Hochschule für Angewandte Wissenschaften Hamburg
*Hamburg University of Applied Sciences*

# Ausarbeitung AW1 - WiSe 2013/2014

**Sebastian Krome**

## Big Data bei unstrukturierten Daten

*Fakultät Technik und Informatik*
*Studiendepartment Informatik*

*Faculty of Engineering and Computer Science*
*Department of Computer Science*

Sebastian Krome

# Big Data bei unstrukturierten Daten

Betreuender Prüfer: Prof. Dr. Kai von Luck , Prof. Dr. Bettina Buth

Eingereicht am: 10. März 2014

# Inhaltsverzeichnis

# 1 Einleitung

Unter dem Begriff *Big Data* wird zum einen die Datenflut, deren Bewältigung uns vor neue Herausforderungen stellt, als auch die damit verbundenen neuen Methoden und Technologien verstanden, die die Verarbeitung und Analyse unterschiedlicher Daten ermöglichen. Eine Besonderheit von *Big Data* sind dabei die Ergebnisse, die durch die gemeinsame Analyse von unterschiedlich strukturierten Daten, die bisher nicht aufeinander bezogen wurden, erzielt werden.

## 1.1 Der Begriff Big Data

Der Begriff *Big Data* ist nicht eindeutig definiert, eine mögliche, weit gefasste Definition ist jedoch folgende von Zikopoulos et al. [IZE11]:

> Big Data applies to information that can't be processed or analyzed using traditional processes or tools.

Auch wenn der Begriff *Big Data* impliziert, dass die einzige Herausforderung der Umgang mit einer äußerst großen Masse von Daten ist, so ist dieses nur eine der durch das 3-V-Modell beschriebenen Eigenschaften und Herausforderungen von *Big Data*. Dieses geht auf einen Forschungsbericht von Laney zurück, in welchem er die zukünftigen Herausforderungen des Datenwachstums und der Datenverarbeitung anhand von drei Dimensionen beschreibt[Lan01]:

- Data Volume
- Data Velocity
- Data Variety

Diese sollen im Folgenden kurz beschrieben werden.

**Volume** beschreibt die Eigenschaft, dass heutzutage immer mehr Daten erfasst und persistent gespeichert werden. In zahlreichen Gebieten des Alltags werden Daten erhoben und gespeichert. Ein Beispiel für das erzeugte Volumen ist Facebook, dessen Nutzer pro Minute 650.000 verschiedene Inhalte generieren, oder ca. 35.000 „Likes“ verteilen [KTGH13].Vom Volumen her noch bedeutender sind Multimediadaten, die ca. 60 % des Internet-Traffics ausmachen[Smi13].

Alleine diese große Menge an Daten stellt für traditionelle Datenbanksysteme eine Herausforderung dar. Weiterhin muss abgewogen werden, ob der Wert der zu speichernden Daten die Kosten für große Datenbanksysteme aufwiegt [KTGH13].

**Velocity** kann auf zwei unterschiedliche Weisen betrachtet werden: Zum einen ist mit Velocity die hohe Frequenz gemeint, mit der Daten erzeugt werden. Die andere Betrachtungsweise ist die Geschwindigkeit, mit der Daten verarbeitet werden müssen, damit die aus ihnen gewonnene Information noch relevant ist [KTGH13]. In diesem Kontext ist auch die hohe Alterungsgeschwindigkeit der Quelldaten ein erwähnenswerter Aspekt von Velocity.

**Variety** nimmt Bezug darauf, dass bei herkömmlichen Analyse-Systemen nur relationale Daten verarbeitet werden. Im Rahmen von *Big Data* sollen nun jedoch auch Informationen aus semistrukturierten und unstrukturierten Daten wie z.B. Webseiten, Inhalte von Social Media Plattformen, Emails, Fotos etc. analysiert werden. Diese stark unterschiedlichen, unstrukturierten Daten stellen jedoch für traditionelle Datenbanksysteme eine Herausforderung dar. Ziel bei der Verarbeitung dieser unterschiedlichen Daten ist, diese mit den vorhandenen relationalen Daten zusammenzufassen und gemeinsam zu analysieren.

Ein weiteres Attribut, das nicht im herkömmlichen 3-V-Modell aufgeführt ist, jedoch auch wiederholt zu den Eigenschaften von *Big Data* gezählt wird, ist der vom IBM geprägte Begriff *veracity*.

**Veracity** bezeichnet die Eigenschaft, dass den Daten, die in der Regel aus unterschiedlichen Quellen stammen, in der Regel eine gewisse Unsicherheit anhaftet. Dies ist damit begründet, dass bei *Big Data* Daten im Gegensatz zu traditionellen Data Warehouse Systemen, die Daten aufwändig bereinigen, Daten kaum qualitativ aufbereitet oder kontrolliert werden. Dies wird damit argumentiert, dass es bei den großen Datenmengen und der hohen Geschwindigkeit, mit der Daten verarbeitet werden müssen, nicht mehr wirtschaftlich sei [IZE11].

## 1.2 Datenverarbeitungsprozesse

Um aus Daten für Menschen nachvollziehbare Informationen zu gewinnen, müssen die vorhandenen Rohdaten – gerade bei großen Datenmengen – aufbereitet werden. Eine solche Aufbereitung findet zum Beispiel im Bereich der Visualisierung, der grafischen Darstellung von Daten, statt. Auf diese Weise sollen Daten für Menschen verständlicher gemacht werden, da Grafiken für Menschen leichter zu verstehen sind als reine Datensammlungen.

Für diese Aufbereitung stellt Fry in *Visualizing Data* einen siebenstufigen Prozess vor, indem die Rohdaten aufbereitet und letztendlich grafisch dargestellt werden. Die einzelnen Schritte des Prozesse sind [Fry07]:

1. aquire: Das Beschaffen der Rohdaten.
2. parse: Das Strukturieren und Kategorisieren der Daten.
3. filter: Das Entfernen aller irrelevanten Daten.
4. mine: Das Erkennen von Mustern in den Daten durch statistische Verfahren oder Verfahren aus dem Bereich des Data Minings.
5. represent: Das Wählen eines Basisvisualisierungsmodells (z.B. Bar Chart, Baum etc.).
6. refine: Die Verfeinerung des Basisvisualisierungsmodells zum besseren Verständnis des Dargestellten.
7. interact: Das Hinzufügen von Methoden, um die Daten zu manipulieren oder bestimmte Eigenschaften auszublenden.

Ein anderer Prozess zur Datenaufbereitung findet im Bereich des *Knowledge Discovery in Databases (KDD)*, der automatischen Analyse von Daten zur Wissensfindung, statt. Dabei steht im Bereich KDD das Data Mining, das eigentliche Auffinden von Mustern in den Daten, im Zentrum des Interesses. Es steht jedoch erst am Ende eines aus mehreren Prozessschritten bestehenden Prozesses, in dem die Rohdaten für das Data Mining qualitativ aufbereitet werden [BKI06].

Dieser Prozess besteht aus ähnlichen Prozessschritten, wie der von Fry vorgestellte Prozess zur Datenvisualisierung, jedoch ohne die letzten drei auf speziell für die Datenvisualisierung definierten Schritte: Es wird eine Menge von Daten als Untersuchungsobjekt festgelegt, diese werden bereinigt (z.B. Filtern von Rauscheffekten und Festlegung der Datentypen) und komprimiert, bevor während des Data Minings Muster in Daten gefunden werde sollen[1].

Auch bei dem Thema *Big Data* sollen aus einer großen Zahl von Rohdaten Informationen gewonnen werden. Auch hierfür muss es einen Prozess geben, der den beiden bereits vorgestellten Prozessen ähnelt: Es müssen zunächst Daten gesammelt werden, diese müssen für das Data Mining aufbereitet werden, in dem anschließend Muster in den Daten gefunden werden sollen, aus denen sich neue Erkenntnisse ableiten lassen.

Ein wichtiger Schritt in der Vorverarbeitung der Daten ist im KDD-Prozess der Prozessschritt *Datenbereinigung*. In diesem werden Ausreißer aus der Datenbasis entfernt, Datentypen festgelegt und die Behandlung fehlender Daten wird geklärt[BKI06]. Der entsprechende Schritt im Datenvisualisierungsprozess ist der Prozessschritt *parse*

In Bezug auf das Thema *Big Data* und die Eigenschaft *variety* besteht ein wesentlicher Unterschied zum klassischen KDD darin, dass bei *Big Data* auch Informationen aus semistrukturierten und unstrukturierten Daten gewonnen werden sollen. Um diese Daten nutzen zu können, müssen sie zunächst in eine von Maschinen verarbeitbare Form gebracht werden, indem sie strukturiert

[1] Für eine genaue Beschreibung der einzelnen Schritte des KDD-Prozesses siehe [BKI06]

werden. Dieser Arbeitsschritt ist ebenfalls dem Prozessschritt *Datenbereinigung* (KDD) bzw. *parse* (Datenvisualisierung) zuzuordnen.

Wie diese Aufbereitung bei Multimediadaten und Textdaten aussehen kann, wird in den folgenden Abschnitten beschrieben.

# 2 Verarbeitung von Multimediadaten

Multimediadaten machen zur Zeit ca. 70% aller vorhandenen unstrukturierten Daten aus. Ihre Rolle für den Bereich *Big Data* wird von Smith folgendermaßen beschrieben [Smi13]:

> „Multimedia clearly is 'big data'. But, it is big data not just because there is a lot of it. Multimedia is big data because increasingly it is becoming a valuable source for insights and information. Multimedia data can tell us about things happening in the world, point out places, events or topics of interest (memes), give clues about a person's preferences and even capture a rolling log of human history."

Multimediadaten sind eine neue Form der Daten, die im Rahmen von *Big Data* gemeinsam mit anderen zur Verfügung stehenden Daten analysiert werden sollen. Die Bedeutung und den Inhalt dieser Daten maschinell zu erfassen ist jedoch ein schwieriges Gebiet und es werden komplizierte Algorithmen benötigt, um die Eigenschaften von Fotos, Audio- oder Videodaten zu erfassen und zu extrahieren [Smi12]. Diese extrahierten Eigenschaften können im Anschluss analysiert und ausgewertet werden.

Um einen Einblick zu geben, auf welche Weise die vorhandenen Informationen aus Multimediadaten gewonnen werden können, werden in die folgenden beiden Abschnitten Arbeiten vorgestellt, die unterschiedliche Verfahren hierzu entwickelt haben.

## 2.1 Erfassung der Bedeutung von Multimediadaten durch automatisiertes Tagging

Ein weit verbreiteter Ansatz zur Erfassung der Bedeutung von Multimediadaten ist beispielsweise die Zuordnung von Tags zu Bildern oder Videos, mit deren Hilfe Nutzer von Plattformen wie Flickr den in Bildern dargestellten Inhalt beschreiben. Auf diese Weise werden Bilder z.B. besser für andere Nutzer suchbar gemacht.

Mit der automatischen Annotation von Bildern, bei der neuen Bildern anhand von aus Trainingsdaten gelernten Mustern Tags automatisch zugeordnet werden, haben sich Ma et al. beschäftigt [MZLK10].

Dazu wird ein Bild zunächst anhand eines 297-dimensionalen Eigenschaftsvektors beschrieben, welcher aus den Low-Level Eigenschaften des Bildes gewonnen wird. Hierzu werden beispielsweise Farbeigenschaften (z.B der durchschnittliche Wert eines Farbkanals oder die Varianz eines Farbkanals) und Kanteneigenschaften (das Bild wird in ein Graustufenbild umgewandelt. Anschließend wird mit Hilfe des Canny-Algorithmus[1] ein Histogramm der Kantenrichtungen berechnet) genutzt.

Anhand des extrahierten Eigenschaftsvektors lassen sich die Ähnlichkeiten zwischen verschiedenen Bildern mit Hilfe des Cosinus-Maßes[2] bestimmen. Mit Hilfe der bestimmten Ähnlichkeiten zwischen den Bildern wird in einem nächsten Schritt ein Graph (*Bildähnlichkeitsgraph*) aufgebaut, in dem ein Bild mit seinen k ähnlichsten Bildern verbunden wird. Das Kantengewicht ist dabei die berechnete Ähnlichkeit zwischen den Bildern (Ein Beispiel eines solchen Graphen mit k=1 ist in Abbildung 2.1(a) gezeigt).

Neben diesem *Bildähnlichkeitsgraphen* wird ein weiterer Graph aufgebaut, der die Beziehung zwischen einem Bild und einem Tag angibt. Dieser ist in Abbildung 2.1(b) gezeigt. $d_1$ bis $d_4$ stellen dabei Bilder dar, $t_1$ bis $t_3$ stehen für die assoziierten Tags.

Das Kantengewicht einer Kante, die von einem Bild d zu einem Tag t verläuft, wird mit Hilfe der Zahl n aller Tags, mit denen das Bild getaggt ist, normalisiert: $Kantengewicht_{dt} = \frac{1}{n}$.

Das Kantengewicht einer Kante, die von einem Tag t zu einem Bild d verläuft, wird mit Hilfe der Zahl m normalisiert, mit der das Tag allen Bildern zugeordnet wurde: $Kantengewicht_{td} = \frac{1}{m}$.

In einem letzten Schritt wird aus den beiden erzeugten Graphen ein *Hybridgraph* gebildet, der sowohl die Ähnlichkeiten zwischen den Bildern, als auch die Beziehung zwischen Tags und Bildern widergibt (siehe Abbildung 2.1(c)).
Neue Bilder lassen sich nun taggen, indem bei ihnen ebenfalls ein Vektor der Low-Level-Eigenschaften bestimmt wird, anhand dessen sich die 40 ähnlichsten bekannten Bilder, die sich bereits im *Hybridgraphen* befinden, bestimmen lassen. Durch das Bestimmen der ähnlichsten Bilder im *Hybridgraphen*, kann das neue Bild nun ebenfalls in diesen eingehängt werden und die Beziehungen zu vorhandenen Tags lassen sich berechnen.

[1] Canny-Alogorithmus: Algorithmus zur Kantendetektion, der anhand von verschiedenen Faltungsoperationen ein Bild liefert, welches im Idealfall nur noch die Kanten des ursprünglichen Bildes enthält. Weiterhin lassen sich pro Pixel die Richtungen der potentiellen Kante durch diesen Pixel berechnen.

[2] Cosinus-Maß: Ein Ähnlichtkeitsmaß, welches die Ähnlichkeit zweier Vektoren anhand des Winkels zwischen diesen Berechnet sim_cos $(d_p, d_q) = \frac{\vec{d_p} * \vec{d_q}}{|\vec{d_p}| * |\vec{d_q}|}$

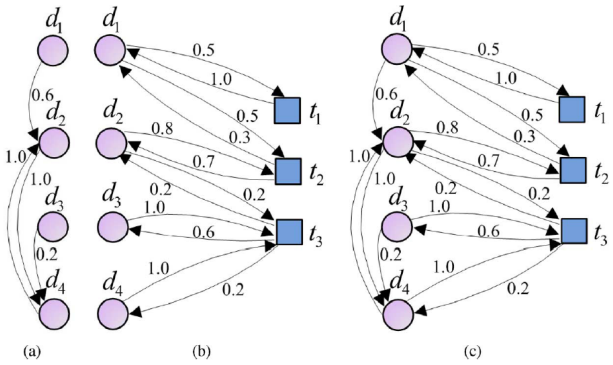


Abbildung 2.1: Konstruktion des *Hybridgraphen*, Quelle: [MZLK10]

## 2.2 Entwicklung von Ontologien für Multimedia

Ein weiterer Bereich zum Thema Erfassung der Semantik von Multimediadaten ist die Entwicklung von Ontologien, mit denen audio-visuelle Konzepte[3] und Beziehungen zwischen diesen dargestellt werden. Diese bilden einen wichtigen Baustein und tragen wesentlich zum Training von semantischen Klassifikatoren bei.

Ein Beispiel für eine solche Ontologie ist LSCOM [NST+06] (*Large-Scaling Concept Ontology for Multimedia*) - Naphade et al. x. Diese besteht aus 1000 Konzepten, mit denen Nachrichtenvideos beschrieben werden sollen, und welche für das Training von semantischen Klassifikatoren verwendet werden können (z.B. für das Tagging von Videos).

# 3 Strukturieren von Textdaten

Eine andere Art von unstrukturierten Daten sind Textdaten, wie z.B. Emails, Webseiten oder Social Media Daten. Diese enthalten ebenfalls relevante Informationen, die es im Rahmen von *Big Data* zu nutzen gilt. Da Textdaten ebenso wie Multimediadaten unstrukturiert sind, müssen diese ebenfalls zunächst in eine strukturierte, von Maschinen verarbeitbare Form gebracht werden.

[3] Konzept: Ein Konzept c ist eine einstellige Funktion $c : M \rightarrow \{0, 1\}$, die über eine Grundmenge M von Beispielen angibt, ob ein Beispiel zum Konzept gehört oder nicht.

Dabei werden durch Methoden des Text Minings verschiedene Informationen aus einem Text extrahiert. Hier kann in verschiedene Verfahren unterschieden werden:

**Statistische Verfahren:** Texten werden unter Ausnutzung statistischer Gesetzmäßigkeiten nach verschiedenen Kriterien Eigenschaften zugeordnet [HQW08].

**Musterbasierte Verfahren:** Innerhalb vorgegebener Texte werden weitgehend allgemeingültige Muster herausgefunden und als Regel definiert (z.B. könnte eine Regel besagen, dass ein Name folgende Form hat: <Anrede><Vorname><Nachname>)[HQW08].

**Nutzung eines Wörterbuchs:** Wörterbücher können genutzt werden, um bestimmte Wörter, die in den Wörterbüchern vorkommen, mit bestimmten Tags zu versehen. Auf diese Weise könnten z.B. Vornamen anhand eines Vornamen-Wörterbuchs extrahiert werden [FS06].

Im Folgenden soll nun anhand von verschiedenen Arbeiten ein Einblick in die Extrahierung von Informationen aus Textdaten gegeben werden. Dabei beziehen sich alle Arbeiten auf die Informationsgewinnung aus Twitter Tweets, ein Bereich der viele Methoden des klassischen Text Minings adaptiert hat, sich dabei jedoch auf eine relativ neue Informationsquelle bezieht.

## 3.1 Extrahierung von Ereignissen

In diesem Abschnitt soll die Arbeit von Ritter et al. [RMEC12] vorgestellt werden, die mit Hilfe von verschiedenen Verfahren des Text Minings Ereignisse und an Ereignissen beteiligte Entitäten aus Tweets extrahiert haben. Das von ihnen entwickelte System TwiCal extrahiert aus einem Strom von Tweets Ereignisse, welche durch ein 4 Tupel beschrieben werden. Die dabei extrahierten Eigenschaften sind:

**Entität:** Eine Entität, die in Zusammenhang mit dem Ereignis steht.

**Ereignis Phrase:** Eine Phrase, die das Ereignis in dem Tweet beschreibt.

**Datum:** Ein eindeutiges Datum, an dem sich das Ereignis ereignet (hat).

**Typ:** Eine Kategorie, der das Ereignis zuzuordnen ist. Diese Eigenschaft wird im Gegensatz zu den anderen nicht aus dem Tweet extrahiert, sondern leitet sich aus den anderen Eigenschaften ab. Aus diesem Grund soll diese Eigenschaft im weiteren Verlauf dieser Ausarbeitung nicht weiter betrachtet werden.

Ein Beispiel eines extrahierten Ereignisses ist in Tabelle 3.1 gezeigt.
Betrachtet man die in Abbildung 3.1 gezeigte Gesamtübersicht des Systems, so ist dort zu erkennen, dass sich das System aus mehreren Komponenten zusammensetzt, die die einzelnen in Tabelle 3.1 aufgeführten Eigenschaften des Ereignisses extrahieren.

Tabelle 3.1: Beispiele von extrahierten Ereignissen, Quelle: [RMEC12]

| Entität | Ereignisphrase | Datum | Typ |
|---|---|---|---|
| Steve Jobs | died | 10/6/11 | DEATH |

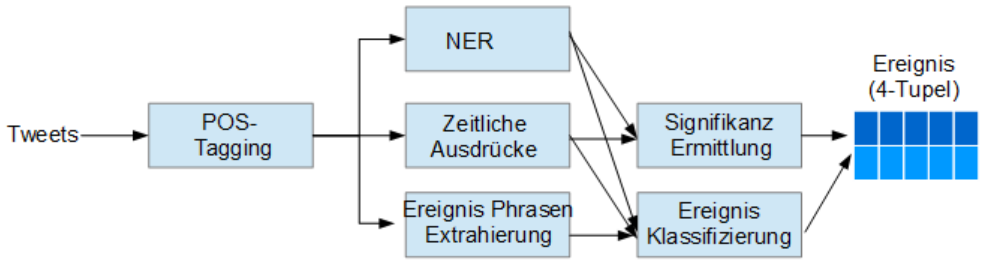


Abbildung 3.1: Gesamtübersicht des Systems TwiCal, angelehnt an [RMEC12]

Von diesen sollen nun die wichtigsten Komponenten kurz erläutert werden. Allen gemeinsam ist, dass sie bereits auf die Ergebnisse eines vorgeschalteten POS-Taggers[1] zugreifen, der speziell für Twitterdaten trainiert wurde. Für eine genauere Beschreibung des POS-Taggers sei an dieser Stelle auf [RCME11] verwiesen.

**Named Enitity Recognition (NER):** Diese Komponente ist dafür zuständig, aus einem Tweet die Named Entities zu extrahieren. Sie wurde speziell für die besonderen Eigenschaften von Tweets entwickelt. So klassifiziert sie zunächst, ob die Groß-und Kleinschreibung (eine wichtige Eigenschaft im Bereich NER) in einem Tweet zuverlässig ist und verwendet auf Tweets trainierte Sequenzmodelle[2], um mögliche Kandidaten von Named Entities zu extrahieren. Diese Eigenschaften werden anschließend zusammen mit Informationen der Ontologie *Freebase* zur Klassifizierung (z.B. Person oder Ort) der Entitäten genutzt. Eine genaue Beschreibung dieser Komponente wird in [RCME11] gegeben.

**Ereignisphrasen Extrahierung** : Diese Komponente ist dafür zuständig, die Ereignisphrasen aus einem Tweet zu extrahieren. Hierzu wurde ein ähnliches Vorgehen wie bei der Named Enitiy Recognition gewählt: Es wurden in 1000 Tweets manuell Ereignisphrasen (wie z.B. *died*) annotiert und für das Training von Sequenzmodellen (Conditional Random Fieds) genutzt. Da bei Sequenzmodellen Eingabesequenzen und nicht einzelne Tokens betrachtet werden,

[1] Part-of-speech (POS) Tagging bezeichnet die Zuordnung von Wörtern zu einem Label mit der Wortart des Wortes.
[2] Sequenzmodelle: Stochastische Modelle, die für eine Eingabesequenz (z.B. Wörter) eine Wahrscheinlichkeitsverteilung für mögliche Ausgabesequenzen (z.B. Sequenz von Labeln) berechnen und anschließend die beste mögliche (wahrscheinlichste) Ausgangssequenz wählen. Eine gute Einführung in Sequenzmodelle wird in [HQW08] gegeben.

eignen sie sich besonders zur Extrahierung von Phrasen, die aus mehreren Tokens (Wörtern) bestehen können. Weitere Informationen, die zur Extrahierung der Ereignisphrase genutzt wurden, sind die Tags der vorgeschalteten POS-Taggers und Informationen, die aus einem Dictionary stammen [RMEC12].

**Zeitliche Ausdrücke:** Diese Komponente ist dafür zuständig, zeitliche Ausdrücke wie *tomorrow* zu extrahieren und in ein eindeutiges Datum (z.B. 13.01.2014) umzuwandeln. Hierfür wurde TempEx [MW00] verwendet. TempEx funktioniert zu großen Teilen regelbasiert und wurde nicht speziell für die Anwendung auf Twitterdaten entwickelt. Da zeitliche Ausdrücke jedoch relativ eindeutig sind, wird trotzdem auch hier eine gute Leistung erreicht.

## 3.2 Extraktion von Entitäten mit Hilfe einer Wissensdatenbank

Neben der Ausnutzung von Sprachstatistik und der Einführung von Regeln zur Extraktion bestimmter Informationen ist das Nutzen eines Wörterbuchs eine weitere Möglichkeit, um bestimmte Informationen aus Text zu extrahieren. Ein Beispiel hierfür ist die im vorherigen Abschnitt vorgestellte Koponente *NER*, die eine Ontologie nutzt, um den Typ einer Named Enitity zu bestimmen. Ein ähnliche Methode haben Gattani et al. [GLG+13] gewählt, um Named Entities aus Tweets zu extrahieren.

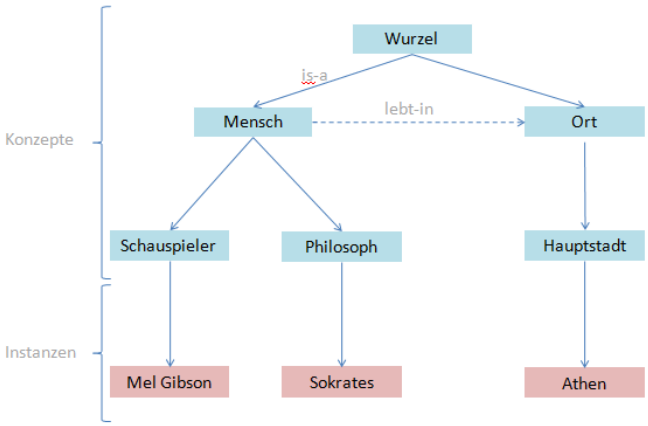


Abbildung 3.2: Beispiel einer kleinen Wissensdatenbank, angelehnt an [GLG+13]

Bei ihrer Methode werden Named Enities anhand einer Wissensdatenbank aus einem Text extrahiert und mit einem Knoten aus der Wissensdatenbank verbunden. Die Wissensdatenbank enthält eine Menge von Konzepten und eine Menge von Instanzen, die miteinander verbunden sind (ein Beispiel einer kleinen Wissensdatenbank ist in Abbildung 3.2 zu sehen). Die Wissensdatenbank wurde

dabei basiererend auf Wikipedia geschaffen, was den Vorteil bietet, dass dort die meisten wichtigen Konzepte und Instanzen enthalten sind. Ein weiter Vorteil ist die hohe Aktualität von Wikipedia.

Nach einigen Vorverarbeitungsschritten, die auf dem Tweet durchgeführt werden, werden die Entitäten aus dem Text extrahiert. Hierzu werden in dem Text Zeichenketten gesucht, die mit Knotennamen der Wissensdatenbank übereinstimmen. Weiterhin werden durch eine Verlinkung zu den Knoten in der Datenbank weitere Informationen zur extrahierten Entität hinzugefügt.

Ein Nachteil dieser Methode ist, das lediglich Entitäten, die in der Datenbank enthalten sind, extrahiert werden. Auch Synonyme können nur erkannt werden, indem ein Knoten in der Datenbank mit möglichst vielen Synonymen verlinkt wird (z.B. Barack Obama mit Obama, BO, Barack, etc.).

# 4 Zusammenfassung

Der Begriff *Big Data* wird im Allgemeinen nur mit der stetig steigenden Datenflut verbunden, die es mit neuen Informationssystemen zu verarbeiten gilt. Sehr gut lässt sich *Big Data* jedoch mit dem 3-V Modell beschreiben, bei dem das steigende Datenvolumen nur einen Aspekt der Herausforderungen ausmacht. Die anderen beiden Aspekten sind die steigende Geschwindigkeit, mit der Daten erzeugt werden und verarbeitet werden müssen, und die Vielfalt (Variety) der zu analysierenden Daten. Auf den letzteren Aspekt wurde im weiteren Verlauf dieser Arbeit der Schwerpunkt gelegt.

Da im Rahmen von Variety auch Informationen , die in semistrukturierten und unstrukturierten Daten vorhanden sind, gemeinsam mit relationalen Daten verarbeitet werden sollen, müssen die in ihnen enthaltenen Informationen zunächst in eine von Maschinen verarbeitbare Form gebracht werden.

Hierzu wurden im Bereich von Multimediadaten Arbeiten vorgestellt, die sich mit dem automatischen Tagging von Bildern und der Ausarbeitung einer auf Nachrichtensendungen ausgerichteten Ontologie für das Training von semantischen Klassifikatoren beschäftigt haben.

Im Bereich Text wurden Arbeiten genannt, die sich mit der Extrahierung von Informationen aus Twitter Tweets beschäftigt haben. Hier wurden bekannte Verfahren aus dem Bereich des Text Mining adaptiert und an die besonderen Eigenschaften von Tweets angepasst.

Insgesamt sind in beiden Bereichen eine Vielzahl von Verfahren bekannt, um unstrukturierte Daten zu strukturieren. Bis zum jetzigen Zeitpunkt, ist die Erfassung der Semantik jedoch aufwändig und nur begrenzt möglich.

# Literaturverzeichnis


[BKI06] Christoph Beierle and Gabriele Kern-Isberner. *Methoden wissensbasierter Systeme*. Vieweg, Wiesbaden, 2006.

[Fry07] Ben Fry. *Visualizing Data*. O'Reilly Media, 2007.

[FS06] Ronen Feldman and James Sanger. *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. Cambridge University Press, 2006.

[GLG+13] Abhishek Gattani, Digvijay S. Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, Sanjib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. *Proc. VLDB Endow.*, 6(11):1126–1137, August 2013.

[HQW08] Gerhard Heyer, Uwe Quasthoff, and Thomas Witting. *Text Mining: Wissensrohstoff Text, Konzepte, Algorithmen, Ergebisse*. W3L-Verlag, Bochum, 2008.

[IZE11] IBM, Paul Zikopoulos, and Chris Eaton. *Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data*. McGraw-Hill Osborne Media, 1st edition, 2011.

[KTGH13] Dominik Klein, Phuoc Tran-Gia, and Matthias Hartmann. Big Data. http://www.gi.de/nc/service/informatiklexikon/detailansicht/article/big-data.html, 2013. abgerufen am: 19.01.2014.

[Lan01] Doug Laney. 3d data management: Controlling data volume, velocity, and variety. 2001.

[MW00] Inderjeet Mani and George Wilson. Robust temporal processing of news. In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, ACL '00, pages 69–76, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[MZLK10] Hao Ma, Jianke Zhu, M. R.-T. Lyu, and I. King. Bridging the semantic gap between image contents and tags. *Trans. Multi.*, 12(5):462–473, August 2010.

[NST+06] Milind Naphade, John R. Smith, Jelena Tesic, Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Alexander Hauptmann, and Jon Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, July 2006.

[RCME11] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[RMEC12] Alan Ritter, Mausam, Oren Etzioni, and Sam Clark. Open domain event extraction from twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1104–1112, New York, NY, USA, 2012. ACM.

[Smi12] J.R. Smith. Mindingthegap. 2012.

[Smi13] John R. Smith. Riding the multimedia big data wave. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '13, pages 1–2, New York, NY, USA, 2013. ACM.