Die Algorithmik hinter der (Web)suche

Studiengang: BSc in Informatik
Betreuer: Prof. Dr. Erich Baur
Experte: Dr. Andreas Spichiger

Das Suchen oder auch „Googeln“ im World Wide Web (Web) ist ein fester Bestandteil unseres heutigen Alltags. Suchmaschinen für die Websuche sind daher allgegenwärtig. Eine einzige Websuche liefert in der Regel bereits unzählige Resultate. Trotzdem wird man im Normalfall unter den ersten Suchresultaten fündig. Doch wie funktioniert die Websuche und wie wird entschieden, was zuoberst angezeigt wird?

Ziel der Arbeit

Die Arbeit hat zum Ziel, die algorithmischen Aspekte der Websuche zu untersuchen. Der Fokus liegt dabei auf dem Ranking. Die folgenden Unterfragen werden beantwortet:

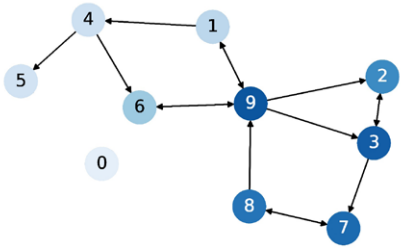
- Welche Grundlagen aus der Mathematik und dem Information Retrieval werden für das Verständnis benötigt?
- Wie funktioniert der PageRank-Algorithmus und welche Konzepte stecken dahinter?
- Welche Anwendungen gibt es für den PageRank-Algorithmus ausserhalb der Websuche und wie kann eine solche implementiert werden?

Einleitung

Das Web umfasst nach Schätzungen mittlerweile über eine Milliarde Webseiten. Um in dieser Masse die gewünschten Informationen zu finden, werden Suchmaschinen benötigt. Im Optimalfall deckt eine Suchmaschine den Informationsbedarf des Nutzers vollständig ab. Die Architektur einer Suchmaschine umfasst verschiedene Elemente wie das Crawling, die Indexierung, die Suche und das Ranking.

PageRank

Der PageRank-Algorithmus verhilft Google bis heute zu grossem Erfolg. Mittels Linkanalyse werden Webseiten anhand der Verlinkungsstruktur des gesamten Webs bewertet. Eine Webseite ist wichtiger, wenn sie von vielen anderen Webseiten verlinkt wird. Dabei fällt nicht nur die Anzahl der eingehenden Links ins Gewicht, sondern auch die Qualität der Links. Zur Berechnung wird das Verhalten eines Surfers im Web als Markov-Kette modelliert. Dabei wird das Web als Graph dargestellt, mit Webseiten als Knoten und Links als Kanten. Mit einigen Anpassungen an der Adjazenzmatrix des Webgraphs ist garantiert, dass die daraus resultierende Markov-Kette zu einer eindeutigen stationären Verteilung konvergiert. Diese Verteilung entspricht dem Ranking und weist jeder Webseite eine Wahrscheinlichkeit zu. Um so höher die Wahrscheinlichkeit, um so besser das Ranking.



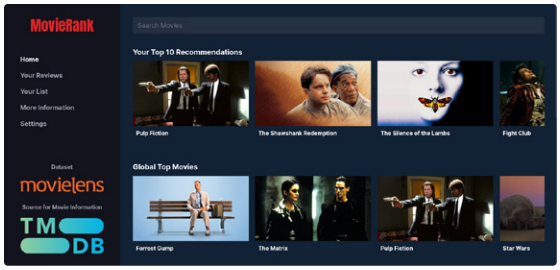
Graph, auf welchem der PageRank-Algorithmus ausgeführt wurde: Dunklere Knoten haben höheres Gewicht.

Anwendungen

Ursprünglich ausgelegt für das Web, wird PageRank in diversen anderen Gebieten modifiziert angewendet. Die Abwandlung ItemRank ermöglicht die Umsetzung eines kollaborativen Empfehlungsdienstes. Dieser erstellt personalisierte Empfehlungen für Objekte wie Filme oder Kochrezepte. Die Empfehlungen basieren auf den Präferenzen von anderen Nutzern, welche ein ähnliches Bewertungsverhalten aufzeigen.

Ergebnisse

Die Arbeit führt den Leser in den PageRank-Algorithmus und in die dazu benötigten Grundlagen der Mathematik und des Information Retrievals ein. PageRank, ItemRank und unsere Abwandlung davon wurden analysiert und mit Python implementiert und bilden die Grundlage für weitere Arbeiten. Ein Empfehlungsdienst in Form einer Django-Web-App erlaubt das interaktive Bewerten von Filmen und gibt dem Nutzer personalisierte Empfehlungen, die aus den Bewertungen berechnet werden.



Implementierung eines kollaborativen Empfehlungsdienstes für Filme mit PageRank.

Pius Loosli
IT Security

Kay Mattern
Data Engineering