

Master's Thesis

Visualizing Scientific Conference Tweets using Clustering



David Pocivalnik

Knowledge Technologies Institute

Graz University of Technology

A thesis submitted for the degree of

Master of Science (MSc)

October 2013

Supervisor: Univ.-Prof. Dipl.-Inf. Dr. Stefanie Lindstaedt

Advisor: Mag.rer.soc.oec. Peter Kraker, BSc

Abstract

Twitter data has been used in research for various applications. Lately the scientific community has recognized the potential of Twitter and makes use of it during scientific conferences. This may indicate that interesting information is provided by the scientific community during scientific conferences. Nevertheless, it is almost impossible to read all tweets published during a conference, and furthermore to manually extract the interesting information out of all the tweets. For instance, during the WWW2012 conference 6901 tweets have been published with the conferences' designated hash-tag *#www2012* during the days the conference was held.

This work describes the implementation and evaluation of a system that clusters tweets published in the context of a scientific conference. The resulting clusters are then visualized in order to make them easily understandable for humans. The conducted evaluation of the system used the tweets published during the WWW2012 conference. The evaluation results indicate that the results produced by the system not only support topic extraction, but also organizational event extraction. Moreover, the results revealed the need for a refined clustering technique and additional processing in order to visualize relations in between clusters.

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)

Contents

Glossary	v
1 Introduction	1
1.1 Motivation	1
1.2 Objective	2
1.3 Outline	4
2 Twitter	5
2.1 Conventions	5
2.2 The role of Tweeters	6
2.3 The role of Twitter	8
2.4 Quality of Twitter data	9
2.5 Twitter in Research	10
2.6 Twitter API	12
3 Clustering	13
3.1 Concepts related to Clustering	13
3.1.1 Weighting	13
3.1.2 Similarity Measures	14
3.1.3 Labeling	14
3.2 Overview of Clustering Algorithms	15
3.3 k -Means	16
3.4 k -Means Extensions and Variations	17
3.5 Clustering Twitter Data in Research	18

4	Visualization	20
4.1	Visualization types	21
4.1.1	Chart	21
4.1.2	Map	21
4.1.3	Network	23
4.1.4	Time	24
4.1.5	Hierarchy	25
4.2	Interactive Visualization	28
5	Overview of Practical Work	30
6	System Design & Implementation Basis	34
6.1	System Design Overview	34
6.2	The Crawler	36
6.2.1	Tweet crawling and pre-processing	36
6.2.2	Solr	37
6.2.3	Solr index fields	37
6.3	Preparatory Work	38
7	Implementation	42
7.1	Tweclu Architecture	42
7.1.1	Prerequisites	42
7.1.2	3 rd party Components & Libraries	44
7.2	The Tweclu Pipe	46
7.2.1	Query	47
7.2.2	Cluster	48
7.2.3	Visualization	53
7.3	Tweclu components	55
7.3.1	tweclu-sui	57
7.3.2	tweclu	57
7.3.3	tweclu-core	59
7.4	Notes on the Implementation	59

8	Evaluation and Results	61
8.1	Expert Interviews	61
8.2	The Data set	62
8.2.1	Tweet Selection	62
8.2.2	The World Wide Web Conference	63
8.3	Evaluation Participants	63
8.4	Evaluation Setup	63
8.4.1	Evaluation Procedure	63
8.4.2	Participant Sheet Contents	66
8.4.3	Evaluation Data & Settings	66
8.4.4	Evaluation Analysis Procedure	71
8.5	Evaluation Results	72
8.5.1	Dataset Analysis	72
8.5.2	Participants	73
8.5.3	Evaluation Analysis	75
9	Discussion & Conclusions	86
	Appendices	92
	Appendix A: Solr Index Fields	93
	Appendix B: Sample Solr Entry	95
	Appendix C: Participant Sheet	100
	List of Figures	101
	List of Tables	103
	References	104

Glossary

API	Application Programming Interface
GUI	Graphical User Interface
HAC	Hierarchical Agglomerative Clustering
HTTP	Hypertext Transfer Protocol
JSON	JavaScript Object Notation
JVM	Java Virtual Machine
PAM	Partitioning Around Medoids clustering algorithm
POM	Project Object Model
REST	REpresentational State Transfer
SMS	Short Message Service
TEL	Technology Enhanced Learning
TF-IDF	Term Frequency-Inverse Document Frequency
UI	User Interface
W3C	World Wide Web Consortium
WWW	World Wide Web
YAML	YAML Ain't Markup Language

1

Introduction

1.1 Motivation

The amount of information is growing at a high rate (Ganth et al., 2008). With that the number of scientific papers. According to de Solla Price (1963), the growth of scientific papers is exponential. This increase makes it almost impossible to keep up with scientific innovations. Even the amount of papers presented during a scientific conference is difficult to read, much less to derive all the topics and their (potential) coherences. The same is true for tweets published during a scientific conference.

Lately participants of scientific conferences use Twitter in order to publish real time updates about presentations, sessions, discussions, and workshops of that conferences. Some examples of conferences with large twitter audiences are the World Wide Web (WWW) conference¹, EdMedia², and the iKnow³. The conferences usually propose a hash-tag which is used by the participants in order to associate a tweet with the conference itself. (Reinhardt et al., 2009)

Having the tweets tagged, it is possible to follow a conference in real time. This opens the opportunity to participate online, or simply stay up-to-date on events one can not attend. Thus, questions can be asked and problems can be

¹<http://www.wwwconference.org>

²<http://www.aace.org/conf/edmedia/>

³<http://i-know.tugraz.at/>

discussed in a bigger audience, which may strengthen the point, or provide further directions a problem can go that has not been thought of yet. On the one hand this emphasizes the need to analyze Twitter data, as it can be seen as a source for valuable information. On the other hand, the amount of tweets published in the course of scientific conferences is not to neglect.

Various data sets of scholarly communication have been used and analyzed in research, be it bio-medical data, conference papers, or data gathered from social networks. For example, a lexical analysis of the Ed-Media 2000 and 2008 white papers was conducted and presented in Wild et al. (2010). The representation of the data presents a different view on the content of the white papers. Without having read the white papers it is possible to get an overview of the main topics discussed. Furthermore, the graph presents the inter-connections of the papers, and how the set of white papers forms a unity that was not obviously visible before the analysis.

The data sets just listed have one important issue in common: all the data is well defined in their specific domain; e.g. conference papers are written according to the standards of the language with respect to grammatical and lexical boundaries to the language used. On Twitter, sentences are usually shortened. Also abbreviations are used in order to meet the 140 character limitation. By resolving this problem, the results of a Twitter analysis would benefit in the additional insights provided in the discussions on questions and problems.

Existing systems using Twitter data are presented e.g. in Kraker et al. (2011). One system visualizes trending topics over time in a streamgraph. The other system is used to visualize both the co-occurrences of two hash-tags as well as the frequency of a hash-tag.

1.2 Objective

The objective of this work is to derive topics and possible coherences contained in a set of tweets, with respect to scientific conferences. In order to achieve this goal the following research questions (RQ) have to be dealt with.

RQ1 - How to cluster tweets?

This question has various challenges to overcome. Due to the limitation of at most 140 characters for one tweet, tweeters shorten sentences and use abbreviations of words and phrases; e.g. *fyi* represents the phrase *for your interest*; such abbreviations are widely used in order to save characters but still getting the message across.¹ The nature of tweets makes it more difficult to gain useful information. Therewith, the problem with analyzing tweets lies in the nature of tweets and has to be kept in mind in order to choose a good analyzing technique. In summary, the challenges to overcome are as following:

- "[u]nlike normal documents, these text & Web segments are usually noisier, less topic-focused, and much shorter, that is, they consist of from a dozen words to a few sentences" (Phan et al., 2008)
- "informal writing style (a poor grammatical structure) with many out of vocabulary words" (Perez-Tellez et al., 2010)

RQ2 - How to visualize results?

Having decided on how to cluster tweets, a visualization is needed in order to present the results. The challenges at this point are firstly to find a visualization that summarizes the results and gives an overview over the results in order to make it understandable and readable by humans. Secondly, the user shall be able to explore the results by navigating through them.

Based on the answers from both RQ1 and RQ2, a system is implemented. The resulting system is used in order to process the data and present a visualization. The resulting visualizations are evaluated in order to answer the following research questions, RQ3 and RQ4.

¹The use of abbreviations is also reported for people using the Short Message Service (SMS), cf. Cherry et al. (2010).

RQ3 - How well do the results represent certain aspects of the conference?

This question aims at validating if the objective was reached or not. In order to be able to answer this question an adequate evaluation technique has to be identified. The chosen evaluation technique shall allow the conductor the freedom to interact with the participant, and the participant to explore the visualized results and express his/her findings. Furthermore, the evaluation chosen is used to identify the relevance of the output according to the data set.

RQ4 - How usable are the presented results?

The last research question aims at collecting feedback on the usability of the presented visualizations. The challenge is to determine whether any result presented is usable for the purpose intended or not. This question needs to be taken into account when identifying an evaluation technique in order to be able to get an answer.

1.3 Outline

The thesis is structured into nine chapters. Chapters 2 through 4 provide the theoretical background for this thesis; chapter 2 focuses on Twitter and related research, chapter 3 on clustering and related research with respect to tweets, and chapter 4 on visualizations. Chapter 5 presents the decisions made in order to implement the system and evaluate the results. Chapters 6 and 7 present the basis for the implementation, and the details on the components used and implemented. Information on the conducted evaluation and the results are presented in chapter 8. Chapter 9 presents a discussion of the results and concluding remarks including suggestions for future work.

2

Twitter

Twitter¹ is a micro-blogging system, and sometimes also referred to as a social network, launched in 2006 (Golder et al., 2010). It allows its users to publish a message with up to 140 characters, called a tweet. The contents of a tweet vary; examples are: a message may contain personal experience or personal interests; companies subscribing to Twitter may publish new developments, innovations, or updates about current services. Due to the limited post length compared to blogging and social networks such as Facebook² or LinkedIn³, Twitter enforces a much faster mode of communication, and furthermore a much higher frequency of updates. (Finin and Tseng, 2007) (Guidi et al., 2012)

2.1 Conventions

The service of writing tweets is only allowed for users that have created an account and authenticated themselves to the service. By default, all tweets are visible to all Twitter users as well as the rest of the world almost instantly⁴. If the user decides to restrict visibility of his/her tweets to only directly connected Twitter users, the user has to explicitly state this in the settings.

¹<http://www.twitter.com>

²<https://www.facebook.com>

³<http://www.linkedin.com>

⁴as stated in the Terms of Service: <https://twitter.com/tos>

The syntax used in tweets is kept to a minimum. The following enumeration shows the most common conventions used (Golder et al., 2010):

- *@user* indicates a message directed at another Twitter user
- *RT @user* indicates a forwarded tweet originally posted by some other user
- *#tag* indicates a hash-tag in order to specify the topic of a tweet

Hash-tag

As just listed, hash-tags can be used to categorize tweets. It is important to mention that the hash-tag is community-driven and assigned by the Twitter users themselves. Thus, one topic can be represented with various hash-tags, e.g. #g+ vs. #google+, and a hash-tag can also be ambiguous, e.g. #bones may refer to the TV-show Bones as well as to the biological bone. (Nishida et al., 2011)

Not only topics can be expressed with a hash-tag; hash-tags serve a dual role. On the one hand, a hash-tag can represent a topic, or can be used in affiliation with an event; e.g. for the I-Know 2012¹ the hash-tag #iknow2012, or for the World Wide Web (WWW) Conference in 2012² the hash-tag #www2012 was used in order to indicate the affiliation to the event. On the other hand, a hash-tag serves as "a symbol of membership of a community" (Yang et al., 2012). A community is defined by its hash-tag and brings together users with the same background and interests (e.g., #android), as well as users who are involved in the same tasks (e.g., #gogreen). Simply by including a hash-tag in a tweet, a user joins a community. (Yang et al., 2012) (Letierce et al., 2010)

2.2 The role of Tweeters

The two main research areas when dealing with users on Twitter are: a user's intention on the one hand, and a user's role on the other hand.

¹<http://i-know.tugraz.at/>

²<http://www2012.org/>

Intention

Users have different intentions when publishing a tweet. Java et al. (2007) found *daily chatter*, *conversations*, *sharing information/URLs*, and *reporting news* to be the most dominant intentions in their analyzed data set. *Daily chatter* represents tweets containing statements about people's current doings, and is the most common intention. Making *conversation*, commenting on or replying to another user's update, is also found to be an important use. A significant amount of tweets contains URLs pointing to additional information and was therefore listed as a dominant intention, *sharing information/URLs*. *Reporting news* represents both reporting news by users and agents, as well as commenting on events.

Classes of users

Like users' intentions, users' roles can vary extensively. Twitter users were first classified by Java et al. (2007) into three categories: *information source*, *friends*, and *information seeker*. An *information source* has a large number of followers, and an update is of valuable nature. Users that follow other users on a regular basis, but aren't actively updating their status are represented as *information seekers*. The class of *friends* represents users that are mainly connected to people they already know from their real life.

A more recent study conducted by Tinati et al. (2012) divides Twitter users into five classes: *idea starter*, *amplifier*, *curator*, *commentator*, and *viewer*. This categorization is based on the users' influence, whereas the influence was measured by the number of re-tweets a user's tweet obtained. *Idea starters* are conversation starters and are highly engaged. Furthermore, they tend to have limited, but high quality connections. *Amplifiers* have a larger amount of connections and aim at sharing ideas and opinions of others. Amplifiers are recognized to be trusted within their connections and reach a wider community than idea starters. *Curators* question and challenge the ideas of *idea starters* and *amplifiers*. Thus, they have influence on the way a conversation goes further. *Commentators* aim to add their own insights to the conversation, but without seeking recognition by the community. *Viewers* consume the information but don't take part in the actual

online conversation. The information is usually shared in their offline network though.

2.3 The role of Twitter

Twitter is a medium that is used in various domains. For instance, Retelny et al. (2012) use Twitter in order to give students the possibility to get involved in developing lectures by asking questions and giving examples that would be reflected in the lecture. Because of the tweets it was easier for the teachers to clarify misunderstandings. Furthermore, the students felt more motivated and arrived in class better prepared. Governments and government organizations use Twitter as well, for instance for real time updates on traffic situations, or for locating people in emergency situation (Wigand, 2010).

Most often however, Twitter is referred to being an information medium and evolving to or already being a news media, cf. Rosoff (2012), Filloux (2012), Nichols et al. (2012), De Francisci Morales et al. (2012), Hu et al. (2012), Kwak et al. (2010) and André et al. (2012).

Regarding events, for instance sports events or political debates, people tweet about those events in real time. This data is considered additional information about the event. De Francisci Morales et al. (2012) harness this additional information in order to generate a summary of events. By concatenating the information retrieved, a reasonable summary can be generated automatically. This again is of benefit for people who were not able to follow the event. Furthermore, there would be no delay in presenting the summary of the event. (Nichols et al., 2012) The focus herein lies in summarizing a single event with a linear sequence of specific events (e.g. a football match), and not events run in parallel. While Nichols et al. (2012) only use Twitter as the source for summarizing events, De Francisci Morales et al. (2012) use both microblogs as well as news streams. By analyzing both sources, personalized news recommendations are being generated.

At this point it is worth mentioning that, as investigated in Hughes and Palen (2009), during events the importance of information broadcasting is more

significant than in times of no events. This indicates that more information is published during events than in other times.

2.4 Quality of Twitter data

As Twitter is a medium that is also used to report breaking events world wide, and its steady growth of popularity, spammers also get attracted more and more to use the service. Knowing the approaches of spammers rises the question on how to identify credible tweets or detect spam, cf. Castillo et al. (2011) and Benevenuto et al. (2010) respectively.

Castillo et al. (2011) analyze the credibility of newsworthy tweets of events that occurred over a time span of two month. Tweets that contained personal opinions or are part of a chat were filtered out and not considered part of the data set that was further analyzed. The approach used was supervised learning. The results indicate that a credible tweet has the feature of including Uniform Resource Locators (URLs), but also that the author of the tweet is one with influence (cf. subsection 2.2). (Castillo et al., 2011) (Gupta and Kumaraguru, 2012) (Xia et al., 2012)

Whereas Castillo et al. (2011) consider data from events of various nature, Letierce et al. (2010) and Weller et al. (2011) study tweets within the course of scientific conferences. Letierce et al. (2010) try to understand how researchers use Twitter to spread information. Therefore Twitter data from different conferences were gathered for further analysis. In the process of answering their research questions, Letierce et al. (2010) discover two exceptional facts when analyzing their data set: firstly, the number of re-tweets is around five times higher, and secondly, the use of hash-tags around three times higher, compared to general Twitter data. "The use of hash-tags and of re-tweet practices reveal a strong desire of the user tweeting during scientific conferences to emphasize particular messages" (Letierce et al., 2010). Weller et al. (2011) discover accordingly that, to their two data sets analyzed, more than 25% of the tweets contain URLs, and the number of re-tweets is around 50%.

2.5 Twitter in Research

Research shows that Twitter is used for various purposes and in many fields, e.g. real time news generation (cf. section 2.3), for protective security, surveillance, or to prevent terrorist attacks (cf. Heverin (2011) and Cameron et al. (2012)). Twitter is not only used during emergencies, popular sports events, or personal issues, but also in the scientific community during scientific conferences. As stated in Reinhardt et al. (2009) and Letierce et al. (2010), Twitter is considered to be a service that is accepted in the scientific community, and not only during conferences.

There are different reasons why Twitter is used during conferences, as stated by Reinhardt et al. (2009) and Ebner and Reinhardt (2009):

- Organizers use Twitter in order to keep attendees informed of any changes, and to spread reminders.
- Attendees write down notes and follow parallel events that are of their interest, but that they can not attend. They also use Twitter in order to ask other attendees questions. Furthermore, attendees not only exchange additional resources to events or during conversations, but also share plans for social activities. There are also researchers who only participate online and start a discussion.
- Presenters use Twitter to ask questions, with the goal to increase discussion participation and information exchange.

Reinhardt et al. (2009) conducted an on-line survey in order to gain insight on the usage of Twitter during scientific conferences. Results show that the main purposes of attendees using Twitter are to share resources and to communicate with others. Further purposes identified include writing down notes and participating in parallel discussions.

The findings of Reinhardt et al. (2009) mostly concur with the findings presented in Ebner and Reinhardt (2009) and Letierce et al. (2010). Ebner and Reinhardt (2009) additionally state that Twitter is also used for *off topic* purposes, such as social activities, or arranging additional meetings.

Scientific conferences are usually divided into various sub-events, for instance tracks, sessions, talks, workshops, etc. The task of automatically aligning tweets to their corresponding sub-event is attempted by Rowe and Stankovic (2010) and Rowe and Stankovic (2011). The approach uses the idea of auxiliary data. Auxiliary data is gathered by finding DBpedia concepts for the tweets with the Zemanta API¹ and storing the returned DBpedia concepts. In their study, three different feature sets are created to which the tweets are aligned to. Two feature sets include the meta data of the event itself, and the meta data of the events parent respectively. The third feature set includes the events' concepts, again retrieved with the Zemanta API. The aligning process was performed with two different settings of k -means, one using the Manhattan distance, and one using the Euclidean distance, as well as a generative model (Naive Bayes). The results showed that k -means using Euclidean distance led to better alignment results than using Manhattan distance. The best results when aligning tweets to their corresponding event were achieved with the generative model.

Twitter data from scientific conferences has also been analyzed in order to detect trends. For instance, Ebner and Reinhardt (2009) use the Yahoo Term Extraction Web Service² in order to "extract the most relevant terms or phrases" contained in a tweet. The results are sorted according to their frequency. The resulting ordered list is displayed in a word cloud, where terms or phrases with high frequency are visualized with a bigger font size than terms or phrases with low frequency. The resulting word cloud represents the trends and important topics of the conference tweets analyzed.

Kraker et al. (2011) analyze Twitter data from scientific conferences in order to detect trends as well. The approaches used, with respect to Ebner and Reinhardt (2009), differ though. Kraker et al. (2011) developed two different visualizations that are based on the same data. On the one hand, a streamgraph (cf. subsection 4.1.4) is used to visualize trending topics over time. On the other hand, a weighted graph (cf. subsection 4.1.3) is used to visualize both the co-occurrences of two hash-tags as well as the frequency of a hash-tag. The evaluation conducted

¹the Zemanta API is discussed in more detail in section 7.1.2

²<http://developer.yahoo.com/search/content/V2/termExtraction.html> last accessed October 2012

pointed out that both visualizations are easy to understand and complement each other. One disadvantage mentioned was that users wanted to see more details on demand. For instance, when clicking on a node, meta data should be shown, and the list of tweets shall be filtered according to the selection.

2.6 Twitter API

In order to interact with Twitter, for instance to collect or post tweets, Twitter offers an Application Programming Interface (API)¹. With this API it is possible to crawl tweets and collect data that is of interest for the application gathering tweets from Twitter. The API has implementations available for various programming languages such as PHP, Java, and Python. It is important to notice that tweets from other users can not be collected or crawled unless the tweets are made public, cf. section 2.1.

¹<https://dev.twitter.com/>

3

Clustering

Clustering is a technique in order to group collections together into clusters without supervision. Collections that are more similar to each other are more likely to end up in the same cluster than collections that are not that similar. (Jain et al., 1999)

Whereas clustering is the way of unsupervised classification, discriminant analysis is a method for supervised classification. In supervised classification a collection of already pre-classified, labeled data exists. This labeled data is used to learn the descriptions of classes. (Jain et al., 1999)

Within the context of this thesis only unsupervised classification is considered. Before diving into clustering algorithms some basic concepts related to clustering are discussed.

3.1 Concepts related to Clustering

3.1.1 Weighting

Weighting methods are used in order to calculate the importance of a term within a specified corpus and are usually based on statistical methods.

TF-IDF is a well known weighting method that has been applied on collections of documents successfully, cf. Manning et al. (2009) and Moh and Bhagvat (2012).

3.1 Concepts related to Clustering

The abbreviation is short for *term frequency-inverse document frequency*. The idea is to characterize a document by its word appearances. A document is more relevant if a term appears more often, whereas the term itself is only rarely used in a set of documents. Considering this thesis for instance, the word *the* will most likely be of less importance than the word *clustering*. (Nanas et al., 2003) (Manning et al., 2009) (Buckley, 1993)

3.1.2 Similarity Measures

Similarity measures are used to calculate the affinity of terms. The most common way to calculate similarity is to "calculate the dissimilarity between two patterns using a distance measure defined on the feature space". (Jain et al., 1999)

The Euclidean distance between two points is the "ordinary" length of the line connecting those points together and is defined as $\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$. The Euclidean distance is an established similarity measure and is used in the k -means algorithm. (Jain et al., 1999) (Xu and Wunsch, 2005) (Manning et al., 2009).

3.1.3 Labeling

Labeling is an important factor for humans in order to understand resulting clusters. The label of a cluster ought to be representative for its contents. One approach to label clusters is to just focus on the contents of a cluster, cluster-internal. Terms within a cluster can be frequent without being representative or helpful for the understanding of the contents. For instance, having the term *year* amongst the most frequent ones in a cluster is not helpful when having a data set specific to the topic *privacy*. Nevertheless, the labels provide a general description of the cluster's contents. This approach is also referred to as *descriptive* labeling. On the contrary, *discriminative* labeling ensures the uniqueness of labels. Thus, it provides a more detailed description of the contents. (Manning et al., 2009) (Kulkarni and Pedersen, 2005b) (Kulkarni and Pedersen, 2005a)

3.2 Overview of Clustering Algorithms

A taxonomy of clustering approaches can be found in Jain and Dubes (1988), see figure 3.1. According to this definition, in the first step, clustering can be divided into hierarchical and partitional clustering.

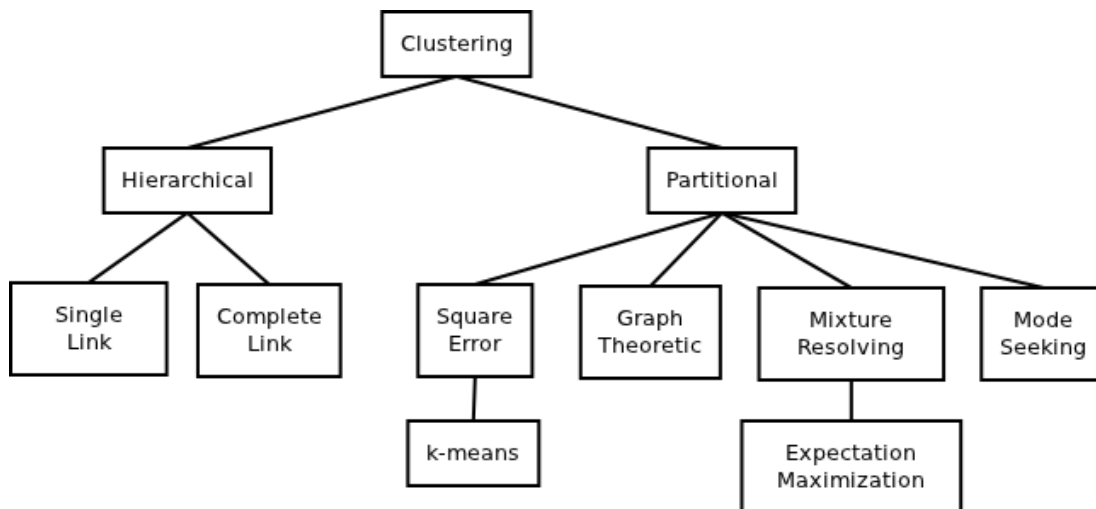


Figure 3.1: A taxonomy of clustering approaches - taken from Jain et al. (1999)

Hierarchical clustering creates a tree of clusters, a cluster hierarchy respectively. Partitioning algorithms on the other hand do not create any hierarchical structure. Instead, the collections are directly divided into clusters. (Dalal and Harale, 2011) (Xu and Wunsch, 2005)

There are some additional issues that may affect all of the different clustering approaches.

Hard vs. fuzzy: Hard clustering algorithms assign a collection exactly to one cluster, whereas with fuzzy algorithms each input collection is assigned to several clusters with a distinct probability¹. (Jain et al., 1999) (Manning et al., 2009)

Incremental vs. non-incremental: Incremental algorithms allow the adding and removing of collections to an already clustered result. The result is

¹fuzzy algorithms are sometimes also referred to as *soft* algorithms

adapted accordingly to the changes caused by the action performed. Non-incremental clustering algorithms have to re-compute the complete data set before changes may take effect. Furthermore, non-incremental algorithms are very likely to yield to different results with the same input. (Jain et al., 1999) (Witten and Frank, 2005)

Following the taxonomy illustrated in figure 3.1 further down, various partitioning clustering approaches exist, with several different implementations each. Detailed information on the approaches can be found in Jain et al. (1999), for instance. This thesis focuses on squared error algorithms, k -means especially, as it is described as "the simplest and most commonly used algorithm employing a squared error criterion" (Jain et al., 1999).

3.3 k -Means

k -means is a classic clustering technique first described in detail in Macqueen (1967). The idea of the algorithm, as described in Witten and Frank (2005) and Jain (2010), is to

1. specify the number k of clusters in advance
2. choose k points (seeds) as random cluster centers
3. assign each instance to the closest cluster center, according to the distance metric chosen
4. calculate the mean of each instance in each cluster
5. set the new points calculated in the previous step as centers
6. repeat until a minimum is reached

This results in k clusters where all points have reached minimum of the total squared distance to its cluster center. Nevertheless, even if a minimum is reached, it could just be a local minimum; reaching a global minimum with k -means is not guaranteed. The fact that the resulting clusters may vary on different executions

3.4 k -Means Extensions and Variations

of the algorithm, as the points chosen in step two may be set differently, may help finding a *better*¹ clustering result: re-running "the algorithm several times with different initial choices and [choosing] the best final result - the one with the smallest total squared distance". (Witten and Frank, 2005) (Jain, 2010)

According to Xu and Wunsch (2005), other important issues have to be taken under consideration when applying k -means:

- Step one states to specify the initial number of clusters, but usually the number of clusters is not known in advance.
- k -means doesn't consider the fact that data may have outliers, or may be noisy, which leads to distortion of clusters as each point will be forced into a cluster.

Despite the problems mentioned, k -means' time and space complexity have fairly good values: k -means' time complexity is $O(Nkd)$, where one can consider d and k to be constant (Jain et al., 1999) (Xu and Wunsch, 2005); its space complexity is $O(Nk)$, k being considered constant again (Xu and Wunsch, 2005). Therefore, k -means is an algorithm that can be used to cluster large data sets.

3.4 k -Means Extensions and Variations

Since the basic k -means algorithm has been around for a long time, extensions and variations dealing with the problems have come up. Early contributions to improve k -means are Partitioning Around Medoids (PAM) and ISODATA. PAM considers outliers, and ISODATA considers both outliers and finding a good amount of clusters (see Ball and Hall (1967) and Kaufman and Rousseeuw (1990) for further information on ISODATA and PAM, respectively). (Xu and Wunsch, 2005) (Jain, 2010) In the following, more recent improvements are discussed in more detail.

¹*better* according to the calculated total squared distance, which might be minimized

Seeds

As stated in the algorithm description in section 3.3, the seeds are chosen randomly. Arthur and Vassilvitskii (2007) propose to choose the seeds randomly only in the first iteration, but further iterations take the previously chosen ones into account and maximize the distance. This method, introduced as k -means++, has proven to be both faster and more accurate than the original k -means.

Growing k -means

Growing k -means makes use of k -means++. In addition, it uses splitting and merging in order to get the best amount of clusters. In order to do so, constraints defining the minimum and maximum amount of clusters are used. The final number of clusters is determined by using split and merge of clusters in order to determine the optimal minimum. (Muhr et al., 2010)

3.5 Clustering Twitter Data in Research

Publications in this area try to classify tweets, or to find related tweets in general. For instance, the goal of Rosa et al. (2011) is to automatically classify tweets into different categories. The categories were pre-defined and existing hash-tags were assigned to the specified categories. The approach is based on the assumption that a tweet's hash-tag is a good indication for the topic of the tweet. The data set was collected accordingly; tweets containing the pre-defined hash-tags were crawled. Both unsupervised and supervised clustering algorithms were used and the results evaluated. With unsupervised clustering, the resulting classification was not correlated on topics, but instead based on language similarity. On the other hand, with supervised learning and a test set close to the training set, the resulting clusters were of topical coherence.

Nishida et al. (2011) attempt to find related tweets, when given a topic. The technique used is adapted from the field of data compression. The proposed approach does leverage the fact that a file containing more similar chunks can be compressed more efficient than a file containing less or even no similar chunks.

3.5 Clustering Twitter Data in Research

The evaluation of the technique shows that even short texts can be classified effectively, but further evaluation has to be done in order to improve accuracy and speed.

Another approach for improving clustering results of tweets is the use of auxiliary data. Auxiliary data is used in order to enrich the originally small data corpus. For instance, Phan et al. (2008) use the most relevant search results that were achieved by querying the tweet in a search engine, whereas Sahami and Heilman (2006) make use of public knowledge repositories such as Wikipedia. Nevertheless, Liu et al. (2011) claim that "[t]hey generally make the implicit assumption that the auxiliary data are semantically related to the input short texts, which is hardly true in practice[...]" Liu et al. (2011) introduce a topic model approach, similar to Phan et al. (2008), which basically creates two sets of topics: the first set is based on the target, the tweet, and the second set is based on the auxiliary data. The results show that their approach improves the clustering results, although irrelevant text was still present in the auxiliary data.

The goal of the method proposed by Perez-Tellez et al. (2010) is more specific in its domain. The authors focus on tweets relating to companies. Tweets are getting clustered into two clusters: one representing tweets referring to a company, and the second one representing tweets not referring to a company. The clustering algorithm used was *k*-means, as "it is a well-known method, it produces acceptable results and [the] approaches may be compared with future implementations". Based on *k*-means, Perez-Tellez et al. (2010) make use of various auxiliary data sets. For instance, one method uses Wikipedia enriched data, which was introduced by humans. Another method creates and uses a thesaurus in order to enrich the data fully unsupervised. The results show that the unsupervised approach provides the best results.

Guidi et al. (2012) use Twitter in order to prove "a relation between users' behavior on the Social Networks and their real life chores". For this, the tweets of Barack Obama within the period of April 29, 2007 and September 6, 2010 were collected and processed. The procedure was (a) to select the most representative keywords with lexical tools, and (b) to cluster the results of (a). The resulting clusters seem to mirror Barack Obama's public work and political agenda.

4

Visualization

According to Ware (2000), visualization in the information age is commonly referred to as *information visualization*, which is defined as "[...] a graphical representation of data or concepts". (Keim et al., 2006) (Ware, 2004)

A visualization that represents data graphically enables a human to compare the data more efficiently; for instance to spot trends, recognize patterns, and identify outliers. It enables humans to get a better insight into the underlying data by exploiting the humans' perceptual system. The vision of visualization is to tell a story about the data it represents, cf. Hearst (2003) and Heer et al. (2010).

With the current and increasing overload of information, visualization has become an important research area for all kinds of disciplines. Visualization has evolved into a technique to present a big amount of data in a compact way, but still preserving the relevant information. This makes it easier and more interesting for humans to actually perform analysis tasks. (Hearst, 2003)

Another important fact, reported by Miller (1956), is that humans can process more information when using their visual abilities.

Visualization is closely linked to the information it has to present, and the goal it has to fulfill. Depending on the amount of data, preprocessing algorithms might be applied on the data. These data mining techniques are used to extract valuable information automatically; clustering, as discussed in section 3, is one example.

4.1 Visualization types

This subsection provides some examples and explanations on different visualization types. A more extensive overview of the various visualization types can be found in e.g. Andrews (2012) and Heer et al. (2010).

Based on the data, and the desired output, different visualization types exist. As discussed in Keim et al. (2010), a good visualization depends on the data; both on the data that is available and has to be visualized, and the quality of the data. Thus, "[o]ne must determine which questions to ask, identify the appropriate data, and select effective visual encodings to map data values to graphical features such as position, size, shape, and color" (Heer et al., 2010).

In the following, examples of visualizations are shown, depending on the *question* that has to be answered.

4.1.1 Chart

The basic types of visualizations are charts. Pie charts, bar charts, etc. are simple, and most users are familiar with charts as they have been around for visualizing two dimensional data for quite some time. Nevertheless, the drawback with charts is that "it [...] limits the types of possible visualizations". (Heer and Shneiderman, 2012)

When dealing with a large data set that might also include multiple dimensions, further visualization types have evolved. The visualization types are based on the data that needs to be visualized. The categorization is based on the one used by Heer et al. (2010).

4.1.2 Map

According to Heer et al. (2010), maps are used in order to display data that can relate to geographical information. Furthermore, as maps are of a familiar type for humans, *redrawing* the borders of an actual map still keeps the basic idea of a map up. This idea is visualized in figure 4.1:

4.1 Visualization types

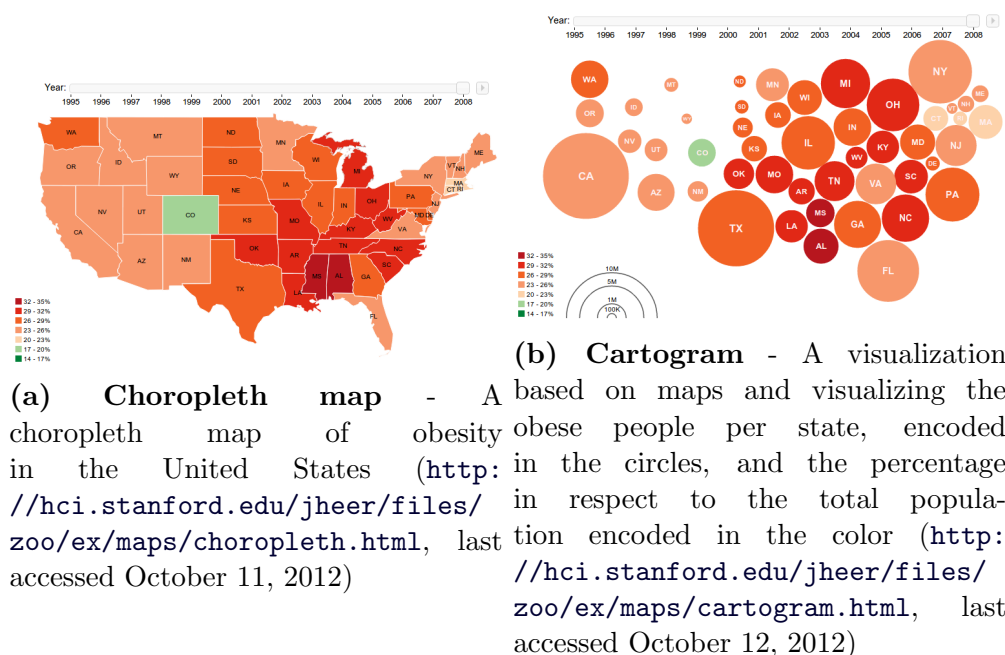


Figure 4.1: Examples of map visualizations

- Figure 4.1a displays the percentage of each state's (of the United States) population that are classified as *obese*. The colors indicate the total percentage of the state, as shown in the legend in the lower left corner of the figure.
- Whereas the color encoding stays the same for figure 4.1b, a second dimension is introduced, representing the actual number of obese people in the state.

The idea of using maps is also re-used in information landscapes. Information landscapes are used in order to visualize high-dimensional data. An information landscape facilitates both topical connections as well as topical peaks. Figure 4.2 illustrates an information landscape of documents on climate change. Topical peaks are represented by a hill, indicating that a group of documents deal with the same subject. For instance, the hill in the upper right corner represents documents dealing with *ice*, *arctic*, and *past*. The visual representation of the hill's height indicates that a great amount of documents deals with the named subjects. But the number of documents isn't as high as for the subjects represented by the

hill labeled with (1). Topical connections are represented in the proximity on the landscape. For instance, the group of documents represented by the hill in the upper right corner have more in common with documents in the center than the ones represented in the lower left corner. (Sabol et al., 2010)

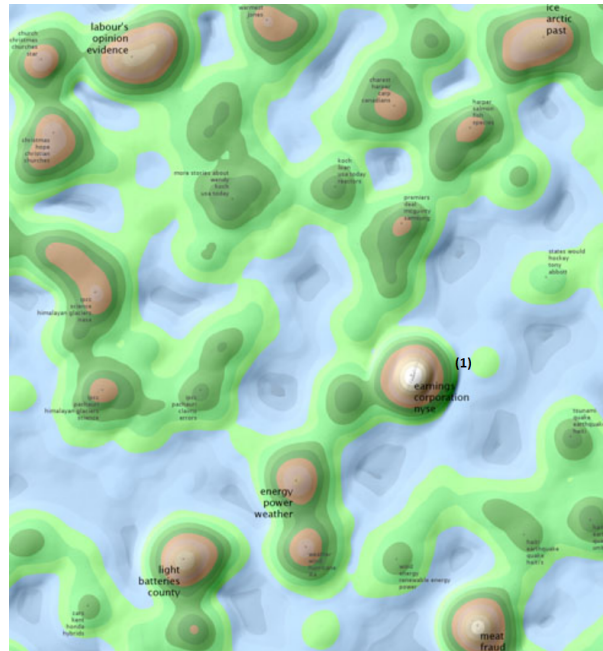


Figure 4.2: Information Landscape - An information landscape of documents on climate change (from Sabol et al. (2010))

4.1.3 Network

Visualizing networks has the focus on the connections represented in the data. Graphs may be used in order to display relationships amongst entities within a network. The relationships within a network can be defined in different ways. For instance, considering users of a social network with friendship relations amongst the users representing a network that is defined by nature. But, the graph can also be used for displaying the most common terms in a data set and their relations to each other. (Heer et al., 2010)

Figure 4.3 presents a graph of the latter; nouns that occurred the most in relation with the term *edchat* in the time frame from October 5, 2012 until Oc-

tober 12, 2012. The nodes indicate the total number of occurrences, and the lines between the nodes indicate the number of combined occurrences of the two connected nodes.

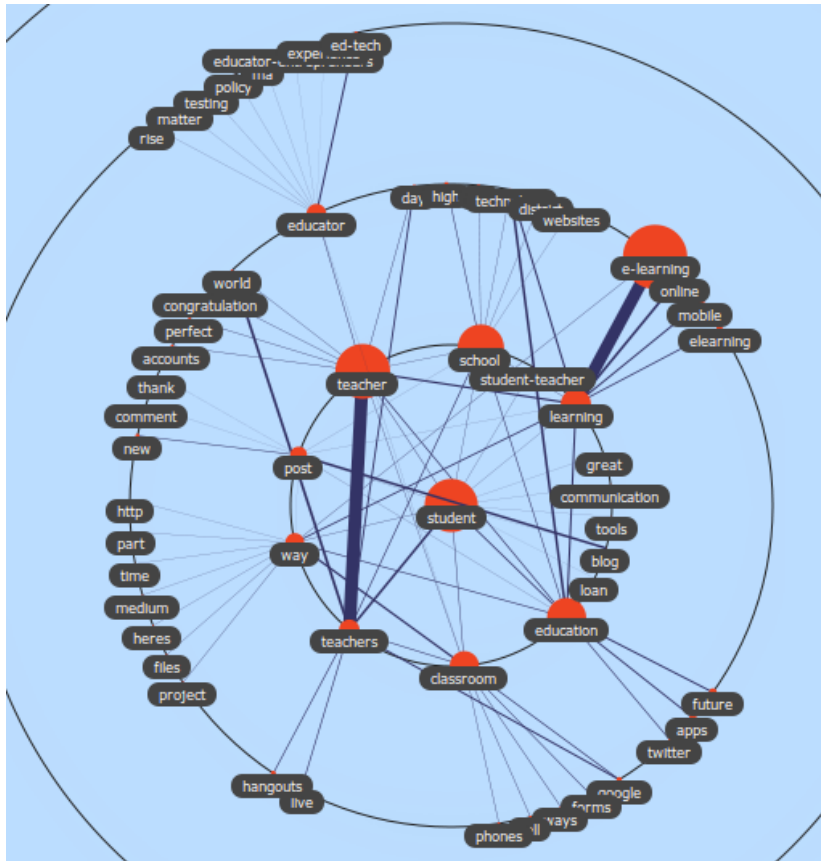


Figure 4.3: Weighted graph - A weighted graph highlighting popular nouns and their relations to each other (http://stellar.know-center.tugraz.at/vis/weighted_graph/index.html, last accessed October 12, 2012)

4.1.4 Time

Time is a feature of data sets that is of importance to various fields of expertise. In finance, for instance, visualizing the value of stocks over time reveals more information than just at one specific date. One possibility for visualizing data values over time is the *Streamgraph*. (Heer et al., 2010) (Byron and Wattenberg, 2008)

Figure 4.4 illustrates a streamgraph that shows the nouns that occurred with the term *edchat* in the time frame from October 5, 2012 until October 12, 2012.

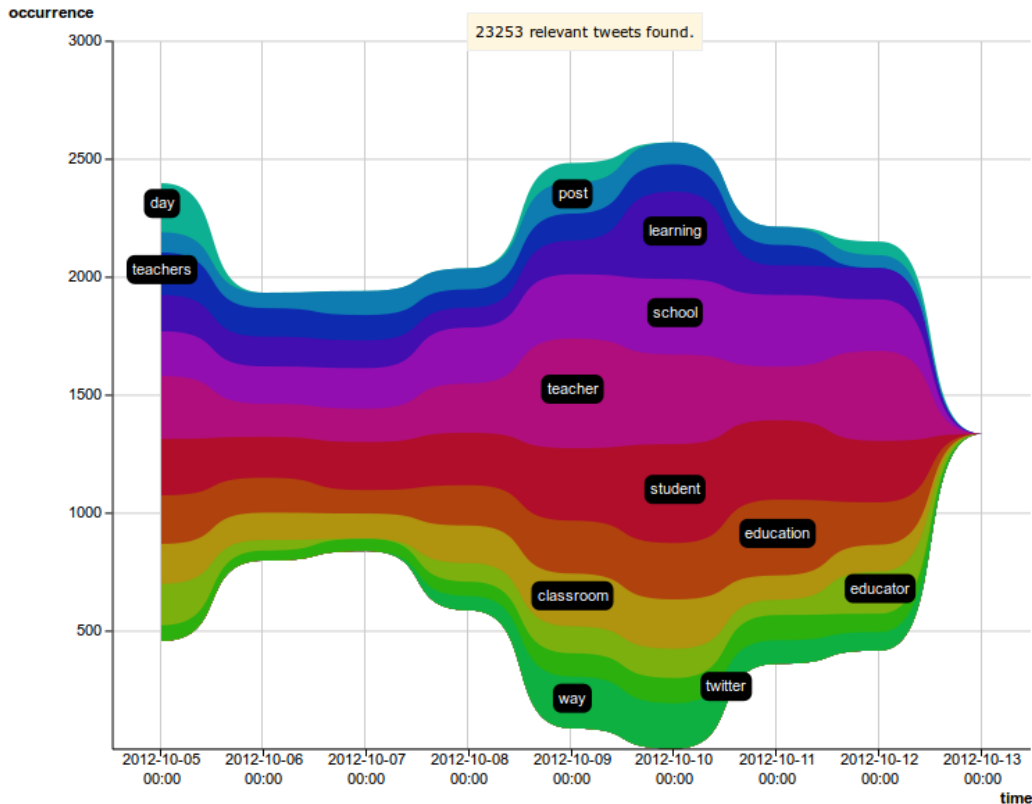


Figure 4.4: Streamgraph - the nouns that occurred with the term *edchat* (<http://stellar.know-center.tugraz.at/vis/streamgraph/index.html>, last accessed October 12, 2012)

4.1.5 Hierarchy

Visualizing data according to their hierarchy is the last type discussed in this thesis. Two types of hierarchies can be distinguished. On the one hand, hierarchies may be existing in the representing data from the beginning, for instance when considering file systems or package structures. One well known visualization of trees is the file explorer present in almost all operating systems. The package tree of the library Flare¹ is shown in figure 4.5 in order to exemplify another type for

¹<http://flare.prefuse.org/>

visualizing hierarchies. For instance, the package `interpolate` which is located in `flare.animate.interpolate` is represented in a the bubble `animate`, whose parent package is the bubble `flare`. (Heer et al., 2010)

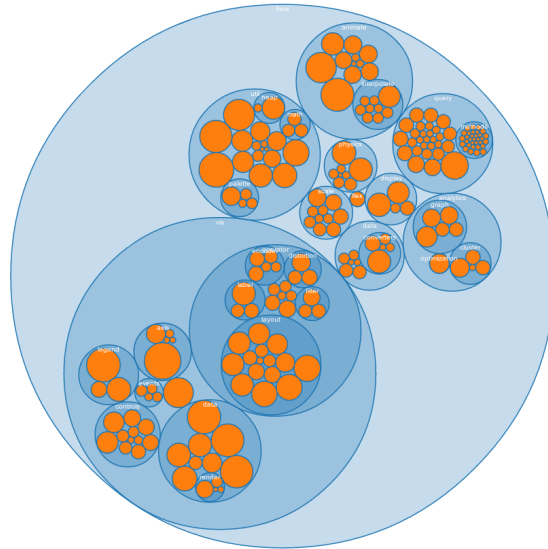


Figure 4.5: Hierarchy representation - the package tree of Flare (<http://flare.prefuse.org/>) packed into circles recursively (<http://hci.stanford.edu/jheer/files/zoo/ex/hierarchies/pack.html>, last accessed October 11, 2012)

On the other hand, hierarchies may be a result of data processing techniques, intended, for instance, on summarizing or merging data. Clustering may be one technique that can be used (see chapter 3 for further information on clustering). (Heer et al., 2010)

According to Delort (2010), Voronoi polygons have also been used for cluster representations successfully; applications can be found in e.g. Pinho et al. (2006) and Granitzer et al. (2004). Granitzer et al. (2004) adapt the idea of Okabe et al. (2000), in order to represent the number of documents contained in a polygon appropriately. Figure 4.6 presents a resulting visualization of large, hierarchically structured document repositories. The various Vornoi polygons represent different groups of documents. For instance, the polygon in the upper left groups documents dealing with geographic entities such as *continents* (*Kontinente*), *regions* (*Regionen*), and *states* (*Staaten*).

4.1 Visualization types

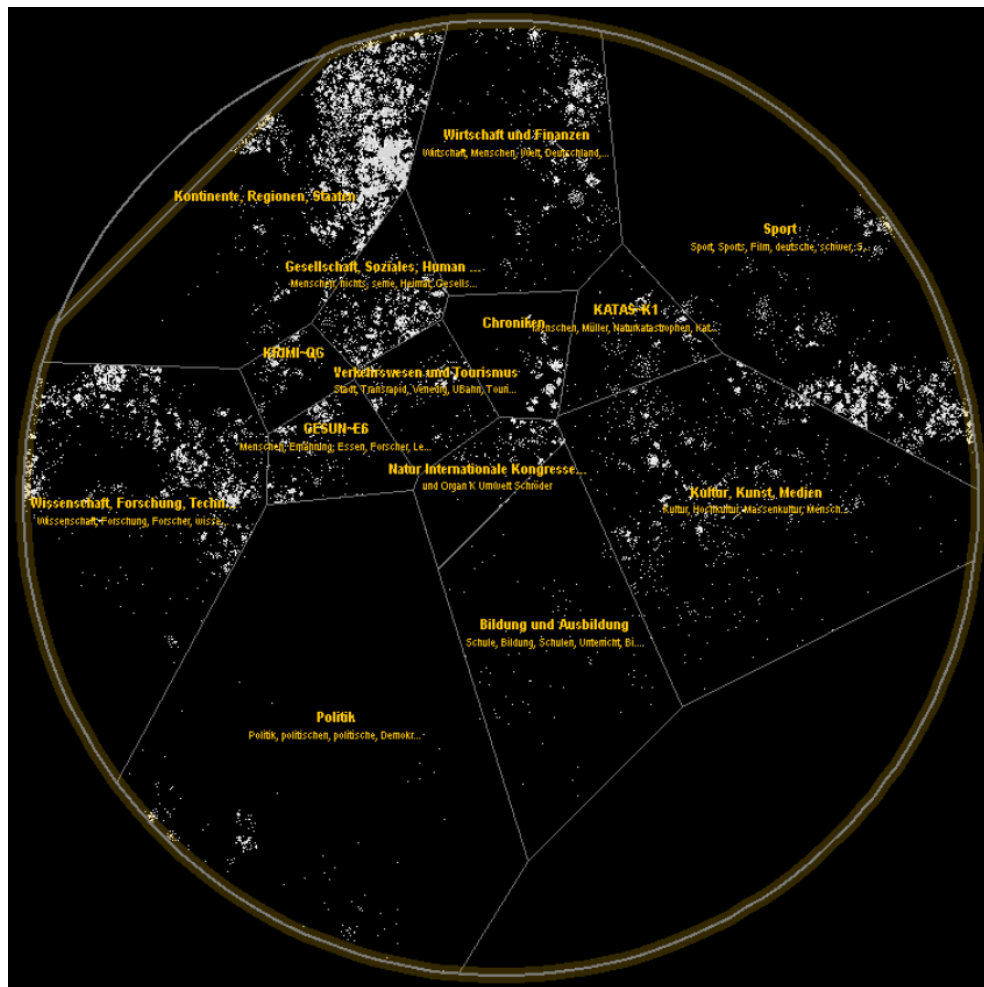


Figure 4.6: InfoSky - the InfoSky visualization (from Granitzer et al. (2004))

4.2 Interactive Visualization

Another important factor for visualization is interactivity, cf. Keim et al. (2006) and Keim et al. (2010). Some of the examples presented in the chapter have the ability of interaction implemented when viewed on-line. For instance, the on-line version of the visualization presented in figure 4.3 allows to (1) zoom in and out with the mouse scroll wheel and (2) click on a node, which moves the clicked node to the center. In order to explore the galaxy shown in figure 4.6, the implementation provides not only zooming, but also highlighting of results of an executed search, cf. Andrews et al. (2002) and Granitzer et al. (2004).

Without the possibility of interacting with the data, users first specified the search query which was executed by a system. The results returned by the system were then shown to the user. With the introduction of interactivity, the systems update the results almost immediately and according to user actions. (Brodbeck et al., 2009)

According to Schneiderman (1998), interactive visualization systems aim at: "Overview first, zoom and filter, then details-on-demand". This means that the user first gets an overview of the underlying data. Using zooming and filtering provides the user with functionalities to explore the data. These interaction techniques allow the user to gain insight into the data that may not have been revealed before. Based on the newly revealed insights, new conclusions can be drawn from the data. Also details may be presented to the user on demand.

Figure 4.7 presents a screenshot from Gapminder World¹, an interactive visualization. The example puts life expectancy into relation with income per capita. The changes over the years from 1937 until 2011 for Japan, Austria, and the United States are shown. Countries are represented by a bubble for each year. In particular, one can see that life expectancy for Japan was drastically lower during and shortly after the second world war, but increased rapidly until it caught up with both Austria and the United States in the 1960s.

¹<http://www.gapminder.org/world/>

4.2 Interactive Visualization

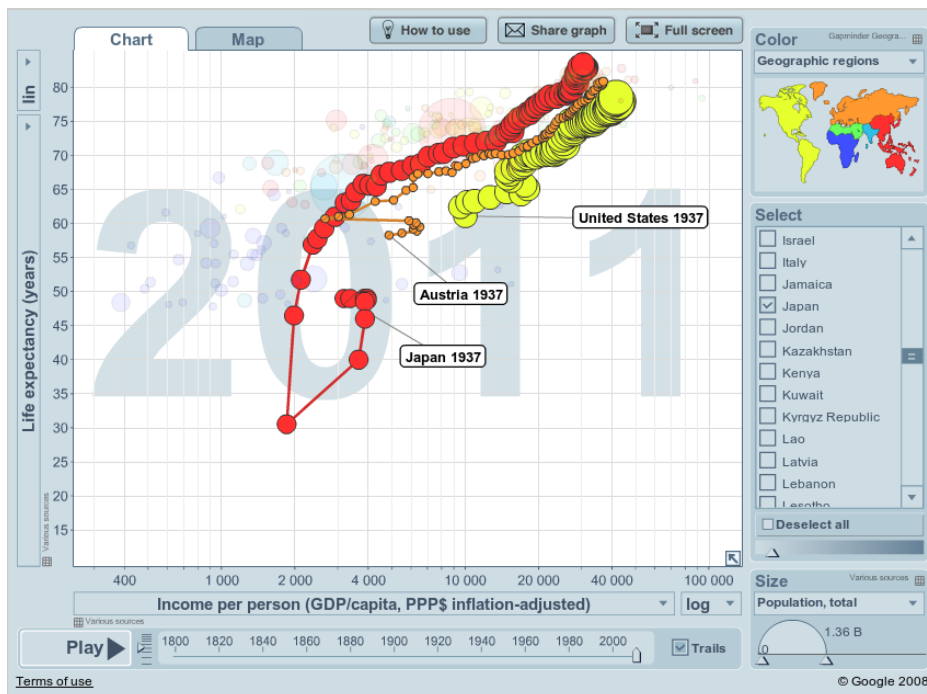


Figure 4.7: Interactive Visualization Example - An example for an interactive visualization. visualization from Gapminder World, from www.gapminder.org

5

Overview of Practical Work

Twitter is an accepted medium in the scientific community. It is also used to discuss and/or present problems and ideas presented during scientific conferences. These discussions bring additional, and valuable information to the topics presented at a scientific conference.

The objective of this thesis is to extract and visualize valuable information of tweets published in the course of a scientific conference. Therefore, four research questions are defined (cf. chapter 1):

- **RQ1** How to cluster tweets?
- **RQ2** How to visualize results?
- **RQ3** How well do the results represent certain aspects of the conference?
- **RQ4** How usable are the presented results?

In the first step a system is developed addressing the challenges of RQ1 and RQ2. The results of the system are evaluated in order to get answers for RQ3 and RQ4.

System

The main result of this work is a system that gathers and processes tweets, and finally visualizes the results. The system works on any data set; as a use case tweets from the World Wide Web conference in 2012 (WWW2012) were taken. In order to collect the data from Twitter a system already used in Kraker et al. (2011) was re-used.

Clustering is used in order to extract information from the data set. As presented in chapter 3, clustering has already been successfully employed with short texts for various purposes. The chosen clustering algorithm is **growing k -means**, which is an extension of k -means (cf. section 3.4). k -means has been around for quite some time, and during that time variations and extensions have been implemented in order to improve the algorithms' flexibility and performance. Growing k -means is one extension. It applies splitting and merging of clusters in order to find a feasible amount of clusters for the data set provided, without having a pre-defined amount of clusters. Furthermore, by choosing the seeds not randomly, the cluster centers are calculated faster. Thus, growing k -means was chosen in order to be used to process the data.

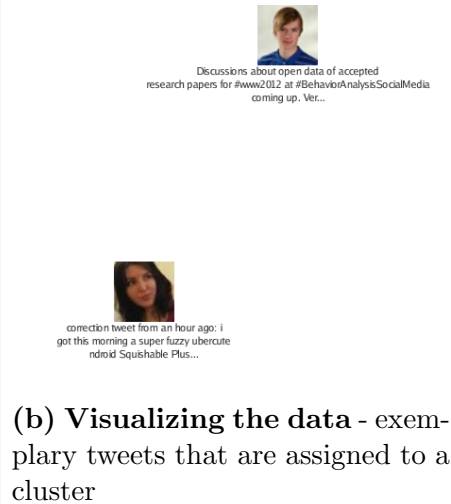
In a next step the results are visualized. A visualization makes it easier for humans to read the data. In order to bring the most benefit to the user, an interactive visualization is chosen, cf. section 4.2. With the help of an interactive visualization, the user is also able to explore the results. Using a graphic library, the resulting clusters are displayed as Voronoi polygons, which have already been used for visualizing clusters successfully (cf. subsection 4.1.5). Furthermore, the tweets that have been assigned to a cluster can be viewed in the resulting visualization. Figures 5.1a and 5.1b illustrate an overview of the resulting clusters represented in Voronoi polygons, and tweets that have been assigned to a cluster respectively.

Evaluation

The purpose of the evaluation is to assess the results of the system by getting answers to the research questions RQ3 and RQ4. In order to do so expert inter-



(a) Visualizing the resulting clusters - the overview of resulting clusters visualized



(b) Visualizing the data - exemplary tweets that are assigned to a cluster

Figure 5.1: Exemplary visualization results

views are conducted. The interviews aim to validate the results with respect to the topics and events at the conference. identifying the relevance of the output according to the data set. Furthermore, the expert’s opinions on usability and usefulness of the presented system is of importance as well. Therefore, pre-defined questions are defined in order to have a common basis for all participants. The following questions are defined to assess if the results represent certain aspects of the conference:

1. How well are thematic topics covered?
2. How well are organizational topics/events covered?
3. How well are highlights represented?

Determining whether the results presented are usable or not, following questions are defined:

1. Do you think the presented visualizations are better than just a list of tweets?

-
2. What did you like and not like about the visualizations presented?
 3. Do you think the visualization would be helpful for other purposes?
 4. Which visualization of the presented did you like most?

During these interviews, three different settings have been provided to the participants for evaluation. Two of the visualizations presented were results of the implemented system, and one was a Weighted Graph visualization implemented by Kraker et al. (2011). The interviews have been qualitatively analyzed with the help of coding. The analysis revealed that the system mainly supports topic extraction, and to some extent organizational event extraction. The results of the analysis also pointed out the need for a refined clustering technique, and additional processing in order to visualize relations in between clusters.

6

System Design & Implementation Basis

This chapter gives an overview of the complete system design in section 6.1. Section 6.2 highlights details on the data, including conducted pre-processing, information on Solr and its configuration. Section 6.3 presents preparatory work which was conducted in the course of the master's project.

6.1 System Design Overview

Figure 6.1 illustrates the complete system overview. Starting at (1) through (3), first Twitter is crawled in order to save the tweets for later processing (see section 6.2). The components implemented in the course of this thesis are represented within the rectangle titled *tweclu*. The flow of the data is to first gather the data from the storage system (5), and second to cluster the data (6). The last step is to visualize the results (7). Each components' dependency to a third party library is illustrated outside the rectangle with a line to the concerning component within the rectangle. For instance, in order to get the data from the storage system, Solr in particular, the library *SolrJ* (4) is used (cf. subsection 7.1.2). A detailed description on the components and the third party libraries is provided in chapter 7.

6.1 System Design Overview

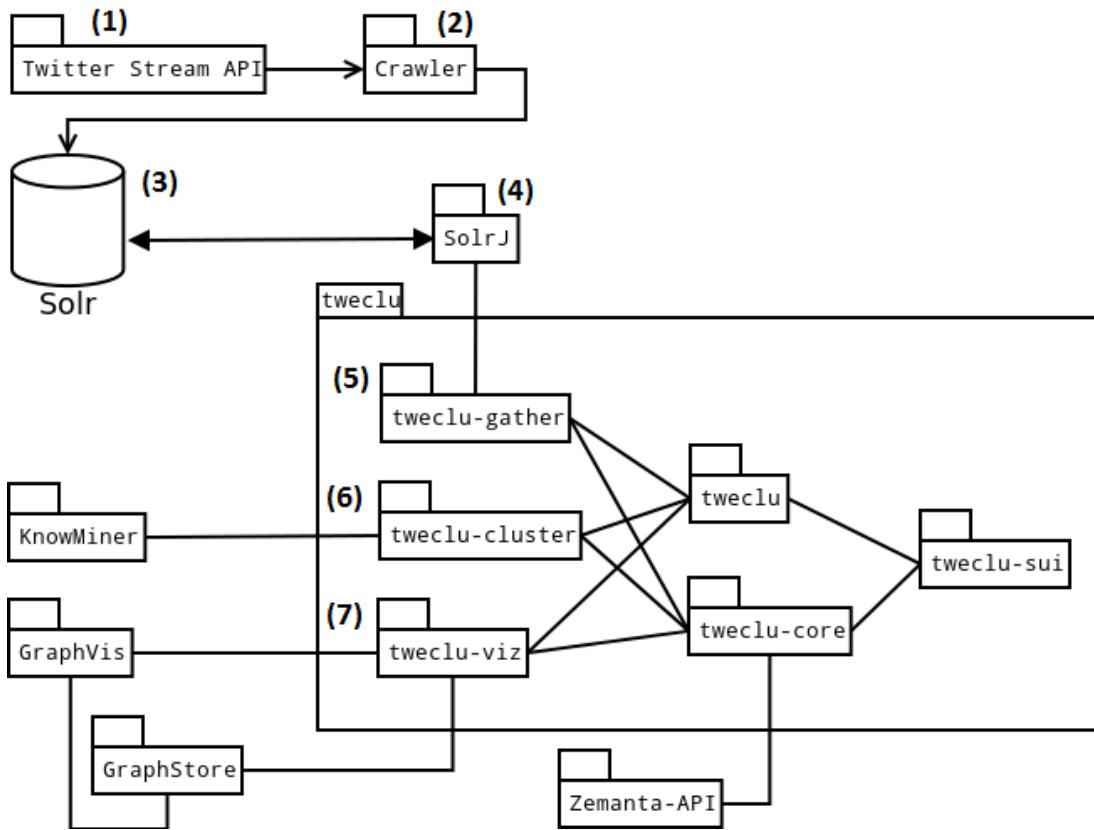


Figure 6.1: System components overview - The system components and their relations

6.2 The Crawler

Figure 6.2 illustrates the basic overview of the complete system with a focus on the tweet processing. This section describes the process from gathering the tweets from Twitter to the storage into Solr. Furthermore the description of the data is presented. It is important to note that this process was taken as is, as it was used in other projects conducted at the Knowledge Technologies Institute and the Know-Center already, cf. Kraker et al. (2011).

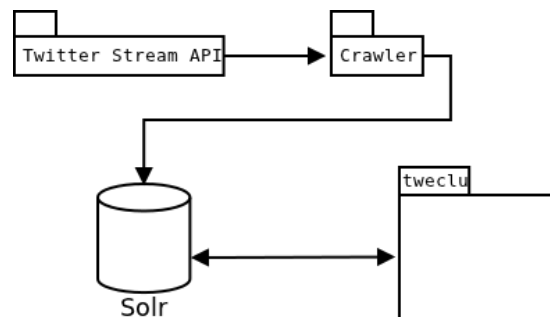


Figure 6.2: System overview (data) - The system overview with a focus on the data gathering

6.2.1 Tweet crawling and pre-processing

In the course of Kraker et al. (2011), a Twitter crawler was developed that makes use of the Twitter Streaming API¹. The crawler is adaptable ”to any domain, by either (a) specifying a taxonomy of keywords, (b) specifying a list of users, or (c) a combination of both”. (Kraker et al., 2011).

The tweets logged by the Twitter crawler are then processed in order to clean up the contents and generate additional meta data. Therefore, tokens with informative value were extracted from the tweet using TreeTagger (cf. Schmid (1994)), a part-of-speech (POS) tagger. The tokens of interest are especially *nouns* and *hash-tags* contained in the tweet. (Kraker et al., 2011)

¹<https://dev.twitter.com/docs/streaming-apis>

6.2.2 Solr

Solr¹ is a product maintained by Apache² and is based on Apache Lucene³. Lucene offers to index documents and to execute search queries based on the generated indices. Furthermore, various search algorithms are available. Solr serves as an additional interface to the functionalities offered by Lucene. Solr is a server which can be used within a servlet container, Apache Tomcat for instance. Additionally, the query language offered by Lucene is extended in Solr, for more detailed search queries. The main benefits of Solr, with respect to this thesis, are the REST and JSON application programming interfaces (APIs). With these APIs, applications querying the index do not need to run on the same machine.

6.2.3 Solr index fields

Besides the cleaned tweets, additional data is stored in the Solr index. For instance, meta data about the tweet itself, meta data about the user, URLs contained in the tweet, etc. The Solr index fields of relevance are as following⁴:

`tweet_id` unique identification of the tweet

`tweet_timestamp` time stamp when the tweet was published

`tweet_text` original content of the tweet

`tweet_text_reduced` the content of the tweet without hash-tags and mentions

`tweet_text_topics` the categories for the tweet returned by Zemanta API⁵

`tweet_terms_hashtags` the extracted hash-tags only, if applicable

`tweet_terms_nouns_lemma` the lemmatized nouns contained in the tweet, if applicable

¹<http://lucene.apache.org/solr/>

²<http://www.apache.org/>

³<http://lucene.apache.org/>

⁴A complete list of the Solr index fields can be found in Appendix A, cf. page 93.

⁵This field was introduced in the course of this work and is not represented in the Solr instance used in Kraker et al. (2011)

`tweet_terms_multi` both the extracted hash-tags as well as lemmatized nouns, if applicable

`tweet_user_screen_name` the screen name of the user who tweeted the tweet

`tweet_user_profile_image_url` the URL to the profile image of the tweeter

The fields `tweet_user_screen_name` and `tweet_user_profile_image_url` are not used for analysis, but are of relevance for visualizing the results.

A sample Solr entry can be found in Appendix B, cf. page 95.

6.3 Preparatory Work

This thesis builds on the master's project, also performed at Knowledge Technologies Institute of Graz University of Technology¹. During the master's project the following tasks were performed:

1. research of available visualization libraries and their supported visualization styles
2. evaluation of visualization library and selected visualization styles with test data

The visualization library chosen was JIT² as it offers various visualization styles. Four different styles were chosen to be evaluated in more detail: Sunburst³, RGraph⁴, Icicle⁵, and SpaceTree⁶.

The results are illustrated in table 6.1. They show that all four visualization styles are interactive. But neither Icicle nor SpaceTree offer a complete overview of the data set. Furthermore, it is not possible to add a quantity to the nodes and/or edges; thus Icicle and SpaceTree dropped out, which left Sunburst and

¹<http://kti.tugraz.at/>

²<http://thejit.org>

³<http://thejit.org/static/v20/Jit/Examples/Sunburst/example2.html>

⁴<http://thejit.org/static/v20/Jit/Examples/RGraph/example1.html>

⁵<http://thejit.org/static/v20/Jit/Examples/Icicle/example2.html>

⁶<http://thejit.org/static/v20/Jit/Examples/Spacetree/example1.html>

6.3 Preparatory Work

Feature / Style	Sunburst	RGraph	Icicle	SpaceTree
quantifiable	terms	terms, edges	-	-
complete overview	yes	yes	-	-
interactive	yes	yes	yes	yes

Table 6.1: Jit Visualization styles and their features - a comparison

RGraph. Both Sunburst and RGraph do offer a complete overview of the data even with a fairly big amount of data. With Sunburst however the *slices* become very small, which leads to non-readable terms on first sight. Furthermore, Sunburst also lacks in support of inter-term relations. While visualizing a lot of terms in RGraph can become messy, RGraph has the big advantage of firstly inter-term relations, and secondly the option to quantify both terms and edges.

With RGraph, a network analysis on tweets of the scientific conference *Alpine Rendez-vous 2011*¹ was performed.

Depending on the data, the graph produced by the system introduced by Kraker et al. (2011) includes terms and/or hash-tags that blur the results. The terms in question are mainly added during further iterations of the algorithm. Thus, the process of the network analysis differs from the research conducted by Kraker et al. (2011) in the way that only 1st degree co-occurrences are taken into account. Through further iterations no new nodes are added, only edges in between existing nodes are added or updated.

An exemplary result is illustrated in figure 6.3. It shows both nouns and hash-tags that occurred with the term *arv11* within the time range from 2011-03-27 until 2011-03-31². The results have not been satisfying; depending on the amount of tweets the graph either has too many nodes in order to get a clear picture, or too few nodes in order to be able to form an opinion. In order to regulate the number of nodes, two threshold values were added to the algorithm: the first threshold x is responsible for only adding nodes that occur at least x times, when processing the first results of Solr; the second threshold y indicates that a term

¹<http://www.stellarnet.eu/programme/wp3/rendez-vous>

²The implementation results are also available on-line at <http://stellar.know-center.tugraz.at/vis/jit-rgraph/>.

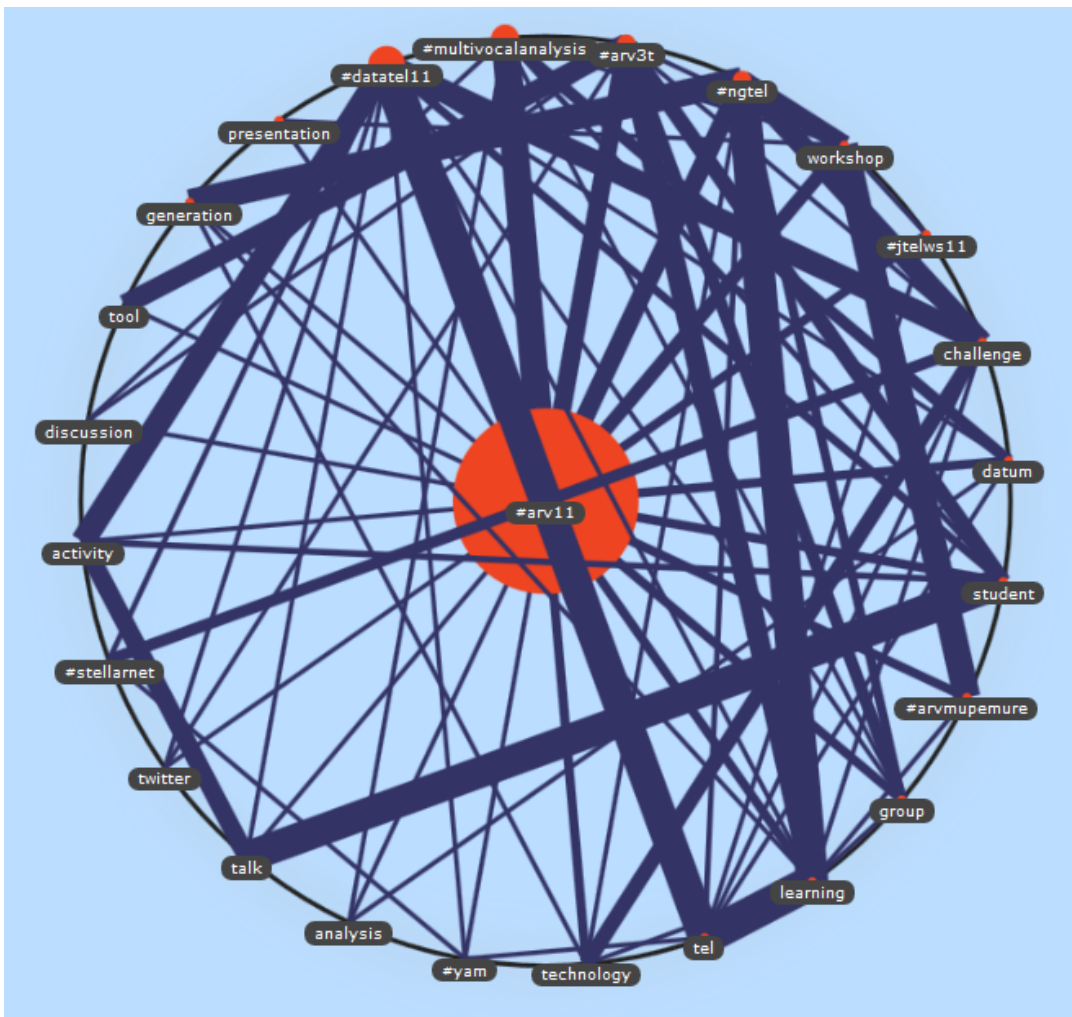


Figure 6.3: Network analysis - showing both nouns and hash-tags that occurred with the term *arv11* from 2011-03-27 until 2011-03-31

6.3 Preparatory Work

has to occur at least y times in total. Depending on the amount of tweets that need to be processed, both threshold values can be manually set accordingly.

As illustrated in figure 6.3 the individual workshops are represented with their hash-tags, for instance *#arv3t*, *#datatel11*, etc. But, the graph is too cluttered and not enough information is presented in order to derive the topics of these workshops, not enough information is presented in the resulting graph. Thus, a system in order to test and evaluate clustering results of tweets was implemented. This system is described in the next chapter.

7

Implementation

This chapter describes the implementation of the system, explaining the architecture, its internal interfaces, and its dependencies to 3rd party libraries.

7.1 Tweclu Architecture

The system implemented is divided up into six components. The components and its dependencies to 3rd party libraries and other components are illustrated in figure 7.1.

Before describing the idea behind the architecture and the workflow itself, some prerequisites, and the essential 3rd party libraries are discussed.

7.1.1 Prerequisites

Maven

Maven¹ is an open source tool maintained by Apache². It is used to manage software projects, including the build process, reporting, and documentation. All the details needed in order to manage a software project, e.g. dependencies or packaging, are defined in a project object model (POM). Based on the POM, the

¹<http://maven.apache.org/>

²<http://www.apache.org/>

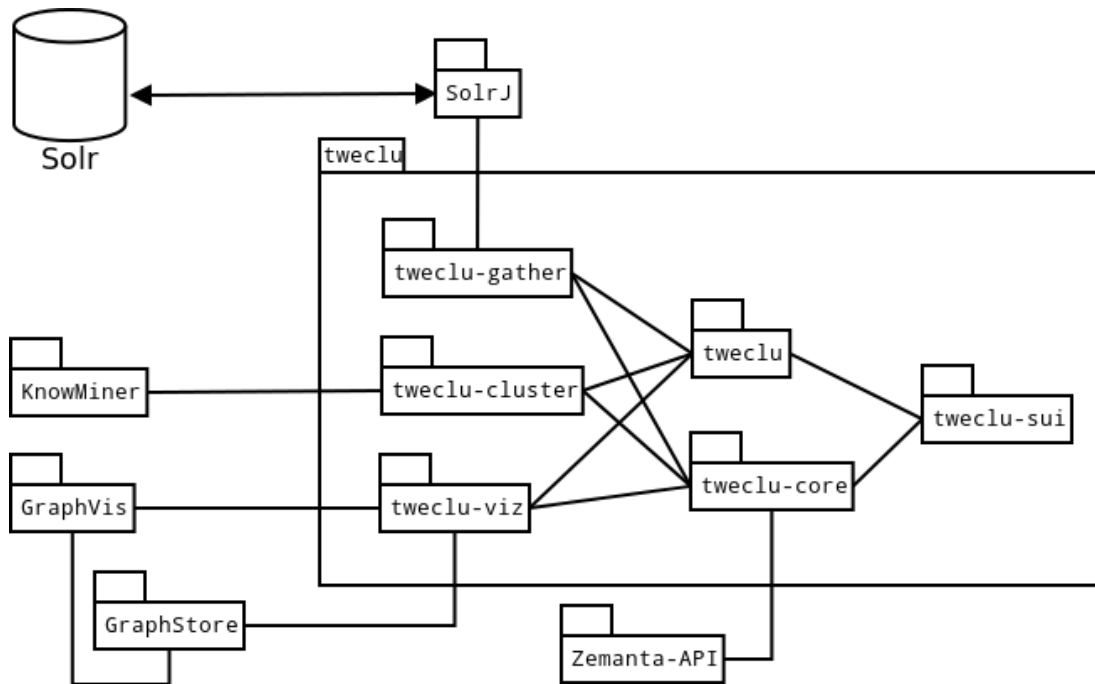


Figure 7.1: System components overview - The system components and their relations, dependencies to 3rd party libraries/components

artifact is built and deployed to the local maven repository. In order to re-use the artifact in another project the `groupId`, `artifactId`, and the `version` are needed, which are all defined in the POM.

A maven server is used in order to manage artifacts. The public maven central repository¹ is a server accessible for everyone and provides many open source projects². This allows the easy use of 3rd party libraries.

Maven was chosen as build tool because of its easy dependency declarations, and also because all the libraries needed for the system were available on maven repositories.

¹<http://search.maven.org/>

²<http://blog.sonatype.com/people/2010/12/now-available-central-download-statistics-for-oss-projects/>

Java

As a programming language Java¹, a product of Oracle², is used. The reasons for choosing Java are:

- a Java application can run on any platform where a Java Virtual Machine (JVM) can be executed
- an acceptable amount of machine learning libraries and frameworks are available and offer an application programming interface (API) for Java
- a Java client to query Solr is available, provided by Apache, named Solrj³

7.1.2 3rd party Components & Libraries

This section captures basic information on the 3rd party components and libraries used, as illustrated in figure 7.1.

SolrJ

SolrJ offers the possibility to execute operations on a Solr server out of a Java application. Operations include adding, removing, and deleting of documents, but also performing query operations.

Zemanta API

The API offered by Zemanta⁴ is an engine suggesting related content from various sources in the Internet, with respect to the data that is sent for analyzing. The data sent is the text of interest that needs to be analyzed⁵. The content suggested is linked to, for instance, Wikipedia⁶, YouTube⁷, and IMDB⁸. If applicable, a

¹<http://www.java.com/>

²<http://www.oracle.com>

³<http://wiki.apache.org/solr/Solrj>

⁴<http://developer.zemanta.com/docs>

⁵http://developer.zemanta.com/media/files/docs/zemanta_api_companion.pdf

⁶<http://www.wikipedia.org/>

⁷<http://www.youtube.com/>

⁸<http://www.imdb.com/>

response also contains the entity type, and indicates to what extent the result is of relevance.

KnowMiner

KnowMiner¹ is a knowledge discovery framework developed over several years at the Know-Center². It is "a service oriented architecture designed with the primary goal to [...] provid[e] a rich set of knowledge discovery functionalities for very different application scenarios" (Klieber et al., 2009). According to Klieber et al. (2009), KnowMiner includes the following services:

- Import
- Information extraction
- Feature extraction
- Information retrieval
- Association
- Summarization
- Clustering
- Classification

For this thesis the clustering service is of importance. According to Klieber et al. (2009), k -means, ISODATA, and other clustering algorithms are implemented. The KnowMiner API only exposes a subset of the implemented clustering algorithms, for instance k -means, Affinity Propagation and Hierarchical Agglomerative Clustering (HAC). The KnowMiner API further disguises any use of other services to the programmer that are used in combination when invoking a clustering algorithm.

KnowMiner was chosen to be the framework for this thesis because of its extensive offer of services, and the facts of being open source and providing a

¹<http://knowminer.know-center.tugraz.at/>

²<http://know-center.tugraz.at/>

Java API. Also, the additional services may be of relevance when extending the system with additional features.

GraphStore & GraphVis

GraphStore is a Storage API for graph data as a basis for scalable graph visualization. It can be used to store graphs, for instance hierarchical relationships. GraphStore furthermore supports to store meta-data to nodes and links. These meta-data can be, for instance, references to individual image icons that shall be displayed next to the actual data.

GraphVis is a hierarchical graph visualization based on hierarchically clustered graph datasets. It reads the data to render from a GraphStore object. Based on the data in the GraphStore, GraphVis calculates the node layout, edge aggregation, grid generation, the edge routing, and the edge bundling. (Kienreich et al., 2012)

Both GraphStore and GraphVis are in active development at the KnowCenter, but have already been applied successfully in various projects, cf. Granitzer et al. (2004) and Muhr et al. (2010).

As discussed in section 4.1.5, Voronoi polygons have been used successfully to visualize clusters. GraphVis supports layouting of clusters in Voronoi polygons. Furthermore, it supports interactive exploration of the data visualized. Because of the benefits mentioned, and the flexibility in customizing individual elements of the resulting visualization, GraphVis was chosen for visualizing the results.

7.2 The Tweclu Pipe

This section provides detailed descriptions of the work-flow represented in the User Interface (UI) and the settings for each step of the Pipe. The Pipe is defined in the *tweclu* component and is the only point with access to the three main components processing the data: *tweclu-gather*, *tweclu-cluster*, and *tweclu-viz*. The processing of the data is defined by each components' interface, as illustrated in figure 7.2.

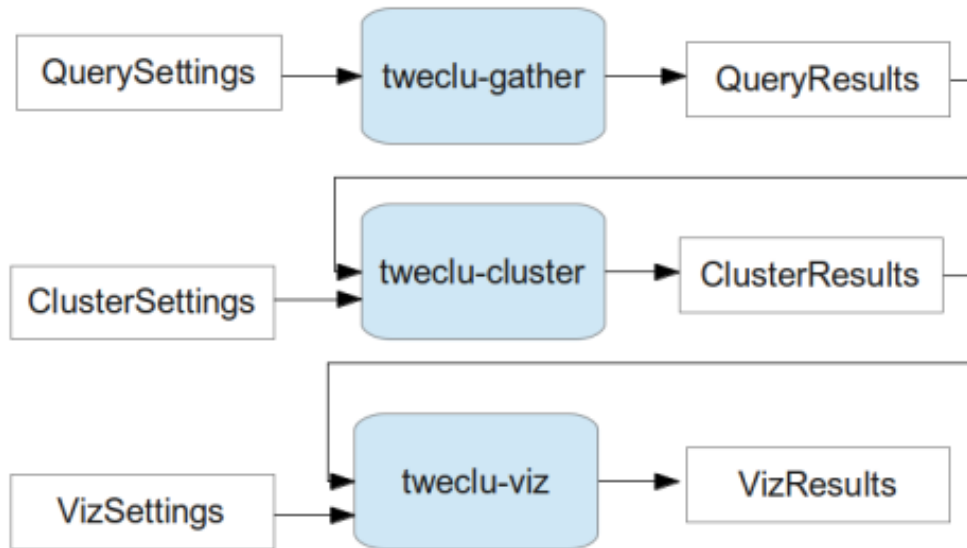


Figure 7.2: Components' Interfaces - In- and Outputs of the modules

Figure 7.2 displays the expected inputs, as well as the generated output of each component. The components illustrated, *tweclu-gather*, *tweclu-cluster*, and *tweclu-viz*, take their generally defined input and convert it further, depending on the selected *components'* defined input. These components are defined with the properties set in the `Query/Cluster/Viz-Settings` class defined for each component.

In order to keep track of the data belonging to one Job, a `Job` class was implemented. The in- and outputs of the various components are reflected in the `Job` class.

7.2.1 Query

This component is usually the first being triggered. According to the `QuerySettings` passed, a Solr query is created and sent to the Solr instance configured via `SolrJ` (cf. subsection 6.2.2). The results retrieved from Solr are then converted into the defined output of this module. The `QueryResults` object contains, in simplified terms, a list of documents that are filled up with the data retrieved from Solr. Before returning the results, individual fields of a document can be sent to Ze-

manta in order to enhance the document with further data. This additional data can also be used for analysis in the next module, *tweclu-cluster*.

The basic settings to provide for a query are illustrated in figure 7.3. In order to process a user's request, he/she is required to fill at least the text fields represented in area (2) with data. The text fields illustrated include the *search term*, the *start* and the *end date*. In order to make it more useable, users can choose to fill the text fields with pre-defined values via the drop-down choice, represented in area (1).

With respect to the settings provided, a Solr query is generated and executed twice. By nature, Solr only returns either a pre-defined number of results (default is set to ten), or the number of results set to the Query object. The object that gets returned from Solr includes a field indicating the total number of results: `totalResults`. Hence, when first executed, the number of results to return is set to zero and the total number of results is read out of the field. The second execution of the query is set to return `totalResults` results. The results are then converted to a `QueryResult`, and all results are added to `QueryResults`, the output as illustrated in figure 7.2.

Area (3) in figure 7.3 provides additional options that can be included in the process of gathering the data from Solr. For instance, for each individual tweet, concepts can be gathered via the Zemanta API. Therefore, each tweet is sent to Zemanta individually. The returned concepts are added to the corresponding `QueryResult`. Querying Zemanta is time consuming due to the limitation defined by the Zemanta API, which only allows one query every second. Thus, it is possible to store the concepts to the Solr index for later use. Furthermore it is possible to just call Zemanta and save the results to a file without proceeding to clustering. This option was introduced to investigate the concepts returned for the tweets.

7.2.2 Cluster

After the query is executed and the `QueryResults` generated, clustering can be performed on the results of the data gathering process. The possible settings are presented in figure 7.4 and its various options are described as follows:

7.2 The Tweclu Pipe

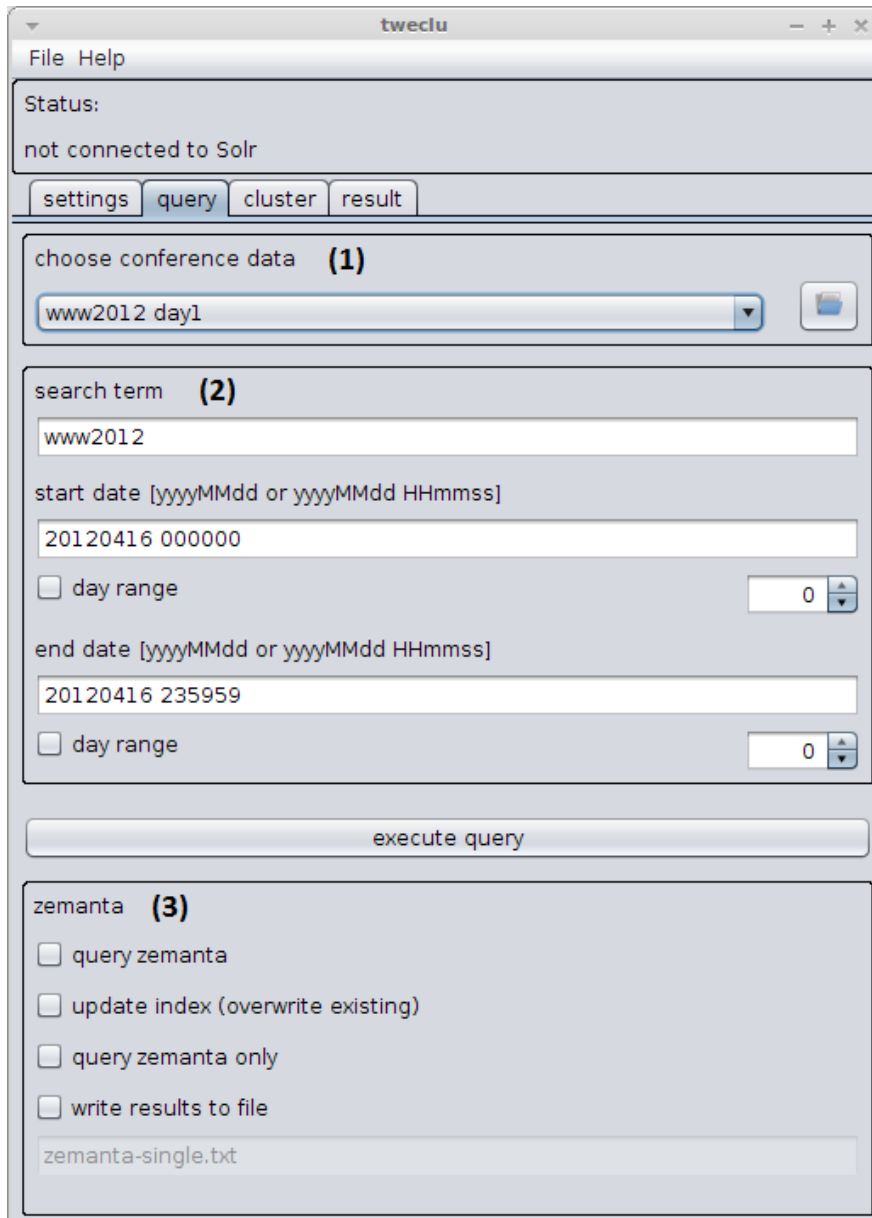


Figure 7.3: UI - Query tab - The query tab providing options to set the QuerySettings

7.2 The Tweclu Pipe

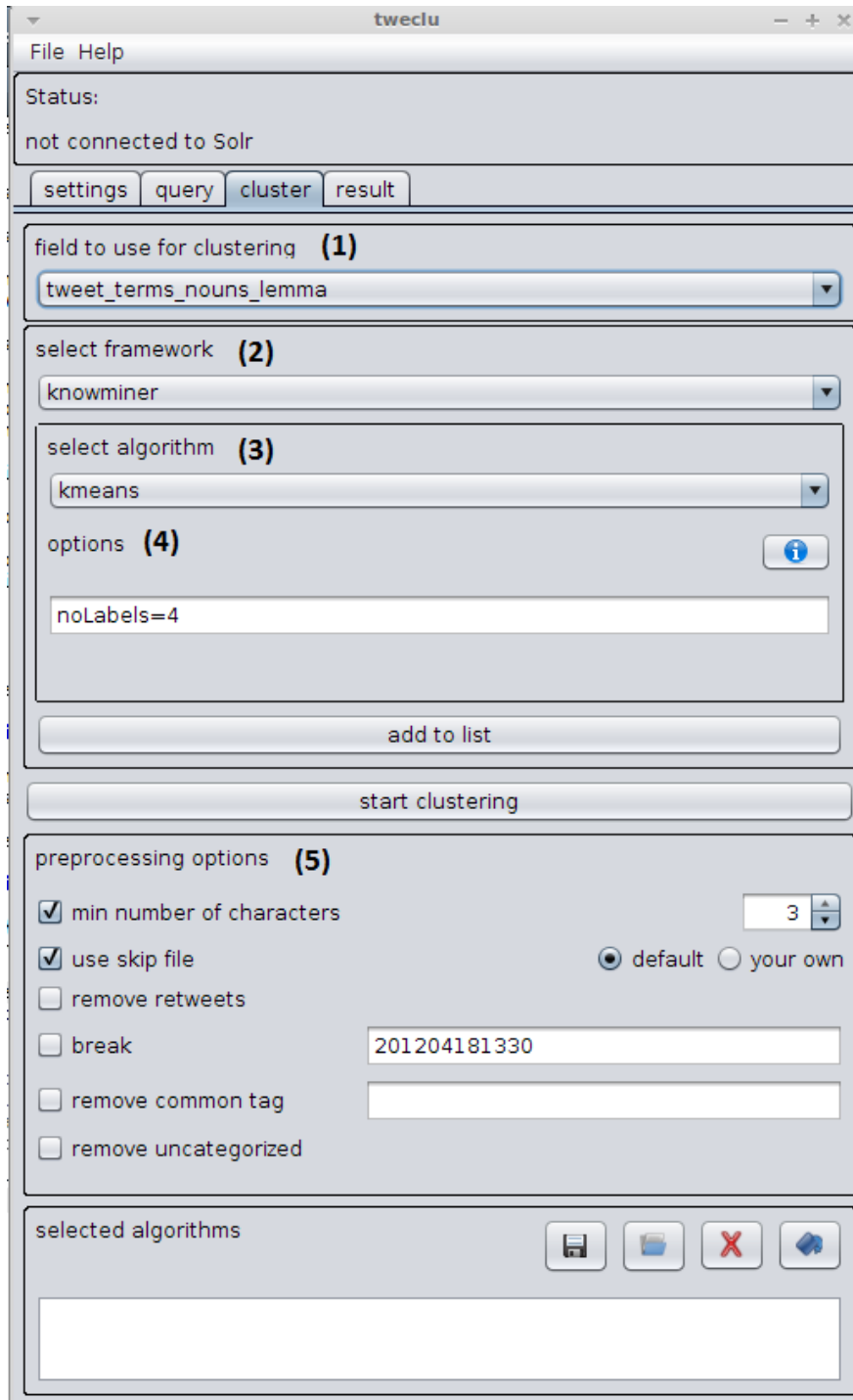


Figure 7.4: UI - Cluster tab - The cluster tab providing options to set the ClusterSettings

- (1) the field representing the data being sent to the clustering algorithm for processing has to be specified. The user can choose from the following fields: `tweet_terms_multi`, `tweet_text_topics`, `tweet_terms_nouns_lemma`, `tweet_text_reduced` (see subsection 6.2.3 for a description of the fields' contents).
- (2) the clustering framework, KnowMiner, has to be selected.
- (3) an algorithm can be selected. The algorithms are restricted to those offered by the selected API (cf. subsection 7.1.2).
- (4) options for the clustering algorithm can be provided.

The options that can be set to an algorithm are displayed in figure 7.5 and are described as follows:

noLabels defines the number of labels for one cluster

noClustersMin defines the minimum amount of clusters

noClustersMax defines the maximum amount of clusters

hi indicates if the results shall be a hierarchy instead of a flat clustering result (cf. chapter 3)

discr indicates to use discriminative labeling method instead of descriptive (cf. subsection 3.1.3)

dist defines the distance function that shall be used, whereby the user can choose from three options: *euclidean* for Euclidean distance, *hamming* for Hamming distance, and *dublin* for Dublin Core distance (cf. subsection 3.1.2)

Furthermore, the user can select some pre-processing steps that will be applied on the data before the data is processed with the clustering algorithm. The possible options are displayed in the *pre-processing options* area, (5), illustrated in figure 7.4. Its impacts are described as follows:

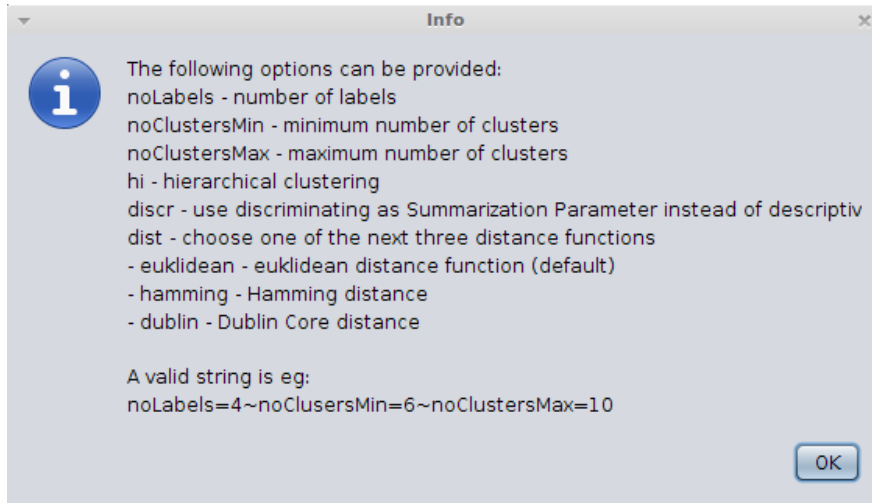


Figure 7.5: Cluster options - Additional options for KnowMiner's clustering algorithm

min number of characters Enabling this option and setting a value `minChars` greater than zero, terms that have fewer characters as `minChars` will be removed before processing the data.

use skip file As tweets may contain terms that are not of relevance (e.g. *haha, omfg, woo*) a *skip* file, indicating terms that shall be skipped, can be used. These *skip* terms depend on the data corpus that is analyzed, thus the user can either change the contents of the default *skip* file, or provide his or her own. If this option is selected, the terms contained will be removed before processing the data.

remove re-tweets This option allows to remove re-tweets. Re-tweets are only recognized by the character sequence *RT @*, thus not all kinds of re-tweets are considered (cf. Weller et al. (2011)). Re-tweets are used more often during conferences, and mostly indicate a popular tweet, cf. section 2. This option was introduced in order to check the influence of re-tweets on the clustering results.

break This field indicates to separate the tweets into one data set before the break, and one after. When enabled, the data sets are processed individually, but added to one `ClusterResult`, which will be displayed in one graph.

remove common tag Enabling this option allows the user to remove a common hash-tag in order to check the influence on the clustering results. This option is only relevant when selecting `tweet_terms_multi` as field for the data source.

remove uncategorized This option is only relevant when considering the field `tweet_text_topics` as data source. Enabling this option removes all tweets where no category was detected by Zemanta (cf. subsection 7.2.1).

When all preferences are set, the `QueryResult` is first processed according to the pre-process settings provided by the user. Then, clustering is initialized and its options set. The pre-processed data is clustered and finally converted into `ClusterResults`, which is the output of this component, as illustrated in figure 7.2. The `ClusterResults` are temporarily serialized onto the local hard drive, using a unique file name for later re-use.

7.2.3 Visualization

The last tab in the collection of tabs provides input elements in order to set the `VizSettings`. The possible settings are presented in figure 7.6 and are described as follows:

remove RT Re-tweets might be of relevance when processing the data, but would overcrowd the resulting visualization. Thus, this option allows to remove all re-tweets from being visualized.

view data selecting this option adds the tweets to the corresponding cluster. The tweets then represent the last layer, which allows the user to view the tweets contained in a cluster, the details, on demand (cf. section 4.2).

draw links enabling this option will draw links from one cluster to another. The system considers clusters to be linked when both clusters have at least one label in common.

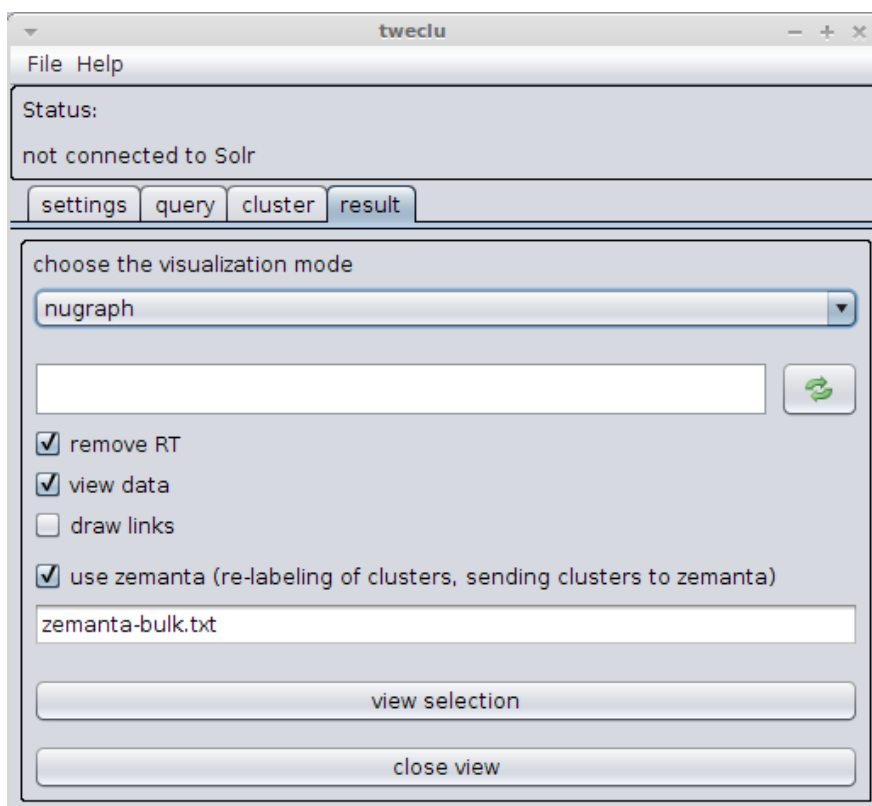


Figure 7.6: UI - Result tab - The results tab providing options to set the VizSettings

use zemanta this option is used in order to generate new labels for the clusters.

Therefore, all data contained in one cluster will be sent to Zemanta. The resulting concepts are sorted, and the top concepts, based on the amount of occurrences, used as labels for the clusters.

The `VizProcessor` performs the operations according to the settings. It also downloads the profile images of all users contained in the data set, and saves them to a `temp`-folder. These images have to comply a specific dimension, thus, some are resized in order to fulfill the dimension criteria.

`GraphVis` is used in order to display the `GraphStore` (cf. subsection 7.1.2). When initializing `GraphVis`, the API provides settings in order to customize the resulting visualization. These settings include, for instance, the minimum and maximum font size, edge colors, and node picture. If the mouse cursor hovers over a node, different settings can be provided. Another feature is to display the cluster borders in the layers below. While exploring the visualization, the user is able to realize which tweet belongs to which cluster. Furthermore, `GraphStore` offers to set images to nodes. To take advantage of this feature, nodes representing a cluster are displayed with the Twitter logo, and the tweets within the cluster display the profile image of the user next to the tweet itself. In order to comply with the regulations of Twitter¹, which demand that the tweet, the user name, as well as the time stamp of the tweet are displayed, this data is stored in the `GraphStore`. Figure 7.7 illustrates the layer of tweets that belong to one cluster, and figure 7.8 displays an excerpt of clusters.

7.3 Tweclu components

The components *tweclu-gather*, *tweclu-cluster*, and *tweclu-viz* have been discussed in section 7.2 in detail, thus only the remaining components implemented in the course of this thesis, as illustrated in figure 7.1, are described in this section.

¹<https://twitter.com/logo> last accessed October 20, 2012

7.3 Tweclu components



Figure 7.7: tweet (data) layer - An exemplary result of a tweet layer

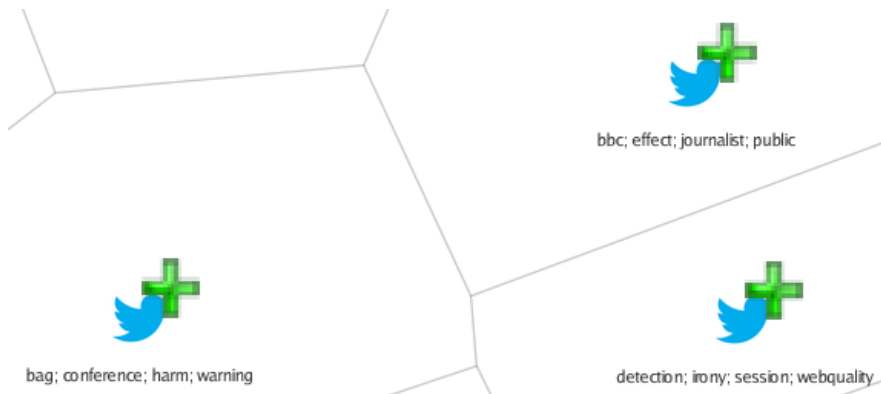


Figure 7.8: Cluster layer - An exemplary result of a cluster layer

7.3.1 tweclu-sui

This component represents the User Interface (UI) of the system, which is implemented with the Java Swing graphical user interface (GUI) widget toolkit. The application can distinguish two modes of operations: **CLUST** and **VIEW**. When starting with the **CLUST** mode, the UI is organized into a status panel indicating the connection to the data source, and four tabs: *Settings*, *Query*, *Cluster*, *Results*, whereas the latter three are shown in figures 7.3, 7.4, and 7.6 respectively. The *Settings* tab, shown in figure 7.9, offers to reload the configuration and to change the URL to Solr. The next three tabs represent the steps of the *Pipe* and provide the user to specify the settings for each step, as explained in section 7.2.

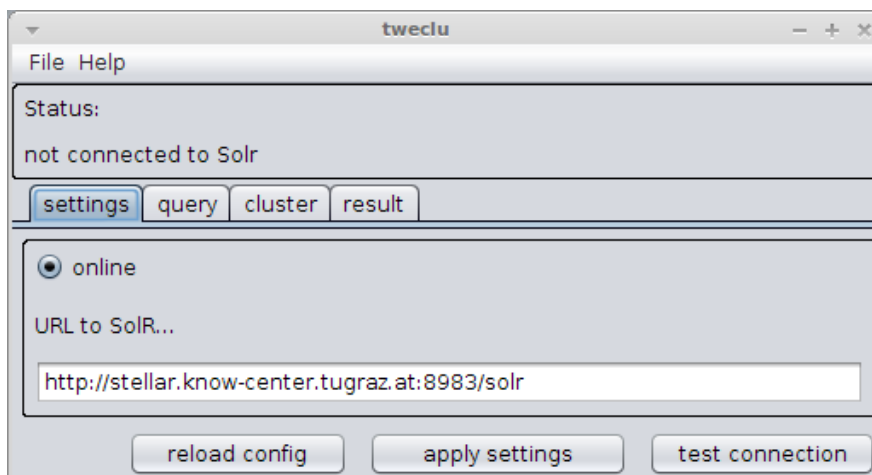


Figure 7.9: UI - Settings tab - The settings tab providing to reload the configuration and change the URL to Solr

When starting the UI in **VIEW** mode, only one tab is shown, as illustrated in figure 7.10. This one tab provides the option to load a (previously stored) *GraphStore* and display the results.

7.3.2 tweclu

As mentioned in the beginning of section 7.2, *tweclu* is considered the heart of the application. Thus, no direct access to the components actually performing

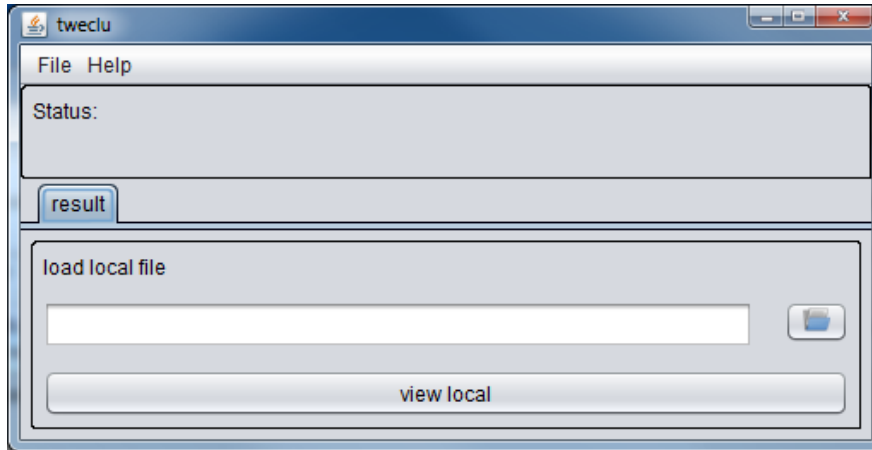


Figure 7.10: UI - VIEW mode - The application started in VIEW mode

operations on the data, *tweclu-gather*, *tweclu-cluster*, and *tweclu-viz*, is considered. This module offers implementations of the interface `IPipe`. A concrete implementation can define the steps it wants to perform. In the course of this thesis it is used to perform one step after another, as the user has to provide the settings for each step sequentially. Assuming that the settings for each step are provided in advance, a pipe that performs all steps at once, without waiting for any user input, can be used as well.

Tweclu also loads and performs checks of the basic configuration during start-up. An important configuration parameter is the path to the *data folder*. The *data folder* has to exist and read and write permissions have to be provided in order to continue with the start-up. The data folder was introduced because (a) various files have to be written to the hard disk, and (b) additional settings can be defined in separate files (cf. subsection 7.2.1). These additional settings files were introduced as these preferences are not likely to change much over time. Therefore, if added as additional controls to the user interface (UI), it would only have lead to an overloaded UI, and would probably have disrupted the work flow, and the user. The data folder consists of three used sub-folders:

data stores GraphStore elements and corresponding serialized objects in order to get reloaded later on

store contains additional settings and images used for visualization

`tmp` is used to store temporary downloaded files from the Internet

7.3.3 tweclu-core

tweclu-core represents the core of the system. This package contains the classes that are generated as output or expected as input for the various modules, cf. for instance figure 7.2. Furthermore, all configuration and settings classes are defined in this core component, utilities that are of general use, as well as an interface to the Zemanta API (cf. section 7.1.2).

7.4 Notes on the Implementation

- Because of the input and output interfaces defined for each component the components can also be used individually, with respect to their expected inputs.
- The components are designed to be extensible with a logical structure behind them. In the case that more complex logic is added to the implemented components the architecture should be revised. Each module should get its own project and invoking of the desired module can be achieved via injection in order to keep it coupled loose.
- The final implementation only supports Solr as a data source for querying data, but the architecture allows to add support for other data sources as well.
- More than one algorithm can be added to an existing `Job`. This functionality was added as users might want to process the same data with different algorithms, or different algorithm settings, and compare the results side by side. Unfortunately `GraphVis` does not support loading two `GraphStores` and displaying them side by side, nor is it possible to start two instances of `GraphVis` side by side. Thus adding more than one algorithm to a `Job` is not used in the course of this system.

7.4 Notes on the Implementation

- Besides storing to GraphStore, also writing a JavaScript Object Notation (JSON) file is supported.

8

Evaluation and Results

Having decided on *how to cluster tweets* (RQ1), and *how to visualize results* (RQ2), this chapter presents the chosen procedure and analysis in order to get answers on *how well the results represent certain aspects of the conference* (RQ3), and *how usable the presented results are* (RQ4).

Expert interviews were chosen in order to evaluate the results. Thus, the data set was specified to be the tweets from the World Wide Web conference in 2012 (WWW2012). The participants of the evaluation were experts in the area covered during the WWW2012, but also attended the conference and interacted with Twitter in the course of the conference. Section 8.1 illustrates background information on expert interviews.

A description of the data set and the participants is provided in sections 8.2 and 8.3. Section 8.4 illustrates the setup of the expert interviews, and section 8.5 finally presents the results.

8.1 Expert Interviews

The evaluation was based on expert interviews, cf. Meuser and Nagel (2005). Experts are selected by the interviewer based on the research area and especially the research question he/she is dealing with. Expert interviews are open interviews in order to get the expert's impressions and opinions. Open guidelines ensure that

the experts get to share their knowledge, but also to keep comparability between the results. The evaluation of the results aims at comparing the answers of all experts and finding results that are common for all experts, stressing interpretations and interpretive patterns, and finding structures of relevance. (Meuser and Nagel, 2005)

8.2 The Data set

8.2.1 Tweet Selection

Tweets are published for various events and various topics, thus hash-tags are used to roughly assign a tweet to a topic and/or event (cf. chapter 2). Hence, tweets considered to be part of the data set have to meet specific criterias. The selection of tweets is defined as following:

- tweets have to be within a specific domain, for the purpose of this thesis it shall be scientific conferences
- tweets have to be published to the public, and not only as private posts (as they would not be available in our data set)
- tweets have to be published within the days the conference actually took place
- tweets have to have a *hash-tag* indicating that the message is actually supposed to be part of the conference

The data collection process and the description of the data stored for further processing is explained separately in section 6.2. In the course of this work the World Wide Web (WWW) conference¹ was chosen. During the first two days of the conference, workshops took place. These two days are considered to be part of the scientific conference as well, although the main conference started on day three.

¹<http://wwwconference.org/>

8.2.2 The World Wide Web Conference

The WWW conference is an annual event which started in 1994¹. The goal is "to provide the world a premier forum for discussion and debate about the evolution of the Web, the standardization of its associated technologies, and the impact of those technologies on society and culture." The conference has become very popular over the years. (WWW, 2012)

8.3 Evaluation Participants

The data set chosen is specific to scientific conferences. Thus, the evaluation participants had to be experts in the field the conference was aimed at. In order to be considered an expert, the participant had to meet the following criteria:

- the participant had to be an active visitor at the conference chosen
- the participant had to have experience in the domain of the conference chosen

In total, four experts evaluated the results of the system. The experts will be referred to as E0, expert zero, through E3, expert three. Subsection 8.5.2 presents details on the participants.

8.4 Evaluation Setup

An evaluation was performed once for each participant, and the duration was set to a maximum of one hour. Furthermore, the evaluation was conducted in German, as all participants' first language is German.

8.4.1 Evaluation Procedure

The procedure was the same for each evaluation and is defined as follows:

¹in the first two years it was held twice

1. ***greeting*** the evaluation started with an informal greeting where it was explained what this evaluation is about and what to expect from the evaluation.
2. ***statistics*** at this point the expert filled out and signed the participant sheet, which was used to gather background information on the experts
3. ***motivation, refresh*** as some time had passed since the conference took place, some questions were asked in order to help refresh the memory and also to get individual information on the expert with regard to the conference. This part was dominated by open questions.
4. ***introduction to tools*** in this step, visualizations were loaded with data different from the actual evaluation data in order to let the user get a feeling on how to handle the visualizations. The exploration was limited to a maximum of five minutes.
5. ***specific refresh*** at this point the expert was told the specific day that was selected to be evaluated. In order to remember the specifics about that day the expert was shown the agenda, as well as the Twitter stream of the particular day. This step was limited to a maximum of five minutes as well.
6. ***evaluation*** in this step each visualization was shown to the expert successively and for each visualization the questions were asked, again successively. The time limitation was set to ten minutes for each visualization.
7. ***feedback*** at the end of the visualization the expert was asked to provide feedback about the evaluated visualizations. This part was dominated by open questions again.

In order to get insights on the participants' Twitter usage behaviour and feedback on the visualizations shown, pre-defined questions were formed.

Motivation, Refresh

During the third step of the evaluation procedure, the participants were asked four questions in order to make the participant recap the conference on a general level, but also to get background information on the Twitter use of each individual participant. The questions asked were:

- *Q3-1* How did you like the conference?
- *Q3-2* What did you expect from the conference?
- *Q3-3* Did you follow the Twitter stream of the conference?
- *Q3-4* Did you tweet as well?

Evaluation

As part of the actual evaluation the following questions were defined to be answered for each visualization presented (cf. section 1.2):

- *Q6-1* How well are thematic topics covered?
- *Q6-2* How well are organizational topics/events covered?
- *Q6-3* How well are highlights represented?

Feedback

During the last phase of the evaluation the participant was asked to provide feedback of the presented visualizations in order to collect feedback on the usability and applicability of the visualizations, but also their personal point of view on Twitter visualizations. Therefore, the following questions were asked:

- *Q7-1* Do you think the presented visualizations are better than just a list of tweets?
- *Q7-2* What did you like and not like about the visualizations presented?

- *Q7-3* Do you think the visualization would be helpful for other purposes?
- *Q7-4* Which visualization of the presented did you like most?

8.4.2 Participant Sheet Contents

In the participant form the experts were asked to provide their name, gender, highest academic degree, current status, and age. Furthermore the discipline they are specialized in had to be selected, as well as the frequency of social media applications had to be stated. The possibilities printed on the sheet representing the stages of frequency concur with the stages stated in Bortz and Döring (2001): always, often, sometimes, rarely, and never. The complete participant sheet can be found in Appendix C, cf. page 100

8.4.3 Evaluation Data & Settings

For the evaluation only one day was chosen to be visualized: day three, the first main conference day of the WWW2012 conference. The total number of tweets of that day is 1838, cf. table 8.2. Based on the selected day, three different visualizations have been prepared for the participants to evaluate. Two visualizations were generated using the system implemented in the course of this work, but with different settings. The third one was the weighted graph from Kraker et al. (2011) in order to provide a different visualization to compare to.

The resulting graph is illustrated in figure 8.1. It contains the nouns that occurred with the term *www2012* and its relations to one another. Unfortunately, labels are overlapping in the center. This is because the label *track* has connections to other labels which only have a connection to the one label, *track*. This causes the labels to overlap. The overlapping labels didn't have any influence on the evaluation results, because participants understood the reason.

The other two visualizations were, as mentioned earlier, generated with the system implemented, but with different settings. The settings both resulting Voronoi visualizations share, are the following:

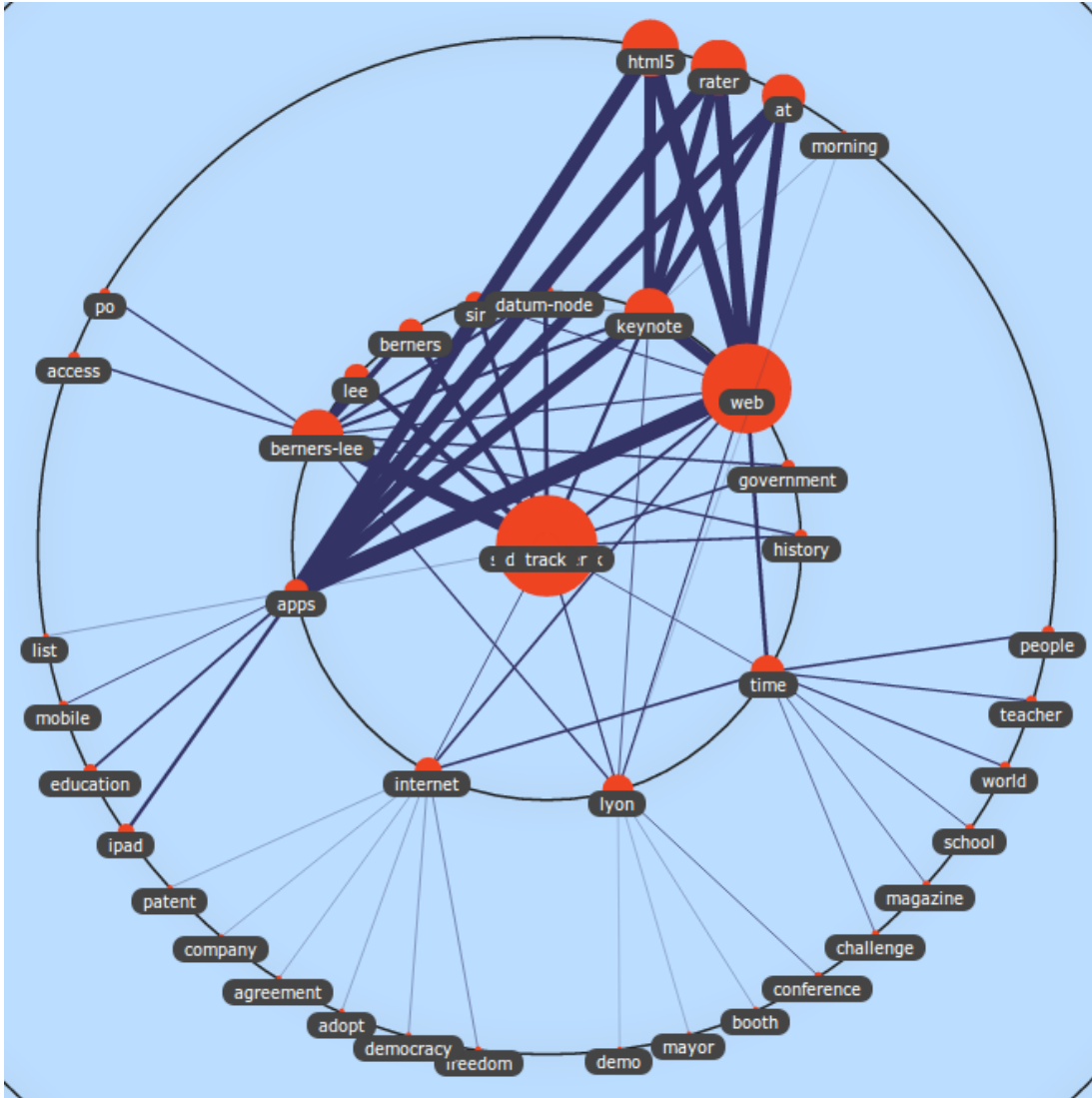


Figure 8.1: Results - Weighted Graph - the resulting visualization for the selected conference day with weighted graph

- the Solr field `tweet_terms_multi` was used as data field for further processing (which contains both the nouns and hash-tags of the tweet, cf. subsection 6.2.3)
- KnowMiner’s k -means clustering algorithm was used in order to process the data and the Euclidean distance was chosen to be used as distance measure
- all re-tweets were used for clustering, but were removed before the Graph-Store was generated in order to keep the resulting visualization clear
- the tweets are displayed as the last layer in the visualization with the user profile images next to the tweet content and its meta data (as illustrated in figure 7.7)
- the boundaries between the clusters were set to be displayed in order not to lose track when exploring the tweets within a cluster (cf. figures 7.7 and 7.8)
- the green ”plus” sign was introduced to illustrate that data is available in the corresponding cluster
- the size of the Twitter bird and the green ”plus” sign indicated the size of the cluster, if more or less tweets belonged to the cluster respectively

The difference in the settings lies in the label generation. The resulting visualization of the first setting used is illustrated in figure 8.2 and uses the KnowMiner’s label generation, which was set to be discriminative (cf. subsection 3.1.3). The setting will be referenced to as *Voronoi multi*.

The resulting visualization of the second setting used is illustrated in figure 8.3. Instead of using the most descriptive words of the cluster, all tweets belonging to a cluster were sent to Zemanta. The concepts found by Zemanta were sent back. The top concepts have been considered to describe the cluster (cf. subsection 7.2.3). This setting will be further referenced as *Voronoi Zem*.

The three results were evaluated successively. In order to avoid results based on the presented sequence, the order of the visualizations were commuted for each participant. Table 8.1 shows the final variations for each of the four experts, E0

8.4 Evaluation Setup

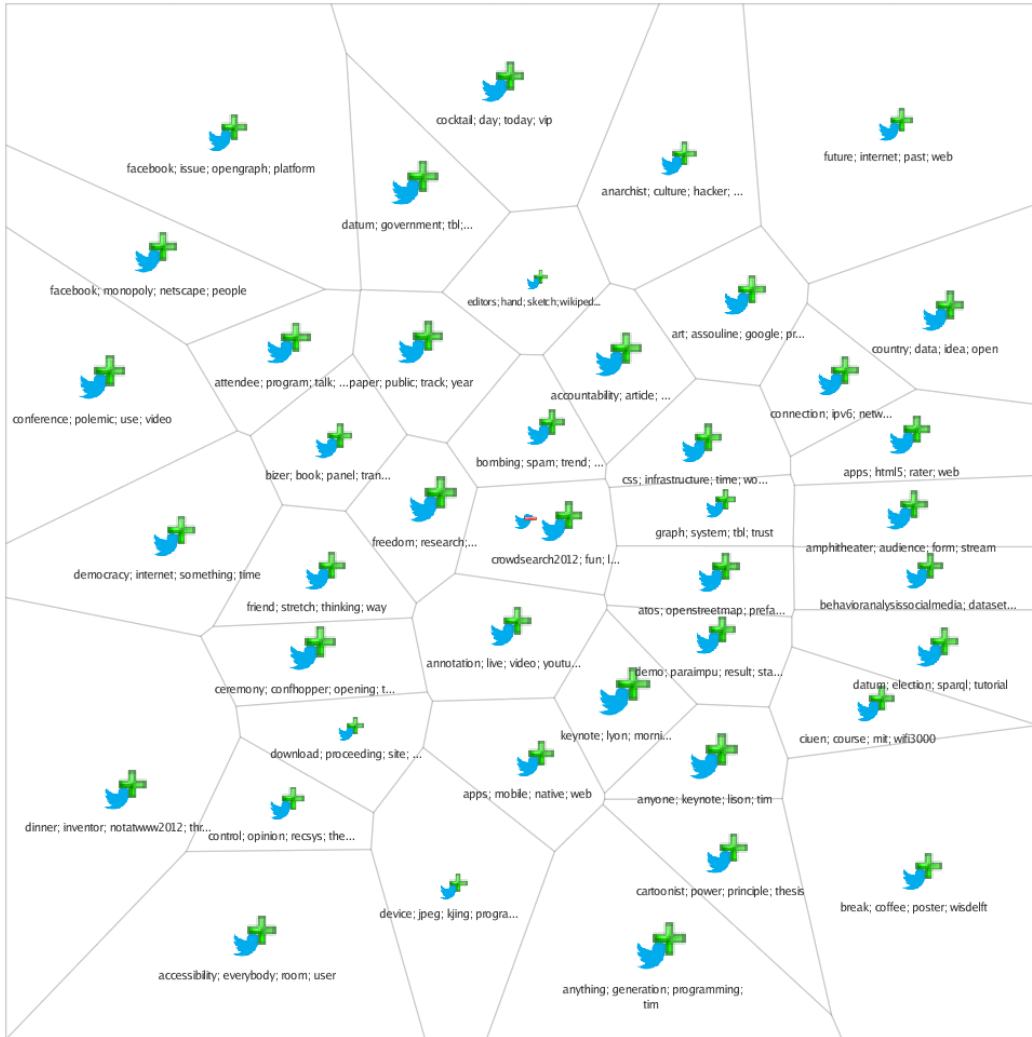


Figure 8.2: Results - *Voronoi multi* - the resulting visualization for the selected conference day with *Voronoi multi* settings

8.4 Evaluation Setup



Figure 8.3: Results - Voronoi Zemanta - the resulting visualization for the selected conference day with *Voronoi Zem* settings

	E0	E1	E2	E3
#1	nouns	WG	Zem	multi
#2	WG	multi	WG	Zem
#3	multi	Zem	multi	WG

Table 8.1: Orders of the visualizations for each participant

through E3; *WG* refers to the weighted graph, *Zem* to the setting using Zemanta, and *multi* to the setting using KnowMiner’s labeling.

As illustrated in table 8.1, E0 stands out as it uses one setting that was not used in further evaluations, using the Solr field *tweet_terms_nouns_lemma* as the data field. This setting will not be covered when analyzing the results.

8.4.4 Evaluation Analysis Procedure

The evaluations were held as semi-structured interviews and have been recorded on audio. Some parts of the evaluation procedure, as described in subsection 8.4.1, were accompanied by pre-defined questions which served as jump start, in order to get the participant talking and, on occasion, to ask follow-up questions.

The evaluations were conducted in German, thus the recordings have been transcribed and translated to English language. The rules for the transcription process were defined as following:

- filler words have been removed
- irrelevant statements have been removed, for instance small talk
- duplicate statements have been removed
- unfinished statements, sentences have been completed using conversational context

The transcribed answers were qualitatively analyzed using coding, a reducing and interpreting approach. The coding scheme was defined with categories capturing the responses related to the questions *Q6-1* to *Q6-3* (cf. subsection 8.4.1). Furthermore, codes to capture responses to the individual visualizations were introduced, in order to capture the responses to questions *Q7-1* to *Q7-4* (cf.

subsection 8.4.1). After analyzing the data more carefully, codes for Twitter visualizations in general and ideas for improvements and ideas in general, were introduced. In order to get both positive and negative responses, sub codes were introduced for the main code categories mentioned. In the next step, statements assigned to the same code were summarized. Therefore, similar phrases were merged in order to remove redundancy.

As for evaluating what aspects of a conference are represented, the questions *Q6-1* to *Q6-3* (as presented in subsection 8.4.1) are analyzed separately. Also RQ4 is analyzed separately, summarizing the questions *Q7-1* to *Q7-4* for each visualization.

8.5 Evaluation Results

8.5.1 Dataset Analysis

In order to conduct experiments and an evaluation, tweets from the World Wide Web (WWW) conference 2012¹ were crawled. The conference took place in Lyon, France, and lasted from April 16 until April 20, 2012. The hash-tag assigned to the chosen scientific conference was *#www2012*. The conference is focused in the field of computer science and had 1497 attendees². Over the course of the five days the conference was held, 6901 tweets could be crawled, and transferred into the Solr index. The tweets gathered all contained the official conference hash-tag, *#www2012*. Table 8.2 lists the number of tweets gathered for each conference day separately in the first two columns.

Further analysis of the data set confirms the findings of Weller et al. (2011) (cf. subsection 2.5) that at least 25% of the tweets contain uniform resource locators (URLs), and around 50% of the tweets are re-tweets. Table 8.2 contains the breakdown of the tweets based on the conference day, whereas column two represents the total number of tweets, column three the number of tweets containing a URL, and column five the number of re-tweets. Furthermore, out of

¹<http://www2012.wwwconference.org> last accessed October 20, 2012

²<http://internetetmoi.fr/www2012/> last accessed October 20, 2012

8.5 Evaluation Results

Day	# of tweets	# of links	%	# of retweets	%
1	1190	590	49.6	505	42.4
2	990	525	53.0	396	40.0
3	1838	686	37.3	844	45.9
4	1762	636	36.1	938	53.2
5	1121	455	40.6	547	48.8
Total	6901	2892	41.9	3230	46.8

Table 8.2: The test dataset for tweets with conference hash-tag #www2012 and the number of tweets with URLs, and re-tweets respectively

the 1838 tweets of day three, 849 (46,19%) were tweeted during the keynote or annotated with the hash-tag *keynote*. This great amount of tweets dealing with the keynote is reflected in the results. Figure 8.4 presents exemplary tweets from the data set, whereas figure 8.4a shows a re-tweet, figure 8.4b a tweet containing an URL, and figure 8.4c a re-tweet containing an URL.

8.5.2 Participants

The evaluation was conducted with four experts. The results from the participant sheet were as follows: three of the participants were, at the time of the evaluation, active PhD students, whereas one participant currently undertook post-doctoral research, but all participants were mainly active in the discipline of computer science. The age range was from 24 until 32 years. According to the participants' statements the most used social media tool was Facebook, closely followed by Twitter. Further social media affine platforms are used by individual participants.

For one participant, the WWW2012 was the first conference he/she attended, the other participants have attended up to ten conferences before. All participants liked and enjoyed the conference very much; *"[the WWW2012] has been the best conference I've attended so far."* Furthermore, all participants did expect interesting presentations of current research, but also an opportunity for networking and exchanging ideas and information with like-minded people.

Moreover, in order to get some idea of the Twitter usage behavior by the participants during the conference, the participants were asked about their tweeting,



RT @pmika: @rtroncy @moustaki And I assume all the people who actually care about httprange-14 are in that room #philoweb #www2012 @p_ansell tweeted on 2012/04/18 04:08:40

(a) **Re-tweet** - an example of a re-tweet



Would be nice if the #www2012 talks on <http://t.co/5nXCfvnW> would include a *link* to the papers on the proceedings @pautasso tweeted on 2012/04/18 16:23:57

(b) **Tweet including URL** - an example of a tweet including an URL



RT @calexy: @bbfish is giving a talk on distributed and secured social Web applications #webid #www2012 <http://t.co/oWpLCYb> @bbfish tweeted on 2012/04/18 17:27:27

(c) **Re-tweet including URL** - an example of a re-tweet including an URL

Figure 8.4: Examples of tweets from the dataset

and tweet reading behavior. They stated that they tweeted more than usual during the conference. Two participants stated that they checked the Twitter stream several times a day, whereas the other two only checked it occasionally. One of the latter stated that he/she checked the updates from people he/she followed more than the stream itself, "*as [there] were too many tweets*".

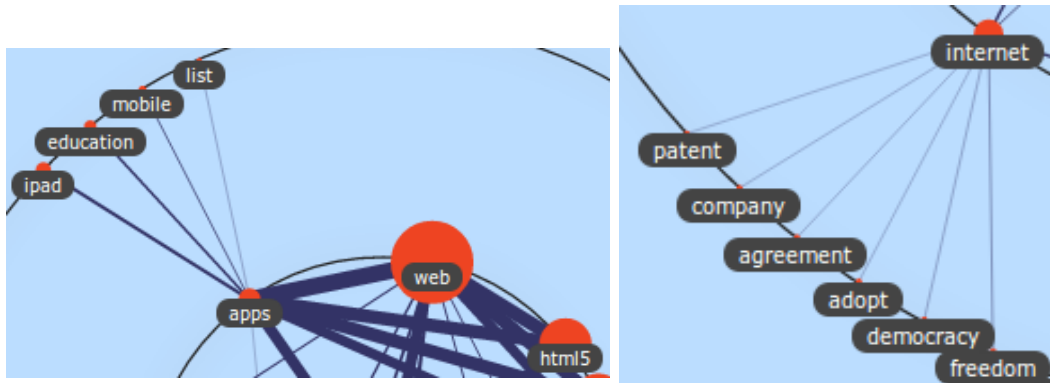
8.5.3 Evaluation Analysis

Q6-1 How well are thematic topics covered?

In the case of the *Weighted Graph*, thematic topics are partly covered. As the tweets in the data set were mainly about the keynote, the topics covered in the keynote are visualized best, as one participant stated: "*the topics from the keynote are depicted fairly good, although [the vision] gets blurred by labels that are not really topics.*" Figure 8.5 presents selections from the result identified as being part of the keynote. Furthermore, it contains labels that are neither topics, nor related to the conference. For instance, the terms *list* and *adopt*, illustrated in figures 8.5a and 8.5b, are labels that may have appeared in the data set more often than other words, but are of no relevance in the context of the conference. Thematic topics from the sessions and developer tracks are not covered at all in the resulting visualization. One participant, who was explicitly looking for a specific session he/she participated in the afternoon, stated explicitly "*the topics not covered are social media analysis, security in the Internet, and user behavior*".

As for *Voronoi multi*, thematic topics are covered very well. Again, the results showed that the topics covered in the keynote were dominant. The terms mentioned the most, with regard to the keynote, were *HTML5*, *democracy*, and *privacy*. Also the topics from the sessions in the afternoon are partly covered. Participants were able to identify the topics of the sessions they participated in the afternoon. The identified sessions differed as their interests lied in different areas. An example of a topic dealt with in a session is illustrated in figure 8.6, where a label represented the session dealing with *behavioural analysis* is shown.

Whereas most of the topics were found on the cluster label layer, one participant recognized a tweet, illustrated in figure 8.7a, that was published in the



(a) **General Terms** - a selection of terms from the *Weighted Graph* visualization

(b) **Further General Terms** - another selection of terms from the *Weighted Graph* visualization

Figure 8.5: Selection of terms from *Weighted Graph*



Figure 8.6: Identified Session with *Voronoi multi* - exemplary cluster labels including the topic of a session

course of a session. *"This was the session I was in that dealt with link farming. The labels don't represent [the session], but the contents do, at least partly."* In order to illustrate the specific situation, figure 8.7b shows the corresponding cluster labels where the session itself, or the greater topic *security*, is not represented. Furthermore, figure 8.7c illustrates that tweets in the cluster reference to the session mentioned, as the one in the lower right. The others are mirrored in the labels, and seem to be partly connected at some point.

In the case of *Voronoi Zem*, thematic topics are mainly covered at an abstract level only. Figure 8.8 illustrates the problem mentioned, as, for instance, *HTML* is more general than *HTML5*, which has been covered in the other resulting visualizations. Thus, the labels retrieved from Zemanta are too abstract. They also contain many names and other entities not related to computer science, the main field of the conference. One participant stated: *"the visualization contains more nouns, many persons appear, and 'sunny california', which is a bit irritating."*

Some labels represent the exact topic from sessions though. For instance, one participant immediately recognized the security session he/she found with *Voronoi multi* as well, which was just described before. Also *recommender system* was associated to a session in the afternoon of that day. The mentioned clusters are illustrated in figure 8.9a and figure 8.9b respectively.


Q6-2 How well are organizational topics/events covered?

Whereas all participants related terms like *coffee*, *break*, and *lunch* as organizational topics/events, one participant also related the sessions depicted from the previous question as an organizational event.


As for *Voronoi multi*, organizational events are represented well. One participant stated that *"organizational topics are clearly evident"*. Besides that, the terms mentioned most often were *coffee* and *VIP*, with the relation to *coffee breaks* and the *VIP cocktail* respectively. Figure 8.10 illustrates a cluster containing the organizational event *VIP cocktail* depicted with both the terms *VIP* and *cocktail* in the same cluster, as they appeared together very often.

In the case of the *Voronoi Zem* visualization, the results depended on the visualization evaluated before. In the case of *Voronoi Zem* being the first visual-

8.5 Evaluation Results



very interesting talk on link farming
on Twitter was presented right now #www2012
- the room seems to be t...



bombing; spam; trend; twitter

(a) **Identified tweet** - the tweet that was identified as being part of the participated security session

(b) **Corresponding Cluster Labels** - the cluster labels of the identified tweet do not depict the session



OH by @vrando: a conversation: twitter
but offline #www2012 #fb
@juansequeda
tweeted on 2012/04/18 17:54:09



I am thinking of applying for the contract.
My joke of a medical report will be online
tonight. going to GMC wit...

(c) **Selected tweets within the Cluster** - selected tweets within the same cluster only partly fitting



Criminal twitter accounts seem to be
socially connected and form a small
world #SecuritySocialNetworks #www2012
@ph_singer
tweeted on 2012/04/18 14:44:20

Figure 8.7: Identified Session with *Voronoi multi*

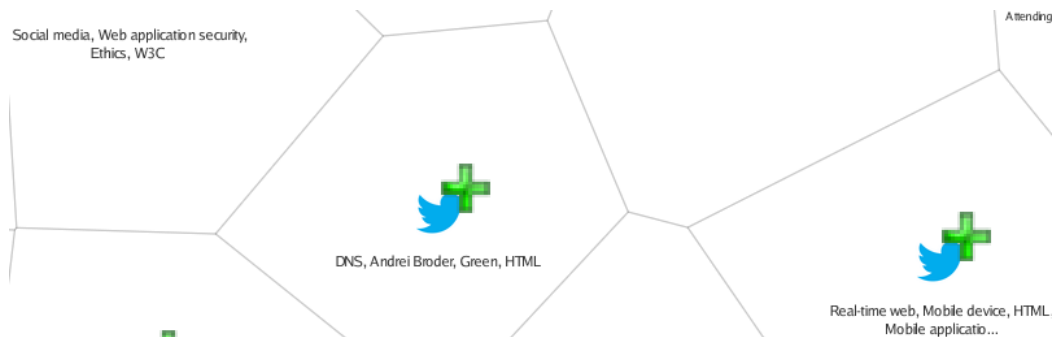


Figure 8.8: General Terms with *Voronoi Zem* - An excerpt from the results showing that terms are more general than specific

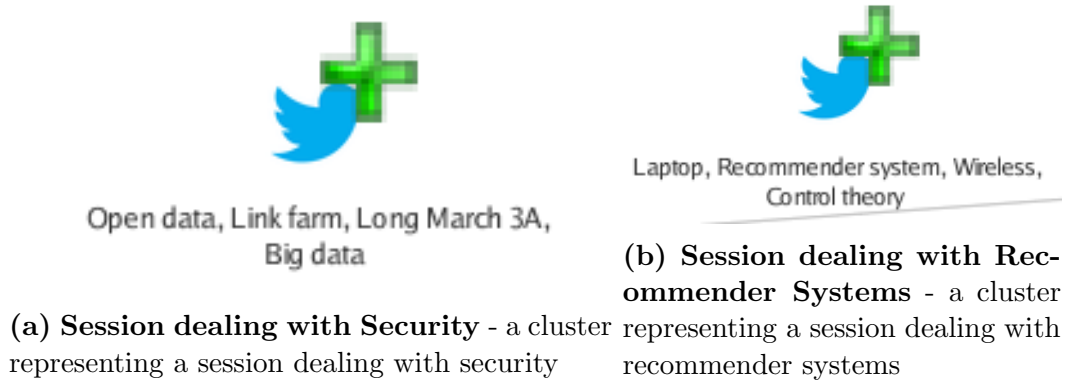


Figure 8.9: Identified Session with *Voronoi Zern*

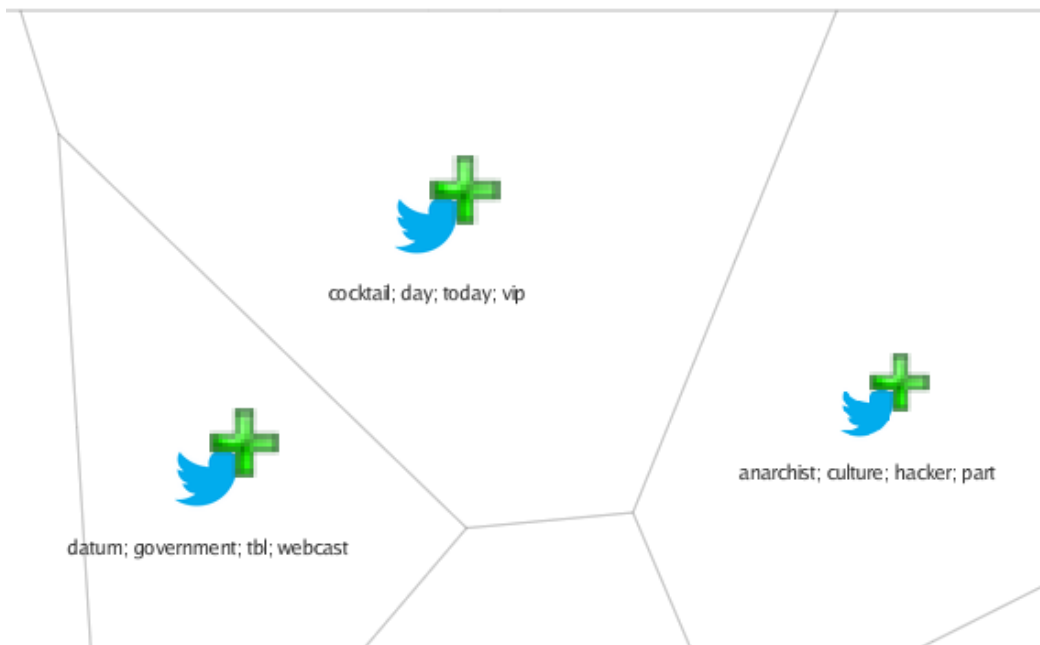


Figure 8.10: Organizational topics/events with *Voronoi multi* - exemplary excerpt from clusters representing organizational events/topics

8.5 Evaluation Results

ization to evaluate, no organizational topics were found. In the case of evaluating *Voronoi multi* beforehand, the participants stated the ones mentioned before as missing. As one participant put it appropriately, "*I'm missing the cocktail reception, and I also don't see the [term] keynote.*" But besides that, other organizational events/topics were found. For instance, the label *Hotel de Ville* was associated with "*the opening ceremony, where the mayor held a speech*"; the related cluster is illustrated in figure 8.11. Also, as with this setting more names appeared as labels, some names were associated with either local public figures, or people who held a speech. In general, organizational topics/events are mainly represented based on location and/or names of persons or buildings, which are represented by named entities.



Figure 8.11: Organizational topics/events with *Voronoi Zem* - exemplary cluster representing the hotel where the opening ceremony was held

As for the *Weighted Graph*, organizational events/topics were very rare, and very abstract. Participants stated that they related the term *keynote* to an organizational event, and also the term *conference*, representing the conference *WWW2012* in its entirety. Besides that, no social events were recognized, as participants stated; "*I'm missing social events*", or "*I'm missing coffee breaks, the cocktail, and also the lunch.*"

Q6-3 How well are highlights represented?

The results showed that the keynote was the big highlight for all participants. Besides that, participants mentioned the following being personal highlights as

8.5 Evaluation Results

well: technologies, and the sessions participated and identified in the course of the first question. Participants remembered the findings from the first question, and answered this question based on the previously achieved knowledge. This led to a much faster execution than presumed. But still, while quickly exploring the visualization, participants were mainly looking for one term and/or cluster representing one of their highlights.

As for the *Weighted Graph*, all agree that the keynote is depicted by the term *keynote* itself. Besides the keynote, *HTML5* and *apps* were mentioned the most being a highlight that was represented. Depending on the evaluation order presented in table 8.1, participants who had found and indicated sessions as their highlights before, stated that they were not present, based on the knowledge achieved from the previously presented results, and the questions asked beforehand.

In the case of *Voronoi multi*, the participants agreed that the keynote is only depicted by its contents, not the term itself. Furthermore, participants stated that "*[the keynote] is distributed over several clusters*". This made it more difficult to name the keynote as being represented, as they were first looking for the term itself. Figure 8.12 illustrates a selection of clusters containing topics covered in the keynote, and distributed over several clusters. In contrast to the *Weighted Graph* results, participants were able to identify the sessions and stated their presence, again based on the knowledge achieved from the previous questions. *HTML5* and *apps* were again mentioned the most as present highlights.

In the case of *Voronoi Zem*, the participants agree that the term *keynote* is not covered. Furthermore, as the contents are more general ones, they answered the question with "no". This was again stated based on the knowledge achieved by answering the first two questions. Furthermore, as with *Voronoi multi*, the terms that were somehow related to the keynote were distributed, as illustrated in figure 8.13. Also for the sessions, participants answered the question based on the knowledge achieved by answering the previous questions.

8.5 Evaluation Results

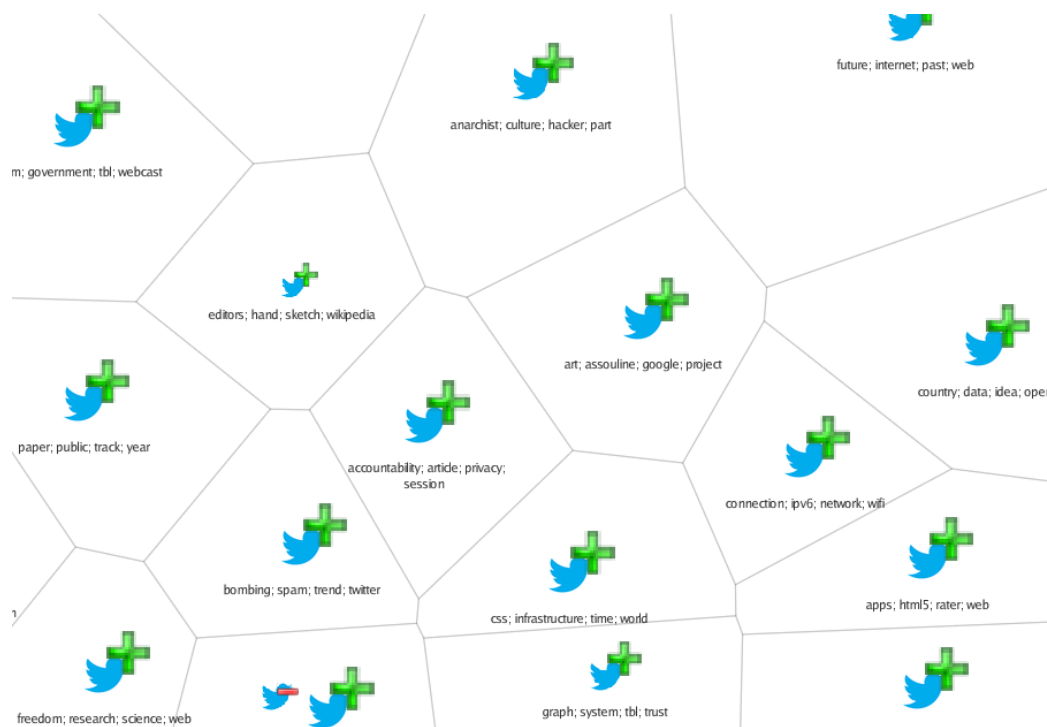


Figure 8.12: Keynote in *Voronoi multi* - exemplary clusters indicating that the keynote is represented in various clusters

8.5 Evaluation Results

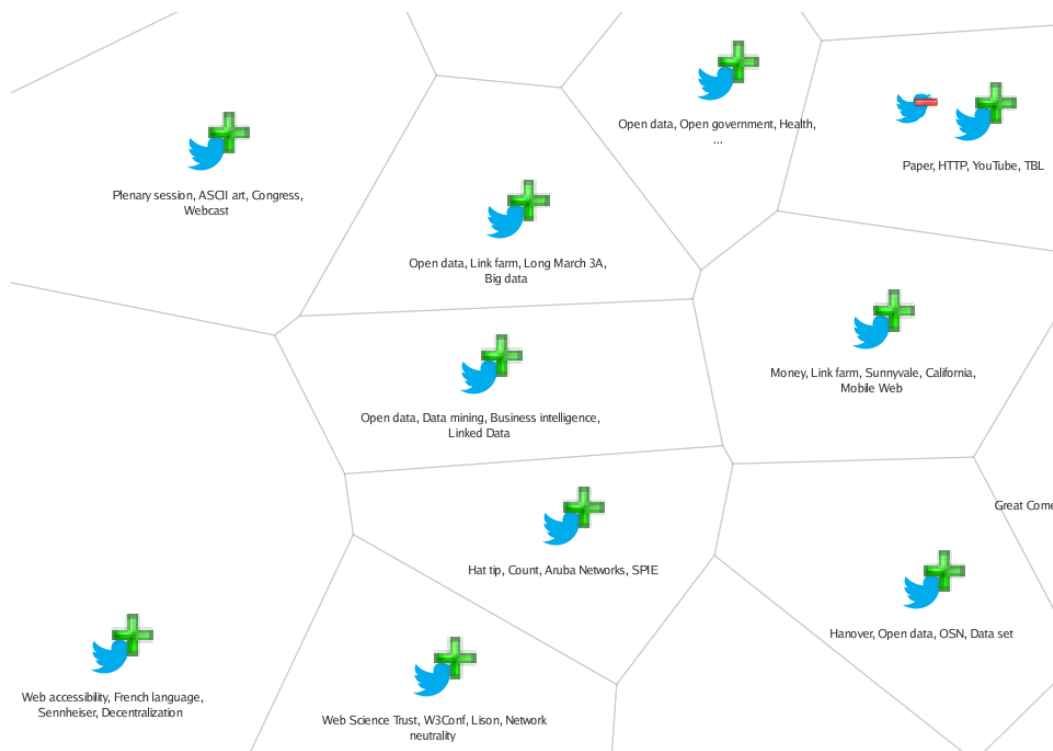


Figure 8.13: Keynote in *Voronoi Zem* - exemplary clusters indicating that the keynote is represented in various clusters

RQ4 - How usable are the presented results?

The participants liked the idea behind the visualizations. They noted that, especially after some time has passed, it helps one to remember topics of a conference day and the conference in its entirety more easily. Furthermore, they enjoyed the interactivity with the presented data.

In general, participants noted that visualizations have merits, especially in the case when people want to get an overview of topics dealt with on a conference. Journalists and marketing were mentioned as further use cases.

Weighted Graph

In the case of the Weighted Graph, participants especially stated that *"the line thickness was good"*, which represented the amount one term occurs with another. People noted that the edges are useful as one *"can see the relations between the terms"*, and build its own abstraction. Besides the positive feedback also negative issues were pointed out. First and foremost, more than one label was placed in the center, thus labels were covered by others. As a consequence, once a participant clicked on a centered label the graph had to be reloaded in order to function correct again. Participants also mentioned that the result *"gets blurred by labels that are not really topics."* In summary, participants liked its clearness and the relations in between the terms.

Voronoi

As for the Voronoi visualizations, participants were very pleased with the handling and the possible interactivity with the results. Also positively mentioned was the fact that the tweets were presented in the corresponding cluster. Alongside participants enjoyed the tweeters profile image displayed next to the tweet contents, which helped the participants identify friends and or commonly known Twitter users. Despite the positive feedback also several shortcomings were pointed out. Firstly, *"it would be a benefit having a bigger cluster size in case the topic is more important"*. The cluster size was only reflected by the size of the Twitter bird, which was irritating as they were in the center and the Voronoi areas in

8.5 Evaluation Results

the center occupied less space than the Voronoi areas on the side. They found that events, especially the keynote, were split over several clusters, without any relation showing up in the visualization. Furthermore, "*similar or equal labels are present in more clusters, which does not help with getting an overview*" They also found that tweets within a cluster do not always correspond to the labels of the cluster.

In summary, participants preferred the results of *Voronoi multi* over *Voronoi Zem*. With *Voronoi multi* some labels seem to be off-topic or not appropriate. Nevertheless, the terms could be found within the clusters' tweets. In contrast, with *Voronoi Zem* more labels seemed to be not appropriate, as there were too many general terms that needed more interpretation in order to assign them to topics dealt with.

9

Discussion & Conclusions

The evaluation results presented in section 8.5 indicate that, especially with *Voronoi multi*, topics that have been dealt with in a scientific conference can be extracted with clustering, and also organizational topics/events. As for the usefulness of using Twitter data, participants stated that the visualizations helped them to remember topics more easily. Furthermore, the results wouldn't just be of benefit for the research community, but also for journalism and marketing.

Nevertheless, the results also revealed the need to (1) either adapt the layouting algorithm of the Voronoi areas, or look for similar alternative visualization frameworks to use, (2) introduce relations in between clusters, and (3) evaluate other clustering algorithms. These shortcomings need to be addressed in order to improve the results. Thus, this chapter provides suggestions on possible future steps based on the shortcomings mentioned in the evaluation results. The participants also provided ideas and personal needs on what they would expect from a visualization of tweets from scientific conferences during their evaluation. These insights and ideas will be covered as well.

Improving the Clustering Results

The fact that tweets are noisier, shorter, and don't follow the usual grammatical level used in normal documents, makes it more difficult to analyze them (cf.

chapter 1. In this thesis, clustering was chosen in order to process tweets. The algorithm chosen was k -means, as it performs well with a big amount of data, and it has been used in various other research dealing with short texts, cf. section 3.5. The evaluation results showed that not all tweets assigned to a cluster fit the labels of the cluster itself. In order to address this issue, different clustering algorithms could be implemented and their performances evaluated. One idea of a clustering algorithm that was already applied in research on Twitter data successfully would be Affinity Propagation, cf. Rangrej et al. (2011) and Kang et al. (2010). Affinity Propagation is a graph based clustering algorithm that is also used in document processing.

A further idea would be the use of auxiliary data, as presented in section 2.5. Whereas the research presented in section 2.5 mainly deals with auxiliary data from search results or DBpedia, the data that could be used in this specific context could be specific as well. For instance, the abstracts of the conference papers provided, or the conference papers themselves, or even the slides used during the presentations, or the contents of the web pages linked to in various tweets.

As presented in the results, the labeling with Zemanta didn't satisfy the experts, as the labels were not as precise as with the labeling offered by KnowMiner. With the help of auxiliary data, that is less noisy and follows grammatical rules, the labels would probably improve for both settings. Due to the fact that more data could be sent to Zemanta, the concepts returned might be more specific, than with the data-set which just consisted of tweets, and also be of better relevance to the content.

Another interesting task would be to map the tweets to their corresponding events/sessions (cf. section 2.5). This could also be achieved with the help of the abstracts of the conference papers presented during the sessions, which usually are published in advance to the event itself. By processing the abstracts in advance, a more detailed hierarchy would be possible, as the event itself is dealing with a specific topic, and the abstracts could provide the sub areas. Furthermore, when assigning the tweets to an event, the additional information could, for instance, generate a new sub area, not only with respect to the intended sub areas.

Improving the Visualization

The Voronoi visualization chosen for presenting the results of the data processing got positive feedback, as presented in the evaluation results. For instance, the interactivity with the data, being able to move around, zoom in and out, being able to view the tweets assigned to a cluster, and the tweeter's profile image next to the tweet content, were well received. But, the participants also mentioned shortcomings of the visualization.

First and foremost, participants noted that the size of the cluster being represented by the size of the bird instead of the occupied area of the Voronoi area was misleading. Thus, an interesting task would be to evaluate alternative visualizations that support the representation of the cluster size for the user on first sight, but still maintain the positive features of the Voronoi visualization framework used, or at least provide adequate alternatives.

Another drawback mentioned by a participant was the availability of the implementation, as it is a desktop application. Using an implementation of a visualization that can be displayed in a Web Browser, for instance with the use of HTML5 Canvas¹, or WebGL², would increase the availability, as also modern mobile devices provide support of the mentioned Web technologies. Furthermore, the current implementation can also be used as a service, and writing the results to JavaScript Object Notation (JSON) is already supported.

Adding an Additional Dimension

In the current implementation, the tweets are processed at once. This means, that even when clustering the tweets during the event, only tweets that have been gathered before the processing started are taken into account. In order to update the results with the newly added/gathered tweets, the process has to be started all over again. An interesting improvement would be to add time as an additional dimension. In order to update the visualization the moment a new tweet is published, or at least in reasonable intervals, incremental clustering

¹<http://www.w3.org/wiki/HTML/Elements/canvas>

²<http://www.khronos.org/webgl/>

could be used, cf. section 3.2. Implementing this idea would increase the usage as the overview would grow with the progress of the conference, it would be literally a living overview. Also, some participants stated that they would more likely use a visualization that updates itself with the latest tweets, or further information, during a conference, than a static generated visualization that needs to be triggered for an update.

A further task could be to add geographic information as additional dimension, based on the country set in a users Twitter account settings, not on the geographical information that may be provided with the tweet itself. By analyzing geographic information it would be possible to realize what topics are of more importance, more popular respectively, to specific regions in the world.

Further Ideas

As presented in the evaluation results, participants stated that the benefit of the weighted graph visualization is the connections in between the labels, which help in getting an idea on the affiliation of the labels, topics respectively. The participants also stated that the topics from the keynote, as presented with the Voronoi visualizations, are spread over multiple clusters. Thus, an interesting task would be to calculate connections of clusters, either based on the generated labels, or on the contents of the clusters in order to highlight affiliations in between clusters.

Participants also stated that, as they are experts, they would like to not only get an overview, but to dive into the results and have the sources presented to them in detail. Especially when adding auxiliary data, as suggested before, this task would be interesting not only from the point of view of the processing, but also from the viewpoint of usability, in order to not smite the user with too much information.

One participant mentioned that it would be very useful to show popular tweets. Based on this idea it would be an interesting task to provide the user with an interface where both an overview and popular tweets are presented. Additionally, indicating the cluster the tweet is contained in would increase the readability

for the end user. This could be achieved with highlighting of the corresponding areas.

Figure 9.1 illustrates a basic sketch of an idea for a future user interface (UI), which shares some ideas as presented in Kraker et al. (2013). The UI contains the user input area in the top (1), the overview presenting the clusters in the center (2), and the details based on the selected cluster in the bottom and the left (3). The sketch is based on the idea of using auxiliary data, the abstracts, but also calculating popular tweets. The highlighted cluster, *CL2*, and the corresponding tweets in the *popular* area, as well as in the details displayed in the bottom, shall indicate that all parts of the visualization are connected and update itself on a selection by the user.

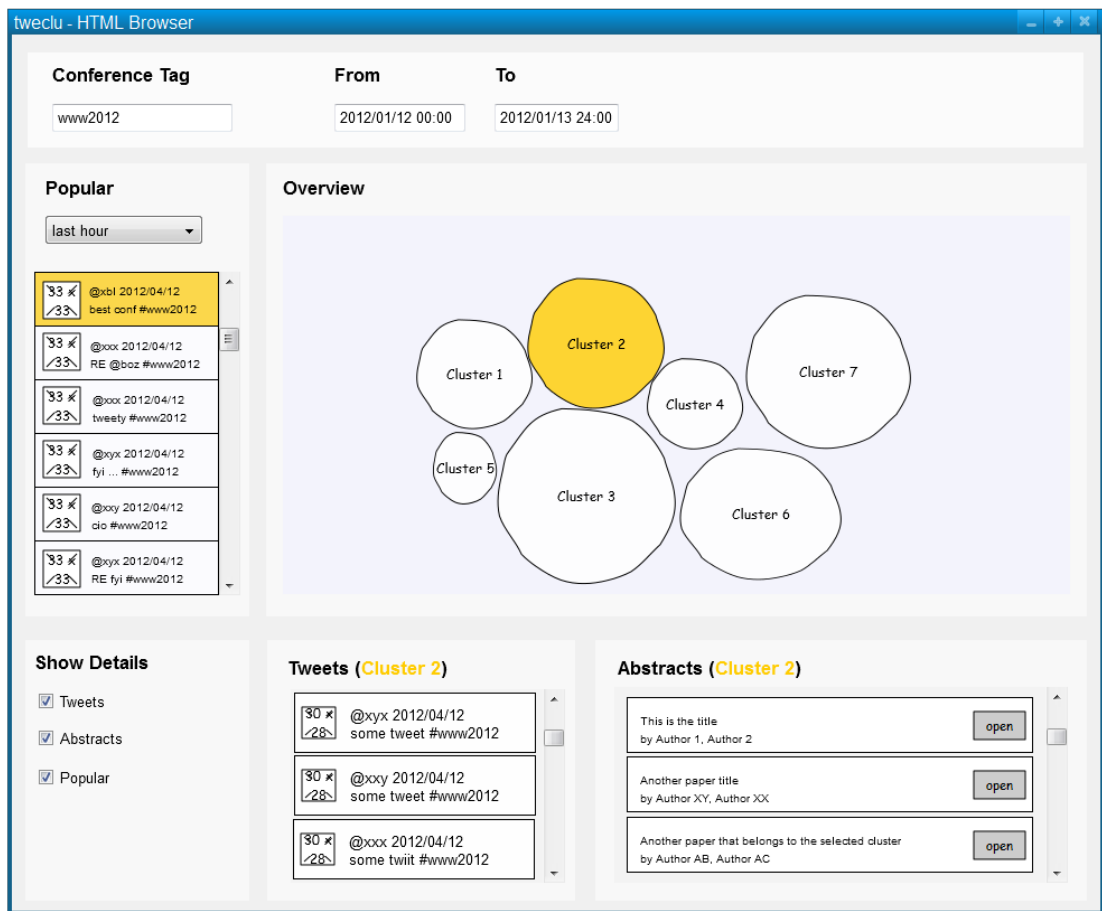


Figure 9.1: Improvement Proposal - A sketch of a possible UI including some of the ideas mentioned

Appendices

Appendix A: Solr Index Fields

The search index of Solr is defined in a configuration file named *schema.xml*. The following listing contains all fields defined for the Solr index used throughout this thesis.

```
<field name="tweet_id" type="string" stored="true" />
<field name="tweet_timestamp" type="date" indexed="true"
  stored="true" />
<field name="tweet_text" type="text_general" indexed="true"
  stored="true" termVectors="true" />
<field name="tweet_text_reduced" type="text_general"
  indexed="true" stored="true" termVectors="true" />
<field name="tweet_user_id" type="string"/>
<field name="tweet_user_name" type="string" stored="true"/>
<field name="tweet_user_screen_name" type="string"
  stored="true" indexed="true"/>
<field name="tweet_user_url" type="string" stored="true"/>
<field name="tweet_user_profile_image_url" type="string"
  stored="true" />
<field name="tweet_user_location" type="string" stored="true"/>
<field name="tweet_coordinates" type="string" stored="true"/>
<field name="tweet_coordinates_long" type="double" stored="true"/>
<field name="tweet_coordinates_lat" type="double" stored="true"/>
<field name="tweet_timeinserted" type="date" stored="true"/>
<field name="tweet_raw" type="string"/>
<field name="tweet_text_topics" type="text_ws" indexed="true"
  stored="true" multiValued="true" termVectors="true" />
<field name="tweet_nouns_topics" type="text_ws"
  indexed="true" stored="true" multiValued="true"
  termVectors="true" />
<field name="tweet_hashtags_topics" type="text_ws" indexed="true"
  stored="true" multiValued="true" termVectors="true" />
<field name="tweet_terms_hashtags" type="text_ws" indexed="true"
  stored="true" multiValued="true" termVectors="true"/>
<field name="tweet_terms_users" type="text_ws" stored="true"
  multiValued="true"/>
<field name="tweet_terms_retweets" type="text_ws" stored="true"
  multiValued="true"/>
<field name="tweet_terms_links" type="text_ws" stored="true"
```

```
    multiValued="true"/>
<field name="tweet_terms_uris" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_urls" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_keywords" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_nouns" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_nouns_lemma" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_verbs" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_verbs_lemma" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_adverbs" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_adverbs_lemma" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_terms_multi" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_user_multi_mentions" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_user_multi_retweets" type="text_ws" stored="true"
    multiValued="true"/>
<field name="tweet_source" type="string" stored="true" />
<field name="tweet_language" type="string" stored="true"/>
```

Appendix B: Sample Solr Entry

The following listing shows a data entry found in the Solr index that contained the hashtag #www2012, a link and is a retweet.

```
<doc>
  <str name="tweet_coordinates"/>
  <double name="tweet_coordinates_lat">0.0</double>
  <double name="tweet_coordinates_long">0.0</double>
  <str name="tweet_id">192554774397325313</str>
  <str name="tweet_language">en</str>
  <str name="tweet_raw">
    {"text":"RT @thetarro: The main conference #www2012 starts now!
    @ WWW2012 http://t.co/TL8Bj6zO",
    "possibly_sensitive_editable":true,"id_str":"192554774397325313",
    "coordinates":null,
    "created_at":"Wed Apr 18 10:06:50 +0000 2012",
    "in_reply_to_status_id_str":null,"favorited":false,
    "source":"\u003Ca href=\"http://twitter.com/download/android\"
    rel=\"nofollow\"\u003ETwitter for Android\u003C/a\u003E",
    "in_reply_to_user_id_str":null,"entities":{"urls":[{"
    indices":[66,86],"url":"http://t.co/TL8Bj6zO",
    "display_url":"instagr.am/p/JjTNhFtyjq/"}],
    "expanded_url":"http://instagr.am/p/JjTNhFtyjq/"}},
    "user_mentions":[{"indices":[3,12],"id_str":"6206062",
    "screen_name":"thetarro","name":"Salvatore Scellato",
    "id":6206062}], "hashtags":[{"text":"www2012","indices":[34,42]}]},
    "possibly_sensitive":false,"contributors":null,"place":null,
    "in_reply_to_screen_name":null,"in_reply_to_status_id":null,
    "geo":null,"user":{"is_translator":false,"statuses_count":5205,
    "time_zone":"London","profile_background_color":"CODEED",
    "id_str":"14771839","follow_request_sent":null,"verified":false,
    "profile_background_tile":false,"created_at":
    "Wed May 14 11:20:30 +0000 2008",
    "profile_sidebar_fill_color":"DDEEF6",
    "default_profile_image":false,
    "notifications":null,"friends_count":310,
    "url":"http://www.cl.cam.ac.uk/~nkl25/",
    "description":"Researcher @ Cambridge Computer Lab",
    "favourites_count":52,"profile_sidebar_border_color":"CODEED",
    "followers_count":651,
    "profile_image_url":
```

```

" http:\\\\a0.twimg.com\\profile_images\\935406753
\\profile_normal.jpg", "screen_name": "neal_lathia",
" profile_use_background_image": true,
" profile_background_image_url_https":
" https:\\\\si0.twimg.com\\images\\themes\\theme1\\bg.png",
" location": "City of London", "contributors_enabled": false,
" lang": "en", "geo_enabled": false,
" profile_text_color": "333333", "protected": false,
" profile_image_url_https":
" https:\\\\si0.twimg.com\\profile_images\\935406753\\
profile_normal.jpg",
" listed_count": 71, "profile_background_image_url":
" http:\\\\a0.twimg.com\\images\\themes\\theme1\\bg.png",
" name": "Neal Lathia", "profile_link_color": "0084B4",
" id": 14771839, "default_profile": true,
" show_all_inline_media": false, "following": null,
" utc_offset": 0, "retweeted": false, "id": 192554774397325313,
" retweeted_status": { "text": "The main conference
#www2012 starts now! @ WWW2012 http:\\\\t.co\\TL8Bj6zO",
" possibly_sensitive_editable": true,
" id_str": "192505590260563968",
" coordinates": { "type": "Point",
" coordinates": [4.85649236, 45.78504376] },
" created_at": "Wed Apr 18 06:51:23 +0000 2012",
" in_reply_to_status_id_str": null, "favorited": false,
" source": "\u003Ca href=\"http:\\\\instagr.am\"
rel=\"nofollow\" \u003EInstagram\u003C/a\u003E",
" in_reply_to_user_id_str": null, "entities":
{ "urls": [ { "indices": [52, 72], "url": "http:\\\\t.co\\TL8Bj6zO",
" display_url": "instagr.am\\p\\JjTNhFtyjq\\",
" expanded_url": "http:\\\\instagr.am\\p\\JjTNhFtyjq\\" } ],
" user_mentions": [], "hashtags": [ { "text": "www2012", "indices":
[20, 28] } ] },
" possibly_sensitive": false, "contributors": null, "place":
{ "bounding_box": { "type": "Polygon", "coordinates":
[[ [4.7718312, 45.7073626], [4.8983666, 45.7073626],
[4.8983666, 45.8082829], [4.7718312, 45.8082829] ] ] },
" place_type": "city", "country": "France", "url":
" http:\\\\api.twitter.com\\1\\geo\\id\\d091189af463dd4a.json",
" country_code": "FR", "attributes": {}, "full_name":
" Lyon, Rh\u00f4ne", "name": "Lyon", "id": "d091189af463dd4a" },

```

```
"in_reply_to_screen_name": null ,
"in_reply_to_status_id": null , "geo": {" type": " Point" ,
"coordinates": [45.78504376 , 4.85649236] } , "user": {
" is_translator": false ,
" statuses_count": 9930 , " time_zone": " London" ,
" profile_background_color": " 131516" , " id_str": " 6206062" ,
" follow_request_sent": null , " verified": false ,
" profile_background_tile": true , " created_at":
" Mon May 21 18:27:28 +0000 2007" ,
" profile_sidebar_fill_color": " efefef" ,
" default_profile_image": false , " notifications": null ,
" friends_count": 770 , " url":
" http://www.cl.cam.ac.uk/~ss824/" , " description": " Sicilian
engineer , studying online location-based social networks for
a PhD in Computer Science. Former intern at Google , still Coffee
Czar." , " favourites_count": 31 , " profile_sidebar_border_color":
" eeeee" , " followers_count": 808 , " profile_image_url":
" http://a0.twimg.com/profile_images/728549602/
my_eye_large_normal.jpg" ,
" screen_name": " thetarro" , " profile_use_background_image": true ,
" profile_background_image_url_https":
" https://si0.twimg.com/images/themes/theme14/bg.gif" ,
" location": " Cambridge , UK." ,
" contributors_enabled": false , " lang": " en" ,
" geo_enabled": true , " profile_text_color": " 333333" ,
" protected": false , " profile_image_url_https":
" https://si0.twimg.com/profile_images/728549602/
my_eye_large_normal.jpg" ,
" listed_count": 47 , " profile_background_image_url":
" http://a0.twimg.com/images/themes/theme14/bg.gif" ,
" name": " Salvatore Scellato" , " profile_link_color": " 009999" ,
" id": 6206062 , " default_profile": false ,
" show_all_inline_media": true ,
" following": null , " utc_offset": 0 } , " retweeted": false ,
" id": 192505590260563968 , " retweet_count": 0 ,
" in_reply_to_user_id": null , " truncated": false } ,
" retweet_count": 0 , " in_reply_to_user_id": null ,
" truncated": false }
</str>
<str name="tweet_source">
KCS-STELLAR/kc_crawler_1/science20_final.twitter.track
```

```
</str>
<arr name="tweet_terms_adverbs">
  <str>main</str>
  <str>now</str>
</arr>
<arr name="tweet_terms_adverbs_lemma">
  <str>main</str>
  <str>now</str>
</arr>
<arr name="tweet_terms_hashtags">
  <str>www2012</str>
</arr>
<arr name="tweet_terms_links">
  <str>http://t.co/TL8Bj6zO</str>
</arr>
<arr name="tweet_terms_multi">
  <str>#www2012</str>
  <str>conference</str>
  <str>www2012</str>
</arr>
<arr name="tweet_terms_nouns">
  <str>conference</str>
  <str>www2012</str>
</arr>
<arr name="tweet_terms_nouns_lemma">
  <str>conference</str>
  <str>www2012</str>
</arr>
<arr name="tweet_terms_retweets">
  <str>thetarro</str>
</arr>
<arr name="tweet_terms_urls">
  <str>http://instagr.am/p/JjTNhFtyjq/</str>
</arr>
<arr name="tweet_terms_verbs">
  <str>starts</str>
</arr>
<arr name="tweet_terms_verbs_lemma">
  <str>start</str>
</arr>
<str name="tweet_text">
```

```
RT @thetarro: The main conference #www2012 starts now! @
WWW2012 http://t.co/TL8Bj6zO
</str>
<str name="tweet_text_reduced">
  The main conference starts now! @ WWW2012
</str>
<arr name="tweet_text_topics">
  <str>uncategorized</str>
</arr>
<date name="tweet_timeinserted">2012-04-18T10:08:24Z</date>
<date name="tweet_timestamp">2012-04-18T10:06:50Z</date>
<str name="tweet_user_id">14771839</str>
<str name="tweet_user_location">city of london</str>
<arr name="tweet_user_multi_retweets">
  <str>neal_lathia</str>
  <str>@thetarro</str>
</arr>
<str name="tweet_user_name">Neal Lathia</str>
<str name="tweet_user_profile_image_url">
  http://a0.twimg.com/profile_images/935406753/profile_normal.jpg
</str>
<str name="tweet_user_screen_name">neal_lathia</str>
<str name="tweet_user_url">
  http://www.cl.cam.ac.uk/~nk125/
</str>
</doc>
```

Appendix C: Participant Sheet

Participant Form

Participant: _____ Gender: female male

Highest academic degree: _____ in _____

Institution: _____

Age: _____

Discipline:

- Computer Science
- Psychology
- Business Administration
- Social Science
- Education
- Cognitive Science
- Human Computer Interaction
- Technology Enhanced Learning
- Other: _____

Social Media usage (frequency):

	always	often	sometimes	rarely	never
Facebook					
LinkedIn					
XING					
Google+					
Twitter					
Mendeley					
Blog					
delicious					

I understand that audio and video recordings will be made of this session. Recordings will only be used for research purposes and never be handed on to third parties. Anonymised transcripts, however, might be provided to others.

Date: _____

Signature: _____

Figure 2: Participant sheet - The contents of the participant sheet

List of Figures

3.1	A taxonomy of clustering approaches	15
4.1	Examples of map visualizations	22
4.2	Information Landscape	23
4.3	Weighted graph	24
4.4	Streamgraph	25
4.5	Hierarchy representation	26
4.6	InfoSky	27
4.7	Interactive Visualization Example	29
5.1	Exemplary visualization results	32
6.1	System components overview	35
6.2	System overview (data)	36
6.3	Network analysis	40
7.1	System components overview	43
7.2	Components' Interfaces	47
7.3	UI - Query tab	49
7.4	UI - Cluster tab	50
7.5	Cluster options	52
7.6	UI - Result tab	54
7.7	tweet (data) layer	56
7.8	Cluster layer	56
7.9	UI - Settings tab	57

LIST OF FIGURES

7.10	UI - VIEW mode	58
8.1	Results - Weighted Graph	67
8.2	Results - <i>Voronoi multi</i>	69
8.3	Results - <i>Voronoi Zemanta</i>	70
8.4	Examples of tweets from the dataset	74
8.5	Selection of terms from <i>Weighted Graph</i>	76
8.6	Identified Session with <i>Voronoi multi</i>	76
8.7	Identified Session with <i>Voronoi multi</i>	78
8.8	General Terms with <i>Voronoi Zem</i>	78
8.9	Identified Sessions with <i>Voronoi Zem</i>	79
8.10	Organizational topics/events with <i>Voronoi multi</i>	79
8.11	Organizational topics/events with <i>Voronoi Zem</i>	80
8.12	Keynote in <i>Voronoi multi</i>	82
8.13	Keynote in <i>Voronoi Zem</i>	83
9.1	Improvement Proposal	91
2	Participant sheet	100

List of Tables

6.1	Jit Visualization styles and their features - a comparison	39
8.1	Orders of the visualizations for each participant	71
8.2	The test dataset for tweets with conference hash-tag #www2012 and the number of tweets with URLs, and re-tweets respectively .	73

References

Paul André, Michael Bernstein, and Kurt Luther. Who gives a tweet?: evaluating microblog content value. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work, CSCW '12*, pages 471–474, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1086-4. doi: 10.1145/2145204.2145277. URL <http://doi.acm.org/10.1145/2145204.2145277>. (Cited on page 8.)

Keith Andrews. Information visualisation, 2012. URL <http://courses.iicm.tugraz.at/ivis/ivis.pdf>. Lecture notes on Information Visualisation, Version of 09 Mar 2012, retrieved on Oct 12, 2012. (Cited on page 21.)

Keith Andrews, Wolfgang Kienreich, Vedran Sabol, Jutta Becker, Georg Droschl, Frank Kappe, Michael Granitzer, Peter Auer, and Klaus Tochtermann. The infosky visual explorer: exploiting hierarchical structure and document similarities. *Information Visualization*, 1(3/4):166–181, December 2002. ISSN 1473-8716. doi: 10.1057/palgrave.ivs.9500023. URL <http://dx.doi.org/10.1057/palgrave.ivs.9500023>. (Cited on page 28.)

David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, SODA '07*, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics. ISBN 978-0-898716-24-5. URL <http://dl.acm.org/citation.cfm?id=1283383.1283494>. (Cited on page 18.)

Geoffrey H. Ball and David J. Hall. A clustering technique for summarizing multi-

REFERENCES

- variate data. *Behavioral Science*, 12(2):153–155, 1967. ISSN 1099-1743. doi: 10.1002/bs.3830120210. URL <http://dx.doi.org/10.1002/bs.3830120210>. (Cited on page 17.)
- Fabrício Benevenuto, Gabriel Magno, Tiago Rodrigues, and Virgílio Almeida. Detecting spammers on twitter. In *Proceedings of the 7th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, 2010. (Cited on page 9.)
- Jürgen Bortz and Nicola Döring. *Forschungsmethoden und Evaluation: für Human- und Sozialwissenschaftler*. Springer Medizin Verlag Heidelberg, 3rd edition, 2001. ISBN 3-540-41940-3. (Cited on page 66.)
- Dominique Brodbeck, Riccardo Mazza, and Denis Lalanne. Interactive visualization - a survey. In *Human Machine Interaction*, LNCS 5440, pages 27–46, Berlin/Heidelberg, 2009. Springer-Verlag. URL <http://diuf.unifr.ch/people/lalanned/Articles/visualization-v05.pdf>. (Cited on page 28.)
- Chris Buckley. The importance of proper weighting methods. In *Proceedings of the workshop on Human Language Technology, HLT '93*, pages 349–352, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics. ISBN 1-55860-324-7. doi: 10.3115/1075671.1075753. URL <http://dx.doi.org/10.3115/1075671.1075753>. (Cited on page 14.)
- Lee Byron and Martin Wattenberg. Stacked graphs & geometry & aesthetics. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1245–1252, November 2008. ISSN 1077-2626. doi: 10.1109/TVCG.2008.166. URL <http://dx.doi.org/10.1109/TVCG.2008.166>. (Cited on page 24.)
- Mark A. Cameron, Robert Power, Bella Robinson, and Jie Yin. Emergency situation awareness from twitter for crisis management. In *Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion*, pages 695–698, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. doi:

REFERENCES

10.1145/2187980.2188183. URL <http://doi.acm.org/10.1145/2187980.2188183>.

(Cited on page 10.)

Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web, WWW '11*, pages 675–684, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963500. URL <http://doi.acm.org/10.1145/1963405.1963500>. (Cited on page 9.)

Colin Cherry, Alan Ritter, and Bill Dolan. Unsupervised modeling of twitter conversations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, number June in HLT '10, pages 172–180. Association for Computational Linguistics, 2010. URL <http://dl.acm.org/citation.cfm?id=1857999.1858019>. (Cited on page 3.)

M. A. Dalal and N. D. Harale. A survey on clustering in data mining. In *Proceedings of the International Conference & Workshop on Emerging Trends in Technology, ICWET '11*, pages 559–562, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0449-8. doi: 10.1145/1980022.1980143. URL <http://doi.acm.org/10.1145/1980022.1980143>. (Cited on page 15.)

Gianmarco De Francisci Morales, Aristides Gionis, and Claudio Lucchese. From chatter to headlines: harnessing the real-time web for personalized news recommendation. In *Proceedings of the fifth ACM international conference on Web search and data mining, WSDM '12*, pages 153–162, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124315. URL <http://doi.acm.org/10.1145/2124295.2124315>. (Cited on page 8.)

Derek John de Solla Price. *Little science, big science*. Columbia University Press, 1963. ISBN 0-231-08562-1. (Cited on page 1.)

Jean-Yves Delort. Hierarchical cluster visualization in web mapping systems. In *Proceedings of the 19th international conference on World wide web, WWW '10*,

REFERENCES

- pages 1241–1244, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772892. URL <http://doi.acm.org/10.1145/1772690.1772892>. (Cited on page 26.)
- Martin Ebner and Wolfgang Reinhardt. Social networking in scientific conferences
Twitter as tool for strengthen a scientific community. In *ECTEL 2009*, pages 1–8, 2009. (Cited on pages 10 and 11.)
- Frédéric Filloux. Lessons from the osama bin laden coverage, 2012. URL <http://www.guardian.co.uk/technology/2011/may/09/lessons-from-bin-laden-coverage>. last accessed: Oct 12, 2012. (Cited on page 8.)
- Tim Finin and Belle Tseng. Why We Twitter : Understanding Microblogging. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007. ISBN 1595934448. doi: <http://doi.acm.org/10.1145/1348549.1348556>. URL <http://doi.acm.org/10.1145/1348549.1348556>. (Cited on page 5.)
- John F. Ganth, Christopher Chute, Alex Manfrediz, Stephen Minton, David Reinsel, Wolfgang Schlichting, and Anna Toncheva. The diverse and exploding digital universe. Technical report, IDC White Paper - sponsored by EMC, March 2008. URL <http://www.emc.com/collateral/analyst-reports/diverse-exploding-digital-universe.pdf>. last accessed: May 27, 2012. (Cited on page 1.)
- Scott Golder, Gilad Lotan, and Danah Boyd. Tweet , Tweet , Retweet : Conversational Aspects of Retweeting on Twitter. *Sciences-New York*, pages 1–10, 2010. URL <http://www.danah.org/papers/TweetTweetRetweet.pdf>. (Cited on pages 5 and 6.)
- Michael Granitzer, Wolfgang Kienreich, Vedran Sabol, Keith Andrews, and Werner Klieber. Evaluating a system for interactive exploration of large, hierarchically

REFERENCES

- structured document repositories. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 127–134, Washington, DC, USA, 2004. IEEE Computer Society. ISBN 0-7803-8779-3. doi: 10.1109/INFOVIS.2004.19. URL <http://dx.doi.org/10.1109/INFOVIS.2004.19>. (Cited on pages 26, 27, 28 and 46.)
- Marco Guidi, Igor Ruiz-agundez, and Izaskun Canga-sanchez. Knowledge Mining from the Twitter Social Network . The case of Barack Obama. In Ajith Abraham, editor, *Computational Social Networks: Mining and Visualization*, pages 211–229. Springer, 2012. URL <http://paginaspersonales.deusto.es/igor.ira/publications.html>. (Cited on pages 5 and 19.)
- Aditi Gupta and Ponnurangam Kumaraguru. Credibility ranking of tweets during high impact events. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, PSOSM '12, pages 2:2–2:8, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1236-3. doi: 10.1145/2185354.2185356. URL <http://doi.acm.org/10.1145/2185354.2185356>. (Cited on page 9.)
- Marti A. Hearst. Information visualization: Principles, promise, and pragmatics, 2003. URL http://www.cs.princeton.edu/courses/archive/spr06/cos323/notes/lecture12_vis/cos323_s06_lecture12_vis.ppt. Handouts of the tutorial at CHI 2003 Conference on Human Factors in Computing Systems. (Cited on page 20.)
- Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Commun. ACM*, 55(4):45–54, April 2012. ISSN 0001-0782. doi: 10.1145/2133806.2133821. URL <http://doi.acm.org/10.1145/2133806.2133821>. (Cited on page 21.)
- Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59, June 2010. ISSN 00010782. URL http://portal.acm.org/ft_gateway.cfm?id=1743567&type=html. (Cited on pages 20, 21, 23, 24 and 26.)

REFERENCES

Thomas Heverin. Microblogging for distributed surveillance in response to violent crises: ethical considerations. In *Proceedings of the 2011 iConference*, iConference '11, pages 827–828, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0121-3. doi: 10.1145/1940761.1940924. URL <http://doi.acm.org/10.1145/1940761.1940924>. (Cited on page 10.)

Mengdie Hu, Shixia Liu, Furu Wei, Yingcai Wu, John Stasko, and Kwan-Liu Ma. Breaking news on twitter. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '12, pages 2751–2754, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4. doi: 10.1145/2207676.2208672. URL <http://doi.acm.org/10.1145/2207676.2208672>. (Cited on page 8.)

A.L. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009. (Cited on page 8.)

A. K. Jain, M. N. Murty, and P. J. Flynn. Data Clustering: A Review. *ACM Comput. Surv.*, 31(3):264–323, 1999. doi: <http://doi.acm.org/10.1145/331499.331504>. URL <http://doi.acm.org/10.1145/331499.331504>. (Cited on pages 13, 14, 15, 16 and 17.)

Anil K. Jain. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8):651–666, June 2010. ISSN 01678655. doi: 10.1016/j.patrec.2009.09.011. URL <http://dx.doi.org/10.1016/j.patrec.2009.09.011>. (Cited on pages 16 and 17.)

Anil K. Jain and Richard C. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1988. ISBN 0-13-022278-X. (Cited on page 15.)

Akshay Java, Xiaodan Song, Tim Finin, and Belle 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*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA, 2007. ACM. ISBN

REFERENCES

- 978-1-59593-848-0. doi: 10.1145/1348549.1348556. URL <http://doi.acm.org/10.1145/1348549.1348556>. (Cited on page 7.)
- Jeon Hyung Kang, Kristina Lerman, and Anon Plangprasopchok. Analyzing Microblogs with Affinity Propagation. In *Proceedings of the First Workshop on Social Media Analytics*, pages 67–70, 2010. ISBN 9781450302173. doi: <http://doi.acm.org/10.1145/1964858.1964868>. URL <http://doi.acm.org/10.1145/1964858.1964868>. (Cited on page 87.)
- Leonard Kaufman and Peter J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 9th edition, March 1990. ISBN 0471878766. URL <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471878766>. (Cited on page 17.)
- Daniel A. Keim, Florian Mansmann, Jorn Schneidewind, and Hartmut Ziegler. Challenges in visual data analysis. In *Proceedings of the conference on Information Visualization, IV '06*, pages 9–16, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2602-0. doi: 10.1109/IV.2006.31. URL <http://dx.doi.org/10.1109/IV.2006.31>. (Cited on pages 20 and 28.)
- Daniel A. Keim, Joern Kohlhammer, Geoffrey Ellis, and Florian Mansmann. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, November 2010. URL <http://www.vismaster.eu/book/>. (Cited on pages 21 and 28.)
- Wolfgang Kienreich, Ralph Wozelka, Vedran Sabol, and Christin Seifert. Graph visualization using hierarchical edge routing and bundling. In *Proceedings of the 3rd international Eurovis workshop on visual analytics (EuroVA 2012)*, pages 97–101, 2012. doi: 10.2312/PE/EuroVAST/EuroVA12/097-101. (Cited on page 46.)
- Werner Klieber, Vedran Sabol, Markus Muhr, Roman Kern, Georg Öttl, and Michael Granitzer. Knowledge discovery using the knowminer framework. *IADIS*, Single: 307–314, 2009. (Cited on page 45.)

REFERENCES

- Peter Kraker, Claudia Wagner, Fleur Jeanquartier, and Stefanie Lindstaedt. On the Way to a Science Intelligence: Visualizing TEL Tweets for Trend Detection. In *Proceedings of the 6th European Conference on Technology Enhanced Learning*, page in press, 2011. URL http://know-center.tugraz.at/download_extern/papers/science_intelligence.pdf. (Cited on pages 2, 11, 31, 33, 36, 37, 39 and 66.)
- Peter Kraker, Kris Jack, Christian Schögl, Christoph Trattner, and Stefanie Lindstaedt. Head start: Improving academic literature search with overview visualizations based on readership statistics. In *Web Science*, New York, NY, USA, 2013. ACM. (Cited on page 90.)
- Anagha Kulkarni and Ted Pedersen. Name discrimination and email clustering using unsupervised clustering and labeling of similar contexts. In *IICAI*, pages 703–722, 2005a. (Cited on page 14.)
- Anagha Kulkarni and Ted Pedersen. Senseclusters: Unsupervised clustering and labeling of similar contexts. In *ACL*, 2005b. URL <http://acl.ldc.upenn.edu/P/P05/P05-3027.pdf>. (Cited on page 14.)
- Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter , a Social Network or a News Media ? Categories and Subject Descriptors. In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM, 2010. ISBN 9781605587998. doi: <http://doi.acm.org/10.1145/1772690.1772751>. URL <http://doi.acm.org/10.1145/1772690.1772751>. (Cited on page 8.)
- Julie Letierce, Alexandre Passant, John Breslin, and Stefan Decker. Understanding how Twitter is used to widely spread Scientific Messages. In *Proceedings of the WebSci10: Extending the Frontiers of Society On-Line*, March 2010. (Cited on pages 6, 9 and 10.)
- Nathan N Liu, Clear Water Bay, Hong Kong, and Kai Zhao. Transferring Topical Knowledge from Auxiliary Long Texts for Short Text Clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*,

REFERENCES

- pages 775–784. ACM, 2011. ISBN 9781450307178. doi: <http://doi.acm.org/10.1145/2063576.2063689>. URL <http://doi.acm.org/10.1145/2063576.2063689>. (Cited on page 19.)
- J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967. (Cited on page 16.)
- Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2009. ISBN 0521865719, 9780521865715. URL <http://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf>. (Cited on pages 13, 14 and 15.)
- Michael Meuser and Ulrike Nagel. ExpertInneninterviews - vielfach erprobt, wenig bedacht. Ein Beitrag zur qualitativen Methodendiskussion. In Alexander Bogner, Beate Littig, and Wolfgang Menz, editors, *Das Experteninterview: Theorie, Methode, Anwendung*, volume 2, pages 71–94. VS Verlag, Wiesbaden, 2005. (Cited on pages 61 and 62.)
- George Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, 63:81–97, 1956. URL <http://www.musanim.com/miller1956/>. (Cited on page 20.)
- Teng-Sheng Moh and Surya Bhagvat. Clustering of technology tweets and the impact of stop words on clusters. In *Proceedings of the 50th Annual Southeast Regional Conference*, ACM-SE '12, pages 226–231, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1203-5. doi: 10.1145/2184512.2184566. URL <http://doi.acm.org/10.1145/2184512.2184566>. (Cited on page 13.)
- Markus Muhr, Vedran Sabol, and Michael Granitzer. Scalable recursive top-down hierarchical clustering approach with implicit model selection for textual data sets. In *Proceedings of the 2010 Workshops on Database and Expert Systems Applications*,

REFERENCES

- DEXA '10, pages 15–19, Washington, DC, USA, 2010. IEEE Computer Society. ISBN 978-0-7695-4174-7. doi: 10.1109/DEXA.2010.25. URL <http://dx.doi.org/10.1109/DEXA.2010.25>. (Cited on pages 18 and 46.)
- N Nanas, V. Uren, A. De Roeck, and J. Domingue. A comparative study of term weighting methods for information filtering. Technical report, Knowledge Media Institute, The Open University, 2003. <http://kmi.open.ac.uk/publications/papers/kmi-tr-128.pdf>. (Cited on page 14.)
- Jeffrey Nichols, Jalal Mahmud, and Clemens Drews. Summarizing sporting events using twitter. In *Proceedings of the 2012 ACM international conference on Intelligent User Interfaces*, IUI '12, pages 189–198, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1048-2. doi: 10.1145/2166966.2166999. URL <http://doi.acm.org/10.1145/2166966.2166999>. (Cited on page 8.)
- Kyosuke Nishida, Ryohei Banno, Ko Fujimura, and Takahide Hoshide. Tweet Classification by Data Compression. In *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, pages 29–34. ACM, 2011. ISBN 9781450309622. doi: <http://doi.acm.org/10.1145/2064448.2064473>. URL <http://doi.acm.org/10.1145/2064448.2064473>. (Cited on pages 6 and 18.)
- Atsuyuki Okabe, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*. Probability and Statistics. Wiley, Chichester, UK, second edition, 2000. (Cited on page 26.)
- Fernando Perez-Tellez, David Pinto, John Cardiff, and Paolo Rosso. On the Difficulty of Clustering Company Tweets. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 95–102. ACM, 2010. ISBN 9781450303866. doi: <http://doi.acm.org/10.1145/1871985.1872001>. URL <http://doi.acm.org/10.1145/1871985.1872001>. (Cited on pages 3 and 19.)
- Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections. In

REFERENCES

- Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM, 2008. ISBN 9781605580852. doi: <http://doi.acm.org/10.1145/1367497.1367510>. URL <http://doi.acm.org/10.1145/1367497.1367510>. (Cited on pages 3 and 19.)
- Roberto Pinho, Maria Cristina Ferreira de Oliveira, Rosane Minghim, and Marinho G. Andrade. Voromap: A voronoi-based tool for visual exploration of multi-dimensional data. In *Proceedings of the conference on Information Visualization, IV '06*, pages 39–44, Washington, DC, USA, 2006. IEEE Computer Society. ISBN 0-7695-2602-0. doi: 10.1109/IV.2006.131. URL <http://dx.doi.org/10.1109/IV.2006.131>. (Cited on page 26.)
- Aniket Rangrej, Sayali Kulkarni, and Ashish V. Tendulkar. Comparative study of clustering techniques for short text documents. In *Proceedings of the 20th international conference companion on World wide web, WWW '11*, pages 111–112, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963249. URL <http://doi.acm.org/10.1145/1963192.1963249>. (Cited on page 87.)
- Wolfgang Reinhardt, Martin Ebner, Günter Beham, and Cristina Costa. How People are using Twitter during Conferences. 2009. URL <http://apo.org.au/node/14601>. (Cited on pages 1 and 10.)
- Daniela Retelny, Jeremy Birnholtz, and Jeffrey Hancock. Tweeting for class: using social media to enable student co-construction of lectures. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion, CSCW '12*, pages 203–206, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1051-2. doi: 10.1145/2141512.2141578. URL <http://doi.acm.org/10.1145/2141512.2141578>. (Cited on page 8.)
- Kevin Dela Rosa, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. Topical Clustering of Tweets. In *Proceedings of the ACM SIGIR Special Interest Group on Information Retrieval's 3rd Workshop on Social Web Search and Mining*

REFERENCES

- (*SIGIR: SWSM 2011*), 2011. URL <http://www.cs.cmu.edu/~kdelaros/>. (Cited on page 18.)
- Matt Rosoff. Twitter just had its cnn moment, 2012. URL <http://www.businessinsider.com/twitter-just-had-its-cnn-moment-2011-5>. last accessed: Oct 12, 2012. (Cited on page 8.)
- Matthew Rowe and Milan Stankovic. Mapping tweets to conference talks: A goldmine for semantics. In *Social Data on the Web Workshop, International Semantic Web Conference*, 2010. URL <http://ceur-ws.org/Vol-664/paper4.pdf>. (Cited on page 11.)
- Matthew Rowe and Milan Stankovic. Aligning tweets with events: Automation via semantics. *Semantic Web Journal*, 2011. URL <http://www.semantic-web-journal.net/content/aligning-tweets-events-automation-semantics>. (Cited on page 11.)
- Vedran Sabol, Kamran Ali Ahmad Syed, Arno Scharl, Markus Muhr, and Alexander Hubmann-Haidvogel. Incremental computation of information landscapes for dynamic web interfaces. In *Proceedings of the IX Symposium on Human Factors in Computing Systems, IHC '10*, pages 205–208, Porto Alegre, Brazil, Brazil, 2010. Brazilian Computer Society. URL <http://dl.acm.org/citation.cfm?id=1999593.1999619>. (Cited on page 23.)
- Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web, WWW '06*, pages 377–386, New York, NY, USA, 2006. ACM. ISBN 1-59593-323-9. doi: 10.1145/1135777.1135834. URL <http://doi.acm.org/10.1145/1135777.1135834>. (Cited on page 19.)
- Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994. (Cited on page 36.)

REFERENCES

- Ben Schneiderman. *Designing the User Interface*. Addison-Wesley, 1998. ISBN 0321537351. (Cited on page 28.)
- Ramine Tinati, Leslie Carr, Wendy Hall, and Jonny Bentwood. Identifying communicator roles in twitter. In *Proceedings of the 21st international conference companion on World Wide Web, WWW '12 Companion*, pages 1161–1168, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1230-1. doi: 10.1145/2187980.2188256. URL <http://doi.acm.org/10.1145/2187980.2188256>. (Cited on page 7.)
- Colin Ware. *Information Visualization – Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2000. ISBN 1558605118. URL <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1558605118>. (Cited on page 20.)
- Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2nd edition, 2004. ISBN 1558608192. URL <http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/1558608192>. (Cited on page 20.)
- Katrin Weller, Evelyn Dröge, and Cornelius Puschmann. Citation analysis in twitter: Approaches for defining and measuring information flows within tweets during scientific conferences. In Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, and Mariann Hardey, editors, *Making Sense of Microposts (#MSM2011)*, pages 1–12, May 2011. URL http://ceur-ws.org/Vol-718/paper_04.pdf. (Cited on pages 9, 52 and 72.)
- F. Dianne Lux Wigand. Twitter takes wing in government: diffusion, roles, and management. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, dg.o '10, pages 66–71. Digital Government Society of North America, 2010. ISBN 978-1-4503-0070-4. URL <http://dl.acm.org/citation.cfm?id=1809874.1809889>. (Cited on page 8.)

REFERENCES

- Fridolin Wild, Chris Valentine, and Peter Scott. Shifting interests: changes in the lexical semantics of ed-media. *International Journal on E-Learning*, 9(4):549–562, 2010. URL <http://oro.open.ac.uk/25146/>. (Cited on page 2.)
- Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005. ISBN 0120884070. (Cited on pages 16 and 17.)
- WWW. History of the www, 2012. URL <http://www2012.wwwconference.org/about/history/>. last accessed December 24, 2012. (Cited on page 63.)
- Xin Xia, Xiaohu Yang, Chao Wu, Shanping Li, and Linfeng Bao. Information credibility on twitter in emergency situation. In *Proceedings of the 2012 Pacific Asia conference on Intelligence and Security Informatics, PAISI'12*, pages 45–59, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 978-3-642-30427-9. doi: 10.1007/978-3-642-30428-6_4. URL http://dx.doi.org/10.1007/978-3-642-30428-6_4. (Cited on page 9.)
- Rui Xu and Donald Wunsch. Survey of clustering algorithms. *IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council*, 16(3): 645–78, May 2005. ISSN 1045-9227. doi: 10.1109/TNN.2005.845141. URL <http://www.ncbi.nlm.nih.gov/pubmed/18252358>. (Cited on pages 14, 15 and 17.)
- Lei Yang, Tao Sun, Ming Zhang, and Qiaozhu Mei. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web, WWW '12*, pages 261–270, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1229-5. doi: 10.1145/2187836.2187872. URL <http://doi.acm.org/10.1145/2187836.2187872>. (Cited on page 6.)