# Masterthesis: Bestimmung der Dokumentenähnlichkeit basierend auf Bayessche Statistik für eine Big-Data Information Retrieval Lösung (Eng: Determination of Document Similarity based on Bayesian Statistics for a Big-Data Information Retrieval Solution)

1. Einleitung
   1. Motivation

Das Thema Big Data gewinnt immer mehr an Bedeutung, bzw. in Unternehmen. Fasel und Meier zitieren nach [Mer11], dass Big Data definiert wird als Daten, die in ihrer Größe klassische Datenhaltung, Verarbeitung und Analyse auf konventioneller Hardware übersteigen ([FAM16], S.5).

* 1. Problemstellung
  2. Zielsetzung

1. Grundlagen, Analyse von Use Cases zur Dokumentähnlichkeitsbestimmung sowie auf Bayessche Statistik basierende Ansätze zur Dokumentähnlichkeitsbestimmung
   1. Grundlagen der Informationsrückgewinnung
      1. Definition

Die Informationsrückgewinnung ist eine Aktivität bei der Material (Dokumente z. B) in einer unstrukturierten Natur (Text Beispielsweise) gefunden wird, um ein Informationsbedürfnis innerhalb von großen Ansammlungen, die auf einem Speicher gespeichert sind, zu erfüllen [MRS08]. Bei der Informationsrückgewinnung werden Informationsobjekte abgebildet, gespeichert, gestaltet und zugegriffen [BRI99].

* + 1. Nutzen

Da das Internet immer mehr genutzt wird und die meisten Anwender bei Suchmaschinen oder E-Mail Dokumente abrufen wollen wird die Informationsrückgewinnung immer mehr gebraucht und angewendet. Diese unterscheidet sich von der traditionellen Datenbanksuche und wird beliebter als Form der Informationszugriff. Die Informationsrückgewinnung wird angewendet um Daten- und Informationsprobleme zu lösen. Sie vereinfacht sie die Semi-strukturierte Datensuche wie Z.B das Finden eines Dokuments wo die Überschrift Java enthält und der Inhalt Threading[MRS08].

Die Informationsrückgewinnung wird auch angewendet um Benutzer zu unterstützen bei der Durchsuchung sowie Filterung von Dokumentensammlungen und um den Satz abgerufener Dokumente weiter zu verarbeiten.

* + 1. Mechanismus

Die Informationsrückgewinnung erfolgt durch eine Software, die für den entsprechenden Zweck hergestellt wurde. Mithilfe einer Software Architektur kann das Mechanismus der Informationsrückgewinnung beschrieben werden. Wichtige Elemente dieser Software Architektur sind eine Datenbank, die die Materialen (Dokumente mit Texten als Inhalt) enthält, ein Datenbank Manager Modul, eine Anwenderschnittstelle, Text Operationen, Abfrageoperationen, die Suche, der Rang und die Indexierung [BRI99].

Der Prozess der Informationsrückgewinnung beginnt sobald der Anwender sein Bedarf an Informationen in Form einer textuellen Abfrage(Schlüsselwörter) spezifiziert hat durch die Anwenderschnittstelle. Dieser textuellen Abfrage werden analysiert und durch die Text Operationen transformiert. Text Operationen generieren daneben eine logische Sicht der textuellen Abfrage. Die textuelle Abfrage wird danach von Datenbank Manager indexiert, bzw. es wird eine sogenannte invertierte Datei erzeugt. Die vorverarbeitete Abfrage wird von Abfrageoperationen in eine Systemebene Darstellung weiter transformiert. Die Informationsrückgewinnung beginnt dann sobald die textuelle Abfrage indexiert wurde. Es erfolgt dafür die Ausführung der textuellen Abfrage über eine Dokumentenquelle um den Abruf einer Menge relevanter Dokumente. Die Abfrageverarbeitung kann schnell erfolgen mithilfe der zuvor aus den Dokumenten in der Dokumentquelle erstellte Indexstruktur. Die abgerufenen Dokumente werden entsprechend Ihrer Relevanz geordnet bevor sie zum Anwender gesendet werden. Eine Untersuchung des Satzes von rangierten Dokumenten auf nützliche Informationen sowie die Erstellung eines Anwender Feedback können dann vom Anwender durchgeführt werden. Die Abbildung 1 zeigt der Mechanismus der Informationsrückgewinnung, wo das Zusammenspiel von Software Architektur Komponente dargestellt ist.

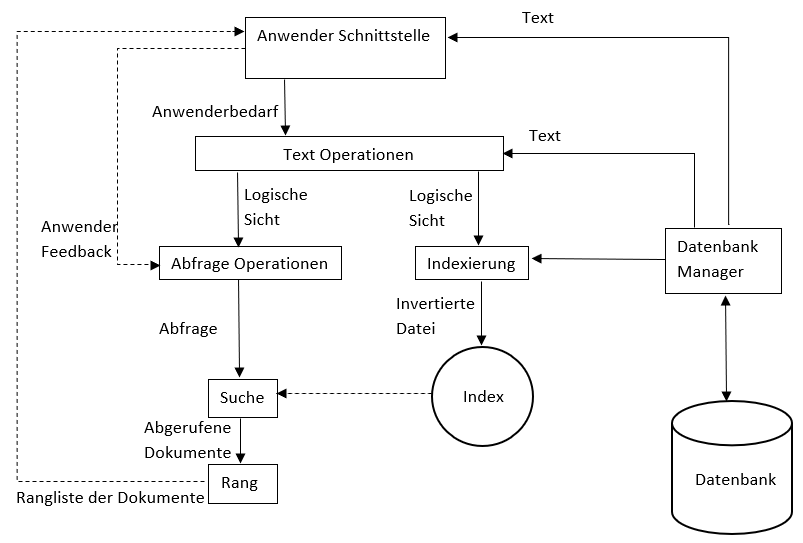


Abbildung : Der Mechanismus der Informationsrückgewinnung[BRI99].

Die Informationsrückgewinnung wird auf Basis von Modelle durchgeführt. Was das Konzept dieser Modelle ist und wie diese funktionieren wird es zunächst erwähnt.

* + 1. Informationsrückgewinnungsmodelle

Modelle sind charakterisiert durch einen Zweck sowie eine Art und werden definiert als Abbild eines realen Systems oder Problem. Im Fall der Informationsrückgewinnung werden Modelle benutzt um Vereinfachungen zu finden und abstrakt die Zusammenfassung oder die Vernachlässigung von Elemente zu bestimmen. Es bestehen drei Informationsrückgewinnungsmodelle, nämlich das boolesche Modell, das Vektorraummodell und die probabilistischen Modelle.

1. Das boolesche Modell

Das boolesche Modell, ist ein Modell, das beruht auf Mengenlehre und boolesche Algebra. Detaillierter ausgedrückt, bei dem booleschen Modell Abfragen können in Form eines booleschen Ausdruckes von Termen formuliert werden, bzw. Terme werden mit den Operatoren UND (Konjunktive Abfrage), ODER (Disjunktive Abfrage) oder NICHT (Negative Abfrage) verbunden. Wegen dem intuitiven Charakter des Konzeptes einer Menge, stellt das boolesche Modell einen leicht zu verstehenden Rahmen für einen gewöhnlichen Benutzer. Des Weiteren Abfragen werden als boolesche Ausdrücke mit präzise Semantik beschrieben. Da das boolesche Modell einfach und formell ist, wurde er in den vergangenen Jahren wahrgenommen und von vielen der frühen kommerziellen bibliografischen Systeme angeeignet. In das boolesche Modell werden Dokumente als eine Menge von Indexbegriffe bezeichnet. Außerdem Indexterme Gewichte sind binär. Das boolesche Modell entscheidet ob ein Dokument relevant oder irrelevant ist für eine gegebene Abfrage. Eine teilweise Übereinstimmung mit dem Dokument wird nicht toleriert.

1. Das boolesche Modell bietet den Vorteil einfach zu sein und ein sauberer Formalismus in seiner Struktur.
2. Ausdrücke haben eine präzise Semantik, die sie für strukturierte Abfragen geeignet macht, die von "Experten" -Benutzern formuliert werden. [CBB13].

Es bestehen jedoch auch Nachteile für dieses Modell. Der Hauptnachteil ist, dass die Eigenschaft der totalen Übereinstimmung bei der Informationsrückgewinnung dazu führt, entweder wenige oder zu viele Dokumenten als Ergebnis der Suche zu bekommen. Etwas, das den Anwender die Formulierung guter Abfragen erschwert. Eine Lösung zu dem o. g Nachteil ist die Koordinationsstufe. Mit der Koordinationsstufe können Aussagen aus atomaren Aussagen nichtbinäre Werte haben, bzw. Ranking Dokumente über die reale Linie[MEL15]. Das heißt die binäre Werte des booleschen Modells kein Dokumentenranking ermöglichen bei abwesender Koordinationsstufe.

Die Koordinationsstufe ist ein Maß für den Grad zu dem jedes zurückgegebene Dokument der Abfrage entspricht [MEL15]. Diese stellt eine Punktzahl für das Ranking der Dokumente bereit. Dieses Ranking ermöglicht es dem Benutzer eine Entscheidung über die Menge der zu prüfenden Dokumente zu treffen und gibt dem System die Möglichkeit, den Dokumentenwert, der nach unten gerankt ist, von der Liste abzuschneiden. Die Kalkulation der Koordinationsstufe erfolgt durch die Transkription einer booleschen Abfrage in Konjunktive normale Form (KNF), bzw. durch die Konjunktionsoperator gebundenen Propositionen Liste. Jeder dieser Propositionen präsentiert sich als Disjunktionen von Atom Vorschläge. Die Koordinationsstufe eines Dokuments entspricht die Anzahl der Vorschläge, aus denen eine KNF durch das Dokument zusammengesetzt ist [MEL15].

Bei der booleschen Informationsrückgewinnung wird die Gewichtung von Termen nicht bestimmt. Es ergibt sich infolgedessen zu kleine oder zu große Ausgabe [BRI99]. Wegen diesem Problem wird das boolesche Model in moderne Informationsrückgewinnung nicht mehr eingesetzt. Eine Alternative ist die Erweiterung des booleschen Modells um die Funktionalität von teilweise Übereinstimmung und Begriffsgewichtung. Um die zukünftige Berücksichtigung der variablen Größe zu ermöglichen, wurde eine Variation der Koordinationsstufe dessen Name gewichtete Koordinationsstufe ist, eingesetzt. Bei der gewichteten Koordinationsstufe erfolgt die Zuweisung eines anderen Gewichts abhängig vom Vorschlag, anstatt der Zuweisung konstantes Gewicht zu jedem vom Dokument wahr gemachter Vorschlag (?Satz Formulierung MEL15 S8??).

Andere Erweiterungen des booleschen Modells unterstützen die Informations-Nähe und Distanz. Mit diesen Erweiterungen kann spezifiziert werden, ob zwei Begriffe in einer Abfrage in einem Dokument nahe beieinander erscheinen dürfen. Es ist möglich die Nähe zu messen durch Begrenzung der Anzahl von dazwischenliegenden Wörtern oder durch Referenzieren auf eine strukturelle Einheit wie ein Satz oder ein Paragraph (Rock NEAR Roll) [CBB13].

1. Das Vektorraummodell

Das Vektorraummodell wird charakterisiert durch den Begriff Ähnlichkeit. Um den grenzenden Aspekt der binären Gewichteinheiten zu überwinden, bietet das Vektorraummodell einen Rahmen, bei dem die teilweise Abstimmung toleriert wird. Diese teilweise Übereinstimmung erfolgt durch die Zuweisung nicht binäre Gewichtseinheiten zu Indexbegriffe in Abfragen und Dokumenten. Die Begriffsgewichte berechnen den Ähnlichkeitsgrad von jedem Dokument, das gespeichert ist im System und die Anwenderabfrage. Die Berücksichtigung der nur teilweise mit Abfragebegriffen übereinstimmende Dokumente erfolgt beim Ähnlichkeitsgrad durch das Sortieren in absteigender Reihenfolge der Dokumente, die abgerufen wurden. Was sich daraus ergibt ist eine viel genauere Ranglisten-Antwortmenge als die Dokumentenantwortmenge, die mit dem booleschen Modell abgerufen ist. Das Vektorraummodell bewertet das Ähnlichkeitsgrad eines Dokuments in Bezug auf die Abfrage als die Korrelation zwischen zwei Vektoren. Die Quantifikation dieser Korrelation erfolgt beispielsweise durch den Kosinus des Winkels zwischen diesen zwei Vektoren[BRI99]. Anstatt eine Vorhersage über die Relevanz eines Dokuments durchzuführen, das Vektorraummodell geht mit der Klassifizierung der Dokumente entsprechend ihrem Ähnlichkeitsgrad mit der Abfrage vor. Der Abruf eines Dokuments erfolgt dann nur wenn die Bedingung der teilweisen Übereinstimmung erfüllt ist. Die Festlegung eines Schwellenwertes für Ähnlichkeit ist beispielsweise möglich, und den Abruf der Dokumente mit einem Ähnlichkeitsgrad über diesem Schwellenwert ist machbar. Die Berechnung einer Rangliste ist möglich nur wenn die Indexbegriffsgewichte erhalten sind. Diese Indexbegriffsgewichte werden erhalten durch welche begriffsgewichte Techniken, die einen Bezug auf die Clustering-Techniken unterstützende Grundprinzipien.

Die Clustering-Techniken werden wie folgend beschrieben. Ein einfacher Clusteralgorithmus hat als Ziel die Trennung einer Sammlung S von Objekten in zwei Mengen, mit gegebener Sammlung S von Objekten und einer vagen Beschreibung einer Menge M: eine Menge, die Objekte mit Bezug auf die Menge M enthält und eine andere bestehend aus nicht mit der Menge M verwandte Objekte. Die Bedeutung von vage Beschreibung ist, dass keine vollständige Informationen vorhanden sind, um die Entscheidung über anwesende Objekte in der Menge M zu treffen. Es ist möglich für anspruchsvollere Cluster-Algorithmen die Trennung von Objekte einer Sammlung in verschiedene Cluster nach ihren Eigenschaften zu bestreben. Bei einem Clusterproblem kommen zwei Hauptprobleme in Frage, nämlich die muss Ermittlung von Merkmale, die besser beschrieben werden von den Objekten, erfolgen und die Ermittlung von Merkmale, die die Objekte in der Menge M besser von den übrigen Objekten in der Sammlung S differenzieren. Das eine Merkmal setzt eine Quantifizierung der Intra-Cluster-Ähnlichkeit frei und das andere Merkmal erlaubt die Quantifizierung der Ungleichheit zwischen den Clustern.

Die Quantifizierung in die Intra-cluster-Ähnlichkeit erfolgt durch die Messung der Rohhäufigkeit eines Ausdrucks innerhalb eines Dokuments. Die Bezeichnung von so eine Begriffshäufigkeit heißt Faktor und die Begriffshäufigkeit misst wie gut die Beschreibung des Dokumenteninhalts ist. Die Quantifizierung der Ungleichheit zwischen den Clustern erfolgt durch Messung der Umkehrungshäufigkeit eines Ausdrucks unter den Dokumenten in der Sammlung. Laut [BRI99] das Vektorraummodell bietet die folgenden Vorteile an:

1. Sein Begriffsgewichtungsschema verbessert die Suchleistung.
2. Seine Strategie der teilweisen Übereinstimmung ermöglicht das Abrufen von Dokumenten, die den Abfragebedingungen angenähert sind.
3. Die Cosinus-Rangliste-Formel sortiert die Dokumente nach ihrem Ähnlichkeitsgrad mit der Suchanfrage

Beim Vektorraummodell besteht auch einen Nachteil, nämlich, dass die Index Begriffe als voneinander unabhängig gelten. Die mögliche Beeinträchtigung der Gesamtleitung erfolgt aufgrund der wahllosen Anwendung von Indexbegriffe auf alle Dokumente in der Sammlung [BRI99].

1. Die probabilistischen Modelle

Mit den probabilistischen Modellen werden Informationsrückgewinnungsprobleme mithilfe der Wahrscheinlichkeitstheorie bestimmt. Diese werden benötigt, um das Problem der schwierigen Anwendung des booleschen Modells in Informationsrückgewinnungsaufgaben zu überwältigen. Ein Beispiel von Probleme wäre fehlendes Ranking für Forscher oder fehlendes oder überlastetes Output für den Endbenutzer [MEL15]. Mit dem Vektorraummodell konnte durch Ranking die Verbesserung der Benutzererfahrung erfolgen. Jedoch fehlt immerhin die Ermittlung offenen linearen Koeffizienten. Die probabilistischen Modelle unterstützen ein System dabei die Dokumentdarstellung auf Relevanz zu prüfen und hilft mit Prinzipien bei der Bereitstellung und Verwendung von Gewichte der Koordinationsstufe. Probabilistische Modelle umfassen die Wahrscheinlichkeits-Ranking-Prinzip, das binäre Unabhängigkeitsmodell[MRS08], Sprachmodelle, das Relevanz Modell[MEL15]. Bevor das Thema des Wahrscheinlichkeits-Ranking-Prinzips behandelt wird, ist es wichtig das Thema Wahrscheinlichkeitstheorie zu besprechen.

**Wahrscheinlichkeitstheorie [MRS08]:** Diese ist ein Feld der Stochastik, wo zufällige Ereignisse beschrieben und modelliert werden. Sie beginnt mit als Mengen aufgefasste und Wahrscheinlichkeiten zugeordnete Ereignisse. Die Wahrscheinlichkeiten in diesem Fall entsprechen reelle Zahlen zwischen 0 und 1. gegeben werden zwei Variablen A und B, die Ereignisse Repräsentieren, wobei die Wahrscheinlichkeit für die jeweiligen Ereignisse 0≤ P (A) ≤ 1 und 0≤ P (B) ≤ 1 abgefragt wird, da es unsicher ist ob, diese Variablen wahr sind in der reellen Welt. Das gemeinsame Ereignis der beiden Ereignisse wird durch die gemeinsame Wahrscheinlichkeit P (A, B) ausgedrückt. Der Ausdruck P(A|B) bedeutet die Wahrscheinlichkeit des Ereignisses A beim Auftritt des Ereignisses B. Die Kettenregel liefert die grundlegen Beziehung zwischen Kettenregel und die bedingte Wahrscheinlichkeit:

(2.1)

Die Wahrscheinlichkeit eines gemeinsamen Ereignisses ohne Annahmen ist gleich die Wahrscheinlichkeit eines der Ereignisse multipliziert mit der Wahrscheinlichkeit des anderen Ereignisses unter der Bedingung, dass der erste Ereignis passiert ist [MRS08]. Das Schreiben von um ein Ereignis zu ergänzen führt uns zu der folgenden Formel:

(2.2)

Die Partition Regel ist ein Teil der der Wahrscheinlichkeitstheorie, angibt dass, wenn ein Ereignis in eine erschöpfende Menge von disjunkten Unterfälle unterteilt wird, die Wahrscheinlichkeit von B ist die Summe der Wahrscheinlichkeiten der Unterfälle [MRS08]. Die folgende Formel drückt ein Sonderfall dieser Regel aus:

(2.3)

Eine Ableitung des Satzes von Bayes aus dieser Sonderfallregel ist möglich:

(2.4)

Die Aktualisierung von Wahrscheinlichkeiten kann durch diese Gleichung durchgeführt werden. Im Satz von Bayes P(A), die vorherige Wahrscheinlichkeit (engl. Prior probability), ist die erste Schätzung des Ereignisses A wenn keine anderen Informationen vorhanden sind. P(A|B) wird „posterior probability“(engl.) genannt und bezeichnet die Wahrscheinlichkeit nachdem den Beweis von B herausgestellt wurde auf Basis der Wahrscheinlichkeit, des Auftritts von B in den zwei Fällen. Ein Ereignis hat unter anderem auch Gewinnchancen. Diese Gewinnchancen G (A) liefern Informationen bezüglich möglicher Änderungen von Wahrscheinlichkeiten durch folgende Formel:

(2.5)

**Wahrscheinlichkeits-Ranking-Prinzip (WRP):** Wenn die Antwort eines Referenzabrufsystems auf jede Anfrage eine Rangfolge der Dokumente in der Sammlung ist, in der Reihenfolge der abnehmenden Wahrscheinlichkeit der Relevanz für den Benutzer, der die Anfrage eingereicht hat, werden die Wahrscheinlichkeiten so genau wie möglich auf der Grundlage der Daten geschätzt dem System für diesen Zweck zur Verfügung gestellt wird, ist die Gesamteffektivität des Systems für seinen Benutzer die beste, die auf der Grundlage dieser Daten erhältlich ist [MEL15]. In einem einfachsten binären Fall des Wahrscheinlichkeits-Ranking-Prinzips, dessen Name 1/0 Verlust (engl. 1/0 loss) ist, bestehen keine Wiederauffindungskosten oder andere Versorgungssorgen zur unterschiedlichen Gewichtung von Fehler oder Aktionen. Das Prinzip dieses Falls ist einfach so, dass ein Punkt verloren wird, wenn ein nicht relevantes Dokument zurückgegeben wird oder ein relevantes Dokument nicht zurückgegeben wird. Gezielt wird die Rückgabe der bestmöglichen Ergebnisse als oberste Dokumente, die für den Nutzer wählbar sind. Nach dem Wahrscheinlichkeits-Ranking-Prinzip, die Einordnung der Dokumente muss in absteigender Reihenfolge erfolgen. Falls es um die Rückgabe einer Reihe von Abrufergebnisse anstatt einer Bestellung geht, die Bayes Optimale Entscheidungsregel minimiert das Verlustrisiko beim Zurückgeben von eher relevante als nicht relevante Dokumente. Das WRP ist wichtig, dass es verbindet der prinzipielle Ansatz zum Ranking und den Effektivitätsmaßen. Die Risiken, die im WRP sind probabilistische Definitionen von Rückruf oder Abfallquote. Ob die Maximierung des Rückrufs im Falle einer gegebenen maximal tolerierten Abfallquote relevant ist, wird es determiniert bei der Maximierung der Wahrscheinlichkeit Mithilfe dieser probabilistischen Sichtweise. Das WRP verweist auf die Optimierung der Wiedergewinnungseffektivität sobald der Rückruf für jedes feste Abfallquote-Kosten maximal wird. Das WRP kann auch mit Rückholkosten umgesetzt werden, wo die Modellierung der Differenzkosten von Falschpositiven und Falschnegativen durchgeführt wird.

**Das binäre Unabhängigkeitsmodell[MRS08]:** Traditionell erfolgt die Anwendung dieses Modells mit dem WRP. Durch Einführung einfacher Annahmen, ermöglicht das binäre Unabhängigkeitsmodell das Schätzen einer Wahrscheinlichkeitsfunktion zu konkretisieren. Binär bedeutet auch boolesch, wobei die Darstellung von Dokumente und Abfragen erfolgt als binäre Begriffsanfall Vektoren. Das heißt die Darstellung eines Dokuments erfolgt durch einen Vektor, der das Wert 1 hat Falls der Begriff im Dokument vorhanden ist oder hat das Wert 0 Falls der Begriff nicht vorhanden ist im Dokument. Das Wort „Unabhängigkeit“ drückt das unabhängige Vorkommen von Begriffe in Dokumenten aus. Die Erkennung einer Assoziation zwischen Begriffe bestehen bei dem Modell nicht. Trotz ihre Unkorrektheit, bietet diese Annahme befriedigende Ergebnisse und entspricht die Annahme von Naive-Bayes-Modellen. Diese Annahme ist auch gleichwertig mit einer Annahme des Vektorraummodells, wo jeder Ausdruck eine zu allen anderen Begriffe orthogonale Dimension entspricht. Die Verfeinerung des Informationsbedarfs von Benutzer erfolgt durch die Anzeige einer Reihe von Ergebnisse. Für eine präzise probabilistische Suchstrategie soll die Abschätzung des Beitrags zur Relevanz von Begriffe in Dokumenten erfolgen. Im binäre Unabhängigkeitsmodell erfolgen die Ableitung einer Ranking-Funktion für Abfragebegriffe, sowie die theoretischen und praktischen Wahrscheinlichkeitsschätzungen. Eine Erweiterung, dessen Name Baumabhängigkeitsmodell ist, versucht die Modellierung von Abhängigkeiten erster Ordnung zwischen Begriffe durch die Verwendung[MET11] .

**Sprachmodelle [MEL15]**

**Das Relevanz Modell[MEL15]**

Probability theory and ranking, binary independance model

* + 1. Web Informationsrückgewinnung
  1. Grundlagen der Bayessche Statistik
     1. Definition und Hintergrund

Die Bayes’sche Statistik ist eine quantitative Methode, die dazu dient empirische Befunde im Falle einer Gleichheit bei Einheiten zu charakterisieren und auszuwerten mithilfe vom Bayesschen Gese. Diese wurde von einem bekannten Mathematiker und englische Minister, Thomas Bayes (1702-1761) entwickelt und publiziert in dem Buch „*An Essay towards solving a Problem in the Doctrine of Chances“* im Jahr 1764. In diesem Buch ist die sogenannte bayessche Formel oder Satz von Bayes beschrieben. Die bayessche Formel wurde am Anfang aufgrund fehlendes Verständnis für eine lange Zeit unbeachtet, bis zu der Zeit zwischen 1812 und 1814, wo die Darstellung der Wahrscheinlichkeitsrechnung von Pierre Simon de Laplace veröffentlicht wurde[LER17]. Es ist nur erst Ende des 20. Jahrhunderts, dass diese Formel bedeutender geworden ist für logisch-statische Wahrscheinlichkeitsargumentationen. Im Jahr 1992 erfolgte die Gründung der International Society for Bayesian Analysis, eine Gelehrtengesellschaft, die als Ziel hat die Förderung der Benutzung entsprechender Theorien und Methoden in der Industrie und Wissenschaft[LER17].

Nachdem eine Definition der Bayessche Statistik sowie ihr Hintergrund gegeben wurden, gilt es zunächst zu erklären in welchen Gebieten diese genutzt wird und was der Nutzen der Bayessche Statistik ist.

* + 1. Anwendungsgebiete und Nutzen

So wird die Bayessche Statistik in medizinischen Bereich (Aus einem oder mehreren positiven medizinischen Test wird die Existenz einer Krankheit deduziert), im Informatik Bereich (Abhängig von bestimmten Wörtern in einer E-Mail werden die Emails als Spam klassifiziert), im Künstliche Intelligenz Bereich, im Qualitätsmanagement (die Aussagekraft von Testreihen wird beurteilt), im Informationsökonomik Bereich (erwarteten Werte von zusätzlichen Informationen werden bestimmt.)

* + 1. Modelle, Parameter und Überzeugungen

Während einer zufallsbedingte Erfahrung die Ergebnisse sind das Resultat eines Zufalles und diese können unterschiedlich sein. Die zufallsbedingte Erfahrung zeichnet sich dadurch, dass diese wiederholt werden kann, hat ein präzises Ziel, alle möglichen Ergebnisse können beschrieben werden und das Ergebnis kann nicht vorhergesagt werden. Ein Beispiel von zufallsbedingter Erfahrung ist der Einwurf einer Münze und sich die erscheinende Seite einprägen. Für die Entscheidung bezüglich wer das Fußball Spiel starten wird der Schiedsrichter entscheiden, dass ein Münzeneinwurf stattfindet. Die Münze in sich hat zwei Seiten: die Vorderseite der Münze hat das Wert der Münze (z. B: 2 Euro) und die Rückseite der Münze hat ein Reichsadler. Sind die zwei Team überzeugt, dass die Münze fair ist? Ob die Münze fair ist oder nicht weißt niemand. Um es zu erfahren wird die Münze zehn Mal eingeworfen. Angenommen kommen folgende Ergebnisse: 8 Mal die Rückseite gegen 2 Mal die Vorderseite[KRU10]. Diese Ergebnisse lassen vermuten, dass die Münze eine Neigung hat. Zu der Zeit wo es entschieden wurde die Münze auszunutzen, die Vermutung war, dass die Münze fair ist (Schiedsrichter). Diese Vermutungen bezüglich des Einwurfs werden als Modell Teams Überzeugungen bezeichnet und können auch mathematisch dargestellt werden. Um mehr Präzision zu erziele werden mathematische Formeln benutzt bei der formalen Mathematik.

Sei das Modell der Wahrscheinlichkeit, dass ein Student ein Fach in einem bestimmten Studiengang besteht. Dieses Modell in sich ist eine Formel, die eine numerische Wahrscheinlichkeit generiert als Ausgabe. Die Eingabe des Modells in diesem Fall ist wie gut der Student sich auf die Prüfung vorbereitet hat, da diese stark die Note beeinflusst. Wie genau sich die Eingabe mit der Ausgabe bezieht kann durch ein Wert moduliert werden. Dieser Wert wird Parameter genannt. Mit dem Parameter wird quantifiziert wie viel der Einfluss zwischen Ein – und Ausgabe ist.

Mit dem Beispiel von einer Urne, die insgesamt 10 Kugeln, darunter 3 grüne Kugeln, 2 rote Kugeln, 5 gelbe Kugeln enthält, bei der eine Person mit einer Hand mischen wird und zufälligerweise eine Kugel rausziehen wird, kann ein Parameter ausgenutzt werden.

Vor dem Münzeneinwurf, zwei Überzeugungen kommen in Frage. Es wird angenommen, dass die Münze ist fair und wird deshalb die Vorderseite mit einer Wahrscheinlichkeit von 50% nach dem Einwurf zeigen. Oder es wird angenommen, dass die Münze unfair ist und wird deshalb 8 Mal oder 2 Mal von 10 Einwürfe die Vorderseite zeigen. Also eine Wahrscheinlichkeit mit einem Verhältnis von 80% für die Vorderseite gegen 20% für die Rückseite oder umgekehrt. (Seite 14). Welche Überzeugung stimmt, wird es nicht gewusst bis der Münzeneinwurf stattfindet. Nach dem Münzeneinwurf wird das Ergebnis die Überzeugungen stärker beeinflussen. Beispielsweise nach dem Ergebnis von 8 Mal gezeigte Vorderseiten von 10 Einwürfe, die Überzeugung, dass die Münze unfair ist mit 80% gezeigte Vorderseite mehr Sinn macht. Die Überzeugung, die vor dem Einwurf entstehen werden vorherige Überzeugungen genannt, da diese vor einige Beobachtung entstehen. Die Überzeugung, die nach dem Einwurf entstehen werden spätere Überzeugungen genannt, da diese nach einigen Beobachtungen entstehen.

* + 1. Die Wahrscheinlichkeit
    2. Der Satz von Bayes

Der Satz von Bayes ist einen mathematischen Satz, der abgeleitet von der Wahrscheinlichkeitstheorie ist und eine Beschreibung der Berechnung vorausgesetzte Wahrscheinlichkeiten liefert. Der Satz von Bayes wird angewendet zur Bestimmung des Wahrscheinlichkeitsmodells für eine vorgegebene Reihe von Daten. Das Modell stellt die Wahrscheinlichkeit der Daten, gegebenen spezifische Parameterwerter und die Modellstruktur [KRU10].

Die bedingte Wahrscheinlichkeit wird definiert mit den Gleichungen 3.11 und 3.12 als folgende Formel: p(y|x) = p (y, x) /p(x). Die Formel besagt, dass die Wahrscheinlichkeit von y gegeben x ist die Wahrscheinlichkeit, dass sie zusammen auftreten relativ zu der Wahrscheinlichkeit, dass x überhaupt passiert [KRU10]. Wenn die Definition algebraisch manipuliert wird, dann werden ein Paar Ergebnisse erhalten. Erstens wird eine Multiplikation beider Seiten von p (y | x) = p (y, x) / p (x) mit p (x), um die Erhaltung von p (y | x) p (x) = p (y, x) zu erzielen. Danach kann die Manipulation der Definition analog erfolge mit p (x | y) = p (y, x) / p (y), um die Erhaltung von p (x | y) p (y) = p (y, x) zu erzielen. Es gibt nun zwei unterschiedliche Ausdrücke, die gleich p (y, x) sind. Dann wird es wie folgt ausgedrückt: p (y | x) p (x) = p (x | y) p (y). Eine die Division des letzten Ausdrucks durch p (x) ergibt

. (4.1)

Der Nenner kann noch in p (x | y) umgeschrieben werden. Dafür dass p(x) noch p(x) = ∑y p(x, y) geschrieben werden von der Gleichung 3.9 in Teil Wahrscheinlichkeit und p(x, y) = p(x|y)p(y), können beide Gleichungen kombiniert werden in p(x) = p(x, y) = p(x|y)p(y). Eine Substitution des Wert von p(x) im Nenner der Gleichung 4.1 ergibt folgendes

. (4.2)

In Gleichung 4.2 ist der Zähler ein bestimmter fester Wert im Nenner  
ist eine Variable, die auf alle möglichen Werte von y über die Summe hinweg auftritt [KRU10]. Die Gleichungen 4.1 oder 4.2 werden Bayes Regel genannt [KRU10].

Wie der Satz von Bayes funktioniert und was es bedeutet kann es durch das Beispiel mit Kartenspiel gezeigt werden. Gesucht wird der König von Herzen. Die Wahrscheinlichkeit, dass es ein König gibt bei den Herzen ist p(K|H) = 1/13 und die Wahrscheinlichkeit, dass es ein Herz gibt bei den Königen ist p(H|K) =1/4. Es besteht im Ganz Kartenspiel die Wahrscheinlichkeit von nur Könige p(K) = 4/52 und die Wahrscheinlichkeit, dass es nur Herzen Karten gibt ist p(H) = 13/52. Ob diese bedingten Wahrscheinlichkeiten der Satz erfüllen, kann es geprüft werden mit der folgenden Kalkulation: p(H|K) p(K)/p(H)= (1/4) (4/52) / (13/52) = 1/13= p(K|H).

Eine Bestätigung des Satzes von Bayes für zwei abhängige Attributen kann durch das Beispiel des Münzeneinwurfs erhalten werden. Die Tabelle 1 ist eine Übersicht der Wahrscheinlichkeiten für den Münzeneinwurf. Mit 3 Einwürfe, wird die Anzahl an Vorderseite und die Anzahl an Schalter zwischen Vorder- und Rückseite gezählt.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Anzahl Schalter | Anzahl Vorderseite | | | | Marginal (Anzahl Schalter) |
| 0 | 1 | 2 | 3 |
| 0 | 1/8 | 0 | 0 | 1/8 | 2/8 |
| 1 | 0 | 2/8 | 2/8 | 0 | 4/8 |
| 2 | 0 | 1/8 | 1/8 | 0 | 2/8 |
| Marginal (Anzahl Vorderseite) | 1/8 | 3/8 | 3/8 | 1/8 |  |

Tabelle 1: Wahrscheinlichkeiten für den Münzeneinwurf [KRU10].

Aus der Tabelle 1 können folgenden Wahrscheinlichkeiten berechnet werden: p(2S|2V) = p (2S, 2V) / p(2V) = (1/8) / (3/8) = 1/3 und p(2V|2S) = p (2V, 2S) /p(2S) = (1/8) / (2/8) = 1/2. Die Wahrscheinlichkeit p(2S|2V) und p(2V|2S) sind nicht gleich, bzw. drücken nicht das gleiche aus. Deshalb kann der Satz von Bayes geprüft werden durch p(2S|2V) p(2V) / p(2S) = (1/3) (3/8) / (2/8) = 1/2. Dieses Ergebnis von p(2S|2V) p(2V) / p(2S) = 1/2 ist gleich p(2V|2S); ein Beweis, dass der Satz von Bayes funktioniert. In diesem Beispiel wurden Werte normalisiert. Unter Normalisierung von Werte wird die Teilung dieser Werte durch deren Summe verstanden. Die bedingten Wahrscheinlichkeiten der Spalten werden erhalten durch die Normalisierung der Wahrscheinlichkeit in der Zeile mit gegebenem Zeilenwert. Die Gesamtwahrscheinlichkeit einer Spalte muss auf 1.0 summiert werden. Dieses wird erreicht, indem die Zellwahrscheinlichkeit in der 2-Schalter Zeile geteilt wird durch seine ursprüngliche Zeile. Die bedingte Wahrscheinlichkeit für die Anzahl an Vorderseite ist beispielsweise erhalten durch die Normalisierung der 2 Schalter-Spalte: p (0V|2S) = 0 / (2/8) = 0, p (1V|2S) = (1/8) / (2/8) = 0.5, p (2V|2S) = (1/8) / (2/8) = 0.5, and p (3V|2S) = 0 / (2/8) = 0.

Angenommen, dass ein Ereignis Ri der Tabelle 2 ist passiert und der Spaltenwert unbekannt ist. Es bleiben nur die Werte der Zelle Ri und wird nur auf die i-te Zeile fokussiert. Da Ri wahr ist, die Summe der Wahrscheinlichkeiten in dieser Zeile muss 1 sein. Eine Resolution für den Satz von Bayes liegt in der Definition der bedingten Wahrscheinlichkeit. Dabei soll befolgt werden, dass die gemeinsame Wahrscheinlichkeit in der i-ten Zeile (Ri) und in der j-ten Spalte(Cj) kann p (Ri | Cj) p (Cj) oder als p (Cj | Ri) p (Ri) zum Ausdruck gebracht werden [KRU10]. Die Tabelle 2 präsentiert die alternativen Ausdrücke der gemeinsamen Wahrscheinlichkeit p (Ri, Cj).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Zeile** | **Spalte** | | |  | **Marginal** |
| … | **j** | … |
| … |  |  |  |  |  |
| **i** | … | p(Ri,Cj)  = p(Ri|Cj) p(Cj)  = p(Cj|Ri) p(Ri) | … |  | P(Ri) |
| … |  | … |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Marginal:** |  | P(Cj) |  |  |  |

Tabelle : Tabelle zur Herstellung des Satzes von Bayes [KRU10].

Der Satz von Bayes wurde oben im Kontext der diskreten Werten vorgestellt. Er kann auch für kontinuierliche Variablen gelten. Der unterschied dabei ist, dass Wahrscheinlichkeitsmassen werden zur Wahrscheinlichkeit Dichten und Summen werden zu Integralen [KRU10]. So wird der Satz von Bayes in diesem Kontext folgend geschrieben:

(4.3)

In der Gleichung 4.3 ist es hinzuweisen, dass y im Zähler repräsentiert ein spezifisch fixer Wert und ist im Nenner eine auf allen möglichen Werten stattfindende Variable.

Der Satz von Bayes wird auch auf Modell und Daten angewendet. Dass die Zeilen- und spaltenvariablen Datenwerte sind, ist ein wichtiger Aspekt für die Anwendung des Satzes von Bayes. Laut [KRU10] ein Modell spezifiziert die Wahrscheinlichkeit bestimmte Datenwerte, die durch die Modellstruktur und bestimmte Parameterwerte gegeben sind. Das heißt allgemein, dass ein Modell die folgendes genau bezeichnet:

p (Datenwerte | Parameterwerte und Modellstruktur).

Der Satz von Bayes wird dann angewendet um das oben genannte umzuwandeln in das gesuchte Wissen. Die Überzeugung bezüglich des Modells könnte mit gegebene Daten folgend aussehen:

p (Parameterwerte und Modellstruktur | Datenwerte).

Ein Beispiel kann mit dem Münzeneinwurf gezeigt werden. Dabei besagt das Modell, dass p (Datum = V | θ) = θ und p (Datum = R | θ) = 1 − θ. Der Satz von Bayes und seine Anwendung bei Daten und Modelle ist bei der Tabelle 3 dargestellt. Dabei entsprechen die Spalten der Tabelle 3 bestimmte Werte des Modellparameters und die Zeilen entsprechen Datenwerte [KRU10].

Es wird dabei die Wahrscheinlichkeit p (D, θ) der Kombination von Datenwert und Parameterwert über alle möglichen Kombinationen zur Erhaltung von Datenwerte und Parameterwerte gegeben. Wobei die gemeinsame Wahrscheinlichkeit der spezifischen Kombination von Parameterwert θ und Datenwert D in Zellen gefasst wird.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Daten** | **Modell Parameter** | | |  | **Marginal** |
|  | θ Wert |  |  |
|  |  | … |  |  |  |
| D Wert | … | p (D, θ)  = p(D|θ) p(θ)  = p(θ|D) p(D) | … |  | p(D) |
|  |  | … |  |  |  |
| **Marginal:** |  | p(θ) |  |  |  |

Tabelle 3: Die Anwendung des Satzes von Bayes auf Modelle und Parameter[KRU10].

Die Prior Wahrscheinlichkeit p (θ), ist die Wahrscheinlichkeit vor der Datenannahme. Das heißt die Wahrscheinlichkeit für alle mögliche Werte von θ. Entnahme der posterior Verteilung wird erreicht, indem die gemeinsamen Wahrscheinlichkeiten der Zeile durch die marginale Zeile p(D) geteilt werden. Die Faktoren von Bayes Regel werden folgend genannt:

= / (4.4)

p (D)

Normalisierungsfaktor

p (θ | D)

posterior

p (θ)

prior

p (D | θ)

Likelihood

Der Faktor dabei ist (Aus dem Nenner von Gleichung 4.3):

(4.5)

Das „prior“ ist die Stärke der Überzeugung auf θ ohne die Daten D. Das „posterior“, p (θ | D) ist die Stärke der Überzeugung auf θ mit Berücksichtigung der Daten D. das „Likelihood“ p (D | θ) ist die Wahrscheinlichkeit, die die durch das Modell mögliche erzeugten Daten mit Parameterwerte θ ausdrückt. Der Normalisierungsfaktor p(D) ist die Wahrscheinlichkeit der Daten nach dem Modell. Der Normalisierungsfaktor (Noch wechsel zum Begriff Evidence) wird bestimmt durch eine über alle möglichen Parameterwerte gewichtete Summierung (???S48).

Das Modell ist wichtig, da dieses den Parameter bedeutend macht. Der Parameter dient dazu das Modell im Satz von Bayes noch klarer auszudrücken. Da Mit einem gegebenen Modell M beispielsweise besteht die Definition aller Wahrscheinlichkeiten, kann die Gleichung 4.4 wieder geschrieben werden wie folgt:

= / (4.6)

p (θ | D, M)

posterior

p (D | θ, M)

Likelihood

p (θ |M)

prior

p (D| M)

Normalisierungsfaktor

Und der Normalisierungsfaktor ist:

. (4.7)

Die Gleichung (4.6) eignet sich, wenn mehrere Modelle in Frage kommen und die Bestimmung der Überzeugungsstärke mithilfe von Daten durchgeführt wird. In einem Fall wo zwei Modelle M1 und M2 in Frage kommen, wird der Satz von Bayes   
p(M1|D) = p(D|M1) p(M1) / p(D) and p(M2|D) = p(D|M2) p(M2) / p(D) formuliert,

wobei p(D) = ∑i p (D | Mi) p(Mi). Das Verhältnis dieser zwei Gleichungen ergibt

(4.8)

**Bayes Faktor**

Das Verhältnis der Normalisierungsfaktoren wird Faktor von Bayes genannt.

* 1. Analyse von Use Cases für die Dokumentähnlichkeitsbestimmung

Die Dokumentähnlichkeitsbestimmung ist ein Verfahren, bei dem Dokumente auf Ähnlichkeit geprüft werden. Welches Ziel wird von diesem Verfahren verfolgt, hängt stark ab von der Problemstellung, bzw. was gesucht wird und wozu. Es bestehen dafür unter anderem ein Paar Use Case (Anwendungsfälle) und Bereiche bei denen das Verfahren angewendet und gebraucht wird. Unter diesen Use Cases können Finden von Dokumenten mit ähnlichen Inhalten (Duplikate), die Verwendung von a priori Wissen, die Systemübergreifende „Fremdschlüssel“, das Profil Matching und die E-Mail Klassifikation (Spamfilter). Es werden zunächst die oben genannt Use Cases vorgestellt, wobei die aktuelle Situation, Problemstellung sowie Lösung erwähnt werden.

* + 1. Finden von Dokumenten mit ähnlichen Inhalten (Duplikaten Findung)

In einem Rechner kann abhängig von Speicherkapazität eine Unmenge von Daten gespeichert werden. Diese Speicherung besteht über Jahren im Computer und der Nutzer kann aus Versehen immer Ähnliche Dokumente in dem gleichen Rechner speichern. Es stellt sich danach ein Speicherplatz Problem, da der Rechner voll belegt mit Dateien aller Art ist. Um mehr freier Speicherplatz zu erwerben, wird beim Nutzer den Bedarf bestehen, nach Duplikate zu suchen und zu löschen im Rechner. Der Einsatz einer Information Retrieval Softwarelösung kann dieses Problem lösen. Es werden mit der Software Lösung alle Duplikate gesucht und gelistet, sodass der Nutzer die Möglichkeit hat, Dateien individuell zu löschen.

Eine andere Variante der Findung von Dokumenten mit Ähnlichkeit gibt einem Anwender die Möglichkeit unterschiedliche Dokumente über ein Thema zu finden, die Beispielsweise der gleiche Anfang und das gleiche Ende haben aber ein unterschiedlicher Inhalt. Der Anwender kann dadurch neue Themen, die zu den Themen seiner aktuellen Dokumente verwandt sind, entdecken.

* + 1. Verwendung von a priori Wissen

In einer Firma besteht eine Datenbank die die Wetterberichte für jeden Tag enthält und eine andere Datenbank, die die Verkaufszahlen pro Artikeln enthält. Um seine Produkte zu verkaufen und schon Voraus konsequent zu produzieren, möchte diese Firma, die beispielsweise Regenschirme verkauft, im Voraus wissen wann das Wetter günstig ist (schlechtes Wetter), um große Verkaufszahlen zu erhalten. Eine Information Retrieval Softwarelösung könnte dabei helfen ähnlichen Schwankungen der beiden Datenbanken oder von Filesysteme/Intranet der Firma herauszustellen und den Mitarbeiter dieser Firma verfügbar machen.

* + 1. Systemübergreifende „Fremdschlüssel“

In einer Firma kommt es vor, dass Kundendaten gespeichert werden für unterschiedliche Abteilungen, wobei jede Abteilung eine eigene Datenbank hat, die unterschiedlich sein kann von anderen Abteilungsdatenbanken. Dieser Unterschied besteht auch bei Abfragesprachen dieser Datenbanken, sowie Primärschlüssel. Die Herstellung von Kundenverträge auf Basis einer Kundennummer kann problematisch sein, da Kunden nicht per Fremdschlüssel in beiden Datenbanken identifizier bar sind. Der Einsatz einer Information Retrieval Lösung wäre optimal, indem ähnliche Kundendaten in zwei verschiedene Datenbanken herausgefunden und deshalb zusammengebracht werden. Dadurch wird die Durchführung einer Datenbankmigration, die zeit- und ressourcenaufwändig ist, erspart.

* + 1. Profil Matching

Bei einer Firma, die nach Fachkräfte sucht, kommt es vor, dass nachdem eine Stellenausschreibung veröffentlicht wird, viele Arbeitssuchenden sich darauf bewerben. Die Bewerbungen werden auf die Firma Datenbank gespeichert. Es geht meistens und Word oder PDF Dateien. Nun ist es so, dass die Firma sich Zeit sparen will und gleichzeitig die passenden Bewerber haben möchte. Eine Information Retrieval Softwarelösung löst dieses Problem, indem bestimmten Schlüsselwörter in Bewerberprofile durchgesucht werden um einen Match mit den Anforderungen der Ausschreibung zu finden.

Dieses Match kann auch genutzt werden um intern bei einer Firma geeignete Mitarbeiter für ein gegebenes Projekt zu finden.

* + 1. E-Mail Klassifikation (Spamfilter)

Spam-Mail sind E-Mails, die gesendet werden, ohne vorher verlangt zu sein. Diese können unerwünschte Werbungsemail sein, worauf Anwender kein Interesse haben. Die E-Mail Spam werden von besonderen Programme gesendet, wobei die gezielte Email Adressen von einer Software zufälligerweise erzeugt werden, oder aus Sammelprogramme. Eine Lösung für Spam E-Mail wäre ein Spam Filter oder eine Information Retrieval Lösung, die Techniken zur Erkennung von Spam nutzen wie die IP-Adresse, der Inhalt oder die Filterlisten [COM18].

IP-Adresse: Da die Spam-Emails vom Computer mit bestimmten IP-Adresse gesendet werden, können durch Erkennung dieser IP-Adresse die E-Mails als Spam bezeichnet und sortiert werden.

Inhalt: Die Spam E-Mail können auf Inhalt geprüft werden. Dafür werden die Betreffzeile und die Nachricht in sich auf bestimmte Schlagwörter überprüft.

Filterlisten: Spam Filter sind mit Schwarze und Weiße Listen versehen, bei denen Spamfilter-merkmale Schlagwörter und IP-Adresse sind. Die E-Mails, die Merkmale aus der Schwarze Liste ausstellen werden gelöscht und in einem Spam-Ordner verlagert oder als Spam markiert. Anderenfalls (Weiße Liste) werden sie im Posteingang verschoben. E-Mails mit Absender dessen Adresse in Adressenbuch gespeichert sind werden automatisch im Posteingangsordner verschoben, da diese darauf hinweist, dass die Absender und Empfänger gerne kommunizieren wollen.

[FAM16] D. Fasel, A. Meier (Hrsg.), Big Data, Edition HMD, Springer Fachmedien, Wiesbaden 2016

[GAB17] <http://wirtschaftslexikon.gabler.de/Definition/big-data.html>

[MER11] Merv, A.: It’s going mainstream, and it’s your next opportunity. Teradata Magazine, 01, (2011)

[LER17] <https://www.lernhelfer.de/schuelerlexikon/mathematik-abitur/artikel/thomas-bayes> (Letzter Abruf: 30.12.2017)

[COM18] <http://www.computerbild.de/artikel/cb-Ratgeber-Internet-So-schuetzen-Sie-sich-vor-der-E-Mail-Spam-3574047.html> (Letzer Abruf:05.02.2018)

[KRU10] J. Kruschke, Doing Bayesian Data Analysis: A Tutorial with R and BUGS,Academic press,Amsterdam 2011

[BRI99] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, New York 1999.

[MRS08] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press, New York 2008.

[CBB13] Stefano Ceri, Alessandro Bozzon, Marco Brambilla, Emanuele Della Valle, Piero Fraternali, Silvia Quarteroni, Web Information Retrieval Springer-Verlag Berlin Heidelberg 2013.

[MEL15] Massimo Melucci, Introduction to Information Retrieval and Quantum Mechanics, Springer-Verlag GmbH Berlin Heidelberg 2015.

[MET11] Donald Metzler, A centric view of information retrieval, Springer-Verlag GmbH Berlin Heidelberg 2011