Institut für Visualisierung und Interaktive Systeme
Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Diplomarbeit Nr. 3366

Extraktion und Visualisierung von Kommunikationsnetzwerken in Twitter

Johannes Dilli

Studiengang:	Softwaretechnik
Prüfer:	Prof. Dr. Thomas Ertl
Betreuer:	Dipl.-Inf. Dennis Thom Steffen Lohmann, M. Sc.
begonnen am:	13. Juli 2012
beendet am:	11. Januar 2013
CR-Klassifikation:	H.3.3, H.5.2

Kurzfassung

Soziale Netzwerke, wie Twitter und Facebook, spielen eine immer größere Rolle im alltäglichen Leben. Dies wirkt sich auch auf das Kommunikationsverhalten der Nutzer aus. Das soziale Netzwerk Twitter konzentriert sich hauptsächlich auf das Schreiben und den Austausch von Kurznachrichten über das Internet. Hierbei kann Twitter sowohl als Broadcast-Medium genutzt werden, um Informationen an ein möglichst breites Publikum zu verbreiten als auch zur direkten Kommunikation zwischen Nutzern. Durch die Kommunikation zwischen Nutzern entstehen in Twitter Kommunikationsnetzwerke.

In dieser Diplomarbeit werden Verfahren entwickelt, die es ermöglichen Kommunikationsnetzwerke in Twitter zu finden und zu extrahieren. Des Weiteren wird eine Visualisierung der gefundene Daten entwickelt. Dabei werden sowohl das Netzwerk als auch die behandelten Themengebiete und Inhalte dargestellt. Durch die Visualisierung und geeignete Filtermöglichkeiten werden Anwender bei der Analyse der Daten unterstützt.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Verzeichnis der Algorithmen

1 Einleitung

Noch vor wenigen Jahren wurde das Internet weitgehend zum Lesen von Websites, zur Suche von Informationen und zum Versenden von E-Mails verwendet. Es diente hauptsächlich als passives Medium zur Informationsgewinnung. Die aktive Teilnahme, z. B. an Forendiskussionen, bildete die Ausnahme. Heutzutage bietet es den Nutzern verschiedenste Möglichkeiten, um selbst aktiv zu werden und Inhalte erstellen zu können. Insbesondere Soziale Netzwerke (auch Online-Communities genannt) spielen eine immer größere Rolle. Soziale Netzwerke bieten ihren Nutzern verschiedene Möglichkeiten, um mit anderen Nutzern zu interagieren. Oftmals können Mitglieder Inhalte mit anderen teilen und sich mit anderen Nutzern verknüpfen (Freundschaften schließen), so dass deren Inhalte gesehen werden können. Zusätzlich können für gewöhnlich auch Nachrichten mit anderen Nutzern ausgetauscht werden.

> Online-Communities zählen bei Jugendlichen zu den drei am häufigsten ausgeübten Anwendungen im Internet und werden von insgesamt 87 Prozent der Internet-Nutzer zumindest selten genutzt.[...] 79 Prozent der Internet-Nutzer loggen sich mindestens mehrmals pro Woche auf den Seiten eines Sozialen Netzwerks ein. [Sü12]

Die Nutzerzahlen Sozialer Netzwerke steigen in den letzten Jahren beständig. Mit steigender Nutzerzahl steigt auch die Anzahl der möglichen Kommunikationspartner. Zusätzlich steigt auch die Zeit, die in Sozialen Netzwerken verbracht wird. Beides führt zu einer Zunahme der Kommunikation zwischen Nutzern über diese Plattformen.

Auf vielen Plattformen, wie Facebook[1], studiVZ[2] oder auch Google+[3] ist diese Kommunikation nicht oder nur teilweise öffentlich. Häufig können nur Freunde sehen was ein Nutzer schreibt, so dass es nicht einfach möglich ist auf diesen Plattformen Daten über die Kommunikation der Mitglieder zu sammeln.

Im Gegensatz dazu bietet das Soziale Netzwerk Twitter[4] eine weitgehend öffentliche Kommunikationsplattform, so dass im Normalfall die Inhalte von jedem eingesehen werden können. Auf Twitter können sich sowohl normale Nutzer, Prominente als auch Unternehmen untereinander austauschen und Inhalte mit anderen teilen. Neben dem Teilen von Inhalten ist es auch möglich direkt mit anderen Nutzern der Plattform in Interaktion zu treten und

[1] http://www.facebook.com
[2] http://www.studivz.net
[3] http://plus.google.com
[4] http://www.twitter.com

sich mit diesen auszutauschen. Durch den Austausch zwischen Twitter-Nutzern bilden sich Kommunikationsnetzwerke innerhalb von Twitter.

1.1 Aufgabenstellung

Im Rahmen dieser Diplomarbeit sollen Kommunikationsnetzwerke in Twitter untersucht werden. Hierbei sollen nicht Netzwerke auf der Basis der „Follower"-beziehungen auf Twitter betrachtet werden, sondern hauptsächlich Kommunikation zwischen Twitter-Nutzern als Datengrundlage herangezogen werden. Um an diese Daten aus Twitter zu erhalten, sollen Verfahren und die passenden Werkzeuge hierfür erstellt werden. Diese sollen es ermöglichen gezielt nach Kommunikationsnetzwerke in Twitter zu suchen und diese zu extrahieren. Nach der Gewinnung der Daten sollen diese aufbereitet und visualisiert werden. Ein Anwender der Software soll durch geeignete Visualisierungen und Filtermöglichkeiten unterstützen werden diese Kommunikationsnetzwerke zu analysieren. Hierbei soll sowohl die Kommunikationsstruktur als auch Kommunikationsinhalte und besprochene Themengebiete dargestellt werden. Zusätzlich zur Entwicklung der Werkzeuge wird auch versucht Erkenntnisse über das Kommunikationsverhalten auf Twitter zu erlangen.

Neben der Extraktion und der Visualisierung der Daten soll auch versucht werden Erkenntnisse über das Kommunikationsverhalten in Twitter zu erlangen: Wie kommunizieren Nutzer untereinander und wie bilden sich Netzwerke bilden.

1.2 Aufbau der Arbeit

Um einen Überblick über die Arbeit und die Vorgehensweise zu geben, wird im Folgenden kurz auf deren Gliederung sowie auf den Zweck und Inhalt der einzelnen Kapitel eingegangen:

Kapitel 1 – Einleitung: Das erste Kapitel führt kurz in das Thema der Arbeit ein, zeigt die Problemstellung auf und bietet einen Überblick über den Aufbau des Dokuments.

Kapitel 2 – Grundlagen: Dieses Kapitel bietet einen Überblick über verschiedene Themen, auf die in dieser Arbeit Bezug genommen wird. Diese dienen als Grundlage für die weitere Arbeit und werden zum Verständnis benötigt. Darüber hinaus werden Arbeiten mit ähnlichen Problemstellung oder die als Grundlage für diese Arbeit dienen, vorgestellt.

Kapitel 3 – Konzept: Ausgehend von den Grundlagen und verwandten Arbeiten wird in diesem Kapitel ein Konzept zum Sammeln und Visualisieren von Twitter-Daten erstellt. Zuerst wird darauf eingegangen, was Kommunikation auf Twitter bedeutet und wie es möglich ist, Kommunikationsnetzwerke zu finden und zu extrahieren. Anschließend wird ein Konzept entwickelt, um die gesammelten Daten zu visualisieren und den

Benutzer dabei zu unterstützen die Daten zu verstehen. Dieses wird anhand von Skizzen vorgestellt.

Kapitel 4 – Implementierung: In diesem Kapitel wird die Implementierung des zuvor entwickelten Konzepts beschrieben. Es werden mehrere Methoden zum Sammeln von Daten auf Twitter vorgestellt und deren Umsetzung sowie deren Vor- und Nachteile beschrieben. Im Weiteren wird auf die verschiedenen Aspekte bei der Implementierung der Visualisierung der gesammelten Daten eingegangen.

Kapitel 5 – Anwendungsfälle: Um die Anwendbarkeit der entwickelten Werkzeuge zu zeigen, werden in diesem Kapitel Anwendungsfälle beschrieben und erklärt, welche Schlussfolgerungen man aus den einzelnen Szenarien ziehen kann.

Kapitel 6 – Zusammenfassung und Ausblick: Dieses Kapitel fasst die Ergebnisse der Arbeit zusammen: Es wird beschrieben, welche Schlussfolgerungen man für das Kommunikationsmedium Twitter ziehen kann. Zusätzlich bietet dieses Kapitel einen Ausblick auf eine mögliche Weiterentwicklung des Ansatzes.

2 Grundlagen

In diesem Kapitel werden die Plattform Twitter und deren Schnittstellen, weitere grundlegenden Themen, sowie verwandte Arbeiten vorgestellt.

2.1 Twitter

Twitter ist ein Soziales Netzwerk, das den Schwerpunkt auf das Schreiben von kurzen Nachrichten in Echtzeit setzt. Twitter kann als eine Art eigener Blog verwendet werden indem nur Nachrichten veröffentlicht werden oder als Kommunikationsplattform und Soziales Netzwerk indem Nachrichten an andere Nutzer geschrieben werden. Zusätzlich zu den öffentlichen Nachrichten können auch private Nachrichten zwischen zwei Nutzern ausgetauscht werden. Die meisten Nutzer verwenden alle drei Arten der Kommunikation.

Twitter wurde 2006 gegründet und hat inzwischen mehr als 500 Millionen aktive Nutzer. [Dug12]

In den folgenden Abschnitten werden kurz die wichtigsten Begrifflichkeiten erklärt, die benötigt werden, um sich auf Twitter zurechtzufinden (vgl. [twic]). Des Weiteren wird der Zugriff auf Twitter-Daten beschrieben.

2.1.1 Tweet

Nachrichten auf Twitter werden als Tweet bezeichnet. Ein Tweet kann maximal 140 Zeichen lang sein. Tweets sind im Normalfall öffentlich einsehbar. Ein Nutzer kann aber die Sichtbarkeit all seiner Tweets auch auf seine Follower (s. Unterabschnitt 2.1.2) beschränken.

Um anderen Twitter-Nutzern schreiben zu können oder Stichworte hervorzuheben, gibt es auf Twitter eine eigene Syntax:

@-Notation: Um eine Nachricht an einen oder mehrere Nutzer zu adressieren, verwendet man die @-Notation, indem man an den Anfang der Nachricht @+Nutzername schreibt. Ein Nachricht an den Nutzer @example könnte somit lauten:

> @example wie geht es dir?

Ist ein Tweet an mehrere Nutzer gerichtet, so werden diese am Anfang der Nachricht aufgezählt:

> @example1 @example2 wie geht es euch?

Wird eine Nachricht nicht direkt an einen anderen Nutzer adressiert, sondern nur über diesen geschrieben, so kann die @-Notation auch mitten im Text verwendet werden:

> Jetzt gehe ich mit @example1 und @example2 zum Fußballspiel!

Hashtags: Wichtige Begriffe in einer Nachricht können durch Hashtags hervorgehoben werden. Hashtags beginnen immer mit einem #. Beispiel:

> Heute #backen wir einen #Kuchen.

Durch Hashtags wird die Suche nach Tweets mit bestimmten Themen auf Twitter erleichtert. Gleichzeitig bietet Twitter eine Übersicht der beliebtesten Hashtags an, so dass man sich schnell einen Überblick über wichtige Themen verschaffen und diese Tweets gezielt betrachten kann.

An einen Tweet kann zusätzlich noch der aktuelle Ort angehängt werden, so dass für jeden sichtbar ist von wo die Nachricht versendet wurde.

2.1.2 Follower

Interessiert man sich für Nachrichten eines anderen Nutzers, so kann man dessen Inhalte abonnieren. Dieser Vorgang nennt sich Folgen, man gehört damit zu den Followern dieses Nutzers. Interessiert man sich nicht mehr für die Inhalte eines Nutzers, so kann man diesem auch wieder entfolgen (engl. unfollow). Im Gegensatz zu manchen anderen Sozialen Netzwerken ist die Follow-/Folgen-Beziehung nicht symmetrisch, so dass jeder Twitter-Nutzer eine Liste an Nutzern hat, denen er folgt und ebenso eine Liste an Nutzern, die ihm folgen.

2.1.3 Retweet

Interessante oder wichtige Tweets kann man an seine eigenen Follower verbreiten indem man diese wiederholt. Dies wird als Retweeten bezeichnet. Bei einem Retweet ist der ursprüngliche Autor der Nachricht weiterhin erkennbar. Die Anzahl der Retweets ist ein Zeichen für die Wichtigkeit einer Nachricht und wie weit sich diese verbreitet hat.

2.1.4 Favoriten

Zusätzlich zum Retweeten einer Nachricht kann man diese zusätzlich als Favorit kennzeichnen. Wie oft eine Nachricht als Favorit markiert wurde, ist neben den Retweets eine weitere Kennzahl, um deren Wichtigkeit zu bewerten. Viele Internetseite, die Statistiken zu Twitter anbieten, wie z. B. Favstar[1], werten die Beliebtheit von Tweets anhand der Favoriten aus.

[1] http://favstar.fm

2.1.5 Timeline

Jeder Nutzer sieht auf seiner Twitter-Startseite die neuesten Tweets aller Nutzer, denen er folgt. Diese Liste an Tweets wird Timeline genannt. Tweets, die direkt an einen anderen Nutzer gesendet wurden, werden in der eigenen Timeline nur angezeigt, wenn man dem Adressat auch folgt. Zusätzlich werden alle Retweets der Nutzer, denen man folgt, in der Timeline angezeigt.

2.1.6 Direct Messages

Direct Messages sind nicht-öffentliche Nachrichten an andere Twitter-Nutzer. Sie können nur an die eigenen Follower versendet werden. Direct Messages tauchen nicht in der Timeline auf.

2.1.7 Zugriff auf Daten

Twitter bietet über verschiedene Schnittstellen Zugriff auf seine Daten. Die Schnittstellen dienen zum Einen zur Verwaltung des eigenen Twitter-Accounts, zum Anderen dem Zugriff auf Tweets. Der Zugriff auf Tweets kann dabei als Suche auf den vorhandenen Tweets geschehen oder es kann auf Live-Daten zugegriffen werden, so dass man Informationen darüber erhält, was in diesem Moment auf Twitter geschieht. Die verschiedenen APIs (application programming interface) unterscheiden sich in der Art in der sie angesprochen werden. Das Antwortformat der APIs ist im Aufbau ähnlich, so dass sich die Verarbeitung der Antwort nur minimal unterscheidet. Im folgenden werden die Grundlagen der REST-API ([twia]) und der Streaming-API ([twib]) beschrieben.

REST-API

Die Twitter REST-API erlaubt den Zugriff auf verschiedene vorhanden Daten. Auf manche Daten kann nur als authentifizierter Nutzer zugegriffen werden. Authentifizierte Nutzer können zusätzlich über die REST-API verschiede Daten ihres Accounts ändern.

Timelines: Authentifizierte Nutzer können hierüber die eigene Timeline, die eigenen Tweets, Tweets, in denen der eigene Twitter-Name auftaucht wird und Retweets der eigenen Nachricht abrufen.

Tweets: Über die ID eines Tweets kann hier der Tweet und dessen Retweets abgerufen werden. Authentifizierte Nutzer können zusätzlich Tweets absenden und eigene Tweets löschen sowie andere Tweets retweeten.

Direct Messages: Authentifizierte Nutzer können neue Direct Messages senden und gesendete und erhaltene Direct Messages abrufen oder löschen.

Friends & Followers: Es können die Follower eines Nutzers und wem dieser Nutzer folgt abgerufen werden. Für zwei Nutzer kann der Status der Freundschaft überprüft werden: Entweder folgen sich beide Nutzer gegenseitig, nur ein Nutzer folgt dem andern oder beiden folgen sich nicht. Authentifizierte Nutzer können zusätzlich anderen Nutzern folgen oder entfolgen.

Users: Die Profile von Nutzern können hierüber abgerufen werden. Authentifizierte Nutzer können die Daten des eigenen Profils ändern und andere Nutzer blockieren, so dass deren Tweets nicht mehr in der Timeline auftauchen und dass eine Kommunikation von und zu diesem Nutzer blockiert ist.

Suggested Users: Twitter schlägt interessante Nutzer vor. Diese können hier nach Kategorie oder speziell für einen authentifizierten Nutzer abgefragt werden.

Favorites: Die letzten 20 Tweets eines Nutzers, die von anderen Nutzern als Favoriten markiert wurden, können hier abgerufen werden. Authentifizierte Nutzer können zusätzlich Tweets als Favorit markieren oder diese Markierung wieder entfernen.

Lists: Authentifizierte Nutzer können hiermit Listen mit Twitter-Nutzern anlegen, um so eine bessere Übersicht zu erhalten. Für jede Liste kann eine eigene Timeline abgerufen, Nutzer hinzugefügt oder entfernt und die Liste gelöscht werden.

Search: Die Suche bietet Zugriff auf die Tweets der letzten sechs bis neun Tage. Eine Suchanfrage kann hierbei aus einfachen Wörtern oder komplexeren Suchoperatoren bestehen. Suchbegriffe werden automatisch UND-verknüpft. Eine ODER-Verknüpfung lässt sich durch die Verwendung von „OR“ zwischen den Begriffen erreichen. Soll ein Ausdruck genau wie angegeben vorkommen, so muss dieser in Anführungszeichen gesetzt werden, z. B. "happy hour". Begriffe, die nicht im Ergebnis vorkommen sollen, wird ein „-“ vorgestellt. Soll ein Tweet einen bestimmten Sender oder Empfänger haben, so kann dieser mit „from:“ bzw. „to:“ angegeben werden. Mentions oder Hashtags werden ganz normal mit „@“ bzw. „#“ angegeben. Tweets von bestimmten Orten können mit „place:“ gesucht werden. Soll der Zeitbereich eingeschränkt werden, so können Tweets nach einem Datum mit „since:“ und dem Datum im Format YYYY-MM-DD gesucht werden, z. B. since:2012-10-25. Tweets bis zu einem Datum können entsprechend mit „until:“ gesucht werden. Tweets aus einer bestimmten Quelle, z. B. der Tweet Button einer Website oder ein bestimmter Twitter-Client können mit „source:“ gesucht werden.

Zusätzlich können der Suche noch folgende optionale Parameter angehängt werden:

geocode Hierüber kann die Suche auf einen Radius um die gegebenen geographischen Koordinaten eingeschränkt werden. Die Koordinaten werden im Format „latitude,longitude,radius“ angegeben.

lang Hierüber können die Suchergebnisse auf eine Sprache eingeschränkt werden. Die Sprache wird als ISO 639-1 Code angegeben.

result_type: Gibt an, ob die Suche aktuelle („recent“) oder beliebte („popular“) Tweets enthalten soll oder beides („mixed“).

count Hiermit kann die Anzahl der Ergebnisse pro Seite angegeben werden.

unitl Beschränkt die Suche auf Tweets vor dem angegebenen Datum (im Format YYYY-MM-DD).

since_id Es werden nur Tweets mit einer ID höher als die gegebene im Suchergebnis beachtet.

max_id Hiermit kann die höchste Tweet-ID angegeben werden. Tweets mit einer höheren ID kommen nicht im Ergebnis vor.

Beim Abfragen der Suchergebnisse ist zu beachten, dass die Daten seitenweise zurück geliefert werden. Der Parameter „count“ gibt dabei die Anzahl der Ergebnisse pro Seite an. Standardmäßig werden nur 15 Ergebnisse pro Seite zurückgeliefert. Um möglichst schnell viele Daten zu erhalten, bietet es sich daher an, diesen Wert auf das Maximum von 100 Ergebnissen pro Seite zu setzen.

Saved Searches: Authentifizierte Nutzer können Suchen zur späteren Wiederverwendung abspeichern, die Suchergebnisse abrufen und die Suche löschen.

Places & Geo: Tweets, die eine Ortsangabe angehängt haben, können hiermit gezielt gesucht werden. Außerdem können Orte, die in der Twitter-Datenbank hinterlegt sind sind, abgefragt werden oder neue Orte erstellt werden.

Trends: Twitter berechnet anhand von aktuellen Tweets aller Nutzer beliebte Themen für Orte. Die Orte und die Themen für diese können abgerufen werden.

Spam Reporting: Auch Twitter hat mit Spam zu kämpfen. Über diese Schnittstelle können Nutzer, die Spam verbreiten gemeldet werden.

OAuth: Anwendungen, die auf Twitter zugreifen, müssen sich über die OAuth-Schnittstelle authentifizieren.

Help: Über die Help-Schnittstelle können die aktuelle Konfiguration von Twitter, die unterstützten Sprachen, die Datenschutzrichtlinie, die Nutzungsbedingungen und das aktuellen Rate Limit (Begrenzung der Zugriffshäufigkeit) abgerufen werden.

Streaming-API

Die Streaming API-bietet Zugriff auf den globalen Strom an Tweets auf Twitter. Man erhält Zugriff auf die aktuellen Livedaten.

Public streams: Die öffentlichen Streams bieten Zugriff auf alle öffentlichen Tweets, entweder kann man auf alle Tweets zugreifen; auf eine von Twitter zufällig erstellte Auswahl oder man kann eigene Filter definieren, um die Anzahl an Tweets einzuschränken. Hierbei kann man nach User-IDs, Stichworten oder nach einem Zielgebiet, das durch zwei geographische Koordinaten beschränkt wird, filtern.

User streams: Der User Stream gibt alle aktuellen Daten und Ereignisse für den authentifizierten Nutzer. In der Standardeinstellung werden auch die Tweets der Nutzer, denen der authentifizierten Nutzer folgt, ausgeliefert. Setzt man den Parameter „with“ auf „user“, so werden nur noch die Ereignisse des Nutzers berücksichtigt.

Site streams: Site Streams sind den User Streams ähnlich, jedoch liefern sie Ergebnisse für mehrere Nutzer. Die Anzahl der Nutzer ist anfangs auf 100 beschränkt, kann aber auf bis zu 1000 Nutzer erhöht werden.

2.2 Tag Cloud

Mit Tag Clouds oder auch Schlagwortwolken sind eine beliebte Visualisierung um Schlagworte (Tags) mit unterschiedlicher Wichtigkeit darzustellen. Die Schriftgröße eines Schlagwortes spiegelt hierbei dessen Wichtigkeit wieder. Somit lässt sich mit einem einfachen Blick feststellen, welche Schlagworte wichtig und welche weniger wichtig sind. Um sich leichter innerhalb der Liste orientieren zu können, werden die Schlagworte oftmals alphabetisch sortiert.

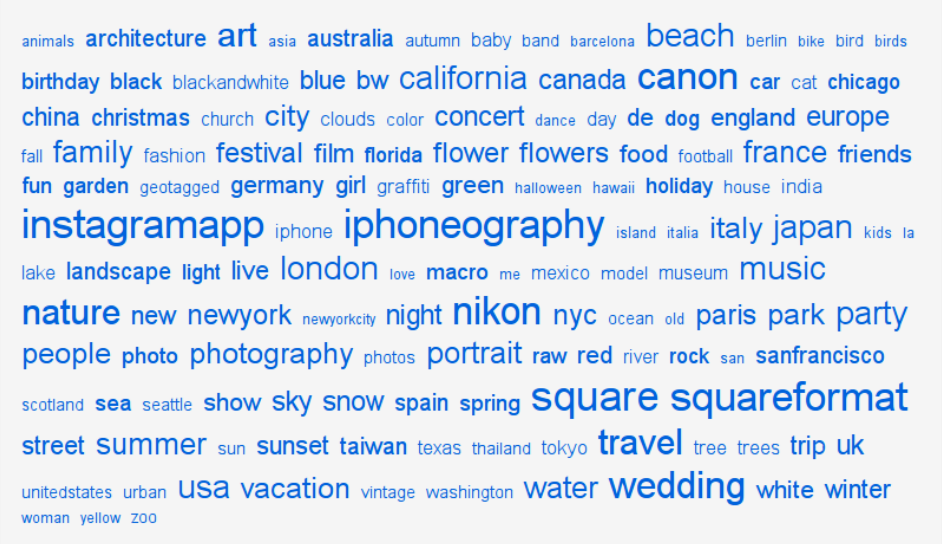


Abbildung 2.1: Beispiel einer Tag Cloud [fli]

Abbildung 2.1 zeigt ein Beispiel einer Tag Cloud der Online-Fotoplattform Flickr, die die häufigsten Schlagworte von Bildern anzeigt. Die Größe der einzelnen Schlagwortes repräsentiert wie oft dieser auf Flickr verwendet wird. „iphoneography“ wird z. B. deutlich öfter verwendet als „zoo“.

2.3 Spline

Splines oder auch Polynomzüge sind Funktionen, die abschnittsweise aus Polynomen zusammen gesetzt sind. Der Begriff stammt aus dem Schiffsbau. Dort wurden glatte Kurven mit biegsamen Latten, so genannten „Straklatten“ (engl. spines), konstruiert.

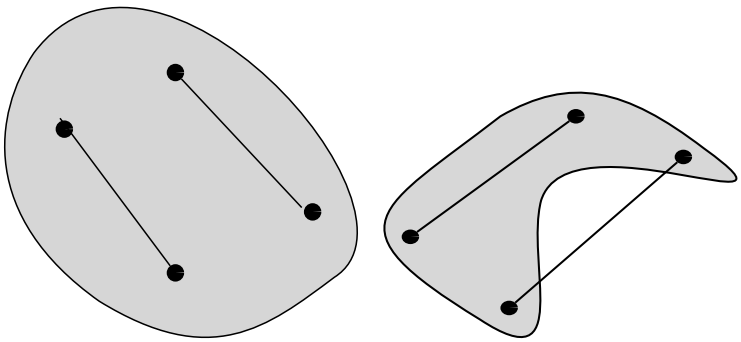
Splines eignen sich zur Interpolation von beliebigen Funktionen anhand von gegebenen Stützstellen.

2.3.1 kubische Splines

Kubische Splines sind glatten Kurven, die aus kubischen Polynomen ($ax^3 + bx^2 + cx + d$) zusammengesetzt sind und durch gegebene Stützstellen verlaufen. Glatte Kurve bedeutet, dass die Teilstücken sowohl am gleichen Punkt aufeinander treffen, die selbe Steigung als auch die selbe Krümmung haben.

2.4 konvexe Hülle

„Die konvexe Hülle einer Teilmenge ist die kleinste konvexe Menge, die die Ausgangsmenge enthält.“ [wik12a] „Eine geometrische Figur oder allgemeiner eine Teilmenge eines euklidischen Raums heißt konvex, wenn für je zwei beliebige Punkte, die zur Menge gehören, auch stets deren Verbindungsstrecke ganz in der Menge liegt.“ [wik12b]

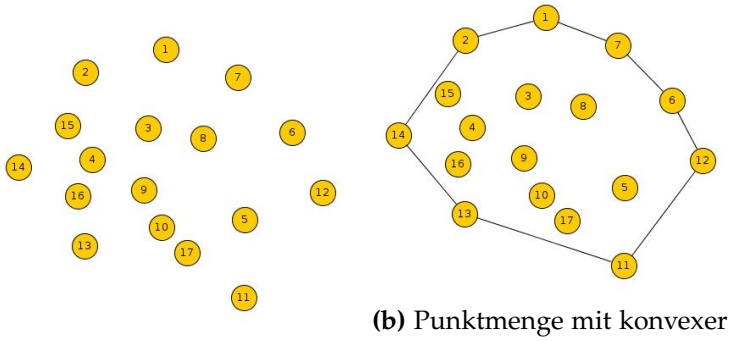


(a) konvexe Menge **(b)** nicht-konvexe Menge

Abbildung 2.2: Konvexe und nicht-konvexe Menge

Abbildung 2.2 zeigt ein Beispiel. Der graue Bereich mit Rand stellt dabei die Menge dar. Anhand der vier Beispielpunkte lässt sich erkennen, dass es sich bei Abbildung 2.2a um eine konvexe Menge handelt, bei Abbildung 2.2b nicht.

Abbildung 2.3 zeigt zur Veranschaulichung eine Menge an Punkten und deren konvexe Hülle.



(a) Punktmenge

(b) Punktmenge mit konvexer Hülle

Abbildung 2.3: konvexe Hülle einer Punktmenge

2.4.1 Graham-Scan-Algorithmus

Um die konvexe Hülle einer endlichen Menge von Punkten in der Ebenen zu berechnen gibt es verschiedenen Ansätze. Im folgenden wird der Graham-Scan-Algorithmus ([Gra72]) vorgestellt.

Gegeben sei eine endliche Punktmenge $S = \{P\}$.

1. Finde den Punkt mit dem kleinsten y-Wert. Gibt es mehrere, wähle den mit dem kleinsten x-Wert. Die Suche kann in $\mathcal{O}(n)$ durchgeführt werden. Der gefundene Punkt bildet den Startpunkt P_0.
2. Sortiere die restlichen Punkte P nach aufsteigendem Winkel, den sie mit dem Punkt P_0 und der x-Achse einschließend. Haben zwei Punkte den gleichen Winkel, so wird der Punkt, der näher an P_0 liegt, entfernt. Die Sortierung kann z. B. mit Quicksort in $\mathcal{O}(n \log n)$ durchgeführt werden.
3. Füge die ersten beiden Punkte zur konvexen Hülle hinzu.
4. Betrachte den nächsten Punkt P_k. Aufgrund der Sortierung liegt dieser zunächst außerhalb der konvexen Hülle. Liegt P_k links des Vektors $\overrightarrow{P_{k-2}P_{k-1}}$, so kann P_k direkt zur konvexen Hülle aufgenommen werden. Liegt P_k rechts des Vektors $\overrightarrow{P_{k-2}P_{k-1}}$, so muss zuerst P_{k-1} aus der konvexen Hülle entfernt werden, da es innerhalb der neuen konvexen Hülle liegen würde, bevor P_k hinzugefügt wird. Die Lage von P_k bezüglich des Vektors kann mit Hilfe der Determinante bestimmt $T(P_{k-2}, P_{k-1}, P_k)$werden.
5. Wiederhole Punkt 4 für die restlichen Punkte.

Determinante $T(A,B,C)$:

$$
\begin{aligned}
T(A,B,C) &= \begin{vmatrix} 1 & x_A & y_A \\ 1 & x_B & y_B \\ 1 & x_C & y_C \end{vmatrix} = \begin{vmatrix} 1 & x_A & y_B \\ 0 & x_B - x_A & y_B - y_A \\ 0 & x_C - x_A & y_C - y_A \end{vmatrix} = \begin{vmatrix} x_B - x_A & y_B - y_A \\ x_C - x_A & y_C - y_A \end{vmatrix} \\
&= (x_B - x_A)(y_C - y_A) - (x_C - x_A)(y_B - y_A) \\
&= \begin{cases} < 0, & \text{wenn C rechts von } \overrightarrow{AB} \text{ liegt.} \\ = 0, & \text{wenn C auf } \overrightarrow{AB} \text{ liegt.} \\ > 0, & \text{wenn C links von } \overrightarrow{AB} \text{ liegt.} \end{cases}
\end{aligned}
$$

Der Graham-Scan-Algorithmus ist somit ein einfaches und schnelles Verfahren($\mathcal{O}(n)$), um die konvexe Hülle für eine Punktmenge in der Ebene zu berechnen.

2.5 Visualisierung von Graphen

Ein Graph G ist als ein Tupel (V,E) definiert. Hierbei ist V eine Menge von Knoten und E eine Menge von Kanten. Für ungerichtete Graphen ohne Mehrfachkanten ist E eine zweielementige Teilmenge von V, für gerichtete Graphen ohne Mehrfachkanten eine Teilmenge von $V \times V$. Graphen mit Mehrfachkanten (Multigraph) sind für den weiteren Verlauf uninteressant und werden deshalb hier nicht definiert oder weiter betrachtet.

Der Grad $d_G(v)$ eines Knoten v ist die Anzahl der Kanten, die diesen mit anderen Knoten verbinden. Der Maximalgrad von Graphen G ist der größte Grad eines Knotens in G. Entsprechend ist der Minimalgrad von G der kleinste Grad eines Knotens in G.

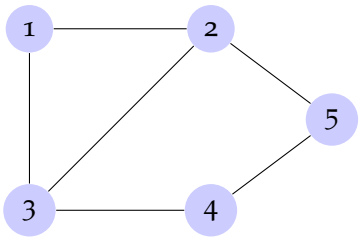


Abbildung 2.4: Beispielgraph

Der Beispielgraph in Abbildung 2.4 ist wie folgt definiert:

$$V = \{1,2,3,4,5\}$$

$$E = \{\{1,2\},\{1,3\},\{2,3\},\{2,5\},\{3,4\},\{4,5\}\}$$

Somit ergibt sich:

$$G = (V,E) = (\{1,2,3,4,5\},\{\{1,2\},\{1,3\},\{2,3\},\{2,5\},\{3,4\},\{4,5\}\})$$

Der Grad des Knotens 1 ist $d_G(1) = 1$, für Knoten 2 entsprechend $d_G(2) = 3$. Der Maximalgrad von G ist drei, der Minimalgrad zwei.

Im Allgemeinen werden die Knoten eines Graphen als Punkte und die Kanten als Verbindungen zwischen diesen Punkten dargestellt. Die Anordnung der Knoten und Kanten kann dabei variiert werden. Im Folgenden werden verschiedene Layoutmethoden vorgestellt:

2.5.1 Hierarchisches Layout

Das hierarchische Layout versucht eine Hierarchie in die Menge von Knoten zu bringen. Im einfachsten Fall ist ein gerichteter Graph mit einem Quelle (keine eingehenden Kanten) und mehreren Senken (keine ausgehenden Kanten) gegeben. Abhängig von der Entfernung zur Quelle können alle Kanten in Äquivalenzklassen eingeteilt werden. In komplizierteren Fällen muss ein geeigneter Algorithmus gefunden werden, um die Knoten in Äquivalenzklassen einzuteilen. Alle Knoten einer Äquivalenzklasse werden auf einer Ebene gezeichnet. Diese Ebene kann z. B. die Höhe oder die Entfernung vom Mittelpunkt des Graphen sein. Abbildung 2.5a zeigt ein Beispiel für ein einfaches hierarchisches Layout.

2.5.2 Orthogonales Layout

Beim orthogonalen Layout werden alle Kanten aus Teilstücken, die entweder vertikal oder horizontal verlaufen, zusammengesetzt. Abbildung 2.5b zeigt ein Beispiel für ein einfaches orthogonales Layout.

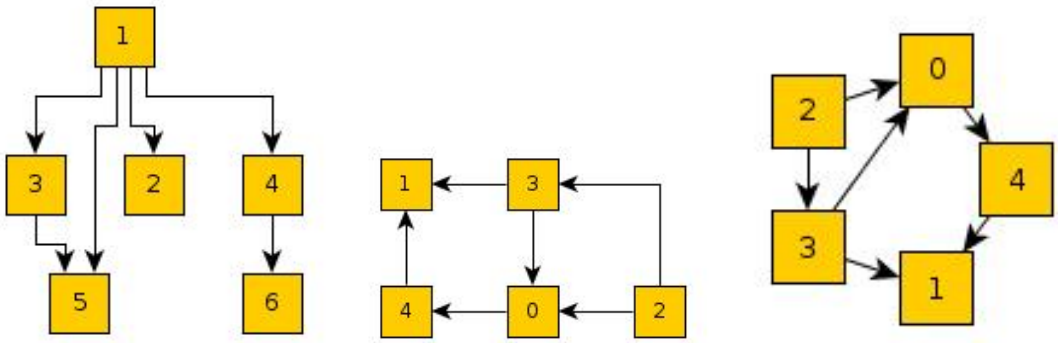
2.5.3 Zirkuläres Layout

Ein zirkuläres Layout ordnet die Knoten auf einem Kreis an. Dabei wird versucht, dass die Kanten möglichst auf dem Kreis selbst liegen und möglichst wenige Kanten durch das Innere des Kreises verlaufen. Abbildung 2.5c zeigt ein Beispiel für ein einfaches zirkuläres Layout.

2.5.4 Force-directed Layout

Die grundlegende Idee des Force-directed Layout ist es, physikalische Kräfte zwischen den Knoten zu simulieren mit dem Ziel, alle Knoten so zu positionieren, dass alle Kanten möglichst die selbe Länge haben und es so wenig wie möglich sich kreuzende Kanten gibt. Jeder Knoten wird als positiv geladenes Teilchen simuliert, so dass sich alle Knoten nach dem coulombschen Gesetz gegenseitig abstoßen. Jede Kante wird als Feder nach dem hookeschen Gesetz simuliert.

Somit ergibt sich die Kraft auf einen Knoten als Summe der wirkenden elektrischen Kräften und der Federkraft zwischen verbundenen Knoten. Die elektrische Kraft sorgt für eine



(a) hierarchisches Layout **(b)** orthogonales Layout **(c)** zirkuläres Layout

Abbildung 2.5: verschiedene Layouts für Graphen

gleichmäßige Abstoßung zwischen den Knoten. Die Federkräfte wirken diesen entgegen und erhalten dadurch die Struktur des Graphen.

2.6 Stemming

Als Stemming werden Verfahren bezeichnet die verschiedene Varianten eines Worts zurück auf ihren gemeinsamen Wortstamm führen. So sollte ein Stememer für die deutsche Sprache die Worte „Aufenthalt“, „Aufenthaltes“ und „aufenthalten“ auf den Wortstamm „aufenthalt“ zurückführen.

2.6.1 Porter-Stemming

1980 wurde von Porter ([Por97]) ein Stemming-Algorithmus entwickelt, der sich zum De-facto-Standard entwickelte. Der Algorithmus wurde ursprünglich für die englische Sprache entwickelt, lässt sich aber leicht für andere Sprachen anpassen.

Der Algorithmus basiert auf Regeln zur Verkürzung von Worten. Jedes Wort lässt sich als eine Sequenz aus Konsonanten und Vokalen darstellen. C steht für eine Folge von Konsonanten und V für eine Folge von Vokalen. Somit lässt sich jedes Wort in der Form $[C]VCVC\ldots[V]$ oder auch $[C](VC)^m[V]$ darstellen. m ist das Maß eines Wortes, z. B.:

$m = 0$ tr-ee, t-o

$m = 1$ w-**eb**, tr-**oubl**-e

$m = 2$ tr-**oubl-es**

Die Regeln um Suffixe zu entferne bestehen aus einer Bedingung und einer Ableitung in der Form $S_1 \rightarrow S_1$. Dies bedeutet, wenn ein Wort auf S_1 endet und der Wortteil vor S_1 die Bedingung erfüllt, wird S_1 durch S_2 ersetzt, z. B. verkürzt die Regel $(m > 0)\ EED \rightarrow EE$ „agreed“ zu „agree“. Die Regeln sind in Gruppen angeordnet. Aus jeder Gruppe darf jeweils nur eine Regel verwendet werden.

2.7 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) ist ein Wahrscheinlichkeitsmodell für Dokumente wie z. B. Texte und Bilder. Es wurde 2003 von Blei et al. vorgestellt ([BNJ03]). Jedes Dokument wird als eine Mischung verschiedener zugrundeliegender Themen betrachtet. LDA ermöglicht es die Ähnlichkeit zwischen Dokumenten anhand der zugrunde liegenden Themen zu erklären. Die Anzahl der Themen ist zu Beginn festgelegt, die Themen selbst ergeben sich jedoch aus der LDA.

2.8 Verwandte Arbeiten

Im folgenden werden verschiedene Arbeiten, die als Anregung für diese Arbeit dienten, vorgestellt.

2.8.1 Vizster: Visualizing Online Social Networks

Heer stellt mit Vizster ([HB05]) ein Werkzeug vor, dass es ermöglicht das soziale Netzwerke Friendster[2] zu visualisieren und zu untersuchen. Friendster wurde als Online-Dating-Plattform mit ausführlichen Nutzerprofilen entwickelt. Beziehungen zu anderen Nutzern werden als Freundschaften dargestellt. Freundschaften kommen nur zustande, wenn sie von beiden Seiten bestätigt werden. Zusätzlich ist es möglich die Freundschaft bzw. den anderen Nutzer zu beschreiben. Die vorhandenen Freundschaften wurden auf den jeweiligen Nutzerprofilen angezeigt. Dies sollte die Qualität der Nutzerprofile erhöhen. Friendster wurde von den Nutzern nicht als Dating-Plattform verwendet, jedoch wurde es als Kommunikationsplattform genutzt. Da Friendster die Sichtbarkeit von Profilen auf einen Freundschaftsgrad von vier (also Freunde von Freunden der Tiefe 4) beschränkte, waren die Nutzer versucht, möglichst viele Freundschaften zu schließen. Dadurch entstand schnell ein großes Netzwerk aus Freundschaften.

Das Ziel von Vizster ist es ein System zur Visualisierung von Freundschaftsnetzwerken zu entwickeln, das es auch Endnutzern einfach ermöglicht, Netzwerke zu untersuchen. Abbildung 2.6 zeigt ein Netzwer in Vizster, das mithilfe eines Graphen dargestellt wird. Der Graph wird mittels eines Force-directed Layout (s. Unterabschnitt 2.5.4) ausgerichtet. Die

[2] http://www.friendster.com/

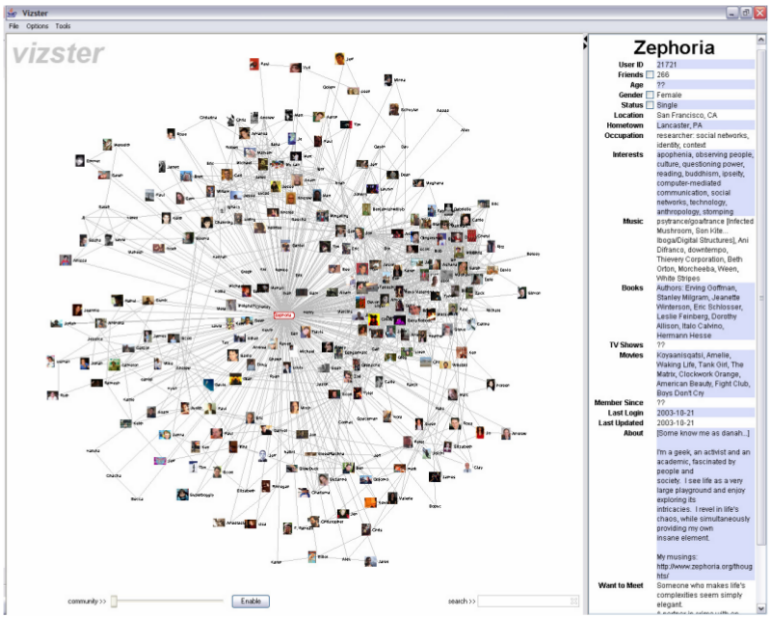


Abbildung 2.6: Vizster Grundansicht[HB05]

Federkraft ist abhängig von der Anzahl der Freundschaften eines Nutzers, so dass Nutzer mit wenigen Freunden stärker an ihre Freunde angezogen werden und Nutzer mit vielen Freunden sich weiter voneinander entfernen.

Durch Anklicken eines Nutzerknotens wird dessen Profil in der rechten Spalte des Programms angezeigt. Wird ein Nutzerknoten mit der Maus überfahren, so werden dessen Freunde hervorgehoben. Über die Suchleiste rechts unten kann direkt nach Nutzern und Stichworten auf deren Profil gesucht werden. Zusätzlich ist es möglich einzelne Untergruppen anhand der Verbindungen zwischen Nutzern zu berechnen und anzeigen zu lassen (s. Abbildung 2.7).

Vizster bietet somit eine einfache Möglichkeit sich einen Überblick über ein Freundesnetzwerk in Friendster zu verschaffen, sich Informationen zu Nutzern anzeigen zu lassen und Untergruppen/-netzwerke zu finden.

Vizster baut im Gegensatz zu dieser Arbeit das Netzwerk aus Freundschaftsbeziehungen auf. Diese Arbeit verwendet hierzu Kommunikation zwischen Nutzern.

2.8.2 Analyzing (Social Media) Networks with NodeXL

Smith et al. präsentieren mit NodeXL [SSMF+09] eine Erweiterung für Microsoft Excel 2007, das es ermöglicht verschiedenste Netzwerkdaten zu visualisieren und zu analysieren. Um die Fähigkeiten von NodeXL vorzustellen werden Daten eines Sozialen Netzwerks eines Unternehmens analysiert. Ähnliche Knoten können zu Clustern zusammengefügt werden, so dass diese gemeinsam untersucht werden können. Die Darstellung des Netzwerks kann mit

Abbildung 2.7: Vizster Untergruppen[HB05]

Hilfe verschiedener Filter angepasst werden. So kann z. B. die Größe der einzelnen Knoten an verschiedene Werte gekoppelt werden. Auch kann das Layout des Graphen ausgetauscht werden.

NodeXL kombiniert bekannten Datenverwaltung mit Excel mit einem System zur Visualisierung für Daten von Netzwerken . Im Gegensatz zu dieser Arbeit ist NodeXL nicht auf spezielle Daten angepasst, so dass für jeden Datensatz eine passende Visualisierung gefunden werden muss.

2.8.3 Conversation Map

Mit Conversation Map [Saco0] stellt Sack ein Newsgroup Browser vor, der die Inhalte von Newsgroup Nachrichten analysiert und anschließend eine Übersicht der beteiligten Nutzer und die behandelten Themen darstellt. Aus den beteiligten Nutzern wird eine Soziales Netzwerk berechnet, das anschließend als Graph dargestellt wird. Die behandelten Themen werden als Stichwortliste dargestellt. Ein Klick auf ein Stichwort hebt die beteiligten Konversationen hervor. Die Nachrichten der Diskussionen werden als Baum dargestellt, wobei die erste Nachricht die Wurzel ist und die Verbindungen zwischen Antworten die Äste. Wird ein Knoten im Baum angeklickt, so wird die zugehörige Nachricht angezeigt. Zusätzlich wird ein semantische Netzwerk aus Begriffen der Nachrichten berechnet. Verbunden Begriffe wurden in ähnlicher Weise verwendet. Ein Doppelklick auf einen Begriff zeigt alle Nachrichten, in denen er verwendet wird.

Conversation Map baut ähnlich wie diese Arbeit ein Soziales Netzwerk aus Nachrichten auf. Ebenso werden behandelte Themen als Stichworte angezeigt. Allerdings handelt es sich bei den Nachrichten um Newsgroup Einträge.

3 Konzept

Zur Erstellung eines Konzepts wird zuerst geklärt, wie Daten aus Twitter extrahiert werden können und welche Daten für ein Kommunikationsnetzwerk wichtig sind. Nach dem dies in den ersten Abschnitten behandelt wurde, kann in den folgenden Abschnitten ein Konzept zur Darstellung des Netzwerkes und der Inhalte entwickelt werden. Im letzten Abschnitt wird die entstandene Oberfläche und deren Interaktionsmöglichkeiten beschrieben.

3.1 Wie können Daten aus Twitter extrahiert werden?

Twitter bietet über die beiden APIs (Unterabschnitt 2.1.7) Zugriff auf verschiedene Daten: Es können sowohl Live-Daten als auch Bestandsdaten durchsucht werden. Die Bestandsdaten können über die REST-API (Abschnitt 2.1.7) abgerufen werden und bieten schnell Zugriff auf eine große Menge an Tweets, die über einen Zeitraum verteilt sind. Allerdings ist dieser Zeitraum von Twitter vorgegeben. Zur Zeit reicht dieser sechs bis neun Tage in die Vergangenheit. Möchte man über einen längeren Zeitraum als neun Tage Daten erhalten, so muss man diese Daten über die Streaming-API (Abschnitt 2.1.7) in Echtzeit sammeln, d. h. der Sammelprozess läuft den gesamten Zeitraum über und sammelt Tweets. Beide Zugriffsarten bieten die Möglichkeit nach Orten, Begriffen und Nutzern zu filtern.

Um Daten von Twitter zu sammeln und zu visualisieren bieten sich prinzipiell zwei verschiedene Möglichkeiten an: Entweder werden die Daten gesammelt und während des Sammelprozesses direkt angezeigt oder der Sammelprozess wird von der Visualisierung der Daten getrennt, so dass zwei getrennte Werkzeuge entstehen. Mit dem einen können die Daten gesammelt und mit dem anderen visualisiert werden. Werden die Daten direkt während des Sammelprozesses angezeigt, so erhält der Benutzer ein direktes Feedback zu den Daten. Außerdem ist es möglich in den laufenden Sammelprozess einzugreifen, Parameter zu ändern und somit den Sammelprozess zu beeinflussen. Da es durchaus vorkommen kann, dass ein Sammelprozess über Stunden oder sogar Tage läuft – unabhängig davon, ob über die REST-API oder die Streaming-API auf Twitter zugegriffen wird – muss die Visualisierung auch über den gesamten Prozess aktuell gehalten werden und ständig an die veränderten Daten angepasst werden. Ein mehrstufiger Sammelprozess erschwert die Visualisierung zusätzlich. Von einem mehrstufigen Sammelprozess kann man gesprochen werden, wenn z. B. erst Twitter-Daten gesammelt und diese in einem weiteren Schritt verändert werden. Wird der Sammelprozesses von der Visualisierung getrennt, so treten diese Probleme nicht auf, da die Visualisierung erst erfolgt, nachdem alle Daten vollständig gesammelt sind. Somit ist es unwichtig, ob und wie sich die Daten während des Sammelns verändern. Allerdings erhält man durch die Trennung keine direktes graphisches Feedback über die gesammelten

Daten und ein Eingreifen in den laufenden Prozess gestaltet sich ebenfalls schwierig. Ein weiterer Vorteil ist, dass das Werkzeug zum Sammeln der Daten ausgetauscht werden kann, ohne dass die Visualisierung hiervon betroffen ist.

	direkte Visualisierung	**getrennte Visualisierung**
Feedback	direktes Feedback mit Visualisierung der Daten	Visualisierung erst nach Abschluss des Sammelprozesses
Eingreifen in den Sammelprozess	Manipulation des Suchprozesses während des Sammelvorgangs	Manipulation nur schlecht möglich
Visualisierung bei mehrstufigem Sammelprozess	Mehrstufiger Sammelprozess schwierig zu Visualisieren	nicht nötig
Berechnung der Visualisierung	Visualisierung muss während des gesamten Sammelvorgangs aktualisiert werden	Visualisierung wird einmalig beim Laden der Daten berechnet
Nutzung mehrerer Computer	Sammeln der Daten und Visualisierung muss auf dem selben Computer erfolgen	Sammeln der Daten und Visualisierung kann auf getrennten Computern erfolgen
Austausch von Komponenten	schwer möglich	einfach

Tabelle 3.1: Vergleich der Ansätze

Tabelle 3.1 zeigt den direkten Vergleich der Vorteile und Nachteile der beiden Ansätze: Trotz des direkten Feedbacks und der Möglichkeit bei einer direkten Visualisierung der Daten in den Sammelprozess einfach eingreifen zu können, fällt die Wahl auf die Trennung des Sammelprozesses von der Visualisierung. Die Möglichkeit, Twitter-Daten unabhängig von der Visualisierung sammeln und verschiedene Vorgehensweisen beim Sammeln von Daten einfach austauschen zu können, erlauben eine flexible Gestaltung der Werkzeuge. Außerdem bietet sich die Möglichkeit, unabhängig von einer funktionierenden Visualisierung bereits Daten zu sammeln.

Als Schnittstelle zwischen Sammelprozess und Visualisierungswerkzeug dient ein vorgegebenes Speicherformat auf der Festplatte. Der Sammelprozess speichert die Ergebnisse dort ab. Von dort können sie vom Visualisierungswerkzeug ausgelesen werden. Die Daten auf der Festplatte enthalten nur die Tweets mit allen relevanten Inormationen. Das entstehende Netzwerk wird dabei nicht gespeichert, sondern wird erst beim Laden der Daten erechnet. Dadurch ist es möglich einfache Sammelwerkzeuge zu entwickeln, die die Struktur des Netzwerks nicht beachten müssen. Die Berechnung der Struktur des Netzwerks geschieht erst im Visualisierungswerkzeug und zwar beim Einlesen der Daten. Dies ermöglicht außerdem das Netzwerk gegebenenfalls an die Visualisierung anzupassen.

3.2 Welche Daten sind für ein Kommunikationsnetzwerk relevant?

Ein Kommunikationsnetzwerk zeichnet sich durch Nutzer, die miteinander kommunizieren, aus. Die Follower-Beziehung auf Twitter – welcher Nutzer welchem anderen Nutzer folgt – zeigt an, dass ein Nutzer Interesse an den Inhalten des Anderen hat. Folgen sich zwei Nutzer gegenseitig, so kann vermutet werden, dass sich beide Nutzer kennen und möglicherweise auch miteinander kommunizieren. Über die Follower-Beziehung lässt sich bereits das Soziale Netzwerk zwischen Nutzern als Graph darstellen. Durch die gerichteten Kanten des Graphen lassen sich leicht wichtige Nutzer, die viele Follower haben, finden. Diese Nutzer haben durch ihre vielen Follower ein großes Publikum und können über dieses Informationen weit verbreiten. Nutzer mit wenig Followern, die aber selbst vielen Nutzern folgen, nehmen eine passivere Rolle ein und konsumieren mehr Nachrichten.

Jedoch ist die Anzahl der Follower eines Nutzers nicht ausreichend, um zu beurteilen wie häufig ein Nutzer mit anderen Kommuniziert. Die Zahl der Follower ist somit kein Maß für die Kommunikationzwischen Nutzern.

> Kommunikation (lateinisch communicatio = Mitteilung, Unterredung; Verb communicare = teilhaben, mitteilen gebildet; Adjektiv communis = gemeinsam) bedeutet Verständigung untereinander; zwischenmenschlicher Verkehr besonders mithilfe von Sprache, Zeichen. Synonyme sind Informationsaustausch, Kontakt, Verständigung. [dud]

Kommunikation setzt also voraus, dass eine Mitteilung ausgetauscht wird. Für Twitter bedeutet dies, dass Nutzer Tweets an andere Nutzer adressieren (s. Unterabschnitt 2.1.1). Antwortet dieser Nutzer auf diesen Tweet, so entsteht Kommunikation zwischen beiden Nutzern. Kommunikation auf Twitter lässt sich somit als gegenseitiger Austausch von Nachrichten definieren.

Um Kommunikationsnetzwerke auf Twitter finden zu können, müssen die versendeten Tweets betrachtet werden. Wie in Abschnitt 2.1 beschrieben, liefert die Twitter-API Daten in JSON. Für das reine Netzwerk sind nur Absender und Empfänger eines Tweets von Interesse. Der Absender eines Tweets erhält man aus dem Wert des „user“-Schlüssels (Streaming-API, Abschnitt 2.1.7) bzw. aus den Werten der „from_user“- und „from_user_id“-Schlüssel (REST-API, Abschnitt 2.1.7). Der erste Empfänger eines Tweets steht in dem Wert des Schlüssels „in_reply_to_user_id“. Die weiteren Empfänger werden nur in „user_mentions“ aufgelistet. Jedoch werden dort alle Nutzer, die in einem Tweet erwähnt werden, aufgelistet. Diese sind nicht zwingend auch ein Empfänger des Tweets. Wie in Unterabschnitt 2.1.1 beschreiben, stehen die Empfänger am Anfang einer Nachricht. Da die Twitter-API es nicht ermöglicht alle Empfänger direkt auszulesen, muss zusätzlich auch der Text des Tweets betrachtet werden, um aus diesem die Empfänger des Tweets zu extrahieren. Hat man für jeden Tweet den Sender und den bzw. die Empfänger, so kann man daraus ein Netzwerk aufbauen.

Um Kommunikationsnetzwerke in Twitter zu finden, kann entweder themenbezogen gesucht oder von bestimmten Nutzern aus gestartet. Eine themenbezogene Suche kann Erfolg versprechen, da Unterhaltungen über bestimmte Themen gehen. Allerdings lässt sich ein

Tweet nicht zwingend dem Thema einer gesamten Unterhaltung zuordnen, wenn nur der Tweet eigenständig betrachtet wird. Von daher ist es auch möglich, sich interessante Twitter-Nutzer zu suchen und deren Kommunikationsnetzwerk zu betrachten.

3.3 Darstellung des Netzwerks

Für die Darstellung des Netzwerks sind hauptsächlich die Nutzer und Verbindung zwischen diesen interessant, da diese die Kommunikation darstellen. Daher wird das Soziale Netzwerk als Graph dargestellt. Der Graph dient nicht zur Darstellung der Inhalte der Nachrichten zwischen Nutzern, sondern soll einen Überblick über die Struktur des Netzwerkes und die beteiligten Nutzer bieten. Die Knoten des Graphs repräsentieren jeweils einen einzelnen Twitter-Nutzer. Der Knoten wird mit dem Namen des Nutzers versehen.

3.3.1 Darstellung von Knoten und Kanten

Die Größe eines Knotens ist abhängig von der Anzahl der versandten Nachrichten – NodeXL (s. Unterabschnitt 2.8.2) ermöglicht eine ähnliche Darstellung. Je mehr Nachrichten ein Nutzer gesendet hat, desto größer wird der Knoten dargestellt. Dabei ist zu beachten, dass sich die Größe der Knoten in einem vorgegebenen Rahmen bewegen muss. Kein Knoten darf so klein werden, dass er übersehen wird. Gleichzeitig darf ein Knoten auch nicht so groß werden, dass er andere Knoten überdeckt und nur noch dieser eine Knoten sichtbar ist.

Die Kanten im Graph stehen für Kommunikation zwischen Nutzern. Die Dicke der Kanten repräsentiert die Anzahl der ausgetauschten Nachrichten. Auf der Kante wird dabei auch die Anzahl der ausgetauschten Nachrichten angezeigt. Um einfach sehen zu können in welche Richtung wie viele Tweets in einer Kommunikation gesendet wurden und somit wie stark ein Nutzer an einer Kommunikation beiteiligt war, soll der Anteil der Kommunikation jedes Nutzers auf der Kante angezeigt werden.

3.3.2 Layout

Für das Layout des Graphen wird ein Force-directed Layout (s. Unterabschnitt 2.5.4) verwendet. Allerdings wird – ähnlich Vizster (s. Unterabschnitt 2.8.1) – die Länge der Federn nicht für alle Kanten konstant gewählt, sondern in Abhängigkeit vom Grad der beiden Knoten, die über die Kante verbunden sind ab. Der kleinere Grad der beiden bestimmt die Länge der Feder. Je Größer dieser ist, desto länger ist auch die Feder. Da der Grad der Knoten eines Netzwerks nach oben nicht beschränkt ist, kann die Feder theoretisch beliebig lang werden. Um dies zu verhindern, wird die Federlänge auf den Maximalgrad des Graphen normalisiert. Somit gruppieren sich Knoten mit Grad eins direkt um den Knoten, mit dem sie verbunden sind. Je höher der Grad zweier Knoten ist, desto größer ist auch die Entfernung zwischen diesen. Wobei jeweils der kleinere Grad der beiden Knoten betrachtet wird.

Zusätzlich zu einem normalen force-directed Layout, sollen die Knoten vom Benutzer verschoben werden können, so dass die Knoten umsortiert werden können. Wenn das automatische Layout bei der Analyse des Netzwerks stört, so soll dieses auch deaktiviert werden können. Dies kann z. B. der Fall sein, wenn mehrere isolierte Netzwerke vorhanden sind und sich diese immer weiter voneinander entfernen. Ist das automatische Layout deaktiviert, so bleiben alle Knoten an ihrer aktuellen Position, können aber von Hand verschoben werden.

3.4 Darstellung von Tweets

Neben den Nutzern und ihren Kommunikationspartnern sollen auch die Inhalte der Nachrichten dargestellt werden. Für die Darstellung eines Tweets sind nur der Sender, der Inhalt und das Sendedatum interessant. Der Empfänger kann direkt aus dem Inhalt herausgelesen werden. Somit muss für jeden Tweet nur der Absender, das Datum und der Inhalt angezeigt werden. Ein Tweet kann wie folgt dargestellt werden:

> *Montag, 6. August 2012 13:32* **Beispielnutzer**: @example wie geht es dir?

Um die Übersicht über die Tweets zu behalten, werden diese zeitlich sortiert.

Tweets können sowohl für einzelne Nutzer oder Themengebiete, als auch für Konversationen zwischen zwei Nutzern angezeigt werden. Für Nutzer und Themengebiete ist eine einfache Liste ausreichend. Wird eine Konversation angezeigt, so orientiert sich Liste an Tweets an der Darstellung von Konversationen auf Smartphones (s. Abbildung 3.1). Die Tweets des einen Nutzers werden linksbündig, die des anderen Nutzers rechtsbündig dargestellt.

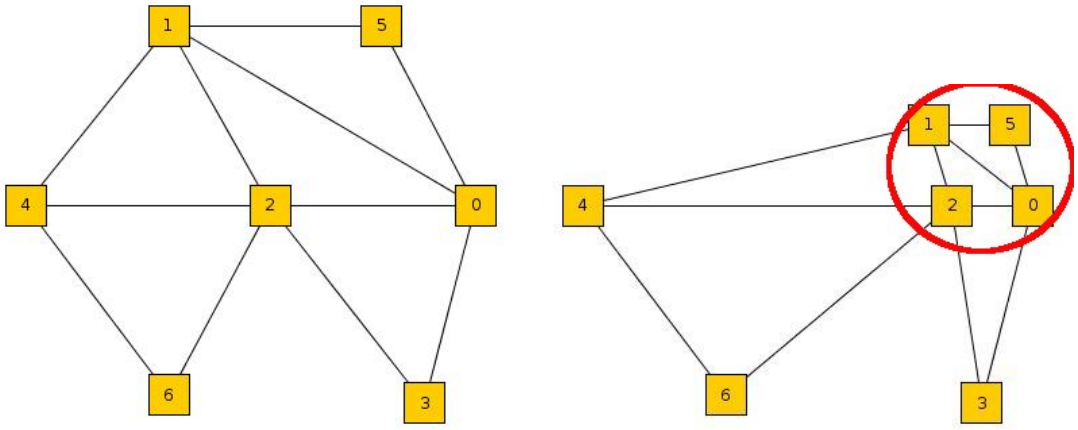
Abbildung 3.1: Konversation auf einem Smartphone

3.5 Themengebiete

Neben dem Kommunikationsnetzwerk und den Inhalten der Tweets sind auch zusätzlich größere Themengebiete interessant. Hierzu müssen aus allen Tweets Themen extrahiert

werden. Diese geschieht in mehreren Schritten: Zuerst werden die häufigsten Worte jedes Tweets bestimmt. Diese werden vollständig in Kleinschreibung überführt und anschließend so genannte Stop-Words entfernt. Dabei handelt es sich um häufig vorkommende Worte, die nicht zum bestimmen von Themen hilfreich sind und deshalb nicht beachtet werden sollen. Dann werden die Worte mittels Porter-Stemming (s. Unterabschnitt 2.6.1) auf ihren Wortstamm gebracht, so dass dann mittels LDA (s. Abschnitt 2.7) Themen aus den Tweets berechnet werden können.

Nach der Extraktion der Themen kann eine Liste mit Themen und den dazugehörigen Stichworten angezeigt werden. Die Darstellung der Themen mit Stichworten erfolgt über eine Tag Cloud (s. Abschnitt 2.2). Die Anzeige der Themen allein ist nicht ausreichend; zusätzlich muss es auch möglich sein, die Nutzer mit den entsprechenden Nachrichten, mit denen sie an einem Thema beteiligt sind, anzuzeigen. Diese Darstellung erfolgt direkt im Kommunikationsnetzwerk, wenn ein Thema aus der Liste ausgewählt wird: Um die beteiligten Nutzer wird ein farblich hervorgehobener Bereich gezeichnet. Gleichzeitig wird das Layout des Graphen so angepasst, dass Nutzer, die an dem ausgewählten Thema beteiligt sind, sich stärker anziehen. Somit wird die Zusammengehörigkeit der Nutzer sowohl über die farbliche Hervorhebung als auch über die stärkere Anziehung zum Ausdruck gebracht. Abbildung 3.2 zeigt wie sich das Layout des Graphen verändert, wenn ein Thema ausgewählt und somit hervorgehoben wird.



(a) normales Layout **(b)** Layout mit hervorgehobenem Thema

Abbildung 3.2: Hervorheben eines Themas

Die Anzahl der Themen für LDA ist fest vorgegeben. Somit kann es geschehen, dass einzelne Themen zu klein und somit nicht aussagekräfitig sind. Umgekehrt kann es auch vorkommen, dass ein Theme zu groß ist und dadurch Stichworte zu verschiedenen Themen enthält. Deshalb soll es möglich sein die Anzahl der Themen zu ändern und diese neu berechnen zu lassen.

3.6 Zeitlicher Filter

Da sich je nach Sammelprozess die Daten über einen gewissen Zeitraum erstrecken können, sollte es möglich sein, auch nur Teile des Zeitraums anzuzeigen, um so den zeitlichen Verlauf von Konversationen und die Entwicklung des Netzwerkes analysieren zu können. Für den angezeigte Zeitraum soll die Zeitdauer, sowie Start- und Endzeit frei ausgewählt werden können.

Die einfachste Lösung hierfür wären zwei Textfelder, in die das Start- und Enddatum des gewünschten Zeitraums eingetragen wird. Allerdings sind hier leicht Fehleingaben möglich und der Benutzer erhält keine Übersicht über den ausgewählten Zeitraum im Verhältnis zur gesamten Zeitspanne der Daten. Um eine einfache graphische Filtermöglichkeit zu bieten, soll ein Zeitleiste, die die gesamte Zeitspanne der geladenen Daten abdeckt und die der Zeitanzeige gängiger Audio- oder Videoabspielsoftware gleicht, angezeigt werden. Über zwei Regler, die unabhängig voneinander bewegt werden können, wird das Start- und Enddatum des gewünschten Zeitraums ausgewählt. Um Veränderungen innerhalb gleich langer Zeitabschnitte einfach vergleichen zu können, sollte es auch möglich sein, die beiden Zeitschieber gleichzeitig zu bewegen. Dies ermöglicht z. B. einen Zeitraum von einem Tag auszuwählen und dieses Zeitfenster anschließend über der Zeitleiste zu verschieben.

Wird der angezeigte Zeitraum geändert, müssen die angezeigten Daten natürlich diesem angepasst werden. Im Kommunikationsgraph (s. Abschnitt 3.3) muss sowohl die Größe der Knoten als auch die Dicke der Kanten angepasst werden, da diese von der Anzahl der Tweets abhängt. Die Themengebiete (s. Abschnitt 3.5) können sich durch die veränderte Anzahl an Tweets ebenfalls ändern. Somit müssen die Themen bei einer Änderung des angezeigten Zeitraums neu berechnet werden. Eventuell im Graph eingezeichnete Themengebiete müssen ausgeblendet werden, da sie ungültig sind. Werden in diesem Moment die Tweets eines Nutzers oder einer Konversation angezeigt, so muss diese Ansicht ebenfalls aktualisiert werden, so dass nur noch die Tweets angezeigt werden, die innerhalb des neuen Zeitraums liegen. Werden die Tweets eines Themas angezeigt, so ist diese Auswahl, wie oben beschrieben, ungültig und muss nicht aktualisiert werden, es werden somit keine Tweets angezeigt.

3.7 Ausblenden von Kanten und Knoten

Aus verschiedenen Gründen kann es vorkommen, dass in den Daten einseitige Kommunikationen vorhanden sind. Also dass Tweets von einem Nutzer an einen anderen vorhanden sind, aber keine Antworten des anderen Nutzers. Dies kann z. B. durch einen Sammelprozess, der auch einseitige Kommunikation sammelt, geschehen. Auch wenn einseitige Kommunikation im Sammelprozess verhindert wird, kann es durch das zeitliche Filter(s. Abschnitt 3.6) zu einseitiger Kommunikation kommen, wenn die Antworten außerhalb des ausgewählten Zeitraums liegen.

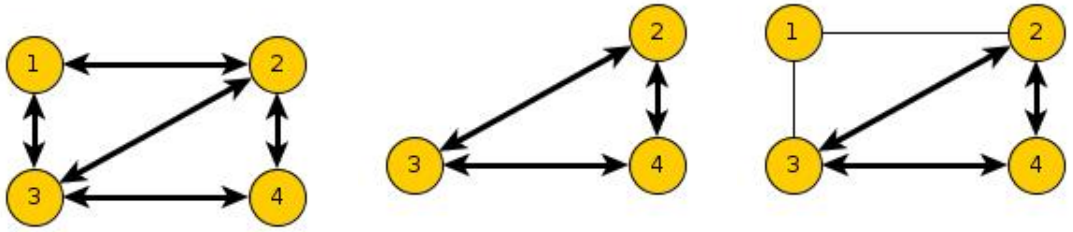
Sollte einseitige Kommunikation nicht erwünscht sein, so soll es die Möglichkeit geben, diese einfach auszublenden. Durch das Ausblenden von einseitiger Kommunikation und

durch das zeitliche Filtern, kann es vorkommen, dass ein Nutzer an keiner Kommunikation mehr beteiligt ist. Entweder, weil er keine Tweets im ausgewählten Zeitabschnitt hat oder weil er nur noch einseitig kommuniziert und einseitige Kommunikation ausgeblendet ist. Dadurch wird der Nutzer vom restlichen Graph isoliert. Dies führt dazu, dass sich dieser immer weiter von den restlichen Knoten entfernt.

Um dies zu verhindern, solle es zwei Möglichkeiten geben:

1. Nutzer, denen keine Tweets zugeordnet sind oder die an keiner Kommunikation beteiligt sind, werden, wie die Kommunikationen, ebenfalls ausgeblendet.
2. Ausgeblendete Kommunikationen werden nicht vollständig ausgeblendet, sondern als dünne Linien trotzdem eingezeichnet. Dadurch werden die Nutzer nicht vom Graph getrennt.

Beide Möglichkeiten sollen unterstützt werden und es soll möglich sein, zwischen diesen umzuschalten.



(a) Ausgangsgraph

(b) Ausblenden der Kanten und Knoten

(c) dünnes Einzeichnen der Kanten

Abbildung 3.3: Ausblenden von Kanten und Knoten

Abbildung 3.3 zeigt den Unterschied zwischen beiden Methoden: Wird aus dem Ausgangsgraph(Abbildung 3.3a) alle Kommunikation zwischen Knoten 1 und 2, sowie zwischen Knoten 1 und 3 ausgefiltert, so entsteht durch die erste Methode Abbildung 3.3b. Die zweite Methode liefert Abbildung 3.3c.

3.8 Übersicht über die Benutzeroberfläche und Interaktionsmöglichkeiten

Als ersten Schritt läd der Benutzer die zuvor gesammelten Daten. Dies geschieht über einen einfachen Dateidialog, mit dem das Lucene-Repository geöffnet wird.

Nach dem Laden der Daten erscheint für den Benutzer eine Oberfläche, die die folgenden Komponenten enthält:

1. Kommunikationsgraph: Wie in Abschnitt 3.3 beschrieben, wird hier der Kommunikationsgraph dargestellt. Zusätzlich ist hier in Rot bereits ein Thema eingezeichnet.

2. Themenliste: Wie in Abschnitt 3.5 beschrieben, werden hier Tag Clouds zu den gefundenen Themen dargestellt.
3. Schieber zum Ändern der Anzahl der Themen und Button zum Neuberechnen der Themen.
4. Zeitfilter: s. Abschnitt 3.6
5. Tweet-Anzeige: Hier werden wie in Abschnitt 3.4 beschriebe Tweets zu Nutzern, Konversationen oder Themen angezeigt.
6. Einstellungen zum Ausblenden von Kanten und Knoten(s. Abschnitt 3.7)

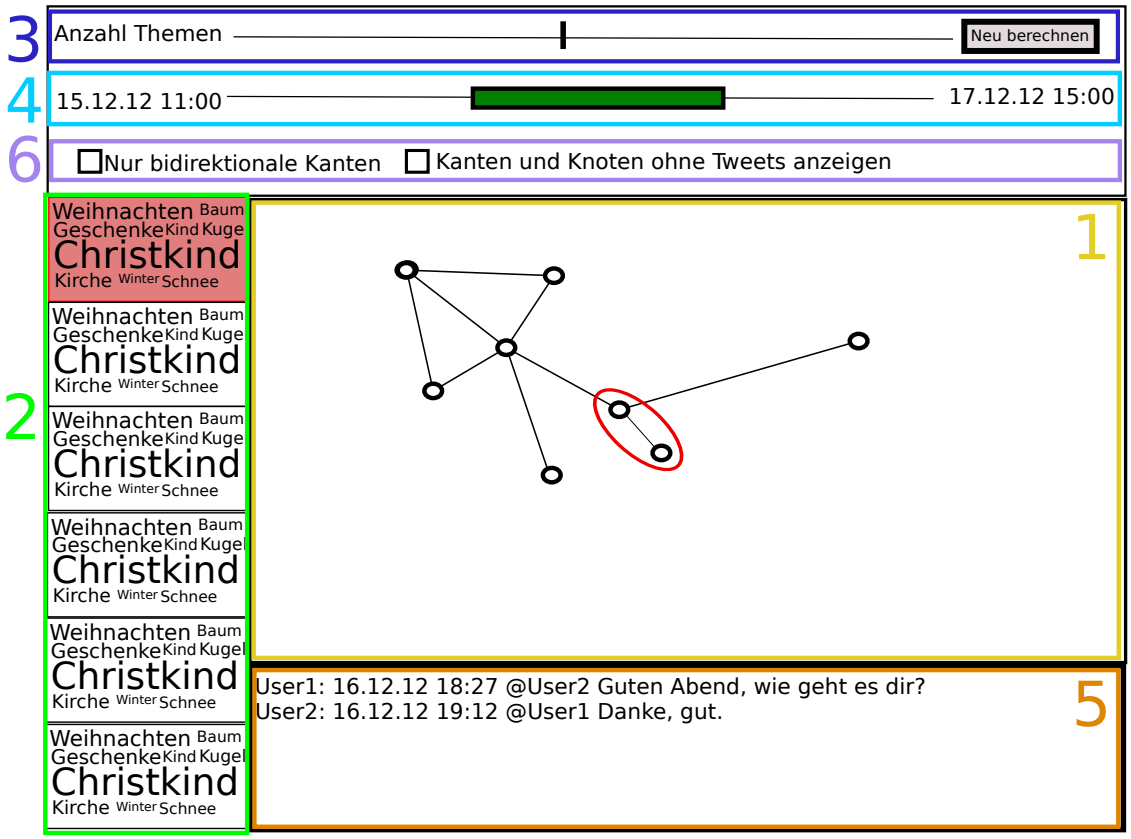


Abbildung 3.4: Prototyp der Benutzeroberfläche

Abbildung 3.4 zeigt einen Prototyp der Oberfläche, in dem die einzelnen Komponenten farblich markiert und ensprechend beschriftet wurden.

Da der Kommunikationsgraph die meisten Informationen anzeigt, wird diesem in der Standardansicht auch der meiste Platz eingeräumt. Im Kommunikationsgraph werden die Benutzer und die zugehörigen Kommunikationen gezeichnet. Zusätzlich können Themengebiete eingezeichnet werden, die sich, wie in Abschnitt 3.5 beschrieben, auf das Layout des Graphen auswirken. Der Benutzer kann einen Knoten des Graphen mit der Maus auswählen, so dass dieser farblich hervorgehoben wird und die Tweets des Nutzers in der Tweet-Anzeige

sichtbar sind. Wird eine Kante angeklickt, so wird diese hervorgehoben und die Kommunikation in der Tweet-Anzeige dargestellt. Werden Themen im Graph angezeigt, so können diese ebenfalls angeklickt werden und es werden alle Tweets, die an dem ausgewählten Thema beteiligt sind, in der Tweet-Anzeige dargestellt. Außerdem werden alle Nutzer, die an dem Thema beteiligt sind, im Graph hervorgehoben. Die Tweets in der Tweet-Anzeige sind immer zeitlich sortiert. Die Trennleiste zwischen Graph und Tweet-Anzeige kann verändert werden, so dass mehr oder weniger Tweets sichtbar sind. Bei Bedarf kann die Tweet-Anzeige auch minimiert werden, so dass der Graph möglichst viel Platz zur Verfügung hat.

Die Liste der Tag Clouds zeigt die mittels LDA gefundenen Themen an. Sollte der vorhandene Platz nicht reichen um alle Tag Clouds anzuzeigen, so sollte es möglich sein durch die Liste zu scrollen. Über die Liste kann die Anzeige der Themen im Kommunikationsgraph verändert werden. Durch Klick auf eine Tag Cloud wird diese im Graph eingezeichnet. Zusätzlich wird das Thema farblich hervorgehoben, wie in Abbildung 3.4 beim ersten Thema bzw. bei der ersten Tag Cloud zu sehen. Dadurch wird schnell sichtbar, welche Themen gerade eingezeichnet sind. Ein erneuter Klick blendet sie wieder aus. Es können auch mehrere Themen gleichzeitig eingeblendet werden. Die Stichworte eines Themas können ebenfalls durch einen Klicken im Graph eingeblendet werden. Genau wie komplette Themen wird das Stichwort farblich hervorgehoben und kann durch einen erneuten Klick wieder ausgeblendet werden. Die Anzeige eines Themas bzw. eines Stichwortes im Graphen ist nicht nur auf eines beschränkt. Es können mehrere Themen und Stichworte gleichzeitig angezeigt werden, so dass sie miteinander verglichen werden können. Um noch vor dem Einblenden eines Themas oder Stichwortes sehen zu können, wie viele Tweets daran beteiligt sind, wird diese Zahl als Tooltip in der Tag Cloud angezeigt. Für das gesamte Thema wird die Anzahl aller involvierten Tweets als Tooltip dargestellt. Dieser wird angezeigt, wenn der Mauszeiger über einer freien Stelle der Tag Cloud ruht.

Da die Anzahl der Themen für LDA fest vorgegeben werden muss, ist es möglich über einen Schieber die Anzahl der Themen zu ändern. Dies kann z. B. der Fall sein, wenn nur eine geringe Anzahl an Themen wirklich von den Twitter-Nutzern behandelt wird, aber deutlich mehr Themen für LDA vorgegeben sind, so dass verschiedene Themen ähnliche Stichworte enthalten und besser zusammengefasst gehören. Durch ändern der Anzahl der Themen, kann der Benutzer dies so anpassen, dass die gefundenen Themen sinnvolle Gebiete umfassen. Durch eine Verringerung der Themen ist es außerdem möglich einen groben Überblick über die behandelten Themen zu erhalten. Wird die Anzahl der Themen erhöht, so werden die einzelnen Themengebiete kleiner und man erhält einen genaueren Einblick. Nachdem die Anzahl der Themen angepasst wurde, werden die Themengebiete neu berechnet.

Wie in Abschnitt 3.6 beschrieben, ermöglicht der Zeitfilter den beobachteten Zeitraum einzuschränken und zu verändern. Der hier grün dargestellte Balken stellt den ausgewählten Zeitraum dar. Die Ränder des Balken können mit gedrückter Maustaste verschoben werden, um das Anfangs- und Enddatum zu verändern. Der gesamte Balken kann mit gedrückter Maustaste ebenfalls verschoben werden. Der Zeitraum zwischen Anfangs- und Enddatum bleibt dabei gleich. Nur Tweets, die innerhalb des gewählten Zeitraums liegen, werden in der Visualisierung beachtet und dargestellt. Eine Ausnahme bilden die Tag Clouds. Diese werden durch das Ändern des angezeigten Zeitraums ungültig. Eine Interaktion mit ihnen

ist dann nicht mehr möglich. Auch werden im Graph keine Themen mehr angezeigt. Erst durch die Betätigung des Buttons „Neu berechnen“ werden die Themen und somit auch die Tag Clouds neu berechnet und können wieder verwendet werden. Dieser Ansatz wird gewählt, da das Berechnen der Themen etwas Zeit in Anspruch nimmt und während dessen keine Interaktion mit der Oberfläche möglich ist. Würde bei jeder kleinen Änderung des Zeitfilters die Oberfläche blockiert, wäre es schwer möglich die Änderungen innerhalb des Graphen zu beobachten, wenn der Zeitraum geändert wird. Durch eine Entkopplung der Themenberechnung und des Zeitfilters kann dieser verändert werden, ohne dass dabei die Oberfläche blockiert wird. Wurde der gewünschte Zeitraum gefunden, können anschließend die Themen über den Button neu berechnet werden.

4 Implementierung

Im Folgenden wird die Implementierung das erarbeiteten Konzepts beschrieben.

4.1 Speicherung der Daten

Wie in Abschnitt 3.1 beschrieben, werden die Daten des Sammelprozesses auf der Festplatte zwischengespeichert. Für die Speicherung der Daten wird die Bibliothek Lucene verwendet. Lucene bietet eine einfache Möglichkeit Daten zu speichern, zu indizieren und zu durchsuchen. Lucene verwaltet Daten in so genannten Repositories. Diese sind ein einfacher Ordner auf der Festplatte; die einzelnen Daten werden in Dokumenten zusammengefasst. Für die Implementierung entspricht ein Dokument einem Tweet mit allen relevanten Daten, wie Sender, Empfänger, Datum und Text. Durch die Definition eines Dokuments ergibt sich ein einheitliches Austauschformat zwischen Sammelprozess und Visualisierung.

Für die Visualisierung werden die Daten aus dem Lucene-Repository geladen und direkt als Kommunikationsstruktur zwischen Nutzern gespeichert. Eine Kommunikation findet immer zwischen genau zwei Nutzern statt. Ist ein Tweet an mehrere Nutzer gerichtet, so wird er auch mehreren Kommunikationen zugeordnet. z. B. wird der folgende Tweet von „FrauMustermann“ geladen:

> @HerrMustermann Wie geht es dir?

Der Tweet wird der Kommunikation zwischen „HerrMustermann“ und „FrauMustermann“ hinzugefügt. Existiert die Kommunikation noch nicht, so wird eine neue zwischen beiden Nutzern erstellt. Hat ein Tweet mehrere Empfänger, so wird er auch zu mehreren Kommunikationen hinzugefügt. Somit ergibt sich eine Datenstruktur, die sich nahe an der späteren Visualisierung des Netzwerks orientiert.

4.2 Sammeln der Daten

Es ist möglich sowohl themen-, als auch nutzerbezogen nach Kommunikationsnetzwerken suchen, wie in in Abschnitt 3.2 beschrieben. Im Folgenden werden verschieden Verfahren zum Extrahieren von Kommunikationsnetzwerken und die Ergebnisse, die man damit erhält, beschrieben:

4.2.1 Sammeln nach Stichworten

Da sich Unterhaltungen meist um ein bestimmtes Thema drehen, bietet es sich an für ein gewähltes Thema eine Stichwortliste zu erstellen und Twitter nach dieser zu durchsuchen. Hierfür wurde die Streaming-API (Abschnitt 2.1.7) von Twitter verwendet und ein Filter für die Stichworte der Liste erstellt. Es werden nur Tweets gesammelt, die eines der Stichworte enthalten. Somit sollte es möglich sein, Unterhaltungen zu bestimmten Themengebieten zu finden.

Ergebnis

Nach mehreren Durchläufen zeigt sich ein recht klares Bild: Wie in Abbildung 4.1 zu sehen, ergeben sich praktisch keine größeren zusammenhängende Netzwerke. In den meisten

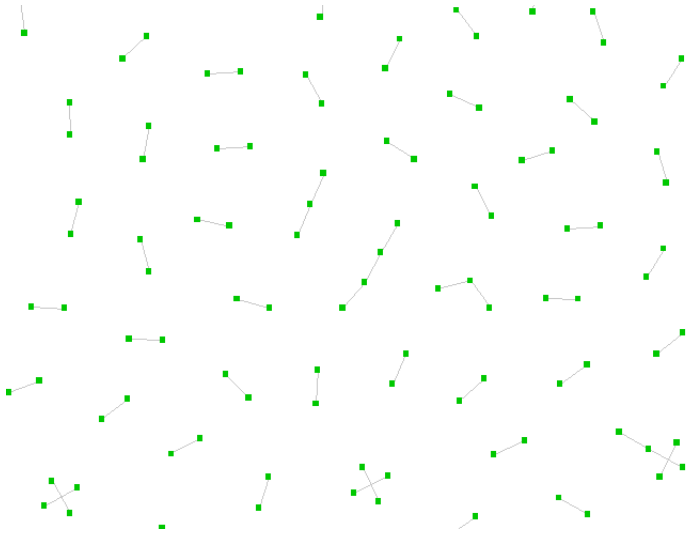


Abbildung 4.1: Ergebnis des Sammelprozesses nach Stichworten

Fällen gibt es nur Verbindungen zwischen zwei Twitter-Nutzern, selten Ketten aus mehreren Nutzern und praktisch nie mehrere Twitter-Nutzer, die alle miteinander schreiben. Die Erklärung hierfür lässt sich im Sammelansatz finden: Es werden nur Tweets mit bestimmten Stichworten gesammelt. Da die Antwort auf diesen Tweet nicht zwingend dieses Stichwort auch enthält wird diese beim Sammelprozess übersehen. Auch Tweets an andere Nutzer, die das Stichwort nicht enthalten, werden nicht gesammelt. Zusätzlich wird durch die Verwendung der Streaming-API nur Tweets gesammelt, die während des Sammelvorgangs versendet werden. Somit werden Antworten oder Tweets, die anderen Tweets vorausgehen, aber vor dem Start des Sammelvorgangs gesendet wurden, nicht erfasst. Dieser Ansatz ist somit kaum geeignet, um Konversationen zwischen Nutzern zu finden.

4.2.2 Ausgehend von einem Nutzer

Da sich im obigen Ansatz gezeigt hat, dass eine themenbezogene Suche nur einen Teil der Tweets, die von Interesse sind, gefunden werden, wird in diesem Verfahren bei einem Twitter-Nutzer gestartet und von diesem aus gesammelt mit wem er kommuniziert. Da sich zusätzlich gezeigt hat, dass die Streaming-API den Nachteil hat, dass nur Live-Daten gesammelt werden, wird hierfür die REST-API (Abschnitt 2.1.7) verwendet, so dass man in der Vergangenheit suchen kann.

Wie in Abschnitt 3.2 definiert, ist Kommunikation nicht einseitig. Somit muss beim Sammelvorgang beachtet werden, dass nur Verbindungen betrachtet werden, auf denen in beide Richtungen geschrieben wird. Für die Nutzer, die man anschließend erhält, wird das Verfahren wiederholt. Für die sich daraus ergebenden Nutzer ebenfalls, usw. Somit ergibt sich ausgehend vom ersten Nutzer eine Baumstruktur. Diese Struktur muss jedoch während des Sammelvorgangs nicht explizit gespeichert werden. Abbildung 4.2 zeigt einen möglichen Nutzerbaum. Ausgehend von Nutzer „User“ wurden dessen Tweets, die an andere Nutzer adressiert sind, gesammelt. Für jeden Adressat eines Tweet wird nun überprüft, ob dieser dem Nutzer „User“ geschrieben hat. Trifft dies zu, so wird der Nutzer in die Suche aufgenommen. Im Beispiel sind das die Nutzer „User1“, „User2“, „User3“ und „User4“. Für diese wird der Vorgang wiederholt. Um diesen Vorgang nicht endlos zu wiederholen, wird die Suchtiefe und somit die Anzahl der Iterationen vorgegeben. Im gegebenen Beispiel wurde die Suche auf die Tiefe 2 beschränkt. Sollte es vorkommen, dass ein Nutzer mehrfach gefunden wird, z. B. weil sowohl „User1“, als auch „User3“ diesem schreiben, so wird dieser nur einmal betrachtet, da beim ersten Vorkommen des Nutzers bereits alle Tweets gesammelt werden.

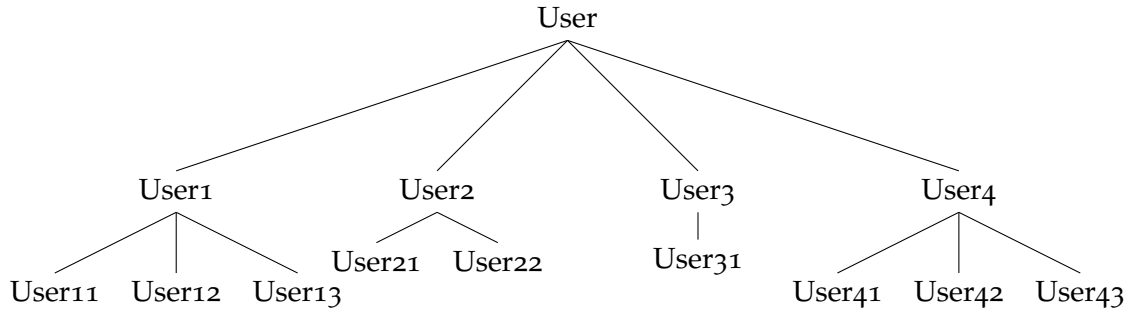


Abbildung 4.2: Beispiel für einen Nutzerbaum

Ergebnis

Die Baumstruktur, die sich aus dem Suchvorgang ergibt, zeigt sich auch deutlich in dem erhaltenen Kommunikationsnetzwerk: Nur wenn Nutzer zusätzlich außerhalb der Baumstruktur mit anderen Nutzern schreiben, ergibt sich ein echtes Kommunikationsnetzwerk. Damit Kommunikation außerhalb des Nutzerbaums zustande kommt, muss ein Nutzer einem anderen Nutzer schreiben, der bereits gefunden wurde, sich folglich entweder auf

der gleichen oder einer geringeren Tiefe im Baum befindet. Schreibt z. B. sowohl „User1“ als auch „User31“ mit „User12“, so entstehen Querverbindungen und es ergibt sich ein kleines Unternetzwerk mit drei Mitgliedern, die alle untereinander Kommunizieren. Die Wahrscheinlichkeit, dass solche Verbindungen auftreten, steigt mit einer größeren Suchtiefe. Allerdings steigt mit der Suchtiefe auch die Anzahl der Knoten und somit der Nutzer exponentiell an. Durch die erhöhte Nutzerzahl sinkt gleichzeitig die Übersicht im Netzwerk. Wird die Suchtiefe verringert, so steigert sich zwar die Übersichtlichkeit, jedoch sinkt auch die Anzahl der Querverbindungen. Somit muss für dieses Verfahren ein Kompromiss aus Querverbindungen und Übersichtlichkeit gefunden werden. Es zeigt sich, dass eine Suchtiefe von 2 bis 3 gute Ergebnisse liefert.

4.2.3 Ausgehend von einem Nutzer mit Stichworten

Um die Suchtiefe steigern zu können und gleichzeitig die Menge der Nutzer gering zu halten, werden die ersten beiden Verfahren kombiniert. Wie im ersten Verfahren wird eine Liste mit Stichworten vorgegeben. Gleichzeitig geht man aber, wie im zweiten Verfahren, von einem Nutzer aus und sucht dessen Kommunikationspartner. Es werden allerdings nur Tweets gesammelt, in denen mindestens eines der vorgegebenen Stichworte enthalten ist.

Ergebnis

Wie im obigen Verfahren zeigt sich auch die Baumstruktur im gesammelten Netzwerk. Allerdings ist wie gewünscht die Anzahl der Nutzer reduziert. Dadurch wird es möglich ohne die Übersicht zu verlieren die Suchtiefe zu erhöhen, um dadurch mehr Querverbindungen zu erhalten. Durch das Filtern nach Stichworten werden allerdings die Kommunikationen lückenhaft, da nicht jeder Tweet erfasst wird.

4.2.4 Ausgehend von mehreren Nutzern

Um gezielt nach zusammenhängenden Kommunikationsstrukturen suchen zu können, werden Verbindungen zwischen mehreren Twitter-Nutzern gesucht. Die Verbindungen müssen nicht zwingend direkt zwischen diesen Nutzern vorhanden sein, sondern können sich auch über mehrere Nutzer ergeben. Das Vorgehen ähnelt der bidirektionale Suche.

Zuerst wird für jeden Ausgangsnutzer ein Nutzerbaum (wie in Unterabschnitt 4.2.2) aufgebaut. Allerdings wird in diesem Falle die Baumstruktur während des Sammelns gespeichert. Um zu verhindern, dass Nutzer doppelt bearbeitet werden, muss eine Liste von bereits bearbeiteten Nutzern global gespeichert werden. Sollte ein Nutzer in einem Nutzerbaum ein zweites Mal auftauchen, so wird für diesen kein neuer Knoten angelegt, sondern eine Querverbindung zu dem bereits vorhandenen Knoten gespeichert. Um Arbeitsspeicher zu sparen, werden die gesammelten Tweets direkt in Lucene gespeichert und nur die Nutzerbäume im Arbeitsspeicher zwischen gespeichert.

Nachdem der Sammelvorgang abgeschlossen ist, wird jeder entstandene Nutzerbaum beginnend mit den Blattknoten bearbeitet. Für jeden Knoten wird überprüft, ob er Verbindung außerhalb der Baumstruktur besitzt, also ob Querverbindungen von oder zu diesem Knoten vorhanden sind. Sind keine Querverbindungen vorhanden, wird der Knoten gelöscht und der nächste Knoten bearbeitet. Beim Löschen eines Knoten ist zu beachten, dass auch alle zugehörigen Tweets aus Lucene entfernt werden müssen. Der Baum wird somit von unten her gekürzt und es bleiben nur Knoten übrig, die mehrere Verbindungen zu anderen Nutzern haben und eine Verbindung zwischen den Ausgangsnutzern herstellen.

Um den Vorgang zu verdeutlichen, wird dieser im folgenden Anhand eines einfachen Beispiels erläutert. Ausgehend von „root1“, „root2“ und „root3“ werden Nutzerbäume inklusive Querverbindungen und somit ein Netzwerk aus Nutzern erstellt. Dieses ist in Abbildung 4.3 dargestellt. Die Verbindungen außerhalb der Nutzerbäume sind grün und fett hervorgehoben. Im Beispiel ist dies z. B. die Verbindung zwischen „1-2“ und „2-3“.

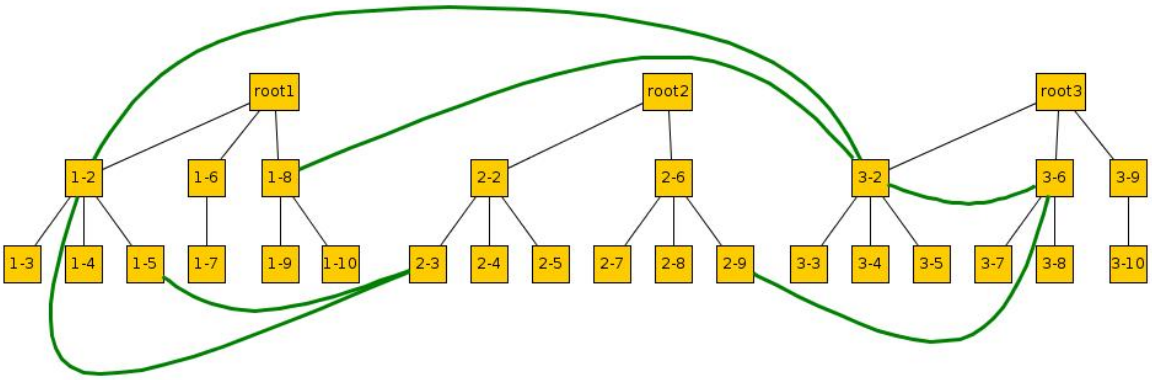


Abbildung 4.3: Sammelprozess: Vor dem Entfernen der Knoten

Anschließend werden alle Blattknoten, die keine Verbindungen außerhalb des Nutzerbaums haben, entfernt. In Abbildung 4.4 sind diese Knoten rot markiert. Nachdem diese Knoten

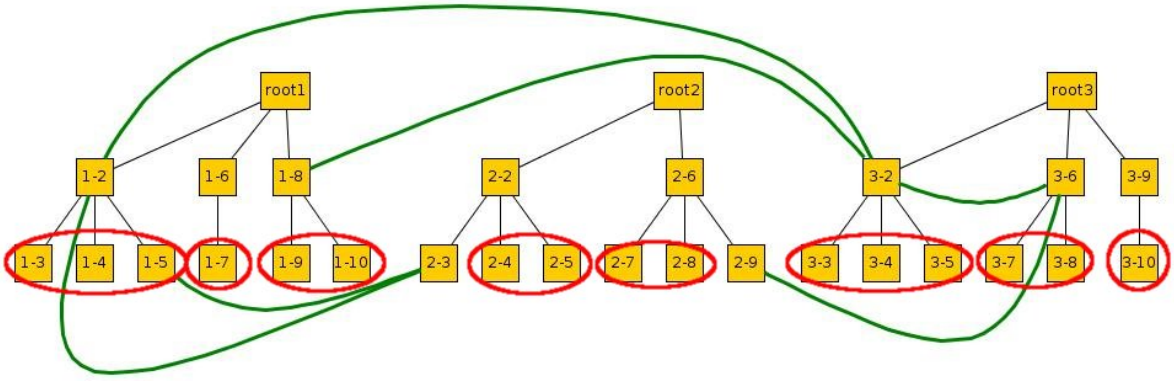


Abbildung 4.4: Sammelprozess: Zu entfernende Knoten markiert

entfernt wurden ergibt sich ein Netzwerk, wie in Abbildung 4.5. Der Vorgang wird so lange für alle Blattknoten wiederholt, bis die Wurzelknoten erreicht sind. In Abbildung 4.5 sind die Knoten der Tiefe 1, die anschließend entfernt werden, markiert. Nach dem Entfernen gelangt

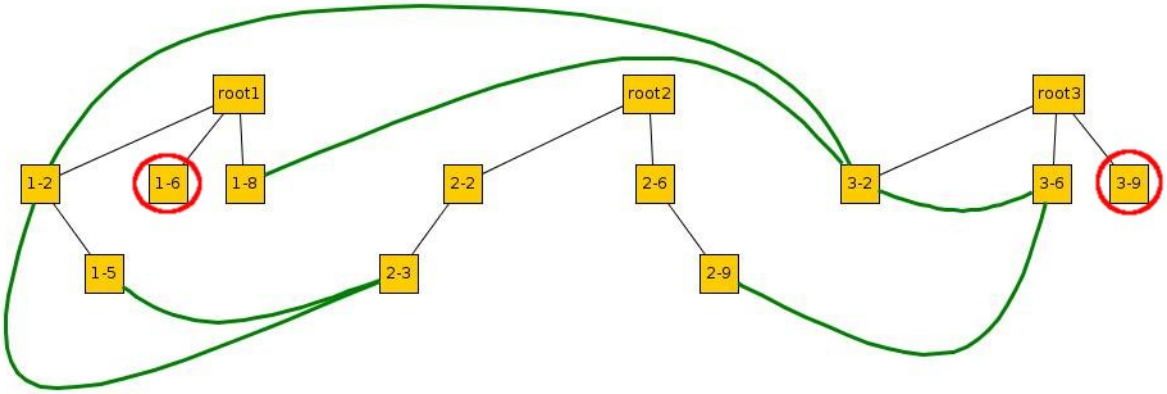


Abbildung 4.5: Sammelprozess: Erste Knoten entfernt und Knoten in Tiefe 1 markiert

man im nächsten Schritt zu den Blattknoten, somit ist der Vorgang für dieses Netzwerk abgeschlossen und man erhält ein Ergebnis wie in Abbildung 4.6.

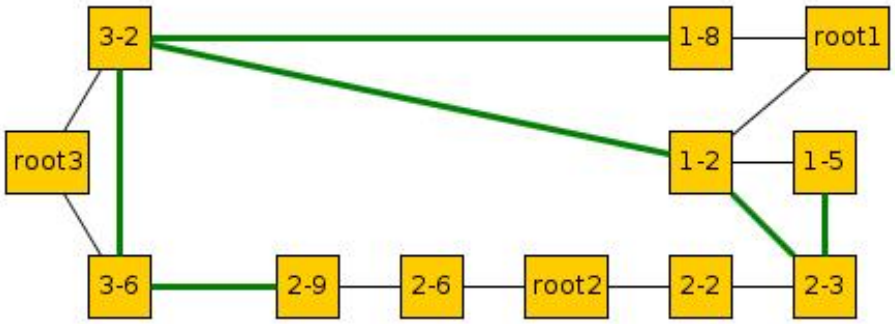


Abbildung 4.6: Sammelprozess: Ergebnis

Ergebnis

Das Verfahren liefert wie erhofft Verbindungen zwischen den Ausgangsnutzern. Zusätzlich zeigen sich häufig weitere Querverbindungen zwischen den Nutzern. Diese Verbindungen sind zwar bei anderen Verfahren vorhanden, jedoch werden sie nicht so deutlich sichtbar, da sehr viele anderen Nutzer ohne Querverbindungen vorhanden sind. Durch das Entfernen der Nutzer ohne Querverbindungen bleiben nur interessante Nutzer übrig und es ist möglich die Suchtiefe zu erhöhen. Eine höhere Suchtiefe sorgt wiederum für mehr Querverbindungen, diese können jedoch auch wieder die Übersicht beeinträchtigen. Gleichzeitig erhöht eine

tiefere Suche die Laufzeit des Sammelverfahrens deutlich. Wird keine Verbindung zwischen den Nutzern gefunden – entweder weil keine vorhanden oder weil die Suchtiefe zu gering ist – so werden alle gesammelten Nutzer wieder gelöscht.

4.2.5 Bewertung der Verfahren

Da die Verfahren deutliche Unterschiede in den Ergebnissen zeigen, sollen diese nun direkt miteinander Verglichen werden:

Das erste Verfahren aus Unterabschnitt 4.2.1 ermöglicht es, durch die Nutzung der Streaming-API, in kurzer Zeit viele Tweets zu sammel, jedoch sind die gesammelten Kommunikationen oftmals nicht vollständig und es gibt nur wenige vernetzte Nutzer.

Das zweite Verfahren (Unterabschnitt 4.2.2) orientiert sich an der Kommunikation der Nutzer, so dass sich ein Netzwerk ergibt und die Kommunikation zwischen den Nutzern vollständig erfasst wird. Jedoch bietet es keine Möglichkeit gezielt nach gut vernetzten Nutzern zu suchen. Die Anzahl dieser kann nur durch eine Erhöhung der Suchtiefe vergrößert werden, was allerdings zu Lasten der Übersichtlichkeit geht. Im Vergleich zum vorherigen Verfahren ist der Sammelvorgang langsamer, da die REST-API von Twitter verwendet wird.

Die Erweiterung des vorherigen Verfahrens in Unterabschnitt 4.2.3 ermöglicht es gezielt nach Themen zu suchen und dadurch die Suchtiefe zu erhöhen, so dass mehr gut vernetzte Nutzer gefunden werden können. Durch das vorherige Filtern der gesammelten Daten, werden nicht alle Kommunikationen vollständig erfasst.

Das letzte Verfahren (Unterabschnitt 4.2.4) erhöht die Anzahl der gut vernetzten Nutzer deutlich, indem alle anderen Nutzer aussortiert werden. Da erst mehrere komplette Nutzerbäume aufgebaut werden müssen, bevor diese gekürtzt werden können, benötigt dieses Verfahren mehr Zeit als die vorherigen. Um sicher Verbindungen zwischen den Nutzern zu finden, empfiehlt es sich die Suchtiefe zu erhöhen, was zusätzlich mehr Zeit benötigt. Der erhöhte Zeitaufwand wird durch gut verbundene Kommunikationsnetzwerke ausgeglichen.

	4.2.1	4.2.2	4.2.3	4.2.4
Schnelligkeit	+	~	~	-
Vollständige Kommunikation	~	+	-	+
Anzahl der gut vernetzten Nutzer	-	-	~	+

Tabelle 4.1: Vergleich der Verfahren.

Abschließend lässt sich zusammenfassen, dass das erste Verfahren nicht zum Sammeln von Kommunikationsnetzwerken geeignet ist. Das zweite Verfahren ermöglicht es relativ schnell Netzwerke zu sammeln. Das dritte Verfahren ermöglicht zwar eine höhere Suchtiefe, jedoch werden nicht alle Kommunikationen vollständig gesammelt, so dass es nur bedingt geeignet ist. Das letzte vorgestellte Verfahren benötigt viel Zeit zum Sammeln der Daten, jedoch werden dabei gezielt gut vernetzte Nutzer gesammelt. Jedoch gehen dabei schlecht vernetzte Nutzer verloren, die eventuell auch interessante Kommunikationsinhalte bereitstellen.

4.3 Darstellung des Graphen

Wie in Abschnitt 3.3 beschrieben, wird das Kommunikationsnetzwerk als Graph mit force-directed Layout dargestellt. Zur Darstellung wird das Visualisierungs-Toolkit Prefuse verwendet. Prefuse ermöglicht es einfach Daten zu laden und als Graph darzustellen. Für das Layout des Graphen werden mehrere Layouts zu Auswahl gestellt, die sich einfach für eigene Zwecke anpassen lassen.

Der Lade- und Visualisierungsvorgang von Prefuse ist in Abbildung 4.7 dargestellt. Wie in Abschnitt 4.1 beschrieben, werden die Daten aus dem Lucene-Repository in die interne Datenstruktur geladen. Die interne Datenstruktur dient anschließend als Rohdaten (*Raw Data*) für Prefuse.

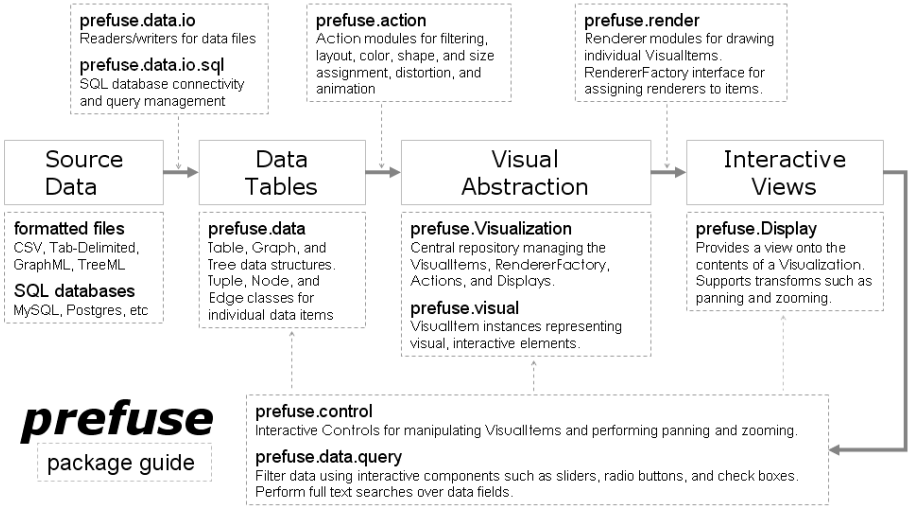


Abbildung 4.7: Prefuse [pre]

Zunächst wird ein Graph in Prefuse erstellt und diesem die Nutzer als Knoten und die Konversationen als Kanten zwischen diesen hinzugefügt, so dass *Data Tables* gefüllt wird. Zu jedem Knoten wird die Nutzer-ID als zusätzliches Attribute gespeichert, so dass alle weiteren Werte schnell aus der internen Datenstruktur ausgelesen werden können. Ähnlich werden für jede Kante die IDs der beiden Nutzer, die an der Konversation beteiligt, gespeichert, so dass auch für die Konversation schnell auf die interne Datenstruktur zugegriffen werden kann. Theoretisch ließen sich auch alle Informationen zu einem Nutzer bzw. zu einer Konversation in den Zusatzattributen speichern. Jedoch werden manche Operationen direkt auf der internen Datenstruktur ausgeführt und dabei deren Daten oder die Sicht darauf verändert. Dies würde dazu führen, dass die Daten in den Attributen auch verändert werden müssen. Um doppelte Datenhaltung zu vermeiden, wird in den Attributen somit nur ein Verweis auf die interne Datenstruktur gespeichert.

Die *Visual Structures* werden mithilfe sogenannter *Actions* aus den *Data Tables* erstellt. In den „Actions“ werden Füll- und Rahmenfarben für alle dargestellten Objekte festgelegt.

Außerdem wird das Layout des Graphen festgelegt. Bevor ein force-directed Layout für den Graph festgelegt wird, wird zuerst ein zufälliges Layout verwendet. Dies beschleunigt die anschließende Berechnung des Layouts, da die Knoten bereits im Raum verteilt sind und nicht mehr alle im Zentrum liegen. (Prefuse setzt alle Knoten beim Hinzufügen ins Zentrum) Das verwendete force-directed Layout ist, wie in *subsec:layout* beschrieben, angepasst, so dass Knoten mit hohem Grad eine größere Entfernung voneinander haben.

Zusätzlich zeigt sich, dass um Knoten, die mit vielen Knoten mit Grad eins verbunden sind, eine dichte Traube aus Knoten bildet. Um diese Traube aufzulockern wird, wenn einer der Knoten eine Grad $>=$ 20 und der andere $<=$ 2 hat, die Länge der Feder minimal erhöht. Außerdem wird die Federlänge verdoppelt, wenn beide Knoten nicht gemeinsam an einem eingezeichneten Thema beteiligt sind. Die Berechnung der Federlänge erfolgt somit wie in Algorithmus 4.1 beschrieben.

Algorithmus 4.1 Berechnung der Federlänge

```
procedure getSpringLength(edge)
    source ← sourceNode(edge)
    target ← targetNode(edge)
    minDegree ← min(degree(source), degree(target))
    maxDegree ← max(degree(source), degree(target))
    if maxDegree >= 20 and minE <= 2 then
        min ← 50
    else
        min ← 30
    end if
    normalized ← normalize(minE)
    length ← calculateLenght(normalized, min, 200)
    if inSameTopic(source, target) then
        factor ← 1
    else
        factor ← 2
    end if
    return length * factor
end procedure

procedure calculateLenght(n, min, max)
    return (n * (max − min)) + min
end procedure
```

Um Knoten, wie in Abschnitt 3.7 beschrieben, bei Bedarf ausblenden zu können, werden mehrere Filter entwickelt, die nacheinander prüfen, ob ein Knoten ausgeblendet wird.

Um aus den *Visual Structures Interactive View* erstellen zu können, werden mehrere *Renderer* definiert werden. Diese legen fest wie Kanten und Knoten dargestellt werden. Die Größe der Knoten und Kanten wird wie in Unterabschnitt 3.3.1 anhand der Tweets eines Nutzer

bzw. einer Kommunikation berechnet. Um den Anteil der Nutzer an einer Kommunikation darstellen zu können wird jede Kanten an ihren Enden mit einem Dreieck versehen, so dass jede Kante einen Doppelpfeil darstellt. Die Größe des jeweiligen Pfeils stellt den Anteil des Nutzers an der Kommunikation und somit die Richtung der Kommunikation dar. Für eine Kante (A,B) zeigt die Pfeilspitze, die näher bei A liegt, den Anteil der Tweet, die B an A versendet hat, an.

4.4 Extraktion und Darstellung der Themen

Im folgenden wird die Umsetzung des Konzepts aus Abschnitt 3.5 beschrieben. Zuerst werden mit Hilfe von Lucene häufige Stichworte aus allen Tweets extrahiert, diese werden anschließend für LDA verwendet. Das Pakets Mallet [mal] bietet eine umfangreiche Bibliothek an Werkzeugen zum Analyse von natürlicher Sprache und zur Klassifikation von Dokumenten. Mit dessen Hilfe werden sowohl Stemming, als auch LDA durchgeführt. Aus den erkannten Themengebieten werden anschließend Tag Clouds berechnet. Werden diese ausgewählt, so wird die Auswahl in das Netzwerk eingezeichnet.

4.4.1 Tag Cloud

In Abschnitt 2.2 wurde bereits die grundlegende Idee einer Tag Cloud beschrieben, hier soll nun auf die konkrete Implementierung eingegangen werden.

Um eine Thema als Tag Cloud darstellen zu können, wird zuerst für jedes Schlagwort des Themas anhand dessen Gewichtung die Schriftgröße wie folgt berechnet. Hierzu wurde die Formel aus [wik12c] verwendet und an die Bedürftnisse entsprechend angepasst.

$$s_i = \left\lceil \left(\frac{w_i - w_{min}}{weight_{max} - weight_{min}} \right) (f_{max} - f_{min}) + f_{min} \right\rceil$$

s_i: anzuzeigende Schriftgröße

f_{max}: maximale Schriftgröße

f_{min}: kleinste Schriftgröße

w_i: Gewichtung des Schlagwortes

w_{min}: insgesamt kleinste Gewichtung

w_{max}: insgesamt höchste Gewichtung

Anschließend werden alle Schlagwörter in einem Rechteck mit vorgegebener Größe platziert. Die Höhe und Breite eines Schlagwortes lassen sich über die Schriftgröße, Schriftart und Länge des Wortes berechnen. Das erste Schlagwort wird in der linken oberen Ecke des Rechtecks platziert. Die restlichen Schlagwörter werden wie folgt platziert:

1. Platziere das Schlagwort rechts neben dem zuletzt gesetzten Schlagwort.

2. Ragt das Schlagwort über den Rand des Rechtecks hinaus, so wird es in der nächsten Zeile am linken Rand des Rechtecks positioniert.

Reicht der Platz innerhalb des Rechtecks nicht, um alle Schlagwörter zu positionieren, so wird der gesamte Vorgang mit einer kleineren maximalen und minimalen Schriftgröße wiederholt.

4.4.2 Visualisierung im Graph

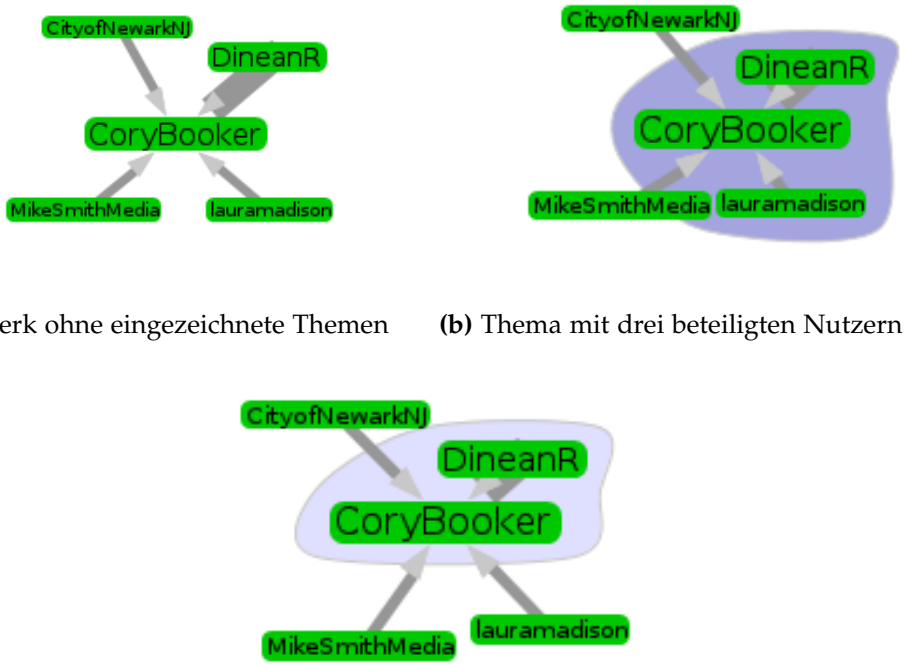
Es ist sowohl möglich, eine komplette Tag Cloud auszuwählen, als auch ein einzelnes Schlagwort innerhalb einer Tag Cloud. Wird die gesamte Tag Cloud ausgewählt, so werden Nutzer, die Nachrichten geschrieben bzw. erhalten haben, die an diesem Thema beteiligt sind, hervorgehoben. Wird nur ein einzelnes Schlagwort ausgewählt, so werden nur jene Nutzer hervorgehoben, die Nachrichten geschrieben bzw. erhalten haben, die an dem Thema beteiligt sind und das ausgewählte Schlagwort enthalten. Es kann durchaus möglich sein, dass es weitere Nutzer gibt, in deren Nachrichten das ausgewählte Schlagwort vorkommt. Sind diese Nachrichten jedoch nicht an der Bildung des Themas beteiligt, so werden die zugehörigen Nutzer auch nicht hervorgehoben.

Um die Themen bzw. Stichwörter mit Hilfe von Prefuse in den Graph einzeichnen zu können, wird das in Abschnitt 4.3 beschriebene Verfahren erweitert: Es werden ein zusätzliches Layout und ein Renderer zum Einzeichnen der Themen und Schlagwörter eingeführt. Soll ein Thema bzw. Schlagwörter gezeichnet werden, so wird zunächst das Layout festgesetzt. Hierzu wird aus der Menge an beteiligten Nutzern bzw. deren Knoten im Graph die konvexe Hülle bestimmt. Eingezeichnet wird diese anschließend durch den Renderer, indem mit Hilfe von kubischen Splines (s. Unterabschnitt 2.3.1) die Punkte der konvexen Hülle verbunden werden. Somit werden alle beteiligten Nutzer durch eine farbliche Markierung umrahmt. Abbildung 4.8b und Abbildung 4.8c zeigen diese Markierung.

Da es vorkommen kann, dass sich Nutzer, die nicht hervorgehoben werden sollen, innerhalb dieser Markierung befinden, wird zusätzlich das Layout des Graphen verändert: Knoten, die an den selben Themen beteiligt sind, werden näher beieinander dargestellt. Da die abstoßenden Kräfte zwischen den Knoten global für alle Knoten berechnet wird, werden die Federlänge für die betreffenden Kanten verkürzt. In Abbildung 4.8b und Abbildung 4.8c ist zu sehen, dass die beteiligten Knoten im Vergleich zu Abbildung 4.8a näher beisammen liegen.

4.5 Darstellung von Tweets

Tweets werden wie in Abschnitt 3.4 beschrieben dargestellt. Um für Konversationen eine Darstellung wie in Abbildung 3.1 zu erhalten, wird die Anzeige der Tweets in drei Bereiche von fester Breite unterteilt: Im ersten und dritten Bereich wird der Twitter-Name des Absenders des Tweets dargestellt. Die Breite wird durch die Länge des Namens vorgegeben. Wird eine Nachricht vom Nutzer, der die erste Nachricht zur Konversation beigetragen hat,



(a) Netzwerk ohne eingezeichnete Themen **(b)** Thema mit drei beteiligten Nutzern

(c) Thema mit drei beteiligten Nutzern

Abbildung 4.8: Einzeichnen von Themen in den Graph und Anpassen des Layouts

dargestellt, so wird dessen Name im ersten Bereich dargestellt. Der dritte Bereich bleibt leer. Dementsprechend bleibt der erste Bereich leer, wenn eine Nachricht des anderen Nutzers dargestellt wird, der Name steht im dritten Bereich. Im mittleren Bereich steht immer das Datum und der Inhalt des Tweets. Diese nehmen die restliche Breite ein. Sollte der Platz nicht ausreichen, so wird der Inhalt des Tweets auf mehrere Zeilen umgebrochen. Somit ergibt sich eine Anzeige der Tweets wie in Abbildung 4.9.

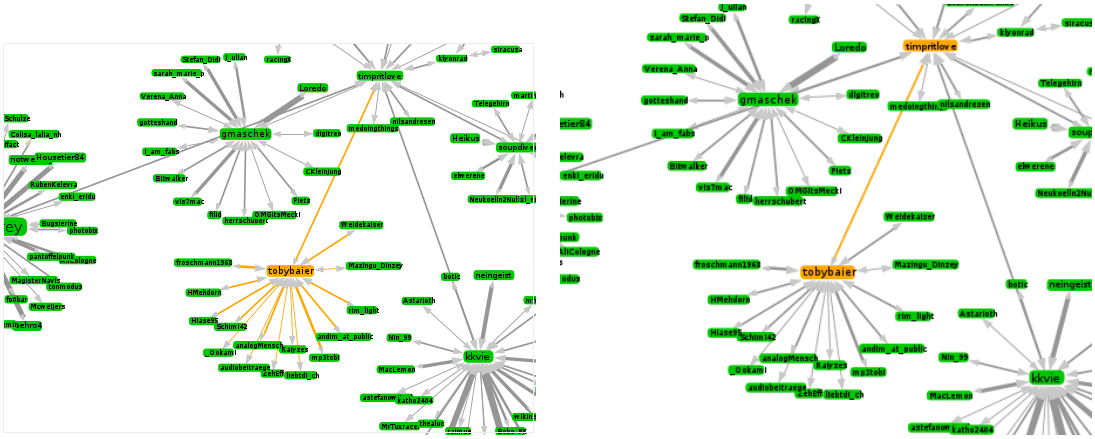
Kay(kkvie)	**2012-12-31 01:11:00**	@timpritlove Also vom @MetalabVienna gibt es einen Flo (@Overflo) und einen Wizard (@Wizard23), die waren beide beim Hardware Hacking aktiv.	
Kay(kkvie)	**2013-01-03 01:21:00**	@timpritlove Hat sich das aufgeklärt?? Auf meinen Tweet kam nie eine Reply. Thx.	
	2013-01-03 11:00:00	@kkvie Bisher nicht.	**Tim Pritlove(timpritlove)**
Kay(kkvie)	**2013-01-03 14:27:00**	@timpritlove War es viell. einer der beiden von mir genannten/hast Du sie deswegen kontaktiert? (Weiß leider nicht, woher das Ulm kam..)	

Abbildung 4.9: Anzeige der Tweets

4.6 Auswählen von Objekten im Graph

Objekte können im Graph durch anklicken ausgewählt werden. Es kann immer nur ein Objekt gleichzeitig ausgewählt sein. Um deutlich zu machen, welches gerade ausgewählt ist

und was somit in der Tweet-Anzeige sichtbar ist, wird das ausgewählt Objekt im Graphen hervorgehoben.



(a) Hervorheben eines Knoten **(b)** Hervorheben einer Kante

Abbildung 4.10: Hervorheben der Auswahl

Wird ein Knoten ausgewählt, so wird dieser, wie in Abbildung 4.10a, farblich markiert. Zusätzlich werden alle Kante, die zu diesem Knoten führen bzw. von diesem abgehen, ebenfalls markiert.

Beim Auswählen einer Kante wird diese, wie in Abbildung 4.10b zu sehen, ebenfalls markiert. Zusätzlich wird der Start- und Zielknoten der Kante markiert.

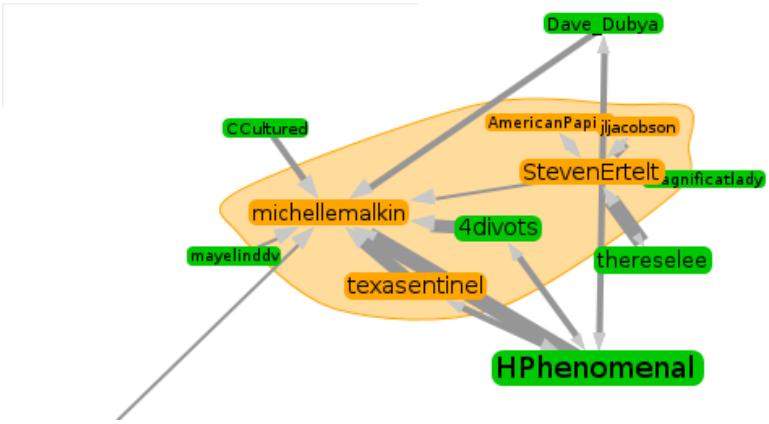


Abbildung 4.11: Hervorheben eines Themas

Wird ein Thema ausgewählt, so wird dieses markiert. Da es, wie in Abschnitt 3.5 beschrieben, trotz Anpassung des Layouts vorkommen kann, dass sich Knoten innerhalb der Markierung des Themas befinden, die nicht dazu gehören, werden alle Knoten des Themas ebenfalls markiert (s. Abbildung 4.11). Dadurch lässt sich eindeutig das Thema und die zugehörigen Knoten bzw. Nutzer erkennen.

5 Anwendungsfälle

Der praktische Nutzen der entwickelten Werkzeuge wird im folgenden anhand von Anwendungsfällen beschrieben.

5.1 Support-Accounts

Viele Unternehmen bieten inzwischen Support per Twitter an. Um die Qualität des Support und die Probleme der Kunden schnell bewerten zu können, können diese Daten mit den entwickelten Werkzeugen gesammelt und angezeigt werden.

Im folgenden Fall wird der Mobilfunkanbieter Fonic (@fonic_de) betrachtet. Mit dem Verfahren aus Unterabschnitt 4.2.2 wurde mit Suchtiefe eins Daten gesammelt. Eine Suchtiefe von eins wurde bewusst gewählt, um nur die Kommunikation des einen Accounts zu erhalten.

Nach dem Laden sieht man, wie in Abbildung 5.1, das Kommunikationsnetzwerk des Nutzers @fonic_de und die behandelten Themen. Zusätzlich ist der Nutzer @fonic_de ausgewählt, so dass dessen Tweets angezeigt werden. Es finden sich einige Nutzer, die den Stand ihrer Bestellung abfragen wollen. Weitere Nutzer haben per E-Mail oder Kontaktformular keine Antwort auf ihre Anfrage erhalten oder die Hotline nicht erreichen. In allen Fällen, in denen Daten des Nutzers benötigt werden, fordert der Fonic-Support die Nutzer auf ihm ihre Rufnummer per Direct Message (s. Unterabschnitt 2.1.6) zu senden. Der weitere Kontakt finden anschließend per Direct Message statt, so dass er nicht mehr eingesehen werden kann. Werden keine persönlichen Daten benötigt, so wird in den meisten Fällen versucht dem Nutzer zu helfen oder ihn an die entsprechenden Stellen zu verweisen. Dass der Support mindestens in einigen Fällen erfolgreich war, kann man daran sehen, dass sich mehrere Nutzer für die schnelle Hilfe bedankt haben.

5.2 Landesparteitag und Dreikönigstreffen der FDP

Das Dreikönigstreffen der Liberalen fand jährlich am 6. Januar im Stuttgarter Staatstheater statt. Es stellt den politischen Jahresauftakt der Freien Demokratische Partei (FDP) dar. Am Tag zuvor fand der 109. Ordentlicher Landesparteitag der FDP Baden Württemberg statt.

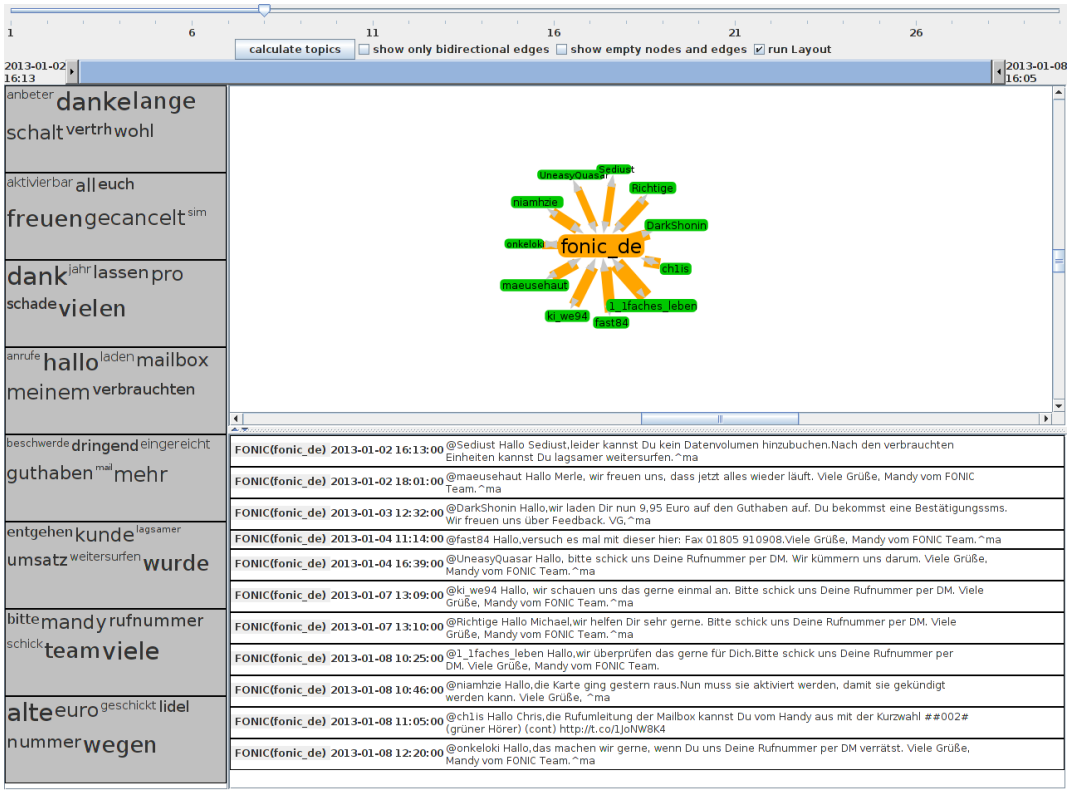


Abbildung 5.1: Ansicht @fonic_de

Auf Twitter wurden Berichte über das Ereignis mit dem Hashtag #3k13 gekennzeichnet. Über die Twitter-Suche[1] können so leicht Tweets und zugehörige Nutzer gefunden werden.

5.2.1 Sammeln der Daten ausgehend von einem Nutzer

Über die Twitter-Suche stößt man auch auf den Nutzer @the_necrosis[2]. Er war an beiden Tagen in Stuttgart vor Ort und twitterte live über den Landesparteitag und das Dreikönigstreffen. Um die Kommunikationen vor, während und nach dem Ereignis einzufangen, wurde am 7. Januar gegen 13 Uhr ein Sammelprozess wie in Unterabschnitt 4.2.2 ausgehend von @the_necrosis mit einer Suchtiefe von 2 gestartet. Das Sammeln der Daten dauerte bis ca. 17:30 Uhr. Die gesammelten Daten erstrecken sich über einen Zeitraum vom 31.12.2012 bis zum 7.01.2013

Nach dem Laden der Daten sieht man ein großes Netzwerk (Abbildung 5.2). Wie in Abschnitt 4.2.2 beschrieben, zeigt sich, dass sich am Rand des Netzwerkes um viele Nutzer um einzelne Nutzer gruppieren. Die behandelten Themen reichen von Stuttgart 21 (Hashtag

[1] https://twitter.com/search/realtime?q=%233k13
[2] https://twitter.com/the_necrosis

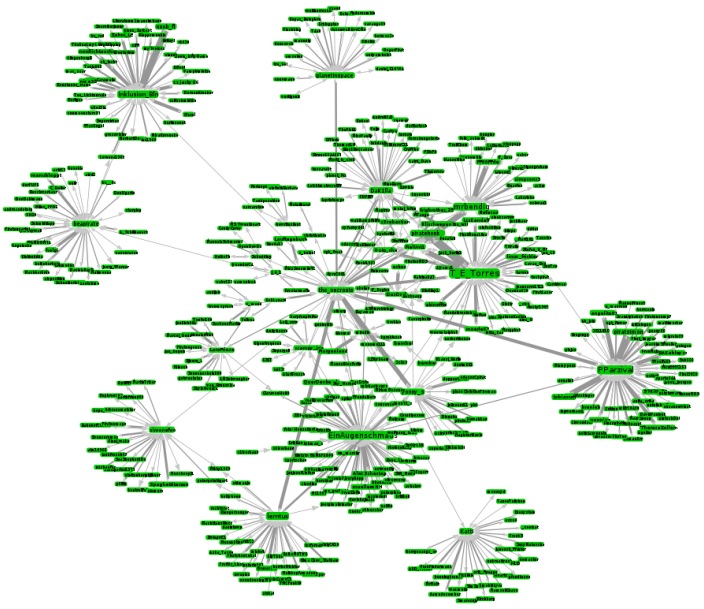


Abbildung 5.2: Netzwerk @the_necrosis

#S21) und Glückwünschen für das neue Jahr (Abbildung 5.5a) bis zu der Piratenpartei und der FDP. Der Flughafen Berlin Brandenburg „Willy Brandt“ stellt ein ständig vorkommenden Thema dar. Sein Eröffnungstermin wurde wiederholt wegen Problemen verschoben, zuletzt am 7. Januar 2013. In Verbindung mit dem Flughafen Berlin Brandenburg taucht auch gelegentlich das Projekt Stuttgart 21 am Stuttgarter Hauptbahnhof auf. Die Glückwünsche zum neuen Jahr sind verständlich, da der Jahreswechsel im Beobachtungszeitraum liegt. Wie erhofft ist auch die FDP ein Thema auf Twitter. Zusätzlich sind auch die Schlagwörter „lpt“ (Landesparteitag) und „stuttgart“ zu sehen.

Interessant ist, dass in den gesammelten Daten die Piratenpartei ein größeres Thema als die FDP darstellt. Verringert man die Anzahl der Themen auf zwei, so erscheint in einem der Themen bereits das Schlagwort „piraten“ auf. Die FDP (Schlagwort „fdp“) erscheint, wenn die Anzahl der Themen auf drei gesetzt wird, gemeinsam mit den Piraten in einer Tag Cloud. Insgesamt betreffen jedoch mehr Tweets die Piratenpartei als die FDP. Andere Parteien, wie die Sozialdemokratische Partei Deutschlands (SPD) oder die Christlich Demokratische Union Deutschlands (CDU), tauchen in den Tag Clouds erst auf, wenn man die Anzahl der Themen auf ca. 15-20 erhöht. Dass die Piratenpartei so ungewöhnlich oft auftaucht, lässt sich mit einem Blick auf das Twitter-Profil von @the_necrosis klären. Dort beschreibt er sich selbst als „Schwerhörig, ADHD, BTA, Pirat, Biologe“. Betrachtet man die Nutzer, mit denen @the_necrosis direkt kommuniziert, so fällt auf, dass sich knapp die Hälfte in ihrem Profil als Pirat bzw. Piratin, also Mitglieder der Piratenpartei, bezeichnen. Weiterhin fällt auf, dass ein weiterer Teil (ca. 20%) sich selbst als gehörlos, schwerhörig oder taubstumm bezeichnen. Betrachtet man mit diesem Wissen die Tag Clouds, so fällt das Schlagwort „untertitel“ auf. Lässt man dieses in den Graphen einzeichnen (Abbildung 5.3), so erkennt man, dass @the_necrosis (rechts im Bild) nicht daran beteiligt ist, sondern ausschließlich

Nutzer um @EinAugenschmaus. Bei @EinAugenschmaus handelt es sich um die gehörlose Lippenleserin Julia Probst, die bekannt wurde, als sie zur Fußball-EM 2012 über Twitter verbreitete was Spieler und Trainer sagten, auch wenn dies für Mikrofone der Reporter zu leise war. Die Diskussionen zum Schlagwort „untertitel“ drehen sich hauptsächlich darum, dass die GEZ-Gebühren für die Untertitelung von Filmen im Fernsehen eingesetzt werden sollen und erstrecken sich über den gesamten Beobachtungszeitraum. Somit hat diese

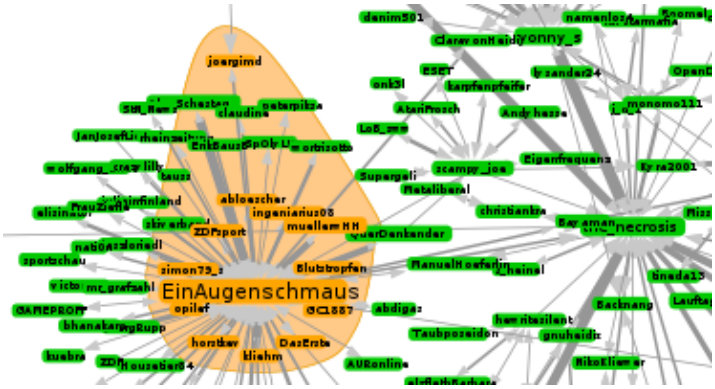
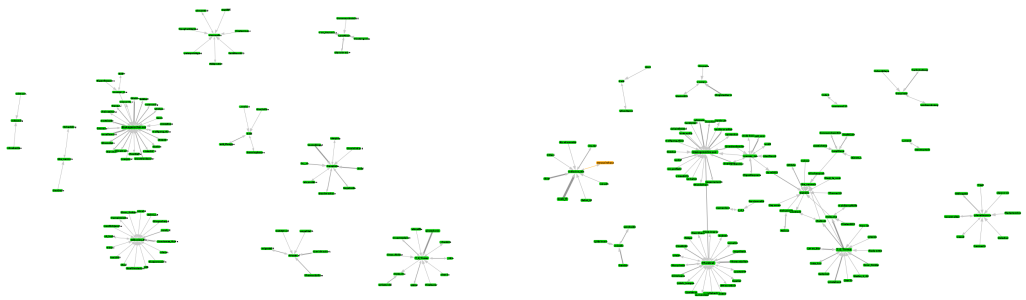


Abbildung 5.3: Schlagwort „untertitel“

Diskussion erstmal weniger mit der FDP zu tun. Betrachtet man die Verbindung zwischen @the_necrosis und @EinAugenschmaus, so sieht man, dass es dabei um die Anwesenheit bzw. das Fehlen von Gebärdensprachdolmetscher auf dem Dreikönigstreffen ging.

Im folgenden wird nun die Entwicklung der Themen und des Netzwerks über den Beobachtungszeitraum beschrieben. Hierfür wird jeweils ein Zeitabschnitt von einem Tag betrachtet:



(a) Netzwerk am 31.12.2012 **(b)** Netzwerk am 01.01.2013

Abbildung 5.4: Kommunikationsnetzwerk

31.12.2012: Kommunikationen finden hauptsächlich unter wenigen Nutzern statt, wie man auch auf Abbildung 5.4a sieht. Der Flughafen Berlin Brandenburg, sowie die FDP und die Piratenpartei spielen eine kleine Rolle. Es dominieren Glückwünsche zum neuen Jahr.

01.01.2013: Es bilden sich größere, zusammenhängende Netzwerke (s. Abbildung 5.4b), für die jedoch keine offensichtlichen Themen erkennbar sind. Es dominieren weiterhin Glückwünsche zum neuen Jahr.

02.01.2013: Die Netzwerke zerfallen wieder in kleinere Teile. Die Zahl der Neujahrsglückwünsche geht zurück und es sind Diskussionen über die Themen der Piratenpartei, sowie der FDP zu finden, jedoch spielt das Dreikönigstreffen noch keine Rolle. Eine Tag Cloud mit den Schlagworten „israel“ und „hama“ (Hamas, durch Stemming verkürzt) fällt ins Auge. Beide Stichworte kommen nur in einer Diskussion zweier Nutzer über den israelisch-palästinensische Konflikt. Weitere Nutzer sind daran nicht beteiligt.

03.01.2013: Die Tag Cloud aus Abbildung 5.5b fällt zusammen mit anderen Schlagworten wie „Diskriminierung“, „Geschlecht“ und „Kernthema“ auf. Dahinter verstecken sich Diskussionen über Sexismus und Feminismus innerhalb der Piratenpartei und ob Feminismus ein Kernthema der Piraten darstellen soll. Des weiteren fällt eine geplante Aktion der Piraten zum Dreikönigstreffen auf: Es soll ein eigenes Plakat im Schlossgarten auf einem Zugangsweg zum Dreikönigstreffen aufgestellt werden. Im weiteren Verlauf stellt sich heraus, dass dies aus Mangel an Teilnehmern nicht stattfinden kanne.

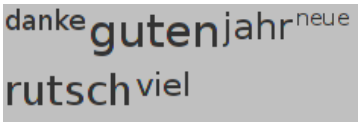
04.01.2013: Es sind keine relevanten Themen zu finden. Jedoch bilden sich erneut Netzwerke mit vielen Nutzern und die Anzahl der kleinen, isolierten Netzwerke sinkt.

05.01.2013: Die FDP und das Schlagwort „lpt“ tauchen passend zum Landesparteitag auf. Zusätzlich finden sich mehrere Diskussion zu verschiedenen Themen: Zu den Schlagworten „gemeinsam“, „Jahre“ und „lernen“ gibt es Diskussionen zum Schulsystem und wie lange Kinder gemeinsam lernen sollen. Zu den Themen in der Tag Cloud in Abbildung 5.5c finden sich weitere Diskussionen zum Bildungssystem, zur Wählbarkeit der verschiedenen Parteien und zum Flughafen Berlin Brandenburg.

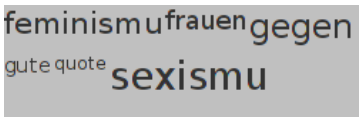
06.01.2013: Zum Dreikönigstreffen prägen Führungsqualitäten der Parteispitze der FDP und ob die FDP für Freiheit steht die Diskussionen. Gleichzeitig überlegen Mitglieder der Piratenpartei, ob FDP-Wähler mit dem Thema Freiheit zum Wählen ihrer Partei gebracht werden können. Abgesehen von den politischen Themen wurde die Qualität des Live-Streams und das fehlen von kostenlosem WLAN vor Ort bemängelt.

07.01.2013: Am 7. Januar wurde bekannt, dass sich die Eröffnung des Flughafen Berlin Brandenburg auf 2014 verschieben wird, so dass sich einige Nutzer hierüber auslassen. Zusätzlich finden sich am Montag nach dem Dreikönigstreffen keine besondere Themen mehr. Nutzer wünschen sich gegenseitig einen guten Morgen und unterhalten sich über Kaffee. Das große Kommunikationsnetzwerk, das sich gebildet hatte, zerfällt in einige kleine Netzwerke.

Es zeigt sich, dass die Auswahl des Nutzers @the_necrosis dazu führte, dass das Thema Piratenpartei an vielen Diskussionen beteiligt war. Am 3. Januar finden sich hauptsächlich Themen zur Piratenpartei und parteiinterne Diskussionen. Im Gegensatz dazu gibt es am 5. und 6. Januar mehr Diskussionen über die FDP und deren Themen. Leider ergibt sich häufig nicht ein zusammenhängendes Netzwerk aus Nutzern, sondern mehrere. Um diese sinnvoll analysieren zu können, muss häufiger die automatische Layoutberechnung abgeschaltet



(a) Tag Cloud zu Neujahr



(b) Tag Cloud zu Feminismus

(c) Tag Cloud zu den Parteien

Abbildung 5.5: Tag Clouds

werden. Insgesamt ist es einfach möglich sich einen Überblick über die behandelten Themen und über die Entwicklung der Kommunikationsstrukturen zu machen.

5.2.2 Sammeln der Daten ausgehend von mehreren Nutzern

Zum Vergleich der beiden Sammelprozesse wurde über einen ähnlichen Zeitraum wie oben ein Sammelprozess ausgehend von mehreren Nutzern (s. Unterabschnitt 4.2.4) gestartet. Hierfür wurde wieder @the_necrosis , sowie @JuLisBW (Junge Liberale Baden-Württemberg), @FDPBW (FDP Baden-Württemberg), @PiratenBW (Piratenpartei) Baden-Württemberg) und @Norberthense (Norbert Hense, Selbstbeschreibung: Pirat, JuPi, Schalker, TV Junkie). Norbert Hense wurde über @the_necrosis gefunden. Der Sammelvorgang wurde am 6. Januar 2013 um 18:00 Uhr gestartet und endete am 7. Januar um 4:40 Uhr. Der erfasste Zeitraum erstreckt sich vom 30.12.2012 bis zum 06.01.2013.

Man erhält aus den Daten ein sehr dichtes Netzwerk mit vergleichsweise weniger Nutzern. Dieses ist in Abbildung 5.6 zu sehen ist. Betrachtet man einzelne Tage, so sind deutlich weniger Nutzer und Verbindungen zwischen diesen vorhanden. Trotzdem bleibt meistens ein großes zusammenhängendes Netzwerk vorhanden und es spalten sich nur wenige kleine Teile ab. Interessanterweise sind die Twitter-Nutzer @JuLisBW, @FDPBW und @PiratenBW im zweiten Teil des Sammelprozesses wieder entfernt worden, so dass sie im Kommunikationsnetzwerk nicht mehr vorhanden sind. Dadurch ergibt sich ein Netzwerk, das hauptsächlich aus Mitgliedern der Piratenpartei, Nutzern die der Piratenpartei nahe stehen und anderen technik- bzw. internetaffine Nutzern besteht. Dies zeigt sich auch in den Themengebieten: So finden sich z. B. am 31.12. Diskussionen über die Größe der verbauten Festplatten, LED-Lampen und Quecksilber in Energiesparlampen und über die Musikabspielsoftware Amarok. Im Vergleich zu Unterabschnitt 5.2.1 wurden nur wenige Neujahrsglückwünsche zwischen den Nutzern ausgetauscht. Stattdessen wird am 01.01. über die Twitter-Bots (Computerprogramme, die automatisch anderen Nutzern folgen) unter den eigenen Followern, Zeichentrickserien und Urzeitkrebse aus dem Yps-Heft diskutiert. Am 05.01. wird über die FDP, den Landesparteitag und das Schulsystem diskutiert. Am 06.01. finden sich praktisch keine Diskussion über Inhalte der FDP.

Der Versuch ausgehend von mehreren Nutzern Diskussionen über die FDP, den Landesparteitag und das Dreikönigstreffen zu finden, ist an der Auswahl der Nutzer gescheitert. Vermutlich hätte Twitter-Accounts von FDP-Mitgliedern anstatt von Organisationen gewählt werden sollen, da die Accounts der Organisationen weniger mit anderen Nutzern in Kontakt treten. Trotzdem zeigt sich, dass das Sammelverfahren wie gewünscht funktioniert und man

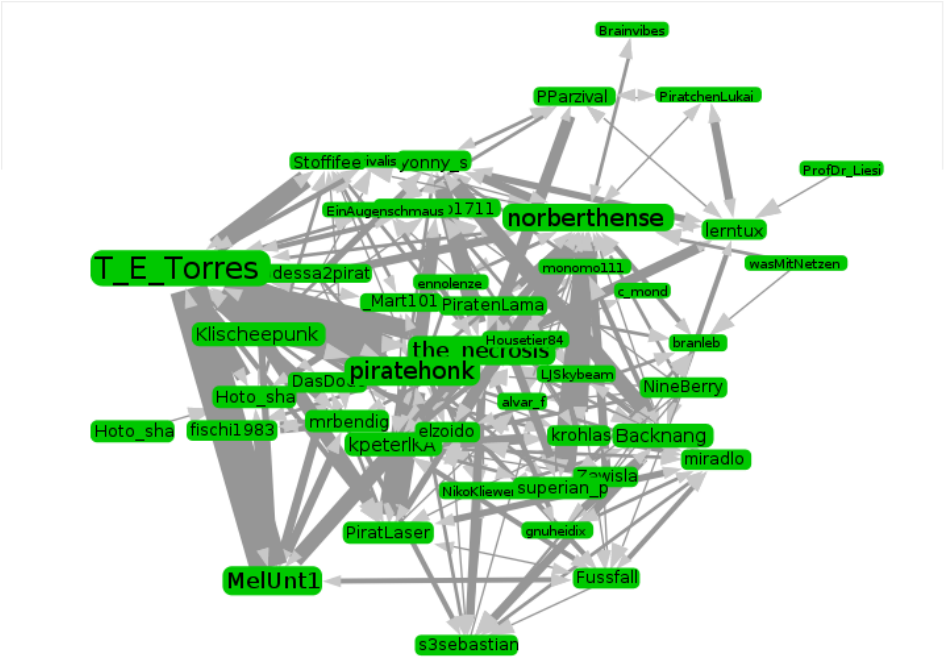


Abbildung 5.6: Sammelvorgang mit mehreren Nutzern

Informationen über die behandelten Themen erhält. Wird der ausgewählte Zeitraum nicht zu klein gewählt, so bleibt auch ein großes, zusammenhängendes Kommunikationsnetzwerk erhalten und es bilden sich nur selten und dann auch nur kleine getrennte Netzwerke. Die Nutzer innerhalb des Netzwerks ändern sich jedoch mit der Zeit. Somit muss zur Analyse des Netzwerkes die Berechnung des Layouts nicht angehalten werden.

6 Zusammenfassung und Ausblick

Im Rahmen dieser Diplomarbeit wurde ein Konzept zum Extrahieren und Visualisieren von Kommunikationsnetzwerken im Sozialen Netzwerk Twitter entwickelt. Die entstandenen Werkzeuge ermöglichen es Kommunikationsnetzwerke in Twitter zu finden und zu analysieren. Eine Trennung des Sammelwerkzeugs von der Visualisierung ermöglichte es verschiedene Konzepte zum Extrahieren von Daten aus Twitter zu erstellen. Zwei der Konzepte stellen sich als verwendbar heraus: Die Extraktion des Kommunikationsnetzwerks ausgehend von einem Nutzer und die Suche von Verbindungen zwischen mehreren Nutzern. In den Anwendungsfälle zeigt sich, dass beide Konzepte auch in realen Szenarien gute Ergebnisse liefern.

Die Visualisierung des Sozialen Netzwerks orientiert sich an bereits vorhanden und bewährten Lösungen: Es wird ein Force-directed Layout mit kleinen eigenen Modifikationen für das Layout des Netzwerks eingesetzt. Die Kommunikationen zwischen Nutzern werden mit Hilfe von Pfeilen darstellt, die Hinweise zur hauptsächlichen Kommunikationsrichtung bieten. Zusätzlich werden aus den Tweets Themengebiete extrahiert und diese mit Hilfe von Tag Clouds dargestellt. Die gefundenen Themengebiete, wie auch einzelne Schlagworte, können in das Netzwerk eingezeichnet werden und beeinflussen dessen Layout. Sowohl für Nutzer als auch für Kommunikationen und Themen können die zugehörigen Tweets betrachtet werden, so dass man einen Überblick über die Diskussionen erhält.

In den Anwendungsfällen zeigt sich, dass das entstandene Konzept in realen Situaten zur Analyse von Kommunikationnetzwerken eingesetzt werden kann. Die entwickelten Werkzeuge ermöglichen die behandelten Themen, sowie die Struktur des Netzwerks in verschiedenen Zeitabschnitten zu analysieren.

6.1 Kommunikation auf Twitter

Wie erhofft lassen sich auf Twitter große, zusammenhängende Kommunikationsnetzwerke finden. Diese bilden sich jedoch häufig spontan und zu bestimmten Themen und sind nicht über einen längeren Zeitraume vorhanden. Aus den Erfahrungen in dieser Arbeit lässt sich schließen, dass es kleine Netzwerke aus Nutzern gibt, die sich regelmäßig untereinander austauschen. Somit haben die meisten aktiven Twitter-Nutzer eine beschränkte Anzahl an Kontakten, mit denen sie oft kommunizieren. Zusätzlich zu diesen festen Kontakten bildet sich weitere Kommunikation durch spontane Reaktionen von Nutzern auf Tweets. Aus diesen Reaktionen können sich Diskussionen mit weiteren Nutzern ergeben. Daraus bilden sich ebenfalls Kommunikationsnetzwerke. Diese Netzwerke ergeben sich für den

Zeitraum der Diskussion und verschwinden danach wieder. Für beliebte Themen können sich jedoch leicht große Netzwerke bilden. Auf Twitter gibt es somit kleine stabile Netzwerke. Zu bestimmten Themen können sich aus diesen jedoch große Netzwerke bilden.

Neben natürlichen Personen sind auch Unternehmen und Organisationen auf Twitter vertreten. Deren Kommunikation unterscheidet sich oftmals von der natürlicher Personen. Zum einen gibt es Service- oder Support-Accounts. Deren Kommunikation ergibt sich weitgehend aus Anfragen von Nutzern und beschränkt sich darauf diesen zuhelfen. Somit ergeben sich selten Diskussionen mit mehreren beteiligten Nutzern. Zum anderen gibt es Accounts, die weitgehend zum Verteilen von Informationen genutzt werden. Diese beschränken sich meistens auf das Schreiben von Tweets. Antworten auf diese werden meist nicht beachtet und somit betreiben diese keine oder nur kaum Kommunikation mit anderen Nutzern. Twitter-Accounts von Unternehmen und Organisationen sind somit in den meisten Fällen nicht in größeren Kommunikationsnetzwerken zu finden, da sie sich nicht an Diskussionen beteiligen.

6.2 Ausblick

Während der Durchführen der Diplomarbeit kamen verschiedene Ideen auf, wie das entstandenen Konzept erweitert und verbessert werden kann. Jedoch konnten nicht alle Ideen im Rahmen dieser Diplomarbeit umgesetzt werden. Im folgenden werden deshalb Ideen zur Erweiterung des Konzepts beschrieben:

6.2.1 Interaktives Sammeln von Daten

Im entwickelten Konzept wird entweder die gesamte Kommunikation eines Nutzers gesammelt oder es wird nach vorgegebenen Themen gefiltert. Eine Bewertung der Daten findet während des Sammelvorgangs nur sehr begrenzt statt. In einer Erweiterung des Konzept können die Daten während des Sammelvorgangs vom Benutzer bewertet werden. Dadurch kann verhindert werden, dass die Kommunikation uninteressanter Nutzer weiter verfolgt wird. Gleichzeitig können interessante Nutzer hervorgehoben werden und so der Sammelvorgang in eine bestimmte Richtung geführt werden. Da ein Sammelvorgang über mehrere Stunden bis hin zu Tagen oder Wochen dauern kann, kann der Benutzer diesen nicht ständig begleiten, so dass ein geeignetes Konzept zur Steuerung des Vorgangs gefunden werden muss.

6.2.2 Automatisiertes Erkennen interessanter Nutzer

Das bisherige Konzept basiert darauf, dass bereits Nutzer bekannt sind, deren Daten gesammelt werden sollen. Die Auswahl der Nutzer beeinflusst deutlich die Brauchbarkeit der gesammelten Daten. Um die Auswahl der Nutzer zu verbessern, kann ein automatisierter

Prozess Tweets zu vorgegebenen Themen sammeln, die beteiligten Nutzer und deren weitere Kommunikation bewerten und daraus eine Liste an potenziell interessanten Nutzern erstellen. Diese können entweder direkt für einen Sammelvorgang verwendet oder zuvor zusätzlich durch einen Nutzer bewertet werden.

6.2.3 Mehrfachauswahl zum Vergleich der Daten

Im bisherigen Konzept ist es nur möglich einen einzelnen Nutzer, eine Kommunikation oder ein Thema bzw. Stichwort auszuwählen. Sollen diese verglichen werden, so müssen diese nacheinander ausgewählt werden. Ein direkter Vergleich ist nicht möglich. Um dies zu ermöglichen, muss ein Konzept entwickelt werden, das es ermöglicht mehrere Themen, Kommunikationen oder Nutzer auszuwählen und gleichzeitig anzuzeigen. Es muss evaluiert werden welche Daten für verschiedene Nutzer verglichen werden können und wie diese visualisiert werden können. Das selbe muss für Kommunikationen und Themen ebenfalls geschehen

6.2.4 Anzeige weitere Daten und Statistiken

Im bisherigen Konzept werden nur Themen, Nutzer und Kommunikationen unter diesen dargestellt. Für die Analyse eines Netzwerks können aber auch weitere Daten von Interesse sein. Hierbei kann es sich um einfach Daten wie das Profil oder die Follower eines Nutzers handeln oder um Statistiken über das Netzwerk und einzelnen Nutzer.

6.2.5 Clustering

Die entwickelte Darstellung eignet sich nur begrenzt für sehr große Netzwerke, da dabei sehr leicht die Übersicht verloren geht. Um große Netzwerke anzeigen zu können, müssen zur besseren Übersicht Teile davon ausgeblendet werden. Hierfür können einzelne Knoten und Teile der verbundenen Knoten eingeklappt werden. Hierfür bietet es sich an Cluster aus Nutzern zu suchen und diese anstatt der Nutzer anzuzeigen. Die Cluster-Knoten können expandiert werden, so dass wieder alle Nutzer angezeigt werden. Außerdem bietet es sich an, Cluster getrennt vom restlichen Netzwerk zu untersuchen.

Literaturverzeichnis

[BNJ03] D. M. Blei, A. Y. Ng, M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003. URL http://dl.acm.org/citation.cfm?id=944919.944937. (Zitiert auf Seite 24)

[dud] Duden | Kommunikation. URL http://www.duden.de/rechtschreibung/Kommunikation. (Zitiert auf Seite 31)

[Dug12] L. Dugan. Twitter To Surpass 500 Million Registered Users On Wednesday. mediabistro, 2012. URL http://www.mediabistro.com/alltwitter/500-million-registered-users_b18842. (Zitiert auf Seite 13)

[fli] Beliebte Tags bei Flickr. URL http://www.flickr.com/photos/tags/. (Zitiert auf den Seiten 7 und 18)

[Gra72] R. L. Graham. An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set. *Inf. Process. Lett.*, 1(4):132–133, 1972. (Zitiert auf Seite 20)

[HB05] J. Heer, D. Boyd. Vizster: Visualizing Online Social Networks. In *Proceedings of the Proceedings of the 2005 IEEE Symposium on Information Visualization*, INFOVIS '05, S. 5–. IEEE Computer Society, Washington, DC, USA, 2005. doi:10.1109/INFOVIS.2005.39. URL http://dx.doi.org/10.1109/INFOVIS.2005.39. (Zitiert auf den Seiten 7, 24, 25 und 26)

[mal] MAchine Learning for LanguagE Toolkit. URL http://mallet.cs.umass.edu/. (Zitiert auf Seite 50)

[Por97] M. F. Porter. Readings in information retrieval. Kapitel An algorithm for suffix stripping, S. 313–316. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997. URL http://dl.acm.org/citation.cfm?id=275537.275705. (Zitiert auf Seite 23)

[pre] Prefuse manual. URL http://prefuse.org/doc/manual/introduction/structure/. (Zitiert auf den Seiten 7 und 48)

[Sac00] W. Sack. Conversation Map: A Content-Base Usenet Newsgroup Browser. S. 233–240, 2000. (Zitiert auf Seite 26)

[SSMF+09] M. A. Smith, B. Shneiderman, N. Milic-Frayling, E. Mendes Rodrigues, V. Barash, C. Dunne, T. Capone, A. Perer, E. Gleave. Analyzing (social media) networks with NodeXL. In *Proceedings of the fourth international conference on Communities and technologies*, C&T '09, S. 255–264. ACM, New York, NY, USA, 2009. doi:10.

1145/1556460.1556497. URL http://doi.acm.org/10.1145/1556460.1556497. (Zitiert auf Seite 25)

[Sü12] M. F. Südwest. JIM 2012 Jugend, Information, (Multi-)Media Basisstudie zum Medienumgang 12- bis 19-Jähriger in Deutschland. 2012. (Zitiert auf Seite 9)

[twia] REST API v1.1 Resources. URL https://dev.twitter.com/docs/api/1.1. (Zitiert auf Seite 15)

[twib] The Streaming APIs. URL https://dev.twitter.com/docs/streaming-apis. (Zitiert auf Seite 15)

[twic] Twitter basics. URL https://support.twitter.com/groups/31-twitter-basics. (Zitiert auf Seite 13)

[wik12a] Konvexe Hülle. Wikipedia, 2012. URL http://de.wikipedia.org/w/index.php?title=Konvexe_H%C3%BClle&oldid=108319315. (Zitiert auf Seite 19)

[wik12b] Konvexe Menge. Wikipedia, 2012. URL http://de.wikipedia.org/w/index.php?title=Konvexe_Menge&oldid=111683899. (Zitiert auf Seite 19)

[wik12c] Tag Cloud. Wikipedia, 2012. URL http://en.wikipedia.org/w/index.php?title=Tag_cloud&oldid=526244996. (Zitiert auf Seite 50)

Alle URLs wurden zuletzt am 09. 01. 2013 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschift