SINNVOLLE DATENERHEBUNG FÜR SMART STUDIES

Lukas Kozlowski,
Hamburg University of Applied Sciences, Dept. Computer Science,
Berliner Tor 7
20099 Hamburg, Germany
lukas.kozlowski@haw-hamburg.de

ABSTRACT

In diesem Paper wird der Einfluss der Digitalisierung mit dem aktuellen Stand der Lehre verglichen. Dabei werden bereits bestehende Lehrplattformen analysiert und diskutiert. Innerhalb dieser Diskussion werden anhand verschiedener Argumente, Defizite in verschiedenen Bereichen der Lehre nachgewiesen. Einer dieser Defizite zielt auf die Individualisierung der Lehre. Auf Grund dieser bestehenden Defizite soll eine neue eigenständige Lösung entwickelt werden. Die Problemlösung bezieht sich auf die Anforderungen der Individualität, der Effizienz und der Unterstützung innerhalb der Lehre. Diese Lösung stellt den Studenten innerhalb seiner Architektur in das Zentrum, als ein zentrales Objekt dar. Dem Studenten sollen neue Möglichkeiten gegeben werden, um dadurch freier, effizienter und gemeinschaftlicher arbeiten zu können. Das Primärziel ist ein individuelles Studium mit adaptiven Vorgaben zu realisieren, um einen großen Freiraum für Produktivität und Effizienz zu generieren. Im weiteren Verlauf soll eine unterstützende Plattform entstehen die, die Sammlung, Aufarbeitung und die Verknüpfung aller im Zusammenhang stehender relevanten Daten ermöglicht, kontrolliert und bezogen auf den einzelnen Studenten auswertet.

1. MOTIVATION UND ANSATZ

Eine Vielzahl an Studiengängen folgt seit längerer Zeit einem Individualisierungsprozess, um eine Vielschichtigkeit innerhalb des Fachgebiets zu ermöglichen. Dieser aktuelle Trend bringt nicht nur positive, sondern auch negative Aspekte mit sich. Der Studiengang der Informatik hat sich Aufgrund dieses starken Individualisierungsprozess gespalten, weiterentwickelt und spezialisiert. Dadurch sind einzelne und feinere Facetten innerhalb der Informatik entstanden wie z.B.:

- Angewandte Informatik
- Wirtschaftsinformatik
- Technische Informatik
- Computer Science

Das Clustering der Informatikstudiengänge ist ein Weg in die richtige Richtung, dennoch führt dieser Weg derzeit zu einer flächendeckenden Ausbildungsstrategie, welche sich negativ auf den Arbeitsmarkt widerspiegelt. Der Arbeitsmarkt wird zum größten Teil von wirtschaftlichen oder wissenschaftlichen Betrieben genutzt um Spezialisten, die sich besonders durch ihr individuelles Fachwissen auszeichnen, zu werben. Daher sollte die Modernisierung der Lehrmethoden der Ausgangspunkt für dieses Ziel sein und die Individualisierung der Lehre wesentlich stärker vorangetrieben werden als es bisher angestrebt wurde, um die Lehre auf eine neue fortschrittlichere Ebene zu befördern.

Feste Strukturen diktieren derzeit den Studienverlauf eines einzelnen Studenten. Der Student hat seine Vorlesungszeit und Klausurtermine, an die er sich strikt halten muss. Der Lehrraum, die Lehrkraft und die vorhandenen Ressourcen sind an eine Fakultät gebunden. Der Lehrstoff wird meistens anhand eines grundsätzlichen Ablaufplans mehrere Semester in Folge vermittelt. Es wird nicht darauf eingegangen in welcher Verfassung, Situation sich der einzelne Student innerhalb seiner Lehrzeit befindet oder ob er irgendwelche spezifischen Stärken oder Schwächen in das Studium mitbringt. Durch diese Struktur tritt die Individualität des einzelnen in den Hintergrund. Synergien werden nicht ermittelt und nicht weiter gefördert oder berücksichtigt.

Im Laufe der Zeit sind diverse technologische Ansätze in der Form von E-Learning-Plattformen entstanden, wie z.B. Wikipedia, Moodle, Emil, usw. Diese Lernplattformen zeichnen sich durch ihre Leitfunktionen aus wie Adaptivität, Interaktivität, Kollaboration, Distribution oder Multimodalität aus. Diese Plattformen sind gut darin jegliche Informationsformen bereitzustellen oder auszutauschen aber folgen festen Konzepten, die eine individuelle Lehre nicht berücksichtigen, sondern mehr auf ein größeres Kollektiv abzielt. Ein anderer Ansatz der Lehrtechnik ist durch Massive Open Online Course (MOOCS) und dessen Weiterentwicklung Small private online course (SPOCS) realisierbar, aber auch diese Ansätze sind bereits zum Teil massiv gescheitert und haben sich derzeit nicht durchgesetzt. [1]

Es muss eine Veränderung innerhalb der Lehre stattfinden, um diese Defizite zu beseitigen. Bei diesen Vorhaben, welches eine Umstrukturierung innerhalb der Lehre hervorruft, fehlt in der derzeitigen Struktur eine unterstützende Lehrplattform. Diese Lehrplattform berücksichtigt wichtige Aspekte wie die Stärken und Schwächen eines Studenten. Dabei werden individuelle Daten erhoben, analysiert, zusammenfasst und auswertet. Dadurch wird die besondere Stärke der Individualität des einzelnen miteinbezogen und mehr gefördert als in den bisherigen passiven Herangehensweisen. Der Student soll aktiver sein und zur Generierung von neuen Inhalten beitragen, die innerhalb dieser Lehrplattform verwendet werden. Der Fokus muss gezielter auf den einzelnen Menschen liegen der sein individuelles Studium beschreitet und um diesen herum muss sich das derzeitige System verändern oder anpassen.

2. ZIELSETZUNG

Das individuelle Studium kann effizienter gestaltet werden. Diese Zielformulierung wird anhand einer Prozesskettenstruktur verdeutlicht siehe Abb.1.

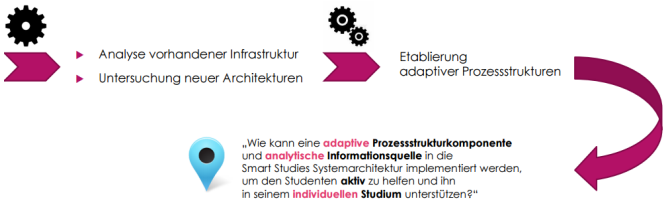

Abb.1 Prozesskette

Am Anfang dieser Kette steht die Frage, wie eine sinnvolle Infrastruktur für das individuelle Studium gestaltet wird. Der darauffolgende Schritt in dieser Kette ist die Entscheidung, ob eine vorhandene Infrastruktur untersucht wird oder ob eine neue Architektur für diese Aufgabenstellung entworfen wird. Innerhalb dieses Prozessschrittes kommt es zur Etablierung von adaptiver Prozessstrukturen. Die wesentliche Ablauf dabei ist der Entwurf einer adaptiven Komponente. Diese Komponente konzentriert sich auf:

- Die Auswertung und Analyse von Informationen
- Studentenunterstützung durch aktive Hilfe
- Berücksichtigung der Individualität

Es soll eine neue Struktur entstehen, die sich von der heutigen Lehrstruktur, in der Art und Weiser der Lehre deutlich unterscheidet. Die Individualität eines einzelnen Studenten soll dabei in den Vordergrund treten. Innerhalb dieser Struktur soll der Student, ohne Einschränkungen und größerer Einarbeitung sich schnell zurechtfinden, indem er weniger festen oder starren Konzepten folgen muss. Der Student wird dabei durch ein aktives Mentoring der anderen erfahrenen Teilnehmer unterstützt werden. Zu diesem Prinzip gehört das Feedback und angepasste Aufgaben die von der Lehrplattform an den Studenten gestellt werden. Diese Struktur soll, als ein eigenständiges System seine Arbeit im Hintergrund verrichten, während der Student sich einzeln und allein auf sein Studium konzentriert.

3. DATENERHEBUNGEN

Der Weg in das Forschungsfeld der Datenerhebung beginnt bei der Datengenerierung. Dieser Abschnitt soll dazu dienen, die Fülle von potenziellen Datenquellen aufzuzeigen, die innerhalb dieses neuen Lehrsystems eingesetzt werden.

Zur Digitalisierung gehört eine immense Informationsverbreitung. Google indexierte 1998 ca. 1.Million individuelle Internetseiten. Im Jahre 2000 waren es 1.Milliarde und 2016 sind es ca. 40 Billionen individuelle Interseiten. Der Suchalgorthimus von Google wird täglich 3.5 Milliarden Mal aufgerufen. Mit dem Internet of Things werden immer mehr Haushaltsgeräte mit Chips bestückt und dadurch kommunikationsfähig mit dem Internet verbunden. Die Quellen, die verschiedenen Informationen generieren, werden hierarchisch gegliedert. [2]

3.1. Datenquellen

Im Jahre 1989 wurde das World Wide Web von Sir Tim Berners-Lee erschlossen. Die erste Informationsnutzung vollzog sich auf reinen statischen HTML Seiten. Die Anzahl der Provider, die eine Internetseite zur Verfügung stellen ist seitdem stark gestiegen, deshalb wird die Anzahl der Internetseiten die wir heute verbuchen täglich neu definiert. CMS-Systeme, Webshops, Blogs, Webapplikationen und intelligente Plattformen, die eigenständige Informationsschnittstellen in der Form von API's anbieten, sind der heutige Standard. Ein anderes Beispiel sind Intelligente Informationssysteme, wie z.B. ein Umweltinformationssystem, welches auf Daten mit einer Vielzahl an Messstationen täglich zugreift, um die Temperatur, den Verkehr oder spezifische Feinstaubinformation abzurufen. All diese Quellen generieren täglich Unmengen an Text-Bild- Video- und Audiodaten, die dem Nutzer zur Verfügung stehen. Dadurch ergibt sich eine Vielzahl an potenziellen Datenquellen. Die Komplexität besteht darin diese Daten in einem logischen Zusammenhang mit der Lehre zu verbinden und zu fördern.

3.1.1. Medien

Der gesamte Datenpool besteht zu einem großen Anteil aus Mediendaten, egal ob im klassischen Sinne durch das Fernsehen, Radio und die Zeitung oder die Weiterentwicklung der Medien durch Social Media wie, Facebook, YouTube oder Twitter. Die Medienkommunikation hat Tag für Tag immer größere Ausprägungen innerhalb der Datenverbreitung angenommen. Twitter verschickt ca. 500 Millionen Tweets täglich. Der Kurznachrichtendienst WhatsApp versendet 64 Milliarden Kurznachrichten am Tag. Auf der Facebookseite werden täglich bis zu 800 Millionen Seiten aktualisiert. Flickr wird 3.5 Millionen Mal täglich für Bilderuploads verwendet die mehrere Megabyte groß sind. YouTube hat ca. 5 Milliarden Videozuschauer an einem Tag. Daraus ergibt sich ein großes Datenpotenzial, die Informationen, die aus Social-Media-Kanälen stammen, logisch zu interpretieren und auszuwerten. Die Meinung der Studenten zu einem bestimmten Lehrthema oder Lehrvideos könnten hierbei eine größere Rolle spielen. [2]

3.1.2. Sensoren

Eine andere immense Datenquelle wird durch die Variation unterschiedlicher Sensoren generiert. Bei der derzeitigen Erdbevölkerung von 7,39 Milliarden Menschen besitzen ca. 2,2 Milliarden Menschen ein Smartphone, also jeder 4 Mensch auf der Welt besitzt bzw. verwendet ein Smartphone. Ein Smartphone besteht durchschnittlich aus 10 unterschiedlichen Sensoren:

Abb.2 Beispiel: Smartphone-Sensoren [3]

Das Sensorenkollektiv generiert bei der Benutzung wertvolle Informationsdaten, die von unterschiedlichen Applikationen interpretiert und ausgewertet werden. Durch die Anzahl der Sensoren und der verschiedenen Möglichkeiten eignet sich ein Smartphone immer als zuverlässige Komponente, die ein wichtiger Bestandteil im neuen Lehrsystem werden sollte. Daher sollte diese Komponente gerade im Bezug auf Individualität hilfreich sein.

3.1.3. Internet of Things

Immer mehr herkömmliche Geräte erhalten einen Chip mit den Auswertungen und die Kommunikation zu dem Gerät selbst ermöglicht wird. Die zukünftige Vision dieser Geräte ist, dass diese selbständig agieren. Dadurch wird ein neues Feld von potenziellen Datenquellen erschlossen. Kann ein Spiegel der mit einen raspberry pi Chip ausgestattet ist in der Zukunft erkennen, ob ein Student krank wird oder sein derzeitiges Befinden auslesen und analysieren? Das Internet of Things (IoT) besitzt bereits derzeit ein großes Potenzial. Es sollte daher weiter analysiert und erforscht werden, ob bereits Bestandteile aus diesem Gebiet eine Verwendung innerhalb der zukünftigen Lehrplattform finden.

Alle diese genannten Quellen erzeugen täglich Daten in einer Größenordnung von Zettabytes. Die jährliche Zuwachsprognose für 2016 beläuft sich auf weitere 50%. Die positive Seite dieser Entwicklung ist, dass der Durchschnittsbürger täglich mehr Zugriff auf Informationsdaten hat, als ein Individuum aus dem 16. Jahrhundert in seinem gesamten Leben. Die negative Seite bezieht sich auf den Aspekt der Data-Pollution und der Data-Exhaustion. Der Informationszuwachs hält sich seit 30 Jahren an das Mooresche Gesetz, das besagt, dass 90% der gesamten Informationsdaten innerhalb der letzten zwei Jahren entstand sind. Sobald die richtigen Datenquellen ausgewählt worden sind folgt als nächster Schritt die Modellierung der Datenstruktur. [2]

3.2. Datenstruktur

Sind profitablen Datenquellen für die zukünftige Lehre ermittelt, findet eine Kategorisierung der Daten statt, in der genauestens auf die Datenform eingegangen wird. Grundsätzlich werden die Daten in zwei verschiedenen Datenkategorien bei einer Datenerhebung separiert.

3.2.1. strukturierte Daten

Diese Datenform ist leichter zu ermitteln und auszuwerten, als die unstrukturierten Daten. Diese Kategorie bezieht sich auf alle Zahlen oder Wörter, die kategorisierbar sind und dadurch schneller analysiert werden. Wie z.B. Transatkionsdaten, die Sensoren eines Smartphones oder Datenbankabfragen. Für die logische Interpretation dieser Daten bedarf es keines größeren Aufwands. Für die Erhebung dieser Daten, reicht die Anwendung von konventionellen Data Mining Werkzeugen.

3.2.2. unstrukturierte Daten

Diese Strukturform bezieht sich auf alle Daten die nicht ermittelbar sind. Diese Daten können nicht numerisch kategorisiert werden, weil die Aussage der Daten vielseitig interpretierbar ist. Es sind Kundenreviews, Photos, Multimediadaten oder Kommentare. An dieser Stelle reichen konventionelle Miningwerkzeuge nicht mehr aus. Es muss jetzt eine Mining Technik spezifiziert werden, wie z.B. Text Mining [4], um einen gezielten Bereich zu fördern. Die Datenabfrage findet an dieser Stelle nicht wie gewohnt über ein Datenbanksystem statt, sondern muss über einen anderen Weg geführt werden. Dieser Weg führt über Inhaltsabfragen durch Suchmaschinen oder CMS-Systeme. Beim Text Mining wird nicht nur der Text nach bestimmten Schlüsselwörtern durchsucht, sondern der Satzbau und die Wortarten durch ein logisches Verfahren ausgewertet und interpretiert. [5]

Diese beiden Datenstrukturen sind in der Lage eine grobe Unterteilung der erhobenen Daten zu ermöglichen. Eine feinere Granularität findet über das Heterogeneous ,Autonomous ,Complex and Evolving (HACE) Theorem statt. Dieses Theorem wird hauptsächlich für Big Data Mining Anwendungen verwendet. Das Big Data Mining ist nur eine Skalierung von Data Mining, also sind alle Dateneigenschaften erblich.

3.2.3. H.A.C.E. Theorem

"Big data starts with large volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data. [6] „

Durch dieses Theorem werden die wichtigsten Eigenschaften innerhalb einer Datenerhebung deutlicher definiert. Dadurch entsteht eine V-Charakteristik, diese wird zur Datenmodellierung und einer feineren Strukturierung der Daten benötigt. Diese Charakteristik unterstützt die weitere Bearbeitung der Daten und bildet die Modellierung der Kerneigenschaften, wie die Heterogenität, die Dezentralisierung von autonomen Datenquellen und die Komplexität, die bei einer Datenerhebung entsteht.

Abb.3 V-Charakteristik

Fünf Eigenschaften bilden ein gemeinsames Datenmodell. Dadurch entstehen gleichzeitig fünf neue Aspekte, die sich selbst auf die Datenerhebung beziehen und somit bei Planung für die zukünftige Lehrplattform mitbedacht werden müssen.

3.2.4. Value

Diese Eigenschaft spiegelt den eigentlichen Wert der Daten wieder. Diese Dateneigenschaft ermöglicht wissenschaftlichen Betrieben und Firmen Entscheidungen zu treffen, die vorher außerhalb der Entscheidungsreichweite lagen, aber durch die Auswertung der Daten ermöglicht wurden. Mit Hilfe dieser Dateneigenschaft und der Einschätzung eines Domänenexperten können brauchbare Daten ermittelt werden. [6]

3.2.5. Velocity

Velocity bezieht sich darauf wie schnell die Daten von A nach B fließen. Eine Aktienbörse generiert täglich 1 TB an Informationsdaten. Diese Daten werden innerhalb eines schnell fließenden Informationsstroms ausgewertet und interpretiert. Für diese Aufgabe sind spezielle NewSQL Datenbanken wie VoltDB notwendig, die innerhalb der Architektur implementiert werden. Diese Datenbanken sind in der Lage schnell fließende Datenströme in kürzester Zeit auszuwerten. Sollte z.B. eine Umfrage innerhalb der Lehrplattform stattfinden an dem mehre tausende Teilnehmer partizipieren, dann spielt diese Eigenschaft eine große Rolle. [6]

3.2.6. Volume

Bezieht sich auf die relative Datengröße, also nicht nur auf wie z.B. Gigabyte, Terrabyte oder Megabyte, sondern auch auf die Angaben wie die Anzahl der Objekte oder die Anzahl der Spalten bei einer Datenbankabfrage. Diese Eigenschaft wird auch als die vertikale Datendimension bezeichnet. [6]

3.2.7. Variety

Diese Eigenschaft bezieht sich auf die Art und Weise, in welcher Form die Daten angefragt werden. Es wird konkretisiert, ob die Daten aus einer relationalen Datenbank entnommen werden oder ob ein NoSQL System zum Einsatz kommt. Bei dieser Eigenschaft steht die Gliederung der Daten im Vordergrund. Es wird an dieser Stelle separiert ob es sich, um strukturierte oder unstrukturierte Daten handelt. [6]

3.2.8. Veracity

Die letzte Eigenschaft bezieht sich auf die Richtigkeit der Daten und prüft dessen Wertgehalt gegen Inkonsistenz. Fehlerhafte Daten erzeugen immer große Probleme bei der weiteren Bearbeitung. Daher ist diese Eigenschaft besonders relevant. [6]

3.3. Die Architketur einer Datenerhebung

Es existieren viele verschiedene Datenförderungsstechnicken, wie Text Mining, Graph Mining etc. um eine bestimmte Datensorte gezielt zu erheben. Jeder dieser Mining Techniken hat andere Anforderungen an die Mining Plattform und fordert nicht nur das Fachwissen über die Miningapplikation, sondern auch wichtige Bestandteile wie den Datenschutz und einen Domänenexperte. Daraus bildet sich eine hierarchische Systemarchitektur, die aus drei unterschiedlichen Ebenen besteht:

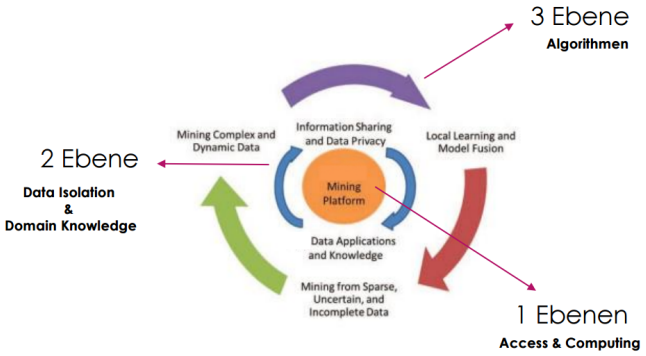

Abb.4 Datenerhebung - Ebenen und Arbeitszyklen [7]

3.3.1. Access & Computing

Im Kern der Mining Plattform, auf der ersten Ebene, findet die Berechnung des Algorithmus statt, wie z.B. durch MapReduce. Es können dabei auch parallele Berechnung stattfinden, wenn eine größere Datenmenge, schneller ausgewertet werden soll. Dabei wird ein Cluster von Computern verwendet, dass sich die Gesamtberechnung aufteilt und somit schneller den laufenden Auswertungsprozess erledigt. [7]

3.3.2. Data Isolation & Domain Knowlegde

In dieser zweiten Ebene wird der Datenschutz berücksichtigt. Dieser kann durch zwei verschiedene Anwendungsformen erreicht werden. Die erste Form ist, dass die sensiblen Daten sofort nach der Erhebung gelöscht werden. Die zweite Form besteht aus einer zertifizierten Zugriffskontrolle von einer bestimmten Gruppe, die die Daten einsehen kann. Der zweite Teil dieser Ebene sind die Domänenkenntnisse, diese werden durch einen Fachmann der Domäne erlangt. Dieser Fachmann kennt sowohl das Einsatzgebiet auf die sich die Datenerhebung bezieht, als auch die relevanten Daten die erhoben werden sollten. Das Wissen des Fachmanns wird bei der Modellierung der Algorithmen verwendet, die bei der Datenerhebung zum Einsatz kommen. [7]

3.3.3. Datenstruktur & Algorithmen

In der dritten Ebene findet die Ausrichtung statt auf die betroffenen Datenquellen, die verwendet werden sollen. Es findet dabei eine Unterscheidung statt, ob es sich bei den erhobenen Daten um strukturierte oder unstrukturierte Daten handelt. Darauf hin wird eine feinere Granulierung durch den Einsatz der V-Charakteristik vorgenommen [7]

3.4. Spezialisierung auf Text Mining

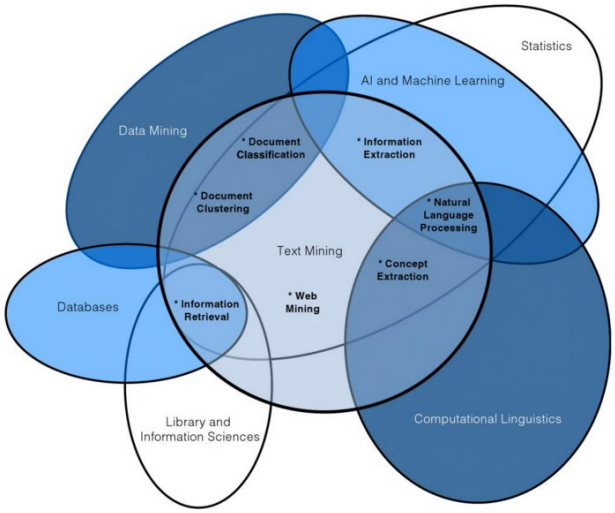

Abb.5 Text Mining Ecosystem [5]

Text repräsentiert Wissen, egal durch welche Inhaltsform ob durch Fachbücher, ein wissenschaftlicher Aufsatz, Lexika und Enzyklopädien, Produktbeschreibungen, Normen, Gesetze, Kommentare oder Verträge. Daten in Textform bilden 80% der gesamten Informationen, die ein wissenschaftliches oder betriebliches Unternehmen täglich verwendet. Aus diesen Grund ist es am sinnvollsten und am effizientesten mit dieser Mining Technik zu beginnen und zu einem späteren Zeitpunkt bei der Weiterentwicklung des Systems andere Mining Techniken wie z.B. Graph Mining zu integrieren. Das Forschungsfeld von Text Mining ist sehr komplex, es existieren dabei Überschneidungen mit verschiedenen anderen Forschungsbereichen (siehe Abb.5), sodass auch diese Bereiche in Planung und Forschung der neuen Lehrplattform miteinbezogen werden. Interdisziplinarität ist ein Bestandteil von Text Mining der bei der Planung und Umsetzung berücksichtigt werden muss.

Die Lehrplattform soll zu Beginn auf der Basis von Text Mining Techniken, als Prototyp realisiert werden, der gezielt brauchbare Daten aus direkten Klausurfragebögen fördert und diese Textinhalte interpretiert. Die Vorteile von Text Mining sind, dass Anwendungsdomäne relevante und spezifische Fachausdrücke identifiziert werden und dadurch logisch sortiert und analysiert werden können. Die angewandte Filterung kann dabei so feingranular sein, dass sogar Ähnlichkeiten innerhalb von Dokumenten oder Wörtern ausgewertet werden können. Ein weiterer Vorteil dieser Technik ist eine saubere Gliederung nach Erläuterungen, Definitionen und Referenzen. Der wichtigste Vorteil ist, dass eine semantische Relation zwischen den einzelnen Strukturen stattfindet. Durch diese Relation können inhaltliche Strukturen innerhalb von verschiedenen Texten ausgewertet oder verglichen werden.

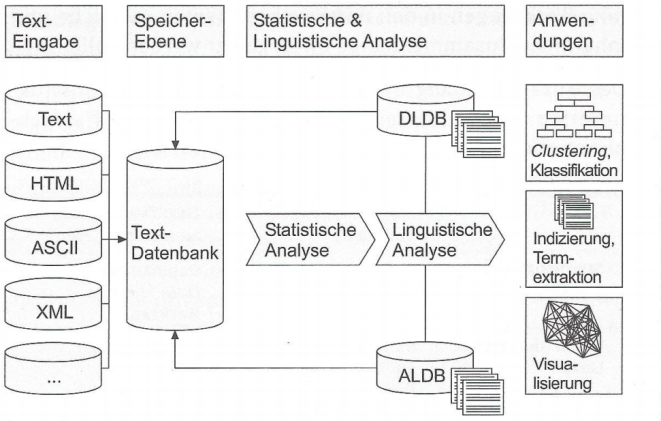

Abb.6 Text Mining Architektur [8]

Die Architektur besteht aus vier verschiedenen Komponenten:

- Eingabe
- Datenspeicherung
- Analysen
- Auswertung

3.4.1. Eingabe

Bei diesem Verarbeitungsschritt können verschiedene Textressourcen von unterschiedlichsten Datenquellen eingebunden werden. Bei diesen Vorgang wird jedes korrekte Textformat wie HTML, XML oder ASCII akzeptiert und innerhalb eines Datenbanksystems abgelegt. Der Text ist dabei weit aus mehr als nur ein reiner *string*, was im späteren Verlauf durch die vielen komplexen Analysen verdeutlicht wird. [8]

3.4.2. Datenspeicherung

Eine klassische Aufteilung eines Datenbanksystems für Text Mining besteht aus drei Teilen. Der erste Teil sind die Datenbanktabellen mit dem die Rohdaten, die durch eine Datenerhebung erhoben wurden. Der zweite Teil sind bestimmte Tabellen aus einer domänenspezifischen linguistischen Datenbank und der dritte Bestandteil sind Tabellen aus einer Allgemeinen linguistischen Datenbank. Durch die im weiteren Verlauf ein Abgleich und eine Analyse erfolgt. [8]

3.4.3. Analysen

Die Komplexität von Text Mining liegt innerhalb zwei unterschiedlichen Analysen. Die erste Analyse ist die statische Analyse. Diese Analyse hat eine Vielzahl von Konzepten. Einer dieser Konzepte basiert auf der Unabhängigkeit und vergleicht, ob bestimme Wörter, Texte oder Kollokationen unabhängig voneinander sind. Ein anderes Konzept der statischen Analyse ist die Variabilität. In diesem Konzept wird die Länge der Sätze verglichen oder welche Wortformen verwendet werden. Es werden daraufhin weitere Stichproben aus dem Text entnommen, ob eine Repräsentativität oder eine Verallgemeinerung stattfindet. Es finden Vergleiche statt, die eine bestimmte Statistik erzeugen, die daraufhin weiterverarbeitet wird. [8]

Die zweite Analyse ist die linguistische Analyse. Innerhalb dieser Analyse werden verschiedene hirarchische Ebenen sog. linguistische Ebenen verwendet, um eine Textrelation herzustellen und um eine automatische Ermittlung semantischer Zusammenhänge zu ermöglichen. Dieses Wissen beruht sich auf Ferdinand de Saussure, der syntagmatische und paradigmatische Relationen beschreibt. Die weiteren drei Relationen die innerhalb einer linguistischen Analyse stattfinden sind die logische, die semantische und die Relation der Fach- und Allgemeinsprache. [8]

3.4.4. Auswertung

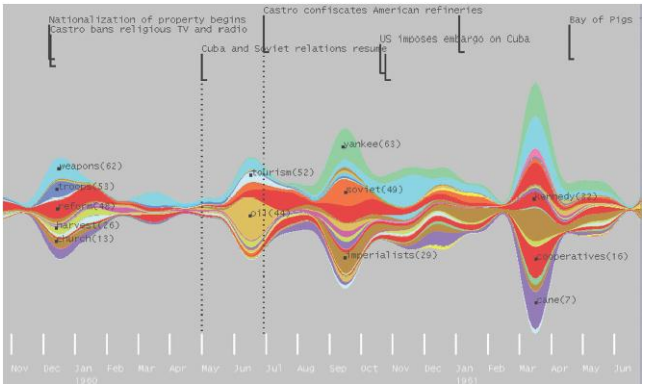

Abb.7 Reden von Fidel Castro 1959 bis 1961 [9]

Bei der Auswertung ist es wichtig, in welcher Form das Ergebnis und für welche Anwendung es bereitgestellt wird. Der erste Anwendugsblock in Abb.6 besteht aus der Klassifikation und des Clusterings, dabei werden Texte automatisch vordefinierten Kategorien zugeordnet. Durch eine Clusteranalyse findet die Segmentierung statt, dabei werden ähnliche Texte bestimmten Gruppen zugeordnet oder zusammengeführt. Der zweite Anwendungsblock ist die Indizierung und Termextraktion. Bei der Termextraktion werden durch eine Abhängigkeitsanalyse die gemeinsamen auftretenden Terme und die Beziehungen zwischen einzelner Dokumente ausgewertet. Bei der Indizierung findet die Trennung zwischen Meta- und Nutzdaten statt, dadurch werden in einem Schritt aus unstrukturierten Dokumenten, strukturierte Metadaten generiert. Die extrahierten Terme dienen dabei als Deskriptoren und werden daraufhin Indexiert. Der letzte Anwendungsblock dient der Visualisierung. An dieser Stelle werden verschiedene Visualisierungstechniken verwendet, um ein Auswertungsergebnis zu repräsentieren. Diese Techniken basieren auf Beschreibungsvektoren, dadurch können Texte in einem multidimensionalen Raum als Grafik dargestellt und interpretiert werden. Ein Beispiel des Potenzials dieser Visualisierung basiert auf der Visualisierungsanwendung ThemeRiver™ [9]. In diesen Beispiel werden alle gesprochenen Reden von Fidel Castro in dem Zeitraum von 1959 bis 1961 analysiert und können zu den jeweiligen Jahren logisch ausgewertet und im Zusammenhang mit anderen zeitlichen Geschehnissen interpretiert werden siehe Abb.7. [8]

4. AKTUELLER FORSCHUNGSSTAND

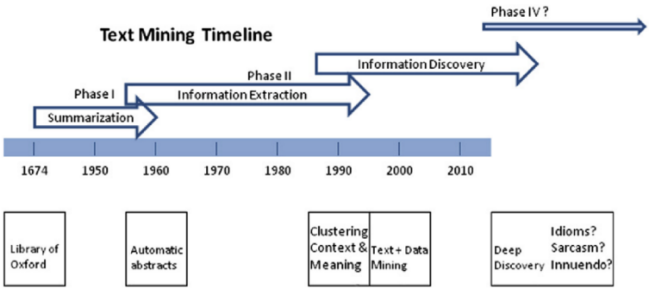

Abb.8 Forschungsstand Text Mining [5]

Text Mining hat bereits eine sehr lange Entwicklung hinter sich. Dieser Weg lässt sich anhand von drei Phasen definieren. Die erste Phase von 1674 bis 1960 galt als die Phase der Zusammenfassung. Diese Phase wurde in zwei weitere Prozesse gegliedert, *information retrival*, dieser Prozess konzertierte sich auf die Unterscheidung unterschiedlicher Dokumente, um eine logische Anfrage zu beantworten. Der zweite Prozess war *information extraction* und diente dazu spezifische Informationen aus einem Dokument zu fördern. Dadurch entstand z.B. in der Bibliothek von Oxford eine der ersten Bibliothekssysteme, dass eine Zusammenfassung und Klassifikation von verschiedenen Büchern unterstützte. Ab 1989 bis ca. 1998 begann die Phase der Informationsförderung. Innerhalb dieser Phase stand das *Web Mining* im Mittelpunkt und der Fokus war auf ein effizientes *Web Crawling* gerichtet. Dabei entstanden die ersten Suchmaschinensysteme. Das Hauptziel war es ein System zu Entwickeln, dass eine Vorhersage treffen konnte, ob bestimmte Links dem gesuchten Thema entsprachen, um effizienter das Internet zu nutzen und um zu verhindern das alle unnötigen Internetseiten, die dem gesuchtem Thema nicht entsprachen, jedes Mal besucht wurden. Die dritte Phase wird als *Information Discovery* bezeichnet und bildet die Anfänge der Forschung von Data Mining und Text Mining. Die Text Mining Forschung hat sich seit dem Jahr 1990 stark verändert. Dabei sind viele neue Mining Techniken entstanden, um Texte zu fördern. [5]

Die erste Innovation galt den Information Extraction Engines (IEE), diese erhielten eine neue Komponente das sog. *tokenization module*. Diese Modul ist in der Lage das eingegebene Dokument aufzuteilen nach Wörtern, Phrasen, Sätzen oder Paragraphen. Die Aufgabe innerhalb dieses Moduls wird als *zoning* bezeichnet. Nach diesen Modul folgt eine weiteres Modul, dass eine *morphological* und eine *lexical* Analyse durchführt. Dieses Modul kennzeichnet alle vorkommenden Textterme und führt eine Eindeutigkeit für alle Wörter und Redewendungen durch. Es folgt das *sementic analysis* Modul. Die Aufgaben dieses Moduls beziehen sich sowohl auf *shallow parsing*,als auch *chunking* bezeichnet, als auch auf *deep parsing*. Bei einem oberflächlichem *parsen*, dem *shallow parsing*, wird ein Satz nach Verben, Begriffen oder Gruppen gefiltert. Das *deep parsing* als die tiefere Analyse, liefert mehr Information zum Satzbau, wie z.B. die Semantik eines Satzes oder Textabschnitts. Sie dient dazu einen tieferen Einblick zu liefern, jenseits dessen was *shallow parsing* liefern kann. Die letzte Komponente bezieht sich auf die Domänenanalyse. Dieses Modul verwendet die *anaphora resolution*. Diese Aufgabe beschäftigt sich mit der Problematik, worauf sich ein Pronomen oder eine Nominalphrase, innerhalb eines Satzes bezieht. All diese Module sind jetzt Bestandteile eines IEE. [5]

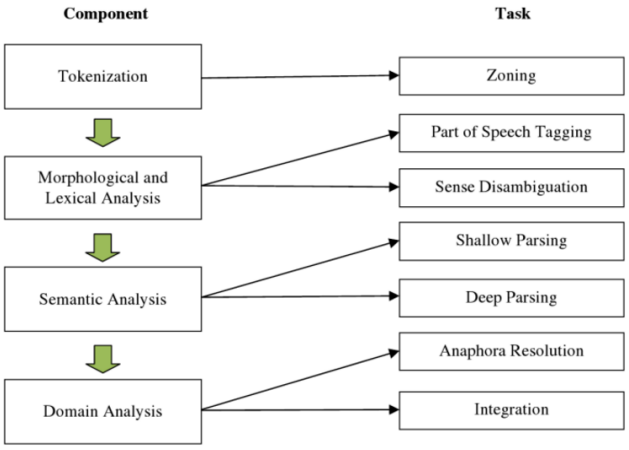

Abb.9 IEE Module [5]

Eines der fortschrittlichsten Konzepte ist die Integration *Machine Learning* Technologie innerhalb der Text Mining Forschung. Um dieses Konzept erfolgreich zu nutzen, muss eine Zusammenstellung von Eingaben definiert werden und alle Dokumente müssen kategorisiert sein. Anschließend kann eine bestimmte Dokumentkategorie dem MI Algorithmus übergeben werden. Nachdem das Model erfolgreich das Training durch die Vorgabe durchlaufen hat, können jetzt im Folgeschritt auch unkategorisierte Dokumente übergeben werden. Diese Dokumente werden jetzt anhand des durchlaufendem Trainings und der neu erlernten Kategorien, systematisch kategorisiert.

Der meist verwendete MI Algorithmus der für Text Mining betrieben wird, basiert auf Entscheidungsbäumen von Neuronalen Netzen, im Zusammenhang mit Support Vector Machines (SVM). Grundsätzlich wird zwischen zwei ML Lernmethoden unterschieden:

- Supervised Learning
- Unsupervised Learning

4.1. Supervised Learning

Überwachte Neuronale Netze versuchen bekannte Werte innerhalb des Datensets anzugleichen. Die Eingabe für eine Textkategorisierung, wird durch eine Wortfrequenz ausgedrückt. Die Ausgaben sind die Textkategorien oder vorhergesagte Werte. Die Ein- und die Ausgabe werden zu Beginn durch den ML Algorithmus mit einem zufälligen Wert in einem Zusammenhang gebracht und gewichtet. Innerhalb des Prozesses werden die Kategorien mit den bekannten Werten verglichen, der Prozess selbst verläuft dabei iterativ. Ein Fehler innerhalb der Kategorisierung oder der Vorhersage wird verwendet, um die Gewichtung zwischen der Ein- und Ausgabe für die folgende Iteration anzupassen. Dieser Prozess verläuft solange, bis die Fehlerzanzahl auf ein Minimum fällt. [5]

4.2. Unsupervised Learning

Unbeaufsichtigte Neuronale Netze werden für die Klassifikation von Dokumenten verwendet, in der die Trainingssets nicht bekannt sind. Dieser Prozess wird als *self-organization-map* bezeichnet. Der Prozess bietet z.B. die Möglichkeit der Darstellung mehrdimensionaler Daten, innerhalb eines niedrigeren ein- oder zweidimensionalen Raums. Dieser Prozess, der sich auf die Reduzierung von Vektoren konzentriert, ist im wesentlichen eine Datenkomprimierungstechnik die als Vektorquantisierung bekannt ist. Darüber hinaus schafft diese Technik ein Netzwerk, das Informationen in einer bestimmten Weise speichert und beliebige topologische Beziehungen innerhalb des Trainingssets aufrecht erhält. [5]

4.3. Text Mining Forschung

Eine vierte Phase innerhalb der Text Mining Forschung soll in naher Zukunft anlaufen (siehe Abb.9). Diese Phase wird als *Deep Discovery* bezeichnet. Ziel ist es innerhalb dieser Phase in der Lage zu sein Sarkasmus, Dialekte oder Anspielungen innerhalb von Textdokumenten zu erkennen. Des Weiteren beschäftigt sich die derzeitige Text Mining Forschung mit [5]:

- Der Analyse von Sozialen Netzwerken
- Der Mehrsprachigkeit beim Text Mining
- Der Klassifizierung von Spam
- Der Erkennung von Anomalien
- Der Trenderkennung
- und der Analyse von *gestreamten* Textdaten

5. HERAUSFORDERUNGEN

Zahlreiche Herausforderung ergeben sich aus dieser Umsetzung. Die erste große Herausforderung ist die Text Mining Technik im richtigen Kontext einzusetzen, um dieses Vorhaben effizient umzusetzen, muss Interdisziplinär gearbeitet und geforscht werden. Dabei werden viele andere Forschungsgebiete wie Machine Learing,A I, Computional Linguistic, Datenbanken, Data Mining und Libary and Information Science angeschnitten. Die Meinung des Domänenspezialisten ist eine weitere Vorgabe. Die Integration von zuverlässigen Datenquellen wird eine weitere Herausforderung. Jede einzelne Datenquelle hat ihre Vor- und Nachteile, die mit in das System einfließen und zusätzlich eine eigenständige Gesetzeslage, die nicht verletzt werden darf. Eine weitere Herausforderung liegt in der Konzeption der Lehrplattform und des Motivationsgehalts der hoch gehalten werden muss, um einen Studenten zu überzeugen diesen neuen Lehransatz zu verfolgen.

6. RISIKEN

Wenn Daten erhoben werden, insbesondere personenbezogenen Daten, dann steht der Datenschutz im Vordergrund. Ein wichtiger Bestandteil des Datenschutz ist die Informationsethik. An dieser Stelle müssen alle problematische Fragen geklärt werden, wie z.B., weit eine Datenerhebung gehen darf und wo die Grenzen sind. Das Bundesdatenschutzgesetz gibt dabei klare Vorgaben durch den Paragraphen § 13, der sich spezifisch auf die Datenerhebung konzentriert vor.

"Das Erheben personenbezogener Daten ist zulässig, wenn ihre Kenntnis zur Erfüllung der Aufgaben der verantwortlichen Stelle erforderlich ist. [BDSG 1, § 13]" [10]

"Das Speichern, Verändern oder Nutzen für andere Zwecke ist nur zulässig, wenn:

- der Betroffene eingewilligt hat,
- offensichtlich ist, daß es im Interesse des Betroffenen liegt, und kein Grund zu der Annahme besteht, daß er in Kenntnis des anderen Zwecks seine Einwilligung verweigern würde,
- die Daten allgemein zugänglich sind oder die verantwortliche Stelle sie veröffentlichen dürfte, es sei denn, daß das schutzwürdige Interesse des Betroffenen an dem Ausschluß der Zweckänderung offensichtlich überwiegt,
- es zur Durchführung wissenschaftlicher Forschung erforderlich ist, das wissenschaftliche Interesse an der Durchführung des Forschungsvorhabens das Interesse des Betroffenen an dem Ausschluß der Zweckänderung erheblich überwiegt und der Zweck der Forschung auf andere Weise nicht oder nur mit unverhältnismäßigem Aufwand erreicht werden kann.

[BDSG 2, § 13] "

Eine zentrale Rolle spielt dabei der Datenschutzbeauftragte, der rechtliche Details genauer erfasst und erläutert und speziell für eine wissenschaftliche Einrichtung, wie die HAW zuständig ist.

7. AUSBLICK

Es soll eine zukünftige Komponente in der Lehre entstehen, die sich innerhalb einer ersten Entwicklungsphase befindet. Diese Komponente ist die neue Lehrplattform. Innerhalb dieses Prozesses wird der erste Schritt der Prozesskette durchgeführt. Dieser Schritt bezieht sich auf die Analyse der vorhandenen Lehrinfrastruktur. Innerhalb dieser Analyse wird besonders auf die bereits bestehenden Lehrplattformen geachtet. Fällt diese Auswertung der Analyse negativ aus, so dass keine Bestandteile der bestehenden Lehrplattformen weiterverwendet werden können, so wird in einem Folgeschritt ein Prototyp einer neuen Lehrplattform konstruiert. Das Primärziel des Prototyps ist, dass dieser in der Lage ist eine Datenerhebung durchzuführen und anschließend auszuwerten. Diese Datenerhebung findet in der Form eines Text Mining Algorithmus statt. Der Text Mining Algorithmus bezieht innerhalb seiner Prototypphase erstmals auf einer Datenerhebung und Auswertung von Umfragebögen. Die erhobenen Daten werden anschließend in einer Datenbank hinterlegt, so dass diese der Lehrplattform für eine weitere Bearbeitung und Auswertung zur Verfügung stehen.

Die Lehrplattform soll durch das Ergebnis der ausgewerteten Fragebögen bereits erste Erkenntnisse liefern und versuchen diese logisch zu interpretieren.Aus diesem Zusammenhang heraus soll eine neue eigenständige Systemkomponente innerhalb der Architektur der Lehre etabliert werden. Diese Komponente soll den Studenten durch eine Vielzahl an Möglichkeiten, wie die individuelle Unterstützung im Studium oder beim Wissensaustausch mit anderen Studenten assistieren und aktiv begleiten. Das System soll die Produktivität und Effektivität des Studenten, durch den Gebrauch von gezielten Datenerhebungen deutlich steigern. Diese Ziele werden anhand von drei Kenngrößen ausgewertet:

- Messbarkeit
- Kontrollierbarkeit
- Organisierbarkeit

7.1. Die Messbarkeit

Das primäre Ziel ist eine Gesamtauswertung, welche den eigentlichen Nutzen und die Vorteile, die dieses Konzept mit sich bringt beurteilt, ermittelt und auswertet. Die Messbarkeit ist letztendlich das Kriterium, dass bewertet, ob das gesamte System eine positive Entwicklung nimmt. Ein Anwendungsbeispiel wäre, wenn die Daten bezogen auf eine bestimmte Semestergruppe anonymisiert verglichen werden. Bei diesem Vergleich werden die Vorteile, Nachteile, Schwächen und Stärken einzelner Studenten hervorgehoben. Es können dabei zwei unterschiedliche Gruppen verglichen werden eine, die die neue Lehrplattform verwendet und eine andere die, die derzeitige klassische Lehre beschreitet.

7.2. Die Kontrollierbarkeit

Das sekundäre Ziel ist die Kontrollierbarkeit. Dieser Bestandteil konzentriert sich auf den Studienverlauf. An dieser Stelle werden die getroffenen Maßnahmen visualisiert. Eine Beispielsituation wäre ein aktives stochastisches Anmeldesystem, dass den Studenten durch Datenerhebung gewonnene Analysen, bei einer Vielzahl von angemeldeten Modulen oder Klausuren die prozentuale Wahrscheinlichkeit errechnet wie hoch seine Erfolgsquoten liegen diese Klausur zu bestehen und wie viele Tage der Studierende, gemessen anhand von durchschnittlichen Werten, für die Lernvorbereitungsphase einplanen sollte. Durch diese zukünftige Anwendung erhält der Student in seiner Anmeldephase, im Kontext auf die einzelnen Module bezogen, ein zusätzliches konstruktives bzw. optionales Feedback, was ihn in seiner Auswahl der Module oder Klausuren unterstützt. Eine andere Feedbackform ist die des Studenten an das System. Indem der Student durch eine gezielte Kommunikation mitteilt, ob die getroffenen Maßnahmen oder besonders gestellte Lehraufgaben an ihn, seiner Vorstellung entsprechen und ob sich sein Lernverständnis dadurch vergrößert hat.

7.3. Die Organisierbarkeit

Das tertiäre Ziel ist die Organisierbarkeit, dieser Bestandteil konzentriert sich auf die Logistik aller möglichen freien und zugänglichen Ressourcen, die innerhalb eines Studiums zur Verfügung stehen. Eine Beispielsituation wäre ein besonders schwieriger Lernabschnitt innerhalb des Moduls, der von einer Vielzahl von Studenten nicht verstanden wird. Für diese betroffenen Studenten, die in dieser Situation eine signalisierte Lernschwäche aufweisen, sollte einen Lehrraum, die passende Zeit und eine Lehrkraft vom neuen zukünftigen System ermittelt werden, um gezielt den nicht verstandenen Lernstoff anders aufbereitet, bezogen auf das Themengebiet, zu lernen. Die Komponente müsste daher in der Lage sein, Teile aus dem Gesamtverlauf des Studiums eines Studenten zu erkennen, zu analysieren und jederzeit auszuwerten.

Diese drei unterschiedlichen Zielsetzungen demonstrieren das Potenzial, welches das neue System in der Lage ist zu nutzen. Es existiert noch eine Vielzahl an anderen Kombinationen von Anwendungsfällen, bzw. Möglichkeiten wie dieses System in der Zukunft innerhalb der Lehre etabliert werden kann. Der Kernbestandteil, den dieses Systems verwenden, stammt aus dem Forschungsfeld der Digitalisierung und bezieht sich auf die Erhebung von Daten.

8. REFERENCES

[1] " Die Zeit: MOOCs Massiv gescheitert," Z2015. [Online]. Available: http://www.zeit.de/2015/44/fernstudium-online-kurse-erfolg-moocs-spocs

[2] W. Fan and A. Bifet, "Mining Big Data : Current Status , and Forecast to the Future," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 2, pp. 1–5, 2013.

[3] Abb.2, *Smart-Phone-Sensoren*. [Online]. Available: http://www.notebookcheck.com/fileadmin/_migrated/pics/Samsung_Galaxy_s4_Sensors.jpgl

[4] R. Feldman and J. Sanger, *Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York, NY, USA: Cambridge University Press, 2006.

[5] D. D. Thomas Hil, John Elder, *Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications*. Academic Press, 2012.

[6] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 97–107, 2014. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6547630

[7] D. S. Tamhane and S. N. Sayyad, "Big Data Analysis Using Hace Theorem," vol. 4, no. 1, pp. 18–23, 2015.

[8] H. Gerhard, Q. Uwe, and W. Thomas, *Text Mining Wissensrohstoff Text*. W3L, 2006.

[9] S. Havre, B. Hetzler, and L. Nowell, "ThemeRiver: visualizing theme changes over time," *Information Visualization, 2000. InfoVis 2000. IEEE Symposium on*, vol. 2000, pp. 115–123, 2000.

[10] B. der Justitz und für Verbraucherschutz, *Bundesdatenschutzgesetz*. [Online]. Available: https://www.gesetze-im-internet.de/bdsg_1990/__13.html

9. ACRONYMVERZEICHNIS

ASCII American Standard Code for Information Interchange

BDSG Bundesdatenschutzgesetz

HACE Heterogeneous ,Autonomous ,Complex and Evolving

HTML Hypertext Markup Language

MOOCS Massive Open Online Course

SPOCS Small private online course

XML Extensible Markup Language

IEE Information Extraction Engines

SVM Support Vector Machines

IoT Internet of Things