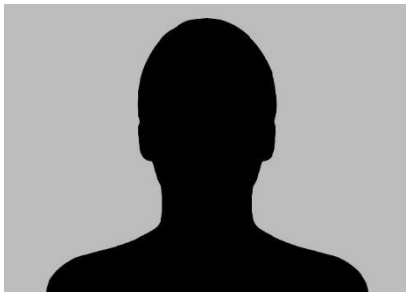
Zürcher Hochschule
für Angewandte Wissenschaften

School of
Engineering

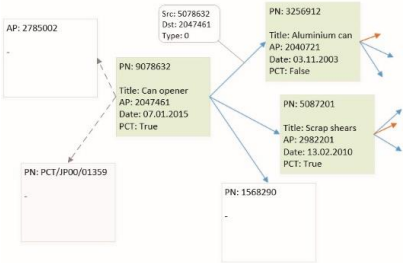
Graph-Analyse mit Big Data Technologien (Scala + GraphX)

Eine Herausforderung von Big Data ist die unstrukturierte Form, in welcher die Daten auftreten. Genau diese macht es aber interessant, sie nicht in eine SQL-Tabelle zu zwängen, sondern in abstrakteren Formen, wie die eines direktionalen Graphen abzubilden. Dies ermöglicht es, Einsichten zu erlangen und Lösungen zu finden, die in der klassischen Form wohl übersehen worden wären.
In dieser Arbeit wird gezeigt, wie eine reelle Big-Data-Quelle, nämlich die Patente des United State Patent and Trademark Office, anhand ihrer Zitate in einem Graphen abgebildet, durchsuchbar gemacht und die Suchresultate mit Relevanz versehen werden. Dabei wird auch auf Herausforderungen eingegangen wie die Architektur, die Ganzheit des Datensatzes oder das Grössenproblem, welches sich stellt, wenn ein Eintrag ein gesamtes Patent ist und nicht nur eine Zeile einer Produkteliste. Für die Relevanzbestimmung werden Indegree bzw. Outdegree der Patente in Bezug auf deren Alter betrachtet sowie der HITS-Algorithmus nach Kleinberg eingesetzt. Weiter wird der Graph in geringem Grad auf Anomalien mittels Strongly Connected Components untersucht. Die Suchresultate werden mittels einer einfachen Form des Known-Item-Retrieval ausgewertet.

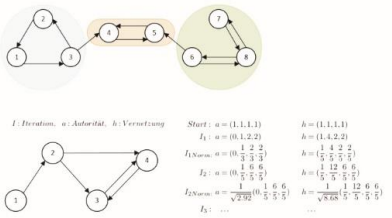


Diplomand
Markus Kessler

Dozent
Kurt Stockinger



Aufbau, um Patente in einem Graphen abzubilden



Strongly Connected Components und die Berechnung des HITS-Algorithmus