Handbuch Internet-Suchmaschinen
D. Lewandowski (Hrsg.)

Ranking-Verfahren für Web-Suchmaschinen

Philipp Dopichaj
Lycos Europe
Carl-Bertelsmann-Straße 29
33311 Gütersloh
dopichaj@acm.org

Abstract. Typische Web-Suchen liefern tausende Ergebnisse. Deshalb ist es von zentraler Bedeutung, dass die Suchmaschine die Ergebnisse sinnvoll sortiert, damit die wichtigsten Ergebnisse auf einen Blick zu erfassen sind. Hierzu wurden im Laufe der Jahre viele Verfahren entwickelt und verbessert. Wir behandeln in diesem Kapitel sowohl Verfahren, die auf dem Text der Webseiten basieren als auch solche, die die Verlinkungsstruktur berücksichtigen. Abschließend betrachten wir mögliche Erweiterungen der Suchverfahren für die Zukunft, ausgehend von technologischen Veränderungen im World Wide Web.

Keywords. Suche, Ranking,Web, Vektorraummodell, PageRank, HITS

Einleitung

Eine WWW-Suchmaschine hat die schwierige Aufgabe, zu einer aus wenigen Begriffen bestehenden Anfrage des Benutzers die passenden Dokumente zu präsentieren. Bei der automatischen Suche kommt es hierbei nicht so sehr darauf an, alle möglicherweise relevanten Dokumente zu finden (im einfachsten Fall alle, die mindestens einen der Suchbegriffe enthalten), sondern aus der Vielzahl potenzieller Treffer die wirklich relevanten herauszufinden. Während es bei kleineren Dokumentsammlungen noch praktikabel sein mag, alle Ergebnisse anzuzeigen, die zu der Suchanfrage passen, ist dies bei den heute üblichen Dimensionen nicht mehr akzeptabel. Der Index von Web-Suchmaschinen umfasst mehrere Milliarden Webseiten, und eine Anfrage liefert daher häufig mehrere tausend oder zehntausend Ergebnisse zurück. Für den Benutzer wäre es offensichtlich unbefriedigend, wenn diese Ergebnisse ohne weitere Unterteilung oder Sortierung präsentiert würden: Um manuell die relevanten Seiten aus der Fülle der Ergebnisse herauszusuchen, wäre ein erheblicher Aufwand nötig.

Um dem Suchenden diese undankbare Aufgabe abzunehmen, werden die Ergebnisse in der Praxis absteigend nach geschätzter Relevanz angeordnet, so dass die wichtigsten Ergebnisse auf einen Blick zu erfassen sind. Da Computer derzeit leider noch nicht in der Lage sind, den Inhalt der Dokumente zu verstehen, müssen die Suchmaschinen auf einfache statistische Verfahren und Heuristiken zurückgreifen, um das Konzept der Relevanz annähern zu können.

Wir gehen in diesem Kapitel von einer eindeutigen, nur von der Suchanfrage abhängigen Relevanz aus. In der Praxis wird man diesen Idealfall nicht erreichen: Zwei Personen, die dieselbe Anfrage stellen, können durchaus verschiedene Informationsbedürfnisse haben, die beispielsweise vom Vorwissen oder von der aktuellen Suchsituation abhängen können. Da die Suchmaschine jedoch keinen Zugriff auf das eigentliche Informationsbedürfnis hat, ist die vereinfachende Annahme einer objektiven Relevanz, die eine eindeutige Ergebnissortierung ermöglicht, nötig. Diese Einschränkung ist nicht so gravierend, wie es zunächst scheinen mag: Zum einen wird es typischerweise für jede Anfrage eine vorherrschende Tendenz des Informationsbedürfnisses geben, zum anderen bleibt dem Suchenden bei einer Fehlinterpretation der Anfrage noch die Neuformulierung der Anfrage als Lösung.

Dieses Kapitel ist wie folgt aufgebaut: Zunächst betrachten wir Standardverfahren für Information Retrieval, wie sie schon vor den WWW-Zeiten zur Textsuche eingesetzt wurden. Anschließend gehen wir auf die Besonderheiten im Web ein und erklären gängige Erweiterungen in diesem Bereich. Da kommerzielle Anliegen im Web eine große Rolle spielen, versuchen unlautere Anbieter, die Suchergebnisse zu ihren Gunsten zu beeinflussen; die Gegenmaßnahmen der Suchmaschinen werden im folgenden Abschnitt behandelt. Schließlich gehen wir noch auf mögliche weitere Entwicklungen der Web-Suche ein.

Wir gehen hier ausschließlich auf Textsuche ein, da hier die Suchverfahren am ausgereiftesten sind. Auch wenn Suchmaschinen die Suche nach Bildern ermöglichen, wird technisch gesehen hierbei eine Textsuche im Umfeld der Bilder eingesetzt, da die Bilderkennungssoftware noch nicht ausgereift genug ist, den Inhalt der Bilder selbst zu erfassen.

Bei den Betrachtungen in diesem Kapitel sollte berücksichtigt werden, dass die konkreten Suchverfahren der Web-Suchmaschinen aus verständlichen Gründen Geschäftsgeheimnisse sind und allenfalls bruchstückhaft bekannt sind. Die Ausführungen stützen sich daher neben den spärlichen öffentlichen Informationen der Suchmaschinen auf Forschungsergebnisse aus dem Bereich der HTML-Suche. Ob und in welcher Form diese in der Forschung entwickelten Verfahren letztlich von den Suchmaschinen umgesetzt werden, ist nicht genau bekannt, aber es ist anzunehmen, dass diese Forschungsergebnisse in die Ähnlichkeitsfunktionen der Suchmaschinen einfließen.

1. Allgemeine Rankingverfahren

Zunächst gehen wir auf die Suche in herkömmlichen Dokumentsammlungen wie Bibliothekskatalogen ein, wie sie seit Jahrzehnten praktiziert wird. Auch wenn viele der Verfahren veraltet scheinen, bauen die aktuellen Verfahren doch darauf auf, so dass diese Verfahren für das Verständnis der Grundlagen hilfreich sind. Weitere Details zu allgemeinen Suchverfahren finden sich bei Baeza-Yates und Ribeiro-Neto [1].

Ausgangspunkt der Suche ist immer ein Mensch mit einem Informationsbedürfnis. Dieses Informationsbedürfnis muss der Benutzer für die Suchmaschine in eine Anfrage umsetzen; die Anfrage besteht hierbei aus einer Menge von Schlüsselwörtern, den Suchbegriffen. (Theoretisch sind auch komplexere Anfragesprachen denkbar; jedoch hat sich in der Praxis herausgestellt, dass schlüsselwortbasierte Suche in den meisten Fällen ausreicht und den Vorteil der einfachen Verständlichkeit hat.)

1.1. Anforderungen an Rankingverfahren

Die allgemeine Aufgabenstellung bei der Suche ist es, dem Benutzer zu seiner Anfrage möglichst passende Dokumente zu präsentieren. Während es bei kleinen Dokumentsammlungen noch ausreichen mag, einfach alle Dokumente aufzulisten, die mindestens einen Term mit der Anfrage gemeinsam haben (sogenanntes *Boole'sches Retrieval*), ist dies schon bei Sammlungen moderater Größe nicht mehr praktikabel, da es zu viele Übereinstimmungen gibt. Im World Wide Web schließlich wäre eine derart einfach gestrickte Suchfunktion vollends zum Scheitern verurteilt – ein einfacher Begriff wie "Auto" taucht derzeit in etwa einer Milliarde Seiten im Google-Index auf. Demzufolge ist es unabdingbar, dass die Suchmaschine den Benutzer weitergehend unterstützt und diese unzähligen passenden Seiten so sortiert, dass die vermutlich wichtigsten als erste präsentiert werden. Idealerweise sollte der Benutzer bereits auf der ersten Ergebnisseite einen passenden Treffer finden.

Ein erster Ansatz ist es, bei mehreren Suchbegriffen solche Treffer höher zu bewerten, die mehr Suchbegriffe abdecken. Bei einer Suche nach "Auto Fahrrad" werden dann solche Seiten höher bewertet, die sowohl den Begriff "Auto" als auch den Begriff "Fahrrad" enthalten; Seiten, die nur einen der Suchbegriffe enthalten, werden zwar gefunden, jedoch danach eingereiht. Diese minimale Erweiterung der Boole'schen Suche ist unter dem Begriff Coordinate-level matching bekannt.

2. Das Vektorraummodell

Coordinate-level matching ist zwar einfach zu implementieren, jedoch bei einer Dokumentensammlung der Größe des World Wide Web längst nicht ausreichend – auch bei Suchanfragen mit vielen Suchbegriffen wird es im Allgemeinen eine sehr hohe Anzahl an Seiten geben, die alle Begriffe enthalten, so dass eine weitere Differenzierung nötig ist. Nach Möglichkeit sollte es hierbei höchst selten vorkommen, dass mehrere Seiten von der Suchmaschine als gleichermaßen relevant bewertet werden. Hierzu wurden in der Forschung verschiedene Verfahren mit unterschiedlichen theoretischen Modellen entwickelt:

- Probabilistische Verfahren [2] versuchen, die Wahrscheinlichkeit der Relevanz für jedes Dokument zu berechnen.
- Das Vektorraummodell [3] versucht mit statistischen Verfahren, die auf der Anzahl der Vorkommnisse der Suchbegriffe in den Dokumenten beruhen, die Ähnlichkeit jedes Dokuments zur Suchanfrage zu berechnen.
- Linguistische Verfahren wie Latent semantic analysis [4] versuchen, automatisch die Texte in begrenztem Umfang zu verstehen, so dass beispielsweise Mehrdeutigkeiten von Wörtern wie "Bank" (Sitzmöbel oder Geldinstitut) aus dem Kontext aufgelöst werden können.

Die meisten dieser Verfahren haben den Schritt zur kommerziellen Verwendung im großen Maßstab wie im World Wide Web nicht geschafft. Als in der Praxis meistverwendetes Vefahren hat sich hierbei das Vektorraummodell herauskristallisiert. Dies ist insofern verwunderlich, da im Gegensatz zu anderen Verfahren keine fundierte Theorie als Grundlage der Entwicklung diente; in der Praxis stellte sich jedoch die Ergebnisqualität als sehr gut heraus. Erst deutlich nach der Etablierung des

Vektorraummodells wurde durch Sparck Jones et al. mit BM25 [5] eine Verbindung von Vektorraummodell und probabilistischen Verfahren hergestellt, so dass nun auch theoretisch nachvollzogen werden konnte, warum das Verfahren so gut funktioniert. Da sich das Vektorraummodell auch sehr effizient implementieren lässt [6], ist es sehr gut für die großen Datenmengen geeignet, wie sie im World Wide Web zu bewältigen sind.

2.1. Lokale und globale Termgewichte

Innerhalb eines Dokuments sind nicht alle Terme gleich wichtig; manche berühren das Hauptthema des Dokuments, während andere nur am Rande erwähnt werden. Um die relative Wichtigkeit der Terme innerhalb eines Dokuments abschätzen zu können, wird die Termhäufigkeit, also die Anzahl der Auftreten eines gewissen Terms (abgekürzt *TF* für *term frequency*), als sogenanntes lokales Termgewicht, verwendet; die Annahme hierbei ist, dass zentrale Begriffe häufiger erwähnt werden.

Bei der Verwendung mehrerer Suchbegriffe in einer Anfrage, zum Beispiel "Haus des Nikolaus", ist eine weitere Verfeinerung der Relevanzabschätzung sinnvoll: Nicht alle Suchbegriffe sind für die Relevanz der Dokumente gleich wichtig. In diesem Beispiel ist es relativ belanglos, ob ein Dokument häufig den Begriff "des" enthält, viel entscheidender ist das Auftreten des selten auftretenden Begriffs "Nikolaus". Die globale Wichtigkeit eines Terms lässt sich durch eine einfache Heuristik recht gut abschätzen: Terme, die in fast allen Dokumenten der Sammlung auftreten - beispielsweise die Artikel "der", "die", "das" - erhalten ein geringes Gewicht, da sie nicht geeignet sind, zwei Dokumente zu unterscheiden. Ein Term hingegen, der in nur wenigen Dokumenten auftritt, erhält ein hohes Gewicht, da er auf sehr spezielle Inhalte verweist, die ansonsten zwischen den häufigeren Termen untergehen könnten.

Ein extremer Ansatz besteht in der Verwendung von Stoppwortlisten, die solche Wörter enthalten, die für die Suche meist nutzlos sind. Solche Wörter werden nicht mit in den Index aufgenommen, weshalb nicht effektiv nach diesen Wörtern gesucht werden kann. Das wird im Normalfall kein Problem darstellen, kann jedoch bei speziellen Anfragen das Auffinden der relevanten Dokumente komplett unmöglich machen; beispielsweise würde die Suche nach "Sein oder nicht sein" bei üblichen Stoppwortlisten komplett ins Leere laufen.

Zur Berechnung wird hier meist die inverse Dokumenthäufigkeit (englisch "inverse document frequency", abgekürzt IDF) verwendet [7]. Hierbei erhalten Terme, die sehr selten sind, ein höheres Gewicht als häufige Terme. In der folgenden Formel gibt $df(t)$ die Anzahl der Dokumente in der Sammlung C an, die den Term t mindestens einmal enthalten

$$idf(t) = \log\left(\frac{|C|}{df(t)+1}\right)$$

Um das Gesamtgewicht eines Terms zu bestimmen, wird dann die Termhäufigkeit mit der inversen Dokumenthäufigkeit multipliziert, auch bekannt als TF-IDF.

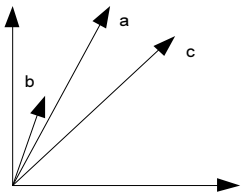
2.2. Ähnlichkeitsfunktionen

Es gibt mehrere Möglichkeiten, auf Basis der lokalen und globalen Termgewichte die Ähnlichkeit eines Dokumentes zur Anfrage zu berechnen. Weitgehend durchgesetzt hat sich die Darstellung der Dokumente und Anfragen als Vektoren reeller Zahlen; hierbei

wird jedem Term ein bestimmter Index in diesen Vektoren zugeordnet, und jede Komponente des Vektors beinhaltet hierbei das lokale Termgewicht (die Termhäufigkeit) des zugehörigen Terms:

$$\vec{d} = (t_1, t_2, \ldots, t_n)$$

Bei der Berechnung der Ähnlichkeit ist es wünschenswert, ein Dokument umso höher zu bewerten, je mehr Terme mit der Anfrage übereinstimmen. Ein erster Ansatz wäre es, einfach das Skalarprodukt der Dokumentvektoren zu verwenden; dieses Verfahren würde jedoch lange Dokumente ungerechtfertigt bevorzugen. Um dieses Problem zu umgehen, wird stattdessen der Winkel zwischen den Dokumentvektoren als Maßstab verwendet – je kleiner er ist, desto ähnlicher sind sich die Dokumente. Aus praktischen Gründen wird der Cosinus verwendet der sich umgekehrt zum Winkel verhält (ein rechter Winkel hat den Cosinus 0, ein Winkel von 0 Grad hat den Cosinus 1):



$$\mathrm{sim}(d_1, d_2) = \cos\left(\vec{d_1}, \vec{d_2}\right) = \frac{\vec{d_1} \cdot \vec{d_2}}{|\vec{d_1}| \cdot |\vec{d_2}|}$$

Abbildung 1.

2.3. Vorteile des Vektorraummodells

Auch wenn das Vektorraummodell im Vergleich zu probabilistischen oder linguistischen Verfahren recht einfach ist und theoretisch nur schwer erklärbar ist, hat es sich in praktischen Experimenten als qualitativ ebenbürtig bewiesen. Neben der guten Ergebnisqualität schlägt auch die Einfachheit des Verfahrens zu Buche, die sich zum einen in einer einfachen Implementierung und zum anderen in der vergleichsweise hohen Geschwindigkeit der Suche niederschlägt [6].

2.4. Weitere Einflussfaktoren

Neben den bisher genannten Faktoren – lokale und globale Termgewichte – können sich auch noch andere Faktoren auf den Ähnlichkeitswert auswirken [8]:

- Position der Suchbegriffe: Falls die Suchbegriffe eher am Anfang des Dokuments vorkommen, werden sie höher gewichtet.

- Nähe der Suchbegriffe untereinander: Wenn in der Anfrage mehrere Begriffe vorkommen, werden Dokumente höher bewertet, in denen die Begriffe nah beieinander auftreten.
- Reihenfolge der Suchbegriffe: Bei einer Anfrage, die aus mehreren Begriffen besteht, werden die Anfragebegriffe, die am Anfang der Anfrage stehen, höher gewichtet.

An dieser Stelle sei noch einmal darauf hingewiesen, dass die tatsächlich verwendeten Verfahren der Suchmaschinen nicht bekannt sind; es ist jedoch anzunehmen, dass sie als Ergebnis jahrelanger Verbesserungen deutlich komplexer sind und mehr Faktoren benutzen, als hier beschrieben wurde.

3. Erweiterungen der Basisverfahren für WWW-Suchmaschinen

Die bisher beschriebenen Suchverfahren gehen von einfach strukturierten "flachen" Texten ohne Auszeichnungen wie Zwischenüberschriften und Fettdruck aus. Die im World Wide Web verwendete Auszeichnungssprache HTML (siehe Abbildung 2) bietet jedoch viele Möglichkeiten zur optischen Gestaltung einer Webseite, und es liegt nahe, diese Auszeichnungen auch für die Suche zu verwenden, um die Ergebnisqualität zu verbessern.

Ferner hat das World Wide Web die Besonderheit, dass die Seiten untereinander über Links verknüpft sind. Da man im allgemeinen davon ausgehen kann, dass ein Autor nur Verknüpfungen auf Seiten anlegt, die er für hilfreich hält, kann man in einem Link eine Art Empfehlung für die Zielseite sehen. Diese Information kann wiederum verwendet werden, um die Qualität einer Website abzuschätzen: Je häufiger auf eine Seite verwiesen wird, desto höher ist ihre Qualität.

3.1. Ausnutzung der Markup-Informationen

Dazu sollte man sich vor Augen führen, wie ein Mensch einen Text inhaltlich einordnet, ohne ihn komplett zu lesen. Beim Überfliegen eines Textes wird er zunächst auf den Titel und die Überschriften achten, die schon einen recht guten Überblick über Inhalt und Struktur geben. Demzufolge sollten Schlüsselwörter, die im Titel und in Überschriften auftreten, ein höheres Gewicht erhalten als solche, die im normalen Fließtext stehen.

```
<body>
   <h1>AG Datenbanken und Informationssysteme</h1>
   <p>Die AG <em>Datenbanken und Informationssysteme</em> wurde ..
</body>
```

AG Datenbanken und Informationssysteme

Die AG *Datenbanken und Informationssysteme* wurde ...

Abbildung 2. HTML-Quellcode und gerenderte Form.

Außerdem können auch noch die Hyperlinks zwischen Dokumenten wichtige Hinweise für relevante Schlüsselwörter geben: Wenn von einem HTML-Dokument auf ein

anderes verwiesen wird, wird dieser Link im Quelldokument mit einem entsprechend markierten Text, dem Anchor-Text, verbunden, zum Beispiel:

```
<a href=”x.html”>Anchor-Text</a>
```

Sinnvollerweise wird der Autor darauf achten, dass dieser Anchor-Text den Inhalt des Zieldokumentes kurz und prägnant beschreibt. Der Suchmaschine wird hierdurch ermöglicht, bei der Indexierung auch die Sicht fremder Autoren zu berücksichtigen; so kann es auch dazu kommen, dass ein bestimmtes Suchergebnis die Suchbegriffe gar nicht enthält – dann nämlich, wenn sie nur im Anchor-Text der Seiten auftauchen, die auf dieses Dokument verweisen.

Verschiedene Untersuchungen haben ergeben, dass hierdurch tatsächlich die Ergebnisqualität verbessert werden kann, und so wurde und wird diese Technik in verschiedenen Varianten von Web-Suchmaschinen eingesetzt.

Cutler et al. ([9,10]) speicherten verschiedene Teile der HTML-Dokumente in separaten Indizes:

- Dokumenttitel (<title>)
- Überschriften höherer Ebene (<h1>, <h2>)
- Überschriften niederer Ebene (<h3> bis <h6>)
- Hervorhebungen (, <b>, , <i>, <u>, <dl>, <ol>, <ul>)
- Anchor-Texte (<a>)
- Normale Texte (Text, der nicht durch obige Liste abgedeckt ist)

Die Texte, die in diesen verschiedenen Kategorien auftauchen, werden auch in verschiedenen Indizes abgespeichert. Bei der Suche wird nun separat für jeden Index die Ähnlichkeit zum Suchbegriff berechnet und die einzelnen Teilähnlichkeiten mit linearer Gewichtung kombiniert. Cutler et al. fanden heraus, dass Hervorhebungen und Anchor-Texte ein achtmal so hohes Gewicht haben sollten wie normaler Text. Das ist insofern bemerkenswert, als die Anchor-Texte kein Bestandteil des eigentlichen Textes sind; offenbar verwenden Autoren bei Querverweisen auf andere Dokumente im Allgemeinen treffende Kurzbeschreibungen. Insgesamt ergaben sich 26 Prozent Verbesserung gegenüber einer Suchmaschine ohne separate Gewichtung dieser Kategorien.

Ähnliche Experimente wurden von Liu et al. [11] im Rahmen von TREC durchgeführt, wobei eine deutlich reduzierte Anzahl von Kategorien verwendet wurde. Das könnte auch die Erklärung dafür sein, dass sich hier nur leichte Verbesserungen gegenüber der Baseline ergaben.

Einen anderen Ansatz verfolgten Robertson et al. [12], die keine separaten Indizes für die verschiedenen Feldtypen verwendeten, sondern lediglich das lokale Termgewicht änderten, so dass zum Beispiel Überschriften ein doppelt so hohes Gewicht hatten wie normaler Fließtext. Diese Variante erwies sich in Experimenten der linearen Kombination – die von Cutler et al. und Liu et al. verwendet wurde – überlegen.

Insgesamt haben sich solche Verfahren in der Praxis bewährt, so dass sie auch heute noch von Web-Suchmaschinen eingesetzt werden. Hierbei ist allerdings unklar, welche Varianten benutzt werden; es ist zu vermuten, dass die hier vorgestellten Möglichkeiten in der Praxis noch deutlich erweitert und angepasst werden.

Neben den oben aufgeführten Kategorien kommen jedoch noch weitere zur Anwendung; so werden beispielsweise teilweise zusätzliche Header-Felder (Meta-Header) ausgewertet, die Schlüsselwörter zum aktuellen Dokument enthalten.

3.2. Ausnutzung der Verlinkung

Neben dem Anchor-Text können Suchmaschinen Hyperlinks auch noch verwenden, indem die Linkstruktur ausgenutzt wird, um hieraus implizite Empfehlungen zu gewinnen.

3.2.1. Global

Neben der dokumentspezifischen Ähnlichkeit einer Seite wird auch noch eine andere Komponente in die Sortierung einbezogen, die von der eigentlichen Suchanfrage unabhängig ist. Hintergrund ist, dass nicht jede Website für den Benutzer den gleichen Stellenwert hat; beispielsweise wird ein großes Nachrichtenportal ein höheres Gewicht haben als eine unbekannte private Seite. Diese intrinsische Wichtigkeit einer Seite lässt sich nicht über den Inhalt bestimmen. Vielmehr wird stattdessen der Bekanntheitsgrad der Seite im Web zu Hilfe genommen, was sich durch eine algorithmische Annäherung aus der Linkstruktur bestimmen lässt.

Eine wichtige Implementierung dieses Prinzips wurde von Page et al. [13] unter dem Namen PageRank als Basis des Google-Suchalgorithmus bekannt.

Es wird ein Benutzer simuliert, der willkürlich auf Links klickt und gelegentlich mit einer gewissen Wahrscheinlichkeit eine Seite besucht, die nicht verlinkt ist (dies ist nötig, da sonst nicht das gesamte Web erfasst werden könnte). Je häufiger eine Seite bei dieser Simulation besucht wurde, als desto wichtiger wird sie angesehen.

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

Hierbei ist $PR(X)$ der PageRank der Seite X, T_i sind die Seiten, die auf Seite A verweisen und $C(T_i)$ die Anzahl der Links ist, die von T_i ausgehen; d ist ein

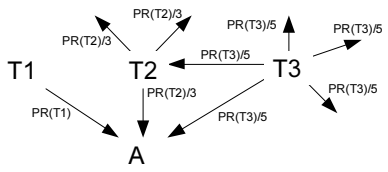


Abbildung 3. PageRank-Beispiel; A hat drei eingehende Links. Der PageRank von T1 wird am höchsten gewichtet, da T1 nur auf A verweist, während zum Beispiel T3 noch auf vier weitere Seiten verweist.

Dämpfungsfaktor, der typischerweise auf einen Wert um 0.85 gesetzt wird (dabei entspricht d im Modell der Wahrscheinlichkeit, dass ein Link besucht wird, und 1-d ist die Wahrscheinlichkeit, dass auf eine beliebige andere Seite gesprungen wird).

Abbildung 3 zeigt ein Beispiel für die Berechnung des PageRanks von Seite A. Gehen wir von einem initialen PageRank von 1 für jede Seite aus, so ergibt sich in der ersten Iteration folgender Wert für den PageRank von A:

$$PR(A) = \frac{PR(T_1)}{C(T_1)} + \frac{PR(T_2)}{C(T_2)} + \frac{PR(T_3)}{C(T_3)} = \frac{1}{1} + \frac{1}{3} + \frac{1}{5} = \frac{23}{15}$$

Offensichtlich können sich bei der Berechnung des PageRank Probleme ergeben, wenn die Linkstruktur Zyklen enthält, also beispielsweise Seite A auf Seite B verweist und umgekehrt. Zur Berechnung des PageRank von Seite A wird der PageRank von Seite B benötigt, der jedoch erst berechnet werden kann, wenn der PageRank von A bekannt ist. In der Praxis wird PageRank als Fixpunktiteration berechnet, das heißt, allen Seiten wird ein initialer PageRank zugewiesen und dann wiederholt die PageRanks für alle Seiten neu berechnet, bis sich keine signifikanten Änderungen mehr ergeben.

Hieraus ergibt sich, dass die Berechnung des PageRank-Wertes für eine große Menge von Seiten – wie es im Web zweifellos der Fall ist – sehr aufwändig ist, so dass sie nicht bei jeder Indexaktualisierung durchgeführt werden kann. Für neue Seiten bedeutet das, dass sie bei der PageRank-Berechnung von vorneherein benachteiligt wären und somit auch kaum eine Chance hätten, so bekannt zu werden, dass viele Seiten auf sie verlinken. Da das offensichtlich nicht erwünscht ist, kann man davon ausgehen, dass neue Seiten bevorzugt behandelt werden [8].

3.2.2. Lokal

Neben PageRank gibt es auch konkurrierende Verfahren wie Kleinbergs Hyperlink-Induced Topic Search *(*HITS*)* [14]. Der grundlegende Unterschied zu PageRank ist hierbei, dass HITS einen anfragespezifischen Qualitätswert für eine Seite berechnet, anstatt einen globalen Wert zu berechnen. Das hat den Vorteil, dass genauer auf die Anfrage reagiert werden kann, aber den Nachteil, dass die zeitintensiven Berechnungen während der Beantwortung der Anfrage durchgeführt werden müssen, was die Antwortzeit deutlich erhöhen kann.

Hierzu werden jeder Seite zwei verschiedene Werte zugewiesen, die deren Qualität als Hub und Authority bewerten. Hubs sind hierbei Seiten, die sich durch gute ausgehende Links auszeichnen (beispielsweise Web-Verzeichnisse), während Authorities als besonders hilfreich für das Anfragethema angesehen werden können.

Bei der Berechnung ergibt sich hier ein Wechselspiel zwischen diesen beiden Werten: Der Hub-Wert ist die Summe der Authority-Werte der Seiten, auf die verwiesen wird, und der Authority-Wert ergibt sich aus der Summe der Hub-Werte der Seiten, die auf die aktuelle Seite verweisen. Demzufolge ist eine Seite ein besonders guter Hub, wenn sie auf viele Seiten mit hohen Authority-Werten verweist, und eine gute Authority zeichnet sich dadurch aus, dass viele gute Hubs auf sie verweisen.

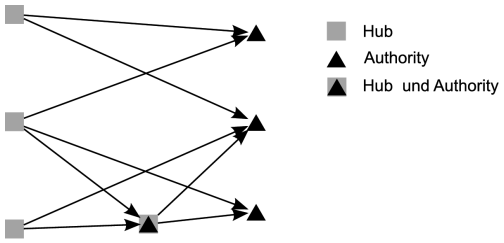


Abbildung 4. Beispiel von Hubs und Authorities: Hubs verweisen auf möglichst viele Authorities, und Authorities werden von möglichst vielen Hubs erreicht. Es kann auch Zwischenstufen geben, zum Beispiel Seiten, die sowohl Hub als auch Authority sind.

Wie schon erwähnt arbeitet HITS nicht auf der gesamten Dokumentsammlung, sondern einer anfragenspezifischen Teilmenge. Da im Allgemeinen die Menge der Dokumente, die die Suchbegriffe erhält, zu groß sein wird, um in akzeptabler Zeit den Algorithmus abzuarbeiten, schränkt Kleinberg die Ausgangsmenge auf die 200 am höchsten von einem Basisalgorithmus gerankten Dokumente ein. Diese Ausgangsmenge wird einen hohen Anteil an relevanten Seiten enthalten, aber im Allgemeinen nicht, wie beabsichtigt, einen Großteil der Autoritäten. Deshalb wird diese Ausgangsmenge noch um die Dokumente erweitert, die über Links mit ihr verbunden sind.

Um den Einfluss von reinen Navigationslinks zu verringern, werden zudem noch Links innerhalb derselben Domain von Links zwischen verschiedenen Domains unterschieden. Links innerhalb derselben Domain werden nun aus dieser Menge entfernt, bevor schließlich die Hub- und Authority-Werte berechnet werden.

Durch die zwei separat berechneten Werte ist es prinzipiell möglich, dem Suchenden je nach den aktuellen Bedürfnissen bevorzugt einzelne Seiten mit passendem Inhalt oder Übersichtsseiten, die auf viele passende Seiten verweisen, zu präsentieren.

Interessanterweise ist der Ursprung der Verfahren, die die Verlinkung ausnutzen jedoch nicht im Information Retrieval zu suchen, sondern stammt aus der Analyse sozialer Netzwerke [15], wo die Algorithmen nützlich sind, den Einfluss einer bestimmten Person in einem sozialen Umfeld zu bestimmen.

4. Spamming

Web-Suchmaschinen sind für viele Benutzer der Einstiegspunkt ins World Wide Web, mit Hilfe der Suchmaschinen finden sie nicht nur allgemeine Informationen wie Nachrichten, sondern auch Produkte. Somit ist es für die Anbieter von großem Interesse, bei den Suchergebnissen an möglichst prominenter Stelle aufzutauchen. Neben legitimen Methoden wie bezahlter Werbung oder Suchmaschinenoptimierung verwenden skrupellose Unternehmen auch unseriöse Tricks, um die Aufmerksamkeit der Benutzer zu erregen. Ziel ist es, die Ranking-Verfahren der Suchmaschinen auszutricksen, um möglichst auf der ersten Ergebnisseite zu erscheinen. In Analogie zu unlauteren Werbeverfahren über E-Mail nennt man diese Verfahren *Spamming*. Gyongyi und Garcia-Molina [16] kategorisierten 2005 die damals aktuellen Arten von Web-Spam:

- Term-Spam (hierbei wird der Seiteninhalt manipuliert):
  - Haupttextspam
  - Titelspam
  - Meta-Tag-Spam (heute nicht mehr aktuell)
  - Anchor-Text-Spam
  - URL-Spam (z. B. buy-canon-rebel-20d-lens-case.example.com)
- Link-Spam (hierbei werden die Querverweise manipuliert):
  - eingehende Links
  - ausgehende Links

Im Folgenden werden wir die wichtigsten Spam-Arten der beiden Hauptkategorien genauer besprechen.

4.1. Tricks der Spammer

Da die Ranking-Verfahren auf den Inhalten der Texte beruhen und hierbei insbesondere das häufige Auftreten einzelner Terme zu einem besseren Ranking führt, liegt es als Manipulationsmöglichkeit nahe, häufig nachgefragte Begriffe nicht nur einmal auf der Seite zu erwähnen, sondern möglichst oft.

Um zu verhindern, dass die Benutzer diesen Trick zu leicht durchschauen, weil sie mit einer mit häufigen Suchbegriffen überfrachteten Seite konfrontiert werden, bedienen sich Spammer häufig des sogenannten Cloakings. Hierbei wird der Suchmaschine ein anderer Seiteninhalt präsentiert als dem Benutzer, und dieser Inhalt ist gezielt darauf ausgerichtet, hoch gerankt zu werden.

Schwieriger ist es, die auf der Linkstruktur basierenden Verfahren auszutricksen, da es hier nötig ist, andere Seiten auf die eigenen zeigen zu lassen. Normalerweise ist es natürlich unmöglich, fremde Seiten zu verändern, um Links einzufügen, aber es gibt Ausnahmen: Wikis sind darauf ausgelegt, von jedermann bearbeitet werden zu können, und ein Großteil der Blogs bietet Besuchern die Möglichkeit an, Kommentare zu hinterlegen. Für die Spammer ist das natürlich von großem Interesse, denn sie können sich hier den hohen PageRank bekannter Wikis und Blogs zunutze machen und auf die eigene Seite abfärben lassen, indem sie hier Links einfügen.

Der Manipulation fremder Seiten sind jedoch enge Grenzen gesetzt; zum einen ist es (auch bei teilweiser Automatisierung) aufwändig, die Links einzutragen, und zum anderen besteht immer die Gefahr, dass der Betreiber der fremden Seite die für ihn unerwünschten Links wieder entfernt. Deshalb bietet sich ein Verfahren an, das in großem Stil den PageRank nach oben treibt: In sogenannten Link-Farmen sind sehr viele Seiten vollständig miteinander vernetzt, was einen maximalen PageRank garantiert. Diese Link-Farmen lassen sich komplett automatisch erstellen, so dass der Aufwand für den Spammer vergleichsweise gering ist.

4.2. Gegenmaßnahmen der Suchmaschinen

Selbstverständlich liegt es im Interesse der Suchmaschinenbetreiber, Spamming zu unterbinden, da es die Ergebnisqualität beeinträchtigt und somit droht, die Werbeeinnahmen zu schmälern. Deshalb werden die Algorithmen so angepasst, dass die Tricks der Spammer möglichst ins Leere laufen.

So führten Gyongyi et al. [17] TrustRank ein, ein Verfahren, das halbautomatisch Spam-Seiten von sinnvollen Seiten unterscheidet. Hierbei muss zunächst von

Menschen eine Menge von Seiten als Nicht-Spam gekennzeichnet werden; ausgehend von diesen Seiten berechnet TrustRank dann Spam-Werte für von diesen Startseiten ausgehenden Seiten durch Links erreichbaren Seiten.

Außerdem können manuell bestimmte Seiten, die unfaire Tricks benutzt haben, für einen gewissen Zeitraum aus dem Index ausgeschlossen werden. Ein prominentes Beispiel war 2006 BMW: Nachdem auf der Hauptseite bmw.de Cloaking betrieben wurde, wurde diese Domain von Google vorübergehend aus dem Index verbannt und verlor auch ihren PageRank.[1]

Insgesamt handelt es sich hierbei jedoch um eine Art Rüstungswettlauf, bei dem die Spammer mit immer neuen Methoden versuchen, die Ranking-Verfahren der Suchmaschinen für ihre Zwecke auszunutzen; die Suchmaschinen ihrerseits müssen dann wiederum reagieren.

5. Zusammenfassung und Ausblick

5.1. Zusammenfassung

Auch wenn Web-Suchmaschinen im Wesentlichen Verfahren des Standard-Information-Retrieval verwenden, gibt es doch verschiedene Möglichkeiten, die Besonderheiten des WWW auszunutzen, um bessere Ergebnisqualität zu bekommen. Neben der Ausnutzung der Textauszeichnung zur besseren Gewichtung der Terme bietet sich hier vor allem die Auswertung der vorhandenen Linkstruktur an, so dass automatisch Qualitätsaussagen über bestimmte Webseiten gemacht werden können.

Darüber hinaus ist es in der unkontrollierbaren Umgebung nötig, den Einfluss unredlicher Organisationen, die nur auf den eigenen Vorteil bedacht sind, in Zaum zu halten, so dass die Suchergebnisse nicht beliebig von außen manipuliert werden können.

Zu diesem Zweck wurden verschiedene Erweiterungen der Standardverfahren aus dem Information Retrieval entwickelt, die auf diese Einflussfaktoren Rücksicht nehmen.

5.2. Ausblick

Auch wenn mit einem guten Rankingverfahren die Bedürfnisse der Benutzer im Allgemeinen gut erfüllt werden können, gibt es noch Verbesserungsmöglichkeiten. Bei einer einfachen Anfrage wie “Apple” hat die Suchmaschine keine Möglichkeit, zu erraten, ob der Benutzer die Frucht, die Computerfirma oder die Plattenfirma meint, und es besteht bei jeder dieser Möglichkeiten die Gefahr, dass die Entscheidung falsch war und dass somit irrelevante Ergebnisse dargestellt werden. Dieses Problem kann durch die Verwendung von Clustering entschärft werden; hierbei werden die Ergebnisse nicht mehr nur linear präsentiert, sondern zusätzlich in thematisch kohärente Gruppen zusammengefasst; für die Anfrage “Apple” schlägt sie Suchmaschine clusty.com beispielsweise die Kategorien “Mac” und “Tree” vor.

1 siehe http://blog.omc.ch/2006/02/05/google-kickt-bmw-de-und-ricoh-de-wegen-suchmaschinenspam-aus-dem-index/

Eine andere Möglichkeit, die passende Interpretation einer Anfrage zu erraten, besteht darin, umfangreiche Benutzerprofile anzulegen. Wenn man anhand der vergangenen Anfragen feststellen kann, dass der Benutzer regelmäßig nach Computerbegriffen gesucht hat, liegt es nahe, die entsprechende Interpretation auch für die ungenaue Anfrage "Apple" zu verwenden. Mit der Anpassung an den Benutzer ergeben sich viele neue Möglichkeiten, allerdings ist diese umfangreiche Erhebung der Daten aus datenschutzrechtlichen Erwägungen nicht unkritisch, so dass unklar ist, wie weit man in dieser Richtung gehen können wird.

5.2.1. Semantische Suche

Die in diesem Kapitel vorgestellten Suchverfahren beruhen allesamt auf einfachen statistischen Verfahren, die keinerlei Verständnis der Inhalte durch die Suchmaschine voraussetzen. Anfragen wie "gib mir alle Seiten aus, die Produktbewertungen zu Kühlschränken von europäischen Benutzern beinhalten" sind so nicht sinnvoll zu bearbeiten, hierfür wäre es nötig, dass die Suchmaschine die Inhalte der Webseiten analysieren und die Inhalte in strukturierter Form extrahieren kann. Das ist beim derzeitigen Stand der Technik im Allgemeinen nicht möglich, und deshalb haben sich Richtungen entwickelt, die die Suchmaschinen (oder allgemein Computeranwendungen) unterstützen wollen, indem die Informationen nicht nur in der rein textuellen HTML-Form, sondern zusätzlich noch in einer strukturierten Form zur Verfügung gestellt werden, die für Programme leicht analysierbar ist.

In dieser Richtung schlugen Berners-Lee et al. [18] 2001 das sogenannte Semantic Web vor, dass auf der Idee basiert, dass Web-Inhalte durch zusätzliche Erläuterungen in einem maschinenlesbaren Format (RDF) ergänzt werden (beispielsweise kann so der Autor einer Seite strukturiert beschrieben werden). Bis heute konnte sich jedoch das Semantic Web in der ursprünglich anvisierten Form nicht durchsetzen, was vermutlich zumindest teilweise an dem zusätzlichen Wartungsaufwand für die Zusatzinformationen liegt. Durch die Trennung der Inhalte von den Metainformationen besteht auch immer die Gefahr, dass letztere veralten – bei Änderungen der Inhalte kann es leicht vergessen werden, die Metadaten zu aktualisieren, und da diese nicht direkt angezeigt werden, wird das nicht unbedingt auffallen.

Um diese Probleme zu umgehen, wurden die Mikroformate (Microformats) erfunden, deren Ziel es ist, die Metadaten wieder mit den Inhalten zusammenzuführen und die Wartung möglichst einfach zu halten. Kernidee hierbei ist es, keine neuen Technologien zu verwenden, sondern so auf bestehende Standards zurückzugreifen, dass die semantischen Informationen direkt mit den Inhalten verknüpft werden. Das wird dadurch erreicht, dass normales HTML-Markup in Kombination mit Klassenattributen verwendet wird.

Im März 2008 kündigte Yahoo an, Mikroformate und RDF bei der Web-Suche zu unterstützen[2]. Vorerst werden sie verwendet, um gewöhnliche Suchergebnisse mit semantischen Informationen anzureichern, eine direkte semantische Suche wird noch nicht unterstützt. Hierbei stellt sich ohnehin die Frage, wie eine brauchbare Anfragesprache für das Semantic Web aussehen könnte – existierende Anfragesprachen wie SPARQL[3] sind nicht für den Endanwender gedacht, da sie zu komplex sind, andererseits wird man mit einfachen Schlüsselwortanfragen die semantischen Informationen nicht vernünftig ausnutzen können.

2 siehe http://www.ylocalblog.com/blog/2006/06/21/we-now-support-microformats/
3 siehe http://www.w3.org/TR/2008/REC-rdf-sparql-query-20080115/

5.2.2. Benutzerspezifisches Ranking

Eine wesentliche Grundlage der Standardverfahren ist, dass zwei Benutzer, die dieselbe Suchanfrage stellen, auch dieselben Ergebnisse erhalten sollten. Dies mag auf den ersten Blick naheliegend erscheinen, hat jedoch den Nachteil, dass sich die Suchmaschine nicht auf die speziellen Bedürfnisse eines Benutzers einstellen kann. Wenn die Suchmaschine die Suchhistorie des Benutzers kennt und beim Ranking berücksichtigt, ist ein hohes Maß an Personalisierung möglich; dies ermöglicht zum Beispiel die Disambiguierung mehrdeutiger Begriffe (wie "apple"). Hierbei sind allerdings auch die gesetzlichen Regelungen zum Datenschutz zu berücksichtigen, die – insbesondere in Europa – die Datensammlung stark einschränken.

5.2.3. Verwendung der Informationen aus Bookmarkdiensten und anderen Quellen

Auf speziellen Seiten wie del.icio.us können Benutzer ihre Bookmark-Sammlung öffentlich im Internet zur Verfügung stellen und mit Tags versehen. Im Allgemeinen bieten die Bookmark-Dienste nur einfache Suchverfahren auf dem Datenbestand (insbesondere basierend auf den Tags) an; in Verbindung einer Volltext-Suchmaschine ist es jedoch möglich, die Empfehlungen, die sich implizit durch das Anlegen eines Bookmarks ergeben, mit einer Schlüsselwortsuche zu kombinieren.

Die Annahme ist hierbei, dass Bookmarks – im Gegensatz zu den normalen Web-Links, wie sie von PageRank und HITS verwendet werden – ausschließlich als Empfehlungen angesehen werden können.

Neben Bookmark-Diensten haben sich Frage-Antwort-Dienste als konkurrenzfähig zu herkömmlicher Suche herausgestellt [19]. Auch hier bietet es sich demnach an, die Ergebnisse mit klassischen Suchergebnissen zu kombinieren, um die Vorteile beider Welten zu erhalten.

5.2.4. Suche von Teildokumenten

Alle gängigen Web-Suchmaschinen haben derzeit die Einschränkung, dass als Ergebnisgranulat Webseiten zurückgeliefert werden; der Suchende muss nun selbst herausfinden, wo genau in den Ergebnisseiten das für ihn relevante Material vorkommt. Dies ist bei kurzen Dokumenten noch problemlos möglich, es gibt mittlerweile jedoch auch viele Dokumente, die viele Bildschirmseiten umfassen. Hier wäre es für den Benutzer sehr hilfreich, wenn die Suchmaschine ihn zu dem relevanten Teil – beispielsweise einem Abschnitt oder Absatz – leiten könnte, der für ihn relevant ist.

Hierzu gibt es in der Forschung bereits Ansätze: Passage retrieval [20] versucht, ohne Berücksichtigung der logischen Struktur des Dokuments passende Textpassagen zu finden, während beim XML-Retrieval [21,22] die Struktur berücksichtigt wird, so dass eine logische Einheit wie ein einzelner Abschnitt gefunden werden kann.

Insgesamt ist es zu erwarten, dass sich im Bereich der Web-Suche in den kommenden Jahren noch vieles tut, um das Sucherlebnis für den Benutzer zu verbessern. Viele der möglichen Erweiterungen werden schon testweise angeboten, werden sich aber vermutlich vor der Massentauglichkeit noch merklich ändern. Da teilweise auch Änderungen an den Web-Seiten vorausgesetzt werden (semantische Suche), hängt der Erfolg wesentlich davon ab, ob die Autoren diese zusätzlichen Möglichkeiten auch annehmen.

Literaturangaben

[1] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley, Harlow, Essex, England, 1999.

[2] Norbert Fuhr. Probabilistic models in information retrieval. The Computer Journal, 35(3):243–255, 1992.

[3] Gerard Salton. The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, 1971.

[4] Michael W. Berry, Zlatko Drmac, and Elizabeth R. Jessup. Matrices, vector spaces, and information retrieval. SIAM Review, 41:335–362, 1999.

[5] Karen Spärck Jones, Steve Walker, and Stephen E. Robertson. A probabilistic model of information and retrieval: development and status. Technical report, Computer Laboratory, University of Cambridge, 1998.

[6] Ian H. Witten, Alistair Moffat, and Timothy C. Bell. Managing Gigabytes. Morgan Kaufmann, 1999.

[7] Stephen Robertson. Understanding inverse document frequency: On theoretical arguments for IDF. Journal of Documentation, 60(5):503–520, 2004.

[8] Dirk Lewandowski. Web Information Retrieval. DGI, 2005.

[9] Michael Cutler, Yungming Shih, and Weiyi Meng. Using the structure of HTML documents to improve retrieval. In Proceedings of the USENIX Symposium on Internet Technologies and Systems, 1997.

[10] M. Cutler, H. Deng, S. S. Maniccam, and W. Meng. A new study on using HTML structures to improve retrieval. In ICTAI 1999 proceedings, pages 406–409. IEEE, 1999.

[11] Yiqun Liu, Canhui Wang, Min Zhang, and Shaoping Ma. Finding ”abstract fields” of web pages and query specific retrieval – THUIR at TREC 2004 web track. In TREC 2004 proceedings, 2004.

[12] Stephen Robertson, Hugo Zaragoza, and Michael Taylor. Simple BM25 extension to multiple weighted fields. In CIKM 2004 proceedings, pages 42–49. ACM, 2004.

[13] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. In Computer Networks and ISDN Systems, 1998.

[14] Jon Kleinberg. Authoritative sources in a hyperlinked environment. In Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1997.

[15] L. Katz. A new status index derived from sociometric analysis. Psychometrika, 18:39–43, 1953.

[16] Zoltan Gyongyi and Hector Garcia-Molina. Web spam taxonomy. In Proc First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb 2005), 2005.

[17] Zoltan Gyongyi, Hector Garcia-Molina, and Jan Pedersen. Combating web spam with trustrank. Technical report, Stanford University, 2005.

[18] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web: a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 284:34–43, 2001.

[19] Dirk Lewandowski and Christian Maaß, editors. Web-2.0-Dienste als Ergänzung zu algorithmischen Suchmaschinen, Berlin, 2008.

[20] Gerard Salton, James Allan, and Chris Buckley. Approaches to passage retrieval in full text information systems. In SIGIR 1993 proceedings, pages 49–58. ACM, 1993.

[21] Norbert Fuhr, Norbert Gövert, Gabriella Kazai, and Mounia Lalmas, editors. Proceedings of the 1st INEX Workshop. ERCIM, 2002.

[22] Norbert Fuhr, Mounia Lalmas, and Andrew Trotman, editors. Comparative Evaluation of XML Information Retrieval Systems. 5th International Workshop of the Initiative for the Evaluation of XML Retrieval, INEX 2006. Springer, 2007.