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

Zusammenfassung:

Es ist wichtig für Unternehmen ihre Chancen zu erkennen um eine bessere Leistung zu erwerben, konkurrenzfähig zu bleiben, ein besseres Profit zu erzielen und Kundentreue beizubehalten. Durch Analyse von Daten ist dies möglich. Das kontinuierliche Wachstum von Informationen zwingt Unternehmen sich an digitalen Wandeln anzupassen und die passende Big Data Technologien(Datalake) zu nutzen. Jedoch kann durch die großen Datenmengen die Suche nach bestimmten Daten für die Unternehmen Mitarbeiter erschwert werden.(Draft)

Das Ziel dieser vorliegenden Arbeit ist es Ansätze basierend auf Bayesscher Statistik zu untersuchen für eine Big Data Information Retrieval Lösung (BigData4Biz), die es ermöglichen die Ähnlichkeit zwischen Dokumenten zu bestimmen. Einen von diesen Ansätzen (BayesLSH) wurde ausgewählt und sein entsprechender Algorithmus(BayesLSH-Lite) wurde implementiert und anschließend evaluiert. Dieser Algorithmus hat als Basis das Locality Sensitive Hashing (LSH), bzw. die Form davon, nämlich Minhash. Bei Minhash wird eine Familie zufälligen Hashfunktion benutzt um einzelne Dokumente in Hash umzuwandeln. Dabei werden nur von jeder eindeutigen Hashfunktion erzeugte minimalen Werte beibehaltet. Die Ähnlichkeitsmessung von zwei Minhashes erfolgt dann Mithilfe der Jaccard Ähnlichkeit. Der BayesLSH-Lite Algorithmus vergleicht eine Menge von Kandidatenpaaren, die unter Minhashes Formen Tauchen, berechnet ihren Jaccard Ähnlichkeiten und unter der Bedingung, dass das Posterior (Satz von Bayes) kleiner ist als einen gegebenen Abrufsparameter, liefert die Liste der ähnlichen Dokumente zurück mit der entsprechenden Jaccard Ähnlichkeit.

Anhand von diesem o. g Algorithmus wurde die Dokumentähnlichkeit zwischen mehreren Dokumenten bestimmt. Darüber hinaus wurde einen Mehrwert erzielt, da die Nutzer von diesem Algorithmus schnell nach bestimmte Informationen suchen oder nach anhand eines Dokuments wissen welche anderen Dokumente fast das gleiche Thema bespricht.

1. Einleitung

Dieses Kapitel befasst sich mit der Motivation, Problemstellung, die Zielsetzung sowie Aufbau dieser vorliegenden Arbeit.

* 1. Motivierendes Beispiel

In dieses Kapitel wird ein motivierendes Szenario beschrieben, welches als Basis genommen wird um das Ergebnis dieser vorliegenden Arbeit zu prüfen. Dieses Szenario basiert auf die Funktionsweise der Plattform BigData4Biz. Das Szenario basiert auf einen Client/Server Modell zur Dokumentähnlichkeitsbestimmung.

Dokumente eines dibuco Kunde können auf unterschiedliche Speicher gespeichert und auf die Plattform BigData4Biz aufgerufen werden. In diesem Szenario dient die Kunden Seite als Client und die BigData4Biz Plattform als Server. Der Kunde gibt eine Abfrage in Form von Phrasen oder Textausschnitte auf die Abfrage-API von BigData4Biz. Bevor der Dokumentenvergleich stattfindet, müssen erstmal alle Dokumente vom Kunden gesammelt werden. Die auf den Kundespeicher existierende Übertragungsschnittstelle und Agenten ermöglichen es Daten in Entitäten umzuwandeln um diese später in BigData4Biz nutzen zu können zur Ähnlichkeitsbestimmung. Agenten können JSON, CSV, oder noch PDF-Formate sein. Die Entitäten können dann durch die Extract-Transform-Load(ETL)-Methode auf die Plattform BigData4Biz übertragen werden. Sobald Die Entitäten auf der Plattform BigData4Biz sind, werden die verarbeitet um zu den dabei bestehenden Algorithmen zu passen. Die verarbeiteten Entitäten können dann untersucht werden bei den Ähnlichkeitsalgorithmen.

* 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 Volum an Daten, die Geschwindigkeit der Datenentstehung, die Vielfalt an Datenquellen und Datenformen sowie die Werte der Daten. Die Speicherung von großen Datenvolumen in Datenquellen kann jedoch 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. Die Einstellung einer Softwarelösung zum Vergleich auf Inhalt von unterschiedlichen Daten aus einem Datalake ist daher denkbar.

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 Datalakes unter Verwendung verschiedener Begriffe der Datenähnlichkeit. Unter Begriffe der Datenähnlichkeit werden verstanden, die in den Daten enthaltene Begriffe, die als Basis für den Datenvergleich genommen werden. Die Datenähnlichkeit kann entweder geschäftsbezogen oder Geschäft agnostisch sein und beruht auf linguistische Aspekte, die berechnet werden unter anderem mit Berücksichtigung der Begriffsfrequenz und -bedeutung [DIB18]. Die Berücksichtigung der Begriffsfrequenz und -bedeutung hilft dabei Informationen im Datalake zu gewinnen und zu entdecken. Durch die Bestimmung der Datenähnlichkeit werden relevante und neue Informationen im Data Lake entdeckt.

Eine Programmierschnittstelle ermöglicht es Abfragen in einer GUI zu formulieren um die Informationen zu gewinnen und zu entdecken. Angesichts der großen Menge eine Datensuche mit Benutzung von Sätzen und Schlüsselwörter für eine Abfrage ist unmöglich.

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 werden die ausgewählten Ansätze implementiert. So können die realisierten Ansätze evaluiert werden auf Skalierbarkeit, das komplette auf Basis von ausgewählten Stichprobendaten. VIP:(Momentan abgeschnitten von der Final Version. Noch klären ob es dabei ist oder nicht)

* 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 Bayesscher 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 Bayesschen 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-Middleware-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 von 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