tweet is discarded without being added to any archives.

Cron

TweetCollector uses a cronjob[6] to guarantee that the collection processes are running when they should be running. In the event of a crash of the webserver or some other error, the PHP processes would stop. Every 15 minutes, the cronjob checks if the four processes are running and restarts them if necessary.

[6] Scheduled tasks on UNIX-like operating systems



TweetCollector

Authentication

You are not logged in - Log in

Archiving

Stop Archiving
Twitter archiving processes are running.

Current database size: 74.1 MB

Keyword/Hashtag Archives

Create new keyword / hashtag archive.

Set all archives to Active / Inactive

ID	Name	Count	Actions		Status	
7	#edmedia14	9	View	Delete	Active	Change
10	#emoocs2014	4438	View	Delete	Active	Change
13	#gadi14	128	View	Delete	Active	Change
6	#gmw14	208	View	Delete	Active	Change
5	#graz	11280	View	Delete	Active	Change

User Archives

Create new user archive.

Set all users to Active / Inactive

User ID	Username	Count	Actions		Status	
9204972	behi_at	2941	View	Delete	Active	Change
6287562	mebner	3865	View	Delete	Active	Change
17515194	tocharius	193	View	Delete	Active	Change
13152942	walthern	1252	View	Delete	Active	Change

Figure 4.1.: TweetCollector index page. Source: http://tweetcollector.tugraz.at

4.2.5. User Interface

The HTML[7] files ”index.html” and ”view.html” are the only GUI[8] a regular user is interacting with. They are supported by the corresponding JavaScript files ”index.js” and ”view.js” to provide the functionality and interactivity. A screenshot of the index page is shown in figure 4.1.

”index.html” is where a user can log in, look at the archiving status and create, delete, activate and deactivate tweet archives. When a user wants to view the tweets in an archive, he or she is taken to ”view.html”. This viewing is just a very basic interface, because the most important functionality of TweetCollector is the API it provides for other tools.

[7] HyperText Markup Language
[8] Graphical User Interface

4.2.6. Helper Functions

This tool uses several small helper functions for managing the archives on the server. This includes creating, deleting and status changing of archives, logging in and out of users, and checking the status of the archiving processes.

4.3. TweetCollector API

TweetCollector provides three different APIs. All of them are PHP based. They accept GET parameters in the URL and return JSON.[9] This API is used by the user interface of TweetCollector itself, as well as by the applications relying on the data TweetCollector provides.

4.3.1. Info

”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.

4.3.2. List

”list.php” does not accept any parameters. This API simply returns a list of all archives in TweetCollector.

[9] JavaScript Object Notation, a data format

4.3.3. Tweets

Like the info API, ″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.

5. TwitterStat

TwitterStat relies on the archiving function of TweetCollector and provides analysis of these archives.

The core principle of TwitterStat is simple: Take the text of each tweet, dissect it into separate words and count how often those words appear in all tweets in the archive. This gives the user a basic understanding of what general topics are discussed in the tweets.

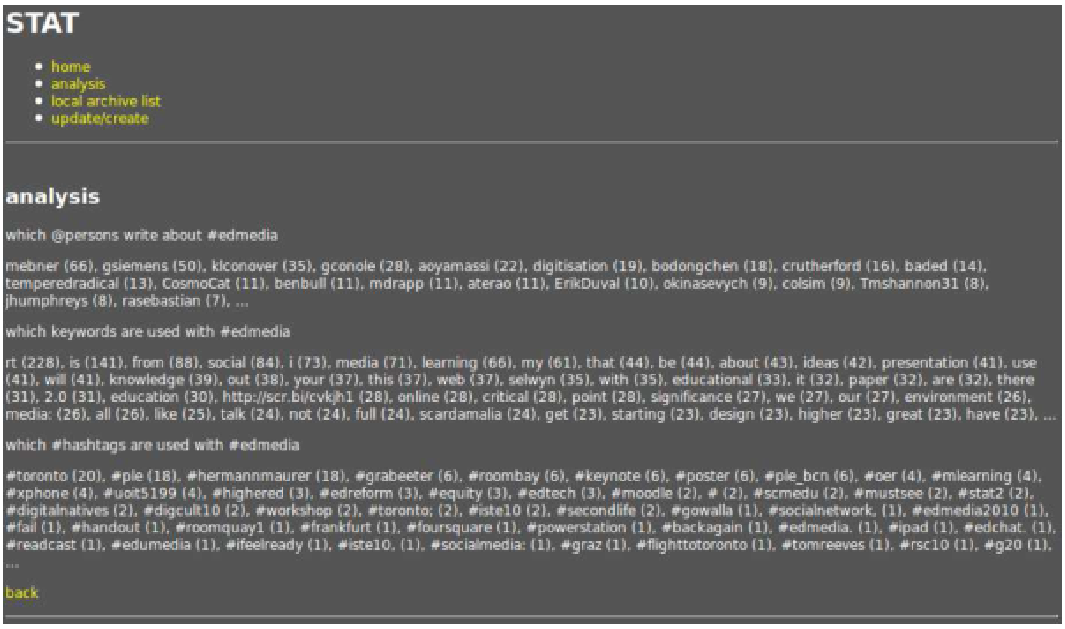
This general principle can be applied to more data points in a tweet archive.

5.1. Development

TwitterStat development started in 2010. Over its lifetime, the software had many different stages that can be distinguished from each other. This section gives a short overview of these stages.

5.1.1. TwapperKepper and Python

TwitterStat started as project for a Bachelor's thesis [Altmann, 2010]. The tweet archives were provided by the TwapperKeeeper web service. Twitterstat used the programming language Python for retrieval and analysis of the tweets. PHP was used for the front-end pages. Figure 5.1 shows a screenshot of this old version.



STAT

analysis

which @persons write about #edmedia

mebner (66), gsiemens (50), klconover (35), gconole (28), aoyamassi (22), digitisation (19), bodongchen (18), crutherford (16), baded (14), temperedradical (13), CosmoCat (11), benbull (11), mdrapp (11), aterao (11), ErikDuval (10), okinasevych (9), colsim (9), Tmshannon31 (8), jhumphreys (8), rasebastian (7), ...

which keywords are used with #edmedia

rt (228), is (141), from (88), social (84), i (73), media (71), learning (66), my (61), that (44), be (44), about (43), ideas (42), presentation (41), use (41), will (41), knowledge (39), out (38), your (37), this (37), web (37), selwyn (35), with (35), educational (33), it (32), paper (32), are (32), there (31), 2.0 (31), education (30), http://scr.bi/cvkjh1 (28), online (28), critical (28), point (28), significance (27), we (27), our (27), environment (26), media: (26), all (26), like (25), talk (24), not (24), full (24), scardamalia (24), get (23), starting (23), design (23), higher (23), great (23), have (23), ...

which #hashtags are used with #edmedia

#toronto (20), #ple (18), #hermannmaurer (18), #grabeeter (6), #roombay (6), #keynote (6), #poster (6), #ple_bcn (6), #oer (4), #mlearning (4), #xphone (4), #uoit5199 (4), #highered (3), #edreform (3), #equity (3), #edtech (3), #moodle (2), # (2), #scmedu (2), #mustsee (2), #stat2 (2), #digitalnatives (2), #digcult10 (2), #workshop (2), #toronto; (2), #iste10 (2), #secondlife (2), #gowalla (1), #socialnetwork, (1), #edmedia2010 (1), #fail (1), #handout (1), #roomquay1 (1), #frankfurt (1), #foursquare (1), #powerstation (1), #backagain (1), #edmedia. (1), #ipad (1), #edchat. (1), #readcast (1), #edumedia (1), #ifeelready (1), #iste10, (1), #socialmedia: (1), #graz (1), #flighttotoronto (1), #tomreeves (1), #rsc10 (1), #g20 (1), ...

Figure 5.1.: The first version of TwitterStat. Source: [Altmann, 2010]

5.1.2. yourTwapperkeeper and AJAX

For a Master's Project at Graz University of Technology, TwitterStat was rewritten from scratch using different technologies.

The closing of the TwapperKeeper web service necessitated the use of the yourTwapperkeeper software to archive tweets. This required the use of a MySQL database.

The analysis API was rewritten in PHP. Instead of plain text, the analysis now returned JSON data, which needed to be processed further to make it easier to read.

This was accomplished by a new front end of HTML pages using JavaScript to dynamically load and process content without the need to reload the page. AJAX[1] was used to achieve this.

5.1.3. Modularization and Dependence on TweetCollector

The current versions of TwitterStat and TweetCollector are separated parts of the TwitterStat version that incorporated yourTwapperkeeper. Starting with the work for this thesis, TwitterStat was broken up into parts to allow for better modularization and re-use of components and APIs.

The yourTwapperkeeper part of TwitterStat became TweetCollector and gained new functionality, better stability and compatibility. Further details about this can be found in chapter 4.

The analysis part of TwitterStat kept that name and was extended and enhanced with new types of analysis and features like returning to subsets of tweets.

The rest of this chapter describes this current version.

[1] Asynchronous JavaScript and XML

5.2. Implementation Details

This section provides in-depth details of the implementation of TwitterStat.

5.2.1. Dependencies

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.

5.2.2. Libraries

TwitterStat uses Bootstrap in version 3.0.2 and jQuery in version 2.0.3. It also uses a Bootstrap plugin called "Bootstrap 3 Typeahead" to provide autocomplete functionality.[2]

5.2.3. Installation and Configuration

TwitterStat requires a running instance of TweetCollector to operate. The URL where the API of TweetCollector can be found needs to be declared in the file config.php. This is the only configuration needed to set up TwitterStat.

[2] https://github.com/bassjobsen/Bootstrap-3-Typeahead, 2014-04-21

5.2.4. TwitterStat API

TwitterStat provides an API for most of its functionality. This API consists of PHP files on the server that return JSON data for specific requests. 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 and Info

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

Analyze

″analyze.php″ is the centerpiece of TwitterStat. It accepts four 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 tweets are then examined in various ways:

- The number and percentage of retweets is calculated.
- The different sources or Twitter clients used to write tweets are counted and ranked.
- The links posted in tweets are counted and ranked.

- The content of each tweet is dissected into individual words. Depending on the type of archive to be analyzed and the presence of the second parameter, different lists of most used words, hashtags, username mentions and other data points are generated.

A more detailed description of possible results can be found in the section ″Analysis Results″.

Tweets

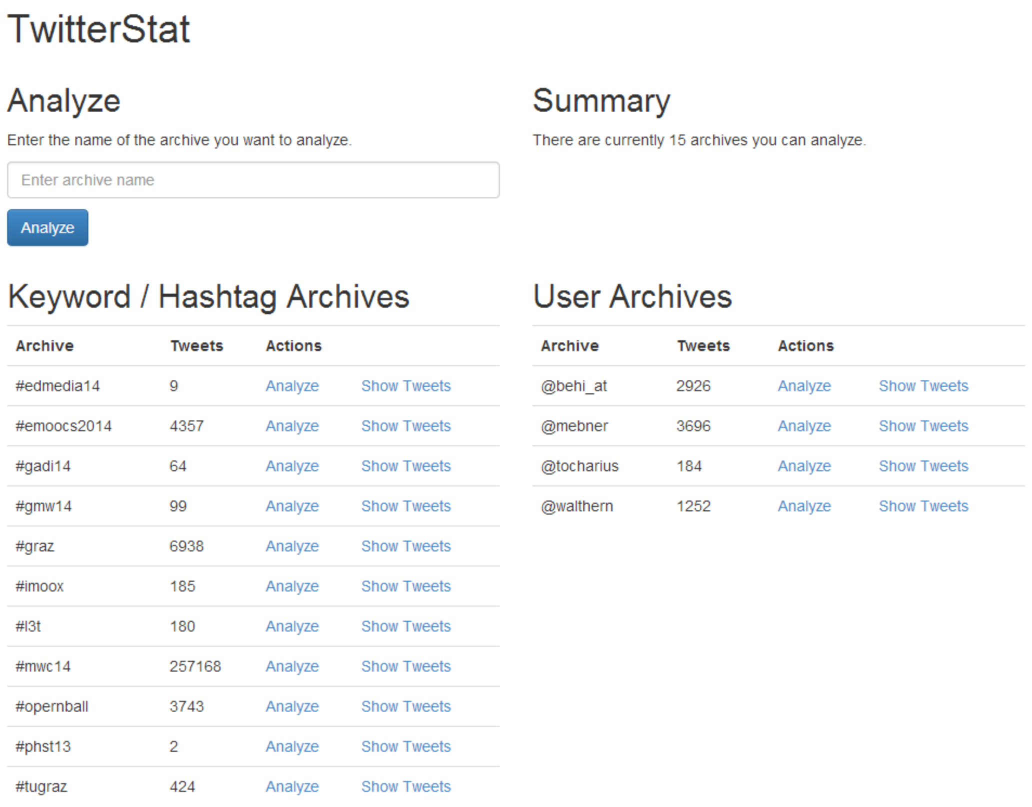
″tweets.php″ returns the tweets of a specific archive. These tweets are retrieved from the TweetCollector API, so all the parameters it supports are present as well:

- ″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.



TwitterStat

Analyze

Enter the name of the archive you want to analyze.

Summary

There are currently 15 archives you can analyze.

Keyword / Hashtag Archives

Archive	Tweets	Actions	
#edmedia14	9	Analyze	Show Tweets
#emoocs2014	4357	Analyze	Show Tweets
#gadi14	64	Analyze	Show Tweets
#gmw14	99	Analyze	Show Tweets
#graz	6938	Analyze	Show Tweets
#imoox	185	Analyze	Show Tweets
#l3t	180	Analyze	Show Tweets
#mwc14	257168	Analyze	Show Tweets
#opernball	3743	Analyze	Show Tweets
#phst13	2	Analyze	Show Tweets
#tugraz	424	Analyze	Show Tweets

User Archives

Archive	Tweets	Actions	
@behi_at	2926	Analyze	Show Tweets
@mebner	3696	Analyze	Show Tweets
@tocharius	184	Analyze	Show Tweets
@walthern	1252	Analyze	Show Tweets

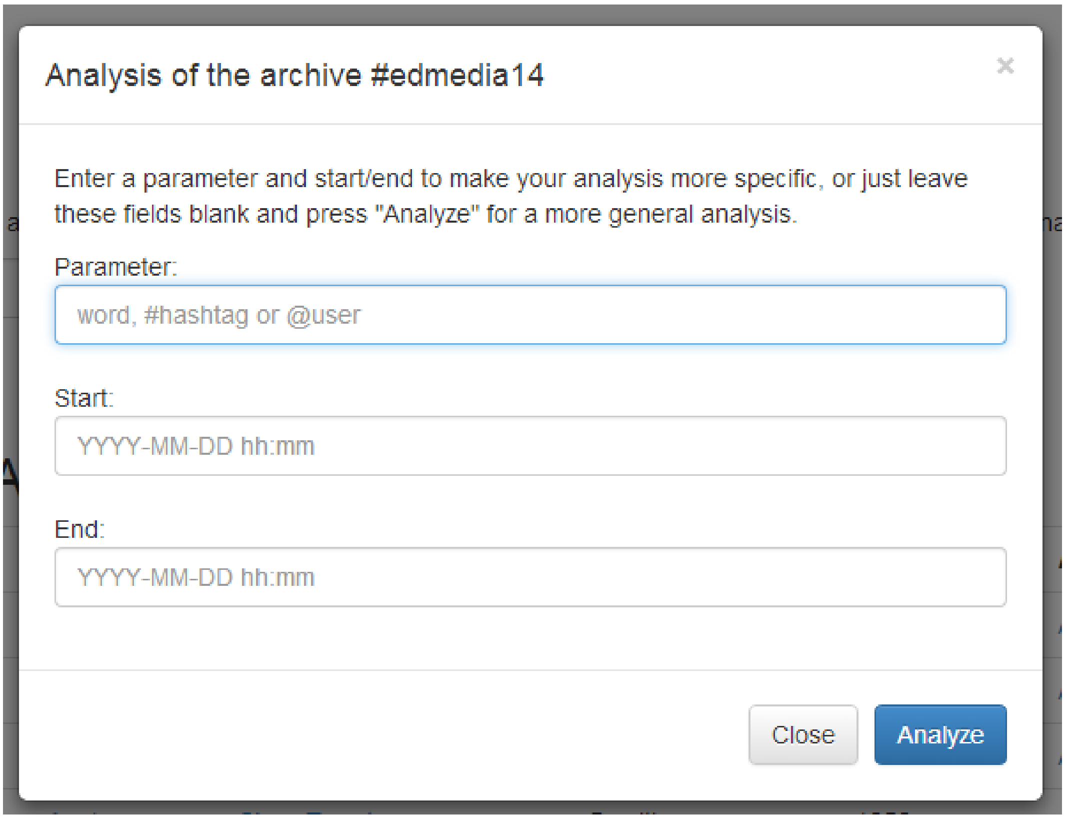
Figure 5.2.: TwitterStat index page. Source: http://twitterstat.tugraz.at

5.2.5. User Interface

Similar to TweetCollector, the TwitterStat user interface consists of HTML pages with JavaScript support. There are three pages a user can interact with.

Index

As shown in figure 5.2, "index.html" and "index.js" provide the homepage of TwitterStat. On this page, all available archives are listed. It also provides



Analysis of the archive #edmedia14

Enter a parameter and start/end to make your analysis more specific, or just leave these fields blank and press "Analyze" for a more general analysis.

Parameter:

word, #hashtag or @user

Start:

YYYY-MM-DD hh:mm

End:

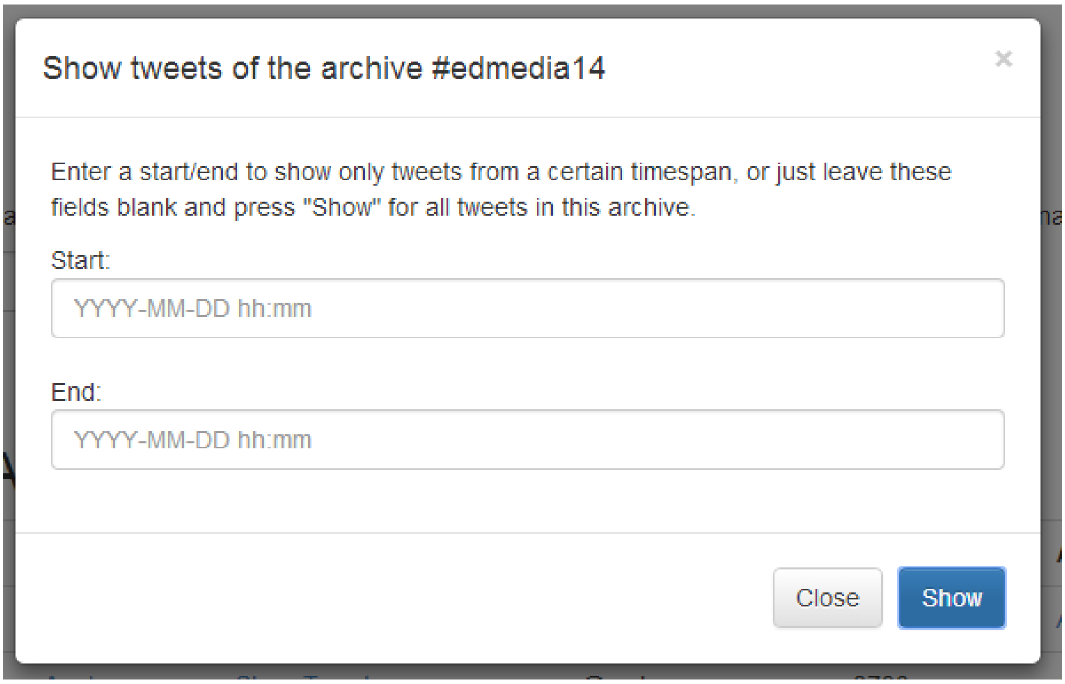
YYYY-MM-DD hh:mm

Figure 5.3.: Analysis dialogue. Source: http://twitterstat.tugraz.at

links to start an analysis or show the tweets of an archive.

When a user wants to analyze a specific archive, the dialogue shown in figure 5.3 asks the user if an additional optional parameter and a start/end date for the analysis should be specified. After this, the analysis is started.

When a user wants to view the tweets in a specific archive, a dialogue asks the user if a start/end date should be specified. More specific subsets of tweets can only be accessed from the analysis page. After clicking "Show", the user is taken to the tweets page to view the specified tweets. Figure 5.4 shows a screenshot of this dialogue.



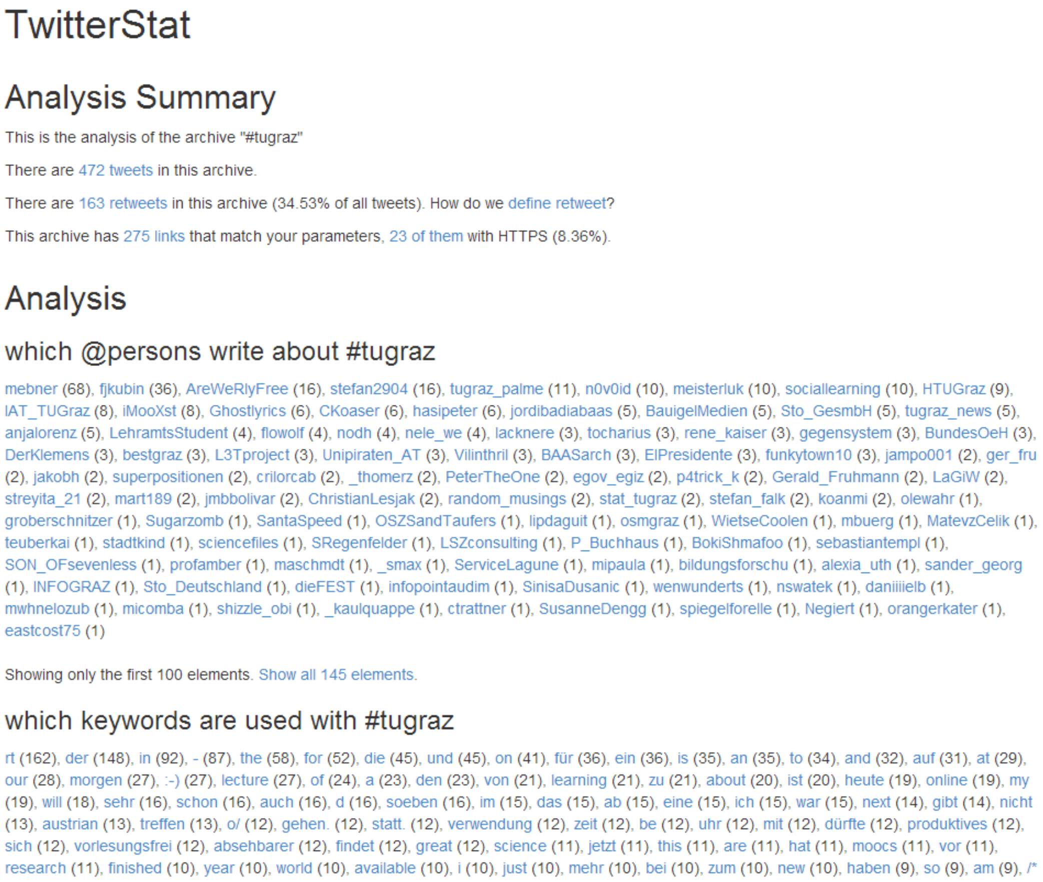
Show tweets of the archive #edmedia14

Enter a start/end to show only tweets from a certain timespan, or just leave these fields blank and press "Show" for all tweets in this archive.

Start: YYYY-MM-DD hh:mm

End: YYYY-MM-DD hh:mm

Figure 5.4.: Tweets dialogue. Source: http://twitterstat.tugraz.at



TwitterStat

Analysis Summary

This is the analysis of the archive "#tugraz"

There are 472 tweets in this archive.

There are 163 retweets in this archive (34.53% of all tweets). How do we define retweet?

This archive has 275 links that match your parameters, 23 of them with HTTPS (8.36%).

Analysis

which @persons write about #tugraz

mebner (68), fjkubin (36), AreWeRlyFree (16), stefan2904 (16), tugraz_palme (11), n0v0id (10), meisterluk (10), sociallearning (10), HTUGraz (9), IAT_TUGraz (8), iMooXst (8), Ghostlyrics (6), CKoaser (6), hasipeter (6), jordibadiabaas (5), BauigelMedien (5), Sto_GesmbH (5), tugraz_news (5), anjalorenz (5), LehramtsStudent (4), flowolf (4), nodh (4), nele_we (4), lacknere (3), tocharius (3), rene_kaiser (3), gegensystem (3), BundesOeH (3), DerKlemens (3), bestgraz (3), L3Tproject (3), Unipiraten_AT (3), Vilinthril (3), BAASarch (3), ElPresidente (3), funkytown10 (3), jampo001 (2), ger_fru (2), jakobh (2), superpositionen (2), crilorcab (2), _thomerz (2), PeterTheOne (2), egov_egiz (2), p4trick_k (2), Gerald_Fruhmann (2), LaGiW (2), streyita_21 (2), mart189 (2), jmbbolivar (2), ChristianLesjak (2), random_musings (2), stat_tugraz (2), stefan_falk (2), koanmi (2), olewahr (1), groberschnitzer (1), Sugarzomb (1), SantaSpeed (1), OSZSandTaufers (1), lipdaguit (1), osmgraz (1), WietseCoolen (1), mbuerg (1), MatevzCelik (1), teuberkai (1), stadtkind (1), sciencefiles (1), SRegenfelder (1), LSZconsulting (1), P_Buchhaus (1), BokiShmafoo (1), sebastiantempl (1), SON_OFsevenless (1), profamber (1), maschmdt (1), _smax (1), ServiceLagune (1), mipaula (1), bildungsforschu (1), alexia_uth (1), sander_georg (1), INFOGRAZ (1), Sto_Deutschland (1), dieFEST (1), infopointaudim (1), SinisaDusanic (1), wenwunderts (1), nswatek (1), daniiiielb (1), mwhnelozub (1), micomba (1), shizzle_obi (1), _kaulquappe (1), ctrattner (1), SusanneDengg (1), spiegelforelle (1), Negiert (1), orangerkater (1), eastcost75 (1)

Showing only the first 100 elements. Show all 145 elements.

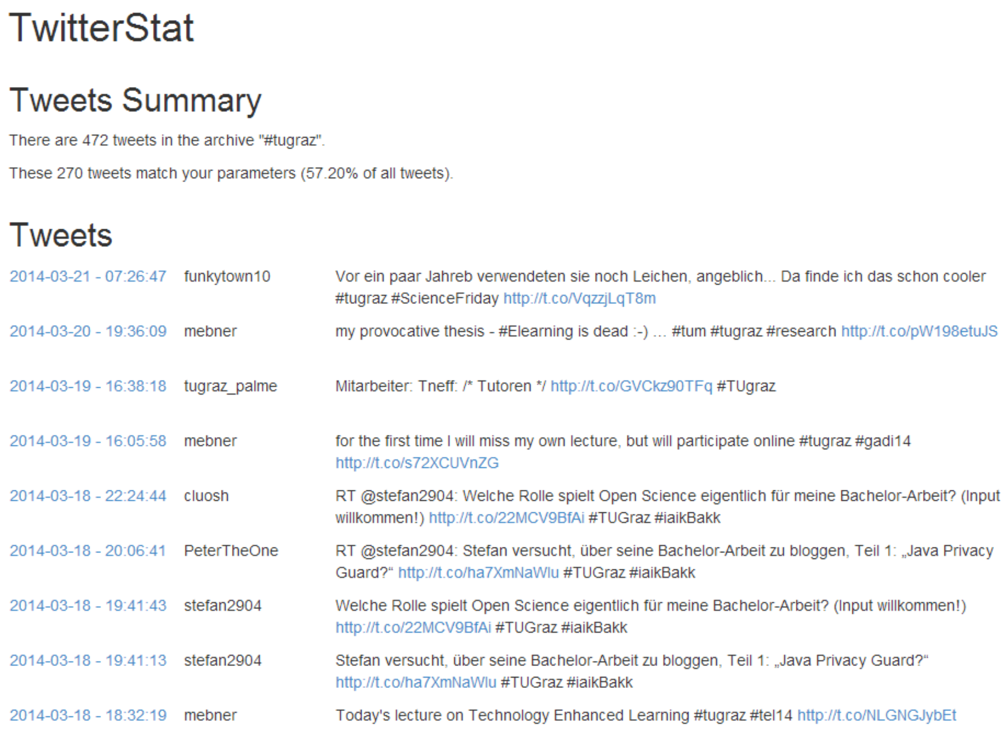
which keywords are used with #tugraz

rt (162), der (148), in (92), - (87), the (58), for (52), die (45), und (45), on (41), für (36), ein (36), is (35), an (35), to (34), and (32), auf (31), at (29), our (28), morgen (27), :-) (27), lecture (27), of (24), a (23), den (23), von (21), learning (21), zu (21), about (20), ist (20), heute (19), online (19), my (19), will (18), sehr (16), schon (16), auch (16), d (16), soeben (16), im (15), das (15), ab (15), eine (15), ich (15), war (15), next (14), gibt (14), nicht (13), austrian (13), treffen (13), o/ (12), gehen. (12), statt. (12), verwendung (12), zeit (12), be (12), uhr (12), mit (12), dürfte (12), produktives (12), sich (12), vorlesungsfrei (12), absehbarer (12), findet (12), great (12), science (11), jetzt (11), this (11), are (11), hat (11), moocs (11), vor (11), research (11), finished (10), year (10), world (10), available (10), i (10), just (10), mehr (10), bei (10), zum (10), new (10), haben (9), so (9), am (9), /*

Figure 5.5.: TwitterStat analysis page. Source: http://twitterstat.tugraz.at

Analysis

"analysis.html" and "analysis.js" present the results of a requested analysis. Depending on the specified parameters, different answers and lists are provided. As shown in figure 5.5, all of the results are links which take the user to the specific tweets that caused a particular result.



TwitterStat

Tweets Summary

There are 472 tweets in the archive "#tugraz".

These 270 tweets match your parameters (57.20% of all tweets).

Tweets

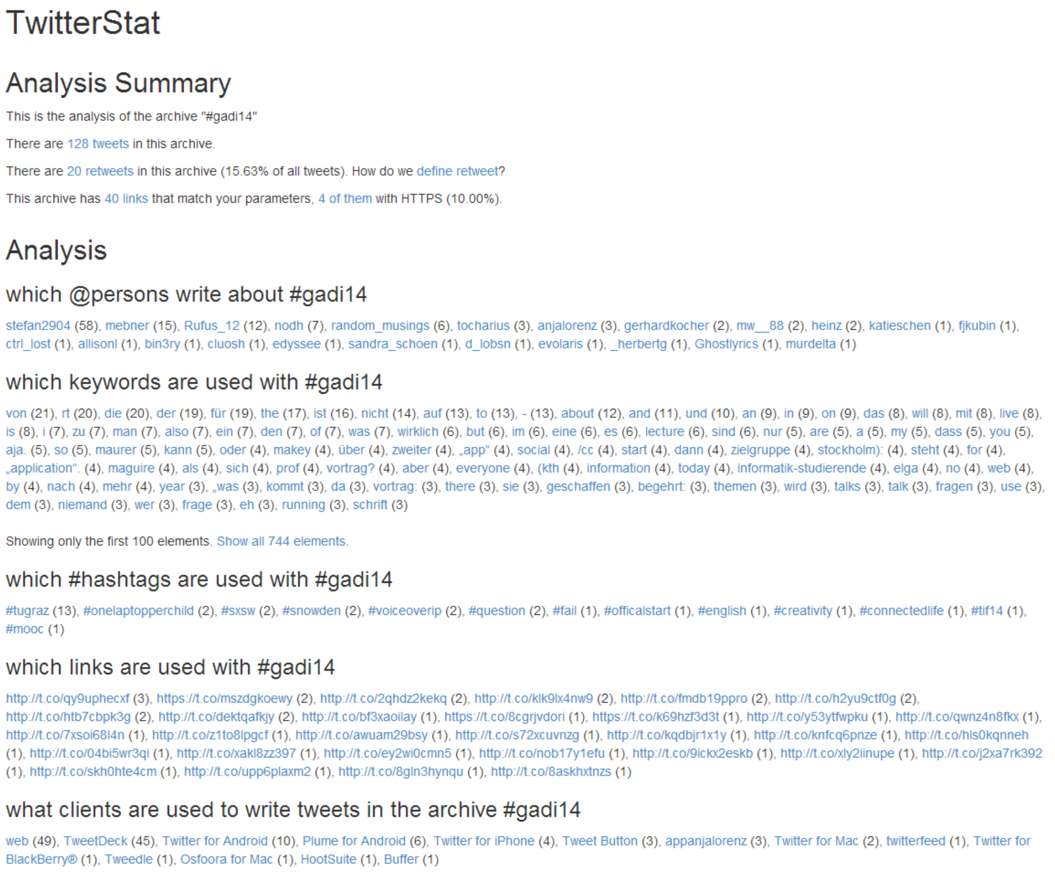
2014-03-21 - 07:26:47	funkytown10	Vor ein paar Jahreb verwendeten sie noch Leichen, angeblich... Da finde ich das schon cooler #tugraz #ScienceFriday http://t.co/VqzzjLqT8m
2014-03-20 - 19:36:09	mebner	my provocative thesis - #Elearning is dead :-) … #tum #tugraz #research http://t.co/pW198etuJS
2014-03-19 - 16:38:18	tugraz_palme	Mitarbeiter: Tneff: /* Tutoren */ http://t.co/GVCkz90TFq #TUgraz
2014-03-19 - 16:05:58	mebner	for the first time I will miss my own lecture, but will participate online #tugraz #gadi14 http://t.co/s72XCUVnZG
2014-03-18 - 22:24:44	cluosh	RT @stefan2904: Welche Rolle spielt Open Science eigentlich für meine Bachelor-Arbeit? (Input willkommen!) http://t.co/22MCV9BfAi #TUGraz #iaikBakk
2014-03-18 - 20:06:41	PeterTheOne	RT @stefan2904: Stefan versucht, über seine Bachelor-Arbeit zu bloggen, Teil 1: „Java Privacy Guard?“ http://t.co/ha7XmNaWlu #TUGraz #iaikBakk
2014-03-18 - 19:41:43	stefan2904	Welche Rolle spielt Open Science eigentlich für meine Bachelor-Arbeit? (Input willkommen!) http://t.co/22MCV9BfAi #TUGraz #iaikBakk
2014-03-18 - 19:41:13	stefan2904	Stefan versucht, über seine Bachelor-Arbeit zu bloggen, Teil 1: „Java Privacy Guard?“ http://t.co/ha7XmNaWlu #TUGraz #iaikBakk
2014-03-18 - 18:32:19	mebner	Today's lecture on Technology Enhanced Learning #tugraz #tel14 http://t.co/NLGNGJybEt

Figure 5.6.: TwitterStat tweets page. Source: http://twitterstat.tugraz.at

Tweets

”tweets.html” and ”tweets.js” display tweets that fit certain criteria. These tweets are retrieved from the TwitterStat API ”tweets.php” and displayed as a list. Additionally, as shown in figure 5.6, this page displays the following information:

- How many tweets are in the specified archive.
- How many tweets match the parameters.
- What percentage of the total tweets in the archive matches the parameters.



TwitterStat

Analysis Summary

This is the analysis of the archive "#gadi14"

There are 128 tweets in this archive.

There are 20 retweets in this archive (15.63% of all tweets). How do we define retweet?

This archive has 40 links that match your parameters, 4 of them with HTTPS (10.00%).

Analysis

which @persons write about #gadi14

stefan2904 (58), mebner (15), Rufus_12 (12), nodh (7), random_musings (6), tocharius (3), anjalorenz (3), gerhardkocher (2), mw__88 (2), heinz (2), katieschen (1), fjkubin (1), ctrl_lost (1), allisonl (1), bin3ry (1), cluosh (1), edyssee (1), sandra_schoen (1), d_lobsn (1), evolaris (1), _herbertg (1), Ghostlyrics (1), murdelta (1)

which keywords are used with #gadi14

von (21), rt (20), die (20), der (19), für (19), the (17), ist (16), nicht (14), auf (13), to (13), - (13), about (12), and (11), und (10), an (9), in (9), on (9), das (8), will (8), mit (8), live (8), is (8), i (7), zu (7), man (7), also (7), ein (7), den (7), of (7), was (7), wirklich (6), but (6), im (6), eine (6), es (6), lecture (6), sind (6), nur (5), are (5), a (5), my (5), dass (5), you (5), aja. (5), so (5), maurer (5), kann (5), oder (4), makey (4), über (4), zweiter (4), „app“ (4), social (4), /cc (4), start (4), dann (4), zielgruppe (4), stockholm): (4), steht (4), for (4), „application“. (4), maguire (4), als (4), sich (4), prof (4), vortrag? (4), aber (4), everyone (4), (kth (4), information (4), today (4), informatik-studierende (4), elga (4), no (4), web (4), by (4), nach (4), mehr (4), year (3), „was (3), kommt (3), da (3), vortrag: (3), there (3), sie (3), geschaffen (3), begehrt: (3), themen (3), wird (3), talks (3), talk (3), fragen (3), use (3), dem (3), niemand (3), wer (3), frage (3), eh (3), running (3), schrift (3)

Showing only the first 100 elements. Show all 744 elements.

which #hashtags are used with #gadi14

#tugraz (13), #onelaptopperchild (2), #sxsw (2), #snowden (2), #voiceoverip (2), #question (2), #fail (1), #officalstart (1), #english (1), #creativity (1), #connectedlife (1), #tif14 (1), #mooc (1)

which links are used with #gadi14

http://t.co/qy9uphecxf (3), https://t.co/mszdgkoewy (2), http://t.co/2qhdz2kekq (2), http://t.co/klk9lx4nw9 (2), http://t.co/fmdb19ppro (2), http://t.co/h2yu9ctf0g (2), http://t.co/htb7cbpk3g (2), http://t.co/dektqafkjy (2), http://t.co/bf3xaoilay (1), https://t.co/8cgrjvdori (1), https://t.co/k69hzf3d3t (1), http://t.co/y53ytfwpku (1), http://t.co/qwnz4n8fkx (1), http://t.co/7xsoi68l4n (1), http://t.co/z1to8lpgcf (1), http://t.co/awuam29bsy (1), http://t.co/s72xcuvnzg (1), http://t.co/kqdbjr1x1y (1), http://t.co/knfcq6pnze (1), http://t.co/hls0kqnneh (1), http://t.co/04bi5wr3qi (1), http://t.co/xakl8zz397 (1), http://t.co/ey2wi0cmn5 (1), http://t.co/nob17y1efu (1), http://t.co/9ickx2eskb (1), http://t.co/xly2iinupe (1), http://t.co/j2xa7rk392 (1), http://t.co/skh0hte4cm (1), http://t.co/upp6plaxm2 (1), http://t.co/8gln3hynqu (1), http://t.co/8askhxtnzs (1)

what clients are used to write tweets in the archive #gadi14

web (49), TweetDeck (45), Twitter for Android (10), Plume for Android (6), Twitter for iPhone (4), Tweet Button (3), appanjalorenz (3), Twitter for Mac (2), twitterfeed (1), Twitter for BlackBerry® (1), Tweedle (1), Osfoora for Mac (1), HootSuite (1), Buffer (1)

Figure 5.7.: Full analysis result of hashtag archive. Source: http://twitterstat.tugraz.at

5.3. Analysis Results

The results page shows the full analysis of a Twitter archive, as shown in figure 5.7.

As mentioned before, the analysis of TwitterStat accepts four parameters: archive, second parameter, start date and end date.

The parameter ”archive” is required to define which archive to analyze. 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 6 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 "#tugraz" 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 #tugraz)

The other results depend on the type of analysis performed. These results are detailed in the following sections. For higher legibility, specific examples are used.

5.3.1. Keyword/Hashtag Archive without Parameter

The archive is "#tugraz". Four results are provided.

- which @persons write about #tugraz
- which keywords are used with #tugraz
- which #hashtags are used with #tugraz

- which links are used with #tugraz

5.3.2. Keyword/Hashtag Archive with Keyword/Hashtag Parameter

The archive is ″#tugraz″, the parameter is ″lecture″. Four results are provided.

- which @persons write #tugraz together with lecture
- which keywords are used with #tugraz and lecture
- which #hashtags are used with #tugraz and lecture
- which links are used with #tugraz and lecture

5.3.3. Keyword/Hashtag Archive with User Parameter

The archive is ″#tugraz″, the parameter is ″@mebner″. Six results are provided. A screenshot of this can be seen in figure 5.8.

- which @persons talk to @mebner about #tugraz
- who does @mebner talk to about #tugraz
- who else is addressed with @mebner about #tugraz
- which keywords are used by @mebner about #tugraz
- which #hashtags are used by @mebner about #tugraz
- which links are used by @mebner about #tugraz

5.3.4. User Archive without Parameter

The archive is ″@mebner″, Four results are provided.

- who does @mebner talk to
- which keywords are used by @mebner
- which #hashtags are used by @mebner
- which links are used by @mebner

5.3.5. User Archive with Keyword/Hashtag Parameter

The archive is ”@mebner”, the parameter is ”#tugraz”. Four results are provided.

- who does@ mebner talk to about #tugraz
- which keywords are used by @mebner with #tugraz
- which #hashtags are used by @mebner with #tugraz
- which links are used by @mebner with #tugraz

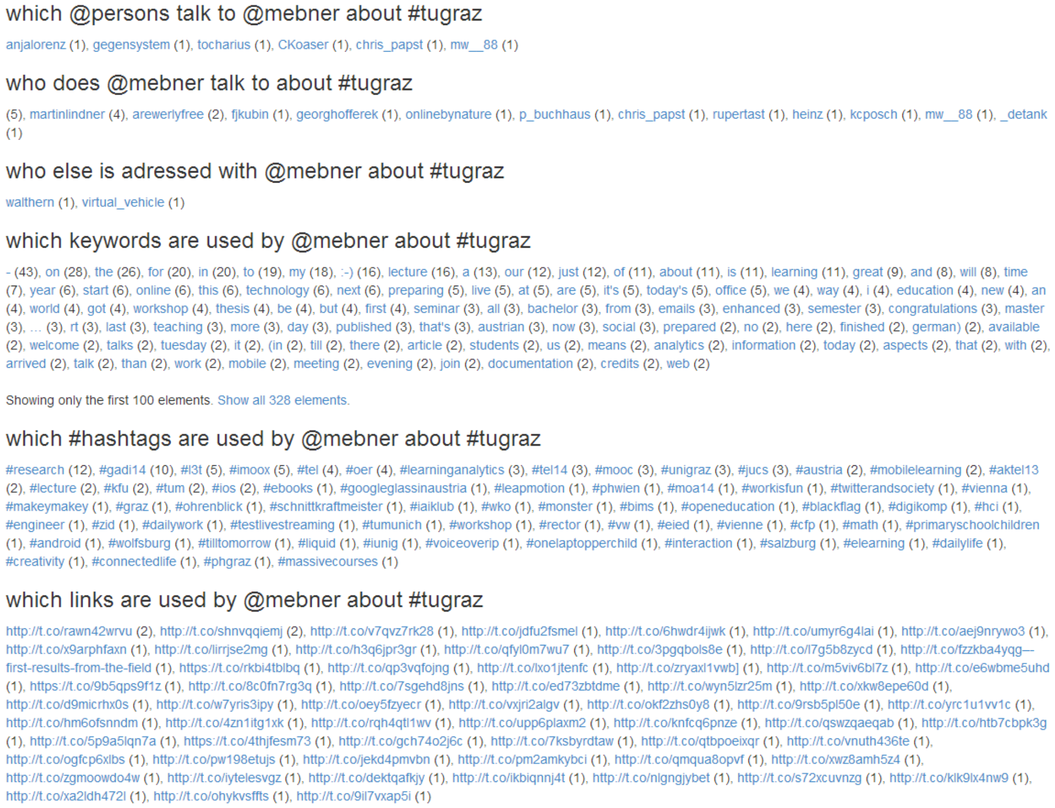
5.3.6. User Archive with User Parameter

The archive is ”@mebner”, the parameter is ”@annebb”. Four results are provided.

- who does @mebner address together with @annebb
- which keywords does @mebner use when talking to @annebb
- which #hashtags does @mebner use when talking to @annebb
- which links does @mebner use when talking to @annebb

5.3.7. Sorted Lists

Each of the parameter-specific results in the sections above 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.



which @persons talk to @mebner about #tugraz

anjalorenz (1), gegensystem (1), tocharius (1), CKoaser (1), chris_papst (1), mw__88 (1)

who does @mebner talk to about #tugraz

(5), martinlindner (4), arewerlyfree (2), fjkubin (1), georghofferek (1), onlinebynature (1), p_buchhaus (1), chris_papst (1), rupertast (1), heinz (1), kcposch (1), mw__88 (1), _detank (1)

who else is adressed with @mebner about #tugraz

walthern (1), virtual_vehicle (1)

which keywords are used by @mebner about #tugraz

- (43), on (28), the (26), for (20), in (20), to (19), my (18), :-) (16), lecture (16), a (13), our (12), just (12), of (11), about (11), is (11), learning (11), great (9), and (8), will (8), time (7), year (6), start (6), online (6), this (6), technology (6), next (6), preparing (5), live (5), at (5), are (5), it's (5), today's (5), office (5), we (4), way (4), i (4), education (4), new (4), an (4), world (4), got (4), workshop (4), thesis (4), be (4), but (4), first (4), seminar (3), all (3), bachelor (3), from (3), emails (3), enhanced (3), semester (3), congratulations (3), master (3), ... (3), rt (3), last (3), teaching (3), more (3), day (3), published (3), that's (3), austrian (3), now (3), social (3), prepared (2), no (2), here (2), finished (2), german) (2), available (2), welcome (2), talks (2), tuesday (2), it (2), (in (2), till (2), there (2), article (2), students (2), us (2), means (2), analytics (2), information (2), today (2), aspects (2), that (2), with (2), arrived (2), talk (2), than (2), work (2), mobile (2), meeting (2), evening (2), join (2), documentation (2), credits (2), web (2)

Showing only the first 100 elements. Show all 328 elements.

which #hashtags are used by @mebner about #tugraz

#research (12), #gadi14 (10), #l3t (5), #imoox (5), #tel (4), #oer (4), #learninganalytics (3), #tel14 (3), #mooc (3), #unigraz (3), #jucs (3), #austria (2), #mobilelearning (2), #aktel13 (2), #lecture (2), #kfu (2), #tum (2), #ios (2), #ebooks (1), #googleglassinaustria (1), #leapmotion (1), #phwien (1), #moa14 (1), #workisfun (1), #twitterandsociety (1), #vienna (1), #makeymakey (1), #graz (1), #ohrenblick (1), #schnittkraftmeister (1), #iaiklub (1), #wko (1), #monster (1), #bims (1), #openeducation (1), #blackflag (1), #digikomp (1), #hci (1), #engineer (1), #zid (1), #dailywork (1), #testlivestreaming (1), #tumunich (1), #workshop (1), #rector (1), #vw (1), #eied (1), #vienne (1), #cfp (1), #math (1), #primaryschoolchildren (1), #android (1), #wolfsburg (1), #tilltomorrow (1), #liquid (1), #iunig (1), #voiceoverip (1), #onelaptopperchild (1), #interaction (1), #salzburg (1), #elearning (1), #dailylife (1), #creativity (1), #connectedlife (1), #phgraz (1), #massivecourses (1)

which links are used by @mebner about #tugraz

http://t.co/rawn42wrvu (2), http://t.co/shnvqqiemj (2), http://t.co/v7qvz7rk28 (1), http://t.co/jdfu2fsmel (1), http://t.co/6hwdr4ijwk (1), http://t.co/umyr6g4lai (1), http://t.co/aej9nrywo3 (1), http://t.co/x9arphfaxn (1), http://t.co/lirrjse2mg (1), http://t.co/h3q6jpr3gr (1), http://t.co/qfyl0m7wu7 (1), http://t.co/3pgqbols8e (1), http://t.co/l7g5b8zycd (1), http://t.co/fzzkba4yqg—first-results-from-the-field (1), https://t.co/rkbi4tblbq (1), http://t.co/qp3vqfojng (1), http://t.co/lxo1jtenfc (1), http://t.co/zryaxl1vwb] (1), http://t.co/m5viv6bl7z (1), http://t.co/e6wbme5uhd (1), https://t.co/9b5qps9f1z (1), http://t.co/8c0fn7rg3q (1), http://t.co/7sgehd8jns (1), http://t.co/ed73zbtdme (1), http://t.co/wyn5lzr25m (1), http://t.co/xkw8epe60d (1), http://t.co/d9micrhx0s (1), http://t.co/w7yris3ipy (1), http://t.co/oey5fzyecr (1), http://t.co/vxjri2algv (1), http://t.co/okf2zhs0y8 (1), http://t.co/9rsb5pl50e (1), http://t.co/yrc1u1vv1c (1), http://t.co/hm6ofsnndm (1), http://t.co/4zn1itg1xk (1), http://t.co/rqh4qtl1wv (1), http://t.co/upp6plaxm2 (1), http://t.co/knfcq6pnze (1), http://t.co/qswzqaeqab (1), http://t.co/htb7cbpk3g (1), http://t.co/5p9a5lqn7a (1), https://t.co/4thjfesm73 (1), http://t.co/gch74o2j6c (1), http://t.co/7ksbyrdtaw (1), http://t.co/qtbpoeixqr (1), http://t.co/vnuth436te (1), http://t.co/ogfcp6xlbs (1), http://t.co/pw198etujs (1), http://t.co/jekd4pmvbn (1), http://t.co/pm2amkybci (1), http://t.co/qmqua8opvf (1), http://t.co/xwz8amh5z4 (1), http://t.co/zgmoowdo4w (1), http://t.co/iytelesvgz (1), http://t.co/dektqafkjy (1), http://t.co/ikbiqnnj4t (1), http://t.co/nlgngjybet (1), http://t.co/s72xcuvnzg (1), http://t.co/klk9lx4nw9 (1), http://t.co/xa2ldh472l (1), http://t.co/ohykvsffts (1), http://t.co/9il7vxap5i (1)

Figure 5.8.: Analysis of hashtag archive with person parameter. Source: http://twitterstat.tugraz.at

6. Twitterwall

TwitterWall builds on the archiving function of TweetCollector and can show different columns of tweets distinguished by the content of the tweets. This can either be used for near real-time tracking to see what people are writing, or to view certain topics of discussion within a tweet archive.

6.1. Implementation Details

The implementation of TwitterWall is very similar to TwitterStat. There is a PHP API to provide server-side access to data and computation and a user interface based on HTML and JavaScript. The dependencies and used libraries are the same. To have a complete overview of the implementation, all the details are listed below.

6.1.1. Dependencies

TwitterWall needs a webserver and PHP. The PHP version required is 5.4 or higher.

6.1.2. Libraries

TwitterWall employs Bootstrap 3.0.2 and jQuery 2.0.3 and the Bootstrap plugin "Bootstrap 3 Typeahead".[1]

[1] https://github.com/bassjobsen/Bootstrap-3-Typeahead, 2014-04-21

6.1.3. Installation and Configuration

Like TwitterStat, TwitterWall relies on a running instance of TweetCollector to supply data. The URL of the TweetCollector API has to be set in the file ″config.php″.

6.1.4. TwitterWall API

TwitterWall uses PHP on the webserver to provide data to the user interface. This data is accessed via AJAX.

″list.php″ provides a list of available archives similar to the other tools discussed in this thesis.

″tweets.php″ provides access to the tweets of a certain archive. A start and end date can be set to limit the timeframe of the requested tweets. This API is polled periodically to provide new tweets for TwitterWall to display.

″time.php″ simply returns the system time of the server. This is used to mitigate incorrect or inaccurate time set on the client computer viewing a TwitterWall.

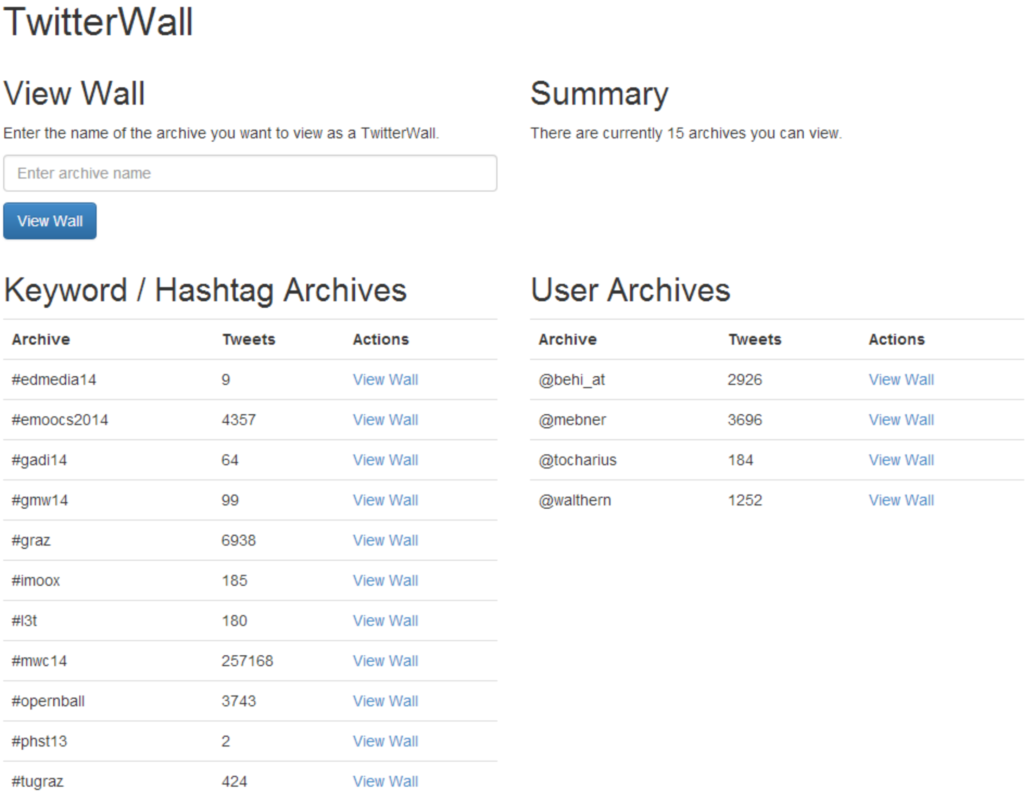
6.1.5. User Interface

There are two JavaScript-supported HTML pages a user can interact with.

Index

″index.html″ and ″index.js″ are the homepage of TwitterWall. Similar to the other tools, a user is presented with the available archives. This is shown in figure 6.1.

When a user picks an archive to display as a TwitterWall, a dialogue is shown. A start and end date as well as the interval in which new tweets are fetched can be entered.



TwitterWall

View Wall

Enter the name of the archive you want to view as a TwitterWall.

Summary

There are currently 15 archives you can view.

Keyword / Hashtag Archives

Archive	Tweets	Actions
#edmedia14	9	View Wall
#emoocs2014	4357	View Wall
#gadi14	64	View Wall
#gmw14	99	View Wall
#graz	6938	View Wall
#imoox	185	View Wall
#l3t	180	View Wall
#mwc14	257168	View Wall
#opernball	3743	View Wall
#phst13	2	View Wall
#tugraz	424	View Wall

User Archives

Archive	Tweets	Actions
@behi_at	2926	View Wall
@mebner	3696	View Wall
@tocharius	184	View Wall
@walthern	1252	View Wall

Figure 6.1.: TwitterWall index page. Source: http://twitterwall.tugraz.at

The user is then taken to the wall display page.

Depending on the time frame the user sets, TwitterWall uses one of several different operation modes:

- No start date, no end date (default): TwitterWall fetches new tweets forever.
- No start date, end date in the future: TwitterWall fetches new tweets until the end date is reached.
- Start and end date in the past: TwitterWall shows old tweets without active fetching.
- Start date in the past, no end date: TwitterWall shows old tweets and fetches new tweets forever.
- Start date in the past, end date in the future: TwitterWall shows old tweets and fetches new tweets until the end date is reached.
- Start date in the future, no end date: TwitterWall waits for start date, then fetches new tweets forever.
- Start date in the future, end date in the future: TwitterWall waits for start date, then fetches new tweets until the end date is reached.

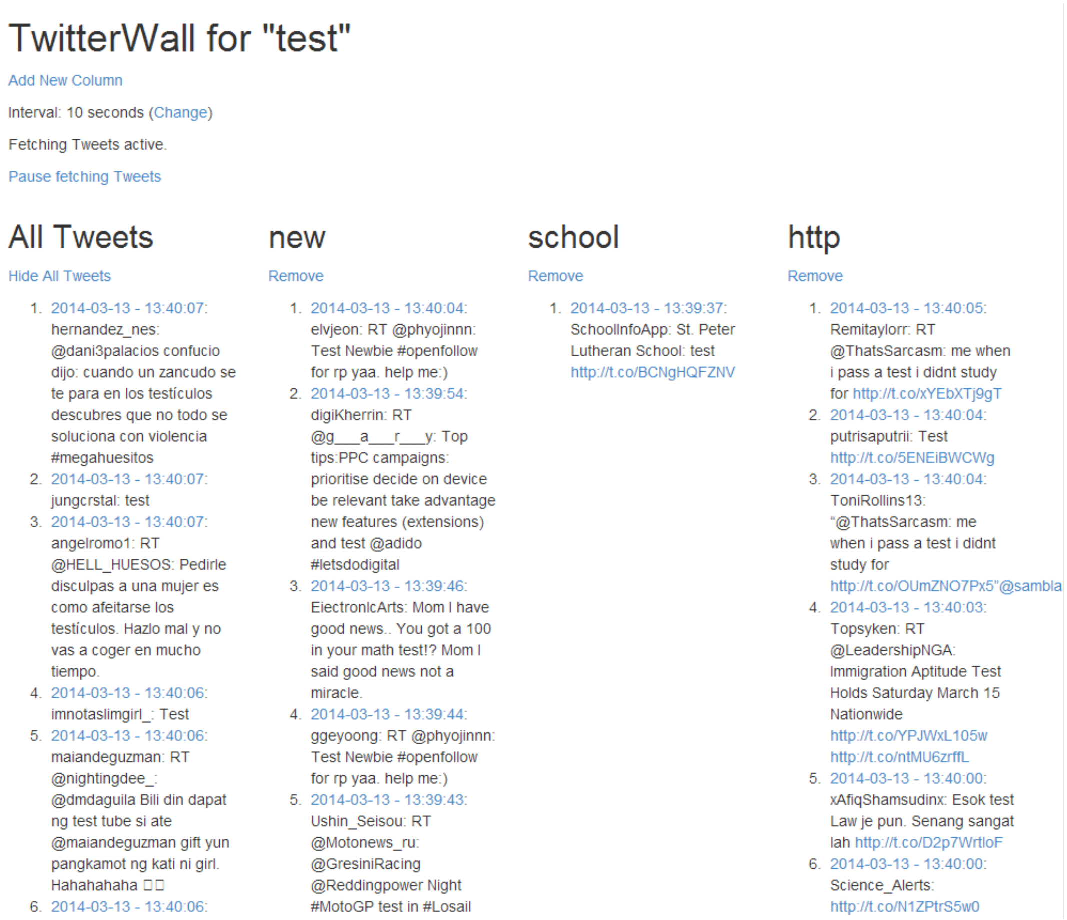
Wall

The wall display page consists of ″wall.html″ and ″wall.js″. Figure 6.2 shows a page displaying a wall for the archive ″test″.

On the top of the page, information about the archive and the status of the fetching mechanism are shown. The interval in which new tweets are fetched is displayed and a link opens a dialogue where the user can adjust the interval.

Below this section are the columns of tweets. When creating a new TwitterWall, only the column ″All Tweets″ is present. This column can be hidden to make room for other filtered columns.

A user can add up to 5 columns by clicking on the link ″Add New Column″. A dialogue pops up where one or more filters can be defined for the new column. The new column is then added to the right side of the existing columns.



TwitterWall for "test"

Add New Column

Interval: 10 seconds (Change)

Fetching Tweets active.

Pause fetching Tweets

All Tweets

Hide All Tweets

1. 2014-03-13 - 13:40:07: hernandez_nes: @dani3palacios confucio dijo: cuando un zancudo se te para en los testículos descubres que no todo se soluciona con violencia #megahuesitos
2. 2014-03-13 - 13:40:07: jungcrstal: test
3. 2014-03-13 - 13:40:07: angelromo1: RT @HELL_HUESOS: Pedirle disculpas a una mujer es como afeitarse los testículos. Hazlo mal y no vas a coger en mucho tiempo.
4. 2014-03-13 - 13:40:06: imnotaslimgirl_: Test
5. 2014-03-13 - 13:40:06: maiandeguzman: RT @nightingdee_: @dmdaguila Bili din dapat ng test tube si ate @maiandeguzman gift yun pangkamot ng kati ni girl. Hahahahaha □□
6. 2014-03-13 - 13:40:06:

new

Remove

1. 2014-03-13 - 13:40:04: elvjeon: RT @phyojinnn: Test Newbie #openfollow for rp yaa. help me:)
2. 2014-03-13 - 13:39:54: digiKherrin: RT @g___a___r___y: Top tips:PPC campaigns: prioritise decide on device be relevant take advantage new features (extensions) and test @adido #letsdodigital
3. 2014-03-13 - 13:39:46: EiectronIcArts: Mom I have good news.. You got a 100 in your math test!? Mom I said good news not a miracle.
4. 2014-03-13 - 13:39:44: ggeyoong: RT @phyojinnn: Test Newbie #openfollow for rp yaa. help me:)
5. 2014-03-13 - 13:39:43: Ushin_Seisou: RT @Motonews_ru: @GresiniRacing @Reddingpower Night #MotoGP test in #Losail

school

Remove

1. 2014-03-13 - 13:39:37: SchoolInfoApp: St. Peter Lutheran School: test http://t.co/BCNgHQFZNV

http

Remove

1. 2014-03-13 - 13:40:05: Remitaylorr: RT @ThatsSarcasm: me when i pass a test i didnt study for http://t.co/xYEbXTj9gT
2. 2014-03-13 - 13:40:04: putrisaputrii: Test http://t.co/5ENEiBWCWg
3. 2014-03-13 - 13:40:04: ToniRollins13: “@ThatsSarcasm: me when i pass a test i didnt study for http://t.co/OUmZNO7Px5”@sambla
4. 2014-03-13 - 13:40:03: Topsyken: RT @LeadershipNGA: Immigration Aptitude Test Holds Saturday March 15 Nationwide http://t.co/YPJWxL105w http://t.co/ntMU6zrffL
5. 2014-03-13 - 13:40:00: xAfiqShamsudinx: Esok test Law je pun. Senang sangat lah http://t.co/D2p7WrtloF
6. 2014-03-13 - 13:40:00: Science_Alerts: http://t.co/N1ZPtrS5w0

Figure 6.2.: TwitterWall wall page. Source: http://twitterwall.tugraz.at

The width of all columns is adjusted dynamically so all of them can fit next to each other. If the screen is less than 960 pixel wide, the columns are displayed below each other.

There are four different kinds of interactions with an active TwitterWall:

- Pause and restart fetching new tweets
- Add and delete filtered columns
- Hide and show "All Tweets" column
- Change interval of tweet fetching

When the fetching timespan is in the future or in the past, the option to pause and restart fetching is not available.

6.2. Differences from TweetDeck

TweetDeck is a Twitter tool for real-time tracking, organizing and engagement.[2] It was already mentioned in chapter 3.

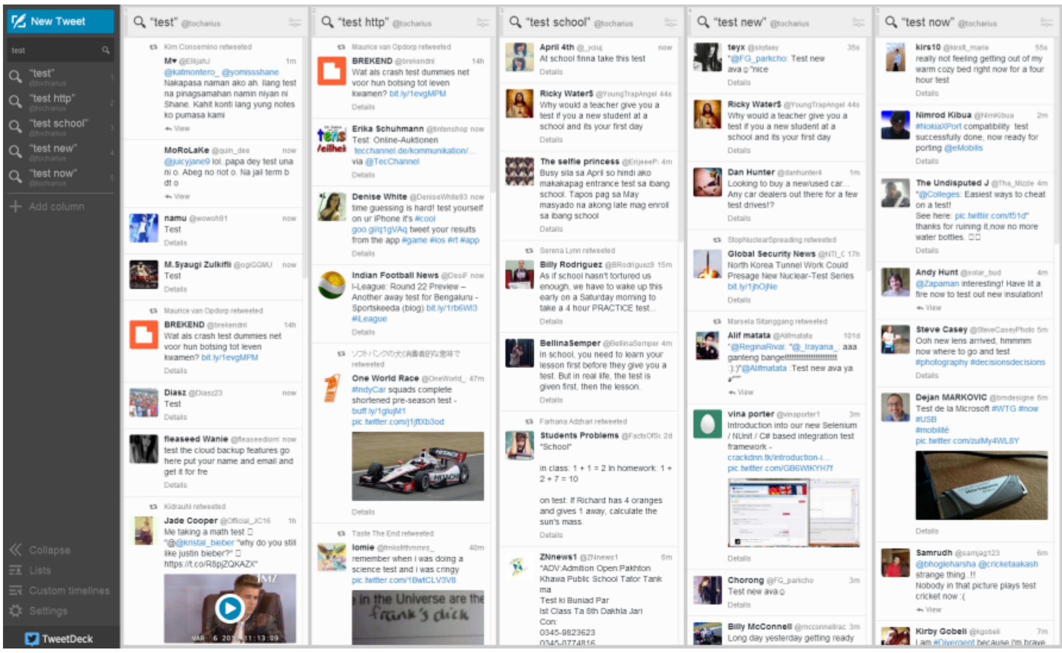
It provides similar features as TwitterWall, most notably the real-time tracking of search results.

A screenhot of TweetDeck in a similar usage mode as TwitterWall is shown in figure 6.3.

6.2.1. Tracking

TweetDeck does not rely on tweet archives like TwitterWall, but instead uses a direct link with the Twitter Streaming API. TwitterWall periodically polls the TweetCollector API for new tweets in an archive. This means that although new tweets are added to TweetCollector archives in real-time, they don't appear on the TwitterWall until the next polling. Therefore, TweetDeck can be classified as real-time, while TwitterWall is only near real-time. Tweets appear individually as they are written in TweetDeck, while they appear in

[2] https://about.twitter.com/products/tweetdeck, 2014-04-21



"test" @tocharius

Kim Consemino retweeted
M♥ @ElijahJ 1m
@katmontero_ @yomissshane Nakapasa naman ako ah. Ilang test na pinagsamahan namin niyan ni Shane. Kahit konti lang yung notes ko pumasa kami
View

MoRoLaKe @quin_dee now
@juicyjane9 lol..papa dey test una ni o. Abeg no riot o. Na jail term b dt o
View

namu @wowoh91 now
Test
Details

M.Syaugi Zulkifli @ogiCGMU now
Test
Details

Maurice van Opdorp retweeted
BREKEND @brekendnl 14h
Wat als crash test dummies net voor hun botsing tot leven kwamen? bit.ly/1evgMPM
Details

Diasz @Diasz23 now
Test
Details

fleaseed Wanie @fleaseedom now
test the cloud backup features go here put your name and email and get it for fre
Details

Kidrauhl retweeted
Jade Cooper @Official_JC16 1h
Me taking a math test □ "@@kristal_bieber "why do you still like justin bieber?" □ https://t.co/R8pjZQKAZX"

"test http" @tocharius

Maurice van Opdorp retweeted
BREKEND @brekendnl 14h
Wat als crash test dummies net voor hun botsing tot leven kwamen? bit.ly/1evgMPM
Details

Erika Schuhmann @tintenshop now
Test: Online-Auktionen tecchannel.de/kommunikation/… via @TecChannel
Details

Denise White @DeniseWhite93 now
time guessing is hard! test yourself on ur iPhone it's #cool goo.gl/q1gVAq tweet your results from the app #game #ios #rt #app
Details

Indian Football News @DesiF now
I-League: Round 22 Preview – Another away test for Bengaluru - Sportskeeda (blog) bit.ly/1rb6WI3 #ILeague
Details

ソフトバンクの犬は消費者的な意味で retweeted
One World Race @OneWorld_ 47m
#IndyCar squads complete shortened pre-season test - buff.ly/1glujM1 pic.twitter.com/j1jftXb3od
Details

Taste The End retweeted
Iomie @fmksfthmmrs_ 40m
remember when i was doing a science test and i was cringy pic.twitter.com/1BwtCLV3V8

"test school" @tocharius

April 4th @_yciuj now
At school finna take this test
Details

Ricky Water$ @YoungTrapAngel 44s
Why would a teacher give you a test if you a new student at a school and its your first day
Details

The selfie princess @ErjeeeP: 4m
Busy sila sa April so hindi ako makakapag entrance test sa ibang school. Tapos pag sa May masyado na akong late mag enroll sa ibang school
Details

Serena Lynn retweeted
Billy Rodriguez @BRodriguez9 15m
As if school hasn't tortured us enough, we have to wake up this early on a Saturday morning to take a 4 hour PRACTICE test…
Details

BellinaSemper @BellinaSemper 4m
In school, you need to learn your lesson first before they give you a test. But in real life, the test is given first, then the lesson.
Details

Farhana Adzhari retweeted
Students Problems @FactsOfSc 2d
"School"

in class: 1 + 1 = 2 In homework: 1 + 2 + 7 = 10

on test: If Richard has 4 oranges and gives 1 away, calculate the sun's mass.
Details

ZNnews1 @ZNnews1 6m
"ADV:Admition Open:Pakhton Khawa Public School Tator Tank ma
Test ki Buniad Par
Ist Class Ta 8th Dakhla Jari
Con:
0345-9823623
0345-0774816

"test new" @tocharius

teyx @skytaey 35s
"@FG_parkcho: Test new ava☺"nice
Details

Ricky Water$ @YoungTrapAngel 44s
Why would a teacher give you a test if you a new student at a school and its your first day
Details

Dan Hunter @danhunter4 1m
Looking to buy a new/used car… Any car dealers out there for a few test drives!?
Details

StopNuclearSpreading retweeted
Global Security News @NTI_C 17h
North Korea Tunnel Work Could Presage New Nuclear-Test Series bit.ly/1jhOjNe
Details

Marsela Sitanggang retweeted
Alif matata @Alifmatata 101d
"@ReginaRivai: "@_Irayana_: aaa ganteng banget!!!!!!!!!!!!!!!!!!!!! :):)"@Alifmatata :Test new ava ya ✌""
View

vina porter @vinaporter1 3m
Introduction into our new Selenium / NUnit / C# based integration test framework - crackdnn.tk/introduction-i… pic.twitter.com/GB6WIKYH7f
Details

Chorong @FG_parkcho 3m
Test new ava☺
Details

Billy McConnell @mcconnellrac 3m
Long day yesterday getting ready

"test now" @tocharius

kirs10 @kirsti_marie 55s
really not feeling getting out of my warm cozy bed right now for a four hour test
Details

Nimrod Kibua @NimKibua 2m
#NokiaXPort compatibility test successfully done, now ready for porting @eMobilis
Details

The Undisputed J @Tha_Mizzle 4m
"@Colleges: Easiest ways to cheat on a test!
See here: pic.twitter.com/f51d"
thanks for ruining it,now no more water bottles. □□
Details

Andy Hunt @solar_bud 4m
@Zapaman interesting! Have lit a fire now to test out new insulation!
View

Steve Casey @SteveCaseyPhoto 5m
Ooh new lens arrived, hmmmm now where to go and test #photography #decisionsdecisions
Details

Dejan MARKOVIC @bmdesigne 6m
Test de la Microsoft #WTG #now #USB #mobilité pic.twitter.com/zulMy4WL8Y
Details

Samrudh @samjag123 6m
@bhogleharsha @cricketaakash strange thing..!! Nobody in that picture plays test cricket now :(
View

Kirby Gobell @kgobell 7m
I am #Divergent because I'm brave

Figure 6.3.: TweetDeck. Source: https://tweetdeck.twitter.com, 2014-04-21

batches on TwitterWall. Either of those can be preferable depending on the volume of new tweets.

6.2.2. Pausing

TwitterWall has the ability to pause fetching tweets, a feature that TweetDeck misses. When fetching is resumed, all tweets since the last check are added at the same time.

TweetDeck fetches new tweets all the time, but the timeline jumps to each new tweet when scrolled all the way to the top. If a user scrolls down, new tweets are added to the top without changing the position in the timeline.

Both of those behaviours have different advantages and disadvantages.

6.2.3. Old Tweets

The area where TwitterWall has a clear advantage over TweetDeck is displaying older tweets. TweetDeck is bound to the same rules as any other Twitter API client, which includes the limit of accessing older tweets. Unless TweetDeck is running when a certain tweet is written, there is no guarantee that the Twitter API can find it again after the fact. When a TweetCollector archive is created soon enough, all tweets are archived and accessible to TwitterWall. This enables TwitterWall to better and more completely show tweets from events in the past. In this regard, TweetDeck can be seen as no more than a more advanced Twitter client, while TwitterWall is a filtering tool that can deal with past archives as well as current tweets.

6.2.4. Events

Due to the different feature sets of TwitterWall and TweetDeck, TwitterWall is more suited for use after events. When tweets from an event like a conference are archived, TwitterWall can provide an overview of what people wrote during the conference. The real-time nature of TweetDeck makes this tool less suited for this task.

7. Use Cases

So far, this thesis has introduced three types of tools: one for Twitter archiving, one for analysis and a third for filtering and tracking. This chapter shows how these tools can be applied to real world use cases and what value and insight can be gained from their usage.

7.1. Analysis of EMOOCS 2014 Conference with TwitterStat

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.

Due to this fact, it can be assumed that when archiving the tweets with a certain conference hashtag, a high percentage of tweets about the conference are caught. Because of the need to start the archiving process early enough, creating an archive as soon as the hashtag is known is preferable. This ensures no tweets are missed.

As with any event, there will be tweets before it begins and after it finishes. This provides the user with the option to analyze all tweets, or limit the date range to just look at tweets before, after or during the conference.

The conference "EMOOCs 2014" was the second European MOOCs Stakeholder Summit.[1] It was held in February 2014 in Lausanne, Switzerland.

The conference aims to be a meeting place of European participants in the MOOC[2] 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 was "#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 thesis are shortened.

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

```
http://twitterstat.tugraz.at/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 followers.

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 tweets contain links.

A user can view the actual tweets containing links:

```
http://twitterstat.tugraz.at/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 then shows several lists:

[1] http://www.emoocs2014.eu, 2014-04-21

[2] Massive Open Online Course

- 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), un (118), business (117), sur (116), students (116), online (113), via (112), this (112), but (111), et (106), european (104), social (102), how (101), one (101), courses (99), learners (98), en (97), who (96), big (95), first (95), can (94), there (91), so (89), course (87), all (87), see (86), university (86), model (85), video (84), or (84), very (82), just (81), interesting (81), 1 (81), presentation (81), they (81), do (80), keynote (80), ? (80), research (80), now (80), also (79), du (79), when (77), à (77), conference (76), day (76), universities (75), need (75), our (75), use (73), their (73), ...
- 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.[3]

[3] http://www.emoocs2014.eu/speaker/simon-nelson, 2014-04-21

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, one 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.

`http://twitterstat.tugraz.at/analysis.html?archive=%23emoocs2014&parameter=@yprie`

- which @persons talk to @yprie about #emoocs2014
  snesterko (1)
- who does @yprie talk to about #emoocs2014
  snesterko (2), tbirdcymru (2), q5x (1), wfvanvalkenburg (1), morgan_it (1), audreyego's (1), limabonas (1)
- who else is addressed with @yprie about #emoocs2014
  tbirdcymru (1)
- which keywords are used by @yprie about #emoocs2014
  rt (39), to (22), the (20), of (17), in (16), a (16), is (14), for (11), mooc (10), not (9), as (7), open (7), universities (6), moocs (6), be (6), - (5), by (5),

about (5), platform (5), google (4), have (4), at (4), from (4), learning (4), are (4), interested (4), i (3), students (3), business (3), courses (3), delivery (3), can (3), stream (3), ...

- 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 first three answers give a more detailed view of the interaction of the Twitter users in the context of the conference: Who talks to whom, who is mentioned and which people are addressed together.

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″.

```
http://twitterstat.tugraz.at/analysis.html?archive=
%23emoocs2014&parameter=%23futurelearn
```

- which @persons write #emoocs2014 together with #futurelearn
  bobreuter (5), FactoryMOOC (3), debrahumphris (2), mLearnopedia (2), yveszieba (2), LT_tech_HE (2), ...

- 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), cost (4), for (4), guideline (4), — (4), £30000 (4), panel (4), webs: (4), 6 (4), week (4), interesting (3), foster (3), many (3), rich (3), sessions (3), all (3), forms (3), their (3), online (3), environment (3), values (3), were (3), conversations (3), ... (3), auch (2), during (2), exact (2), futurelearn (2), one (2), suites (2), openuniversity (2), via (2), relationship (2), best (2), what (2), inge (2), (myblog) (2), aus (2), loneliness. (2), announces (2), barrier (2), big (2), ? (2), collaboration (2), conscious (2), effort (2), needed (2), aspects (2), design (2), ocean (2), ...
- 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 #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.

This is only one possible way in which TwitterStat can be utilized. The same kind of analysis can be applied to person and keyword archives.

7.2. Tracking Tweets during Lectures with TwitterWall

The prime use case for TwitterWall is surveying the reactions of students during a lecture. This was done during lectures of the course ″Social Aspects of Information Technology″ at Graz University of Technology in the summer semester of 2014.

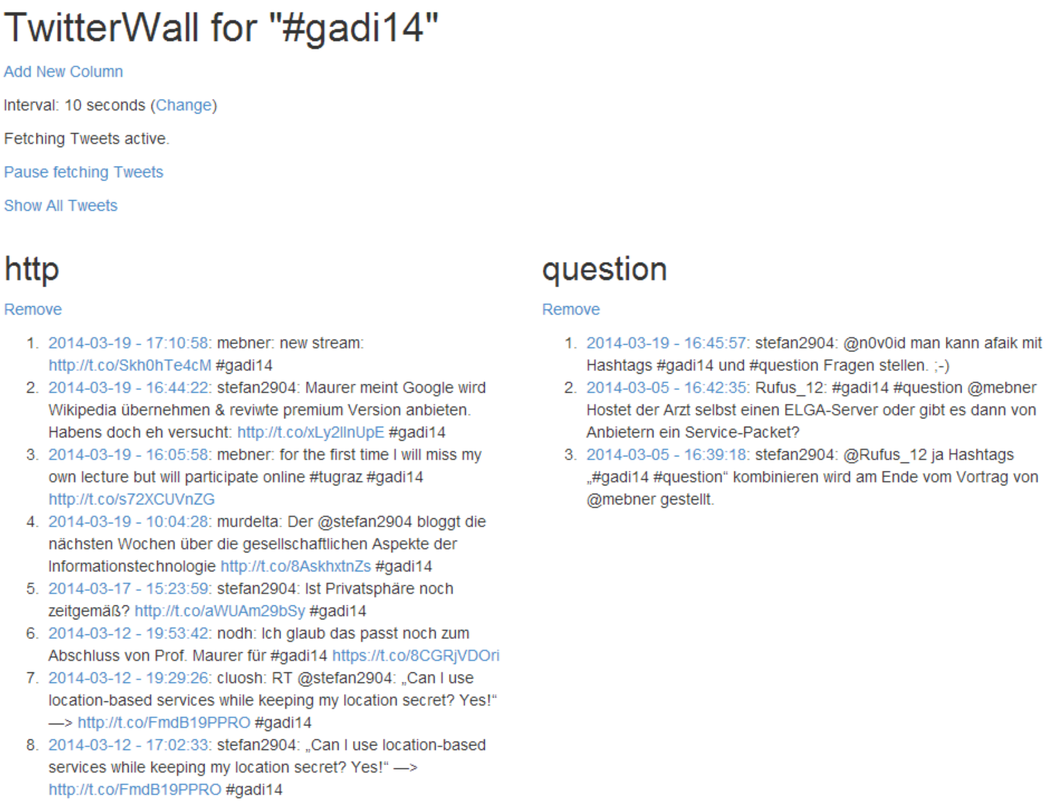
Students were asked to tag their tweets about the lecture with the hashtag ″#gadi14″. If they had questions they wanted to be answered, they could use the additional tag ″#question″.

The lecture consists of guest presentations by various experts. Using TwitterWall, the organizer of the lecture was able to look at questions from students, evaluate them, and ask the experts on behalf of the students.

The fetching of new tweets can be paused if the answering of a question takes longer or more questions are pouring in. After the fetching is resumed, all new tweets are fetched so there are no time gaps for tweets to go missing.

After the lecture is finished, a TwitterWall with a date range matching the lecture times can be created. This wall is static because no new tweets are added. It can be used to look at comments and questions during the lecture and improve it for the next time it is held.

Figure 7.1 shows a TwitterWall during a lecture of ″Social Aspects of Information Technology″. The column ″All Tweets″ is hidden, and the two filtered columns ″http″ and ″#question″ are shown. The first columns displays all links posted with the hashtag ″#gadi14″, while the second columns shows all tweets tagged as questions.



TwitterWall for "#gadi14"

Add New Column

Interval: 10 seconds (Change)

Fetching Tweets active.

Pause fetching Tweets

Show All Tweets

http

Remove

1. 2014-03-19 - 17:10:58: mebner: new stream: http://t.co/Skh0hTe4cM #gadi14
2. 2014-03-19 - 16:44:22: stefan2904: Maurer meint Google wird Wikipedia übernehmen & reviwte premium Version anbieten. Habens doch eh versucht: http://t.co/xLy2llnUpE #gadi14
3. 2014-03-19 - 16:05:58: mebner: for the first time I will miss my own lecture but will participate online #tugraz #gadi14 http://t.co/s72XCUVnZG
4. 2014-03-19 - 10:04:28: murdelta: Der @stefan2904 bloggt die nächsten Wochen über die gesellschaftlichen Aspekte der Informationstechnologie http://t.co/8AskhxtnZs #gadi14
5. 2014-03-17 - 15:23:59: stefan2904: Ist Privatsphäre noch zeitgemäß? http://t.co/aWUAm29bSy #gadi14
6. 2014-03-12 - 19:53:42: nodh: Ich glaub das passt noch zum Abschluss von Prof. Maurer für #gadi14 https://t.co/8CGRjVDOri
7. 2014-03-12 - 19:29:26: cluosh: RT @stefan2904: „Can I use location-based services while keeping my location secret? Yes!“ —> http://t.co/FmdB19PPRO #gadi14
8. 2014-03-12 - 17:02:33: stefan2904: „Can I use location-based services while keeping my location secret? Yes!“ —> http://t.co/FmdB19PPRO #gadi14

question

Remove

1. 2014-03-19 - 16:45:57: stefan2904: @n0v0id man kann afaik mit Hashtags #gadi14 und #question Fragen stellen. ;-)
2. 2014-03-05 - 16:42:35: Rufus_12: #gadi14 #question @mebner Hostet der Arzt selbst einen ELGA-Server oder gibt es dann von Anbietern ein Service-Packet?
3. 2014-03-05 - 16:39:18: stefan2904: @Rufus_12 ja Hashtags „#gadi14 #question“ kombinieren wird am Ende vom Vortrag von @mebner gestellt.

Figure 7.1.: A TwitterWall during a Lecture. Source: http://twitterwall.tugraz.at

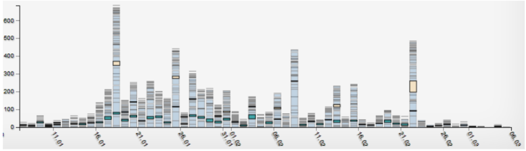


Figure 7.2.: MMIS2 example of tweets over time. Source: MMIS2 2013 report by Group 12

7.3. Providing Data for Visualizations with TweetCollector

TweetCollector is the centerpiece of this collection of tools, but in and of itself its usefulness is limited. The added value from collecting tweets is how these archives are used by tools relying on TweetCollector. This makes the API the most valuable part of TweetCollector.

TwitterStat and TwitterWall are just two ways to utilize this data. The lecture ″Multimedia Information Systems 2″ at Graz University of Technology teaches data visualization. In the years 2013 and 2014, tweet archives from TweetCollector were used as a basis to create visualizations.

The students could choose from four different questions:

- Who talks to whom, and how much?
- What reach do retweets have?
- What are the role of retweets and mentions in communication?
- When are people tweeting the most?

Other than these broad questions, the details of the implementation of these visualizations were up to the students. Examples of implementations are shown in figures 7.2 and 7.3.

TweetCollector enabled the students to perform these tasks on large sets of data without having to crawl a social network for themselves.

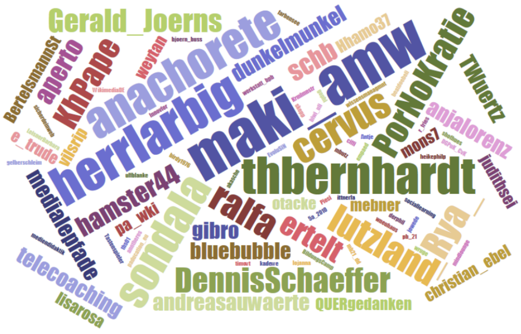


Figure 7.3.: MMIS2 example most active users. Source: MMIS2 2013 report by Group 5

7.4. Use Case Summary

As can be seen in the previous examples, TweetCollector, TwitterStat and TwitterWall can serve in a variety of use cases. The types of usage enabled are very different, yet they all stem from the basic concept of archiving tweets and using the resulting archives in interesting ways.

The results achieved are interesting, yet there is still room for improvement. The next chapter discusses how these results compare to existing research, and how they are relevant for the posed research questions. Chapter 9 suggests some further enhancements for these tools.

8. Discussion

So far, this thesis has given an overview of the state of the art of scientific research concerning Twitter, as well as the existing online tools to conduct Twitter archival and analysis. Then, the tools developed in the scope of this work have been introduced and demonstrated using real examples. With this information, the research questions can be discussed.

8.1. Value of Twitter Archives

The first question was "What value can tweet archives provide?". Primarily, tweet archives enable access to tweets too old to be found by the Twitter search engine. This encompasses more than 3200 tweets for single Twitter users, and tweets older than six to eight days for regular search terms or hashtags.

When looking at the broad spectrum of scientific research discussed in chapter 3, it can be seen that most types of analysis need a corpus of tweets to analyze. This can't be achieved by querying the Twitter API for tweets at the time of analysis due to lack of availability of old tweets.

Most researchers use their own software to crawl and archive tweets. However, the kinds of archives are mostly the same. They are either hashtag archives, or keyword archives. This duplication of effort on different crawling tools wastes a lot of time that could be put to better use analyzing the archived tweets.

The software yourTwapperKeeper tries to solve this problem by providing an open source tool that anyone can use to archive tweets on their own

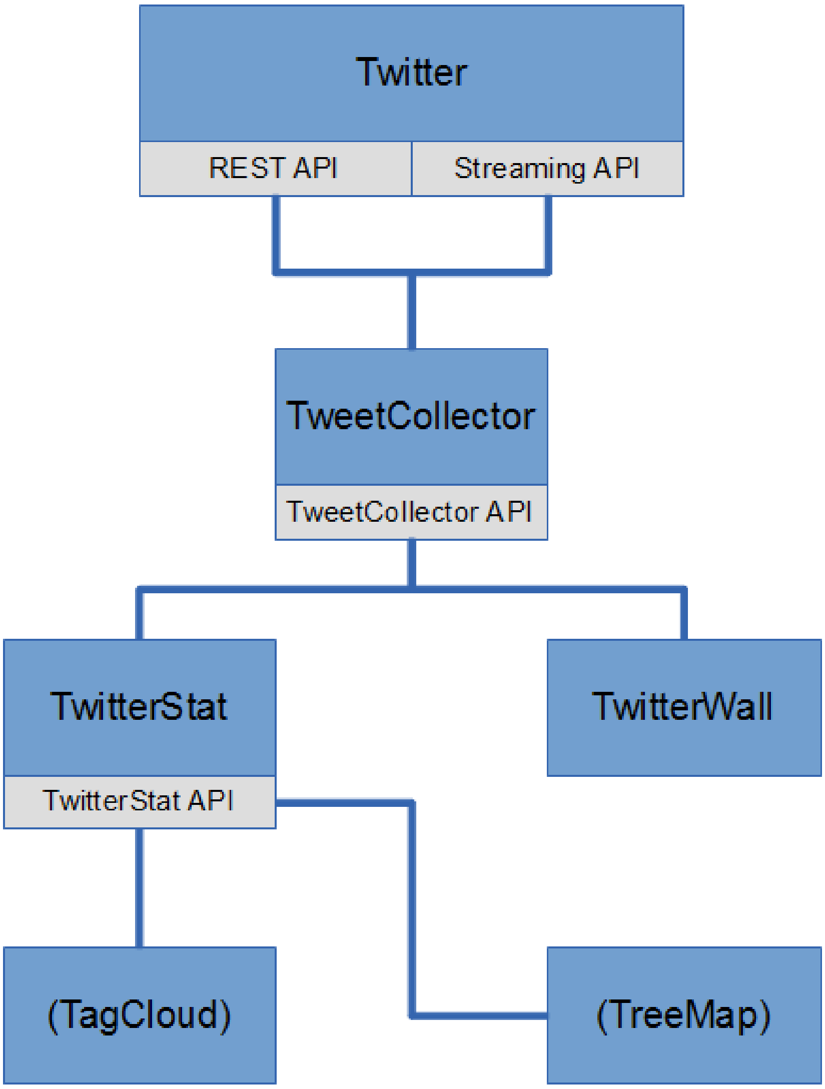


Figure 8.1.: Tree structure of applications using TweetCollector. Source: Own illustration

servers. However, due to changes in the Twitter API and incomplete storage of some tweet metadata, it is not an ideal solution.

TweetCollector tries to improve on this by taking yourTwapperKeeper as the foundation, and building on it. By providing more complete storage of tweets, as well as adding the ability to create person archives, TweetCollector provides more value than yourTwapperKeeper.

The most important part of TweetCollector is the API that provides access to the stored data in a machine readable format. This enables subsequent applications to use the archived tweets in any way desired. Possible uses of this API are demonstrated by the applications TwitterStat and TwitterWall, or the usage of archives for visualization purposes in courses at Graz University of Technology.

Just like the Twitter API enables a whole ecosystem of apps that interact with it, the TweetCollector API enables the same dynamic for tweet archives. If the applications using TweetCollector provide an API of their own, this reuse of data can be replicated again. For example, TwitterStat has an API that provides access to the raw analysis data. This data can be used for visualizations or tag clouds. As depicted in figure 8.1, a whole tree structure of applications can be developed this way, with all of them relying on TweetCollector as the root.

To summarize, tweet archives created by TweetCollector provide the following value:

- Access to old tweets not available through Twitter API
- More complete metadata storage
- Ability to create hashtag, keyword and person archives
- Open API to build applications using tweet archives for analysis, visualization, filtering and other uses

8.2. Value of Context for Analysis

The second question posed in this thesis was ”What value can the context of an analysis provide?”.

When surveying the available literature, one can see 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 can 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.

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.

9. Outlook and Future Works

The tools described in this paper provide a broad spectrum of features useful for retrieval, storage, analysis and visualisation of tweets. The possibilities of this kind of work are not yet exhausted. Due to the modular structure of the toolset and accessible APIs, there is a foundation on which further extensions can be built. This chapter details some of the refinements and enhancements possible.

9.1. General Enhancements

The suggestions listed here apply to TweetCollector, TwitterStat and TwitterWall.

9.1.1. User Interface

All three applications described in this thesis have a very bare-bones user interface. The use of Bootstrap allowed for basic layouting and responsiveness, but other than that no design flourishes were applied.

The reason for this is the focus of these prototypes lies on functionality. To attract more casual users, a more polished user interface would be helpful. Due to the separation of content, logic and styling, applying any design should be an easy task.

9.1.2. Landing Pages

The feature set of these tools is large and not self-explanatory. To make it possible for everyone to understand what the software does and how to use it, landing pages with explanations and examples would be useful.

9.2. TweetCollector Enhancements

TweetCollector is the most complex part of the tools introduced in this thesis. The reason for this is that it is the only part that interacts with the Twitter API itself, and that retrieving and storing tweets continually is difficult. This sections recommends some enhancements to this component.

9.2.1. Entities

TweetCollector only stores a subset of the information the Twitter API provides about each tweet. If more metadata is stored, more information can be provided to subsequent tools.

Entities are types of additional data which have been parsed from the text of a tweet.[1] There are 5 types of entities: media, urls, user_mentions, hashtags, and symbols.

This can be especially valuable for media and URLs, because both of those appear as a `http://t.co` shortened link in the regular text of a tweet. These links makes it difficult to extract the entity behind the shortened link because each link has to be visited to get this information. For use cases like the analysis of TwitterStat, this is infeasible because of the time involved in resolving each link individually. Storing these entities would enable analysis of most tweeted pictures and a better display of the most tweeted links.

The entities user_mentions and hashtags might be useful to simplify the analysis process of most tweeted hashtags and usernames.

[1] https://dev.twitter.com/docs/entities, 2014-04-21

9.2.2. User Management

At the moment TweetCollector supports only very basic user management and authentication. Users need to have Twitter accounts which are hard-coded in the config.php file. They are authenticated with Twitter using OAuth and have to log in again every time they close the browser. There is no distinction between users, so every user can create and delete archives and start/stop the archiving process, regardless of who created a specific archive. This design decision keeps the tool simple, but it necessitates that only trusted users are added to the list.

An advanced user management (whether based on Twitter accounts or not) with different roles for users (administrator, regular user) and an automated sign-up process would allow a single instance of TweetCollector to work for more users without compromising archiving for each of them.

9.2.3. Advanced OAuth Token Usage

If the user management mentioned in section 9.2.2 is implemented and every user has to sign in with his or her Twitter account, the archiving processes could authenticate with the Twitter API using each users credentials. This would allow faster retrieval of tweets, because the API limits would count against multiple users and not just against one user. TweetCollector would also be able to operate more active archives at the same time.

9.2.4. Rewrite in Java

TweetCollector is written in PHP because its predecessor yourTwapperkeeper was written in PHP. Over time, it became obvious that the choice of this programming language was not ideal. TweetCollector relies on four processes that are always running to provide archiving functionality. At the time of this writing, PHP still lacks the robust process management functions of other programming languages.

Various workarounds have been employed to mitigate this. PHP command line interface is used to manually start, stop and check the processes using common UNIX commands. When the webserver has to be restarted for any reason, these processes don't start again by themselves. Therefore, a cron job is run every 15 minutes. This cron job checks if the processes should be running and if they are actually running, and restarts them if they are not.

Both of these workarounds lead to the dependence on a UNIX based operating system. A more elegant solution would be a TweetCollector written in Java, running on a Java server like Tomcat. The processes could be redeployed on every start of the server without reliance on any cron jobs, and Java process management is very mature and refined. It would also enable TweetCollector to run on any operating system where Java servers are available.

Rewriting the entire program would be a massive undertaking and was not realized in the course of this work.

9.3. TwitterStat Filtering

The current version of TwitterStat counts every word it finds. This provides the maximum amount of analysis data, but can obscure the information a user is looking for in between many common words that are used in regular sentences.

Consider this list of the 25 most used keywords from the archive "#emoocs2014":

rt (2365), the (1407), of (1135), to (1113), a (1006), in (856), is (787), and (757), for (683), at (674), moocs (627), mooc (529), on (453), - (370), de (356), are (339), from (324), by (311), not (300), about (296), learning (286), with (275), : (269), la (255), data (224)

"rt" can signify a retweet, but not all mentions of "rt" are retweets that fit our definition. Showing this entry gives the user the option to look at all tweets containing ″rt″, even if they are not retweets. Hiding this entry draws more attention to the other relevant content.

This list also contains the words ″the″, ″of″, ″to″, ″a″, ″in″, ″is″, ″and″, ″for″, and ″at″ before ″moocs″ and ″mooc″. One might argue that these words are common in the English language and can be filtered out. This makes ″mooc(s)″ the most used words in this archive, which is only logical for a conference about massive open online courses. A different argument is that these words do provide some valuable information. The presence of ″is″ and the absence of ″was″ (which is far further down the list with 55 mentions) shows that most tweets are about the present instead of the past. Similar arguments can apply to ″for″ and ″against″, ″are″ and ″were″ and lots of other common words.

If all of those common words are removed from the list, only ″mooc(s)″, ″learning″ and ″data″ are left. This might be preferable to casual observers, but each removal of an element risks hiding data that can lead to valuable insights.

Different people will have different preferences regarding the level of detail wanted in the analysis. Therefore, an optional feature to filter out common words would be a good idea for further improvements of TwitterStat. This could be done before starting the analysis or afterwards in the user interface. A blacklist[2] defined by the user could be employed. If anything, this section shows that there is no definite right or wrong way of doing this, and the choice should, if possible, be left to the user.

9.4. TwitterWall Display of Tweets

At the moment, TwitterWall is mostly a proof-of-concept application. It was implemented in very little time, re-using components from TwitterStat. This shows that a tool building on TweetCollector archives can be functional very quickly, but it also means the user interface is very basic. To attract more casual users, the display of tweet content needs to be more visually appealing. TweetDeck is a good example of how this can be achieved.

[2] A list of disallowed words

9.5. Mobile Applications

Internet usage in general and Twitter usage specifically are getting more mobile. To support this usage scenario, TweetCollector, TwitterStat and TwitterWall all have webpages with responsive layout to make them usable on smaller screens.

An even better user experience can be provided using native applications for mobile platforms like iOS and Android. At the time of this writing, an iOS application for TwitterWall is being developed as part of a Bachelor's thesis at Graz University of Technology. Applications for the other tools and on other platforms are planned.

9.6. Semantic Research

As mentioned in the chapter on existing scientific research, Softic et al. conducted semantic research on Twitter with the help of an older version of TwitterStat [Softic et al., 2010]. This research is still ongoing. With the improvements to TweetCollector and TwitterStat, it is possible that new insights can be gained.

10. Conclusion

The goal of this thesis is to show the potential of Twitter archives. To achieve this, several topics were explored.

The state of the art of current academic research on Twitter, as well as existing tools of Twitter archiving, analysis and filtering 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.

The existing Twitter tools provide different ways of archiving and analyzing tweets. None of these tools fulfilled the specific needs of this work, so a new set of tools was developed.

A tweet archiving tool called TweetCollector was created and presented. 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.

Following TwitterStat, the filtering tool TwitterWall was developed. TwitterWall provides the ability to monitor new tweets in an archive in real-time and filter these tweets according to user-definable criteria.

Several use cases for the application of this suite of tools were covered. Archives from TweetCollector were used to create visualizations in lectures.

TwitterStat was used to analyze tweets from a conference. TwitterWall served as a real-time audience response system during keynotes.

Afterwards, these results were discussed. It was shown that TweetCollector provides value by having more complete metadata storage and more types of available archives than comparable tools. The open API can be used to build application relying on this data. The ”back to tweets” feature of TwitterStat was shown to be valuable for determining context of the original tweets.

Both research questions were answered. It was shown that tweet archives and context analysis provide significant benefits. This validates the application of these tools, because there is an advantage for the user.

Future improvements and extensions of these tools were proposed, to make the prototypes more user-friendly. The archiving capabilities of TweetCollector can be extended further as well.

This thesis shows that the data provided by Twitter itself is not sufficient for many applications. The retrieval and storage of data from Twitter is necessary to create persistent archives of tweets available for further usage.

These tweet archives enable a variety of new 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 thesis can be valuable for a variety of users.

Bibliography

T. Altmann. Erschließung und analyse von twitter analyse tools. Bachelor's thesis, Graz University of Technology, 2010.

J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

d. boyd, S. Golder, and G. Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10. IEEE, 2010.

A. Bruns and S. Stieglitz. Towards more systematic twitter analysis: Metrics for tweeting activities. *International Journal of Social Research Methodology*, 16(2):91–108, 2013.

M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10:10–17, 2010.

M. Ebner. Introducing live microblogging: How single presentations can be enhanced by the mass. *Journal of research in innovative teaching*, 2(1):91–100, 2009.

M. Ebner. The influence of twitter on the academic environment. *Social Media and the New Academic Environment: Pedagogical Challenges. IGI Global*, pages 293–307, 2013.

M. Ebner and H. Maurer. Can weblogs and microblogs change traditional scientific writing? *Future Internet*, 1(1):47–58, 2009.

M. Ebner and W. Reinhardt. Social networking in scientific conferences–twitter as tool for strengthen a scientific community. In *Proceedings of the 1st International Workshop on Science*, volume 2, pages 1–8, 2009.

M. Ebner and M. Schiefner. Microblogging-more than fun. In *Proceedings of IADIS mobile learning conference*, volume 2008, pages 155–159, 2008.

M. Ebner, C. Lienhardt, M. Rohs, and I. Meyer. Microblogs in higher education–a chance to facilitate informal and process-oriented learning? *Computers & Education*, 55(1):92–100, 2010.

M. Ebner, T. Altmann, and S. Softic. @ twitter analysis of# edmedia10–is the# informationstream usable for the# mass. *Form@ re-Open Journal per la formazione in rete*, 11(74):36–45, 2011.

C. Honeycutt and S. C. Herring. Beyond microblogging: Conversation and collaboration via twitter. In *System Sciences, 2009. HICSS'09. 42nd Hawaii International Conference on*, pages 1–10. IEEE, 2009.

B. A. Huberman, D. M. Romero, and F. Wu. Social networks that matter: Twitter under the microscope. *arXiv preprint arXiv:0812.1045*, 2008.

B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

A. M. Kaplan and M. Haenlein. The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons*, 54(2):105–113, 2011.

B. Kelly, M. Hawksey, J. O'Brien, M. Guy, and M. Rowe. Twitter archiving using twapper keeper: technical and policy challenges. In *7th International Conference on Preservation of Digital Objects (iPRES 2010)*. University of Bath, 2010.

B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proceedings of the first workshop on Online social networks*, pages 19–24. ACM, 2008.

H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010.

G. Lotan, E. Graeff, M. Ananny, D. Gaffney, I. Pearce, and d. boyd. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International Journal of Communication*, 5:31, 2011.

H. Mühlburger, M. Ebner, and B. Taraghi. twitter try out# grabeeter to export, archive and search your tweets. *Research 2.0 approaches to TEL*, pages 1–9, 2010.

L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.

A. Pak and P. Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, 2010.

W. Reinhardt, M. Ebner, G. Beham, and C. Costa. How people are using twitter during conferences. *5th EduMedia conference Salzburg*, pages 145–156, 2009.

T. Sakaki, M. Okazaki, and Y. Matsuo. 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, 2010.

S. Softic, M. Ebner, H. Mühlburger, T. Altmann, and B. Taraghi. twitter mining# microblogs using# semantic technologies. In *6th Workshop on Semantic Web Applications and Perspectives, SWAP*, pages 1–12, 2010.

T. Terpstra, A. de Vries, R. Stronkman, and G. Paradies. Towards a realtime twitter analysis during crises for operational crisis management. In *ISCRAM'12: Proceedings of the 9th International ISCRAM Conference*, 2012.

A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

S. Vieweg, A. L. Hughes, K. Starbird, and 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*, pages 1079–1088. ACM, 2010.

D. Zhao and 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*, pages 243–252. ACM, 2009.

Appendix

Appendix A.

API Documentation

A.1. TweetCollector

TweetCollector has three different API endpoints.

A.1.1. info.php

Info supplies information about a specific tweet archive of TweetCollector.

URL

`http://tweetcollector.tugraz.at/api/info.php`

Parameters

- screen_name: User name of a Twitter user that is being archived
- user_id: Twitter ID number of a Twitter user that is being archived
- keyword: Word or Hashtag of a TweetCollector Keyword/Hashtag archive
- id: Numerical ID of a TweetCollector Keyword/Hashtag archive

Usage

Only one of the parameters can be used with a single request. The usage of exactly one parameter is required.

Example: `http://tweetcollector.tugraz.at/api/info.php?keyword=%23gadi14`

A.1.2. list.php

List provides a list of all tweet archives of TweetCollector.

URL

`http://tweetcollector.tugraz.at/api/list.php`

Parameters

none

Usage

No parameters are necessary. The API returns the complete list of available archives.

A.1.3. tweets.php

Tweets either returns all tweets from one archive, or tweets from one archive within a defined timeframe.

URL

`http://tweetcollector.tugraz.at/api/tweets.php`

Parameters

- screen_name: User name of a Twitter user that is being archived
- user_id: Twitter ID number of a Twitter user that is being archived
- keyword: Word or Hashtag of a TweetCollector Keyword/Hashtag archive
- id: Numerical ID of a TweetCollector Keyword/Hashtag archive
- start(optional): start date for the returned tweets as UNIX time stamp
- end(optional): end date for the returned tweets as UNIX time stamp

Usage

Only one of the parameters screenname, user_id, keyword or id can be used with a single request. The usage of exactly one of these parameter is required. The parameters start and end are optional.

Example: `http://tweetcollector.tugraz.at/api/tweets.php?keyword=%23gadi14&start=1399461770`

A.2. TwitterStat

TwitterStat has four different API endpoints.

A.2.1. analyze.php

Analyze returns analysis results for a specified tweet archive.

URL

```
http://twitterstat.tugraz.at/api/analyze.php
```

Parameters

- archive(required): the name of the archive to be analyzed; person archives start with the symbol ″@″
- parameter(optional): a parameter to make an analysis more specific; person parameters start with the symbol ″@″
- start(optional): start date for the analysis as UNIX time stamp
- end(optional): end date for the analysis as UNIX time stamp

Usage

The parameter archive is required, the other three are optional.

Example: `http://twitterstat.tugraz.at/api/analyze.php?archive=%23tugraz&parameter=@tocharius`

A.2.2. info.php

Info supplies information about a specific tweet archive.

URL

```
http://twitterstat.tugraz.at/api/info.php
```

Parameters

- archive(required): the name of the archive to be analyzed; person archives start with the symbol ″@″

Usage

The parameter archive is required.

Example: `http://twitterstat.tugraz.at/api/info.php?archive=%23tugraz`

A.2.3. list.php

List provides a list of all tweet archives.

URL

`http://twitterstat.tugraz.at/api/list.php`

Parameters

none

Usage

No parameters are necessary. The API returns the complete list of available archives.

A.2.4. tweets.php

Tweets returns a subset of tweets from one archive filtered according to the supplied parameters.

URL

`http://twitterstat.tugraz.at/api/tweets.php`

Parameters

- archive(required): the name of the archive; person archives start with the symbol "@"
- start(optional): start date for the returned tweets as UNIX time stamp
- end(optional): end date for the returned tweets as UNIX time stamp
- from(optional): tweets from a specified username
- mention1(optional): tweets where a specified username is mentioned
- mention2(optional): tweets where a second specified username is mentioned
- word1(optional): tweets where a specific word or hashtag is mentioned
- word2(optional): tweets where a second specific word or hashtag is mentioned
- rt(optional): tweets that are retweets
- links(optional): tweets that contain hyperlinks
- safelinks(optional): tweets that contain hyperlinks with encryption (HTTPS)
- source(optional): tweets written with a specified Twitter client

Usage

The parameter archive is required, the other optional parameters are used to filter the returned tweets.

Example: `http://twitterstat.tugraz.at/api/tweets.php?archive=%23tugraz&from=mebner&safelinks=true`

A.3. TwitterWall

TwitterWall has three different API endpoints.

A.3.1. list.php

List provides a list of all tweet archives.

URL

`http://twitterwall.tugraz.at/api/list.php`

Parameters

none

Usage

No parameters are necessary. The API returns the complete list of available archives.

A.3.2. time.php

Time returns the current UNIX timestamp of the server.

URL

`http://twitterwall.tugraz.at/api/time.php`

Parameters

none

Usage

No parameters are necessary. The API returns the current UNIX timestamp of the server.

A.3.3. tweets.php

Tweets either returns all tweets from one archive, or tweets from one archive within a defined timeframe.

URL

`http://twitterwall.tugraz.at/api/tweets.php`

Parameters

- archive(required): the name of the archive; person archives start with the symbol ”@”
- start(optional): start date for the returned tweets as UNIX time stamp
- end(optional): end date for the returned tweets as UNIX time stamp

Usage

The parameter archive is required, the other optional parameters are used to filter the returned tweets.

Example: `http://twitterwall.tugraz.at/api/tweets.php?archive=%23mwc14&start=1399464410&end=1399464420`