Big-Data-Analyse im Informatikunterricht mit Datenstromsystemen: Ein Unterrichtsbeispiel

Andreas Grillenberger und Ralf Romeike[1]

Abstract: Big Data stellt heute ein zentrales Thema der Informatik dar: Insbesondere durch die zunehmende Datafizierung unserer Umwelt entstehen neue und umfangreiche Datenquellen, während sich gleichzeitig die Verarbeitungsgeschwindigkeit von Daten wesentlich erhöht und diese Quellen somit immer häufiger in nahezu Echtzeit analysiert werden können. Neben der Bedeutung in der Informatik nimmt jedoch auch die Relevanz von Daten im täglichen Leben zu: Immer mehr Informationen sind das Ergebnis von Datenanalysen und immer häufiger werden Entscheidungen basierend auf Analyseergebnissen getroffen. Trotz der Relevanz von Daten und Datenverarbeitung im Alltag werden moderne Formen der Datenanalyse im Informatikunterricht bisher jedoch allenfalls am Rand betrachtet, sodass die Schülerinnen und Schüler weder die Möglichkeiten noch die Gefahren dieser Methoden erfahren können.

In diesem Beitrag stellen wir daher ein prototypisches Unterrichtskonzept zum Thema Datenanalyse im Kontext von Big Data vor, in dem die Schülerinnen und Schüler wesentliche Grundlagen von Datenanalysen kennenlernen und nachvollziehen können. Um diese komplexen Systeme für den Informatikunterricht möglichst einfach zugänglich zu machen und mit realen Daten arbeiten zu können, wird dabei ein selbst implementiertes Datenstromsystem zur Verarbeitung des Datenstroms von Twitter eingesetzt.

Keywords: Big Data, Datenanalyse, Data Mining, Assoziation, Clusterbildung, Klassifikation, Datenstromsysteme, Echtzeitverarbeitung, Snap!, Twitter, Unterrichtsbeispiel

1 Einleitung

Die Bedeutung der Verarbeitung von Daten nimmt im Kontext von *Big Data* rapide zu: Nicht nur in der Informatik kommt dem Fachgebiet *Datenmanagement*, das sich durch die Innovationen im Kontext von Big Data aus dem Gebiet *Datenbanken* entwickelt hat, eine steigende Bedeutung zu, sondern auch in Wirtschaft, Politik und Gesellschaft. Gleichzeitig zeichnet sich der Stellenwert dieser Entwicklungen beispielsweise auch dadurch ab, dass in diesem Zusammenhang eine völlig neue Berufsgruppe entsteht: der Datenwissenschaftler ("*Data Scientist*") [DP12]. Neben diesen Einflüssen halten Daten und Datenanalysen jedoch auch immer häufiger Einzug in den Alltag, oft ohne dass dies bewusst wahrgenommen wird: Kreditkartenanbieter analysieren ständig Transaktionen auf mögliche Betrugsfälle, Sensordaten werden ausgewertet um Extremwetterereignisse schnell erkennen zu können und im Produktmarketing werden Echtzeitanalysen eingesetzt, um bei der Einführung eines neuen Produkts einem gegebenenfalls auftretenden unerwünschten Image

[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Didaktik der Informatik, Martensstr. 3, 91058 Erlangen
andreas.grillenberger@fau.de, ralf.romeike@fau.de

möglichst bald entgegenwirken zu können. Heute begegnet daher jeder andauernd den Ergebnissen verschiedener Datenanalysen. Andererseits werden jedoch auch immer größere Datenmengen produziert, nicht nur im Bereich von Wissenschaft, Wirtschaft und Politik, sondern wiederum auch im Alltag: Mobile Geräte wie Smartphones erzeugen unaufhörlich große Datenmengen, beispielsweise durch Übermittlung ihres Standortes (nicht nur aus GPS- sondern auch aus WLAN- und Funkzellendaten) oder indem jedes mit dem Gerät aufgenommene Foto u. a. mit GPS-Daten versehen wird. Durch aktuelle Innovationen wie das *Internet der Dinge* („Internet of Things“, IoT), werden voraussichtlich bald noch reichhaltigere Datenquellen aus vielfältigeren Bereichen des täglichen Lebens zur Verfügung stehen: In der Vision des IoT werden jegliche Geräte und Gegenstände mit denen der Mensch interagiert mit Sensoren und eingebetteten Systemen versehen und erhalten somit eine eigene digitale Identität [Br09]. Durch die tiefe Integration solcher Systeme in das tägliche Leben – die Vision des IoT sieht vor, dass diese Geräte nahezu unbemerkt den Alltag unterstützen – entstehen sowohl Gefahren, beispielsweise für die eigene Privatsphäre, aber auch vielfältige Möglichkeiten mit diesen Daten umzugehen, sie für eigene Zwecke zu nutzen und somit das eigene Leben zu unterstützen.

Obwohl sich das IoT und ähnliche Entwicklungen noch in den Grundzügen befinden, kann heute schon erkannt werden, dass eine immer stärkere *Datafizierung*[3] des gesamten Lebens stattfindet. Gleichzeitig fehlen jedoch einem Großteil der Menschen selbst grundlegende Kenntnisse darüber, welche Möglichkeiten und Gefahren sich insbesondere durch die zunehmenden Möglichkeiten zur schnellen Analyse selbst großer Datenmengen eröffnen, und somit auch wesentliche Kompetenzen, die zum Umgang mit (eigenen und fremden) Daten nötig sind. Obwohl derartige komplexe Entwicklungen sicherlich nicht vollumfänglich im Informatikunterricht abgebildet werden können und sollen, liegt jedoch trotzdem die Vermutung nahe, dass dabei grundlegende Aspekte im Sinne der fundamentalen Ideen nach Schwill [Sc93] enthalten sind (vgl. auch [GR15b]). Die Bedeutung und Tragweite von Big Data und der zugrundeliegenden Konzepte wurden auch von Berendt et. al. [Be14] in einer von ihnen beschriebenen Unterrichtsreihe ausführlich dargestellt. Der dabei genutzte Ansatz, im Unterricht eigene Datenanalysen – bei Berendt et. al. insbesondere Assoziationsanalysen – durchzuführen, scheint dabei für das Verständnis vielversprechend zu sein. Im vorgeschlagenen Unterrichtskonzept wurde dazu der Apriori-Algorithmus diskutiert, was sich jedoch in der Erprobung als problematisch erwiesen hat: „Weniger erfolgreich verlief ein für das Reihenziel bzw. dessen Verständnis unabdingbarer Baustein: der Apriori-Algorithmus“ [Be14]. Dabei liegt die Vermutung nahe, dass der Algorithmus aufgrund seiner Komplexität zu Verständnisschwierigkeiten geführt hat. In diesem Beitrag stellen wir daher eine Unterrichtsidee vor, die am Beispiel der grundlegenden Datenanalysemethoden *Assoziation*, *Clusterbildung* und *Klassifikation* zeigt, wie die Grundlagen von aktuellen Innovationen der Informatik für den Informatikunterricht zugänglich gemacht werden können, indem eine Beschränkung auf die grundlegenden und fundamentalen Aspekte dieser Entwicklungen stattfindet, aber trotzdem der innovative Charakter gewahrt wird. Diese Unterrichtsidee soll insbesondere Lehrerinnen und Lehrern als Anregung und Einstieg in das komplexe Thema *Big Data* dienen und stellt daher kein detailliert ausgearbeitetes und erprobtes Unterrichtsbeispiel dar. Zur Umsetzung der dargestellten Ideen wird

[3] Datafizierung bezeichnet die zunehmende Erfassung vielfältiger Aspekte der Realität in Form von Daten.

zusätzlich ein Werkzeug vorgestellt, mit dem Schülerinnen und Schülern durch eine Erweiterung der universellen Programmierumgebung Snap! [HM14] eigene Datenstromanalysen durchführen und dafür auf den Datenstrom des sozialen Netzwerkes Twitter zugreifen können. Bevor dieses Werkzeug in Kapitel 3 dieses Beitrags vorgestellt wird, skizzieren wir in Kapitel 2 die fachlichen Grundlagen der aktuellen Entwicklungen im Datenmanagement und von Datenstromsystemen. Auf dieser Basis stellen wir in Kapitel 4 die konkrete Unterrichtsidee vor.

2 Unterrichtsgegenstand: Big Data und Datenanalysen

Allen aktuellen Innovationen auf dem Gebiet Datenmanagement liegt die Verarbeitung immer größerer Mengen verschiedenartiger Daten in kurzer Zeit zu Grunde: *Big Data*. Dieser Überbegriff für die aktuellen Entwicklungen wird durch die sog. drei „V"s charakterisiert: große Datenmengen (*volume*), schnelle Datenerzeugung und -verarbeitung (*velocity*) sowie unterschiedliche Strukturierung oder gar Unstrukturiertheit dieser Daten (*variety*). Bei der Verarbeitung von Big Data ist die Gewinnung von neuen Informationen und Erkenntnissen aus oft schon vorhandenen umfangreichen Datensätzen ein wesentliches Ziel: Im Kontext von *Data Mining*, dem „Datenbergbau"[4], werden Daten oft als das Gold des 21. Jahrhunderts bezeichnet, in anderen Kontexten auch als das neue Öl. Die Innovationen im Datenmanagement sind dabei vielfältig einsetzbar und eröffnen verschiedene neue Möglichkeiten, stellen aber gleichzeitig den Datenschutz vor neue Herausforderungen und verursachen neue Gefahren für die Privatsphäre. In diesem Zusammenhang bezeichnet Dittrich [Di13] Daten auch als das neue Uran: Natürlich können Daten sinnvoll und gewinnbringend eingesetzt werden, sie sind aber auch kaum löschbar, können in falsche Hände gelangen, angereichert werden und vieles mehr.

Während zur Speicherung und Verarbeitung von Daten bisher insbesondere Datenbanken zum Einsatz kamen, gewinnen im Kontext von Big Data weitere Ansätze zur Datenverarbeitung deutlich an Bedeutung, wie beispielsweise *Datenstromsysteme* (DSS): Im Gegensatz zu Datenbanken werden die Daten in DSS nicht dauerhaft gespeichert, sondern sofort verarbeitet, wodurch eine wesentlich höhere Geschwindigkeit der Datenverarbeitung ermöglicht wird. Durch die sofortige Verarbeitung kann auf geänderte Daten (beispielsweise Sensor-Messwerte) schnell reagiert werden, wie es z. B. bei der Messung von seismischen Wellen in einem System zur Tsunamierkennung und -vorwarnung nötig ist. Diese Systeme betonen daher insbesondere den Aspekt der *velocity* von *Big Data*. Einen wichtigen Einsatzzweck von DSS stellt daher die Überwachung von Datenquellen („*Monitoring*") dar: Gerade im Kontext des Internets der Dinge und auch bei der Heimautomation ist dieses Prinzip allgegenwärtig, beispielsweise indem Beleuchtung und Heizung eines Gebäude automatisch in Abhängigkeit davon geregelt werden, ob Personen anwesend sind oder Jalousien sich automatisch dem Sonnenstand anpassen.

Eine wichtige Basis stellen in diesem Zusammenhang die grundlegenden Datenanalysemethoden dar. An diesen zeigt sich auch die Möglichkeit, die Grundlagen von komplexen

[4] Obwohl sich der Begriff *Data Mining* durchgesetzt hat, sollte dabei möglicherweise eher vom Informationsbergbau gesprochen werden, da es dabei um die Gewinnung von Informationen, nicht von Daten, geht.

und im Ganzen schwer erfassbaren Datenanalysen für den Unterricht didaktisch reduziert greifbar zu machen. In der in diesem Artikel vorgestellten Unterrichtsidee werden wir uns insbesondere auf die drei wichtigsten Datenanalysemethoden [ES00] „Klassifikation“, „Clusterbildung“ und „Assoziation“ beschränken:

- **Clusterbildung** dient dazu, Gemeinsamkeiten zwischen verschiedenen Daten zu analysieren und Daten nach diesen Merkmalen zusammenzufassen. Für die automatisierte Clusterbildung werden bekannte Verfahren wie beispielsweise der „k-Means-Algorithmus“ eingesetzt.
- **Klassifikation** bezeichnet das Einsortieren von Daten in vordefinierte Klassen anhand vorgegebener Merkmale. Von der Clusterbildung unterscheidet sich diese Methode daher dadurch, dass vordefinierte Klassen verwendet anstatt unbekannte gefunden werden. Zur Klassifikation von Daten werden beispielsweise Bayes-Klassifikatoren oder Entscheidungsbäume eingesetzt.
- **Assoziationen** werden genutzt um Zusammenhänge zwischen verschiedenen Merkmalen eines Datensatzes auszudrücken. Diese typischerweise in der Form „Wenn ... dann ...“ formulierten Zusammenhänge werden häufig eingesetzt, um aus bekannten Merkmalen unbekannte vorherzusagen. Aus der der Assoziationsanalyse stammt auch der von Berendt et. al. [Be14] eingesetzte Apriori-Algorithmus.

Im Unterricht können solche Methoden und die damit einhergehenden Möglichkeiten nur unter Nutzung umfangreicher Datenquellen erprobt werden. Solche Datenquellen stehen heute jedoch aus verschiedenen Bereichen zur Verfügung: Für relativ statische Datenanalysen, beispielsweise mit Datenbanken, können Datensätze herangezogen werden, die im Rahmen von *Open-Data*-Initiativen[5] zunehmend veröffentlicht werden. Gleichzeitig stellen insbesondere soziale Netzwerke einen umfangreichen und vielfältig einsetzbaren Strom von Daten bereit, der gut für dynamische Analysen, beispielsweise im Kontext von DSS, genutzt werden kann, durch den starken Bezug zur Lebenswelt aber meist auch einen motivierenden Charakter aufweist. Daher werden wir uns in diesem Unterrichtsbeispiel auf die Analyse des Twitter-Datenstroms mit Hilfe eines einfachen Datenstromsystems konzentrieren.

3 Datenstromanalyse mit Snap!

In diesem Kapitel werden wir kurz das angesprochene und im folgenden Unterrichtskonzept genutzte Werkzeug zur Analyse des Twitter-Datenstroms mit Snap*!* beschreiben[6]. Obwohl es bereits verschiedene Datenstromsysteme gibt, war für den Unterricht eine Eigenentwicklung auf diesem Gebiet unverzichtbar, da bekannte Systeme auf diesem Gebiet für den Unterricht kaum geeignet sind: Insbesondere ist durch die umfangreichen Möglichkeiten, die Einstiegshürde hoch, während zugleich die zugrundeliegenden Prinzipien nur

[5] Open Data bezeichnet Daten, die der Allgemeinheit zur freien Nutzung und Weiterverbreitung zur Verfügung gestellt werden. Sie stammen oft aus der öffentlichen Verwaltung (vgl. z. B. http://www.govdata.de).

[6] Eine detailliertere Beschreibung kann [GR15a] entnommen werden.

schwer erkennbar sind. Hinzu kommen die derzeitigen Entwicklungen auf diesem Gebiet, die die Auswahl eines Systems deutlich erschweren: Im Gegensatz zu Datenbanken gibt es bei DSS bisher keine universelle Anfragesprache, sodass das erworbene Sprachwissen und die Kenntnisse über die Bedienung sehr anwendungsspezifisch wären. Diese Nachteile können durch die Erweiterung der Programmierumgebung Snap! vermieden werden: Den Schülern ist häufig dieses oder ein ähnliches Werkzeug schon bekannt. Durch die blockorientierte Programmierung wird die Einstiegshürde gesenkt und zugleich wird vermieden, sich auf eine bestimmte Anfragesprache zu spezialisieren, indem die Anfragen auf höherer Abstraktionsebene betrachtet werden.

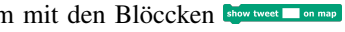
Das implementierte Werkzeug basiert dabei auf zwei Bausteinen: Einerseits musste Snap! um entsprechende Blöcke für die Datenstromanalyse erweitert werden, andererseits musste eine Anbindung an den zu analysierenden Datenstrom geschaffen werden: Eine direkte Anbindung von Snap! an die von Twitter zur Verfügung gestellte API ist aufgrund von Sicherheitsmaßnahmen zur Vermeidung von Cross–Site–Scripting–Attacken in allen üblichen Browsern nicht möglich. Um die Anbindung dennoch zu ermöglichen, musste daher ein Programm implementiert werden, das als Proxy zwischen der Twitter-API und Snap! dient und den Zugriff durch Snap! explizit zulässt. Um die im Folgenden beschriebenen Blöcke in Snap! nutzen zu können, muss daher diese Hilfsanwendung im Hintergrund laufen – es reicht jedoch aus, die Anwendung auf einem PC pro Klassenraum zu starten und Snap! so zu konfigurieren, dass die Daten von diesem PC abgefragt werden.

Zur Erweiterung von Snap! um die für Datenstromanalysen benötigten Blöcke, wurden als Basis die 

und 

Blöcke herangezogen. Die Snap!-Erweiterung ist damit auch auf der offiziellen Snap!-Installation lauffähig, da nur Snap!-eigene Funktionalitäten verwendet werden. Um Daten aus dem Twitter-Datenstrom einzulesen, wurde der Block 

implementiert, der jeweils einen Tweet von der Hilfsanwendung einliest. Innerhalb von Snap! wird das gesamte Tweet-Objekt als JSON-formatierter[7] Text gespeichert. Zum Zugriff auf die einzelnen Attribute ist daher eine weitere Funktion nötig, die den JSON-Text interpretiert und das gewünschte Attribut ausliest. Diese Funktion wird durch den Block 

repräsentiert. Zur ansprechenden Darstellung der Analyseergebnisse wurde zusätzlich die Möglichkeit einer Kartendarstellung bzw. einer Darstellung als Balkendiagramm mit den Blöccken 

bzw. 

implementiert. Von einigen dieser Blöcke, beispielsweise dem Block zum Auslesen von Attributen eines Tweets, wurden weitere Varianten implementiert, beispielsweise 

zum Auslesen der Geokoordinaten.

4 Unterrichtsidee

Im Alltag können Schülerinnen und Schüler Datenanalysen und deren Auswirkungen heute an vielfältigen Stellen begegnen, beispielsweise bei der Verwendung von sozialen Medien oder beim Einkauf im Online-Versandhandel. Um die Möglichkeiten, das Potential und auch die Risiken solcher Analysen erkennen und verstehen zu können, zielt das im

[7] JSON bezeichnet die JavaScript Object Notation, ein gebräuchliches und einfach lesbares Format zum Austausch strukturierter Daten.

Folgenden vorgestellte Unterrichtskonzept darauf ab, die drei grundlegenden Datenanalysemethoden „Klassifikation“, „Clusterbildung“ und „Assoziation“ didaktisch reduziert zu vermitteln. Um einen handlungsorientierten Unterricht zu diesen Themen zu ermöglichen, in dem die Lernenden von Anfang an die Möglichkeit zur Nutzung realer Daten sowie zur interaktiven Entwicklung und Anpassung ihrer Datenanalysen haben, wird das zuvor beschriebene Snap!-basierte Werkzeug eingesetzt, das es den Schülerinnen und Schülern ermöglicht, selbst eigene Datenstromanalysen durchzuführen.

4.1 Randbedingungen und Einsatzkontext

Die im Folgenden beschriebene Unterrichtsidee orientiert sich an keinem speziellen Lehrplan, sondern stellt einen universellen Ansatz dar, Datenanalysemethoden didaktisch reduziert im Unterricht einzusetzen. Damit können Elemente dieser Sequenz in unterschiedlicher Tiefe und Breite in verschiedenen Kontexten eingesetzt werden, wie beispielsweise den Möglichkeiten von Datenanalysen, zu Datenschutz und Privatsphäre oder als Motivation für die Speicherung großer Datenmengen, eingesetzt werden. Das dazu nötige Vorwissen beschränkt sich auf die grundlegende Unterscheidung von Informationen und Daten sowie idealerweise grundlegende Kenntnisse bzw. Erfahrungen mit einer blockbasierten Programmiersprache. Kenntnisse in der Datenanalyse, zu Datenstromsystemen, Datenbanken oder ähnlichem werden dabei zwar nicht zwingend benötigt, können aber förderlich eingebunden werden.

4.2 Unterrichtsverlauf

4.2.1 Einführung

Zur Schaffung eines lebensweltlichen Bezugs wird in diesem Unterrichtsbeispiel exemplarisch der Einsatz von Personalisierungsmöglichkeiten im Online-Versandhandel herangezogen. Jedem Schüler der bereits öfter im Internet eingekauft hat, ist sicherlich aufgefallen, dass sich die Portale der Online-Versandhändler an die eigenen Einkaufsgewohnheiten anpassen, insbesondere in Bezug auf empfohlene Produkte. Während schon für diese Anpassungen grundlegende Datenanalysen genutzt werden, wird es jedoch wesentlich spannender, wenn miteinbezogen wird, dass solche Empfehlungen oft nicht nur mit einem Benutzerkonto, sondern auch mit Merkmalen wie IP-Adresse, Region oder Browser-Identifikation verknüpft werden können und somit auch ohne Benutzerkonto wertvolle Informationen über den Nutzer zur Verfügung stehen oder ermittelt werden können. Das Wissen über den Wohnort der Besucher kann im Zeitalter von Big Data nicht mehr nur eingesetzt werden um regionale Produkte anzubieten, sondern beispielsweise auch um vorherzusagen, wie viel Geld der Kunde zur Verfügung hat oder welchen Beruf er ausübt – wenn auch nur mit einer bestimmten Wahrscheinlichkeit. Dass diese Wahrscheinlichkeitsanalyse für viele Zwecke ausreicht zeigt ein Patent, dass Amazon 2014 zugesprochen wurde: Der Versandhändler plant, Produkte schon in ein naheliegendes Versandzentrum zu senden wenn ein Kunde Interesse daran zeigt, aber noch vor einer konkreten Bestellung

[Sp14]. Um verschiedene Zusammenhänge aufzudecken, reichen einzelne Datensätze oft nicht aus: weitere Informationsquellen müssen herangezogen werden. Als ideale Datenquelle für Marketingzwecke dienen dabei oft soziale Medien wie Twitter oder Facebook: Durch die große Menge an zur Verfügung stehenden Daten können gute Resultate relativ einfach gewonnen werden.

Ziel des im Folgenden vorgestellten Unterrichts ist es daher, dass die Schülerinnen und Schüler die Gewinnung solcher Zusammenhänge nachvollziehen, solche Zusammenhänge an einfachen Beispielen selbst gewinnen und auch kritisch hinterfragen können.

4.2.2 Datenanalysemethoden

Nach einer kurzen Einführung in das Thema Datenanalysen können die Schülerinnen und Schüler mit Hilfe des zuvor erwähnten Werkzeugs einfache Aanalysen selbst durchgeführen und die Funktionsweise der zugrundeliegenden Methoden am Beispiel nachvollziehen.

Klassifikation Anhand einfacher Klassifikationsaufgaben können die Lernenden das zugrunde liegende Prinzip erkennen: die Einteilung von vorliegenden Daten in feste und zuvor definierte Kategorien. Beispielsweise kann dafür die Aufgabe gestellt werden, mit dem zur Verfügung gestellten Tool alle eingehenden Tweets anhand bestimmter Stichworte oder der Sprache in der sie verfasst wurden zu klassifizieren. Unter Nutzung des angesprochenen Werkzeugs kann als erstes Resultat beispielsweise ein Diagramm, wie in Abbildung 1 dargestellt, gewonnen werden.

Abb. 1: Klassifikation der Tweets nach Sprache, dargestellt als Balkendiagramm in Snap!

Bei der Diskussion der möglichen Aussagen, die aus der Klassifikation gewonnen werden können, stellen die Schülerinnen und Schüler die Grenzen dieser Methode fest: Beispielsweise könnte problemlos analysiert werden, welches Produkt beliebter ist als ein anderes, woher die meisten Käufer kommen, und ähnliches. Die zuvor beschriebenen Schlussfolgerungen, die heute aus Daten gewonnen werden können, gehen darüber weit hinaus: Es geht nicht nur darum zu entdecken, welches Produkt am beliebtesten ist, sondern welches Produkt in welcher Region bevorzugt wird – es wird eine zweite Dimension eingeführt. Eine Kategorisierung nach mehreren Dimensionen wäre zwar möglich, die Anzahl der Kategorien explodiert dabei jedoch, da jede Kategorie mit jeder anderen kombiniert werden kann. Zusätzlich können nicht in allen Fällen klare Kategorien schon vor der Analyse festgelegt werden, z. B. muss die Region hier nicht zwingend administrativen Regionen entsprechen, sodass in diesem Fall die Klassifikation an ihre Grenzen gerät.

Clusterbildung An dieser Stelle setzt die Clusterbildung an: Ähnliche Daten werden zusammengefasst und als Gruppe betrachtet. Trotz komplexer mathematischer Grundlagen kann eine einfache Clusteranalyse bereits mit dem zur Verfügung gestellten Werkzeug bewältigt werden: Nach der Visualisierung einer bestimmten Eigenschaft auf der Karte, können Cluster intuitiv bestimmt und so erkannt werden, welche der Ausprägungen dieser Eigenschaft in verschiedenen Regionen vorherrscht. Durch die intuitive Herangehensweise kann die Clusterbildung somit nachvollzogen werden, ohne die mathematischen Grundlagen betrachten zu müssen. In Abbildung 2 wird dieser Vorgang beispielhaft dargestellt.

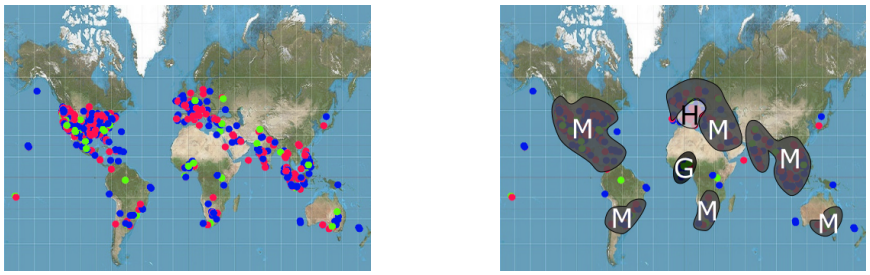


Abb. 2: Visualisierung aller Tweets auf einer Karte in Snap!. Die Farbe der Punkte entspricht der Follower-Zahl (rot: über 500, blau: über 100, grün: unter 100). Die rechte Grafik stellt beispielhafte Cluster mit **g**eringer, **m**ittlerer und **h**oher Followerzahl dar. Karte ©2011 Strebe, CC BY–SA 3.0

Dieses einfache Beispiel zeigt einige der grundlegenden Fragestellungen bei der Verwendung von Clusterbildung: Wie groß sollen die erzeugten Cluster sein? Ab wann wird eine Ausprägung einer Eigenschaft als vorherrschend charakterisiert? Welche Fehler bin ich bereit einzugehen?

Assoziation Gerade wenn Vorhersagen anhand geclusterter Daten getroffen werden sollen, spielen diese Fragen eine entscheidende Rolle: Je größer die Cluster, umso größer ist bei heterogenen Daten der Fehler, der eingegangen werden muss, um diese als ein Cluster betrachten zu können. Dieser Fehler wirkt sich auch auf die Assoziationsanalyse aus, da möglicherweise falsche oder nur teilweise zutreffende Assoziationen gebildet werden. In diesem Kontext sollten daher verschiedene mögliche Assoziationen, die sich aus den obigen Betrachtungen ergeben, mit den Schülern auf ihre Aussagekraft hin untersucht werden: Mit Blick auf die in Abbildung 2 dargestellte Karte scheint die Assoziation „Wer Twitter in Westeuropa, Südostasien oder den USA nutzt, hat mindestens 100 Follower“ mit relativ geringem Fehler zutreffend zu sein. Diese Aussagen vernachlässigen aber jegliche Kausalität: Es wird nicht weiter überlegt, warum beispielsweise in Afrika, Australien oder Russland Twitter anscheinend kaum genutzt wird. Da es sich bei der Auswertung jedoch um eine nicht-repräsentative Stichprobe handelt (es wurden nur Tweets betrachtet, die Geodaten offenbaren), kann kein Rückschluss auf alle Twitter-Nutzer getroffen werden. Möglicherweise hängt die Bereitschaft die eigenen Geodaten zu offenbaren mit der eigenen Aktivität – und damit möglicherweise auch der Follower-Anzahl – zusammen, sodass die getroffene Schlussfolgerung durch die Stichprobenwahl beeinflusst wurde. Diese Grenzen der Aussagekraft von Datenanalysen muss man sich daher unbedingt bewusst machen, um falsche Schlüsse und somit auch das Vorurteil der Allwissenheit und Unfehl-

barkeit von Big Data zu vermeiden, aber auch um die Bedeutung eines möglichst vollständigen Datensatzes bei Big-Data-Analysen nachvollziehen zu können.

4.2.3 Möglichkeiten und Risiken von Datenanalysen

Die obigen Beispiele zeigen deutlich, wie aus einfachen und bedeutungslos wirkenden Daten interessante und je nach Einsatzgebiet sehr wertvolle Daten gewonnen werden können. Übertragen auf den Online-Versandhandel zeigt sich damit die Möglichkeit, eine starke Personalisierung von Produktempfehlungen, Werbung (eigener sowie fremder) etc. durchzuführen, was für beide Seiten – Kunde und Betreiber – wünschenswert ist: Einerseits wird dem Kunden das Auffinden interessanter Produkte erleichtert, andererseits steigert sich der Absatz des Händlers. Neben diesen Möglichkeiten die durch die Prinzipien der Datenanalyse eröffnet werden, müssen jedoch auch immer die damit einhergehenden Gefahren betrachtet werden, um auf diese geeignet reagieren und angemessen damit umgehen zu können. Dieselben Onlinehändler, die derartige Datenanalysen zum beidseitigen Vorteil von Kunde und Händler einsetzen, können diese Daten beispielsweise auch zur Personalisierung des Preises eines Artikels nutzen: Wer aus einem reichen Umfeld kommt, ist sicherlich eher bereit, hohe Preise zu bezahlen, als jemand aus einer ärmeren Wohngegend. Dies zeigt auch die Gefahr einer Stigmatisierung aufgrund von Big-Data-Analysen: Durch die reine Betrachtung von Korrelationen anstatt von Kausalitäten wird ein wichtiger Aspekt vernachlässigt. Es wird nicht mehr überlegt, warum jemand möglicherweise in einer eher ärmeren Wohngegend lebt, sondern daraus beispielsweise nur noch direkt seine Kreditunwürdigkeit gefolgert. Im schlimmsten Fall geschieht dies, ohne dass derjenige auch nur geringen Einfluss nehmen kann, da möglicherweise keinerlei konkrete Daten über diese Person zur Verfügung stehen, sondern nur statistische Analysen der Entscheidung zugrunde liegen.

5 Fazit

Wie in der Unterrichtsidee dargestellt wurde, können die grundlegenden Datenanalysemethoden geeignet didaktisch reduziert betrachtet werden, um sie in der Schule diskutieren zu können. Damit wird das Potential geschaffen, auf dieser Grundlage auch komplexere Beispiele, wie z. B. den Apriori-Algorithmus verstehen zu können. Die drei Methoden *Klassifikation*, *Clusterbildung* und *Assoziation* stellen dabei nicht nur in der Datenanalyse wichtige Grundlagen dar, sondern können auch in anderen Gebieten der Informatik, aber auch in anderen Wissenschaften und im täglichen Leben sinnvoll eingesetzt werden. Diese Methoden werden beispielsweise im Kontext von Datenbanken, Data Warehouses und Datenstromsystemen verwendet, aber auch in der Mustererkennung oder der angewandten Informatik (z. B. medizinische Informatik). Gerade Klassifikationen und Assoziationen können aber z. B. auch beim objektorientierten Entwurf wiedergefunden werden. Obwohl diese Methoden im Zusammenhang mit Big Data und Data Mining als relativ moderne Entwicklungen erscheinen, betonen sie exemplarisch ein wichtiges Prinzip der Informatik: die Reduzierung von Komplexität, indem gleichartige Objekte zusammengefasst (*klassi-*

fiziert), die Ähnlichkeit von Objekten erkannt (*Clusterbildung*) und Regeln/Zusammenhänge zwischen verschiedenen Objekten und deren Verhaltensweisen aufgedeckt werden (*Assoziation*). Auch im Zusammenhang mit Datenanalysen waren diese Methoden schon lange vor dem Schlagwort *Big Data* etabliert und wurden beispielsweise schon 2000 von Ester & Sander [ES00] in der Art beschrieben, in der sie auch heute eingesetzt werden. Auch in Zukunft ist nicht zu erwarten, dass die Bedeutung dieser Methoden abnehmen wird, im Gegenteil ist aufgrund der beschriebenen Entwicklungen aktuell ein Ansteigen ihrer Relevanz zu beobachten und auch zukünftig weiter zu erwarten.

Durch die Thematisierung dieser Methoden im Informatikunterricht bekommen die Schülerinnen und Schüler außerdem die Möglichkeit, Datenanalysen selbst durchzuführen, was durch das dargestellte Werkzeug unterstützt wird. Die Analyse des Twitter-Datenstroms stellt dabei nur eine der vielfältigen Einsatzmöglichkeiten dar: Mit geringem Aufwand kann Snap! beispielsweise auch so erweitert werden, dass statt des Twitter-Datenstroms RSS-Feeds, andere RDF-formatierte Daten und ähnliches eingesetzt werden können. Durch die Nutzung weiterer Werkzeuge, beispielsweise von *import.io*, können auch beliebige Webseiteninhalte für die Datenanalyse vorbereitet werden (*Data Scraping*).

Literaturverzeichnis

[Be14] Berendt, Bettina; Dettmer, Gebhard; Demir, Cihan; Peetz, Thomas: Kostenlos ist nicht kostenfrei. LOG IN, (178/179), 2014.

[Br09] Brand, Leif; Hülser, Tim; Grimm, Vera; Zweck, Axel: Internet der Dinge - Perspektiven für die Logistik. 2009. https://www.vdi.de/fileadmin/vdi_de/redakteur/dps_bilder/TZ/2009/Band%2080_IdD_komplett.pdf, zuletzt geprüft: 27.04.2015.

[Di13] Dittrich, Jens: NSA, Überwachung, Big Data und warum Daten wie Uran sind. 2013. https://youtu.be/gwz6u8kqvSo, zuletzt geprüft: 27.04.2015.

[DP12] Davenport, Thomas H.; Patil, D. J.: Data scientist: the sexiest job of the 21st century. Harvard business review, 90(10):70–77, 2012.

[ES00] Ester, M.; Sander, J.: Knowledge Discovery in Databases:. Springer Berlin, 2000.

[GR15a] Grillenberger, Andreas; Romeike, Ralf: Analyzing the Twitter Data Stream Using the Snap! Learning Environment (in Druck). In: Proceedings of ISSEP 2015, Lecture Notes in Computer Science. Springer International Publishing, 2015.

[GR15b] Grillenberger, Andreas; Romeike, Ralf: Big Data im Informatikunterricht: Motivation und Umsetzung. In: INFOS 2015. Lecture Notes in Informatics (LNI). Köllen Druck+Verlag, Bonn, 2015.

[HM14] Harvey, Brian; Mönig, Jens: Snap! Reference Manual. 2014. http://snap.berkeley.edu/SnapManual.pdf, zuletzt geprüft: 27.04.2015.

[Sc93] Schwill, Andreas: Fundamentale Ideen der Informatik. Zentralblatt für Didaktik der Mathematik, 1993.

[Sp14] Spiegel Online: Neues Patent: Amazon will schon vor der Bestellung liefern. 2014. http://www.spiegel.de/article.do?id=944252, zuletzt geprüft: 27.04.2015.