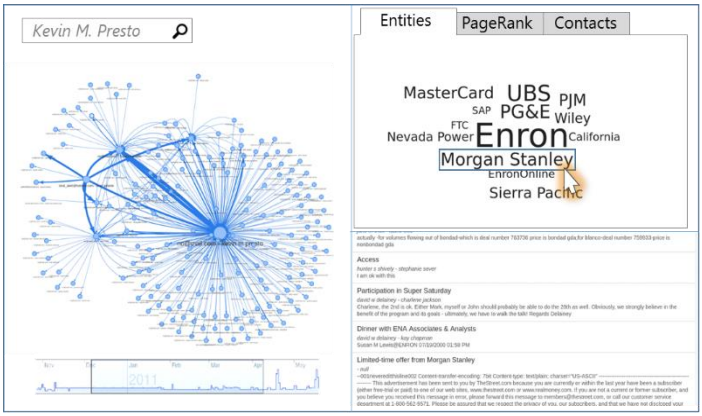
Bachelorprojekt für 2017/2018
Fachgebiet Informationssysteme
Prof. Dr. Felix Naumann

COMMERZBANK

Der Leuchtturm im Datennebel

Exploration großer Dokumentensammlungen



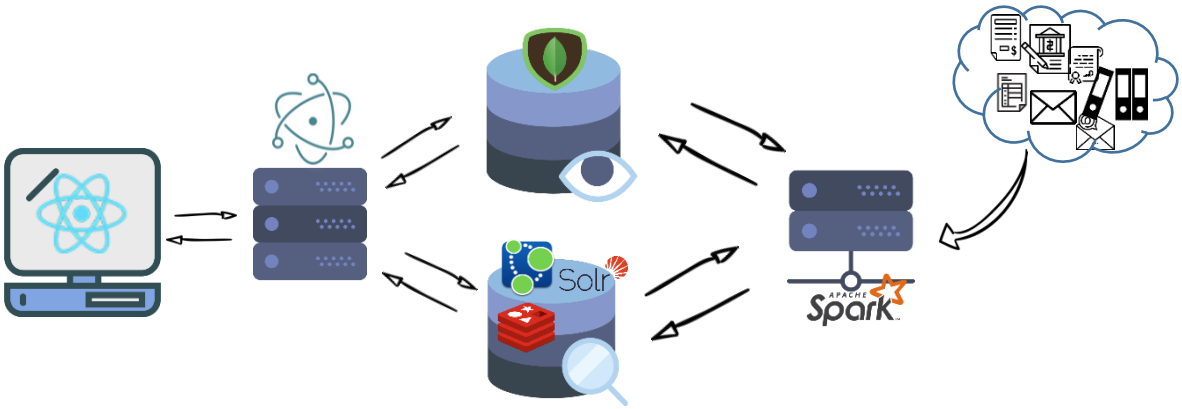
Die Dokumentation von Geschäftsprozessen in Unternehmen ist heutzutage zunehmend digitalisiert und umfasst unüberschaubare Datenmengen in vielfältigen Textformaten wie etwa Verträge, Formulare oder Protokolle. Solch ein Korpus dokumentiert nicht nur das Tagesgeschäft, sondern den Werdegang von Entscheidungsprozessen im Kleinen und Großen. Bei internen Audits werden diese Daten von Spezialisten zur Ursachenfindung herangezogen. Bedingt durch den Umfang von oft hunderten Gigabytes ist eine langwierige Einarbeitung erforderlich um einen Überblick zu bekommen. Vor ähnlichen Problemen stehen Journalisten nach Veröffentlichungen wie den **Panama Papers** durch Wikileaks. In diesem Bachelorprojekt wird ein System entwickelt, welche die Kerninformationen solcher Sammlungen automatisiert **strukturiert und visualisiert**.

Projektbeschreibung

Mithilfe automatisierter Extraktion von Informationen aus unstrukturierten Daten sollen Filterwerkzeuge implementiert werden. Diese erfassen verschiedene Aspekte, wie etwa die (zeitliche) Verteilung von relevanten Schlagwörtern, Entitäten, Themen und Dokumenttypen. Die nun strukturierten Daten werden über eine **interaktive Oberfläche** visualisiert. Benutzer erlangen damit einen Eindruck über die vorherrschenden Inhalte und können den Datensatz schrittweise explorieren.
Insbesondere umfasst das Projekt folgende Bereiche:

- **Phrasen- und Entitätenextraktion, sowie -disambiguierung:** Häufig verwendete Schlagwörter und Phrasen können genutzt werden, um grob die Inhalte der Dokumente zusammenzufassen. Besonderes Augenmerk liegt dabei auf Named Entities und der Zuordnung ihrer Erwähnungen im Text zu konkreten bekannten Datenobjekten.
- **Dokumentklassifikation:** Die Einordnung von Dokumenten in einfache Kategorien, wie etwa Verträge, Gerichtsbeschlüsse oder Rechnungen, kann als Vorsortierung oder Filter im Interface genutzt werden. Nach der Klassifikation können für bestimmte Dokumententypen optimierte Algorithmen angewendet werden.
- **Analyse von Beziehungsnetzwerken:** Metadaten der Dokumente können genutzt werden, um Strukturen in Prozessketten zu identifizieren, bzw. bereits extrahierte Informationen in einen breiteren Kontext zu setzen.
- **Datenvisualisierung:** Erst durch deren effektive Darstellung können die vielen Einzelinformationen sinnvoll genutzt werden. In der Oberfläche werden beispielsweise Schlagwörter in ihrem Kontext dargestellt, der etwa durch die zeitliche Einordnung gegeben ist.

Technologien, Architektur und Vorgehensmodell



Das zu entwickelnde System besteht aus zwei Komponenten. Zunächst gibt es eine Verarbeitungspipeline, die Rohdaten normalisiert in einer Cassandra ablegt. In weiteren Pipelines werden Algorithmen aus dem Bereich des **Textmining und Graphanalyse** implementiert um die Datensammlung anzureichern. Die extrahierten Informationen werden zur späteren effizienten Verknüpfung in Key-Value Stores (Redis) abgelegt. Zur Volltextsuche wird Solr genutzt und Zusammenhänge werden als Graphen in einer Neo4J Datenbank gespeichert.

Dem gegenüber steht eine auf Electron/React basierende **Web-App**, die den Zugriff auf die Daten ermöglicht. Vor allem jedoch steht hier die interaktive Visualisierung der zuvor extrahierten Informationen, basierend auf WebGL, im Vordergrund.

Um eine gemeinsame und flexible Konkretisierung der zu erreichenden Ziele zu ermöglichen, ist ein agiles Vorgehen wünschenswert; als Vorgehensmodell wird Scrum empfohlen.

Projektpartner

Die Commerzbank ist eine führende, international agierende Geschäftsbank mit Standorten in mehr als 50 Ländern. Mit den Geschäftsbereichen Privatkunden, Mittelstandsbank, Corporates & Markets und Central & Eastern Europe bietet sie ihren Privat- und Firmenkunden sowie institutionellen Investoren ein umfassendes Portfolio an Bank- und Kapitalmarktdienstleistungen an. Die Commerzbank finanziert über 30 Prozent des deutschen Außenhandels und ist unangefochtener Marktführer in der Mittelstandsfinanzierung. Mit den Töchtern comdirect und der polnischen mBank verfügt sie über zwei der weltweit innovativsten Online-Banken.

Die Commerzbank betreibt mit 1.100 Filialen sowie rund 90 Geschäftskundenberatungszentren eines der dichtesten Filialnetze der deutschen Privatbanken. Insgesamt betreut sie rund 15 Millionen Privat- sowie 1 Million Geschäfts- und Firmenkunden. Im Jahr 2014 erwirtschaftete sie mit durchschnittlich rund 52.000 Mitarbeitern Bruttoerträge von knapp 9 Milliarden Euro.

Das Projekt für bis zu 6 Studenten beginnt am 9. Oktober 2017 und wird durch Dr. Ralf Krestel, Tim Repke und Prof. Felix Naumann betreut. Es setzt die erfolgreiche Kooperation von Bachelorprojekten der vergangenen Jahrgänge fort. Fragen können gerne an tim.repke@hpi.de gerichtet werden.