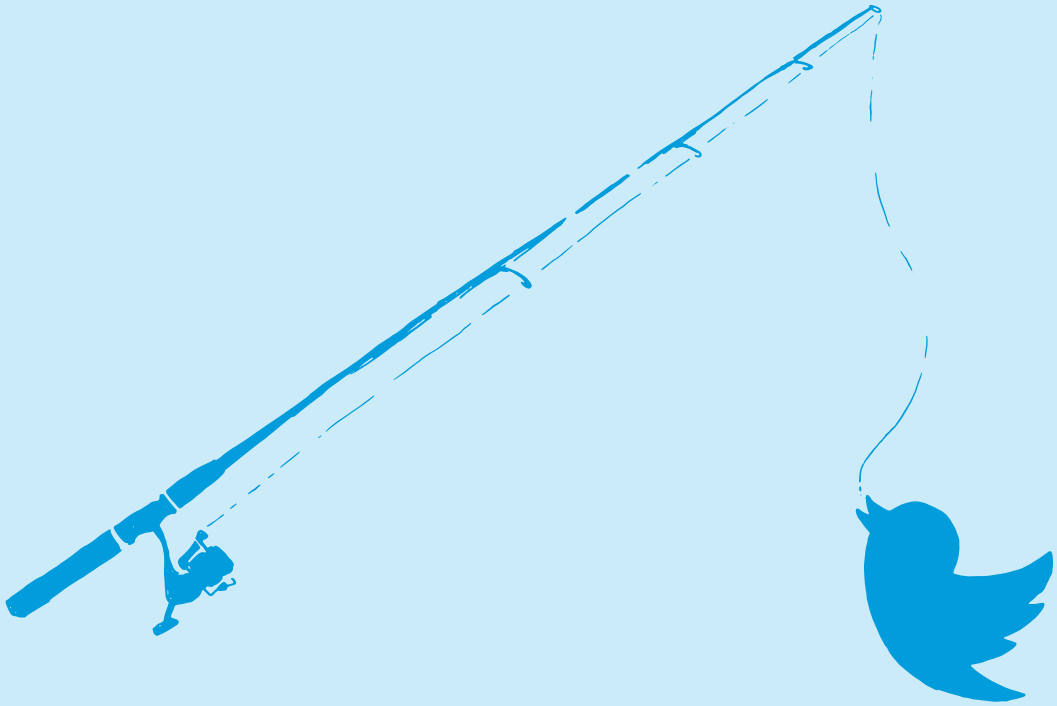


# Im Datenstrom fischen

**JunProf. Dr. Michael Grossniklaus entwickelt und verbessert Systeme zur automatischen Ereigniserkennung in Twitter-Datenströmen**

**earthquake, minute, singapore, strong, prepare**
Wed Apr 11 10:40:13 CEST 2012

**newpope, white, smoke, habemus, vatican**
Wed Mar 13 19:05:55 CET 2013

**gotze, mario, alemania, goal, schurrle**
Sun Jul 13 23:21:44 CEST 2014

**letsgopats, touchdown, patriots, work, stop**
Mon Feb 02 03:45:01 CET 2015

Fünf Wörter, mit einem Zeitstempel versehen. Was auf den ersten Blick wie ein wirres Kauderwelsch erscheint, kann uns sehr viel über die Welt verraten: Was aktuell in aller Munde ist, welche großen Ereignisse weltweit geschehen, welche Gefühlslage zu wichtigen Themen herrscht. Bei den Wörtergruppen handelt es sich um Auszüge aus dem Datenstrom von Twitter. Ein Computer-Algorithmus hat Millionen aktueller Tweets abgeglichen und nach den häufigsten und weitverbreitetsten Schlüsselwörtern gesucht. Die dominantesten Schlagwörter hat er aus dem Datenstrom ausgelesen und mit einer Zeitkennung versehen, wie ein Nachrichten-Ticker. Sie zeigen uns, was die Menschen auf Twitter beschäftigt, und geben uns damit ein digitales Abbild der Welt.

## Ein ‚sozialer Sensor“

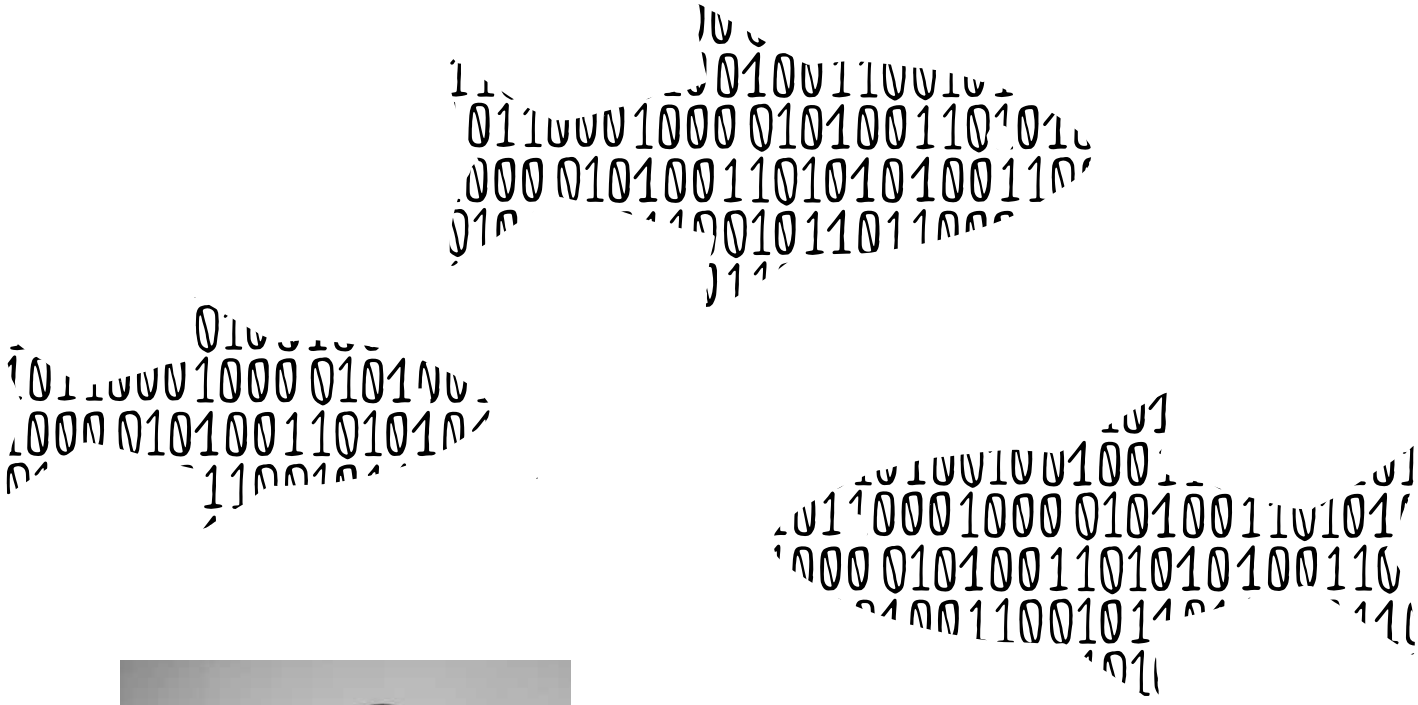
Der Mikroblogging-Dienst Twitter zählt zu den meistfrequentierten Datenströmen im Internet, mit über 500 Millionen Tweets pro Tag von rund 320 Millionen monatlich aktiven Nutzern. Eine gigantische Datenquelle an minutenaktuellen Nachrichten und Informationen. Welches Potential darin steckt, diesen Datenstrom als „sozialen Sensor“ nutzbar zu machen, lassen bereits erste Algorithmen zur automatischen Ereigniserkennung in Twitter-Datenströmen erahnen. Über Twitterdaten ließen sich schon ganze Verläufe von Grippewellen durch den amerikanischen Kontinent nachzeichnen. Bei einem Erdbeben auf den Philippinen wurden Twitterdaten genutzt, um Hilfsdienste zu koordinieren und schnell und gezielt an die Orte zu schicken, wo sie gebraucht wurden. Ein Frühwarnsystem für Naturkatastrophen und terroristische Anschläge wäre auf Grundlage einer Twitter-Ereigniserkennung denkbar.

Doch auch für den alltäglichen Gebrauch bietet die automatisierte Ereigniserkennung zahllose Möglichkeiten, angefangen vom persönlichen Nachrichtenüberblick über Verkehrsanalysen bis hin zu wirtschaftlichen Anwendungen wie zum Beispiel zur Beobachtung und Einschätzung von Börsenkursen. Eine Twitter-Ereigniserkennung ist schneller als redaktionell bearbeitete Nachrichtenübersichten, orientiert sich global und ist vor allem näher dran an den Menschen vor Ort.

> „Bei den Möglichkeiten, was sich damit machen lässt – da ist nur die Phantasie die Grenze.“
>
> JunProf. Dr. Michael Grossniklaus

## Ein Kompromiss aus Schnelligkeit und Qualität

„Bei den Möglichkeiten, was sich damit machen lässt – da ist nur die Phantasie die Grenze“, verspricht Prof. Dr. Michael Grossniklaus, Juniorprofessor für Datenbanken und Informationssysteme an der Universität Konstanz. Gemeinsam mit seinem Team entwickelt und verfeinert der Informatiker Methoden der automatischen Ereigniserkennung in Twitter-Datenströmen. „Unser Ansatz ist: Wir wollten nicht einfach ‚noch ein weiteres Verfahren zur Ereig-



JunProf. Dr. Michael Grossniklaus ist Juniorprofessor für Datenbanken und Informationssysteme an der Universität Konstanz. Zu seinen Forschungsschwerpunkten zählen Anfrageoptimierungen und Datenstromverarbeitungssysteme („data stream management systems“, DSMS).

niserkennung‘ auf den Markt bringen. Wir wollen Ideen geben, wie man Datenanalysesysteme effizienter konstruieren kann, um sie auf die komplexer werdenden Anforderungen der kommenden zehn Jahre vorzubereiten“, erklärt Michael Grossniklaus. Als ersten Schritt seiner Arbeit schuf er daher ein Vergleichsystem, in das die bestehenden, teils sehr unterschiedlichen Verfahren implementiert werden konnten, um ihre Stärken und Schwächen zu überprüfen.

Dabei zeigte sich: Allen Verfahren ist gemeinsam, dass sie einen Kompromiss aus Schnelligkeit und Qualität machen müssen. Sie müssen die enormen Datenmengen von Twitter durchkämmen und auswerten. Eine akribische Auswertung und Überprüfung der gefundenen Schlagworte kostet Zeit, derzeit nicht selten einen ganzen Tag. Für Benachrichtigungsdienste oder sogar Frühwarnsysteme bei Katastrophen ist dies bei weitem zu langsam, die Nachricht ist bis dahin längst veraltet. Eine schnellere Bearbeitung der Daten birgt andererseits die Gefahr, Ereignisse zu übersehen oder vermeintliche Treffer zu produzieren, die für den Nutzer des Dienstes jedoch nicht relevant sind.

„Wie schaffen wir es, ein Verfahren zu entwickeln, das verlässliche Resultate nicht im Tagesrhythmus meldet, sondern jede Minute?“, stellt Grossniklaus die Schlüsselfrage seines Forschungsprojekts. Auf

dem Koordinatenkreuz zwischen Schnelligkeit und Qualität orientieren sich die meisten Verfahren derzeit primär an der qualitativen Achse; der Fokus liegt auf der Optimierung der Präzision der Ergebnisse. Für Michael Grossniklaus ist aber klar, dass der Schlüssel für eine praktikable Anwendung vielmehr auf Seiten der Geschwindigkeit liegt. „Wir müssen, wenn nötig, den Aufwand der Rechercheverfahren kontrolliert zurückfahren – wie mit einem Regler, der das Verhältnis zwischen Aufwand und Schnelligkeit anpasst“, lautet für ihn daher die Antwort. „Wir richten unser Verfahren so aus, dass es die ganz großen Ereignisse aufspürt, aber weniger prägnante – oder sehr themenspezifische – Meldungen ausspart. Im Idealfall könnten wir sehr viel mehr ‚runtime-performance‘ gewinnen, als wir an Qualität verlieren“, schlussfolgert Grossniklaus.

## Ist das Ereignis auch wirklich ein Ereignis?

Wie aber kann geprüft werden, ob das System relevante Ergebnisse liefert? Übersieht es wichtige Tweets? Sind die Ereignisse, die es meldet, auch wirklich ein Ereignis? Michael Grossniklaus nutzt historische Twitterdaten, um die Tauglichkeit seines Verfahrens zu testen: Zehn Terabyte gespeicherte Twittermeldungen, das entspricht zehn Prozent sämtlicher Tweets seit 2012. Der historische Rückblick auf bekannte Daten steckt einen Horizont ab, was das System melden könnte und sollte. Dennoch muss überprüft werden, ob die gemeldeten Ereignisse auch tatsächlich von Relevanz sind. Bislang kam man an einer „händischen“ Überprüfung – üblicherweise mittels Nutzerstudien – kaum vorbei. Dieses Verfahren ist aber langsam und kann angesichts der Datenmengen von Twitter an Grenzen stoßen.

Michael Grossniklaus fand eine praktikablere Lösung: Anstelle einer manuellen Prüfung lässt er die gemeldeten Ereignisse automatisiert mit journalistischen Webseiten abgleichen. Nachrichtenportale wie Reuters, Bloomberg und die New York Times bieten eine tägliche Zusammenfassung ihrer Schlagzeilen. Diese tägliche Sammlung bildet für Grossniklaus den Bewertungsrahmen, was an „großen Nachrichten“ hätte gefunden werden können – und was tatsächlich gefunden wurde.

## Ein ungezogener Datenstrom

Was macht Twitters Datenstrom so relevant für die Forschung? „Es sind die großen Fluktuationen im Datenstrom“, antwortet Michael Grossniklaus. Twitter hat seinen ganz eigenen Zyklus. Der Datenstrom kommt ins Stocken, wenn die Sonne über Ländern aufgeht, in denen Twitter gesperrt ist, und erreicht seinen Höhepunkt bei sozialen Großereignissen wie Weltmeisterschaften und dem Super Bowl. Die sprachlichen Hürden der Tweets erschweren ferner deren automatische Analyse: Börsenkurse wären einfach zu analysieren, denn sie bestehen aus Zahlen. Bei Twitter hingegen haben es die Algorithmen mit verschiedenen Sprachen und Slangs zu tun, mit ungenauer Rechtschreibung und Spam.

„Twitter ist ‚ein sehr ungezogener Datenstrom‘“, schildert Grossniklaus augenzwinkernd. „Es ist ein Text-Stream, der von Max Mustermann geschrieben ist. Das macht ihn so unvorhersehbar und für die Analyse so komplex. Das Hauptproblem ist nicht, dass die Datenmengen größer werden. Wir haben die nötige Kapazität, um dies alles zu speichern. Das Problem ist, dass die Daten unstrukturiert sind und die Anwendungen rasant komplexer werden.“

| gra.

„Wie schaffen wir es, ein Verfahren zu entwickeln, das verlässliche Resultate nicht im Tagesrhythmus meldet, sondern jede Minute?“

JunProf. Dr. Michael Grossniklaus