Über Suchmaschinen und Datenbanken

Mit dem Siegeszug des Web ist der größte Informationsspeicher in der Geschichte der Menschheit entstanden. Suchmaschinen und Datenbanken ermöglichen uns eine einfache und effiziente Suche zu den Informationen des Web. Wie funktionieren Suchmaschinen und Datenbanken eigentlich? In dem folgenden Beitrag werden die Ansätze und Methoden von Suchmaschinen und Datenbanken vorgestellt.

Wolfram Sperber

Die Nadel im Heuhaufen suchen – diese Metapher ist uns aus den Kindertagen für eine schier aussichtslose Suche geläufig, die nur mit übernatürlichen Kräften bewältigt werden kann. Heute heißen unsere Helfer für die Suche nach Informationen Google, Yahoo, Altavista oder in der Mathematik etwa die Datenbank Zentralblatt MATH (ZMATH).
Suchmaschinen und Datenbanken können heute weit mehr als die Nadel im Heuhaufen finden, sowohl was die quantitative als auch, was die qualitative Seite betrifft.
Wie ist das möglich? Nichts geht mehr heute ohne Mathematik – das gilt auch und gerade für Suchmaschinen. Im Folgenden soll am Beispiel von Google's PageRank über die Konzepte und Methoden der Web-Suchmaschinen berichtet und mit den Leistungen fachspezifischer Datenbanken verglichen werden.

Das Web – „unendliche Weiten"

Wie groß ist das Web eigentlich? - Diese Frage kann nicht exakt beantwortet werden. Schon die Frage, was alles zum Web gehört, ist alles andere als einfach zu beantworten.
Eine Abschätzung über die Größe des Web liefern die Suchmaschinen selbst. Um einigen der oben genannten Schwierigkeiten aus dem Weg zu gehen, vereinfacht man die Aufgabenstellung: Man fragt „nur noch" nach der Anzahl der Dokumente, die für die Suchmaschinen zugänglich sind. Der Teil des Web, der für die Suchmaschinen zugänglich ist, wird i.A. als Surface Web bezeichnet. Im Wesentlichen verbirgt sich dahinter die Anzahl der stationären Webseiten, die mit anderen Webseiten verlinkt sind.
Jede Suchmaschine zählt für sich die Anzahl der Dokumente, die über sie suchbar sind. Man stellt dann Testanfragen an die verschiedenen Suchmaschinen und vergleicht die Trefferlisten der verschiedenen Suchmaschinen. Aus dem Vergleich lässt sich dann die ungefähre Größe des Surface Web abschätzen. Anfang 2005 wurde die Größe des Web mit mehr als 20.000.000.000 ($= 2 \times 10^{10}$) Webseiten angegeben, d.h. es gibt wesentlich mehr Webseiten als Bewohner der Erde, und es ist wesentlich größer als ein (normaler) Heuhaufen.

Abb. 16: Google-Anfrage „Mathematik"

Die Nadel im Heuhaufen

Suchen und Finden von Informationen – das beschäftigt die Menschen schon seit langem. Informationen wurden zunächst in Stein gehauen oder geritzt, später auf Papyrus, Pergament und Papier geschrieben und heute zunehmend in elektronischer Form abgespeichert. Als die Anzahl der Papyrusrollen zu groß wurde, begannen die Ägypter, Römer und Griechen, die Papyrusrollen mit zusätzlichen Angaben über den Inhalt, die Autoren, die Herkunft etc. zu beschriften. Damit hatte die Katalogisierung (systematische Erschließung) begonnen. Von Aristoteles sind Überlegungen über die Systematisierung von Systemen (Klassifikation) und erste Klassifikationsschemata bekannt, die sich auch in Bibliothekssystemen niedergeschlagen haben.
Die Erschließung und das Zugänglichmachen der Informationen ging in die Hände von Spezialisten (Bibliothekare, Dokumentare, Archivare, etc.) über. Die zunehmende Anzahl von Publikationen, gerade auch in den Wissenschaften, im 19. und 20. Jahrhundert sprengte die Möglichkeiten der Bibliotheken (die lokalen Bibliotheken waren zunehmend außerstande, alle neu erschienenen Publikationen zu erwerben und zu verwalten) und machte zusätzliche Informationsdienste erforderlich, etwa das Jahrbuch über die Fortschritte der Mathematik oder das Zentralblatt für Mathematik, die von der mathematischen Community initiiert und getragen wurden und werden. Sie sammeln die Publikationen in der Mathematik und erschließen sie systematisch. Experten aus den Fachgebieten erstellen Referate über die Publikationen. Das Zentralblatt für Mathematik und das Jahrbuch über die Fortschritte der Mathematik in der gedruckten Version bzw. heute als Datenbank ermöglichen einen einfachen und schnellen Zugang zu diesen Daten und damit zum aktuellen Wissen in der Mathematik.
Das Web brachte – auch für das Wissensmanagement in der Mathematik – neue Herausforderungen

- Das Web ist, wie oben dargestellt, inzwischen der größte Informationsspeicher der Menschheit. Es werden mehr Informationen bereitgestellt denn je, zu den klassischen Dokumenttypen sind neue hinzugekommen, in der Mathematik etwa Software, Simulationen, etc.
- Die Informationen liegen weltweit verteilt vor: Es gibt für das Web keine zentrale Steuerung. Jeder, der einen Computer und einen Internetanschluss hat, kann beliebige Informationen erstellen und im Netz publizieren. Es gibt damit auch keine zentrale Qualitätskontrolle.
- Die Informationen sind unstrukturiert und i.A. nicht erschlossen: Der Verzicht auf eine zentrale Steuerung und Kontrolle bedeutet, dass der Nutzer nicht nur frei ist zu entscheiden, ob und welche Informationen er bereitstellt, sondern auch über die Art und Weise der Bereitstellung.
- Die Informationen sind dynamisch: Im Gegensatz zu gedruckten Publikationen weist das Web eine hohe Dynamik auf. Dokumente können auf einfache Art und Weise erstellt, aus dem Web entfernt oder an andere Stellen verschoben werden.
- Die Informationen sind verlinkt: Das Vorbild für die Verlinkung sind die Referenzen in den wissenschaftlichen Publikationen. Die Hypertext Markup Language (HTML) macht per Mausklick einen direkten Übergang von einer Seite zu einer anderen Webseite möglich.

Das unglaubliche Wachstum und die Größe des Web machen automatische Methoden erforderlich, um die Informationsflut zu bewältigen. Um eine Suche nach Informationen im Web zu ermöglichen, wurde die klassische Vorgehensweise der Bibliotheken und Referateorgane von den Suchmaschinen adaptiert:

1 Sammeln der Informationen
2 Indexieren
3 Abfrage-Schnittstelle für den Nutzer

Das Sammeln der Informationen übernehmen die so genannten Crawler oder Gatherer. Die Crawler arbeiten sich anhand der Hyperlinks durch das Web (d.h. eine Startmenge von Webseiten wird analysiert, vorhandene Links werden verfolgt und neu gefundene Dokumente für die Auswertung durch die Suchmaschine aufbereitet).
Die Indexierung (Katalogisierung) besteht aus speziellen Methoden für die Auswertung und Analyse der Dateien. Das umfasst die Extraktion der Daten (etwa von Wörtern eines Textdokuments) und das Abspeichern der Daten in verschiedenen Indexen.
Der dritte Schritt, die Abfrageschnittstelle der Suchmaschine, transformiert die Anfrage des Nutzers in eine für die Suchmaschine verständliche Abfrage und leitet sie an die Indexe weiter.

Eine Schwierigkeit bleibt, die jeder Nutzer von Suchmaschinen kennt. Im Allgemeinen ist die Anzahl der Treffer bei einer Anfrage sehr hoch. Der Treffer müssen – in Abhängigkeit von der Suchanfrage – nach ihrer Relevanz (Bedeutung) sortiert werden. Jeder Nutzer möchte die „wichtigsten“ Treffer an erster Stelle sehen.

Und genau an dieser Stelle hat Google mit dem PageRank angesetzt. Die Wurzeln von Google's PageRank liegen in der bibliometrischen Auswertung wissenschaftlicher Literatur.

Impact Factor und Zitationsanalyse

Bibliometrie beschäftigt sich mit der statistischen Auswertung von Texten und Publikationen. Ein solcher Untersuchungsgegenstand sind die Literaturreferenzen in den Publikationen.
Die Zitationsanalyse von Zeitschriften setzt auf dem „Impact Factor“ auf. Der Impact Factor ist ein Maß für Zitationen wissenschaftlicher Zeitschriften und wird häufig als Kriterium für die Bedeutung einer Zeitschrift benutzt. Der Impact Factor einer Zeitschrift wird berechnet als Quotient der Anzahl der Zitierungen eines Artikels einer Zeitschrift in einem bestimmten Zeitraum und der Anzahl aller Artikel dieser Zeitschrift in demselben Zeitraum. Eine Zeitschrift, deren Artikel durchschnittlich einmal zitiert werden, hat also den Impact Factor 1.
Der Impact Factor wertet den gerichteten Graphen, der aus den Dokumenten als Knoten und den Zitationen als Kanten gebildet wird.
Seit Anfang der sechziger Jahre werden die Zitationen von ca. 6000 wissenschaftliche Zeitschriften vom „Institute for Scientific Information“ (ISI) ausgewertet und in den Journals Citation Reports (http://scientific.thomson.com/products/jcr) publiziert.
Das Für und Wider und die Relevanz daraus abgeleitete Bewertungen bibliometrischer Methoden in den Wissenschaften ist ausführlich diskutiert worden. Bibliometrische Maße sind einerseits einfach zu ermittelnde und auch für den Laien verständliche Kriterien für die Bewertung einer wissenschaftlicher Arbeit, beschränken sich aber auf ein einziges Kriterium, nämlich die Analyse der Zitationen. Der Impact Factor macht keine Aussage über die tatsächliche Bedeutung einer Arbeit. Er ist mehr ein Maß für die Nützlichkeit und die Popularität einer Publikation denn für ihre Qualität.
1976 haben Gabriel Pinski und Francis Narin einen rekursiven Impact Factor eingeführt, um Zitierungen in Zeitschriften mit einem hohen Impact Factor höher zu wichten als solche in Zeitschriften mit kleinem Impact Factor. An diese Idee knüpft der PageRank von Google unmittelbar an. Für weitere Informationen sei auf die entsprechende Wikipedia-Seite
http://en.wikipedia.org/wikipedia/impact_factor
verwiesen.

Der PageRank: Sag mir, wer über Dich spricht, und ich sag Dir, wie wichtig Du bist

Google's PageRank basiert - analog den Zitationen für wissenschaftliche Publikationen - auf der Analyse des gerichteten Graphen, der durch die Hyperlinkstruktur des Web definiert ist.

Die Links, die auf eine Webseite zeigen, sind ein Maß für die Popularität einer Web Seite bzw. die Wichtigkeit dieser Webseite angenommen.

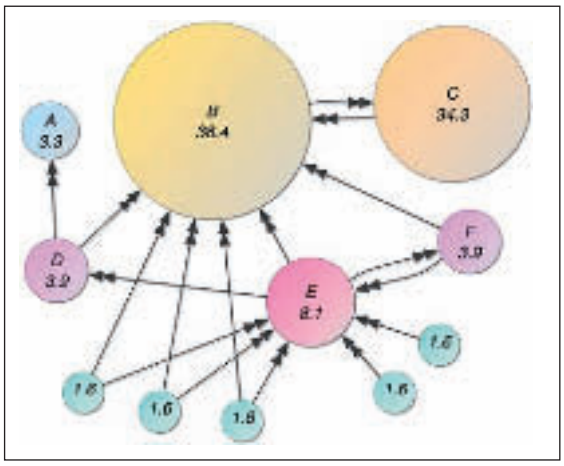
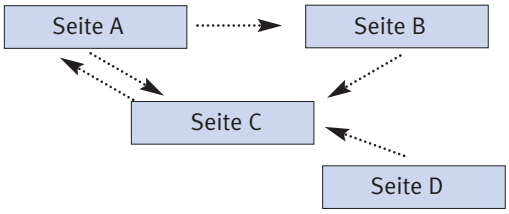


Abb. 17: Linkanalyse (Google’s PageRank)

Als konkretes Beispiel betrachten wir den folgenden Graph, der aus den Dokumenten A, B, C und D besteht



Die Links lassen sich durch folgende Tabelle charakterisieren:

	A	B	C	D	(Ausgangspunkt des Links)
A	-	-	x	-	
B	x	-	-	-	
C	x	x	-	x	
D	-	-	-	-	
(Endpunkt des Links)					

Kein Link entspricht der 0, ein Link von einer Seite auf eine andere wird mit 1/|.| bewertet, wobei |.| die Anzahl aller Links ist, die von der Seite ausgehen (im Beispiel ist |B| = |C| = |D| = 1, |A| = 2). Das „Gewicht einer Seite“ (der PageRank) wird also gleichmäßig auf alle Links aufgeteilt, die von der Seite ausgehen.
Das führt dann zu dem folgendem linearen Gleichungssystem. Die in der Mathematik übliche Matrix-Vektorschreibweise ergibt sich in natürlicher Weise aus der obigen Tabelle:

$$\begin{bmatrix} p(A) \\ p(B) \\ p(C) \\ p(D) \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} p(A) \\ p(B) \\ p(C) \\ p(D) \end{bmatrix}$$

Dabei bezeichnet p(.) den PageRank einer Seite.

Das Gleichungssystem für das gesamte Web ist riesig. Man versucht daher, das Gleichungssystem mit einfachen Methoden zu lösen. Ein solches Lösungsverfahren ist etwa das Iterationsverfahren. Im Beispiel heißt das: Man nehme beliebige Startwerte für p(A), p(B), p(C) und p(D) und setze diese in die rechte Seite des Gleichungssystems (1) ein. Die Berechnung der rechten Seite von (1) liefert neue Werte für p(A), p(B), p(C) und p(D). Diese dienen wiederum als neue Startwerte für die zweite Iteration, etc. Das Iterationsverfahren erfordert pro Schritt nur wenige Multiplikationen und Additionen, lässt sich also einfach auf einem Rechner implementieren.

Einige Bemerkungen über Google's Modell und den PageRank:

1 Einfache Rechnungen zeigen, dass es bei der Iteration Schwierigkeiten geben kann. Die Lösung kann etwa periodisch sein, das Iterationsverfahren konvergiert also nicht.
Lange vor Google haben sich Mathematiker mit dieser Art von Gleichungssystemen (Perron-Frobenius Theorie über nichtnegative Matrizen) beschäftigt und wissen, wie das Gleichungssystem (also die Zahlen in der Koeffizientenmatrix) beschaffen sein muss, damit man es iterativ lösen kann.

2 Page und Brin, die Gründer von Google, sind bei ihren Überlegungen von dem zu Grunde gelegten Modell ausgegangen. Dazu untersuchen sie das Verhalten beim Surfen im Web.
Die Situation: Ein Surfer befindet sich auf einer beliebigen Webseite. Er wählt irgendeinen Link auf dieser Webseite aus. Was ist, wenn diese Webseite keinen Link (man spricht dann von dangling nodes) enthält? Ein weiterer Fall: Der Surfer bricht seinen Spaziergang durch das Web an dieser Stelle ab und kehrt zu einer bekannten Adresse, etwa aus seiner Bookmark-Liste, zurück. Die Berücksichtigung des Surfverhaltens führt mathematisch zu einer Modifikation des Gleichungssystems (1) (auf das mathematische Modell übersetzt heißt das: die Koeffizientenmatrix wird „irreduzibel“ und „primitiv“). Das modifizierte Gleichungssystem lässt sich dann als ein spezieller stochastischer Prozess auffassen, nämlich als Markov-Kette, ein Prozess, der aus der Wahrscheinlichkeitsrechnung gut bekannt ist. Die Eigenschaft einer Markov-Kette besagt, dass der untersuchte Prozess „ohne Gedächtnis“ ist, dass also das Surfverhalten nicht von der Vorgeschichte abhängt. Interessant ist, dass Page und Brin in ihren Publikationen (Brin, & Page 1998; Brin, Page, Motwami, & Winograd, 1999) Markov-Ketten nicht erwähnen, in ihrem Modell des Surfens das stochastische Verhalten aber intuitiv korrekt eingebaut haben.

3 Eine ausführlichere Darstellung des mathematischen Hintergrunds würde den Rahmen dieses Artikels sprengen, dazu sei auf das Buch von (Langville, & Meyer, 2006) verwiesen. Für weitere Publikationen, die sich mit dem PageRank beschäftigen, verwiesen wir auf unsere Datenbank ZMATH, www.zentralblatt-math-org/zmath/en. Fragen Sie doch einfach mal nach „PageRank“ oder „Google“.

4 Google berechnet tatsächlich den PageRank jeder Webseite iterativ aus einem Gleichungssystem. Das Gleichungssystem ist gigantisch, mit heute geschätzten 50 bis 100 Milliarden Seiten (Variablen) ist Google zumindest eines der größten Gleichungssysteme, wenn nicht das größte Gleichungssystem, das jemals gelöst worden ist.

Google's PageRank-Verfahren hat die Informationssuche im Web grundlegend verändert. Mit automatischen Mitteln (der Lösung von Gleichungssystemen) lässt sich ein Ranking erzeugen, das Google binnen kürzester Zeit zur weltweit populärsten Suchmaschine gemacht hat.

Ähnlich wie beim Impact Factor der Zitationsanalysen erhalten bei diesem Vorgehen populäre Webseiten (also solche, auf die viele Links zeigen) einen hohen PageRank[1].

[1]Google berücksichtigt aber bei seinem Ranking weitere Faktoren, etwa den Zeitpunkt der Veröffentlichung, den Dokumententyp oder persönliche Interessen des Anfragenden.

Google und ZMATH

Macht Google ZMATH zukünftig überflüssig? Wohl kaum. Google bewertet weder den Inhalt noch die Qualität von Publikationen. Google's Ziel ist es, möglichst viele Webseiten zu erschließen und dem Nutzer eine einfache Abfrage der Informationen zu bieten. Google beschränkt sich dabei auf die Informationen, die im Web vorhanden sind (das sind in der Mathematik bei weitem nicht alle, wobei der Anteil der Informationen, die zumindest auch im Web vorhanden sind, steigt). Google entwickelt keine speziellen Methoden für mathematische Informationen (etwa eine Systematisierung nach der Mathematical Subject Classification oder die Auswertung von in der Mathematik üblichen Dokumentformaten wie TeX oder MathML) .
Fokus der Datenbank ZMATH sind Publikationen aus der Mathematik und ihren Anwendungsgebieten, diese werden in ZMATH nahezu vollständig erfasst. Die Publikationen sind als Artikel in Zeitschriften, als Bücher oder Beiträge in Proceedings erschienen, sind also bereits wissenschaftlich evaluiert. Die hohe Qualität von ZMATH wird durch die intellektuelle Aufbereitung, Systematisierung und inhaltliche Besprechung der Publikationen geprägt und gesichert. Das Fundament dafür ist die breite Mitarbeit der mathematischen Community bei der Erzeugung der Datenbank. Zunehmend werden in ZMATH auch automatische Verfahren, etwa zur Verlinkung mit Volltexten oder zur Autorenidentifikation genutzt, um

- die weiter steigenden Publikationszahlen in der Mathematik (ZMATH nähert sich 100.000 erfassten Arbeiten p.a.) zu bewältigen
- die Funktionalität und Qualität von ZMATH (neue Such- und Navigationsmöglichkeiten oder die Vernetzung mit anderen relevanten Informationen) auszubauen.

Statt der Alternative „Universelle Suchmaschine" oder „Spezifische Datenbank" erscheint eine kooperative Arbeitsteilung zwischen beiden Klassen von Informationssystemen sinnvoll.
Fachspezifische Datenbanken wie ZMATH können durch zusätzliche Informationen und Dienste die universellen Suchmaschinen wirkungsvoll ergänzen und erweitern und umgekehrt. Der Einsatz automatischer Methoden und Verfahren kann – wie am Beispiel des PageRank für die Websuche angedeutet – sinnvoll sein. Ob darüber hinaus eine engere Kooperation von Betreibern universeller Suchmaschinen und fachspezifischer Datenbanken möglich ist, muss die Zukunft zeigen.

Ein kurzes Fazit: Die Entwicklung der Suchmaschinen im Web hat die Methoden für die Auswertung und die Verarbeitung von Informationen radikal verändert. Die Entwicklung der Suchmaschinen hat gezeigt, dass mit automatischen Methoden und Verfahren große Informationsmengen aufbereitet und ausgewertet werden können. Die intellektuelle Aufbereitung kann damit aber nicht ersetzt werden und ist gerade in den Wissenschaften unverzichtbar.
Die heutigen Suchmaschinen und die Datenbanken sind Teil der Entwicklung intelligenter, effizienter und nutzerfreundlicher Informationsdienste.

Literaturverzeichnis

Brin, S., & Page, L. (1998). The anatomy of a large hypertextual Web engine. Computer Networks and ISDN Systems , S. 107 - 117.

Brin, S., Page, L., Motwami, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. Stanford: Stanford University, Computer Science Department.

Langville, A. N., & Meyer, C. D. (2006). Google's PageRank and Beyond TheScience of Search Engine Rankings. Princeton and Oxford: Princeton University Press.

Dr. Wolfram Sperber
FIZ Karlsruhe, Abteilung Mathematik und Informatik,
Editor Zentralblatt MATH