Suche: Von Grundlagen zur Föderation

Jens Lechtenbörger

November 2018

1 Einleitung

Der Erfolg des Web hängt eng mit der Entwicklung leistungsfähiger Suchmaschinen zusammen. Nachdem zunächst manuell gepflegte Verzeichnisse (in Analogie zu Gelben Seiten für das Telefon) als Einstiegspunkte beim Surfen gedient hatten, entwickelten sich zahlreiche Suchmaschinen mit unterschiedlichen Konzepten für die Definition der Relevanz von Web-Seiten zu gegebenen Suchbegriffen.

Konzeptionell ist die Suche recht einfach: Ausgehend von bekannten URLs (engl. *seed pages*) rufen sogenannte *Crawler* die zugehörigen Inhalte ab, speichern und indexieren sie und extrahieren die URLs aus in den Web-Seiten enthaltenen Verweisen. Sämtliche so gewonnenen URLs bilden die stetig wachsende Grundlage für weitere Crawler-Läufe, wodurch nach und nach immer größere Bereiche des Web erschlossen werden. Eine große Herausforderung besteht bei der Suche von Web-Seiten zu einem Suchbegriff darin, aus der Fülle von Seiten, die den Suchbegriff umfassen, die *relevantesten* möglichst weit oben unter den Ergebnissen anzuzeigen. Wegweisend zeigte sich das Konzept des PageRank, das den Grundstein für den Erfolg von Google legte. (Man beachte, dass „Page“ sowohl „Seite“ bedeutet als auch der Name eines der Erfinder ist. Details zu Google von den Gründern Brin und Page finden sich in diesem Artikel aus dem Jahre 1998.)

Der PageRank einer Seite gibt an, wie wichtig sie unabhängig von ihrem Inhalt ist. Entsprechend können Suchergebnisse nach ihrem PageRank sortiert werden. Die Idee des PageRank beruht darauf, ausgehende Verweise als Empfehlungen anzusehen. (Wenn ich auf eine andere Seite verweise, mache ich das in der Regel, weil ich sie für lesenswert halte.) Entsprechend ist der PageRank einer Seite hoch, wenn viele andere Seiten mit hohem PageRank auf sie verweisen. Im oben genannten Artikel wird gezeigt, dass sich diese Idee durch folgende (rekursive) Gleichung ausdrücken lässt:

$$PR(A) = (1-d) + d\left(\frac{PR(T_1)}{C(T_1)} + \cdots + \frac{PR(T_n)}{C(T_n)}\right)$$

Aufgabe: Erläutern Sie die in dieser Gleichung vorkommenden Komponenten hinsichtlich der im Artikel genannten Intuition des *Random Surfers*.

2 Meta-Suchmaschinen

Um sich beim Suchen nicht einer zentralen Organisation anvertrauen zu müssen, können Meta-Suchmaschinen genutzt werden. Eine Meta-Suchmaschine benötigt keine eigenen Crawler, sondern sendet die empfangenen Suchbegriffe an eine Reihe anderer Suchmaschinen, kombiniert die Ergebnisse dieser Suchmaschinen und gibt das Resultat als eigenes Suchergebnis zurück. Beispiele für Meta-Suchmaschinen, die auf freier Software basieren und daher als dezentrale Instanzen betrieben werden können, sind das von dem deutschen Verein SUMA-EV entwickelte MetaGer und SearX; letzteres wird in diesem Interview mit dem Hauptentwickler vorgestellt. Wenn Sie Suche so spannend finden, dass Sie eigene Untersuchungen anstellen möchten, interessiert Sie vielleicht, dass SUMA-EV Stipendien zur Unterstützung von Abschlussarbeiten vergibt.

Aufgabe: Lesen Sie das oben genannte Interview. Welche Unterschiede sieht der SearX-Entwickler zwischen MetaGer und SearX?

3 Peer-To-Peer-Suche

Um bei der Suche von anderswo kontrollierten Crawler-Ergebnissen unabhängig zu werden, ist es notwendig, eigene Crawler zu betreiben und deren Ergebnisse selbst zu verwalten. Offenbar erfordert dies weitaus größere Ressourcen als der Betrieb einer Meta-Suchmaschine.

Die Suchmaschine YaCy implementiert Crawler und Suche auf Basis freier Software in einem Peer-to-Peer-Netz. Einzelne Peers können sowohl als autonome Suchmaschinen betrieben werden als auch der Föderation des sog. Freeworld-Netzes beitreten, in dem Crawler-Aufgaben gemeinschaftlich übernommen und Suchergebnisse gemeinschaftlich zusammengetragen werden. Jede/r einzelne kann einen YaCy-Peer betreiben (die Installation ist erstaunlich einfach – im Wesentlichen ein Doppelklick – und wird in Lehrfilmen für gängige Betriebssysteme beschrieben) und diesem Peer eine eigene Liste von Seed-URLs mitgeben, die in anderen Suchmaschinen vielleicht unterrepräsentiert sind.

Aufgabe: Welche Vorteile verspricht die Eigendarstellung der Philosophie von YaCy? Sehen Sie Nachteile?

4 Lizenzangaben