**Bestimmung der Dokumentenähnlichkeit basierend auf Bayessche Statistik für eine Big-Data Information Retrieval Lösung**

1. Einleitung
   1. Motivation

Angesicht des operativen Tagesgeschäfts und der Geschäftsentwicklung nach Jahren ist die Datenflut im unternehmerischen Umfeld nicht zu stoppen. Die Speicherung von Großer Datenmengen zur Auswertung und Verarbeitung erfolgt in mehreren konventionellen Datenbanken. Big Data umfasst alle Daten, die sowohl zur Verarbeitung als auch zur Analyse oder Auswertung versehen sind und deren   
Menge konsiderabel wächst. Darüber hinaus ist Big Data charakterisiert durch das große Volumen an Daten, die Geschwindigkeit der Datenentstehung, die Vielfalt an Datenquelle und Datenformen und die Werte der Daten. Die Speicherung von großen Datenvolumen in Datenquellen kann Schwierigkeiten bereiten, bzw. wenn diese auf mehreren physischen Maschinen erfolgt [FAM16]. Es wird dann für die Datenanalyse in Big Data, Datalake eingesetzt.

Unter Datalake wird einen sehr großen Speicher verstanden, der die Daten in ihrem ursprünglichen Rohformat enthält. Das große Datenvolum und die bestehenden strukturierte, semi-strukturierte oder noch unstrukturierte Daten in Datalake können das Finden jeglicher Daten Mühe und Zeitaufwand für die Mitarbeiter eines Unternehmens bereiten. Die traditionelle Suche und unscharfe Informationsrückgewinnung bieten sich an um dieses Problem zu lösen. Diese sind aber nicht optimal, da sie nur als Basis vorhandenes Wissen haben. Die Strukturierung der Information nach Inhalt und Bedeutung stellt sich als notwendig. Daher kann eine Softwarelösung, die Daten aus unterschiedlichen Datenbanken auf Inhalt vergleicht, eingestellt werden.

Eine Information Retrieval Middleware-Lösung deren Namen BigData4Biz ist, wird von der Firma Dibuco GmbH entwickelt. Diese Softwarelösung beruht auf einem Konzept der mehrdimensionalen Strukturierung eines Datalake unter Verwendung verschiedener Begriffe der Datenähnlichkeit. Wobei die Datenähnlichkeit entweder geschäftsbezogen oder Geschäft agnostisch sein kann und beruht auf linguistische Aspekte, die berechnet werden unter anderem mit Berücksichtigung der Begriffsfrequenz und -bedeutung. Die Berücksichtigung der Begriffsfrequenz und -bedeutung hilft dabei Informationen im Datalake zu gewinnen und zu entdecken.

Eine Programmierschnittstelle ermöglicht es Abfragen in einer GUI zu formulieren um die Informationen zu gewinnen und zu entdecken. Angesichts der großen Menge an Daten eine Datensuche mit Benutzung von Sätze und Schlüsselwörter für eine Abfrage ist unmöglich. Unter Annahme, dass die Benutzer ganz unbewusst sind, was die Anzahl und die Art der Informationen und Daten angeht, BD4B bittet die Option für Benutzer bekannte Informationen zu sehen und abzufragen, sodass alternative und ähnliche Informationen als Ergebnisse angezeigt werden, so nach Relevanzgrad und mit einer immer engen Einschränkung bei der Suche.

Im Kontext von großen Datenmengen, die immer mehr wachsen, ist die Skalierbarkeit sehr wichtig. Die vorliegende Arbeit setzt sich an diesem Punkt an und untersucht sowohl die unterschiedlichen Ansätze zur Ähnlichkeitsbestimmung, die auf Bayesschen Statistik basieren als auch Ihre Skalierbarkeit. Ziel ist es der bestmögliche Ansatz zur Dokumentähnlichkeit zu finden um eine optimale und adaptierte Datensuche in BigData4Biz zu ermöglichen.

* 1. Problemstellung

Der Aufwand bei der Bestimmung der Dokumentähnlichkeit für mehrere Quellen ist in der Regel mit hohem Zeitaufwand verbunden. Vor allem die Bestimmung der Ähnlichkeit in einem Haufen von Daten, die sich entweder im Internet befinden oder in gegebenen Datenspeicher. Hier setzt eine Informationsrückgewinnungslösung an. Diese versucht die Bestimmung der Ähnlichkeit zwischen Dokumente durchzuführen. Dafür gibt es mittlerweile eine Vielzahl an Ansätze. Je mehr die Übereinstimmung sein soll, desto höher wird der Aufwand sein. Darüber hinaus ergibt sich, dass für die Bestimmung der Dokumentähnlichkeit zwei sich gegenseitig gegenüberstehende Herausforderungen beeinflussen, nämlich der Bestimmungsaufwand und den Wunsch Dokumente mit der höchsten Übereinstimmung im gesamten Datensee zu finden. Da dadurch die Nutzer zufrieden sein können. Außerdem kann es passieren, dass Dokumente nicht unbedingt zutreffen, obwohl die als ähnlich bestimmt wurden. Faktoren wie den Kontext, die Semantik sowie die Bedeutung müssen berücksichtigt werden. Um dieses Problem zu lösen, soll einen bestimmten Ansatz basierend auf Bayessche Statistik ausgewählt werden, der diese Faktoren berücksichtigt. Die vorliegende Arbeit untersucht auf Bayessche Statistik basierende Ansätze zur Dokumentähnlichkeitsbestimmung.

* 1. Zielsetzung

Das Ziel dieser Masterarbeit ist es, die Dokumentähnlichkeit zu bestimmen für eine Big Data Informationsrückgewinnungslösung auf Basis von Ansätze, die auf Bayesschen Statistik basieren. Für die zuvor genannten Herausforderungen der Informationsrückgewinnungslösung, sollten in der vorliegenden Arbeit unterschiedliche Ansätze untersucht werden. Die in dieser Arbeit Dokumentähnlichkeitsbestimmung soll so erfolgen, dass die oben genannte Probleme gelöst werden.

Die Bayessche Statistik bietet Ansätze zur Dokumentähnlichkeitsbestimmung im Bereich der Informationsrückgewinnung, die angewendet und adaptiert zu gegebenen Fälle werden. Die Big Data Informationsrückgewinnungslösung kann zudem unterschiedlichen Datenquellen benutzen. Für die vorliegende Arbeit wird als Datenquelle hier ein Datenspeicher benutzt. Aus diesem Grund gilt es Ansätze zur Dokumentähnlichkeitsbestimmung basierend auf Bayessche Statistik zu analysieren Nach der Auswahl optimaler Ansätze wird eine Teilmenge der ausgewählten Ansätze implementiert. So können die realisierten Ansätze evaluiert werden auf Skalierbarkeit, das komplette auf Basis von ausgewählten Stichprobendaten.

* 1. Aufbau der Arbeit

Die vorliegende Arbeit besteht aus sieben Kapiteln. Das Kapitel 1 befasst sich mit der Einleitung zum Thema dieser Arbeit, wo die Problemstellung, Zielsetzung sowie der Aufbau der Arbeit erläutert werden.

Das Kapitel 2 behandelt die Grundlagen, die Analyse von Use Cases zur Dokumentähnlichkeitsbestimmung sowie auf Bayessche Statistik basierende Ansätze zur Dokumentähnlichkeitsbestimmung. In Kapitel 2.1 werden Grundlagen der Informationsrückgewinnung gegeben, nämlich die Definition des Begriffs Informationsrückgewinnung. Weitere wichtige Aspekte der Informationsrückgewinnung wie ihr Nutzen, ihr Mechanismus, ihre Modelle sowie ihre Web Variante werden erläutert. Kapitel 2.2 erläutert die Grundlagen der Bayessschen Statistik mit unteren Punkte über die Definition, der Hintergrund, die Anwendungsgebiete, der Nutzen, Modelle, Parameter, Überzeugungen der Bayessche Statistik, sowie die Wahrscheinlichkeit und der Satz von Bayes. Das Kapitel 2.3 beschäftigt sich mit der Analyse von Use Cases für die Dokumentähnlichkeitsbestimmung. Beim Kapitel 2.4 wird die Analyse existierender Ansätze zur Dokumentähnlichkeitsbestimmung basierend auf Bayesschen Statistik.

Das Kapitel 3 stellt die Informationsrückgewinnung-Middlewäre-Lösung vor. Dabei wird ihr Nutzen, Ihre Anforderungen, Fähigkeiten und Funktionsweise erläutert.

Im Kapitel 4 werden Ansätzen zur Dokumentähnlichkeitsbestimmung in der Informationsrückgewinnung-Middlewäre-Lösung ausgewählt. Im Kapitel 4.1 werden Auswahlkriterien für Ansätzen zur Dokumentähnlichkeitsbestimmung erläutert und in Kapitel 4.2 wird die Funktionsweise des ausgewählten Ansatzes erläutert.

Im Kapitel 5 wird die Implementierung einer Teilmenge ausgewählten Ansätze erläutert.

Im Kapitel 6 wird die Skalierbarkeit der realisierten Ansätze evaluiert durch die Benutzung von ausgewählten Probedaten. Das Kapitel 6.1 erläutert die Fähigkeiten und Grenzen dieser Ansätze.

Die vorliegende Arbeit wird mit dem Kapitel 7 durch eine Zusammenfassung und einen Ausblick abgeschlossen.

[ENZ17] <http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/daten-wissen/Datenmanagement/Datenmanagement--Konzepte-des/Big-Data> 09.06.2018