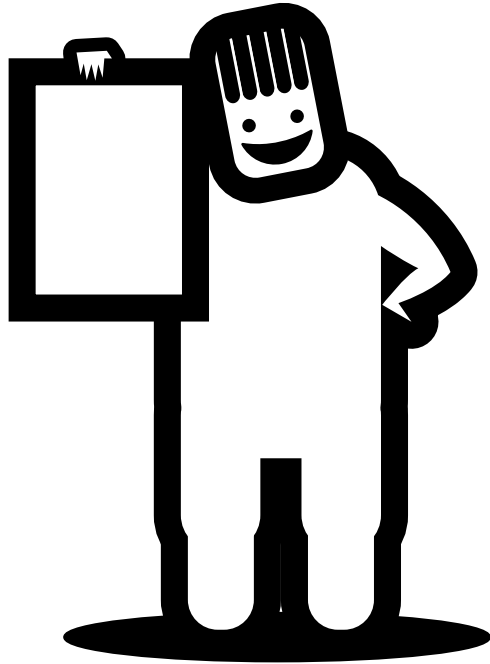




Informationsbeschaffung im Internet

Grundlegende Konzepte verstehen und umsetzen

Werner Hartmann

Michael Näf

Peter Schäuble

Urheberschaft: Werner Hartmann, Michael Näf, Peter Schäuble

Illustrationen: François Chalet

Das vorliegende Dokument ist die kostenfreie Version im Format PDF des 2000 im Orell Füssli Verlag Zürich unter demselben Titel erschienenen Buches, das unterdessen aber vergriffen ist.

Verschiedene Rechte vorbehalten. Jegliche kommerzielle Nutzung dieser Unterlagen ist untersagt. Erlaubt ist die private sowie alle nicht-kommerzielle Nutzung (z. B. an Schulen). Im Detail gelten die Bestimmungen unter <http://www.internet-kompetenz.ch/copyright/>.

Inhalt

Einleitung	6
1 Internet als Informationsquelle	11
Daten, Wissen und Information im Internet	13
Informationsdienste	17
Puzzle-Steine eines Informationssystems	21
Zwei fiktive Dienste	23
2 Moderne Suchmethoden	27
Probleme bei der Informationssuche	29
Relevanz eines Dokuments	32
Gewichtung von Dokumenten nach Relevanz	34
Rangierungsprinzipien	36
Die Rangliste in der Praxis	42
Rangierungsprinzipien in der Praxis	42
Inhaltsunabhängige Bestimmung der Relevanz	43
Web Spamming	44
3 Indexierung von Textdokumenten	47
Eine Beispielindexierung	49
Eigene Experimente helfen weiter	54
4 Funktionsweise von Suchsystemen	59
Der Web-Roboter	62
Der Index	64
Puzzle gelöst: das Suchsystem ist komplett	67
Wie landen Webseiten im Index?	69

Decken reale Suchdienste das ganze Web ab?	70
Wieso erscheint eine Seite nicht im Index?	70
Metasuchdienste	72
5 Metadaten und	
Boole'sche Suchmethoden	77
Boole'sche Operatoren	79
Boole'sche Suche im Dokumentinhalt	79
Metadaten	81
Suche in Teilkollektionen mittels Metadaten	83
Definition von Teilkollektionen in der Praxis	87
Vorsicht: Boole'sche Suche im Dokumentinhalt	89
6 Interaktive Suchtechniken	95
Iterativer Ablauf einer Recherche	97
Suche nach ähnlichen Dokumenten	99
Relevanzrückkoppelung	100
Manuelle Anfrageerweiterung	102
Interaktive Techniken in der Praxis	102
Manuelle Relevanzrückkoppelung	102
Interaktive Suche im Beispiel	103
Manchmal führt ein Umweg schneller zum Ziel	104
7 Katalogdienste	107
Aufbau von Katalogsystemen	109
Manuelle Erstellung	110
Automatische Klassifizierung von Dokumenten	111
Konkrete Katalogdienste im Internet	115
Gegenüberstellung: Such- und Katalogdienste	115
8 Push-Dienste	121
Hol-Prinzip und Bring-Prinzip	122
Push-Dienste der einfachsten Art	123
Push-Dienste der flexibleren Art	124
Push-Dienste in der Praxis	128
Mailing-Listen und Newsgroups	128

9	Evaluation von Informationsdiensten	131
	Kriterien zur Bewertung von Informationsdiensten	132
	Informationsdienste in der Praxis beurteilen	138
10	Suchtipps	141
	Richtige Dokumentensammlung wählen	143
	Richtiges Werkzeug benutzen	144
	Viele präzise Suchbegriffe verwenden	145
	Irrelevante Dokumente in der Rangliste ignorieren	146
	Dem System mitteilen, was man weiss	147
	Bestraungtes Dokument vollstndig untersuchen	148
	Interaktive Techniken anwenden	149
	In Teilsammlungen suchen	150
	Anfragen in verschiedenen Sprachen formulieren	151
	Rechtschreibung und alternative Schreibweisen beachten . .	152
	Stichwortverzeichnis	153

Einleitung

Internet – Netz der Netze, Datenozean, Information Superhighway. Das Internet kennt viele Bezeichnungen und ebenso viele Vorurteile. In erster Linie ist und bleibt das Internet aber eine riesige, globale Datenquelle. Diese Datensammlung kann man für seine eigenen Interessen nutzen. Voraussetzung für die erfolgreiche Informationsbeschaffung im Internet sind allerdings fundierte Kenntnisse der verfügbaren Werkzeuge. Informationssuche im Netz lässt sich nicht mit der Bedienung eines Web-Browsers gleichsetzen, sondern geht weit darüber hinaus.

Informationssuche im Internet ist ein relevantes Thema. Im weltweiten Netz rücken die Anbieter von Daten und die Informationssuchenden immer näher zusammen, sodass sich eine immer grössere Gemeinschaft von Internet-Anwendern mit dem Thema der Informationsbeschaffung im Internet konfrontiert sieht. Für viele Erwerbstätige gehört die Informationssuche zur täglichen Arbeit und damit zu den wichtigen Schlüsselqualifikationen. Schon 1994 schätzte eine Studie den Aufwand für die Informationsbeschaffung in Europa auf eine Million Personenjahre pro Jahr.

Informationssuche im Internet ist ein anspruchsvolles Thema. Die Menge des weltweit verfügbaren Wissens wächst explosionsartig. Das Internet verkörpert ein wenig stabiles Umfeld. Neue Angebote zur Informationssuche schiessen wie Pilze aus dem Boden. Bestehende Angebote sind fortwährenden Änderungen unterworfen.

Hier setzt das vorliegende Buch an. Es vermittelt langlebiges Wissen in der sehr kurzlebigen Welt des Internets. Der Schwerpunkt liegt auf den grundlegenden Methoden, die auch morgen noch gültig sind. Dabei wird darauf verzichtet, Bedienungsanleitungen zu möglichst

allen Werkzeugen aufzuführen. Wichtiger ist es, die allgemein gültigen Prinzipien der Informationsbeschaffung im Internet zu kennen. Daneben wird die *Hilfe zur Selbsthilfe* gross geschrieben, denn mittels geeigneter Anfragebeispiele lässt sich selbstständig herausfinden, welche Funktionen ein Informationsdienst anbietet und welche nicht.

Zielpublikum

Das Buch richtet sich an einen breiten Personenkreis von Internet-Anwendern und -Anwenderinnen, die über grundlegende Fertigkeiten und Erfahrungen mit dem Internet verfügen und sich im Hinblick auf schnelle und effektive Informationssuche weiterbilden möchten. Das Spektrum reicht vom «Internet-Polizisten» oder Datenschutzbeauftragten über Mitarbeiterinnen in politischen Kommissionen, Medienschaffende, Verlagsmitarbeiter, Bibliothekarinnen bis hin zu Wissenschaftlern, Entwicklungsingenieuren, Bankfachleuten, Marketingspezialistinnen, Medizinerinnen, Ausbildungsverantwortlichen oder Lehrerinnen, Studenten und Schülerinnen aller Fachrichtungen. Kurz: Das Buch richtet sich an alle, die sich mit dem Internet zur effizienten und effektiven Informationsbeschaffung auseinander setzen möchten.

Abgrenzung

Der im Buch behandelte Stoff konzentriert sich ausschliesslich auf das zielgerichtete Suchen und Finden von Informationen im Internet. Weitere Aspekte im Zusammenhang mit dem Thema der Informationsbeschaffung wären zum Beispiel das Auswählen und Bewerten der gefundenen Informationen oder Fragen zur Urheberrechtsproblematik. Wir verzichten aber auf die Behandlung dieser Aspekte, um das Buch nicht unnötig zu überladen. Genügend andere Publikationen widmen sich eingehend solchen Fragestellungen.

Aufbau

Dieses Buch ist in erster Linie als Lehrbuch aufgebaut. Die einzelnen Kapitel liefern jeweils die nötigen Voraussetzungen für die folgenden Kapitel. Dank dem ausführlichen Stichwortverzeichnis dient das Buch aber auch als hilfreiches Nachschlagewerk nach der ersten Lektüre.

Alle Kapitel sind gleich aufgebaut. Ausgangspunkt sind jeweils typische Anwenderprobleme, mit denen – konsequent aus der Perspektive des Anwenders oder der Anwenderin – in das Thema eingeführt wird. Darauf werden die grundlegenden Prinzipien und das Hintergrundwissen für das jeweilige Themengebiet vermittelt. Es folgen die vermehrt praxisorientierten Aspekte, die für die Umsetzung der Theorie in die Praxis nötig sind und in vielen Fällen wertvolle Zusatzinformationen liefern. Im Kapitelabschluss werden die einführenden, beispielhaften Anwenderprobleme anhand des erarbeiteten Wissens gelöst.

Links, Links, Links

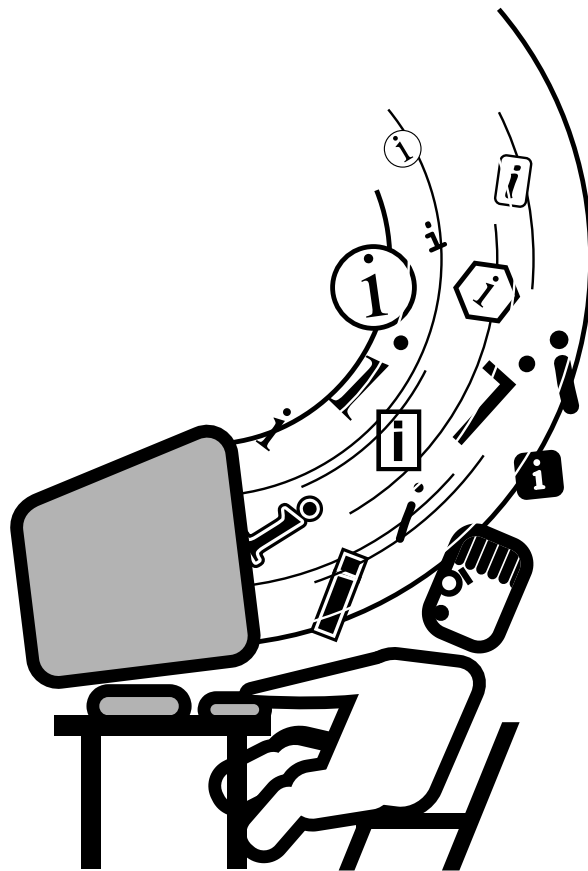
Es wäre ein Leichtes, eine umfangreiche Sammlung von konkreten Internet-Adressen in diesem Buch abzudrucken. Leider wären viele dieser Adressen schon kurz nach der Veröffentlichung des Buches nicht mehr gültig. Aus diesem Grund verzichten wir gänzlich auf konkrete Adressen. Die wenigen angeführten URLs dienen lediglich zur Illustration und erheben keinen Anspruch auf Gültigkeit. Mit einer einzigen Ausnahme:

`http://www.internet-kompetenz.ch/`

Auf den Webseiten unter dieser Adresse werden weiterführende Informationen zum Buch, Linksammlungen und Übungen angeboten.

Kapitel 1

Internet als Informationsquelle





John Cage ist ein Komponist des 20. Jahrhunderts, der häufig neue Wege beschritten hat. In seinem Stück «4'33"» beispielsweise ist während der ganzen viereinhalb Minuten nichts anderes zu hören als das Räuspern und Flüstern des Publikums. Ich möchte nun mehr wissen über das Leben von John Cage. Ein Bekannter rät mir: «Im Internet findest du alles!» Also versuche ich es und besuche einen dieser Suchdienste: OMNISEARCH. Dort gebe ich die Anfrage *cage* ein.

Sofort erhalte ich die Antwort und erlebe gleichzeitig eine herbe Enttäuschung. An erster Stelle wird mir etwas von einem Schauspieler namens Nicolas Cage angeboten. Gefolgt von einem französischen Spielfilm mit dem Titel «La cage aux folles». Dann gibt es irgendwo in den USA den «Mid-West Cage Bird Club» und einige eigenartige Abkürzungen wie «CAGE – Commercial And Government Entity» oder «CAGE – Cyber Art Gallery Eindhoven». Daneben tauchen massenhaft Webseiten von Privatpersonen mit dem Nachnamen Cage sowie kommerzielle Angebote von Herstellern von Vogel- und sonstigen Käfigen auf. Die spärlichen Treffer zum Thema John Cage beschränken sich auf CD-Kataloge oder Ähnliches. So viel also zur grössten Bibliothek der Welt!

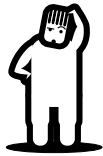


Im Jahre 1997 wurde die Tandem Computers Inc. aufgekauft. Wer übernahm die Firma? Ich versuche es mit OMNISEARCH und der Anfrage *Tandem* und werde auf der Stelle überschwemmt mit etwa 300 000 Treffern. In den meisten der gefundenen Dokumenten geht es um Fahrräder. Immerhin habe ich gelernt, dass begleitete Fallschirmsprünge zu Übungszwecken auch Tandemsprünge genannt werden. Leider hat das nichts mit meiner Frage zu tun – und weit und breit gibt es keine Spur vom Käufer der Firma Tandem.

Zu allem Übel gibt es auch noch technische Probleme. Eines der Dokumente trägt einen viel versprechenden Titel. Doch wenn ich es anwähle, erhalte ich eine Fehlermeldung, und das Dokument wird nicht angezeigt. Ein anderes Dokument wird mir zwar angezeigt, aber erst nach sehr langer Wartezeit. Anscheinend gehört OMNISEARCH zu den langsameren Suchdiensten. Oder wieso dauert es so lange?

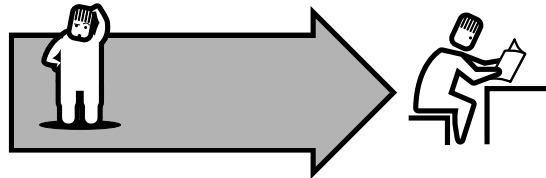


Ende 1998 brachte eine Frau Achtlinge zur Welt. Wie hiess die Frau, und wo fand die Geburt statt? Kein Problem, ich suche einfach mit der Anfrage *Achtlinge* bei OMNISEARCH. Doch was geschieht? Die Rangliste bleibt leer! Funktioniert OMNISEARCH nicht richtig? Oder habe ich etwas falsch gemacht? Ich kann mir einfach nicht vorstellen, dass niemand über diese Geburt berichtet hat.



Ein leidiges Thema für alle Computerbesitzer sind die Viren. Regelmässig taucht die bedrohliche Meldung in den Medien auf, es sei ein neuer Virus im Umlauf. Häufig kommen diese Warnungen aber zu spät. Deshalb möchte ich mich auf eigene Faust auf dem Laufenden halten. Die Anfrage *virus update* bei OMNISEARCH liefert mir brauchbare Dokumente. Also nehme ich mir vor, die Anfrage einmal pro Woche durchzuführen, um so über die neuen Viren informiert zu bleiben. Allerdings ist es mühsam, dass ich immer dran denken muss und dass ich immer wieder Dokumente erhalte, die ich schon einmal durchgesehen habe. Gibt es für mein Anliegen kein besseres Werkzeug?

Nach diesen Problemen aus der Sicht der Internet-Anwender wird es höchste Zeit für etwas Hintergrundinformation. Auf die gestellten Fragen werden wir am Ende des Kapitels zurückkommen ...

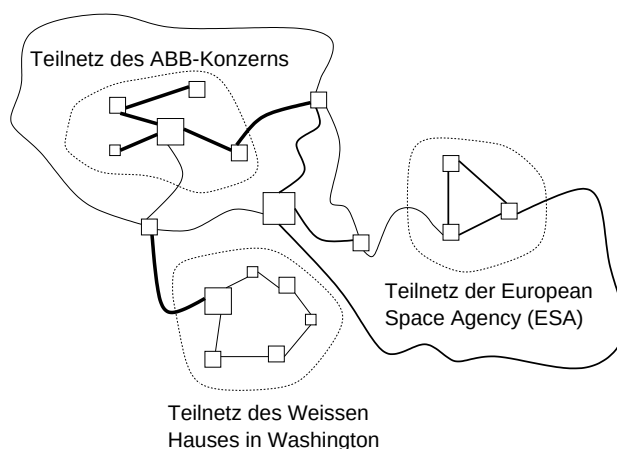


Daten, Wissen und Information im Internet

Im Internet steht eine riesige Menge von *Daten* zu den unterschiedlichsten Themen zum Abruf bereit. Alle im Internet verfügbaren Daten zusammengekommen ergeben das im Internet gespeicherte *Wissen*. Und das spezifische Wissen, das man in einer bestimmten Situation benötigt um beispielsweise ein Problem zu lösen, wird *In-*

formation genannt. Bevor wir uns aber mit der Informationssuche beschäftigen können, fassen wir kurz zusammen, wie man via Internet auf Daten zugreifen kann.

Beginnen wir mit einem kurzen Überblick über die wichtigsten Internet-Eigenschaften. Das Internet ist ein Zusammenschluss von Teilnetzen auf der ganzen Welt. Darum wird es auch das Netz der Netze genannt. Am Internet beteiligen sich Millionen von Servern, und jeder Server bietet einen oder mehrere Dienste wie WWW, E-Mail oder News an. Die Server sind grundsätzlich autonom, können also fast nach Belieben schalten und walten und ihre Angebote selbst bestimmen. Das Internet unterliegt an sich keiner zentralen Verwaltung. Allerdings sollten sich alle am Internet teilnehmenden Rechner an gewisse Standards halten, die von zentralen Stellen wie zum Beispiel dem World Wide Web Consortium empfohlen werden.

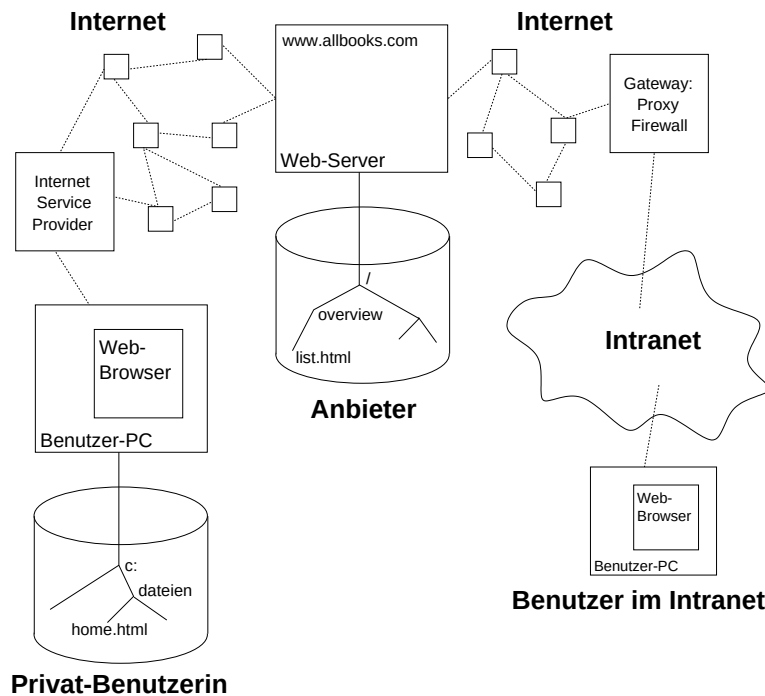


Die unzähligen Netzverbindungen sorgen dafür, dass die Rechner im Internet sich miteinander unterhalten können. Zwischen zwei Rechnern im Netz sind üblicherweise verschiedene Wege möglich. Das Internet sucht sich automatisch eine mögliche Verbindung.

Für einen Benutzer, der auf das Internet zugreifen möchte, sind vor allem vier Punkte wichtig: (1) Bei langen Leitungen dauert die Übertragung von Daten länger als bei kurzen. (2) Dicke Leitungen können mehr Daten pro Zeiteinheit transportieren als dünne. (3) Leistungsfähigere Server antworten rascher als weniger leistungsfähige.

(4) Je grösser die Menge der angeforderten Daten ist, desto länger dauert die Übertragung. Alle vier Punkte beeinflussen die Wartezeit beim Zugriff auf Daten im Internet.

Wie kann nun ein Benutzer konkret auf Daten im Internet zugreifen? Voraussetzung ist natürlich ein passend ausgerüsteter Computer, zum Beispiel ein Laptop. Auf dem Laptop ist ein Web-Browser aufgestartet. Der Web-Browser ist ein Werkzeug zur Ansicht von Daten im Internet und speziell im World Wide Web. Der Browser hat in erster Linie die Aufgabe, Webseiten auf dem Bildschirm darzustellen. Eine Webseite kann beispielsweise lokal in der Datei `c:\dateien\home.html` gespeichert sein.



Interessanter wird es aber, wenn Webseiten von einem anderen Rechner, einem so genannten Web-Server, stammen. Auf der Anbieterseite steht ein Web-Server, der bestimmte Angebote zur Verfügung stellt. Ein solches Angebot könnte ein fiktiver Bücherkatalog namens AllBooks sein. Der Web-Server hat eine eindeutige Adresse

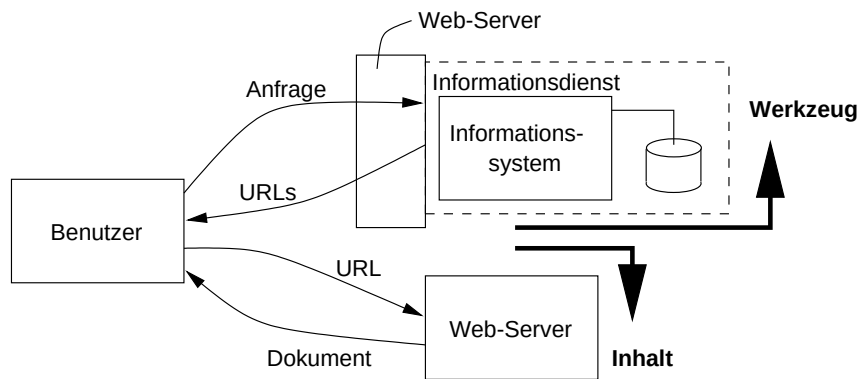
wie zum Beispiel `www.allbooks.com`. Alle vom Web-Server angebotenen Webseiten liegen in einem Verzeichnisbaum bereit und sind über einen *Uniform Resource Locator* (URL) eindeutig ansprechbar. Der URL `http://www.allbooks.com/overview/list.html` identifiziert bei unserem Server die Bücherliste `list.html` im Übersichtsverzeichnis namens `overview`. Der Verzeichnisbaum mit den Webseiten ist frei wählbar, muss aber dem Web-Server bekannt gemacht werden. Alle zusammengehörenden Webseiten auf einem Web-Server bezeichnen wir als eine *Web-Site*. Zum Beispiel besteht die Web-Site von AllBooks aus allen Seiten, die auf dem Server `www.allbooks.com` liegen.

Die Aufgabe des Internets besteht darin, die Kommunikation zwischen Anbieter- und Benutzerseite zu gewährleisten. Der Benutzer schickt den URL der gewünschten Webseite mit Hilfe des Browsers und via Internet zum entsprechenden Web-Server. Anschliessend transportieren die Rechner im Internet die Antwort vom Server zurück zum Benutzerrechner. Die Verbindung zwischen Benutzer-PC und Internet läuft über eine analoge Telefonverbindung, eine ISDN-Leitung, das Kabelnetz, Mobilfunk oder eine andere Kommunikationstechnologie – es gibt zahlreiche Varianten. Bei einem Firmen-Intranet ist zwischen Benutzer-PC und Internet häufig ein Firewall-Rechner zwischengeschaltet, der unerlaubte Zugriffe von aussen verhindert.

Normalerweise sind Webseiten so genannte *statische Seiten*, die permanent unter einem eigenen URL ansprechbar sind. Daneben gibt es auch *dynamische Seiten*, die zum Zeitpunkt des Zugriffs erstellt werden und nur gerade für diesen einen Zugriff existieren. Ein möglicher Weg zum Abrufen beziehungsweise Erstellen von dynamischen Seiten besteht in so genannten CGI-Programmen, die sich häufig schon aufgrund des URL zu erkennen geben: beispielsweise `http://www.clock.nz/cgi-bin/germantime.pl`. Hier liefert ein Programm namens `germantime.pl` im CGI-Verzeichnis des Servers `www.clock.nz` die aktuelle Zeit in Deutschland. Neben den CGI-Programmen gibt es auch andere Methoden zur Erstellung von dynamischen Webseiten.

Informationsdienste

Um mit der Datenflut im Internet fertig zu werden, braucht es Hilfsmittel, die uns bei der Informationsbeschaffung zur Seite stehen. Nicht immer kennt man gerade die richtige Adresse einer Webseite, welche die gewünschte Information enthält; und zielloses Surfen ist in den wenigsten Fällen erfolgreich. Unter all den Rechnern im Internet gibt es Server, die ein Angebot speziell für die Informationsbeschaffung zur Verfügung stellen. Solche Werkzeuge für den Informationszugriff nennen wir *Informationsdienste*. Der Zugriff auf einen Informationsdienst erfolgt häufig in folgenden zwei Schritten:

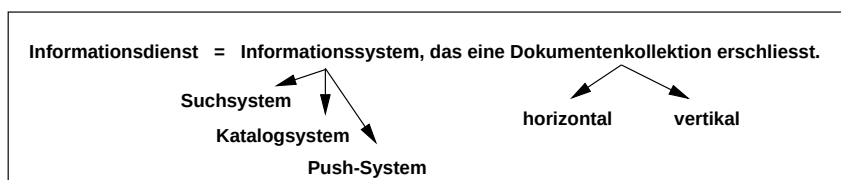


1. Schritt: Ein Benutzer mit einem Informationsbedürfnis konsultiert den Informationsdienst, der sich typischerweise hinter einem Web-Server «versteckt». Als Antwort erhält der Benutzer einen oder mehrere Verweise auf Dokumente, die hoffentlich brauchbare Informationen bereithalten.

2. Schritt: Erst jetzt greift der Benutzer auf den tatsächlichen Inhalt der gefundenen Webseiten zu. Die Webseiten sind nicht beim Informationsdienst, sondern bei normalen Web-Servern gespeichert.

Ein Informationsdienst ist ein *Informationssystem*, das eine *Dokumentenkollektion* erschliesst. Mit Erschliessen von Dokumenten ist das Registrieren aufgrund von in den Dokumenten vorkommenden Begriffen (Wörter, Namen, Zahlen usw.) gemeint. Das Informationssystem ist das eigentliche Werkzeug, ein Softwaresystem, das die

Funktionen für den Informationszugriff bereitstellt. Die Dokumentenkollektion steht unabhängig von einem Informationssystem zur Verfügung und liefert die *Inhalte* zum Informationsdienst.



Die klare Trennung zwischen dem Informationssystem und der damit erschlossenen Dokumentenkollektion ist wichtig! Ein Informationssystem ohne Kollektion ist wie ein Buch, in dem alle Seiten ausser dem Inhaltsverzeichnis und dem Stichwortverzeichnis entfernt wurden. Oder umgekehrt: Eine Kollektion ohne Werkzeug für den Zugriff ist wie eine grosse Bibliothek mit fünf Millionen Büchern auf einem riesigen Haufen, ohne jede Katalogisierung.

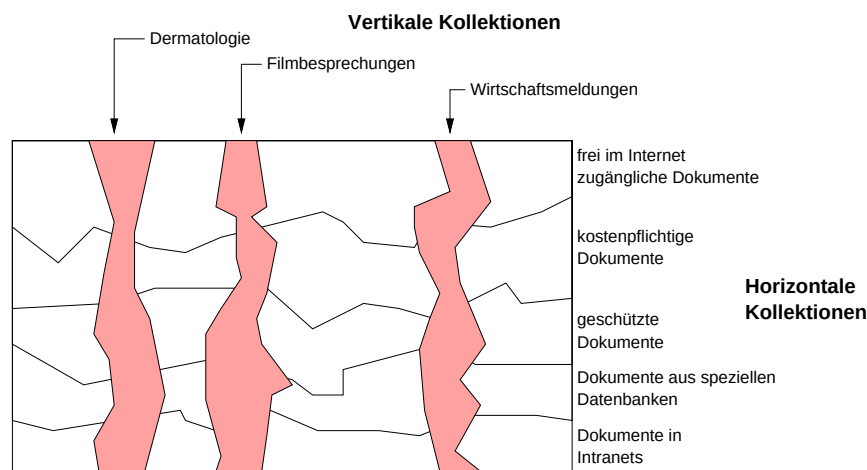
Dokumentenkollektionen

Eine Dokumentenkollektion ist eine Menge von Dokumenten (zum Beispiel Webseiten), die von einem Informationssystem erschlossen werden. Wir beschränken uns im Folgenden auf Textdokumente wie beispielsweise den Text der Schweizerischen Bundesverfassung oder das Bücherverzeichnis mit dem URL <http://www.allbooks.com/overview/list.html>. Grundsätzlich können Dokumente aber auch völlig anderer Natur sein: zum Beispiel eine digitalisierte Kopie von Edvard Munchs «Der Schrei» im GIF-Format oder das Lied «Aqualung» der Band Jethro Tull in einem Audio-Format.

Wir unterscheiden zwei Grundtypen von Dokumentenkollektionen: Eine *horizontale Dokumentenkollektion* versucht möglichst viele der im Internet zugänglichen Dokumente zu erschliessen und die verschiedenen Gebiete in der ganzen Breite abzudecken. Jedes nur denkbare Thema kann auftauchen – vom Programmcode eines vollständigen Betriebssystems bis zu Tipps für Sammler von Kinderüberraschungseiern.

Eine *vertikale Dokumentenkollektion* dagegen konzentriert sich auf Dokumente aus einem mehr oder weniger eng umrissenen Themenbereich, der in seiner ganzen Tiefe erschlossen werden soll.

Horizontale Dokumentenkollektionen decken typischerweise die oberste Schicht der frei zugänglichen Dokumente ab, seltener auch kostenpflichtige Dokumente. Demgegenüber können in vertikalen Dokumentenkollektionen durchaus auch Dokumente aus gezielt ausgewählten Web-Sites, speziellen Datenbanken oder aus einem durch einen Firewall-Rechner geschützten Intranet auftauchen.

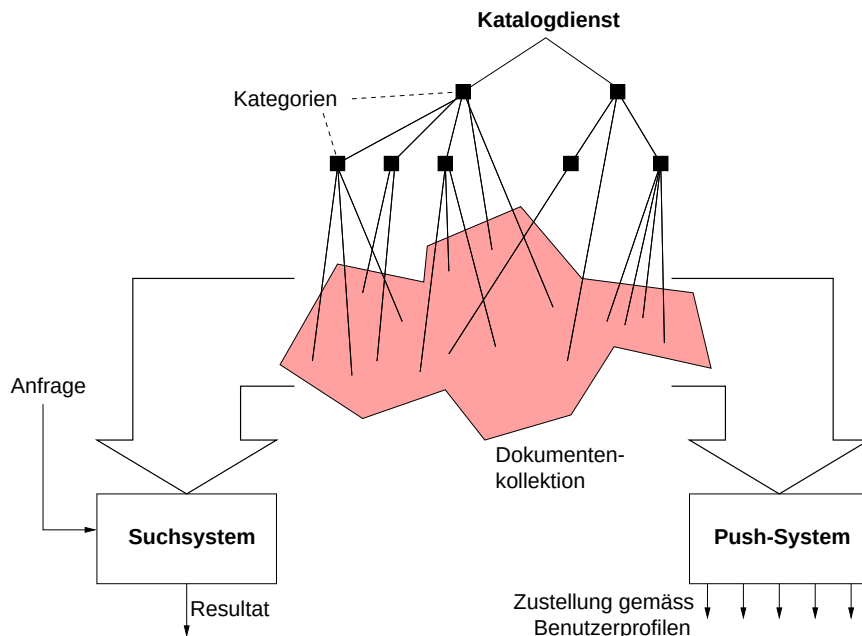


Informationssysteme

Kombiniert man eine Dokumentenkollektion mit einem Informationssystem, so erhält man einen Informationsdienst, der Benutzerinnen beim Informationszugriff unterstützt. Wir behandeln die drei wichtigsten Typen von Informationssystemen:

Suchsysteme (auch Information-Retrieval-Systeme oder Suchmaschinen genannt) stellen Funktionen zur Dokumentensuche bereit. Die Anfrage einer Medizinerin könnte beispielsweise lauten: *neue Methoden zur Behandlung der Sichelzellenanämie*. Daraufhin liefert das Suchsystem eine Liste von Dokumenten, welche die gewünschte Information enthalten sollen.

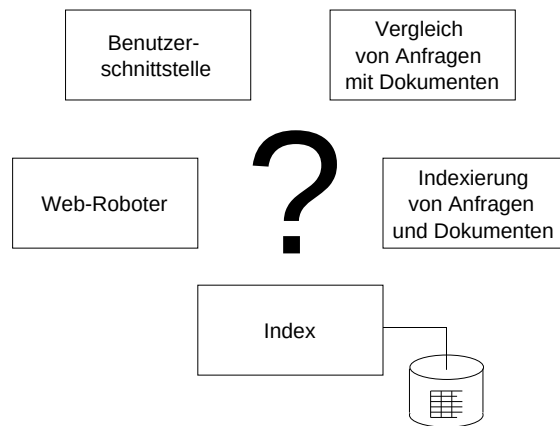
Katalogsysteme kann man mit einem Branchetelefonbuch vergleichen. Die einzelnen Einträge sind in vorgegebene *Kategorien* eingeteilt. Die Kategorien sind häufig hierarchisch geordnet. In der Kategorie *Medizin*, in der Unterkategorie *Anämie* könnten Dokumente zur Sichelzellenanämie eingereiht sein.



Die *Push-Systeme* bieten einen anderen Service an. Während Benutzer von Such- oder Katalogsystemen bei jedem Informationsbedürfnis die Dokumente beim Informationssystem holen müssen (man spricht auch von «Pull-Systemen»), versorgen Push-Systeme ihre Benutzer *aktiv* und fortlaufend mit aktuellen Informationen. Für jeden Benutzer existiert ein so genanntes *Profil*. Das Profil beschreibt das Informationsbedürfnis eines Benutzers, so wie die Anfrage bei einem Suchsystem. Neu im Internet auftauchende Dokumente werden automatisch mit allen gespeicherten Profilen verglichen und bei genügend guter Übereinstimmung dem Benutzer zugestellt. Mittels eines Push-Systems könnte sich die Medizinerin laufend über neue Methoden zur Behandlung der Sichelzellenanämie orientieren lassen.

Puzzle-Steine eines Informationssystems

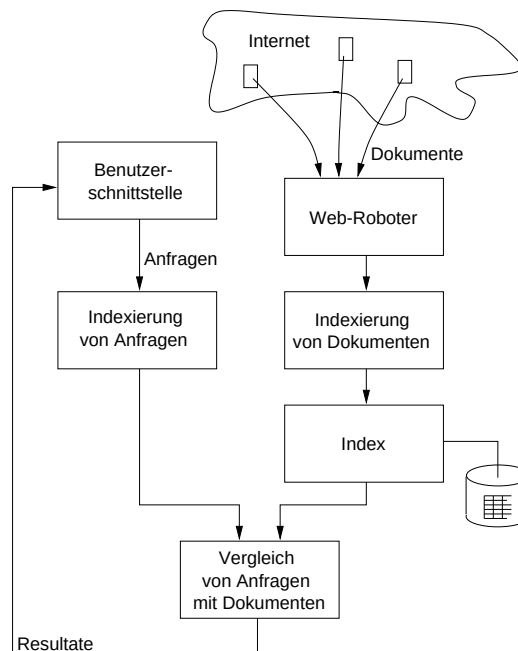
Für den effektiven und effizienten Informationszugriff ist es wichtig, die Funktionsweise des verwendeten Werkzeugs zu kennen. Im Grunde bestehen die meisten Informationssysteme aus fünf wichtigen Komponenten, die wir einzeln und grösstenteils voneinander losgelöst betrachten können. Je nachdem wie die Komponenten dann kombiniert werden, entsteht ein Such-, ein Katalog- oder ein Push-System.



- Die *Benutzerschnittstelle* ist das Einzige, was ein Benutzer vom Informationsdienst zu sehen bekommt. Sie ermöglicht die Kommunikation zwischen Dienst und Anwendern.
- Der *Web-Roboter* besucht regelmässig den ihm zugewiesenen Teil des WWW auf der Suche nach neuen oder geänderten Dokumenten, die er dann dem Informationsdienst zuspielen kann.
- Während der *Indexierung* werden Dokumente oder auch Anfragen untersucht und darin vorkommende Begriffe identifiziert. Das Vorgehen ist also ähnlich wie in einem Sachbuch: Nachdem der Buchinhalt geschrieben ist, identifiziert man im Text die wichtigen Begriffe und fügt diese im Stichwortverzeichnis des Buchs ein. Zu jedem Eintrag im Stichwortverzeichnis wird die entsprechende Seite vermerkt. Eine vergleichbare Funktion übernimmt der Index des Informationsdienstes.

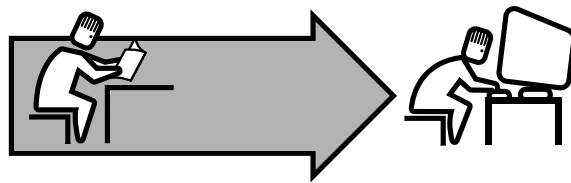
- Der *Index* ist eine Datenstruktur, die das schnelle Vergleichen von Anfragen mit Dokumenten erlaubt. Ein Suchsystem benutzt den Index, um zu einem Begriff alle Webseiten zu finden, in denen das Wort auftaucht.
- Beim *Vergleich von Anfragen mit Dokumenten* wird ermittelt, inwiefern ein Dokument eine Anfrage befriedigt.

Als Beispiel sind in der unten stehenden Grafik die fünf Komponenten zu einem Suchsystem zusammengestellt. Ein kurzer Überblick über das Geschehen: Der Web-Roboter durchpflügt das Internet. Alle gefundenen Dokumente werden zunächst indexiert und anschliessend im Index des Suchsystems abgelegt. Sobald eine Benutzerin mit Hilfe der Benutzerschnittstelle eine Anfrage an das System schickt, wird die Anfrage in einem ersten Schritt ebenfalls indexiert. Daraufhin wird die Anfrage mit den im Index in geeigneter Form gespeicherten Dokumenten verglichen. Zum Schluss erhält die Benutzerin eine Liste mit Verweisen auf die gefundenen Dokumente.



Viel mehr soll an dieser Stelle noch nicht verraten werden. Wir werden in den folgenden Kapiteln die einzelnen Komponenten genauer unter die Lupe nehmen. Dabei konzentrieren wir uns vorerst auf die Suchdienste. Später werden wir die Komponenten auf andere Weise kombinieren und uns mit Katalog- und Push-Diensten auseinander setzen.

Schauen wir nun, was es aus praktischer Sicht zu berichten gibt ...



Zwei fiktive Dienste

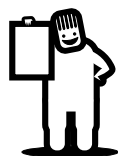
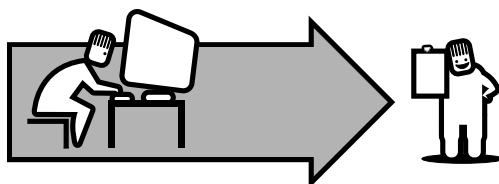
Für die Beispiele haben wir zwei Informationsdienste ausgewählt: Der Suchdienst OMNISEARCH bietet eine horizontale Dokumentensammlung an und möchte damit eine möglichst grosse Zahl von Leuten ansprechen, denn OMNISEARCH finanziert sich ausschliesslich durch Werbung, die auf den Webseiten angezeigt wird. Der Suchdienst NEWSSEEKER erschliesst eine vertikale Dokumentensammlung, die Nachrichtenmeldungen aus aller Welt umfasst.

Im Internet kann man sich selten darauf verlassen, dass etwas morgen noch so funktioniert, wie man es heute erlebt hat. Suchdienste im Internet können sich von Tag zu Tag ändern. Zudem ist die Zahl der verfügbaren Angebote unüberschaubar. Laufend werden neue Systeme angepriesen, andere verschwinden ohne Vorwarnung. Und um alles noch schlimmer zu machen, haben Anwender völlig unterschiedliche Vorlieben und Abneigungen. Aus all diesen Gründen haben wir kurzerhand zwei fiktive Suchdienste ins Leben gerufen. Im Internet gibt es folglich keinen Dienst mit dem Namen OMNISEARCH oder NEWSSEEKER. Das erlaubt uns, in erster Linie auf die

Grundlagen einzugehen, anstatt Handbücher zur Benutzung von AltaVista, Excite, Infoseek, NorthernLight, Google, Euroseek, Fireball, Goto, HotBot usw. abzufragen. Stattdessen sind im Online-Teil unter <http://www.internet-kompetenz.ch/> Hinweise vermerkt, wie man den Schritt von den fiktiven zu den echten Suchdiensten schafft. Dort sind auch Verzeichnisse mit den unterschiedlichsten Suchdiensten aufgeführt.

Übrigens verzichten wir darauf, zusätzlich zu den Suchdiensten OMNISEARCH und NEWSSEEKER auch erfundene Push- oder Katalogdienste einzuführen. Der Grund: Die meisten Beispiele in den folgenden Kapiteln arbeiten mit Suchdiensten. Katalog- und Push-Dienste werden vornehmlich in den entsprechenden Kapiteln behandelt und stützen sich auf dieselben grundlegenden Prinzipien, die auch für Suchdienste gelten.

Zum Abschluss des Kapitels werden die Antworten zu den Fragen aus der Kapiteleinführung präsentiert ...

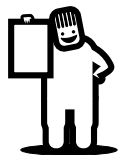


Was muss ich bei der Informationsbeschaffung unbedingt beachten? Primär muss ich die richtige Dokumentensammlung wählen. Horizontale Sammlungen sind sicher oft ein guter Ausgangspunkt für eine Suche. Sollte sich aber für ein bestimmtes Thema eine vertikale Sammlung im entsprechenden Gebiet finden lassen, umso besser. Das ist wie bei Buchhandlungen: Suche ich nach Literatur zur Aufzucht von Bonsais, so werde ich in einer auf Gartenpflanzen spezialisierten Buchhandlung schneller fündig als in einem grossen Buchladen, der alle Themen abdecken will.

Wichtig ist aber auch die Wahl des Werkzeugs. Im Internet gibt es die unterschiedlichsten Angebote, die mich beim Informationszugriff

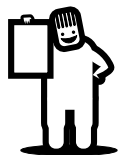
unterstützen. Der Suchdienst OMNISEARCH ist nur eine Möglichkeit. Je nach Fragestellung ist vielleicht ein Katalogdienst oder ein Push-Dienst besser geeignet.

Nun habe ich immer noch das Problem mit all den Käfigen bei meiner Suche nach John Cage. Ich bin an allgemeinen Informationen zu einem bekannten Komponisten interessiert. Unter diesen Voraussetzungen sollte ein Katalogdienst eine passende Kategorie anbieten. Und tatsächlich werde ich in der Kategorie *Entertainment / Music / Genres / Classical / Composers / Modern / Cage, John* fündig. Dort stosse ich auf eine Sammlung von Biografien und Kommentaren zu seinem Werk. In dieser Kategorie werde ich zum Glück von allen Vogelkäfigen und auch von allen anderen Personen namens Cage verschont.



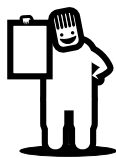
Es ist nicht verwunderlich, dass ich in der horizontalen Dokumentensammlung von OMNISEARCH Probleme hatte mit meiner Anfrage zur Tandem Computers Inc. Mit NEWSSEEKER steht mir ein Suchdienst mit einer vertikalen Kollektion im Bereich von Nachrichtenmeldungen zur Verfügung. Dieser Dienst eignet sich wunderbar für meine Frage, denn die Meldung über den Kauf von Tandem Computers Inc. wird sicherlich auftauchen. Dafür werde ich kaum mit all den Fahrradhändlern belästigt. Problemlos finde ich heraus, dass die Firma von Compaq Computers Corp. aufgekauft wurde.

Zu den technischen Problemen: Eines der Dokumente wurde mir gar nicht angezeigt – bei einem anderen musste ich sehr lange warten. Schuld ist hier nicht der Suchdienst. Vielleicht wurde der Web-Server mit dem verlangten Dokument vorübergehend vom Netz genommen, oder die Verbindung zum Web-Server ist ausgefallen oder momentan überlastet.



OMNISEARCH hat zum Suchbegriff *Achtlinge* kein einziges Dokument gefunden. Offenbar existiert in der ganzen horizontalen Kollektion von OMNISEARCH nicht ein Dokument mit dem Begriff. Eigentlich sollte man aber denken, dass es zu diesem Thema etliche Pressemeldungen gab. Und weil NEWSSEEKER eine vertikale Kollektion im Bereich von internationalen Nachrichtenmeldungen anbietet, wiederhole ich dort

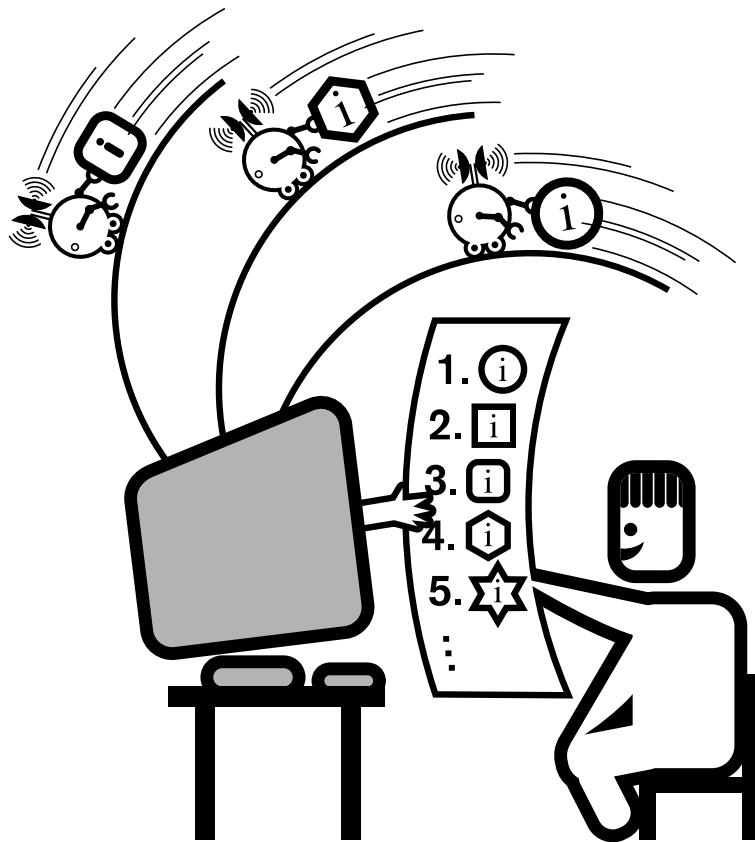
meine Anfrage. Tatsächlich: Im Dezember 1998 gebar Nkem Chukwu Achtlinge im US-Bundesstaat Texas.



Ich möchte fortlaufend über die aktuellsten Computerviren informiert werden. Katalog- und Suchdienste eignen sich in erster Linie für eine einmalige Anfrage. Für mein Problem ist ein Push-Dienst die richtige Lösung, denn er stellt mir nach meinen Wünschen regelmässig neu gefundene Dokumente zu einem Thema zu. Ich mache mich also auf die Suche nach einem geeigneten Push-Dienst im Bereich von Computer und Informatik. Der Push-Dienst soll mir ab sofort alle Dokumente zum Thema Viren schicken.

Kapitel 2

Moderne Suchmethoden





Tag für Tag benutze ich im Büro die kleinen, nützlichen Helfer, die meine losen Blätter zusammenhalten. Die normalerweise vernickelten Drahtgebilde heissen «Büroklammern». Eines Tages fragte ich mich, wer diese unscheinbaren und doch fast unverzichtbaren Werkzeuge erfunden hat.

Also: Auf ins Internet, da wird sich die Antwort wohl finden lassen! Ich benutze OMNISEARCH mit der Anfrage *Büroklammer*. Leider ist die Antwort des Suchdienstes nicht befriedigend. Ich finde Berichte über Büroausstattungsfirmen, die Büroklammern im Sortiment haben. Andere Webseiten enthalten Tipps für Bewerbungen, die nicht einfach mit Büroklammern geheftet werden sollen.

Ich bin enttäuscht. Wieso liefert mir OMNISEARCH nicht die von mir gesuchte Information? Wie wählt ein Suchsystem überhaupt die Dokumente aus, die sie mir präsentiert?



Eines meiner absoluten Lieblingsbücher ist Michael Endes «Die unendliche Geschichte». Ich nehme an, im Internet gibt es zahlreiche Gleichgesinnte, die Webseiten zu diesem Buch veröffentlicht haben. Kurzerhand mache ich mich auf die Suche mit der Anfrage *unendliche Geschichte*. OMNI-

SEARCH versorgt mich mit Tausenden von Hinweisen auf Dokumente, von denen keines viel versprechend aussieht. Meistens geht es um unendliche Geschichten in einem anderen Sinn. Zum Beispiel wird auf einer Webseite mit der besonders langen Lebensdauer von Eternit-Dachplatten geworben. An anderer Stelle geht es um den Recyclingkreislauf, der ebenfalls wie eine unendliche Geschichte anmutet. Was soll ich tun, damit ich *meine* Geschichte finde?

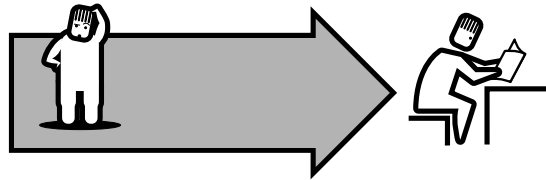
Ich frage mich, wie gross der Nutzen der Suchdienste im Netz überhaupt ist. Auf viele meiner Anfragen werden Millionen von Dokumenten gemeldet. Wie kann ich eine so riesige Menge jemals überblicken und die von mir gesuchte Information finden?



Ich arbeite an einem Bericht zur Geschichte der siebenköpfigen Exekutive der Schweizer Regierung – «Bundesrat» genannt. Für meine Arbeit brauche ich eine Übersicht über alle Bundesräte seit der Gründung des Bundesstaats im Jahre 1848. Ich versuche die Anfrage *Schweiz Bundesrat vollständige Liste*. Das Resultat ist ernüchternd. Ich finde viele Do-

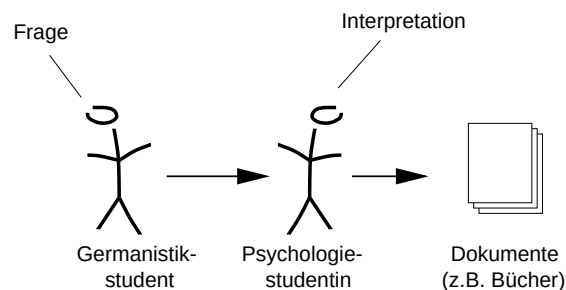
kumente zu den Schweizer Bundesräten, doch nirgends die gesuchte Gesamtübersicht.

Die Probleme drehen sich in erster Linie darum, dass man mit einer Anfrage keine brauchbaren oder viel zu viele Dokumente erhält. Zur Lösung der aufgeworfenen Fragen müssen wir uns zunächst mit den Schwierigkeiten bei der Informationssuche auseinander setzen ...

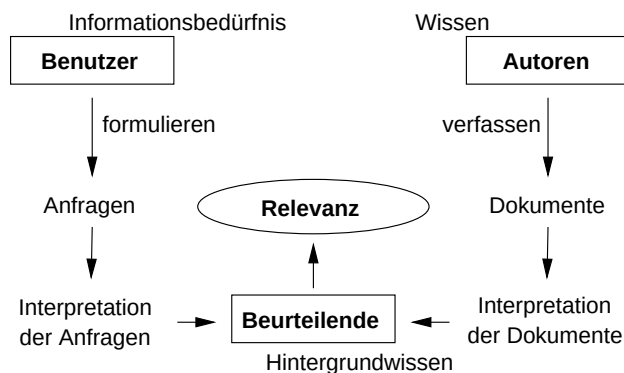


Probleme bei der Informationssuche

Der Begriff *Information Retrieval* – Informationssuche – ist in zweierlei Hinsicht irreführend. Erstens geht es nicht nur um das *Wiederfinden* von Informationen, sondern auch um das *Finden* von Informationen, die man noch nie zuvor gesehen hat. Zweitens erhält man nicht *direkt* die gewünschte Information, sondern *Dokumente*, welche die gesuchte Information hoffentlich enthalten. Eine wichtige Voraussetzung für eine erfolgreiche Informationssuche ist also, dass die Dokumentensammlung Dokumente mit der benötigten Information enthält. Das heisst, irgendjemand muss ein Dokument mit diesen Informationen verfasst und veröffentlicht haben.



Ein erstes Beispiel: Ein Germanistikstudent belegt das Nebenfach Psychologie und muss sich im Rahmen einer Seminararbeit über den Freud'schen Ansatz der Psychoanalyse informieren. Leider kennt der Student die nötigen Quellen zu diesem Thema nicht. Deshalb besucht er eine befreundete Psychologiestudentin und stellt eine entsprechende Frage – zum Beispiel: «Ich muss mich über den Freud'schen Ansatz der Psychoanalyse informieren. Kannst du mir weiterhelfen?» Anschliessend interpretiert die Psychologiestudentin diese Frage und identifiziert die für sie entscheidenden Begriffe. In diesem Fall sind das die Begriffe «Freud'scher Ansatz» und «Psychoanalyse». Den Rest der Frage benötigt sie nicht, um das Informationsbedürfnis ihres Kollegen zu bestimmen. Nun greift die Studentin auf ihre private Büchersammlung zurück und trägt eine Sammlung von Büchern zusammen, die aufgrund des Inhalts mit der Frage in Zusammenhang stehen. Dabei profitiert die Studentin von ihrem umfangreichen Hintergrundwissen im Bereich der Psychologie sowie von ihrer Kenntnis der Büchersammlung. Wichtig: Die Studentin beantwortet *nicht direkt* die Anfrage des Studenten, sondern vergleicht den Inhalt einiger Bücher mit der Anfrage und wählt dann die (hoffentlich) geeigneten Bücher aus.



Mit obigem Beispiel sind wir schon sehr nahe am grundsätzlichen Problem bei der Informationssuche. Auf der einen Seite steht der Benutzer, in unserem Beispiel der Germanistikstudent. Er hat ein Informationsbedürfnis, welches er in einer Anfrage formuliert. Diese Anfrage besteht aus einer Menge von Suchbegriffen. Auf der anderen

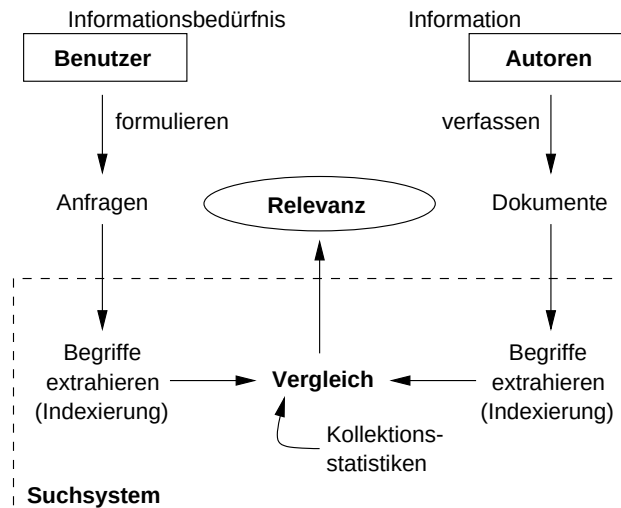
Seite befinden sich die Autoren. Sie verfügen über Informationen und verfassen Dokumente, die auf diesen Informationen basieren.

Zwischen den Autoren und den Benutzern steht eine beurteilende Person. Häufig sind die Beurteilende und die Benutzerin dieselbe Person (im Beispiel war es die Psychologiestudentin); das muss aber nicht unbedingt so sein. In Bibliotheken beispielsweise steht den Benutzern oft eine Bibliothekarin mit Rat und Tat zur Seite. Die Bibliothekarin wird so zum Bindeglied zwischen Autoren und Benutzern. In einem ersten Schritt interpretiert die Beurteilende sowohl die verfügbaren Dokumente (beispielsweise die Bücher in der Bibliothek) als auch die Anfrage. Aufgrund dieser Interpretationen bestimmt die Beurteilende dann, ob und wie stark ein gewisses Dokument relevant für die gestellte Anfrage ist. Die Beurteilung der Relevanz wird stark beeinflusst durch das Hintergrundwissen der Beurteilenden sowie durch die Fähigkeit, das Hintergrundwissen mit dem im Dokument dargestellten Wissen zu verknüpfen.

Das folgende Beispiel erläutert die Rolle des Hintergrundwissens: Ein Schüler stellt erste Nachforschungen für einen Geografievortrag zum Thema «Wo leben Eskimos?» an. Er findet ein Dokument mit dem Inhalt: «Die Inuit bewohnen die nördlichsten Gebiete der USA und Kanadas sowie Grönland und Teile Sibiriens.» Das Dokument ist für den Schüler nur relevant, falls ihm bereits bewusst ist, dass sich Eskimos selber als Inuit bezeichnen. Andernfalls wird er das Dokument wahrscheinlich ignorieren.

Das Suchsystem kommt ins Spiel

Die Beurteilung der Relevanz von Dokumenten soll nun automatisiert werden. Das heisst, das Suchsystem muss die Rolle der Beurteilenden übernehmen und zu einer Anfrage die «passenden» – das heisst die relevanten – Dokumente finden. Anstatt Dokumente und Anfragen zu *interpretieren*, werden lediglich Begriffe im Text identifiziert, gezählt und die Zahlen miteinander verglichen. Gleichzeitig kann auf statistische Angaben über die ganze Dokumentensammlung – auf das «Hintergrundwissen» des Suchsystems – zurückgegriffen werden. Das Resultat des Vergleichs ist ein Relevanzwert, der die Relevanz eines Dokuments gegenüber der Anfrage beschreibt.



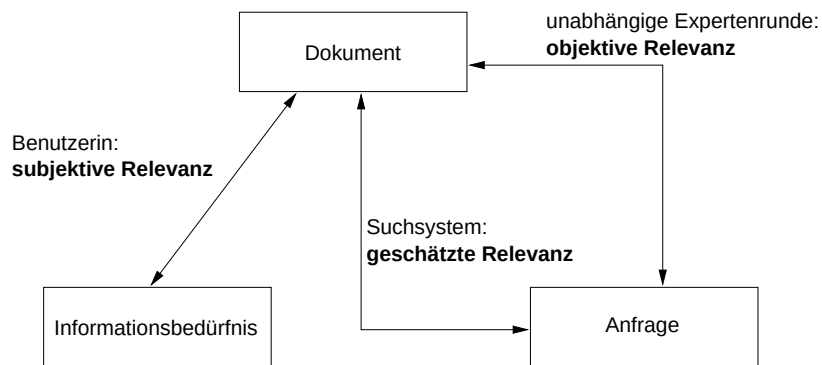
Die Informationssuche ist trotz der Automatisierung von den beteiligten Personen abhängig. Anfragen und Dokumente werden verfasst. Dabei spielen persönliche Vorlieben bezüglich des Schreibstils eine Rolle. Verschiedene Personen beschreiben den gleichen Sachverhalt mit verschiedenen Worten. Das Suchsystem kann diese Wortwahl nicht beeinflussen. Aus diesem Grund lässt sich feststellen: Das perfekte Suchsystem wird es nie geben, weil an mehreren Stellen menschliche Einflüsse eine Rolle spielen.

Relevanz eines Dokuments

Der Begriff der Relevanz ist schon oft vorgekommen, doch fehlt nach wie vor die konkrete Definition. Das soll nun nachgeholt werden: Ein Dokument heisst relevant, wenn es bei einer Informationssuche gefunden werden soll.

Drei unterschiedliche Arten von Relevanz müssen unterschieden werden: *Subjektive Relevanz* ist eine Beziehung zwischen einem Dokument und einem Informationsbedürfnis. Üblicherweise beurteilt eine Benutzerin die subjektive Relevanz eines Dokuments. Für die Informationssuche formuliert die Benutzerin ihr Informationsbedürfnis in einer Anfrage. Leider können die Suchbegriffe in einer Anfrage das

Informationsbedürfnis nur selten perfekt repräsentieren. Die *objektive Relevanz* ist eine Beziehung zwischen einer Anfrage und einem Dokument. Die objektive Relevanz kann zum Beispiel von einer Gruppe unabhängiger Experten auf einem Gebiet bestimmt werden. Ein Suchsystem bestimmt eine dritte Art von Relevanz: Die *geschätzte Relevanz* ist ebenfalls eine Beziehung zwischen einer Anfrage und einem Dokument. Die Relevanz wird in diesem Fall nach bestimmten Regeln berechnet. Bei der Informationssuche kommt es häufig vor, dass die drei Arten der Relevanz nicht übereinstimmen.



Beispiel: Für einen Kurzvortrag sucht eine Schülerin nach einigen Eckdaten zum Vulkan Stromboli in Italien. Die Schülerin wählt einen Suchdienst – zum Beispiel OMNISEARCH – und stellt eine entsprechende Anfrage. OMNISEARCH liefert eine Liste mit Dokumenten, wovon drei herausstechen: Bei einem der Dokumente handelt es sich um die Werbung eines Reisebüros, das Ausflüge zum Vulkan organisiert. OMNISEARCH schätzt das Dokument als relevant zur Anfrage ein (geschätzte Relevanz), obwohl es für die Schülerin nicht relevant ist (keine subjektive Relevanz). Ein anderes Dokument liefert eine hundertseitige Abhandlung mit allen erdenklichen Details zum Vulkan. Eine unabhängige Expertenrunde würde ein solches Dokument sicherlich als relevant zur Anfrage erachten (objektive Relevanz). Doch der Schülerin ist es eindeutig zu ausführlich. Stattdessen wählt die Schülerin ein drittes Dokument, in dem auf einer halben Seite die wichtigsten Angaben zum Vulkan Stromboli zusammengefasst sind (subjektive Relevanz).

Unterschiede zwischen der subjektiven Relevanz und der von einem Suchsystem geschätzten Relevanz treten sehr häufig bei mehrdeutigen Begriffen auf. Beispiel: Für das Jubiläumsbuch eines Tennisclubs wird nach den Biografien von ehemaligen Grössen im Weltennis gesucht, etwa von Yannick Noah. Der Benutzer stellt die simple Anfrage *Noah* an OMNISEARCH. Kurz darauf antwortet OMNISEARCH mit einer Menge von Dokumenten. Viele davon behandeln den Tennisspieler. Viele andere befassen sich mit der Geschichte um Noahs Arche. Für den Benutzer sind die Arche-Noah-Dokumente nicht von Interesse, sie befriedigen sein Informationsbedürfnis nicht und sind folglich subjektiv nicht relevant. Trotzdem kann man dem Suchsystem nichts vorwerfen. Die Anfrage *Noah* steht zweifelsohne sowohl mit den Tennis-Dokumenten als auch mit den Arche-Noah-Dokumenten in Zusammenhang.

Gewichtung von Dokumenten nach Relevanz

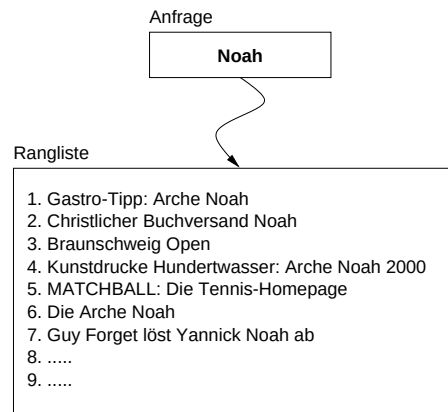
Wir haben einige Probleme bei der Informationssuche kennen gelernt: (1) Häufig ist das Informationsbedürfnis aus der Anfrage eines Benutzers nicht klar ersichtlich. (2) Die Relevanz eines Dokuments bezüglich einer Anfrage ist schwierig zu beurteilen. Was das Suchsystem als relevant betrachtet, kann für den Benutzer subjektiv irrelevant sein. (3) Fehlendes oder nicht aktiviertes Hintergrundwissen erschwert die Informationssuche zusätzlich.

Wir haben festgestellt, dass für die erwähnten Probleme keine perfekte Lösung existiert. Stattdessen wird mit einer Annäherung gearbeitet. Das Zauberwort heisst *Relevance Ranking*. Damit meint man das Anordnen von Dokumenten gemäss absteigender Relevanz bezüglich einer Anfrage. Relevance Ranking läuft in zwei Schritten ab:

Erster Schritt: Nachdem die Benutzerin eine Suchanfrage gestellt hat, werden alle verfügbaren Dokumente mit der Anfrage verglichen. Bei diesem Vergleich weist das System jedem Dokument einen Relevanzwert zu. Mit der Höhe des Relevanzwertes drückt das System die geschätzte Relevanz der jeweiligen Dokumente in Bezug auf die Anfrage aus. Je höher der Relevanzwert ausfällt, desto wahrscheinli-

cher stuft das Suchsystem ein Dokument bezüglich der Anfrage als relevant ein.

Zweiter Schritt: Nun sortiert das Suchsystem die Dokumente aufgrund des Relevanzwertes in absteigender Reihenfolge. Die so entstehende geordnete Liste wird *Rangliste* genannt. Die Rangliste wird der Benutzerin präsentiert, die je nach ihrem Bedürfnis wenige oder viele Dokumente daraus auswählt und genauer betrachtet.



Beispiel: Die Rangliste zur Noah-Recherche könnte die verschiedensten Dokumente enthalten: Einen Tipp aus einem Gastronomieführer, einen Hinweis auf ein Bild von Friedensreich Hundertwasser, religiöse Dokumente, eine Tennis-Homepage und Meldungen, in denen der Name des Tennisspielers auftaucht.

In der Rangliste werden alle Dokumente aufgeführt, welche das Suchsystem als relevant bezüglich der Anfrage erachtet. Auf Platz 1 steht das Dokument mit dem höchsten Relevanzwert, gefolgt von den übrigen Dokumenten, nach Relevanzwerten absteigend geordnet.

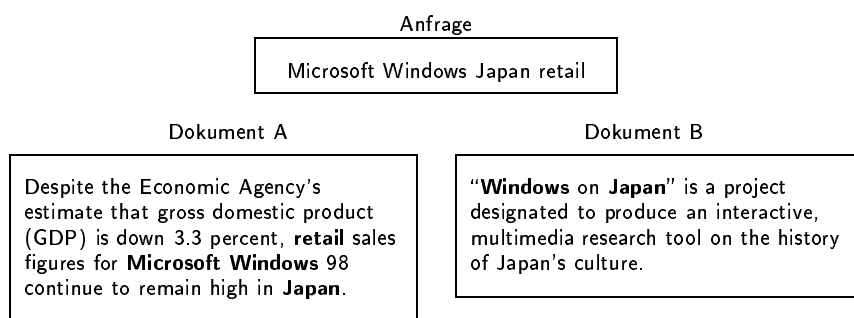
Man kann sich zwei vollkommen unterschiedliche Bedürfnisse bei einer Recherche vorstellen: Der Physiker auf der Suche nach dem Zahlenwert von Pi auf 40 Stellen genau gibt sich beispielsweise mit *einem einzigen* relevanten Dokument zufrieden. Ein Patentanwalt hingegen muss abklären, ob für eine neue Erfindung bereits ein Patent existiert oder nicht. Deshalb möchte er natürlich *möglichst alle* relevanten Dokumente auffinden, die ähnliche Erfindungen beschreiben. Er

wird also einen grösseren Teil der Rangliste in Betracht ziehen als der Physiker. Durch das Sortieren der Dokumente in der Rangliste gemäss ihrer Relevanzwerte wird diesen zwei völlig entgegengesetzten Bedürfnissen zugleich Rechnung getragen.

Rangierungsprinzipien

Wie geht nun ein Suchsystem konkret vor, um die Relevanz eines Dokuments bezüglich einer Anfrage zu berechnen? Das Vorgehen basiert auf einer wichtigen Annahme: Die Vorkommen von Suchbegriffen in einem Dokument geben Hinweise auf die Relevanz dieses Dokuments. Diese Annahme bildet die theoretische Grundlage für wissenschaftliche Modelle zur Berechnung der Relevanz. Die teilweise komplexen mathematischen Hintergründe sollen der Leserschaft hier erspart bleiben. Stattdessen erklären wir anhand von sechs Beispielen und den dazugehörigen Rangierungsprinzipien, welche Kriterien ein Suchsystem bei der Schätzung der Relevanz berücksichtigen kann. Es handelt sich um diejenigen Rangierungsprinzipien, die sich in vielen Fällen als besonders effektiv herausgestellt haben.

Die erste Beispielanfrage zielt auf Dokumente ab, die Angaben über die Verkaufszahlen von Microsoft-Windows-Produkten in Japans Einzelhandel enthalten.



Dokument A liefert Informationen zu diesem Thema und ist darum relevanter als Dokument B, das über ein kulturhistorisches Projekt in Japan informiert. Der offensichtliche Unterschied zwischen den beiden Dokumenten: In A kommen alle vier Suchbegriffe mindestens

einmal vor, während in B nur zwei der Suchbegriffe auftauchen. Diese Erkenntnis führt zu Rangierungsprinzip 1:

Rangierungsprinzip 1

**Je mehr Suchbegriffe in einem Dokument vorkommen,
desto wahrscheinlicher ist das Dokument relevant.**

Bei der zweiten Beispielanfrage geht es um Dokumente über Erdbeben.

Anfrage

Erdbeben

Dokument A

In China ereignete sich ein **Erdbeben** mit Epizentrum 200 km nördlich der Hauptstadt. Das **Erdbeben** dauerte 10 Sekunden. Der durch das **Erdbeben** verursachte Schaden geht in die Millionen.

Dokument B

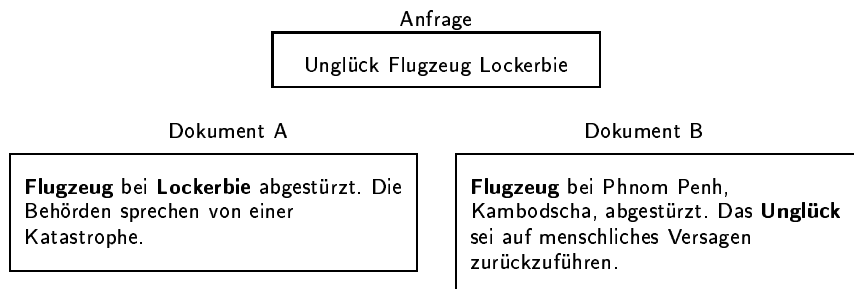
Nach dem **Erdbeben** von 1990 wurden die Schäden zügig repariert, und die Stadt erlebt seither einen wirtschaftlichen Aufschwung.

Dokument A dreht sich eindeutig um ein Erdbeben in China. Der Suchbegriff *Erdbeben* kommt darin dreimal vor. Im Gegensatz dazu geht es in Dokument B in erster Linie um den wirtschaftlichen Aufschwung in einer bestimmten Stadt. Ein Erdbeben wird nur am Rande erwähnt, und der entsprechende Suchbegriff taucht lediglich einmal auf. Deshalb:

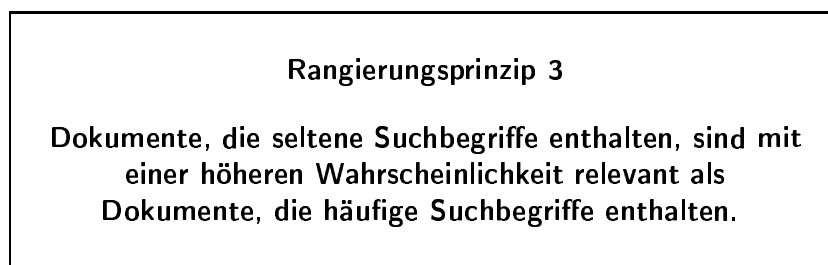
Rangierungsprinzip 2

**Je häufiger ein Suchbegriff in einem Dokument
vorkommt, desto wahrscheinlicher ist das Dokument
relevant.**

In der dritten Beispielanfrage sucht jemand nach Dokumenten über das Flugzeugunglück in Lockerbie.



In beiden Dokumenten tauchen je zwei Suchbegriffe der Anfrage auf; Rangierungsprinzip 1 hilft also nicht weiter. Und auch bezüglich Rangierungsprinzip 2 unterscheiden sich die Dokumente nicht, weil die Suchbegriffe in A und B gleich häufig vorkommen. Schauen wir also genauer hin: Der Begriff «Flugzeug» steht in beiden Dokumenten und bringt uns nicht weiter. Es verbleiben die Begriffe «Unglück» und «Lockerbie». In einer typischen Dokumentensammlung mit internationalen Nachrichten dürfte das Wort «Unglück» häufig auftreten. Das Wort «Lockerbie» dagegen bezeichnet einen spezifischen geographischen Ort und wird bedeutend seltener vorkommen. Dokument A dürfte deshalb bei dieser Anfrage mit grosser Wahrscheinlichkeit das relevantere Dokument sein. Das zugehörige Rangierungsprinzip lautet:



Die vierte Beispielanfrage zielt auf Informationen über das Leben von Nelson Mandela ab.

Anfrage

Nelson Mandela

Dokument A

Friedensnobelpreisträger von 1993

Im Jahre 1993 ging der Friedensnobelpreis an **Nelson Mandela**, geboren am 25. Juli 1918 in Transkei, Südafrika. Er trat 1944 dem Afrikanischen Nationalkongress (ANC) bei und engagierte sich gegen die Apartheid-Politik. 1964 wurde er angeklagt, den Sturz der Regierung geplant zu haben, und verbrachte die Jahre bis 1990 in Haft. 1991 wurde er zum Präsidenten des ANC gewählt.

Dokument B

Nobel Prize Winners

Nobel Prize in Literature

1997 D. Fo

...

1901 S. Prudhomme

Nobel Prize in Peace

1997 J. Williams

...

1993 **Nelson Mandela**, F. W. de Klerk

...

1901 J. H. Dunant, F. Passy

Nobel Prize in Economics

1997 R. C. Merton, M. S. Scholes

...

1969 R. Frisch, J. Tinbergen

Nobel Prize in Physics

1997 S. Chu, C. Cohen, W. D. Phillips

...

1901 W. C. Roentgen

Nobel Prize in Chemistry

1997 P. D. Boyer, J. E. Walker

...

1901 J. H. Van't Hoff

Nobel Prize in Medicine

1997 S. B. Prusiner

...

1901 E. A. von Behring

Die beiden Suchbegriffe *Nelson* und *Mandela* treten in beiden Dokumenten gleich oft auf. Allerdings behandelt Dokument A konkret das Thema in einem kurzen Abschnitt. Dokument B hingegen zeigt eine umfangreiche Liste mit den Namen aller Nobelpreisträger in den verschiedenen Kategorien. Als ein Name unter vielen taucht auch Nelson Mandela auf. Also lautet das entsprechende Rangierungsprinzip wie folgt.

Rangierungsprinzip 4

Ein kurzes Dokument ist mit einer höheren Wahrscheinlichkeit relevant als ein langes Dokument, welches die gleichen Suchbegriffe gleich häufig enthält.

Bei der fünften Beispielanfrage sollen Dokumente zum Big Ben in London gefunden werden.

Anfrage

Big Ben

Dokument A

Big Ben is the name of the 13 ton bell that produces the "Westminster Chime" in the heart of London every hour. The lamp in the spire is lit during House of Common debates.

Dokument B

UltraTV
Unsere Empfehlung für heute:
20:00 **Ben** Hur, Historien-Epos
23:45 Little **Big** Man

In Dokument A werden einige Fakten über die wohl berühmteste Kirchenglocke der Welt geliefert. Die beiden Suchbegriffe tauchen unmittelbar nebeneinander auf. In Dokument B gibt ein Fernsehsender seine Empfehlung für das Tagesprogramm bekannt. Obwohl auch in B beide Suchbegriffe vorkommen, ist weit und breit keine Spur von «Big Ben» zu finden. Die Suchbegriffe werden jeweils in anderem Zusammenhang verwendet. Das Rangierungsprinzip dazu:

Rangierungsprinzip 5

Je näher die Suchbegriffe beieinander liegen, desto wahrscheinlicher ist das Dokument relevant.

Bei der sechsten und letzten Beispielanfrage sind Informationen über den griechischen Philosophen Plato gesucht.

Anfrage

Plato

Dokument A

Plato

Griechischer Philosoph, 427–347 v. Chr.
Schüler des Sokrates. Mit Aristoteles
Begründer der abendländischen
Philosophie, schuf die erste Akademie.

Dokument B

Sokrates

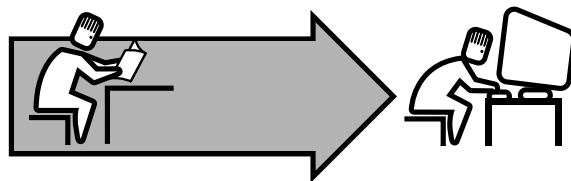
Griechischer Philosoph, 469–399 v. Chr.
In Athen der Gottlosigkeit und
Verführung der Jugend angeklagt und
zum Tod durch Gift verurteilt.
Seine Schüler: **Plato** und Xenophon.

Hier ist auf der Stelle klar, welches Dokument das relevantere ist. Zwar taucht der Suchbegriff *Plato* in beiden Dokumenten genau einmal auf, doch wird Plato in Dokument B erst zum Schluss kurz erwähnt. Der Rest von B behandelt Platos Lehrer Sokrates. Bei Dokument A kommt der Suchbegriff ganz zu Beginn vor, denn der Text handelt von Plato selbst. Es wird häufig beobachtet, dass Autoren die wichtigsten Schlüsselwörter relativ weit oben innerhalb des Texts verwenden. Die allgemeine Regel lautet:

Rangierungsprinzip 6

**Je früher die Suchbegriffe in einem Dokument
vorkommen, desto höher seine Relevanz.**

Mit Hilfe der obigen Rangierungsprinzipien haben wir das Wichtigste zusammengefasst, was Suchsysteme bei der Beurteilung der Relevanz von Dokumenten beachten. Es folgen die praktischen Aspekte ...



Die Rangliste in der Praxis

Ein Suchsystem wählt die Dokumente zu einer Anfrage aufgrund der Rangierungsprinzipien aus. Das Resultat dieser Auswahl ist eine Rangliste, die an die Benutzerin geschickt wird. Die Rangliste ist eine simple Liste von Dokumenten. Normalerweise erhält man von jedem Dokument den Titel als Hyperlink, der direkt zum zugehörigen Dokumentinhalt verweist. Manche Suchsysteme liefern zusätzliche Angaben zum Dokument, beispielsweise einen Auszug aus dem Inhalt, die Grösse, das Änderungsdatum oder auch die Sprache.

Nur die wenigsten Suchsysteme schicken die gesamte Rangliste in einem Stück, denn ein solches Vorgehen wäre bei zehn Millionen gefundenen Dokumenten denkbar ungünstig. Darum wird die Rangliste in gleich grosse Stücke von beispielsweise zehn Einträgen unterteilt. Für die Benutzerin stehen Hyperlinks zur Verfügung, um in der Rangliste vor- und zurückzublättern.

Sehr wichtig ist die Sortierung der Rangliste! Jedes Suchsystem vermerkt diejenigen Dokumente zuoberst in der Rangliste, für die es den höchsten Relevanzwert berechnet hat. Je weiter hinten in der Rangliste ein Dokument auftaucht, desto weniger relevant sollte es für die Anfrage sein. Also braucht es den Anfragersteller überhaupt nicht zu beunruhigen, wenn das Suchsystem eine Rangliste mit Millionen von Einträgen liefert. Entscheidend ist einzig, dass die wirklich relevanten Dokumente weit vorne in der Rangliste erscheinen.

Rangierungsprinzipien in der Praxis

Die sechs Rangierungsprinzipien machen klar, nach welchen Kriterien ein Suchsystem gefundene Dokumente rangieren *kann*. Leider bedeutet das nicht, dass sämtliche Suchsysteme alle Prinzipien tatsächlich anwenden. In der Regel veröffentlichen die Betreiber aus Konkurrenzgründen keine sehr konkreten Angaben über die interne Funktionsweise ihrer Systeme. Wie also soll man herausfinden, welche Prinzipien ein bestimmtes System anwendet?

Als ersten Schritt sollte man auf jeden Fall die Hilfeseiten des Suchdienstes durchlesen. Vielleicht gibt es da bereits erste Hinweise.

Weiter kann man davon ausgehen, dass die meisten Suchsysteme sicherlich die Rangierungsprinzipien 1 und 2 unterstützen. Auch das dritte Prinzip wird häufig unterstützt. Und ebenso Prinzip 6 in der einen oder anderen Variante; zum Beispiel indem Dokumente mit Suchbegriffen im Titel oder in Überschriften als relevanter eingestuft werden. Schliesslich verschafft man sich bei regelmässiger Verwendung einiger weniger Systeme ein Gefühl dafür, wie die Rangierung von Dokumenten funktioniert.

Ein weiteres Problem: Wie gewichtet ein Suchsystem die einzelnen Rangierungsprinzipien? Dokument A enthält vielleicht den Begriff *Währungsunion* dreimal im Text. In Dokument B kommt der Suchbegriff nur einmal vor, dafür an erster Stelle im Titel. Ist A oder B nun das relevantere Dokument? Die Antwort hängt davon ab, wie viel Wert das Suchsystem auf das jeweilige Rangierungsprinzip legt, und auch dafür verschafft man sich bei regelmässiger Anwendung ein Gefühl.

Inhaltsunabhängige Bestimmung der Relevanz

Die sechs besprochenen Rangierungsprinzipien richten sich nach dem Inhalt eines Dokuments. Daneben gibt es Kriterien, die sich vom Dokumentinhalt lösen. Solche Kriterien werden bei manchen Suchdiensten zusätzlich zu den üblichen Prinzipien angewendet.

Anzahl Referenzen

Für jedes Dokument in der Kollektion kann gezählt werden, wie viele andere Dokumente mittels Hyperlink darauf verweisen. Das ergibt die Anzahl der Referenzen von aussen auf ein Dokument. Dokumente mit vielen Referenzen erhalten einen höheren Relevanzwert. Hierbei kommt die subjektive Relevanz ins Spiel, denn Webseiten mit vielen Referenzen sind typischerweise diejenigen, die besonders beliebt sind.

Derselbe Grundsatz spielt auch bei Fachbüchern: Je mehr andere Autoren auf ein bestimmtes Buch verweisen, desto angesehen wird das Buch. Es entwickelt sich mit der Zeit unter Umständen zum Standardwerk auf dem entsprechenden Gebiet.

Anzahl Zugriffe

Ein Suchdienst hat mit recht geringem Aufwand die Möglichkeit zu zählen, wie oft auf ein bestimmtes Dokument zugegriffen wird. Dokumente mit hohen Zugriffszahlen gelten dann als relevanter als solche mit niedrigeren Zahlen. Die Überlegung: Wenn ein Dokument bereits von Millionen von Benutzern angeschaut wurde, so ist es wahrscheinlich relevant. Auch hier spielt die subjektive Relevanz eine Rolle.

Beide besprochenen Kriterien – Anzahl der Referenzen oder Zugriffe – arbeiten mit der Wahrscheinlichkeit, dass ein Dokument für eine *beliebige* Anfrage relevant ist. Es werden keine Suchbegriffe benötigt, um die Relevanzwerte zu bestimmen. Deshalb kombinieren Suchsysteme diese Kriterien mit den sechs Rangierungsprinzipien.

Wenn das Geld regiert...

Was heisst «relevant»? Bisher war ein Dokument immer dann relevant, wenn es die Bedürfnisse eines *Benutzers* befriedigte. Auf der anderen Seite könnte man die Relevanz auch über den *Anbieter* oder die *Autorin* einer Seite definieren.

Es gibt Suchdienste, die genau das tun. Sie wenden keine Rangierungsprinzipien an, sondern versteigern stattdessen ihre Suchbegriffe. Beim Begriff *Airline* beteiligt sich typischerweise eine Reihe von Fluggesellschaften an der Auktion. Die Gesellschaft mit dem höchsten Gebot darf dann bestimmen, welches Dokument für diesen Suchbegriff in der Rangliste an erster Stelle erscheinen wird – vermutlich die Homepage der Fluggesellschaft.

Web Spamming

Immer wieder missbrauchen auch «böse Zeitgenossen» die Rangierungsprinzipien von Suchsystemen mit dem Ziel, eine bessere Rangierung der eigenen Webseiten zu erreichen.

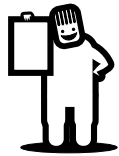
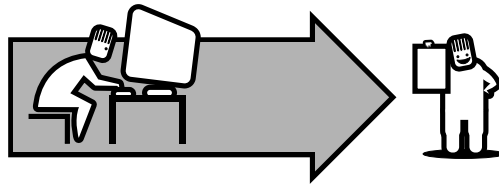
Manche Seiten wiederholen ihre Schlüsselwörter hundert- oder gar tausendfach. Rangierungsprinzip 2 sorgt dafür, dass solche Seiten weit vorne in der Rangliste zu einer entsprechenden Anfrage erscheinen. Damit die Wiederholungen das Aussehen der Seite nicht

beeinträchtigen, werden die Begriffe zum Beispiel in weisser Schrift auf weissem Hintergrund geschrieben.

Andere Seiten benützen Begriffe, nach welchen im Web zwar häufig gesucht wird, die aber nichts mit dem eigentlichen Inhalt der betreffenden Seite zu tun haben. So werden Benutzer in die Irre geführt.

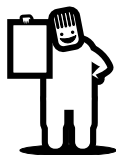
Die meisten Betreiber von Suchsystemen versuchen, Techniken dieser Art – *Web Spamming* genannt – einen Riegel vorzuschieben. Spezielle Software soll den Missbrauch entdecken. Die Urheber werden anschliessend durch den Ausschluss ihrer Seiten aus dem Angebot des betreffenden Suchdienstes bestraft. Trotzdem bleiben viele Fälle von Web Spamming unentdeckt und liefern manchmal die Erklärung für eigenartige Resultate auf Suchanfragen.

Nach den praktischen Hinweisen kommen wir auf die Probleme der Internet-Anwender zurück ...



Im März 1866 wurde Johan Vaaler in Aurskog, Norwegen, geboren. Er machte einen Abschluss in Elektronik und Mathematik und war schon als junger Mann als Erfinder bekannt. 1899 erfand Vaaler die Büroklammer. Das Patent auf die Erfindung musste er in Deutschland anmelden, weil in Norwegen damals noch kein Patentrecht existierte.

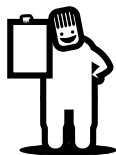
Wie man sieht, habe ich den Erfinder der Büroklammer unterdessen gefunden. Meine ursprüngliche Anfrage bestand aus einem einzigen Begriff – *Büroklammer*. Das ist natürlich zu wenig. Das Suchsystem hat so fast keine Chance, eine vernünftige Antwort zu liefern. Nach kurzem Überlegen sind mir weitere Suchbegriffe eingefallen. Und mit der Anfrage *Büroklammer Briefklammer erfand Erfinder Patent* habe ich schliesslich die Antwort auf meine Frage gefunden.



Was nehme ich aus diesem Kapitel mit? Ich soll für meine Anfragen gute Suchbegriffe verwenden, die für das gesuchte Dokument möglichst charakteristisch sind. Dazu muss ich mich in das gesuchte Dokument eindenken. Welche typischen Begriffe könnten darin vorkommen? So konstruiere ich eine Art «Wunschkokument» als Anfrage.

Ausserdem merke ich mir die Rangierungsprinzipien. Vor allem die ersten drei sind wichtig. Deshalb benutze ich in meinen Anfragen Suchbegriffe, die im gewünschten Dokument häufig und in allen anderen Dokumenten selten vorkommen. Bei der Zusammenstellung der Suchbegriffe denke ich ausserdem daran, welche Synonyme es für einen Begriff geben könnte. Die Synonyme füge ich dann ebenfalls zur Anfrage hinzu.

Mein Problem war die «unendliche Geschichte». Auf meine Anfrage wurden Dokumente geliefert, die nichts mit Endes Buch zu tun haben. Also muss ich weitere charakteristische Suchbegriffe suchen. In diesem Fall ist das leicht – ich zähle einfach einige der Fabelwesen auf, die in der Märchenwelt von Phantasien leben: *unendliche Geschichte Atreju Steinbeisser Fuchur*. Damit gebe ich dem Suchsystem konkrete Hinweise, wonach ich genau suche.

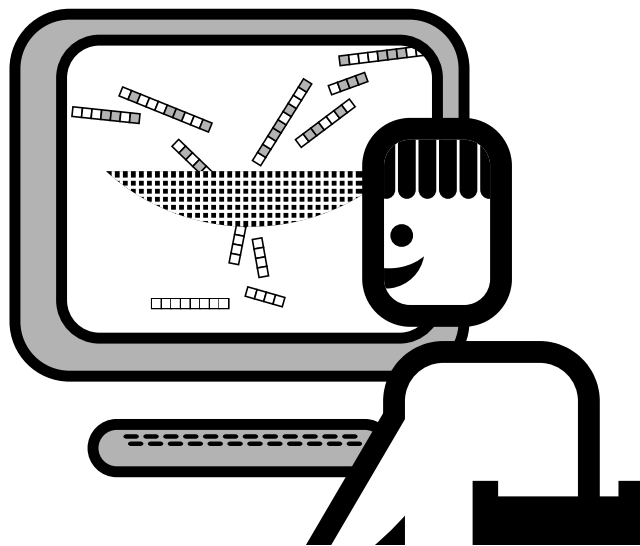


Ich bin immer noch erfolglos auf der Suche nach der Liste mit allen Schweizer Bundesräten. Wie sieht das gewünschte Dokument aus? Es enthält natürlich alle Namen der Bundesräte. Für eine gute Anfrage benutze ich demnach einfach die Namen einiger Bundesräte, die mir einfallen: *Gnägi Schlumpf Stich Ritschard Leuenberger Villiger Egli Metzler*.

Die Anfrage führt problemlos zum Ziel. Ich habe hier von meinem Hintergrundwissen Gebrauch gemacht. Wer mit der schweizerischen Politik nicht vertraut ist, dürfte Mühe haben, mehrere Bundesräte mit Namen zu nennen.

Kapitel 3

Indexierung von Textdokumenten





Seit dem letzten Kapitel habe ich mir angewöhnt, mich in die gesuchten Dokumente «einzudenken». Dabei beachte ich die Rangierungsprinzipien und verwende möglichst viele charakteristische Begriffe in meinen Anfragen. Unterdessen frage ich mich aber: Wie «liest» das Suchsystem die Webseiten und meine Anfrage? Wie muss ich die Begriffe ins Suchformular eingeben?

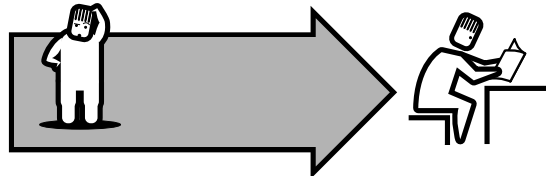
Aufgetaucht sind diese Fragen, als ich herausfinden wollte, welche Staaten an der Europäischen Währungsunion beteiligt sind. Eine mögliche Anfrage bei OMNISEARCH wäre etwa *Mitgliedstaaten Europäische Währungsunion*. Allerdings sind auch Dutzende von anderen Varianten denkbar. Es könnte auch einfach von «Mitgliedern» die Rede sein oder von «Europa» oder von der «Währungs-Union» mit Bindestrich. Unsicher bin ich auch bei der Gross- und Kleinschreibung. Soll ich *Europäische* oder *europäische* schreiben? Zu allem Übel habe ich auch noch gehört, dass bei Computern die deutschen Umlaute häufig für Probleme sorgen. Muss ich also zusätzlich *europaeische* anstelle von *europäische* ausprobieren?



Plasmodiophora brassicae – das ist der Name eines Pilzes, der die so genannte Kohlhernie verursacht. Die Kohlhernie ist eine Krankheit, die vor allem bei Blumenkohl und anderen Kohlarten auftritt. Es kommt zu Wucherungen an den Wurzeln, die Pflanzen welken und gehen ein.

Ich wollte mich bei NEWSSEEKER informieren, ob die Kohlhernie jemals verantwortlich war für einen bedeutenden Ernteverlust in der Blumenkohlproduktion. Dabei habe ich eine erstaunliche Entdeckung gemacht: Auf die Anfrage *Blumenkohl* meldet NEWSSEEKER ein Dokument mit dem Titel «Rot-grüne Mehrheit für einen Machtwechsel in Bonn». Was soll das? Wählt NEWSSEEKER die Dokumente zufällig aus? Ich kann mir das nicht erklären.

Werfen wir zur Klärung der Fragen einen Blick hinter die Kulissen von Suchsystemen und betrachten, wie die Systeme mit Dokumenten und Anfragen umgehen ...



Eine Beispielindexierung

Die Rangierungsprinzipien im letzten Kapitel haben gezeigt: Zur Bestimmung der Relevanz eines Dokuments wird der Dokumentinhalt mit den Suchbegriffen der Anfrage verglichen. Das Vorgehen stützt sich auf den folgenden Grundsatz: Die Vorkommen von Suchbegriffen in einem Dokument liefern Hinweise auf die Relevanz des Dokuments. Aber was ist ein Suchbegriff?

Ein Suchsystem bestimmt die Suchbegriffe, indem es eine Anfrage oder ein Dokument gründlich untersucht. Diese Analyse wird *Indexierung* genannt. Wir werden ein konkretes Beispiel durcharbeiten, um zu verstehen, was bei einer möglichen Indexierung geschieht. Dabei werden die wichtigsten Schritte aufgezeigt. Viele Systeme begnügen sich mit einer bescheideneren Indexierung. Es gibt aber auch Systeme, die wesentlich aufwendigere Indexierungsmethoden anwenden. Ausgangspunkt in unserem Beispiel ist ein Dokument zum Thema Raumfahrt sowie eine entsprechende Anfrage:

Anfrage	Dokument
Wann gelang die erste Mondlandung?	Am 20. Juli 1969 landeten die Amerikaner erstmals auf dem Mond. Neil Armstrong berührte den Mond als Erster und verkündete: "One small step for man, one giant step for mankind."

Bei diesem Beispiel handelt es sich um ein Dokument, das mehrere Sprachen verwendet. Ein Teil der folgenden Indexierung arbeitet abhängig von der Sprache des Dokuments; zum Beispiel mit entsprechenden Wörterbüchern oder sprachabhängigen Regeln. Deshalb werden im ersten Schritt die verschiedenen Sprachen im Text identifiziert (*Sprachidentifikation*). Im Beispiel handelt es sich um Deutsch im ersten Textteil und in der Anfrage, während der zweite Textteil in Englisch geschrieben ist.

Im zweiten Schritt der Indexierung – *Buchstabenumwandlung* genannt – werden die deutschen Umlaute ä, ö, ü durch die Schreibweisen mit zwei Buchstaben ae, oe und ue ersetzt. Dasselbe gilt für andere Sprachen. Zum Beispiel wird bei französischen Dokumenten das é zu e oder das ç zu c umgeschrieben. Im Beispieltext kommt die Buchstabenumwandlung nur zweimal zum Zug, aus «berührte» wird «beruehrte» und statt «verkündete» steht neu «verkuendete»:

Wann gelang die
erste Mondlandung?

Am 20. Juli 1969 landeten die Amerikaner
erstmals auf dem Mond. Neil Armstrong
beruehrte den Mond als Erster und verkuendete:
"One small step for man, one giant step for
mankind."

Es folgt die *Wortextraktion*. In dieser Phase werden die einzelnen Wörter aus dem Text herausgelöst. Dabei gehen vor allem die Interpunktionszeichen verloren. Im Beispiel sind das die Punkte, die Anführungszeichen, das Komma, der Doppelpunkt und das Fragezeichen in der Anfrage.

Wann gelang die
erste Mondlandung

Am 20 Juli 1969 landeten die Amerikaner
erstmals auf dem Mond Neil Armstrong beruehrte
den Mond als Erster und verkuendete One small
step for man one giant step for mankind

Im Anschluss werden durch die *Stoppwortelimination* alle Stoppwörter entfernt. Stoppwörter sind Begriffe, die nichts oder nur sehr we-

nig zur Beschreibung des Inhalts eines Dokuments beitragen. Beispielsweise liefert der weibliche bestimmte Artikel «die» keinen hilfreichen Hinweis auf den Inhalt eines Dokuments. Warum diese Begriffe Stoppwörter genannt werden, kann nur vermutet werden. Eine hilfreiche Assoziation ist der Ausspruch: «Stopp! Dieses Wort wird ignoriert.» Im Beispieltext trifft es Wörter wie «die», «auf», «das» und «for».

gelang erste
Mondlandung

20 Juli 1969 landeten Amerikaner erstmals Mond
Neil Armstrong beruehrte Mond Erster
verkuendete One small step man one giant step
mankind

In der nächsten Phase geht es um Wortzerlegungen und Wortnormalisierungen. Je nach verwendeter Sprache werden während der *Wortzerlegung* alle zusammengesetzten Begriffe in ihre Einzelteile (Komposita) aufgeteilt. Die Wortzerlegung ist in Sprachen wie Deutsch oder Finnisch von grosser Bedeutung. In anderen Sprachen wie Englisch und Französisch wird die Wortzerlegung im Allgemeinen nicht angewendet, weil fast keine zusammengesetzten Wörter vorkommen.

Im Beispiel werden aus dem Begriff «Mondlandung» die beiden Worte «Mond» und «Landung» gebildet. Anschliessend werden alle verbleibenden Wörter auf eine feste Normalform zurückgeführt. Mit einer solchen *Wortnormalisierung* erreicht man, dass Wörter in unterschiedlichen Flexionen (durch Deklination oder Konjugation entstandene Beugungen von Wörtern) als identisch betrachtet werden, obwohl sie anders geschrieben sind. Ausserdem werden während der Wortnormalisierung alle Grossbuchstaben in Kleinbuchstaben umgewandelt.

Je nach Sprache werden für die Wortnormalisierung zwei wichtige Techniken verwendet. In der deutschen Sprache ermittelt man eine *Grundform*, zum Beispiel Nominativ, Singular. Dazu greift das Suchsystem auf ein Wörterbuch zurück. Aus einem Wort wie «Häusern» wird die Grundform «Haus» ermittelt. Andere Sprachen wie Englisch kennen keine Beugungen im Wortinnern. Es kann ein einfacheres Verfahren angewendet werden: Bei der *Wortstammreduktion*

werden aufgrund von Regeln einfach die Suffixe entfernt. Beispiele: «sings» wird zu «sing» (Suffix -s), «invented» wird zu «invent» (Suffix -ed) und «going» wird zu «go» (Suffix -ing).

Die im Dokument verwendeten Sprachen spielen bei der Wortzerlegung und Wortnormalisierung eine grosse Rolle. Suchdienste mit einer vertikalen Dokumentensammlung können üblicherweise eine sorgfältigere Indexierung anbieten, weil oft nur eine oder einige wenige Sprachen benutzt werden. Schwierigkeiten gibt es dagegen bei globalen, horizontalen Kollektionen, wo vielleicht Dutzende von verschiedenen Sprachen im Einsatz sind.

Unser Beispieltext verwendet nur zwei Sprachen und sieht nach der Wortzerlegung und Wortnormalisierung so aus:

geling erst mond land	20 juli 1969 land amerika erst mal mond neil armstrong beruehr mond erst verkuend one small step man one giant step mankind
--------------------------	---

Es bleibt, die einzelnen Begriffe im Dokument zu zählen und eine entsprechende Tabelle aufzustellen. Die Tabelle wird beispielsweise für das Rangierungsprinzip 2 benötigt, welches besagt, dass ein Dokument umso relevanter bewertet wird, je häufiger ein Suchbegriff darin vorkommt. Ausserdem hält die Tabelle die Position der Begriffe innerhalb des Dokuments fest. Mit Hilfe der Position lässt sich die Distanz zwischen zwei Suchbegriffen (Rangierungsprinzip 5) bestimmen.

Begriff	Häufigkeit	Positionen
mond	2	8, 12
erst	2	6, 13
step	2	15, 19
land	1	4
1969	1	3
...	...	...

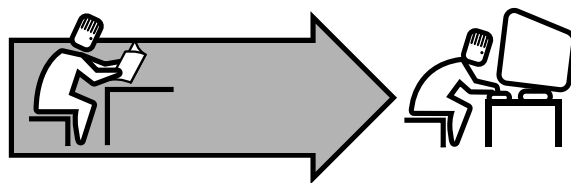
Die Tabelle kann man als vereinfachte Darstellung des Dokumentinhalts ansehen. Dank der Indexierung werden drei Suchbegriffe der Anfrage – «mond», «land» und «erst» – im Dokument gefunden, obwohl in der Originalversion des Dokuments keine einzige Übereinstimmung zwischen Anfrage und Dokument vorlag.

Weitere Merkmale von Dokumenten

Viele Suchsysteme erweitern die Indexierung und ziehen zusätzliche Eigenschaften oder Strukturen in einem Dokument in Betracht. Neben den Begriffen aus dem Inhalt können weitere Merkmale aus dem Dokument extrahiert werden:

- *Dokumenttitel und Überschriften:* In einem korrekt verfassten HTML-Dokument können der Titel des Dokuments sowie die einzelnen Überschriften im Text problemlos identifiziert werden. Leider halten sich bei weitem nicht alle Verfasser an die vorgesehenen Regeln zur Erstellung von HTML-Dokumenten.
- *Hyperlinks:* Verweise auf andere Dokumente (so genannte Hyperlinks) können ebenfalls bestimmt werden.
- *Dokument-Adresse:* Die Adresse des betrachteten Dokuments (der URL) ist natürlich immer verfügbar.
- *Modifikationsdatum:* Teilweise lässt sich das Datum der letzten Änderung und damit das Alter eines Dokuments herausfinden.
- *Meta-Tags:* Die Meta-Tags sind ein Bestandteil von HTML. Sie dienen dazu, zusätzliche Angaben zu einer Webseite festzuhalten, beispielsweise den Autorennamen, Schlüsselwörter oder Zusammenfassungen. Die Angaben innerhalb der Meta-Tags werden von einem Web-Browser nicht angezeigt, können aber von einem Suchsystem berücksichtigt werden.

Es folgen einige Hinweise, wie man in der Praxis durch geeignete Beispielanfragen herausfinden kann, wie sorgfältig ein bestimmtes Suchsystem die Indexierung durchführt ...



Eigene Experimente helfen weiter

Das Wichtigste zuerst: Das soeben vorgestellte Verfahren zur Indexierung ist keineswegs allgemein gültig! Stattdessen handelt es sich um eine Beispielindexierung, welche die wichtigsten Schritte berücksichtigt. Viele Suchsysteme führen eine einfachere Indexierung ohne Buchstabenumwandlung und Wortzerlegung und -normalisierung durch. Ein bestimmtes Suchsystem wendet vielleicht nur die Stoppwortelimination an. Bei einer derart reduzierten Indexierung ändern sich Text und Anfrage nur sehr wenig. Folglich würde in unserem Beispiel die Anfrage zur Mondlandung das relevante Dokument nicht finden, da die Suchbegriffe der Anfrage mit keinem der Begriffe im Dokument übereinstimmen:

gelang erste Mondlandung

20 Juli 1969 landeten Amerikaner erstmals Mond Neil Armstrong berührte Mond Erster verkündete One small step man one giant step mankind

Unsere beiden Suchdienste NEWSSEEKER und OMNISEARCH decken zwei Extreme ab. OMNISEARCH führt lediglich eine sehr einfache Indexierung durch. Es bleibt im Wesentlichen bei der Wortextraktion und der Stoppwortelimination. NEWSSEEKER dagegen wendet alle oben vorgestellten Techniken an.

Man könnte nun natürlich für alle bekannten Suchdienste aufzählen, welche Art der Indexierung sie durchführen. Allerdings wäre das langweilig und wenig hilfreich, denn Suchsysteme können ihre Funktionsweise von heute auf morgen ändern. Stattdessen folgen einige Tipps, wie man bei seinem Lieblingssuchdienst selber herausfinden kann, wie die Indexierung abläuft.

Der einfachste und offensichtlichste Rat zuerst: Unbedingt einen Blick in die Hilfeseiten des Suchdienstes werfen! Vielleicht lässt sich dort bereits alles Wichtige finden. Falls die Hilfeseiten nur mager ausgestattet sind, helfen eigene Experimente weiter. Im Folgenden werden mögliche Experimente beschrieben.

Wortzerlegung und -normalisierung

Man führt eine Anfrage mit einem zusammengesetzten Begriff durch, der in der entsprechenden Dokumentensammlung sehr selten oder gar nie vorkommt. Anhand der gefundenen Dokumente kann man nun erkennen, ob eine Wortzerlegung durchgeführt wurde. Zum Beispiel kann man beim Suchdienst NEWSSEEKER die Anfrage *Schlagsahne* starten und erhält ein Dokument mit dem überraschenden Titel «Erfolgreiches Jahr für die Grossbanken». Zwei Sätze im Text liefern die Erklärung: «Die Banken jammern und *sahnen* Rekordgewinne ab» und «Im Devisenmarkt weht ein rauer Wind, der sich in Verlusten *niederschlägt*». Hinter NEWSSEEKER steckt folglich ein Suchsystem, welche Wortzerlegung und Wortnormalisierung durchführt. Andernfalls wäre das gezeigte Dokument mit der benutzten Anfrage nicht gefunden worden.

Die Wortzerlegung und -normalisierung kann den Benutzerinnen viel Arbeit abnehmen. Trotzdem bieten viele Suchdienste diesen Service aus Effizienzgründen nicht an. Manchmal kann man sich mit den so genannten *Wildcards* behelfen. Wildcards funktionieren als Platzhalter für beliebige Buchstabenkombinationen. Häufig wird der Stern (*) als Platzhalter verwendet. Der Suchbegriff *paint** genügt dann, um alle vier Formen *paints*, *painted*, *paint* und *painting* abzudecken. Leider können die Platzhalter oft nicht im Wortinnern oder am Wortanfang angewendet werden.

Stoppwörter

Suchsysteme können die Stoppwörter auf zwei unterschiedliche Arten festlegen: (1) Die Stoppwörter sind in einer Stoppwortliste fest vorgegeben. Solche Stoppwortlisten sind üblicherweise an die jeweilige Dokumentensammlung und die verwendete Sprache angepasst. (2) Manche Suchsysteme machen es sich einfacher und bestimmen: Bei uns gelten die 300 häufigsten Begriffe in der Dokumentensammlung als Stoppwörter (natürlich können es mehr als 300 oder auch weniger sein).

Viele Suchsysteme führen eine Stoppwortelimination durch. Man kann das leicht selber überprüfen, indem man ein sehr häufig be-

nutztes Wort der entsprechenden Sprache als Suchbegriff verwendet. Beispielanfrage: *to be or not to be*. Viele Suchsysteme finden das Theaterstück von Shakespeare nicht, weil sie sämtliche Suchbegriffe als Stoppwörter betrachten. Doch Achtung! Einige Suchsysteme behandeln gerade diese Anfrage speziell und liefern die erwarteten Resultate. In solchen Fällen helfen an sich sinnlose Anfragen wie ein *the* oder *to* in Englisch beziehungsweise ein *die* oder ein *er* in Deutsch weiter.

Umlaute und Akzente

Hier gilt es herauszufinden, ob das Suchsystem eine Buchstabenumwandlung durchführt. Dazu wählt man einen Begriff mit einem Umlaut (oder einem Akzent) und führt die Anfrage einmal mit und einmal ohne Umlaut durch. Zum Beispiel *Währungsunion* und *Waehrungsunion* oder *Nestlé* und *Nestle*. Liefert das Suchsystem in beiden Fällen dieselben Resultate, so wird die Buchstabenumwandlung durchgeführt. Andernfalls kommt man nicht darum herum, bei Anfragen mit Umlauten oder Akzenten beide Schreibweisen zu versuchen.

Gross- und Kleinschreibung

Auch in Bezug auf die Gross- und Kleinschreibung verhalten sich die Suchsysteme unterschiedlich. Ein einfaches Experiment bringt etwas Licht in die Sache: Man wählt eine gebräuchliche Abkürzung und stellt Anfragen in verschiedenen Variationen. Zum Beispiel mit den drei Anfragen *UNESCO*, *unesco* und *uNEscO*.

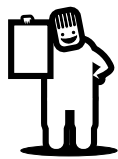
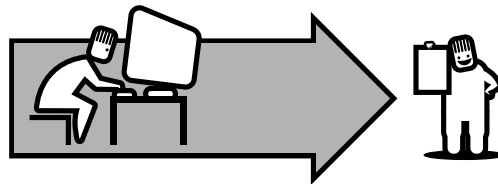
Viele Suchsysteme (zum Beispiel OMNISEARCH) gehen nach folgendem Prinzip vor: Vollständig klein geschriebene Suchbegriffe finden Begriffe in beliebiger Schreibweise. Doch sobald auch nur ein Buchstabe gross geschrieben wird, ist eine exakte Übereinstimmung erforderlich. Andere Suchsysteme – wie beispielsweise NEWSSEEKER – führen eine umfangreiche Indexierung durch und verlassen sich darauf, dass in der Anfrage die üblichen Orthografieregeln verwendet werden. Die linguistischen Komponenten für die Wortzerlegung und -normalisierung können bei korrekter Orthografie zuverlässiger arbeiten. Beispiel: Der Suchbegriff *stelle*, das heisst das Verb «stellen»

in der ersten Person Singular, findet alle Varianten des Verbs wie «herstellen», «unterstellen» oder einfach «stellen». Beim Suchbegriff *Stelle* dagegen ist das Substantiv gemeint. Eine solche Anfrage findet «Lehrstellen», «Tankstellen» und «Baustellen».

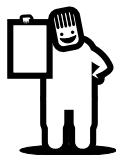
Zusatzstrukturen

Um etwas über die Zusatzstrukturen herauszufinden, müssen die Hilfeseiten zu Rate gezogen werden. Dort sollte beschrieben sein, was der Suchdienst unterstützt und wie die Strukturen angesprochen werden können. Bei OMNISEARCH lässt sich beispielsweise mit *domain:uk* gezielt nach Dokumenten suchen, deren URL als Länderbezeichnung UK enthält.

Nach den Hinweisen für die Praxis sollten wir gewappnet sein für die Lösung der Anwenderprobleme ...



Meine Frage lautete: Welche Staaten beteiligen sich an der Europäischen Währungsunion? Offenbar muss ich bei OMNISEARCH in den sauren Apfel beissen und alle Kombinationen durchspielen. Bei diesem Suchdienst muss ich einen Mehraufwand betreiben, weil mir die Indexierung fast keine Arbeit abnimmt. Immerhin hilft mir manchmal der Platzhalter. So kann ich mit der Anfrage *mitglied* währungsunion europ** zahlreiche Varianten wie «Mitglied», «Mitglieder», «Mitgliedstaaten», «europäisch», «europäische» usw. auf einen Schlag abdecken. Übrigens verwende ich bei OMNISEARCH ab sofort in der Regel klein geschriebene Suchbegriffe, um so beliebige Variationen bezüglich Gross- und Kleinschreibung abzudecken.



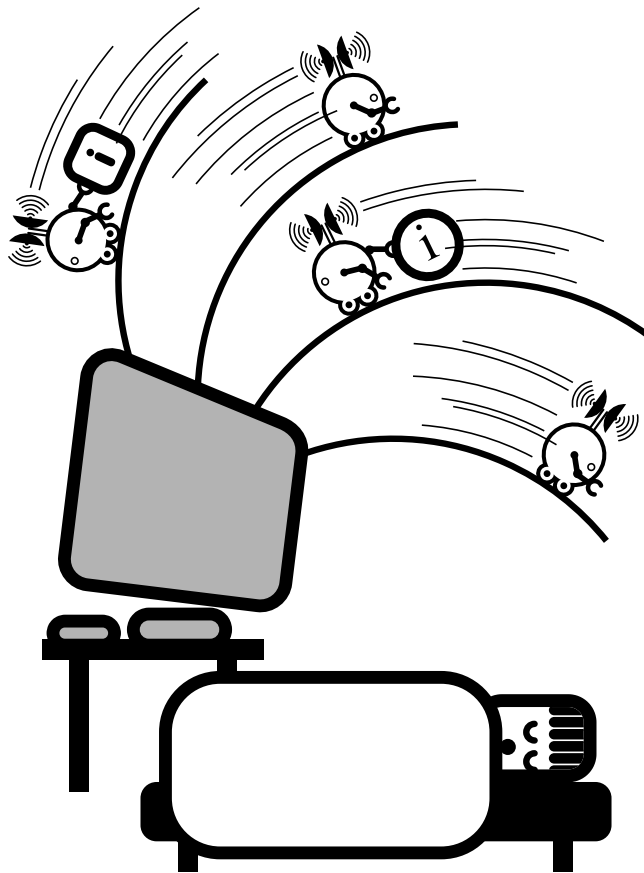
Unterdessen erstaunt es mich nicht mehr, dass ich als Antwort auf die *Blumenkohl*-Anfrage das Dokument mit dem Titel «Rot-grüne Mehrheit für einen Machtwechsel in Bonn» erhalten habe. Im Text geht es nämlich um die Wahl Gerhard Schröders zum Bundeskanzler der BRD.

Abgelöst wurde damit Helmut *Kohl* nach 16-jähriger Amtszeit. Ausserdem wird beschrieben, wie Kohl zum Abschied mit *Blumen* beschenkt wurde. Daraus kann ich schliessen, dass NEWSSEEKER mit Hilfe der Wortzerlegung mit dem Suchbegriff *Blumenkohl* auch die Begriffe *Kohl* und *Blumen* findet. Weiter kann ich schliessen, dass offenbar kein Dokument mit dem Begriff *Blumenkohl* in der Kollektion von NEWSSEEKER existiert. Deshalb wird mir das gefundene Dokument mit den aufgetrennten Begriffen an erster Stelle präsentiert.

Die Situation ist ganz und gar nicht benutzerfreundlich – jeder Suchdienst verhält sich etwas anders. Am besten wäre es, wenn sich alle Suchdienste an gewisse Standards hielten. Leider tun sie das nicht, und man muss die enorme Vielfalt in Kauf nehmen. Trotzdem habe ich mir drei Dinge vorgenommen: Erstens, ich schaue mir die Hilfeseiten der Suchdienste an. Zweitens, ich spiele mit den Systemen und führe eigene Experimente durch. Drittens, ich konzentriere mich auf einige wenige Informationsdienste. So muss ich mir nicht so viele Eigenheiten merken und lerne die Systeme immer besser kennen.

Kapitel 4

Funktionsweise von Suchsystemen





In den frühen 1960ern unterzeichneten zwölf Staaten das Antarktisabkommen. Dutzende weiterer Staaten schlossen sich unterdessen dem Vertrag an. Damit ist die Antarktis der einzige Kontinent, der vollständig durch ein internationales Abkommen regiert wird.

Ich suche gerade nach dem genauen Vertragstext des Antarktisabkommens. Dazu benutze ich NEWSSEEKER und die Anfrage *Antarktisabkommen Frieden Umweltschutz Artikel*. Gleich beim ersten Versuch stosse ich in der Rangliste auf einen sehr viel versprechenden Eintrag: «Die 14 Artikel des Antarktisabkommens». Voller Vorfreude wähle ich den Link an, und was geschieht?

Page not found!

Mit dieser Fehlermeldung beglückt mich mein Browser. Das ist nicht das erste Mal, dass mir so etwas passiert, und jedes Mal ärgere ich mich etwas mehr. Wollen die Suchdienste mich als Benutzer auf den Arm nehmen? Wieso wird mir eine viel versprechende Seite mit perfekter Beschreibung angeboten, die dann offenbar gar nicht mehr existiert?

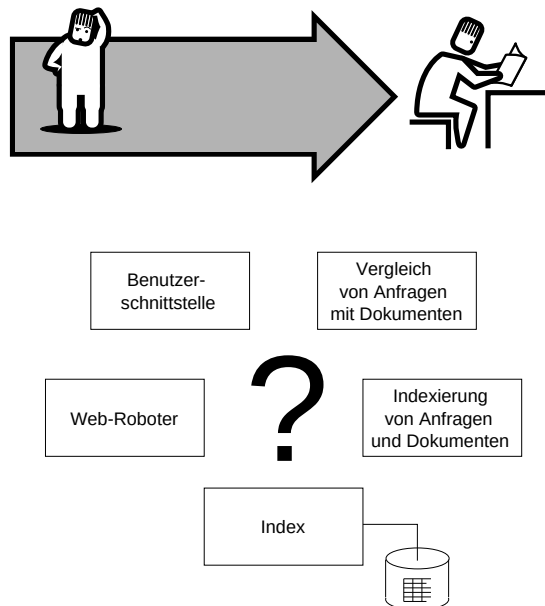


Ein ähnlich frustrierendes Erlebnis: Ich suche nach den aktuellen Fussballresultaten in der britischen Liga. Ich versuche mein Glück mit OMNISEARCH und finde rasch einige relevante Dokumente. Eine Webseite liegt auf dem Server einer englischen Zeitung und hat dort die Adresse `sports/headline.html`. Gemäss Titel geht es um den Sieg von Manchester United gegen Arsenal. Das interessiert mich, also klicke ich auf den Link und lande erfolgreich auf der Seite. Doch dort finde ich kein Wort über Fussball – stattdessen werde ich mit der Schlagzeile «Tour de France with new leader» überrumpelt.

Immer wieder habe ich den Eindruck, dass diese Suchdienste völlig willkürliche Resultate liefern. Irgendwie ist es ja auch gar nicht möglich, dass die Systeme ihre Arbeit wirklich sorgfältig erledigen! OMNISEARCH behauptet, etwa 200 Millionen Webseiten zu durchsuchen. Es ist doch schlicht unmöglich, dass jede einzelne Seite herun-

tergeladen und durchsucht wird. Immerhin erhalte ich meine Antwort jeweils innert Sekunden!

Zur Erklärung dieser eigenartigen Phänomene und der aufgeworfenen Fragen müssen wir uns zunächst mit der Funktionsweise von Suchsystemen auseinander setzen ...



Zwei der Komponenten eines Suchsystems haben wir bereits ausführlich besprochen. Für den Anfrage-Dokumentenvergleich stützt sich das Suchsystem auf gewisse Rangierungsprinzipien, so wie wir sie im zweiten Kapitel beschrieben haben, und produziert eine Rangliste, die an den Benutzer geschickt wird. Eine mögliche Indexierung wurde in Kapitel drei vorgestellt. Es verbleiben zwei ungeklärte Fragen: Wie beschafft sich das Suchsystem die Dokumente aus dem Internet, wie funktioniert der Web-Roboter? Und: Wieso kann ein Suchsystem innert Sekunden eine Rangliste ermitteln, auch wenn die Kollektion Millionen von Dokumenten enthält? Antwort: Der Index ist der Hauptgrund für die enorme Schnelligkeit bei der Beantwortung von Anfragen an das System. Wie aber ist der Index aufgebaut?

Der Web-Roboter

Es gibt Millionen von Web-Servern, die Hunderte von Millionen von Webseiten anbieten. Irgendwann entsteht auf irgendeinem dieser Server eine neue Seite, oder es wird eine der bestehenden Seiten geändert. Wie bringt OMNISEARCH all dies in Erfahrung?

Hier kommt der Web-Roboter (auch Spider oder Crawler genannt) ins Spiel. Der Web-Roboter ist ein Programm mit der Aufgabe, Webseiten zu finden. Dazu nützt der Roboter die Eigenschaft des World Wide *Web* aus, dass die Dokumente über Hyperlinks miteinander verbunden oder eben *verwoben* sind.

Damit ist das Vorgehen eigentlich klar:

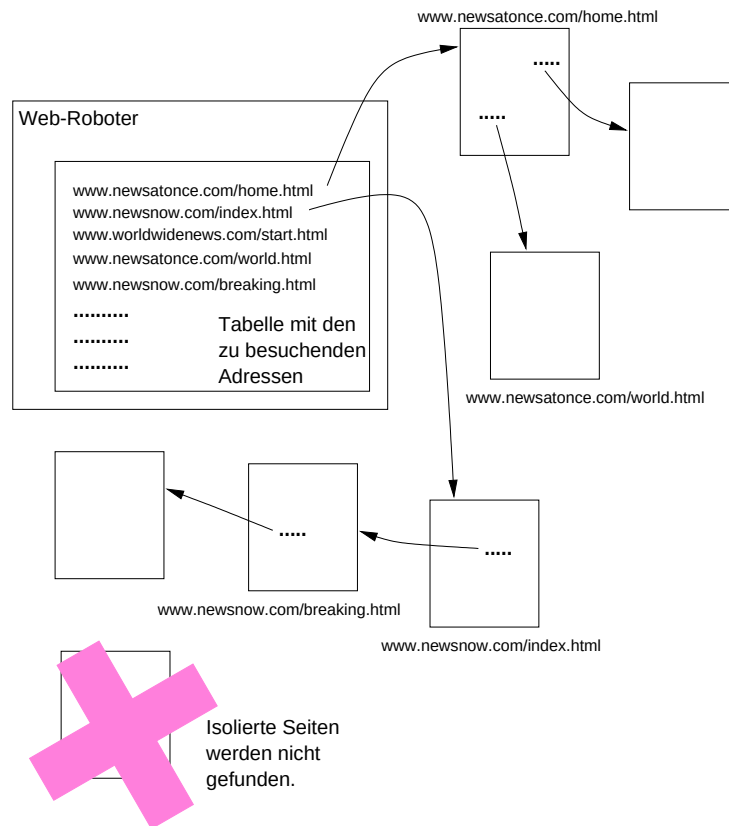
1. In einer Tabelle legt eine für den Suchdienst verantwortliche Person die Startpunkte für die Suche nach Webseiten fest.
2. Der Web-Roboter geht durch diese Liste mit URLs und bezieht die zugehörigen Seiten aus dem Internet. Dann wird jede Seite nach weiterführenden Verweisen (Hyperlinks) untersucht.
3. Die gefundenen Hyperlinks landen ebenfalls in der Tabelle mit den URLs, damit der Web-Roboter über die schon besuchten Seiten Bescheid weiss. Später werden auch die neu eingetragenen Seiten nach weiteren Verweisen untersucht. Auf diese Weise arbeitet sich der Web-Roboter immer weiter in die Tiefen des WWW vor.
4. Der Web-Roboter muss die entstandene URL-Tabelle regelmässig neu durcharbeiten und die entsprechenden Webseiten begutachten. Nur so kann er herausfinden, ob sich der Inhalt einer Seite geändert hat.

Früher oder später findet der Web-Roboter somit alle Seiten, die auf irgendeinem Weg von den Startseiten aus erreicht werden können. Oder umgekehrt: Er findet *keine* isolierten Seiten.

Ein Beispiel

NEWSSEEKER bietet eine vertikale Dokumentensammlung bestehend aus internationalen Nachrichtenmeldungen an. Wir gehen davon aus,

dass die Meldungen von insgesamt drei fiktiven Anbietern im WWW stammen: *NewsAtOnce*, *WorldWideNews*, und *NewsNow*. Also muss der Web-Roboter von NEWSSEEKER die zugehörigen Web-Server besuchen und möglichst alle Dokumente finden:



In der URL-Tabelle stehen zunächst nur die Einstiegsseiten der drei Web-Server. Die Adressen der neu gefundenen Seiten trägt der Roboter ebenfalls in die Tabelle ein, zum Beispiel den URL `http://www.newsatonce.com/world.html`. Falls jemand vergisst, eine Webseite mit einer neuen Meldung durch eine bestehende zu referenzieren, so entsteht eine isolierte Seite, die nicht gefunden wird.

Das Kleingedruckte

Web-Roboter dürfen sich nicht völlig frei im Internet bewegen. Die Suchsystembetreiber können dem Roboter gewisse Regeln vorschreiben, an die er sich zu halten hat. Diese Vorschriften sehen je nach Verwendung des Suchsystems anders aus. Ein Beispiel, wie das bei NEWSSEEKER aussehen könnte:

```
INCLUDE_PATTERN  http://www.newsatonce.com/*
INCLUDE_PATTERN  http://www.worldwideneews.com/*
INCLUDE_PATTERN  http://www.newsnow.com/*
...
EXCLUDE_PATTERN  *.gif
EXCLUDE_PATTERN  *.jpg
...
EXCLUDE_PATTERN  *.wav
EXCLUDE_PATTERN  *.mpg
...
```

Wie erwähnt konzentriert sich NEWSSEEKER auf das Angebot von drei Web-Servern. Deshalb wird hier für *jeden* Hyperlink verlangt, dass eine der drei Server-Adressen im URL vorkommt. Es werden keine Verweise weiterverfolgt, die zu anderen Web-Servern führen. Bei horizontalen Dokumentensammlungen wie derjenigen von OMNISEARCH fällt diese Restriktion weg, weil möglichst alle Seiten im Web angeboten werden sollen.

Weiter wird durch den Regelsatz bestimmt: Alle Verweise auf Bilddateien (GIF, JPEG usw.), Audiodateien (WAV usw.) oder Videosequenzen (MPEG usw.) werden ignoriert.

Der Index

Der Web-Roboter ist dafür verantwortlich, Dokumente im WWW zu finden. Die gefundenen Dokumente reicht der Roboter weiter an die Indexierungskomponente des Suchsystems. Nach der Indexierung werden die Informationen aus jedem Dokument im Index festgehalten. Ein Ausschnitt aus dem Index von OMNISEARCH könnte wie folgt aussehen.

Begriffe		Vorkommen			URL
mars	chocolate.com/ mars.html	geschichte.de/ mars.html	planets.org/ list.html	...	
	120	36	3	...	Häufigkeit
	7, 12, 51, ...	21, 23, ...	12, 33, 40	...	Positionen
pluto	disney.com/ comics.html	planets.org/ list.html	...	...	
	78	15	...	...	
	1, 18, ...	67, 73, ...	...	...	
saturn	sega.com/ consoles.html	cars.uk/ dealers.html	planets.org/ list.html	...	
	99	10	8	...	
	1, 4, 9, ...	51, 126, ...	80, 85, ...	...	

Der Index funktioniert demnach ganz ähnlich wie ein Stichwortverzeichnis in einem Buch. Zu jedem Begriff gibt er Auskunft darüber, in welchen Dokumenten das Wort vorkommt, wie oft es jeweils auftaucht und an welchen Positionen. In einer zweiten Tabelle können zu jedem Dokument zusätzliche Angaben wie URL, Titel, Datum der letzten Änderung sowie allfällige weitere Informationen ausgelesen werden. Mit diesen Zusatzinformationen kann das Suchsystem eine Rangliste erstellen, ohne die vollständigen Dokumente zu haben.

URL	Titel	Datum	
chocolate.com/mars.html	Mars macht mobil	12. 03. 1998	...
geschichte.de/mars.html	Kriegsgott Mars	29. 05. 1997	...
planets.org/list.html	Planetenübersicht	07. 02. 1999	...
disney.com/comics.html	Mickey's Dog	25. 04. 1998	...
sega.com/consoles.html	MegaDrive, Saturn, ...	15. 12. 1998	...
cars.uk/dealers.html	UK Car Dealers	17. 01. 1999	...
...	...	...	...

Wie arbeitet das Suchsystem mit dem Index?

Man kann sich den Index als eine Sammlung von simplen Ein-Wort-Suchanfragen vorstellen. Ein Hobby-Astronom auf der Suche nach Planeten in unserem Sonnensystem stellt vielleicht die Anfrage *mars*. OMNISEARCH muss daraufhin im Wesentlichen nur die entsprechende Liste aus dem Index zurückliefern: eine Seite über Schokoladeriegel, eine andere über den Planeten Mars und eine dritte mit Informationen zum Kriegsgott mit demselben Namen sowie viele andere Webseiten. Für jedes gefundene Dokument sucht sich OMNISEARCH die Zusatzinformationen aus der zweiten Tabelle heraus und stellt damit

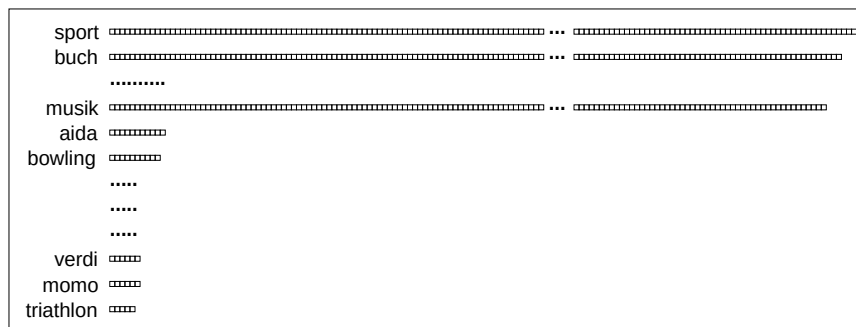
die Rangliste zusammen. Rangierungsprinzip 2 schreibt vor, dass diejenigen Dokumente als die relevantesten betrachtet werden, in denen der Suchbegriff *mars* am häufigsten vorkommt.

1. Mars macht mobil
<http://www.chocolate.com/mars.html>
2. Kriegsgott Mars
<http://www.geschichte.de/mars.html>
3. Alphabetische Planetenübersicht
<http://www.planets.org/list.html>
4. ...
<http://...>

Mit dieser Rangliste ist unser Astronom noch nicht zufrieden, und er überlegt sich dann, wie er die störenden Einträge ohne Bezug zu seinem Thema wegbringen kann. Er versucht es mit mehr als einem Planetennamen: *mars saturn pluto*. In diesem Fall kombiniert das Suchsystem die drei Listen, die zu diesen Begriffen gehören. Rangierungsprinzip 1 sorgt dann dafür, dass diejenigen Seiten in der Rangliste zuerst aufgeführt werden, die alle drei Begriffe enthalten. Im Beispiel ist das die Seite mit der Adresse <http://www.planets.org/list.html>. Die übrigen Dokumente rutschen in der Rangliste auf die hinteren Ränge ab.

Wie können Benutzer die Index-Eigenschaften ausnützen?

Die folgende Grafik zeigt, wie ein konkreter Index aussehen könnte. Jedes kleine Quadrat stellt einen Eintrag dar, das heisst ein Dokument, in dem der entsprechende Begriff mindestens einmal auftaucht.



In einem typischen Index gibt es einige wenige Begriffe, die über eine immense Anzahl an Einträgen verfügen. Diese Begriffe kommen folglich in einem Grossteil aller Dokumente vor. Im Beispiel: *sport*, *musik* und *buch*. Daneben existieren für die weitaus grösste Anzahl der Begriffe nur ganz wenige Einträge. Diese Begriffe treten demnach nur in den wenigsten Dokumenten auf. Im Beispiel: *verdi*, *momo* oder *triathlon*. Diese in der Praxis beobachteten Index-Eigenschaften haben zwei wichtige Konsequenzen:

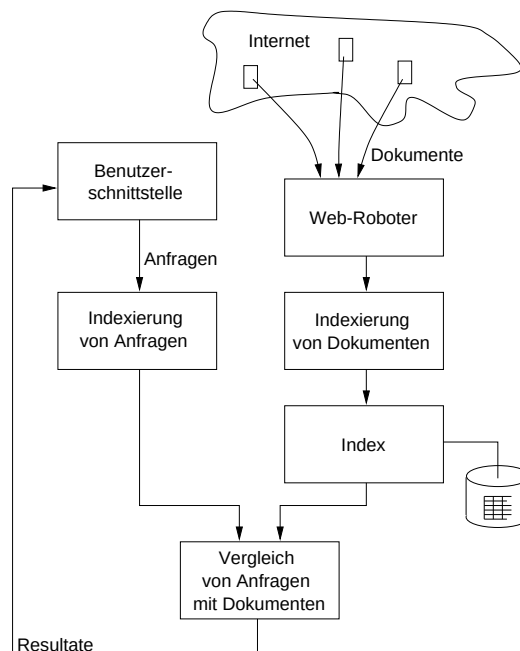
- *Sucheffizienz* beschreibt, wie schnell (effizient) das Suchsystem die Rangliste ermittelt. Für jeden Suchbegriff in der Anfrage muss ein Suchsystem die entsprechende Liste im Index untersuchen. Natürlich braucht das Verarbeiten mehr Zeit für eine lange Liste als für eine kurze. Also: Eine Suchanfrage kann umso schneller beantwortet werden, je spezieller die Suchbegriffe ausfallen.
- *Sucheffektivität* charakterisiert, wie gut (effektiv) das Suchsystem relevante Dokumente findet. Ein Benutzer kann die Sucheffektivität beeinflussen, indem er seltene Suchbegriffe wählt und auf diese Weise eher auf relevante Dokumente stossen sollte. Auch Rangierungsprinzip 3 besagt: Das Auftreten eines seltenen Suchbegriffs führt zu einer höheren Relevanz eines Dokuments.

Zusammenfassend folgt, dass eine ideale Anfrage aus vielen seltenen Suchbegriffen besteht. Damit wird sowohl eine gute Effizienz als auch eine gute Effektivität erzielt.

Puzzle gelöst: das Suchsystem ist komplett

Unterdessen haben wir alle wichtigen Komponenten eines Suchsystems besprochen und können die Funktionsweise zusammenfassend rekapitulieren. Dafür betrachten wir nochmals das Schema aus dem ersten Kapitel und überlegen uns, wie die verschiedenen Bausteine eines Suchsystems zusammenspielen: Der Web-Roboter forscht nach

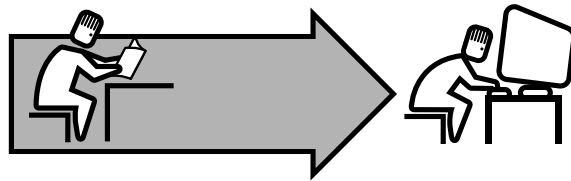
Dokumenten. Bei der Indexierung werden die Begriffe aus den Dokumenten extrahiert und statistische Informationen zusammengestellt. Manche Suchsysteme bieten eine umfangreichere Indexierung an, andere eine eher einfache. Die Informationen zu jedem Dokument werden im Index festgehalten.



Das Suchsystem wartet nun auf Suchaufträge von den Benutzerinnen. Alle Anfragen werden ausschliesslich mit Hilfe des Index ausgewertet. Das heisst, es werden keine Dokumente in Echtzeit untersucht. Der Vorteil: Die Suche erfolgt auf diese Weise viel schneller. Und der Nachteil: Der Index stimmt nicht immer mit der «Realität» überein. Wenn ein Dokument im WWW verändert oder gar gelöscht wird, dauert es eine gewisse Zeit, bis der Web-Roboter die Seite bei einem seiner «Rundgänge» besucht und die Änderung bemerkt. In dieser Zeit spiegelt der Index nach wie vor den ursprünglichen Inhalt der geänderten Seite wider. Für gewisse Anfragen werden deshalb manchmal Dokumente gefunden, die nicht mehr existieren oder rein gar nichts mehr mit den Suchbegriffen zu tun haben.

Sobald eine Anfrage mit Hilfe der Informationen im Index bearbeitet wurde, stellt das Suchsystem eine Rangliste mit den gefundenen, nach absteigenden Relevanzwerten sortierten Dokumenten zusammen. Im letzten Schritt wird die Rangliste der Benutzerin präsentiert.

Damit ist die Theorie im Zusammenhang von Web-Roboter und Index eines Suchsystems abgehandelt. Im Folgenden gehen wir auf einige praktische Aspekte und die Metasuchdienste ein ...



Wie landen Webseiten im Index?

Passive Variante: Ein Webmaster macht eine neue Webseite im World Wide Web zugänglich. *Mindestens eine* andere Seite enthält einen Hyperlink auf die neue Seite. Zudem befindet sich die Seite mit dem Verweis bereits im Index des Suchdienstes. Unter diesen Voraussetzungen kann der Webmaster einfach abwarten. Früher oder später wird der Web-Roboter automatisch über den bestehenden Hyperlink zur neuen Seite gelangen und sie in den Index aufnehmen. Unter Umständen dauert es allerdings mehrere Wochen, bis der Roboter auf die Seite stösst.

Aktive Variante: Die meisten Suchdienste bieten die Möglichkeit, Seiten direkt über eine Funktion wie *Add Page* oder ähnlich anzumelden. Damit kann man den Web-Roboter zum Besuch der neuen Seite auffordern, und das Dokument sollte rascher im Index erscheinen.

Decken reale Suchdienste das ganze Web ab?

Viele der Suchdienste mit horizontalen Dokumentensammlungen wie OMNISEARCH versuchen, einen möglichst lückenlosen Querschnitt durch alle Themengebiete im Internet anzubieten. Gibt es einen Suchdienst, der ausnahmslos alle Dokumente im Index bereitstellt? Die Antwort: Nein! Es gibt immer Seiten, die sich dem Zugriff durch einen Web-Roboter entziehen. Mögliche Gründe werden wir nachfolgend erläutern. Zunächst aber eine wichtige Konsequenz aus dieser Feststellung: Wir müssen davon ausgehen, dass unterschiedliche Suchdienste auch über einen unterschiedlichen Index verfügen. Das heisst: Webseiten, die mit dem einen Suchdienst auffindbar sind, können mit einem anderen nicht gefunden werden. Also kann man durchaus versuchsweise den Suchdienst wechseln, falls man bei einer Recherche ganz und gar nicht weiterkommt.

Wieso erscheint eine Seite nicht im Index?

Es folgt eine Liste mit den wichtigsten Gründen, weshalb ein Dokument nicht im Index eines Suchdienstes auftaucht.

Isolierte Seiten

Seiten ohne Verweis von einem anderen Dokument und ohne explizite Anmeldung bei einem Suchdienst tauchen nicht im Index auf.

Dynamische Seiten

Viele Webseiten sind statischer Natur. Diese statischen Webseiten sind als Dateien gespeichert und bleiben unverändert, bis jemand die Dateien bearbeitet. Es gibt aber auch dynamische Webseiten, die bei jedem Aufruf neu erstellt werden und normalerweise jedes Mal einen neuen Inhalt aufweisen. Viele Web-Roboter ignorieren dynamische Webseiten, weil sie die dahinter liegenden Programme nicht ansprechen können oder weil sie sich die Sisyphusarbeit bei diesen stets ändernden Seiten ersparen möchten. Im ersten Kapitel wurde ein Programm besprochen, das bei jedem Aufruf die aktuelle Uhrzeit

liefert. Der Web-Roboter würde folglich bei jedem Besuch eine neue Seite vorfinden und immer wieder im Index aktualisieren.

Robot Exclusion

Die Betreiberinnen von Web-Servern können Web-Robotern den Zugriff zu ihren Seiten verwehren. Der *Robot-Exclusion*-Standard legt fest, wie dabei vorzugehen ist. Der Standard ist nicht verbindlich, doch die meisten Roboter halten sich daran. Das gehört zum Internet-Knigge.

Aber wieso sollte man überhaupt einen Suchdienst daran hindern wollen, bestimmte Seiten anzubieten? Beispiel 1: Ein Roboter fällt durch schlechtes Benehmen auf, indem er auf eine Webseite viel zu häufig zugreift. Beispiel 2: Manche Web-Sites bieten neben den öffentlichen auch private Bereiche an. Nur befugte Personen können die privaten Seiten anschauen. Also lässt man auch Web-Roboter nur im öffentlichen Bereich zu.

Mangelnde Qualität

Dokumente können von einem Suchdienst ignoriert werden, weil sie bestimmte Qualitätsmerkmale nicht erfüllen. Vielleicht ist die Rechtschreibung mangelhaft, oder die Seite wurde zu lange nicht mehr aktualisiert. Es gibt auch Suchsysteme, die Webseiten mit problematischem Inhalt (zum Beispiel Pornografie, übertriebene Gewalt oder politisch extreme Meinungen) nicht berücksichtigen. Dabei werden beispielsweise Dokumente gesperrt, die Problembegriffe in einer bestimmten Häufigkeit und Kombination enthalten.

Beschränkung durch Suchdienst

Gewisse Web-Roboter legen es gar nicht erst darauf an, tatsächlich alle Webseiten zu finden. Stattdessen werden ihnen von den Betreibern bestimmte Beschränkungen vorgegeben. Es gibt zwei gängige Methoden: (1) Der Web-Roboter darf – ausgehend von einer Homepage – höchstens bis zu einer maximalen Linktiefe (zum Beispiel drei) weitere Seiten in den Index aufnehmen. (2) Es kann auch vorgeschrieben werden, dass von einer Web-Site nur eine bestimmte Höchstzahl

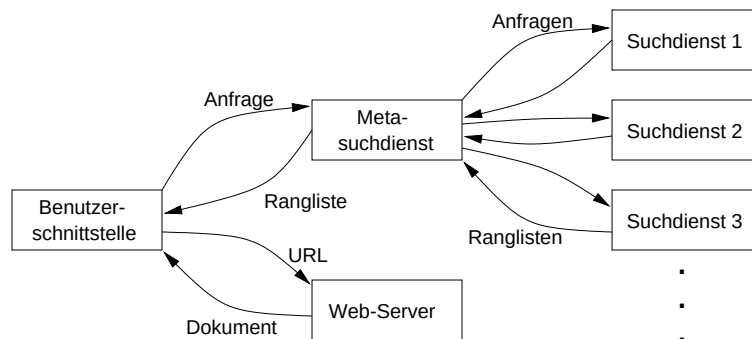
von Webseiten in den Index aufgenommen werden darf. Sobald solche Maximalwerte erreicht sind, werden keine weiteren Seiten von der entsprechenden Web-Site akzeptiert.

Einem Suchdienst ist es übrigens auch freigestellt, wie vollständig er die gefundenen Seiten indexiert. Es gibt Suchdienste, die nur den Titel oder die ersten paar Zeilen eines Dokuments in den Index aufnehmen. Andere Systeme indexieren den vollständigen Text. Entsprechende Hinweise finden sich üblicherweise in den Hilfeseiten.

Metasuchdienste

Nun kennen wir alle wichtigen Komponenten eines Suchsystems und verstehen die verschiedenen Problemarten. Deshalb gehen wir jetzt einen Schritt weiter und betrachten die Metasuchdienste. Metasuchdienste sehen auf den ersten Blick sehr ähnlich aus wie «normale» Suchdienste. Hinter den Kulissen jedoch läuft ein anderer Mechanismus ab. Metasuchdienste arbeiten ähnlich wie Parasiten, indem sie die Angebote bestehender Suchdienste ausnützen, um Suchanfragen zu beantworten. Metasuchdienste lassen somit die anderen Dienste die meisten Arbeiten erledigen. Eine Recherche mit einem Metasuchdienst läuft folgendermassen ab:

1. Die Benutzerin schickt mit Hilfe der Benutzerschnittstelle eine Suchanfrage an den Metasuchdienst.
2. Der Metasuchdienst leitet die Anfrage an eine Serie von Suchdiensten (und vielleicht auch Katalogdiensten) weiter. Manchmal muss die Anfrage dazu erst in das jeweilige Format übersetzt werden.
3. Nun wartet der Metasuchdienst auf Antwort von den angefragten Suchdiensten. Die einzelnen Ranglisten werden bei deren Ankunft gesammelt. Falls die Antwort von einem der Suchdienste über längere Zeit ausbleibt, werden seine Resultate in der Regel schlicht ignoriert.
4. Zum Schluss werden die verschiedenen Ranglisten miteinander kombiniert und der Benutzerin präsentiert.



Vorteile

Metasuchdienste erschliessen die Dokumentensammlungen von verschiedenen Suchdiensten gleichzeitig. Das ist vorteilhaft, wenn man mit einer Recherche bei einem bestimmten Suchdienst nicht weiterkommt. Oder wenn man zu Beginn einer Recherche abschätzen möchte, welcher Dienst wohl die viel versprechendsten Dokumente zum Thema liefern könnte. Natürlich ist man mit einem Metasuchdienst schneller, als wenn man manuell die verschiedenen Suchdienste einzeln konsultieren würde.

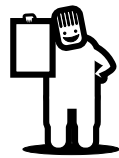
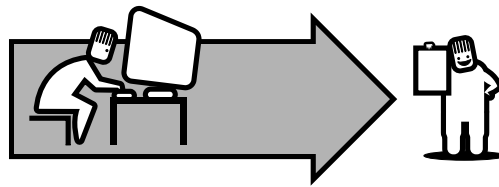
Nachteile und Probleme

Ein Metasuchdienst kann nur diejenige Funktionalität anbieten, die von *allen* angesprochenen Suchdiensten ebenfalls unterstützt wird. Folglich bieten Metasuchdienste einen reduzierten Funktionsumfang an, und von erweiterten Möglichkeiten kann nur selten Gebrauch gemacht werden.

Ein anderes bedeutendes Problem betrifft die Rangliste. Aufgabe des Metasuchdienstes ist es, die Ranglisten der verschiedenen Suchdienste zu einer einzigen zu verschmelzen. Die Schwierigkeit: Die Suchdienste liefern die gleichen Dokumente, aber in unterschiedlicher Reihenfolge. Da die Relevanzwerte nicht normiert sind, wird häufig der Durchschnittsrang ermittelt. Viel schwieriger wird es jedoch, wenn verschiedene Suchdienste verschiedene Dokumente liefern. Der

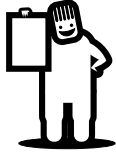
eine Suchdienst ermittelt beispielsweise 20 sehr relevante Dokumente, während der andere nur irrelevante findet. Das Zusammenfügen mehrerer Ranglisten zu einer einzigen unter diesen Bedingungen ist eine schwierige Aufgabe, die nicht perfekt gelöst werden kann. Deshalb ziehen sich viele Metasuchdienste aus der Affäre, indem sie die Resultate nach Suchdiensten getrennt präsentieren.

Wie immer zum Abschluss eines Kapitels folgen nun die Lösungen zu den Anwenderproblemen ...



Der Index ist offensichtlich das Kernstück eines Suchsystems. Dank dem Index und einigen zusätzlichen Daten über das Dokument wie Titel und Datum kann das System eine Anfrage innert Sekundenbruchteilen beantworten und muss nicht jedes Mal alle Webseiten beziehen und untersuchen. Zu diesem Zweck hält der Index den Inhalt der Webseiten in geeigneter Form fest. Das führt aber auch zu Problemen. Ein Suchdienst kann nicht ununterbrochen jede Webseite auf Änderungen hin überprüfen. Darum kommt es manchmal vor, dass die im Index gespeicherte Information nicht mit der entsprechenden Webseite übereinstimmt.

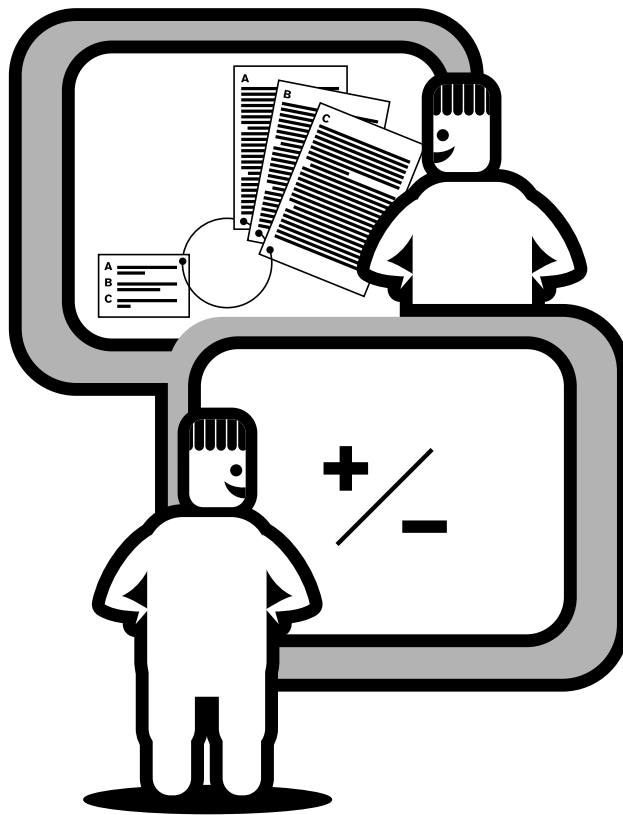
Im Fall des Antarktis-Dokuments beispielsweise wurde offenbar das Dokument vom Netz genommen. OMNISEARCH hat das Fehlen der Seite allerdings noch nicht bemerkt und findet im Index nach wie vor die auf der Webseite verwendeten Begriffe. Also taucht die Seite in der Rangliste auf, doch beim Besuch der Seite werde ich mit einer Fehlermeldung belohnt.



Das Problem mit den Fussballresultaten hat den gleichen Grund. Anscheinend bietet die gefundene englische Zeitung regelmässig – vielleicht täglich – die neuesten Sportschlagzeilen in der HTML-Datei **sports/headline.html** an. Wenn der Web-Roboter von OMNISEARCH die Seite besucht, übernimmt er den jeweiligen Artikel in den Index. Doch schon am nächsten Tag wird der Artikel bei der Zeitung durch eine neue Schlagzeile ersetzt. Deshalb habe ich die Adresse des Dokuments mit meinen Suchbegriffen gefunden, und trotzdem hatte der neue Inhalt nichts mehr mit meiner Anfrage zu tun.

Kapitel 5

Metadaten und Boole'sche Suchmethoden





Ich bin gerade auf der Suche nach der Homepage des WWF, oder ausgeschrieben «World Wildlife Fund». Erfahrungsge-
mäss finde ich die Hauptseiten von derart bekannten Orga-
nisationen ohne grosse Probleme. Häufig findet ein Such-
dienst mit einem einzigen Stichwort die richtige Seite auf
Antrieb. Also benutze ich OMNISEARCH und starte auf gut Glück die
Anfrage *WWF*.

Doch hier mache ich eine überraschende Entdeckung. Die Ab-
kürzung WWF steht auch für «World Wrestling Federation». Und
die Wrestler haben ihre Seiten offenbar für OMNISEARCH optimiert,
denn die Rangliste ist überfüllt mit Seiten von ihrem Web-Server mit
der Adresse www.wwf.com. Ich schaue einen Grossteil der Rangliste
an, doch weit und breit finde ich keine Spur von den Wildtieren. Was
tun?



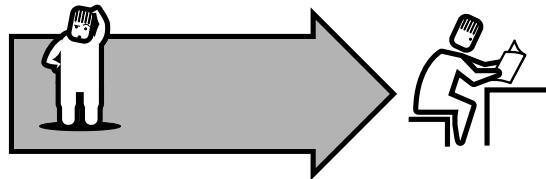
Ich bereite ein kurzes Referat zum Tourismus im deutsch-
sprachigen Raum vor. Dabei möchte ich die Übernachtungszahlen im Gastgewerbe in Österreich, Deutschland
und der Schweiz mit den jeweiligen Einwohnerzahlen ver-
gleichen. Die Einwohnerzahlen sind kein Problem. Die finde
ich zum Beispiel in einem guten Atlas. Aber bei den Übernachtungs-
zahlen hatte ich bisher kein Glück. Auch die Suche im Internet blieb
erfolglos. Immer habe ich viel zu viele nicht relevante Dokumente er-
halten, obwohl ich es mit etlichen Anfragevarianten versucht habe.
Wie gehe ich am besten vor?



Letztes Jahr war ich in der Toskana in den Ferien. Ich hat-
te ein Zimmer in einem Hotel, das mir sehr gut gefallen
hat. Es war das Hotel Aurora in Castiglione della Pescaia.
Das Hotel habe ich nun einem Freund empfohlen und will
ihm die Adresse der zugehörigen Webseite liefern. Ich weiss,
dass die Webseite existiert. Trotzdem kann ich sie einfach nicht fin-
den, obwohl ich sogar fortgeschrittene Techniken angewendet habe.
Diese Techniken wurden mir von einem Bekannten empfohlen. Mit
Hilfe der so genannten Boole'schen Operatoren kann ich Suchbegrif-
fe erzwingen oder zu Phrasen kombinieren. Meine Anfrage: *Aurora*
AND Hotel AND "Castiglione della Pescaia". Damit verlange ich,
dass alle drei Suchbegriffe zwingend in allen gefundenen Dokumen-

ten vorkommen müssen. Ausserdem sollen die Begriffe «Castiglione della Pescaia» exakt in dieser Reihenfolge auftauchen. Aber trotz der eindeutigen Suchbegriffe bleibt meine Recherche erfolglos. Wieso?

Bevor die angesprochenen Probleme gelöst werden können, müssen wir einige weitere Aspekte der Informationssuche kennen lernen ...



Boole'sche Operatoren

Eine unserer Anwenderinnen hat bereits von den Boole'schen Operatoren Gebrauch gemacht. Die wichtigsten Boole'schen Operatoren sind AND, OR und AND NOT. Mit Hilfe der Boole'schen Operatoren lässt sich in einer Suchanfrage eindeutig und zwingend festlegen, welche Bedingungen die gefundenen Dokumente erfüllen müssen.

Mittels AND werden zwei Suchbegriffe verknüpft, wenn beide Begriffe zwingend im Dokument auftauchen müssen. Bei OR genügt einer der beiden Suchbegriffe. Das ist beispielsweise bei Synonymen nützlich. Und mittels AND NOT werden sämtliche Dokumente ausgeschlossen, die den entsprechenden Suchbegriff enthalten. Einige Beispiele: Die Boole'sche Anfrage *Köln AND Dom* findet nur Dokumente, die sowohl den Begriff «Köln» als auch den Begriff «Dom» enthalten. Bei *Inuit OR Eskimo* müssen die gefundenen Dokumente mindestens einen der beiden Begriffe enthalten. Die Anfrage *chocolate AND NOT swiss* findet Dokumente mit dem Begriff «chocolate» und schliesst Dokumente mit dem Begriff «swiss» aus.

Boole'sche Suche im Dokumentinhalt

Bei den obigen Beispielen wurden Boole'sche Operatoren auf Begriffe aus dem Inhalt eines Dokuments angewendet. Das kann gelegentlich

hilfreich, oft aber auch gefährlich sein! Ein Beispiel: Mit der Anfrage *birth AND death AND rate AND italy* werden alle Dokumente gefunden, die genau die vier Begriffe enthalten.

Das Problem: Boole'sche Operationen verlangen exakte Übereinstimmung zwischen den Suchbegriffen der Anfrage und den Begriffen im Dokument. Die obige Anfrage beispielsweise findet keine Dokumente, in denen von der «mortality rate» anstelle der «death rate» die Rede ist. Natürlich kann man das Problem mit der Anfrage *birth AND (death OR mortality) AND rate AND italy* umschiffen. Allerdings ist es schwierig, immer an alle möglichen Synonyme zu denken. Folglich besteht die Gefahr, dass man mit einer Boole'schen Anfrage unbewusst relevante Dokumente ausschliesst.

Noch deutlicher wird die Problematik bei Verwendung der Boole'schen Operation *AND NOT*. Beispiel: Jemand möchte sich über nicht amerikanische Raumfahrtprogramme informieren und stellt die Anfrage *Raumfahrt AND NOT NASA*. Damit werden auch relevante Dokumente ausgeschlossen. Zum Beispiel ein Dokument, in dem es um die europäische Raumfahrtbehörde ESA geht und das auch kurz auf die Zusammenarbeit mit der NASA eingeht.

Nur weil ein Dokument einen nicht relevanten Abschnitt enthält, heisst das nicht, dass das ganze Dokument nicht relevant ist. Man kann sich eine Zeitungsseite mit vielleicht einem Dutzend Artikel vorstellen. Die meisten der Artikel sind für eine bestimmte Anfrage nicht relevant. Das ändert aber nichts an der Relevanz eines interessanten Berichts, der sich ebenfalls auf der Seite befindet. Anstatt im NASA-Beispiel den Begriff *NASA* strikt auszuschliessen, verzichtet man häufig besser auf die Boole'sche Suche im Dokumentinhalt und fügt stattdessen zusätzliche charakteristische Begriffe zur Anfrage hinzu: *Raumfahrt ESA CSA NASDA INPE RKA CNSA*.

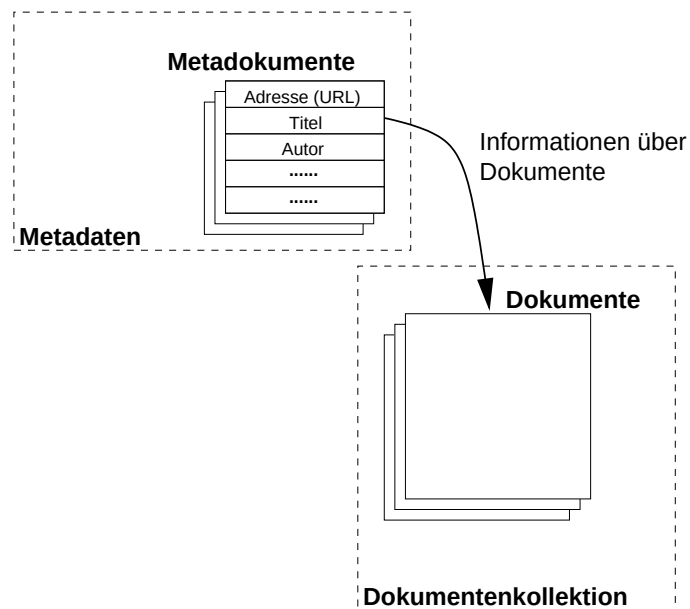
Mit Hilfe der Boole'schen Techniken im Dokumentinhalt wird das Ziel verfolgt, die Rangliste möglichst freizuhalten von nicht relevanten Dokumenten. Wie erwähnt läuft man dabei aber Gefahr, auch relevante Dokumente auszuschliessen. In der Regel ist es weniger problematisch, einige nicht relevante Dokumente in der Rangliste vorzufinden. Solche Dokumente lassen sich einfach ignorieren. Problematischer ist der umgekehrte Fall: Es werden Dokumente ausgeschlossen,

die an sich relevant sind. Man denke beispielsweise an den Patentanwalt. Er ist darauf angewiesen, möglichst alle relevanten Dokumente zu finden und keine Patentanmeldung zu übersehen. Oder ein Arzt möchte alle Dokumente zu einem bestimmten Medikament finden, damit er nicht die allfällige Notiz über eine gefährliche Nebenwirkung verpasst.

Trotz der erwähnten Schwierigkeiten mit den Boole'schen Operatoren im Dokumentinhalt können diese Techniken im Zusammenhang mit Metadaten durchaus Gewinn bringend eingesetzt werden.

Metadaten

Bisher haben wir uns in erster Linie mit Dokumenten beschäftigt, die gewisse Sachverhalte, Fakten, Ereignisse und Ähnliches beschreiben. Daneben gibt es auch Dokumente, die andere Dokumente in einer strukturierten und kompakten Form beschreiben. Es sind dies die so genannten *Metadokumente*. Alle Metadokumente zusammen bilden die *Metadaten*. Die Metadaten enthalten Informationen *über* die Dokumente in einer Kollektion.



Metadaten kommen auch im Alltag zum Einsatz: Der Katalog einer Bücherei beispielsweise ist eine Sammlung von Metadaten. Man kann darin nach Autorenangaben suchen, nach dem Erscheinungsdatum, nach der ISBN-Nummer usw. Metadaten können aber auch Deskriptoren oder Schlagwörter enthalten, die den Inhalt eines Dokuments beschreiben.

Zu Webseiten können ebenfalls Metadaten bereitgestellt werden. Offensichtliche Metadaten für Webseiten sind die Adressen (URLs) der Seiten, eine Inhaltszusammenfassung, der Titel, eine Taxierung des Inhalts (zum Beispiel eine Altersfreigabe) oder das Datum der letzten Inhaltsänderung. In der Praxis werden Metadaten durch eigens dafür geschaffene Beschreibungssprachen definiert. Ein verbreitetes Beispiel für eine solche Sprache ist XML.

Man unterscheidet zwei wichtige Arten von Metadaten:

- *Normalisierte Metadaten* erfüllen Regeln, die eine einheitliche Darstellung und exakte Vergleiche erlauben. Zu den normalisierten Metadaten gehören beispielsweise vierziffrige Jahreszahlen, die nur auf eine einzige Art und Weise geschrieben werden können. So kann man aus dem Vergleich der Zeichenketten «1999» und «2000» schliessen, dass es sich um unterschiedliche Jahreszahlen handelt. Normalisierte Metadaten entstehen auch, wenn ein standardisiertes, kontrolliertes Vokabular verwendet wird. Dann dürfen – zum Beispiel für die Zusammenfassung eines Texts – nur Begriffe aus diesem Vokabular verwendet werden. Begriffe dieser Art werden Schlagwörter genannt.
- *Nicht normalisierte Metadaten* sind nicht eindeutig. Das ist beispielsweise bei URLs der Fall. Ein URL wie `http://www.xyz.com/` kann auch ohne das führende `http://` oder den abschliessenden `/` in einer Webseite auftauchen. Zudem können unterschiedliche URLs auf ein und dieselbe Webseite verweisen.

Ein sehr wichtiges Beispiel für nicht normalisierte Daten ist normaler Text. Ein Sachverhalt kann auf unterschiedliche Arten formuliert werden. Für viele Begriffe gibt es zahlreiche alternative Bezeichnungen und Synonyme.

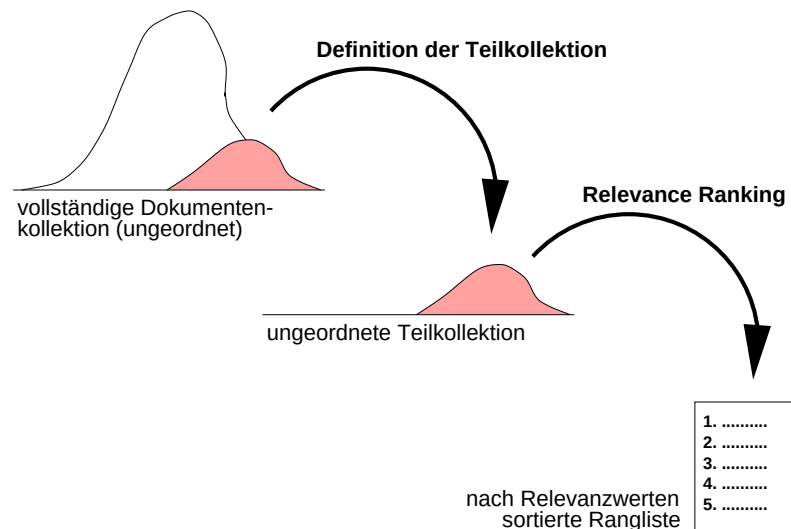
Auch Datumsangaben sind nicht eindeutig. Der amerikanische Nationalfeiertag kann auf viele verschiedene Arten notiert werden: 4. 7., 07/04, 04. 07., 4. Juli, 4th of July usw. Das Problem dabei: Die beiden Zeichenketten «07/04» und «4. Juli» stimmen nicht überein. Daraus kann jedoch nicht geschlossen werden, dass es sich um unterschiedliche Datumsangaben handelt.

Suche in Teilkollektionen mittels Metadaten

Metadaten kann man für die Informationssuche ausnützen, denn mit Hilfe der Metadaten lassen sich *Teildokumentenkollektionen* (kurz: Teilkollektionen) definieren, die dann gezielt durchsucht werden können. Der Vorteil dabei ist, dass man die Suche auf einen Teilbereich innerhalb einer grösseren Menge von Dokumenten beschränkt. Wer zum Beispiel nach amerikanischen Universitäten sucht, die im Bereich der Robotik tätig sind, beschränkt die Suche vorzugsweise auf Webseiten innerhalb der Domain «edu».

Die Suche in Teilkollektionen läuft in zwei Schritten ab:

1. *Definition der Teilkollektion:* Mit geeigneten Hilfsmitteln wird eine Teilmenge der kompletten Dokumentenkollektion festgelegt. Dazu werden die jeweils verfügbaren Metadaten verwendet. Resultat: eine ungeordnete Menge von Dokumenten, die alle dieselben Eigenschaften erfüllen. Beispiel: alle Dokumente, die auf einem Web-Server in der Domain «edu» lagern.
2. *Relevance Ranking:* Nun steht ein ungeordneter Haufen mit Dokumenten zur Verfügung. Über den Inhalt der Dokumente wurde noch nichts ausgesagt. Deshalb wird im zweiten Schritt mit Hilfe von einigen Suchbegriffen Ordnung in die Teilkollektion gebracht. Das geschieht aufgrund der üblichen Rangierungsprinzipien. Resultat: Eine Rangliste bestehend aus den Dokumenten der Teilkollektion, nach Relevanzwerten geordnet. Im Beispiel: Die aus der Domain «edu» stammenden Dokumente werden nach den Suchbegriffen *robotics*, *research*, *engineering*, *laboratory* und *department* rangiert.



Das Vorgehen ist auch im Alltag bekannt: Ein Kunde betritt eine Buchhandlung und fragt nach allen amerikanischen Sciencefiction-Büchern der letzten drei Wochen. Allerdings möchte er nichts von William Gibson, denn von ihm kennt er schon alles. Und am liebsten hätte er Bücher, in denen grüne Ausserirdische vorkommen.

Was macht der Verkäufer? Zunächst definiert er die Teilkollektion. Das heisst, er sucht alle Bücher heraus, welche die gestellten Bedingungen erfüllen: Herkunft USA, Genre Sciencefiction, Datum jünger als drei Wochen, Autorennamen nicht William Gibson. Das lässt sich mit Hilfe von Metadaten im Bücherkatalog erledigen, ohne je ein Buch zur Hand zu nehmen.

Es folgt das «Relevance Ranking». Der Verkäufer trägt die Bücher der Teilkollektion zusammen. Dann studiert er überall die Zusammenfassung auf dem Buchrücken und blättert vielleicht auch kurz durch die Seiten, um Hinweise auf den Inhalt der Bücher zu erhalten. Schliesslich präsentiert er dem Kunden zuerst alle Bücher mit grünen Ausserirdischen, danach diejenigen mit andersfarbigen Ausserirdischen und zum Schluss auch noch die restlichen Bücher.

Aber auch in anderen Bereichen nützt man immer wieder die Möglichkeit zur Bildung von «Teilkollektionen» aus: Beim Schuhkauf trifft man eine Vorauswahl aufgrund von Schuhgrösse, Material, Preis

und anderen «Metadaten». Anschliessend wird in dieser Teilmenge ein ansprechendes Modell ausgewählt.

Ein weiteres Beispiel: Gesucht wird nach einem Schuhregal. Anstatt in einem grossen Möbelgeschäft jedes einzelne Möbelstück anzuschauen, schränkt man die Suche besser ein. Zunächst wird das richtige Stockwerk gewählt. Dort grenzt man die Suche auf die richtige Abteilung ein und sucht schliesslich ein passendes Regal aus.

Voraussetzungen für die Definition von Teilkollektionen

Es kann sehr nützlich sein, eine Recherche auf eine Teilkollektion zu beschränken, anstatt die komplette Dokumentenkollektion zu durchsuchen. Es ist aber darauf zu achten, dass die definierte Teilkollektion möglichst alle relevanten Dokumente enthält. Andernfalls werden relevante Dokumente unbeabsichtigt und unbemerkt ausgeschlossen. Beispiel: Eine Benutzerin ist auf der Suche nach einer Firma in Liechtenstein. Sie schränkt die Suche auf die Teilkollektion aller Web-Server im Fürstentum Liechtenstein mit der Länderkennung «li» ein. Unter Umständen schliesst die Benutzerin so just die relevanten Webseiten aus, weil die gesuchte Firma innerhalb der internationalen «com»-Domain angemeldet ist.

Die Einschränkung der Suche auf eine Teilkollektion ist dann problemlos möglich, wenn drei Bedingungen erfüllt sind:

- Für die Definition der Teilkollektion werden *normalisierte Metadaten* verwendet. Damit wird gewährleistet, dass die Einschränkung eindeutig ist. Beispiel: Eine Journalistin sucht in einem Zeitungsarchiv nach Informationen im Zusammenhang mit dem Fall der Berliner Mauer. Die Zeitungsartikel sind mit einer vierziffrigen Jahreszahl versehen. Also kann die Journalistin die Suche auf alle Artikel seit 1989 einschränken.
- Die für die Einschränkung verwendeten Metadaten müssen für sämtliche Dokumente in der Kollektion *vollständig erfasst* sein. Im Beispiel bedeutet das: Ausnahmslos jeder Artikel im Zeitungsarchiv muss mit einer Jahreszahl versehen sein, sonst kann die Einschränkung vom System nicht exakt durchgeführt werden.

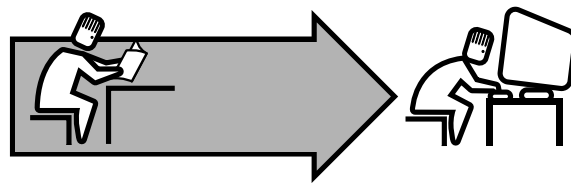
- Die gewählte Einschränkung muss *zwingend* für alle relevanten Dokumente gelten. Nur so ist es möglich, eine Teilkollektion mit allen relevanten Dokumenten zu erstellen. Im Beispiel verlangt die Journalistin, dass die vom System gelieferten Dokumente jünger als 1989 sind. Mit dieser Einschränkung werden allenfalls relevante Dokumente ausgeschlossen, die über die Ereignisse im Vorfeld zur Wiedervereinigung Deutschlands mit der DDR berichten.

Natürlich kann man die Einschränkung einer Recherche auf eine Teilkollektion auch durchführen, wenn nicht alle drei Bedingungen erfüllt sind. Es besteht dann aber die Gefahr, dass man nicht alle relevanten Dokumente berücksichtigt.

Die am Anfang des Kapitels vorgestellten Boole'schen Suchmethoden eignen sich zur Definition von Teilkollektionen innerhalb einer vollumfänglichen Dokumentenkollektion aufgrund von Metadaten. Mit Hilfe der Operatoren AND, OR und AND NOT lassen sich Teilmengen konstruieren und miteinander kombinieren, bis die gewünschte Teilkollektion entstanden ist.

Mit der Operation AND kann eine Teilkollektion immer stärker eingeschränkt werden. Alle mit AND verknüpften Bedingungen müssen zwingend zutreffen. Mit Hilfe von OR lassen sich Alternativen angeben. Es genügt, wenn Dokumente eine einzige der mit OR verknüpften Bedingungen erfüllen. Die Operation AND NOT schliesslich dient zum Ausschluss von Dokumenten mit bestimmten Eigenschaften. Mit Hilfe der Klammern werden Boole'sche Ausdrücke zu Einheiten zusammengefasst.

Nach diesen theoretischen Betrachtungen wird es Zeit für die praktischen Hinweise zur Suche in Teilkollektionen ...



Definition von Teilkollektionen in der Praxis

Wie kann man die Metadaten ansprechen, und wie kann man Bedingungen für die Dokumente in der gewünschten Teilkollektion festlegen? Suchdienste unterscheiden sich oft in beiden Fragen stark. Bei manchen Suchdiensten können die Metadaten von Dokumenten gar nicht angesprochen werden. Bei anderen wiederum kann man auch das letzte Detail festlegen. Auch bei den Boole'schen Operationen ist das Spektrum gross: Bei einigen Systemen wird der volle Funktionsumfang angeboten, andere unterstützen vereinfachte Operationen, wieder andere realisieren Boole'sche Operationen mit Hilfe von umfangreichen Formularen.

Erschwert wird die Angelegenheit dadurch, dass sich von Suchdienst zu Suchdienst sowohl die Syntax der Boole'schen Operatoren als auch die Struktur der Metadaten ändern kann. Als Benutzer eines Suchdienstes muss man sich deshalb auf alles gefasst machen. Man kommt nicht darum herum, die Hilfeseiten anzuschauen.

OmniSearch

Der Suchdienst OMNISEARCH bietet Boole'sche Operationen in vereinfachter Form und einige Möglichkeiten für den Zugriff auf Metadaten an. Das `+` unmittelbar vor einem Suchbegriff verlangt, dass alle Dokumente in der Teilkollektion das entsprechende Kriterium erfüllen. Das `-` bewirkt das Gegenteil und schliesst ein Kriterium aus. Normale Suchbegriffe ohne ein solches Vorzeichen werden für das Relevance Ranking verwendet.

So kann bei OMNISEARCH beispielsweise mit der Anfrage `+domain:gov declaration independence` eine Teilkollektion aller Webseiten bei Regierungsbehörden der USA (Domain «gov») gebildet werden. Innerhalb dieser Teilkollektion wird anschliessend gemäss der Suchbegriffe *declaration* und *independence* eine Rangliste erstellt.

Oder mit der Anfrage `-server:www.netscape.com netscape navigator browser` lässt sich nach Informationen über Netscapes Web-Browser suchen, die nicht vom Hauptserver des Herstellers stammen.

NewsSeeker

Bei NEWSSEEKER werden keine Boole'schen Operationen für die Definition von Teilkollektionen angeboten. Stattdessen wird den Benutzerinnen ein Formular präsentiert, in dem verschiedene Bereiche angewählt werden können; beispielsweise Wirtschaftsmeldungen, Katastrophenmeldungen, Politik usw. Nach dem Aktivieren wird ausschliesslich in den entsprechenden Teilkollektionen gesucht.

News-Artikel

Im Internet gibt es neben den bestens bekannten Diensten E-Mail und WWW eine Reihe weiterer Dienste. Einer davon sind die so genannten *News*. Das System funktioniert nach dem Prinzip der öffentlichen Anschlagbretter. Beliebige Benutzerinnen können in zahlreichen Foren zu den verschiedensten Themen bestehende News-Artikel lesen oder eigene Artikel veröffentlichen.

Es existieren spezielle Suchdienste, mit deren Hilfe Benutzerinnen die gesammelten News-Artikel in allen Foren durchsuchen können. Solche Suchdienste lassen üblicherweise die Einschränkung auf Teilkollektionen mittels Metadaten zu. Allerdings erfolgt die Einschränkung häufig nicht mit Hilfe Boole'scher Operatoren, sondern über ein Formular. In diesem Formular werden die einzelnen Eingabefelder nach den eigenen Wünschen ausgefüllt.

Suchbegriffe	waschmaschine	
Forum	ch.market	
Titel		
Autor		
Datum (TT MM JJJJ)	01 06 1999	01 09 1999
von		bis

Hier suchte offenbar jemand im September 1999 nach einer billigen Waschmaschine. Er versuchte sein Glück in einer Schweizer Newsgroup für An- und Verkäufe verschiedenster Art. Für die Anfrage sind die Metadaten in diesem Fall sehr hilfreich. Die Suche wird auf News-Artikel des Forums **ch.market** beschränkt. Zudem wird mittels der Angaben zum Datum verlangt, dass die Artikel nicht älter als drei Monate sind. Die Felder zum Titel des Artikels und zum Autor sind leer und bleiben somit uneingeschränkt. Als Suchbegriff wird einfach *waschmaschine* gewählt.

Vorsicht: Boole'sche Suche im Dokumentinhalt

Die Boole'schen Suchmethoden sind nützlich für die Definition von Teilkollektionen, wenn die Teildokumentenmenge aufgrund von eindeutigen Kriterien innerhalb der Metadaten definiert werden kann. Problematisch ist die Boole'sche Suche im Dokumentinhalt. Zum Abschluss folgen drei Beispiele, welche die Gefahren und mögliche Lösungen nochmals aufzeigen sollen.

Apfelkuchen für Walnussallergiker

Ein Dessert-Fan ist auf der Suche nach einem tollen Rezept für Apfelkuchen ohne Walnüsse, denn auf Walnüsse ist er allergisch. Er benützt bei OMNISEARCH die Boole'schen Operationen, um alle Walnüsse loszuwerden. Die Anfrage lautet *+apfelkuchen rezept -walnüsse*. In der Rangliste tauchen folglich nur Dokumente auf, die den Begriff «Apfelkuchen» zwingend enthalten und ganz sicher keine Dokumente, in denen das Wort «Walnüsse» vorkommt. Der Suchbegriff *rezept* bleibt ohne Vorzeichen. Das heisst: Kommt zusätzlich in einem Dokument der Begriff «Rezept» vor, so soll es umso relevanter gemeldet werden. Mit der Anfrage kommt der Dessert-Fan vermutlich schadlos zum Ziel. Doch die Sache hat einen Haken.

Boole'sche Operationen kennen keine Gnade! Erfüllt ein Dokument die gewünschten Bedingungen nicht, so wird es in der Rangliste nicht erscheinen. Oft schränkt man die Dokumentenmenge zu stark ein und schliesst relevante Dokumente aus. Und man merkt es nicht

einmal, weil man das relevante Dokument natürlich gar nie zu Gesicht bekommt.

Im obigen Beispiel zum Thema Walnussallergiker und Apfelkuchen: Stellen wir uns vor, es existiere das perfekte Dokument für das Informationsbedürfnis des Dessert-Fans. Der Titel: «Extra für alle Allergiker: Ein Rezept für Apfelkuchen ohne Walnüsse». Mit der oben verwendeten Anfrage wäre das Dokument nicht in der Rangliste erschienen.

In der Regel ist es deshalb angebracht, auf die Boole'schen Operationen im Dokumentinhalt zu verzichten. Meistens kann man das Relevance Ranking ausnützen oder Einschränkungen mit Hilfe der Metadaten verwenden um Probleme zu lösen, die vermeintlich die Anwendung von Boole'schen Methoden im Dokumentinhalt erfordern. Es folgen zwei typische Fälle.

Spezifischer Ausschluss von Seiten

Beispiel: Ein Fussballnarr sucht nach den Resultaten vergangener Fussballspiele. Seine Anfrage lautet *soccer*. Erster Kritikpunkt: Ein Suchbegriff ist zu wenig, er sollte sich unbedingt noch einige weitere Begriffe einfallen lassen. Trotzdem führt er die Anfrage durch. Er erhält eine unbefriedigende Rangliste, denn sie wird dominiert von störenden Einträgen eines Computerspieleherstellers, der ein Fussballspiel namens *Power Soccer* anpreist. Auf der Web-Site des Herstellers unter <http://www.powersoccer.com> gibt es Dutzende von Seiten mit Informationen zum Spiel, und auf jeder Seite wird der Name wiederholt.

«Kein Problem!», denkt unser Fussballnarr, denn er kennt ja nun die Boole'schen Suchmethoden. Damit können die unerwünschten Seiten ausgeschlossen werden. Mit der neuen Anfrage *soccer -power* werden alle Seiten mit dem Namen des Computerspiels ausgeschlossen. Leider unterdrückt diese Anfrage auch alle Seiten, die den Begriff «Power» in anderem Zusammenhang verwenden, weil zum Beispiel von der schwindenden Kraft der Spieler oder von einem kraftvollen Spiel die Rede ist.

Sobald irrelevante Einträge in der Rangliste gehäuft auftreten, kann man oft auch mit Hilfe der Metadaten eine geeignete Teilkol-

lektion definieren. Im obigen Beispiel stammen alle störenden Seiten vom gleichen Web-Server. Also definiert man eine Teilkollektion ohne diesen Server mit der Anfrage *soccer -server:www.powersoccer.com*, sofern der verwendete Suchdienst diese Möglichkeit anbietet.

Konzentration auf ein Thema

Der folgende Fall ist ähnlich wie der soeben besprochene: Eine Anfrage produziert eine Rangliste, die neben dem relevanten Thema auch Dokumente zu ganz anderen Bereichen enthält. In solchen Situationen hilft das Verwenden von zusätzlichen Suchbegriffen, die das gewünschte Thema näher umschreiben.

Im zweiten Kapitel lieferte die Anfrage *Noah* Dokumente zum Tennisspieler und zu Noahs Arche. Wie könnte die Anfrage verändert werden, um mehr Tennis-Dokumente im oberen Bereich der Rangliste zu erhalten? Die riskante Variante: Ausschluss der Arche-Noah-Dokumente mittels einer Boole'schen Anfrage wie *Noah -Arche*. Dabei kann es vorkommen, dass auch relevante Dokumente ausgeschlossen werden, beispielsweise Dokumente über Yannick Noah bei einem ironischen Hinweis auf die Arche. Die weniger gefährliche Variante: Man fügt einen Begriff zur Anfrage hinzu, der in engem Zusammenhang mit dem gewünschten Thema steht. In diesem Fall ist die Lösung offensichtlich – man stellt die Anfrage *Yannick Noah*.

Phrasensuche

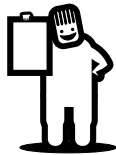
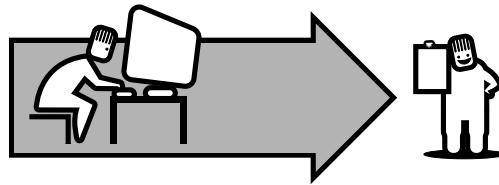
Neben all den Risiken bei der Boole'schen Suche im Dokumentinhalt soll eine letzte, oft hilfreiche Funktion nicht unerwähnt bleiben: Die so genannte *Phrasensuche*. Mit Hilfe dieser Funktion lassen sich zwei oder mehr Suchbegriffe zu einem festen Paket verschnüren. Die Suchbegriffe müssen dann zwingend in der angegebenen Reihenfolge in den gesuchten Dokumenten vorkommen.

Bei OMNISEARCH werden Phrasen festgelegt, indem man die Begriffe mit doppelten Anführungszeichen umschließt. Häufiger Verwendungszweck sind Buchtitel, Filmtitel, Produktnamen, Firmennamen, Zitate, Personennamen oder andere Bezeichnungen, die immer in derselben Art auftreten. Beispiele: *“Big Ben“*, *“Elvis Presley“*,

“Confoederatio Helvetica“, “World Wide Web“, “Cable News Network“, “Ben Hur“ oder “Just do it“.

Mit der Phrasensuche kann man also denjenigen Suchsystemen auf die Sprünge helfen, die das Rangierungsprinzip 5 nicht unterstützen: Je näher die Suchbegriffe beieinander liegen, desto relevanter das Dokument.

Schauen wir nun, wie die Anwender ihre Probleme mit den neuen Techniken lösen ...

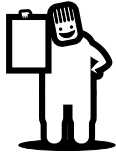


Ich bin immer noch auf der Suche nach dem «richtigen» WWF und habe ein Problem mit all den störenden Dokumenten vom Server der Wrestler unter `www.wwf.com`. Eine mögliche Lösung ist jetzt klar: Ich kann bei OMNISEARCH sehr einfach die irrelevanten Dokumente ausschliessen, indem ich eine Teilkollektion ohne den entsprechenden Server bilde. Die neue Anfrage lautet dann *WWF -host:www.wwf.com*.

Eine andere Möglichkeit ist, auf die Boole'schen Techniken und Metadaten zu verzichten und stattdessen auf die Rangierungsprinzipien zu setzen. Ich kann die mich interessierenden Dokumente in der Rangliste nach vorne bringen, indem ich weitere charakteristische Suchbegriffe benutze. In diesem Fall sind das die Begriffe *World*, *Wildlife* und *Fund*. Und weil die Wörter immer in dieser Kombination auftreten, setze ich sie als Phrase zusammen. Das führt zur Anfrage *WWF "World Wildlife Fund"*.

Es gibt noch eine Möglichkeit: Raten! Wenn es nur darum geht, die Homepage einer Firma oder einer Organisation zu finden, kommt man mit Raten häufig zum Ziel. WWF zum Beispiel ist eine internationale Nonprofit-Organisation. Also hätte ich als ersten Versuch direkt den URL `www.wwf.org` eintippen können. WWF ist aber

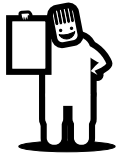
auch ein kommerzieller Wrestlingveranstalter, der natürlich unter dem URL www.wwf.com erreichbar ist. Weitere Beispiele: Der IBM-Konzern ist unter www.ibm.com zu finden, die Schweizerischen Bundesbahnen unter www.sbb.ch und die Seiten des Eiffelturms unter www.tour-eiffel.fr.



Für mein Problem mit den Übernachtungszahlen kann ich die Boole'schen Techniken ausnützen, um eine passende Teildokumentenkollektion zusammenzustellen. Anschließend kann ich in dieser Kollektion gezielt weitersuchen.

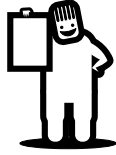
Ich habe mir überlegt, dass ich mit meinem Problem vermutlich bei den statistischen Ämtern von Deutschland, Österreich und der Schweiz am schnellsten zum Ziel komme. Also stelle ich eine Kollektion zusammen, die nur gerade Dokumente von diesen drei Web-Servern enthält. Das erreiche ich mit dem Boole'schen Ausdruck *server:statistik-bund.de OR server:admin.ch OR server:oestat.gv.at*.

Nun kann ich Begriffe wählen, nach denen die Dokumente in der definierten Teilkollektion rangiert werden. Ich versuche es mit *tourismus übernachtung* uebernachtung**. Die Anfrage deckt den Umlaut in beiden Varianten ab und findet mit dem * ausserdem die Mehrzahl von Übernachtung. So komme ich rasch ans Ziel. 1997 zählte man in Österreich 100 und in der Schweiz 66 Millionen Übernachtungen im Gastgewerbe. In Deutschland waren es rund 300 Millionen.



Ich habe mir eine weitere interessante Anwendung der Metadaten überlegt. Seit einiger Zeit unterhalte ich meine eigene Homepage unter der Adresse www.i.ch/home.html. Nun möchte ich wissen, wie beliebt ich im Netz bin. Das

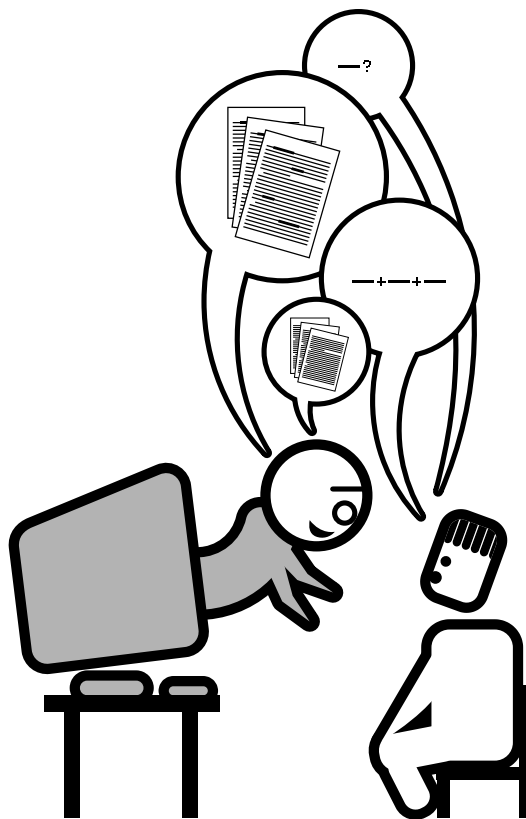
heisst, wie viele andere Webseiten verweisen auf meine Homepage? OMNISEARCH hilft mir weiter: Mit Hilfe der Einschränkung *+link:www.i.ch/home.html* wähle ich alle Seiten aus, die einen Link auf meine Homepage beinhalten. Natürlich möchte ich meine eigenen Verweise nicht mitzählen und schliesse deshalb mit dem Zusatz *-host:www.i.ch* alle Seiten von meinem Server aus. Ausserdem möchte ich die Dokumente an oberster Stelle sehen, die mich im Zusammenhang mit Kunst und Malerei erwähnen, denn ich biete auf meinen Webseiten einige meiner Werke an. Also lautet die vollständige Anfrage *+link:www.i.ch/home.html -host:www.i.ch kunst malerei*.



Bei meiner Suche nach dem Hotel namens Aurora in Castiglione della Pescaia habe ich Boole'sche Methoden im Dokumentinhalt angewendet und bin dabei prompt in die Falle getappt. Meine Anfrage lautete *Aurora AND Hotel AND "Castiglione della Pescaia"*. Auf der gesuchten Homepage taucht der Begriff «Hotel» allerdings nicht auf. Stattdessen ist von der «Albergo Aurora» die Rede. Ohne Boole'sche Operatoren hätte die Anfrage *Aurora Hotel "Castiglione della Pescaia"* das Hotel gefunden. Die Boole'schen Operatoren hätte ich besser dazu verwendet, eine Teilkollektion der Webseiten innerhalb der Domain «it» zu definieren.

Kapitel 6

Interaktive Suchtechniken





Kurz vor zwölf. Sheriff Will Kane erwartet die Miller Gang mit dem 12:00-Uhr-Zug. Er rüstet sich zur Verteidigung seiner Heimatstadt Hadleyville. Niemand will ihm helfen – sein Deputy hält sich zurück, weil dieser es auf den Posten des Sheriffs abgesehen hat, und der ehemalige Sheriff ist zu alt für eine Auseinandersetzung.

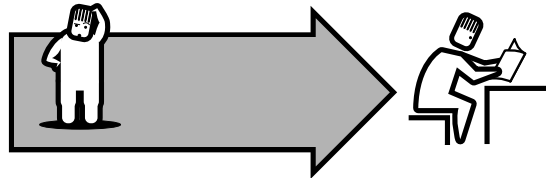
So in etwa gibt sich die Ausgangslage im Kultwestern «High Noon» aus dem Jahre 1952. «High Noon» gewann vier Oscars, einen davon für die Filmmusik. Ich möchte herausfinden, wer für die Musik im Film verantwortlich war. Mein erster Versuch bei OMNISEARCH: *“High Noon“*. Der Erfolg bleibt aus – ich finde etwas über «High Noon Poker Tournaments», «High Noon Wild West Shows» oder eine «High Noon Home Brewery».

Gut, zwei Suchbegriffe sind natürlich auch etwas wenig bei einem Allerweltsbegriff wie «High Noon». Also versuche ich es nochmals mit *“High Noon“ music movie composer*. Ich erhalte andere Resultate, doch auch die führen nicht zum Ziel. Dritter Versuch mit *“High Noon“ sound track compose*. Leider ist die Antwort auf meine Frage immer noch nicht in Sicht. Frustriert gebe ich die Suche auf. Ich hätte doch schon lange ein relevantes Dokument finden sollen!



Kürzlich ist mir ein Prospekt für Kochkurse bei einer bestimmten Schule in die Hände geraten. Das Angebot sieht sehr interessant aus, und ich würde gerne einen solchen Kurs besuchen. Doch bevor ich mich bei dieser Schule anmelde, möchte ich mich über die Konkurrenz informieren. Mich interessiert ein vollständiger Überblick über alle Anbieter von solchen Kursen, was jeweils geboten wird und was ich zu zahlen hätte. Kann mir ein Suchdienst weiterhelfen?

Für diese etwas anspruchsvolleren Problemstellungen brauchen wir leistungsfähige Hilfsmittel. Solche Hilfsmittel werden in den folgenden Abschnitten vorgestellt ...



Iterativer Ablauf einer Recherche

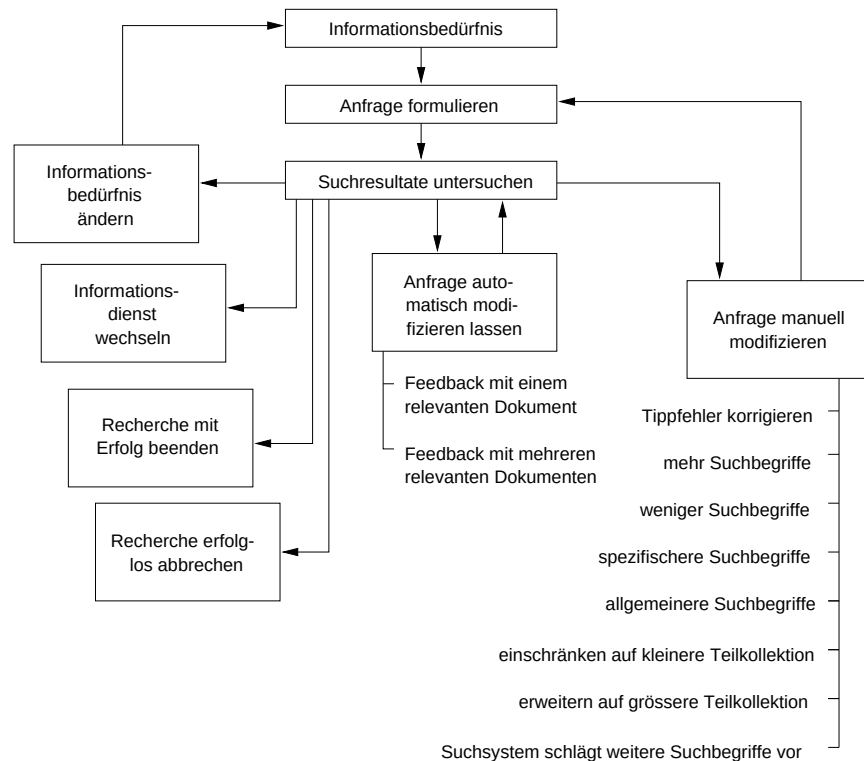
Informationssuche ist ein iterativer, das heisst in mehreren Durchgängen auszuführender Prozess! Diese Aussage gilt in den meisten Fällen, denn nur selten führt die erste Anfrage an einen Suchdienst gleich zum Ziel. Im Normalfall arbeitet man sich schrittweise zum gewünschten Resultat vor. Mit jeder Anfrage lernt man etwas hinzu und verwendet das zusätzliche Wissen für die folgenden Suchanfragen. Bei diesem iterativen Prozess kann man auf mehr oder weniger Unterstützung durch das Suchsystem zählen, je nachdem wie luxuriös der verwendete Suchdienst ausgestattet ist. Einige Techniken der Interaktion mit unterschiedlich ausgeprägter Systemunterstützung werden in diesem Kapitel vorgestellt. Doch überlegen wir uns zuerst, welche Ereignisse während einer Recherche eintreten können.

An erster Stelle steht natürlich immer das Informationsbedürfnis einer Benutzerin. Die Benutzerin entscheidet sich für einen Suchdienst und formuliert eine Anfrage, indem sie geeignete Suchbegriffe zusammenstellt. Die Anfrage wird an den Suchdienst geschickt, und die Benutzerin begutachtet je nach Bedürfnis einige oder viele Dokumente der resultierenden Rangliste.

Nun hat die Benutzerin eine Auswahl an Möglichkeiten. Vielleicht war die Recherche von Erfolg gekrönt. Das Informationsbedürfnis ist befriedigt, die Suche ist beendet. Oder die Benutzerin gibt die Suche erfolglos und frustriert auf. Eine andere Möglichkeit: Die Benutzerin beschliesst, ein anderes Werkzeug für die Informationsbeschaffung im

Internet zu benutzen, oder sie wendet sich sogar einem anderen Medium zu. Auch im Internet-Zeitalter ist ein Telefonat an die richtige Stelle oder der Gang in die Bibliothek manchmal viel versprechender als die endlose Suche im Netz der Netze.

Es kann auch vorkommen, dass eine Benutzerin aufgrund der betrachteten Dokumente das ursprüngliche Informationsbedürfnis an die neuen Gegebenheiten anpasst. Sie wird daraufhin eine neue Anfrage formulieren und die Recherche fortführen, eventuell mit einem anderen Suchdienst.



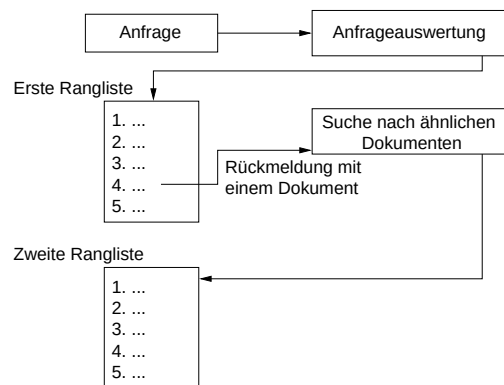
Am häufigsten jedoch wird die Benutzerin aus den bisher gemachten Erfahrungen und aus den untersuchten Dokumenten lernen, ihre Anfrage entsprechend ändern und die Suche mit der neuen Anfrage weiterverfolgen. Die Anpassung erfolgt automatisch oder manuell.

- *Manuelle Modifikation der Anfrage:* Man kann Suchbegriffe zur Anfrage hinzufügen oder aus der Anfrage entfernen. Oder man kann die bestehenden Suchbegriffe anpassen, indem spezifischere (seltener allgemeinere) Alternativen gewählt werden. Bei solchen Entscheidungen wird man sein Hintergrundwissen zum Thema einfließen lassen und auch das Wissen um das verwendete Werkzeug und die Dokumentensammlung. Manche Suchsysteme unterstützen ihre Benutzer bei der manuellen Modifikation der Anfrage.
- *Automatische Modifikation der Anfrage:* Die automatische Modifikation setzt voraus, dass das Suchsystem ein Feedback (eine Rückmeldung) erhält. Mit Hilfe des Feedbacks kann das System selbstständig eine neue Anfrage zusammenstellen und auswerten. Das Feedback an das System kann automatisch erfolgen, indem beispielsweise einige der bestrangierten Dokumente in die Rückmeldung einfließen. In manchen Fällen ist jedoch der Benutzer für die Rückmeldung verantwortlich. Dabei teilt der Benutzer dem System mit, welche Dokumente für ihn relevant waren. Für die *Suche nach ähnlichen Dokumenten* wird ein einziges relevantes Dokument benötigt. Für eine *Relevanzrückkopplung* darf das Feedback aus mehr als einem relevanten Dokument bestehen.

Suche nach ähnlichen Dokumenten

Jemand kommt auf die Idee, ein vollständiges Dokument als Anfrage für einen Suchdienst zu verwenden. Was geschieht? Das Suchsystem geht wie immer vor. Zuerst indexiert es die Anfrage und extrahiert dabei die Suchbegriffe. Gestützt auf die Rangierungsprinzipien sucht das System dann nach Dokumenten, die bezüglich dieser Anfrage relevant sind, und stellt daraus die Rangliste zusammen. Zu den relevantesten Dokumenten gehören natürlich diejenigen, die dem ursprünglichen Dokument sehr ähnlich sind, weil sie dieselben Begriffe in vergleichbarer Häufigkeit enthalten.

Suchsysteme mit dieser Funktion erlauben es der Benutzerin also, ein bereits gefundenes Dokument als neue Anfrage an das System zu stellen. Bei langen Dokumenten werden nur die «wichtigen» Begriffe im Text als Suchbegriffe für die Anfrage verwendet. Wichtig sind gemäss der Rangierungsprinzipien beispielsweise solche Begriffe, die im Dokument häufig, ansonsten aber eher selten vorkommen.

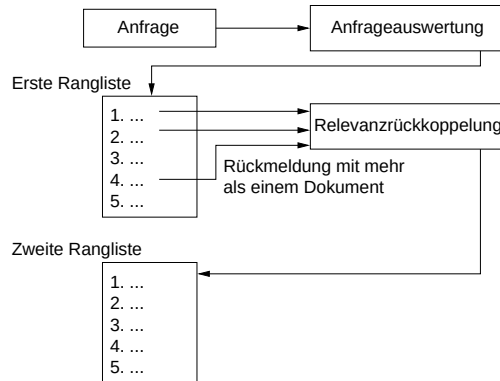


Bei der Suche nach ähnlichen Dokumenten nützt man den Vorteil aus, dass Anfragen mit einer grossen Zahl von Suchbegriffen oft zu besseren Resultaten führen als Anfragen mit nur wenigen Suchbegriffen.

Relevanzrückkoppelung

Die Relevanzrückkoppelung stellt ein mächtigeres Werkzeug dar als die Suche nach ähnlichen Dokumenten: Nach einer ersten Anfrage beurteilt der Benutzer zunächst, welche der Dokumente in der Rangliste für sein Informationsbedürfnis relevant sind. Diese Information teilt er anschliessend dem Suchsystem mit. Daraufhin untersucht das Suchsystem die relevanten Dokumente, identifiziert häufige Begriffe und stellt daraus eine neue Anfrage zusammen. Bestimmte Suchbegriffe werden also als besonders relevant erkannt. Genauso kann das System andere Begriffe als besonders irrelevant erkennen. Schliesslich wird die neue Anfrage verarbeitet und dem Benutzer eine neue Rangliste präsentiert.

Mit Hilfe der Relevanzrückkoppelung lassen sich umfangreiche Anfragen mit vielen Suchbegriffen zusammenstellen, mit denen häufig viele zusätzliche relevante Dokumente gefunden werden. Dokumente werden auch dann gefunden, wenn sie keinen einzigen Suchbegriff der ursprünglichen Anfrage enthalten.



Achtung: Gefahr des Abdriftens!

Manchmal kommt es vor, dass man während einer Recherche vom ursprünglich relevanten Themengebiet in einen Teilbereich oder in ein völlig anderes Gebiet gleitet. Dieses Problem nennen wir Abdriften.

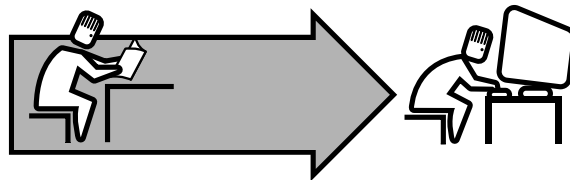
Die Gefahr des ungewollten Abdriftens vom ursprünglichen Informationsbedürfnis während einer Relevanzrückkoppelung stellt sich vor allem bei Themen, die verschiedene Aspekte umfassen. Man sollte darauf achten, jeden Aspekt durch ein relevantes Dokument abzudecken.

Beispiel: Ein Ölscheich ist angsterfüllt auf der Suche nach Informationen über alternative Energiequellen. Er findet einiges über Solarenergie, Windenergie, Gezeitenkraftwerke und Ähnliches. Mit Hilfe einer Relevanzrückkoppelung möchte sich der Scheich weitere Dokumente beschaffen und sollte dazu möglichst zu jeder Art von Energiegewinnung ein Dokument einspeisen. Verwendet er lediglich einige Dokumente zum Thema Solarenergie, wird er in diesen Bereich abdriften und in erster Linie Resultate aus dem Gebiet der Solarenergie erhalten.

Manuelle Anfrageerweiterung

Es handelt sich hier um ein einfaches, halbautomatisches Instrument. Bei der Verarbeitung einer Anfrage versucht das Suchsystem aufgrund der gefundenen Dokumente herauszufinden, welche weiteren Suchbegriffe mit der Anfrage in Zusammenhang stehen könnten. Moderne Suchsysteme erkennen mögliche Beziehungen zwischen Begriffen, weil sie systematisch ermitteln, welche Begriffe häufig zusammen vorkommen. Eine Auswahl dieser zusätzlichen Suchbegriffe wird der Benutzerin präsentiert, die dann einzelne Begriffe zur Anfrage hinzufügen, von der Anfrage ausschliessen oder einfach ignorieren kann. Aufgrund der vorgeschlagenen Begriffe kann nun eine neue Anfrage zusammengestellt werden.

Es folgt ein praktisches Beispiel, das die vorgestellten Techniken nochmals aufgreift ...



Interaktive Techniken in der Praxis

Die Möglichkeiten der interaktiven Suche unterscheiden sich in der Regel von Suchdienst zu Suchdienst. Wie immer sollte man entsprechende Hinweise in den Hilfeseiten finden.

Manuelle Relevanzrückkoppelung

Was tun, wenn ein Suchdienst keines der automatisierten Hilfsmittel anbietet? Dann hilft man sich am besten selber mit einer «manuellen Relevanzrückkoppelung». Man begutachtet die relevanten Dokumente, die auf eine erste Anfrage gemeldet werden. In den Dokumenten

identifiziert man gute Suchbegriffe, die in Zusammenhang mit der ursprünglichen Anfrage stehen. Gute Suchbegriffe sind häufig Fachbegriffe oder auch Eigennamen von Produkten, Firmen, Personen, Orten und dergleichen. Je enger ein Begriff mit dem gesuchten Dokument verknüpft ist, desto besser.

Im zweiten Schritt erweitert man manuell die ursprüngliche Anfrage mit diesen zusätzlichen Suchbegriffen und führt anschliessend die neue Suchanfrage durch.

Interaktive Suche im Beispiel

Gesucht wird nach statistischen Angaben, ob die Zahl der Heiraten in der Schweiz eher zu- oder abnimmt. NEWSSEEKER liefert auf die Anfrage *Anzahl Heiraten Schweiz Statistik* das folgende Dokument:

*Weniger **Heiraten** und Geburten, Stagnation der Scheidungen*
Gemäss Ergebnissen des Bundesamtes für **Statistik** (BFS) hat sich der Rückgang der **Anzahl** Eheschliessungen 1995 in der **Schweiz** fortgesetzt. Es wurden 4% weniger Trauungen verzeichnet als im Vorjahr. Die Scheidungszahlen sind erstmals in den 90er-Jahren gleich geblieben. Etwa 15 000 Ehepaare liessen sich scheiden. Abgenommen hat hingegen die **Anzahl** Geburten. Die Geburtenhäufigkeit sank von 1,49 auf 1,47 Kinder je Frau – ein neuer Tiefststand.

Das Dokument ist relevant für die Anfrage. Allerdings soll nun noch überprüft werden, ob weitere Dokumente jüngerem Datums zum Thema existieren. Das kann man von Hand erledigen, indem man das Dokument untersucht und geeignete Begriffe zur ursprünglichen Anfrage hinzufügt. Welches sind die für das vorliegende Dokument charakteristischen Begriffe? Zunächst fallen verwandte Begriffe wie Ehe, Eheschliessungen, Trauung auf. Ausserdem kann man schliessen, dass in ähnlichen Dokumenten häufig gleichzeitig über Scheidungszahlen und Geburtenhäufigkeiten informiert wird. Die neue Anfrage nach dieser manuellen Relevanzrückkoppelung könnte lauten: *Anzahl Heiraten Schweiz Statistik Ehe Trauung Geburten Scheidungen*.

Die automatisierte Suche nach ähnlichen Dokumenten funktioniert etwas anders. Es werden stur alle Begriffe (mit Ausnahme der

Stoppwörter) aus dem Dokument extrahiert und zu einer neuen Anfrage zusammengestellt. Die neue Anfrage wird einen ähnlichen Effekt haben wie die von Hand zusammengestellte. Es werden weitere Dokumente gefunden, welche dieselben Begriffe verwenden wie im gegebenen Dokument.

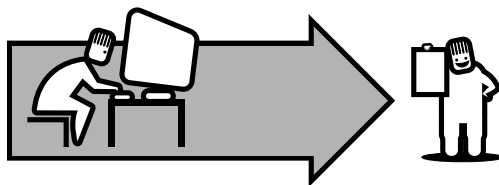
Bei einer Relevanzrückkoppelung kann die Benutzerin dem Suchsystem mehr als ein relevantes Dokument vorgeben. Die einzelnen Dokumente werden indexiert. Daraufhin stellt das System die extrahierten Begriffe zu einer neuen Anfrage zusammen.

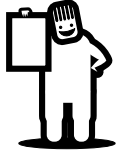
Manchmal führt ein Umweg schneller zum Ziel

Neben den vorgestellten Möglichkeiten der interaktiven Suche sollte man immer auch an die verschiedenen Angebote denken. Oft gibt es für ein Informationsbedürfnis ein spezielles Angebot im Netz oder eine spezifische vertikale Dokumentensammlung. In solchen Situationen kann man über einen Umweg ans Ziel gelangen: Zunächst sucht man zum Beispiel mit OMNISEARCH nach dem Spezialangebot und forscht im zweiten Schritt mit dem neu gefundenen Informationsdienst weiter.

Beispiel: Gesucht ist die deutsche Übersetzung des englischen Begriffs «robin». Es ist recht unwahrscheinlich, dass eine Webseite mit dem Inhalt «die Übersetzung des Wortes <robin> ins Deutsche lautet <Rotkehlchen>» existiert. Viel eher wird man auf Dokumente zu Robin Hood stossen. Darum sucht man vorzugsweise zuerst nach einem geeigneten Wörterbuch, zum Beispiel mit der Anfrage *english german dictionary* bei OMNISEARCH. Wird man fündig, kann man anschliessend innerhalb des Wörterbuches die Übersetzung vornehmen.

Nun sind die Anwender wieder an der Reihe, um die Lösungen zu den anfangs gestellten Fragen zu präsentieren ...

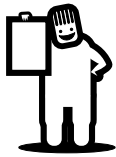




Mein neues Motto lautet: Nicht verzagen, wenn es mal etwas länger geht! Stattdessen lieber zwischendurch die eigene Vorgehensweise und die gewählten Suchbegriffe überdenken, falls man nicht weiterkommt.

Mein Problem war die Filmmusik zu «High Noon». Mit meinen Anfragen bin ich bisher nicht zum Ziel gekommen. Ich könnte selbstverständlich andere Suchbegriffe versuchen. Unterdessen habe ich mir jedoch überlegt, dass es sicherlich irgendwo im Netz ein spezielles Angebot für Filmfans geben muss – so etwas wie eine vertikale Dokumentensammlung im Filmbereich.

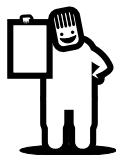
Doch wie finde ich ein solches Spezialangebot? Ich versuche es mit einem dieser Katalogdienste. Tatsächlich: Sehr rasch stosse ich auf die Kategorie *Entertainment / Movies / Databases*. Ich wähle eine der aufgeführten Datenbanken aus und stelle dort die einfache Anfrage *High Noon*. Endlich! Dimitri Tiomkin und Ned Washington waren gemeinsam für die Filmmusik verantwortlich.



Jetzt ist mir klar, wie ich zu meiner Übersicht mit verschiedenen Anbietern von Kochkursen komme. Ich benutze die Kursbeschreibung aus dem Prospekt und suche nach ähnlichen Dokumenten im Internet. Dazu nehme ich einen Leuchtstift und markiere in der Ausschreibung alle Begriffe, die charakteristisch sind für eine solche Kursbeschreibung.

Besuchen Sie unseren **Kochkurs** für Anfänger. Sie als **Teilnehmerin** oder **Teilnehmer** werden in die Kunst des Kochens eingeführt. Aus marktfrischem **Gemüse**, **Fleisch** und **Fisch** werden Sie die köstlichsten **Gerichte** zubereiten. Beim gemeinsamen **Rüsten**, **Kochen** und **Essen** erhalten Sie viele nützliche Tipps. Und als Andenken schenken wir Ihnen eine Sammlung von **Rezepten** aus der italienischen **Küche**.

Dann stelle ich mit den markierten Begriffen eine Anfrage zusammen: *Kochkurs Teilnehmerin Teilnehmer Gemüse Fleisch Fisch rüsten kochen essen Rezepte Küche*. OMNISEARCH liefert mir auf diese Anfrage eine ausführliche Liste mit Anbietern von Kochkursen, und ich kann die verschiedenen Angebote rasch miteinander vergleichen.

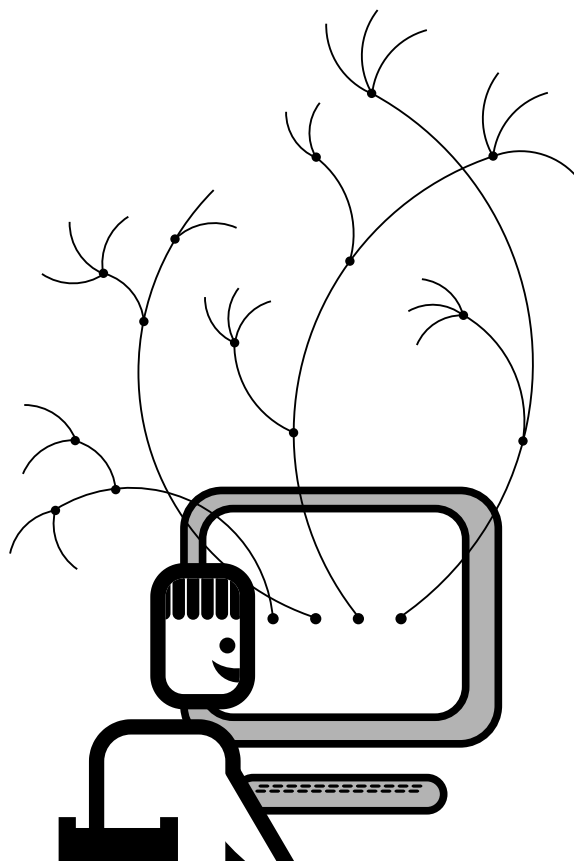


Jahr für Jahr zum Nationalfeiertag sparen unsere Politiker nicht mit ausgefeilten Ansprachen an die Nation. Zudem gibt es eine Menge anderer Gelegenheiten für eine Rede. Da frage ich mich, ob die Politiker ihre Referate immer selber schreiben, oder ob sie nicht auch ab und zu von einer Kollegin etwas kopieren. Mit Hilfe des Internets kann ich meinen Verdacht ohne Probleme überprüfen, denn viele der Ansprachen werden irgendwo im Netz veröffentlicht. Also kann ich die interaktiven Suchtechniken für meine Zwecke ausnützen. Sobald ich eine typische Rede gefunden habe, suche ich damit nach ähnlichen Dokumenten oder führe eine Relevanzrückkoppelung durch. So sollten sich die am häufigsten behandelten Themen, ja sogar allfällige Plagiate prompt entlarven lassen.

Ich habe diese Plagiat-Recherche bei NEWSSEEKER durchgeführt und mir meine Gedanken gemacht. An dieser Stelle möchte ich aber nicht weiter auf dieses heikle Thema eingehen, schliesslich will ich niemandem auf die Füsse treten ... Stattdessen gebe ich lieber einen letzten Hinweis: Man kann diese Technik der Suche nach ähnlichen Dokumenten auch zum Auffinden von Copyright-Verletzungen benutzen. Dazu füttert man ein Suchsystem mit einem eigenen Dokument und überprüft dann, ob im Internet Kopien davon vorkommen.

Kapitel 7

Katalogdienste





Als ich noch ein Kind war, zählten «Die schwarzen Brüder» zu den unangefochtenen Favoriten in meinem Bücherregal. In dem Buch wird eine Gruppe von Kindern aus dem Schweizer Kanton Tessin in Menschenhändlermanier nach Mailand verschleppt. Dort werden sie als Kaminfeger für jede noch so dreckige Arbeit ausgenutzt. Doch zum Schluss siegt der Geist der Solidarität der «schwarzen Brüder» über die egoistischen Machenschaften der Kaminfegerbosse.

Unglücklicherweise kann ich mich nicht mehr an den Namen der Autorin des Bestsellers erinnern. Also mache ich mich in einem Katalogdienst auf die Suche. Die Recherche beginnt viel versprechend – über die Kategorien *Literatur* und *Bücher* lande ich bei *Kinderbücher*. Innerhalb der Kinderbücher-Kategorie kann ich nach den Begriffen *schwarze brüder* suchen. Doch ich finde nur Einträge über die Gebrüder Grimm, weil darin ebenfalls der Suchbegriff *brüder* auftaucht. Ansonsten bleibt meine Suche erfolglos.



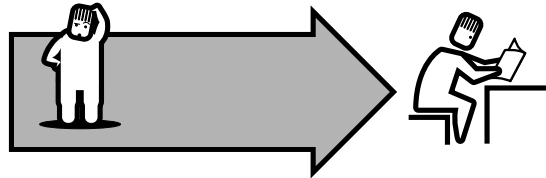
Als migränegeplagter Hobby-Radrennfahrer stellt sich für mich eine wichtige Frage: Welches Medikament kann ich gegen meine Kopfschmerzen nehmen, ohne eine Disqualifikation bei einer allfälligen Dopingkontrolle fürchten zu müssen?

Kurzerhand besuche ich einen Katalogdienst auf der Suche nach einer Liste mit Medikamenten, die gesperrt sind. Natürlich wähle ich die Kategorie *Freizeit / Sport / Radfahren*. In dieser Kategorie finde ich diverse Unterkategorien zu den unterschiedlichsten Facetten des Radsports. Doch zum Thema Doping taucht nichts auf. Wo liegt das Problem?



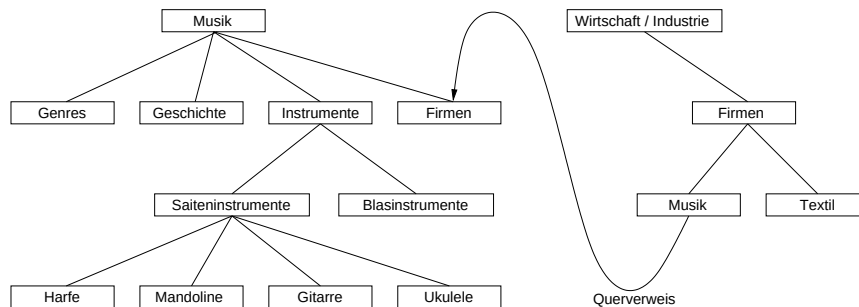
Ich will wissen, wie viele Apache-Indianer es noch gibt. Kein Problem! Im Katalogdienst meiner Wahl kann ich auch nach Stichworten suchen. Ich trage den Begriff *apache* in das Formular ein und starte die Suche. Kurz darauf erhalte ich eine Liste mit den gefundenen Einträgen. Sofort wähle ich den ersten Eintrag an und lande in einer Kategorie, die rein gar nichts mit Indianern zu tun hat. Stattdessen ist von der Apache Software Foundation, Perl, SSL und anderen kryptischen Begriffen die Rede.

Offenbar ist unseren Internet-Anwendern noch nicht ganz klar, wie Katalogsysteme funktionieren. Nehmen wir diese Werkzeuge also etwas genauer unter die Lupe ...



Aufbau von Katalogsystemen

Bei Katalogsystemen funktioniert der Zugriff auf Informationen anders als bei Suchdiensten. Ein Suchsystem bietet den *direkten* Zugriff auf Dokumente in einer ungeordneten Menge von Dokumenten an. Katalogsysteme dagegen versuchen, eine gewisse Struktur und Ordnung in die Dokumentenmenge zu bringen. Ein Katalogsystem funktioniert ähnlich wie eine Bibliothek oder das Branchenverzeichnis im Telefonbuch: Die Einträge sind in *Kategorien* eingeteilt. Ein Katalogsystem unterstützt demnach den *indirekten* Zugriff auf Dokumente via Kategorien.



Ein Katalogsystem ist nichts anderes als eine Hierarchie von Kategorien. Ganz oben in der Hierarchie erscheinen die allgemeinsten Kategorien. Je weiter man hinuntersteigt, desto spezifischer werden die Einträge. Viele Kategorien besitzen eine oder mehrere Unterkategorien, und ausnahmslos jede Kategorie verfügt über einen Namen.

Neben der streng hierarchischen Gliederung gibt es auch Querverweise, die in eine andere Teilhierarchie mit Dokumenten zu einem verwandten Thema führen können.

Die Dokumente innerhalb der Kollektion des Katalogdienstes werden diesen Kategorien als Verweise (Hyperlinks) zugeordnet. Häufig taucht ein Dokument in mehreren Kategorien auf. Zu jedem Dokument existieren im Allgemeinen der Titel und eine kurze Zusammenfassung sowie eventuell weitere Metadaten.

Auch bei Katalogdiensten (kurz: Katalogen) muss klar zwischen dem Werkzeug und der Dokumentenkollektion unterschieden werden. Die Anordnung von Dokumenten in einer Kategorienhierarchie stellt das Prinzip für den Informationszugriff dar. Die tatsächlichen Informationen können den unterschiedlichsten Quellen entstammen. Es gibt Katalogdienste, die eine horizontale Dokumentenkollektion abdecken. Auf der anderen Seite sind vertikale Dokumentenkollektionen genauso denkbar: zum Beispiel ein Katalogdienst im Bereich von Meteorologie oder einer zum Thema Architektur.

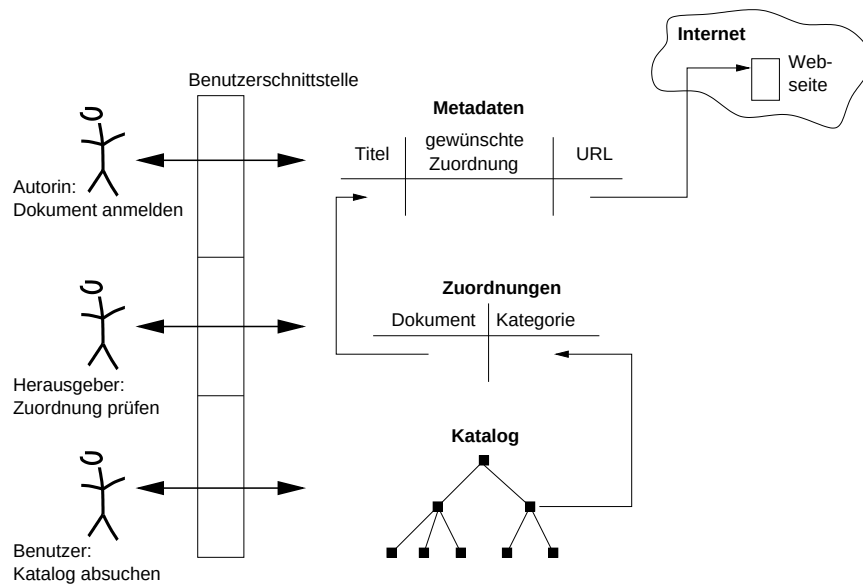
Wie läuft eine typische Recherche in einem Katalogdienst ab? Genauso wie man es von einer Bibliothek her kennt: Man beginnt bei einer der allgemeineren Kategorien und blättert so lange weiter, bis eine geeignete Unterkategorie gefunden wird. Dort sucht man nach einem relevanten Dokument und schaut es an. Häufig kann auch nach Kategorien gesucht werden. Man muss also unterscheiden zwischen der Suche nach Kategorien und der Suche nach Dokumenten.

Manuelle Erstellung

Manche Katalogdienste werden von Menschenhand erstellt. Hinter diesen Diensten steckt eine Gruppe von Herausgebern. Sie legen zuerst die Hierarchie der Kategorien fest. Dann werden neue Dokumente gesichtet und aufgrund des Inhalts einer geeigneten Kategorie zugeordnet.

Die Herausgeber sind auch für das Auffinden neuer Dokumente verantwortlich. Oder sie warten einfach darauf, bis die Besitzer von Webseiten ihre Seiten explizit beim Katalogdienst anmelden.

Bei den manuell erstellten Katalogdiensten findet eine gewisse Qualitätskontrolle statt, weil die Herausgeber die Dokumente und die zugehörigen Metadaten (zum Beispiel eine Kurzzusammenfassung) vor dem Einordnen überprüfen. Ein Herausgeber hat die Möglichkeit, einen Eintrag abzulehnen. Mögliche Gründe: Das Dokument stellt zu wenig Information bereit, um für eine bestimmte Kategorie eine Bereicherung zu sein. Oder ein Dokument passt in keine der verfügbaren Kategorien, und es scheint nicht angebracht, eine neue Kategorie zu eröffnen. Der häufigste Fall ist, dass die Autoren ein Dokument in zu vielen Kategorien anmelden.

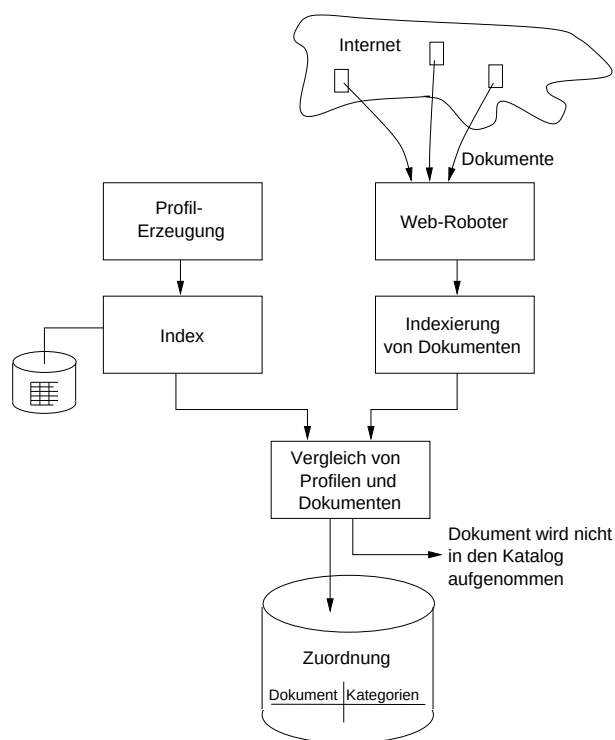


Automatische Klassifizierung von Dokumenten

Klassifizierung, so nennt man das Einordnen von Dingen (zum Beispiel von Dokumenten) in Kategorien. Dieser Vorgang kann automatisiert werden, sodass ein Katalogdienst auch ohne menschliche Betreuung betrieben werden kann. Die meisten für ein automatisiertes Katalogsystem benötigten Bausteine haben wir bereits kennen gelernt. Nun müssen sie nur noch richtig kombiniert werden.

Kernstück eines Katalogsystems sind die Kategorien. Bei einem automatisierten Katalogsystem muss in einem ersten Schritt eine Sammlung von Kategorien erstellt werden. Üblicherweise übernehmen menschliche Herausgeber diese Arbeit.

Die rechte Seite im Schema kennen wir: Ein Web-Roboter beschafft Dokumente aus dem Internet. Die gefundenen Dokumente werden indexiert, und für jedes Dokument werden die üblichen Metadaten wie URL, Titel, Modifikationsdatum usw. vermerkt.



Wie werden Dokumente eingeordnet?

Jedes neu erschlossene Dokument soll automatisch einer Kategorie im Katalogsystem zugeordnet werden. Zu diesem Zweck existiert für alle Kategorien ein eigenes Profil. Das *Profil* ist eine Menge von Begriffen, die das Themengebiet der jeweiligen Kategorie charakterisiert.

Mit diesen Profilen kann für ein neues Dokument geprüft werden, ob es in eine gewisse Kategorie passt oder nicht. Dazu muss lediglich das Profil der entsprechenden Kategorie mit dem Dokumentinhalt verglichen werden. Bei diesem Vergleich kommen die besprochenen Rangierungsprinzipien zur Anwendung. Das Resultat: ein Relevanzwert. Je höher der Wert, desto stärker sind Profil und Dokument verknüpft. Liegt der Relevanzwert über einem bestimmten Schwellenwert und sind allfällige Zusatzbedingungen erfüllt, wird das Dokument der Kategorie zugeordnet. Liegt der Wert hingegen zu tief, so versucht das System, das Dokument einer anderen Kategorie zuzuordnen. Manche Dokumente sind für keines der Profile relevant genug. Solche Dokumente tauchen nicht im Katalogdienst auf, sind aber unter Umständen trotzdem zugreifbar – zum Beispiel über ein Suchsystem.

Wie hält sich der automatische Katalogdienst frisch?

Das Auffinden von neuen Dokumenten im Internet genügt noch nicht. Der Web-Roboter muss zudem periodisch die schon eingeordneten Dokumente besuchen. So kann festgestellt werden, ob sich der Inhalt geändert hat. Ein modifiziertes Dokument wird aus der aktuellen Kategorie entfernt und frisch eingeordnet, falls der Relevanzwert zwischen dem Dokument und dem aktuellen Kategorienprofil zu gering ausfällt. Gelöschte Dokumente müssen aus dem Katalogdienst entfernt und neu entdeckte Dokumente eingeordnet werden.

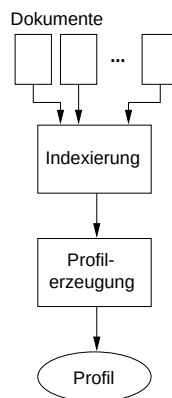
Woher stammen die Kategorienprofile?

Ein Profil besteht häufig aus einer Anzahl von Begriffen und beschreibt damit das Thema in einer Kategorie. Die Begriffe im Profil können von menschlichen Herausgebern beim Aufbau der ganzen Kategorienhierarchie festgelegt werden. Dabei werden direkt die Begriffe definiert.

Denkbar ist aber auch der indirekte Weg, der den Herausgebern die Auswahl von Begriffen für die Profile abnimmt. Ausgangspunkt ist eine Anzahl Dokumente, die inhaltlich das Thema einer Kategorie treffen. Die Wahl der charakteristischen Dokumente ist Aufgabe der

Herausgeber. Alles Übrige hingegen kann automatisiert werden: Zunächst werden die Dokumente indexiert. Anschliessend entsteht das Profil, indem wichtige Begriffe aus den Dokumenten zusammengestellt werden.

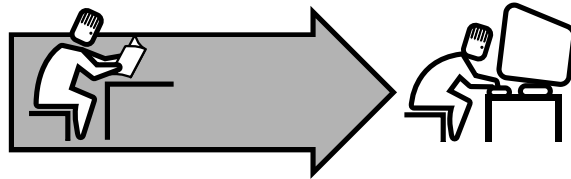
Die Komponente zur Profilerzeugung greift auf die Kriterien der Rangierungsprinzipien zurück. Das Vorgehen ist eigentlich identisch mit dem Verarbeiten einer Relevanzrückkoppelung. Dort wird ebenfalls aus einer Menge von relevanten Dokumenten eine neue Anfrage – ein Profil – zusammengestellt. Die Profile werden anschliessend in einem Index abgelegt. So können neue Dokumente in kürzester Zeit mit sehr vielen Profilen verglichen werden.



Was sieht eine Benutzerin vom Katalogdienst?

Wir wissen nun, wie Dokumente mit Hilfe von Profilen automatisch in den passenden Kategorien abgelegt werden. Eine Benutzerin des automatisierten Katalogdienstes bekommt dasselbe zu sehen wie beim manuell erstellten Katalogdienst. Die internen Datenstrukturen des Katalogsystems werden für die Benutzerin ebenfalls in Form der gewohnten Kategorienhierarchie präsentiert.

Im Praxisteil geht es in erster Linie um eine Gegenüberstellung zwischen Katalog- und Suchdiensten ...



Konkrete Katalogdienste im Internet

Im Internet steht eine Unzahl von Katalogdiensten zur Verfügung. Wiederum hängt die Wahl sehr von den eigenen Vorlieben bezüglich der Benutzerschnittstelle oder von den Interessensgebieten ab.

Gegenüberstellung: Such- und Katalogdienste

Vorteile von Katalogdiensten

Der wohl wichtigste Vorteil von Katalogdiensten: Falls ein Informationsbedürfnis durch eine Kategorie abgedeckt wird, kann die Suche auf diese Kategorie beschränkt werden.

Katalogdienste bieten sich ausserdem an, wenn man sich einen ersten allgemeinen Überblick über ein bestimmtes Thema verschaffen möchte und in diesem Bereich noch über kein ausgeprägtes Hintergrundwissen verfügt. In solchen Fällen sind häufig bereits die Namen der Kategorien von grossem Interesse, weil sie wertvolle Hinweise auf die Terminologie im entsprechenden Wissensgebiet liefern. Beispiel: Welches sind die wichtigsten Stilrichtungen in der Malerei?

Bei den manuell erstellten Katalogdiensten kann man zudem von einer gewissen Qualitätskontrolle der angebotenen Dokumente ausgehen. Ausserdem besteht die Möglichkeit, dass die Herausgeber zu jedem Dokument eine kurze Bewertung angeben. Dadurch treffen manuell erstellte Katalogdienste eine Vorauswahl. Benutzer werden nicht mit einer Fülle von nicht relevanten Daten überschwemmt.

Nachteile von Katalogdiensten

Die meisten Katalogdienste definieren ein möglichst vernünftiges Gerüst von Kategorien und hoffen, dass sich ein Grossteil der Benutzer in der festgelegten Hierarchie orientieren kann. Doch die gewählte Struktur muss nicht für jede Benutzerin die ideale sein. Es kann durchaus vorkommen, dass eine Benutzerin in einem völlig falschen Bereich nach dem gewünschten Dokument fahndet. Nicht selten ist auch der Fall, dass wichtige Webseiten nicht angemeldet sind, während eher unwichtige Seiten überbewertet werden und somit in zu vielen Kategorien vorkommen.

Ein weiteres Problem besteht beim Einordnen von Dokumenten in die Kategorien des Katalogdienstes. Die optimale Zuordnung ist oft unklar. In vielen Fällen empfehlen verschiedene Personen für das gleiche Dokument unterschiedliche Kategorien.

Verschmelzung von Suchdiensten und Katalogdiensten

Der Trend ist offensichtlich: Verschiedene Informationsdienste im Internet verschmelzen immer mehr zu einem einzigen Angebot. Viele Suchdienste bieten eine Mischform an. Sie präsentieren ihre Dokumentensammlung auch in kategorisierter Form. Benutzer haben die Wahl, ob sie im Katalogdienst blättern oder lieber eine Suche im Index durchführen möchten. Ein weiterer Hinweis auf die Verschmelzung zeigt sich bei den Suchmöglichkeiten von Katalogdiensten.

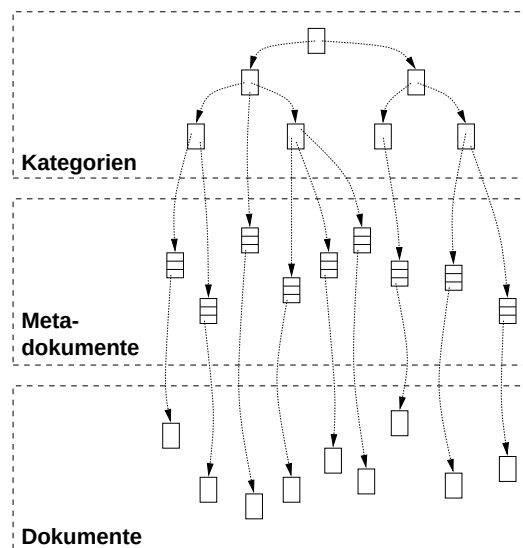
Suche in Katalogdiensten

Auch Katalogdienste stellen üblicherweise eine Funktion zur Stichwortsuche bereit. Die Suche in einem kombinierten Katalog- und Suchdienst kann folgendermassen ablaufen:

- Zuerst werden die Namen der Kategorien durchsucht. Kommt der Suchbegriff in einem Kategoriennamen vor, wird dieser gemeldet. Insofern stellen Kategorien nichts anderes dar als spezielle Dokumente, die auf andere (Unter-)Kategorien oder auf Metadaten verweisen.

- Im zweiten Schritt werden die Dokumenttitel und eventuell deren Kurzbeschreibungen durchsucht und die gefundenen Dokumente angezeigt. Das heisst, es wird nach Metadokumenten gesucht.
- Bleiben die ersten beiden Suchdurchgänge erfolglos, wird die Anfrage an einen bestehenden Suchdienst weitergereicht. Dazu gehen viele manuell erstellte Katalogdienste eine Allianz mit einem Suchdienst ein. Dem Benutzer werden vom Katalogdienst die gleichen Resultate präsentiert, die auch der Suchdienst geliefert hätte. Erst in diesem letzten Schritt werden konkrete Dokumente (zum Beispiel Webseiten) durchsucht.

Beispiel: Die Suche nach *Gitarre* liefert an oberster Stelle vielleicht die Kategorie *Unterhaltung / Musik / Instrumente / Saiteninstrumente / Gitarre* und erst weiter unten in der Rangliste ein Dokument mit dem Titel «Klangtherapie unter Einsatz von Panflöte, Harfe und klassischer Gitarre» in der Kategorie *Medizin / Alternative Heilpraktiken*.

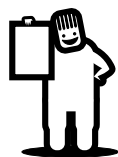
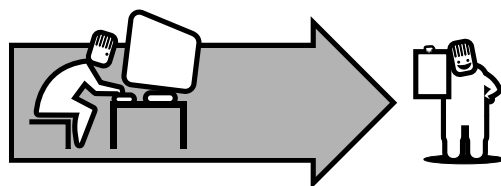


Fazit: Die Suche in einem Katalogdienst erfolgt in drei Ebenen von oben nach unten. Ein Suchdienst hingegen verwendet lediglich die

untersten beiden Ebenen. Durchsucht wird der Dokumentinhalt. In manchen Fällen kann mit Hilfe der Metadaten die Kollektion eingeschränkt werden. Zudem benötigt der Suchdienst die Metadaten, um die Rangliste zu erstellen. Die oberste Ebene mit der Kategorienhierarchie existiert nur bei den Katalogdiensten. Katalogdienste stecken Zusatzinformationen in die angebotene Dokumentenkollektion, indem sie die Dokumente gliedern.

Zu beachten ist dabei, dass die Metadokumente bei manchen Katalogdiensten ganze Web-Sites und nicht einzelne Webseiten beschreiben.

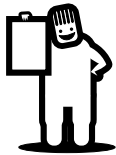
Mit all diesen Informationen sollten die Internet-Anwender nun bereit sein für die Lösung ihrer Probleme ...



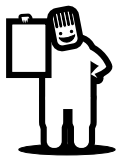
Ich suche den Namen der Autorin des Kinderbuchs «Die schwarzen Brüder». Offenbar ist dieses Informationsbedürfnis zu spezifisch für einen Katalog. Gerade die manuell erstellten Katalogdienste decken niemals dieselbe Menge an Dokumenten ab wie die grossen Suchdienste, sondern beschreiben häufig ganze Web-Sites. Der von mir verwendete Katalog enthält keine Angaben zum gesuchten Buch.

Ich könnte nun mein Glück bei einem anderen Katalog versuchen. Doch vermutlich bin ich für mein Informationsbedürfnis mit einem Suchdienst besser bedient. Tatsächlich – OMNISEARCH liefert mir auf die Anfrage *“die schwarzen brüder“* prompt die gesuchten Informationen. Das Buch wurde von Lisa Tetzner geschrieben. Ausserdem erfahre ich, dass Tetzner mit Kurt Kläber verheiratet war. Kläber, der als geflüchteter Antifaschist das Pseudonym Kurt Held wählte, schrieb den berühmten Jugendroman «Die rote Zora».

So bin ich glücklicherweise direkt zur Antwort auf meine Frage gelangt. Falls ich auch mit dem Suchdienst nicht weitergekommen wäre, hätte ich mich auf die Suche nach einem speziellen Verzeichnis von Kinderbüchern oder einem Bibliothekskatalog gemacht.



Mein Dopingproblem ist gelöst! Ich habe bei meinem ersten Versuch mit *Radfahren* die falsche Kategorie gewählt. Jetzt habe ich mir überlegt, dass eine Liste mit gesperrten Medikamenten vielleicht von einer Ärztevereinigung veröffentlicht wird. In der Kategorie *Medizin / Doping* bin ich dann fündig geworden. Allerdings ist es immer Ansichtssache, wo ein Dokument am besten eingeordnet werden sollte. In einem anderen Katalogdienst ist die Dopingliste möglicherweise unter *Sport / Doping* abgelegt.

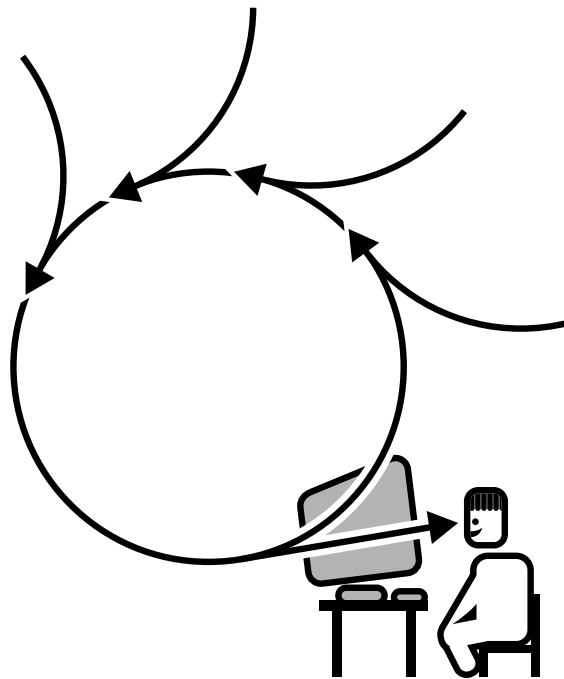


Nun ist mir klar, was mein Problem war. «Apache» ist auch der Name eines Web-Servers. Deshalb wurde ich mit Begriffen wie Perl oder SSL konfrontiert, denn sie stehen im Zusammenhang mit Web-Servern. Bei meiner Stichwortsuche im Katalog hätte ich die Liste der gefundenen Kategorien genauer studieren müssen. Die erste Kategorie lautete *Computers / Software / Internet / World Wide Web / Servers / Unix / Apache*. Da erkennt man sofort, dass es nicht um Indianer geht. Stattdessen wähle ich neu die Kategorie *Society and Culture / Cultures / American / Native American / Tribes, Nations, and Bands / Apache*.

Es gibt auch eine andere Möglichkeit, mein Problem zu umgehen: Ich kann zunächst in der Kategorie *Society and Culture* einsteigen und anschliessend eine Stichwortsuche nach *Apache* ausschliesslich innerhalb dieser Kategorie durchführen. Dadurch bleiben mir der Web-Server und auch allfällige Kategorien über Apache-Hubschrauber erspart.

Kapitel 8

Push-Dienste

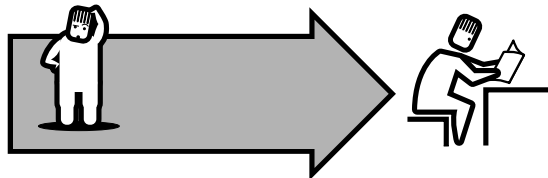




Mich fasziniert das Wachstum und die rasante Entwicklung des Internets. Deshalb will ich fortlaufend über die neuesten statistischen Untersuchungen und demografischen Erhebungen im Zusammenhang mit dem Internet informiert sein. Das Problem dabei ist, dass solche Untersuchungen von den verschiedensten Stellen durchgeführt und veröffentlicht werden – zum Beispiel von Universitäten, von Consulting-Firmen oder von öffentlichen Ämtern. Natürlich habe ich keine Lust, permanent bei diesen Stellen zu prüfen, ob eine neue Veröffentlichung vorliegt.

Zum Glück gibt es die Push-Dienste, die mir diese mühselige Arbeit abnehmen sollen. Folglich benutze ich einen dieser Dienste, die verschiedene Kanäle zu vorgegebenen Themen anbieten. Ich wähle den Kanal «Internet» aus und werde in der Folge mit einem nicht abreissenden Strom von Dokumenten versorgt. Leider erhalte ich aber viel zu viele Dokumente, die alle nur möglichen Aspekte des Internets abdecken! Nur selten ist auch etwas für mich Relevantes dabei. Ist der Push-Dienst vielleicht nicht das richtige Werkzeug für mich?

Schauen wir uns zunächst an, wie Push-Dienste funktionieren ...



Hol-Prinzip und Bring-Prinzip

In diesem Kapitel geht es um ein Bedürfnis, das sich grundlegend von den bisherigen Bedürfnissen unterscheidet. Bisher sollte mittels Katalog- und Suchdiensten ein spontanes Informationsbedürfnis auf der Stelle befriedigt werden. Nun beschäftigen wir uns mit einem über eine bestimmte Zeit unveränderten Informationsbedürfnis, das fortlaufend befriedigt werden soll.

Für dieses andersartige Bedürfnis benötigen wir auch ein anderes Werkzeug. Für den Zugriff auf einen Katalog- oder einen Suchdienst

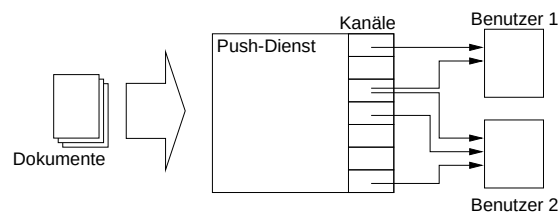
muss die Benutzerin selber aktiv werden. Sie muss den entsprechenden Informationsdienst aufsuchen und eine Anfrage stellen. Das Vorgehen wird *Pull-Prinzip* (Hol-Prinzip) genannt – die Benutzerin *holt* sich die Information. Neu soll die Information jedoch ohne das Zutun der Benutzerin *gebracht* werden. Ein Informationsdienst ist gefragt, der nach dem *Push-Prinzip* (Bring-Prinzip) arbeitet.

Ein nach dem Push-Prinzip funktionierendes Medium aus dem Alltag ist beispielsweise das Fernsehen. Wer den Wetterkanal wählt, wird auf der Stelle mit Informationen zur aktuellen Wetterlage versorgt. Es ist nicht nötig, dem Fernseher zuerst mitzuteilen, was man sehen möchte.

Push-Dienste – so nennen wir also Informationsdienste im Internet, die nach dem Push-Prinzip funktionieren. Für die Push-Dienste sollte etwas schon jetzt klar sein: Sie eignen sich in erster Linie für eine spezifische Art von Informationsbedürfnissen: sich über einen Themenbereich informiert halten. Für die Suche nach der exakten Höhe des Matterhorns sind Push-Dienste ungeeignet. Ausser man ist bereit, so lange zu warten, bis die Information zufällig zugestellt wird.

Push-Dienste der einfachsten Art

Eine einfache Art von Push-Diensten im Internet funktioniert ganz ähnlich wie ein Fernseher. Es wird eine Zahl von fest vorgegebenen Kanälen zu relativ allgemein gehaltenen Themenkreisen angeboten, zum Beispiel Wetter, Sport, Kunst und andere.



Die Benutzer können die für sie relevanten Kanäle abonnieren und erhalten daraufhin die entsprechenden Nachrichten laufend zugestellt. In der Regel erfolgt die Zustellung über spezialisierte Webseiten oder per E-Mail.

Die Benutzer von Push-Diensten der einfachsten Art haben die Freiheit bei der Wahl der Kanäle. Darüber hinaus stehen ihnen aber keine Einflussmöglichkeiten zur Verfügung. Insbesondere können sie die Themen nicht gemäss ihren eigenen Bedürfnissen einengen oder erweitern.

Push-Dienste der flexibleren Art

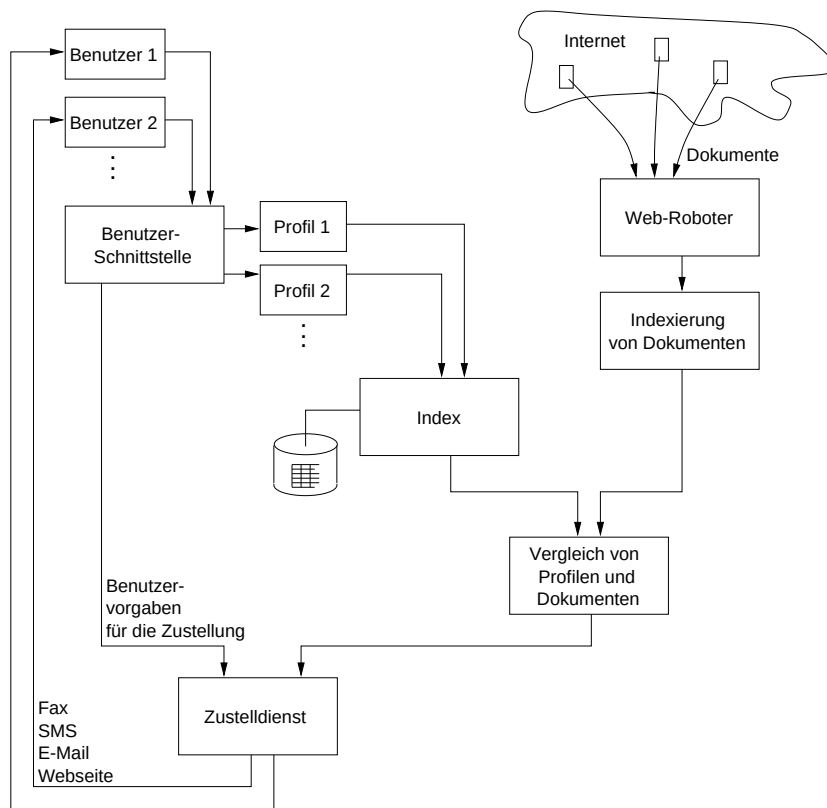
Andere Arten von Push-Diensten erlauben mehr Flexibilität. Die Benutzer haben die Möglichkeit, ihre Informationsbedürfnisse festzulegen und anzupassen. Dazu wählen sie geeignete Suchbegriffe, die das Informationsbedürfnis beschreiben. Mit den gewählten Begriffen entsteht ein so genanntes Benutzerprofil. Oft stehen auch zahlreiche Formen der Zustellung von Dokumenten zur Wahl.

In einem Push-System werden drei Hauptkomponenten unterschieden: Bezug und Erschliessung von Dokumenten, Filterung von Dokumenten anhand von Benutzerprofilen und Zustellung von Dokumenten basierend auf entsprechenden Angaben von den Benutzern.

Dokumentenbezug und Erschliessung

Das Vorgehen ist dasselbe wie bei Such- und Katalogdiensten. Ein Web-Roboter ist dafür zuständig, Dokumente im Internet aufzufinden und an die Indexierungskomponente weiterzureichen. Der Web-Roboter kann auf die unterschiedlichsten Angebote zugreifen: zum Beispiel auf normale Webseiten oder auf Mail- und News-Archive oder auf spezifische Datenbanken.

Auch bei Push-Diensten ist die Trennung zwischen dem Push-System als Werkzeug und der Dokumentensammlung wichtig. Manche Dienste versuchen, das Internet in seiner ganzen Breite an Themen abzudecken, und bieten folglich eine horizontale Kollektion an. Andere Dienste konzentrieren sich auf ein enges Gebiet und benötigen dafür vielleicht sogar Zugriff auf sehr spezifische, kostenpflichtige Datenbanken.



Filterung

Wir haben im Zusammenhang mit den Katalogdiensten Profile kennen gelernt. Diese haben die Aufgabe, das Thema einer Kategorie mit einer Menge von Begriffen inhaltlich zu beschreiben. Auch Push-Dienste benötigen Profile. Sie sollen das Informationsbedürfnis eines Benutzers charakterisieren und stellen somit das Interessenprofil einer Person dar. Im einfachsten Fall bestehen die Profile aus einer Sammlung von charakteristischen Begriffen und einem Schwellenwert. Die Profilinformationen werden für den raschen Zugriff in einem Index abgelegt, so wie bei einem Suchdienst die Dokumente im Index abgelegt sind.

Die Filterung der neu erschlossenen Dokumente erfolgt anhand der Profile. Jedes neue Dokument wird mit allen Profilen verglichen. Der Vergleich stützt sich auf einige oder alle Rangierungsprinzipien. Daraus resultiert ein Relevanzwert. Je enger ein Dokument und ein Profil miteinander verwandt sind, desto höher fällt der Wert aus. Überschreitet der Relevanzwert den im Profil vorgegebenen Schwellenwert, so wird das Dokument an den Zustelldienst weitergereicht.

Zustellung

Eine eigene Komponente im Push-System ist für die Zustellung von Dokumenten an die jeweilige Benutzerin verantwortlich. Nach dem Vergleich zwischen neu erschlossenen Dokumenten und den Profilen erhält der Zustelldienst die Information, welche Dokumente an welche Personen auf welchem Weg weitergeleitet werden sollen.

Eine bequeme Art der Zustellung ist E-Mail. Dabei erhält die Benutzerin in regelmässigen Abständen eine Übersicht über die neu gefundenen Dokumente, die gemäss Profil relevant sind. Die Nachrichten fallen unterschiedlich umfangreich aus. Vielleicht wird der ganze Dokumentinhalt verschickt. Vielleicht erhält die Benutzerin auch nur die Dokumenttitel, eine Zusammenfassung und den URL. Das ist abhängig vom konkreten Push-Dienst oder kann sogar von der Benutzerin selbst festgelegt werden.

Neben E-Mail ist eine ganze Palette anderer Techniken und Medien für die Zustellung denkbar: Dokumente können auf einer spezifischen Webseite aufgelistet werden. Benutzerinnen entscheiden dann selber, wann sie sich informieren möchten. Für dringendere Informationen wie zum Beispiel Börsenberichte eignet sich die Zustellung via SMS (Short Message Service) auf Mobiltelefone. Eine weitere Option sind Nachrichten per Fax.

Interaktionen zwischen Benutzern und dem Push-System

Benutzer können das Verhalten eines Push-Systems in verschiedenen Bereichen beeinflussen:

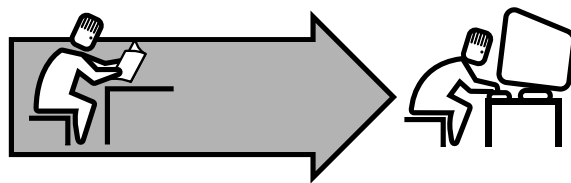
- *Profilerzeugung*: Bevor ein Push-Dienst mit seiner Arbeit beginnen kann, muss er das Informationsbedürfnis eines Benut-

zers kennen. Dazu geben die Benutzer eine Anzahl charakteristischer Begriffe vor und konstruieren so eine Art «Wunschdokument».

Wie bei den Katalogsystemen kann ein Profil auch mittels einer Menge von relevanten Dokumenten definiert werden. Eine Komponente zur Profilerzeugung bestimmt in solchen Fällen automatisch die wichtigen Begriffe in den Dokumenten und stellt daraus ein Profil zusammen.

- *Profilanpassungen:* Die Relevanzrückkoppelung wird bei einem Suchdienst angewendet, um aufgrund einiger relevanter Dokumente eine neue Anfrage zu erstellen. Dieselbe Technik ist auch im Zusammenhang mit Push-Diensten hilfreich. Ein Benutzer teilt dem Push-System mit, welche der zugestellten Dokumente für ihn tatsächlich relevant waren. Aufgrund der Rückmeldung kann das System das Benutzerprofil entsprechend anpassen. Dabei werden wichtige Begriffe aus den relevanten Dokumenten extrahiert und zum Profil hinzugefügt – das Profil eines Benutzers wird also laufend seinen Bedürfnissen angepasst.
- *Vorgaben zur Zustellung:* Benutzer können dem Zustelldienst vorschreiben, in welcher Form, auf welchem Weg und mit welcher Häufigkeit relevante Dokumente auszuliefern sind. Beispielsweise möchte man vielleicht nicht alle zehn Minuten eine E-Mail erhalten, sondern stattdessen lieber die gesammelte Dokumentenliste einmal täglich in einer längeren Nachricht begutachten.

Kommen wir zu den praktischen Hinweisen im Zusammenhang mit Push-Diensten ...



Push-Dienste in der Praxis

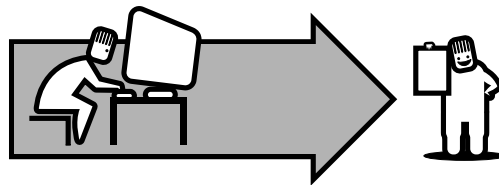
Ein Push-Dienst versorgt seine Benutzer laufend mit den neusten Informationen zu einem Gebiet. Konkrete Push-Dienste unterscheiden sich vor allem in der Wahl der Dokumentensammlung, in den möglichen Zustellungsarten, im Vorgehen für die Festlegung von Profilen und im Umfang der Einflussmöglichkeiten durch die Benutzer.

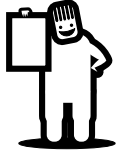
In der Praxis gibt es Push-Dienste, die auf Pull-Diensten basieren. Beispiel: Man installiert auf seinem Privat-PC einen Software-Assistenten, der für die Suche im Internet ausgelegt ist. Dann gibt man dem Assistenten eine Zahl von Suchbegriffen (ein Profil) mit und schickt ihn auf die Reise. Er wird eine Reihe von Such- und Katalogdiensten im Netz besuchen und ihnen jeweils die vorgegebene Anfrage stellen. Zum Schluss gehen die gesammelten Resultate an den Benutzer zurück. Nun kann man den Assistenten anweisen, dieselbe Recherche zweimal pro Woche durchzuführen. Resultat: Es handelt sich um ein aktives Werkzeug, das auf passive Informationsdienste zurückgreift.

Mailing-Listen und Newsgroups

Im weitesten Sinn funktionieren auch die im Internet angebotenen Mailing-Listen und Newsgroups ähnlich wie einfachste Push-Dienste. Man abonniert eines der vorgegebenen Themen und wird fortan mit Artikeln aus diesem Bereich beliefert. Allerdings zielen diese Foren in der Regel eher auf den Dialog ab, sodass die Teilnehmer die Möglichkeit haben, eigene Artikel beizusteuern. Die aktive Teilnahme der Beteiligten ist häufig sehr erwünscht.

Zum Abschluss wird wie gewohnt das Anwenderproblem gelöst ...



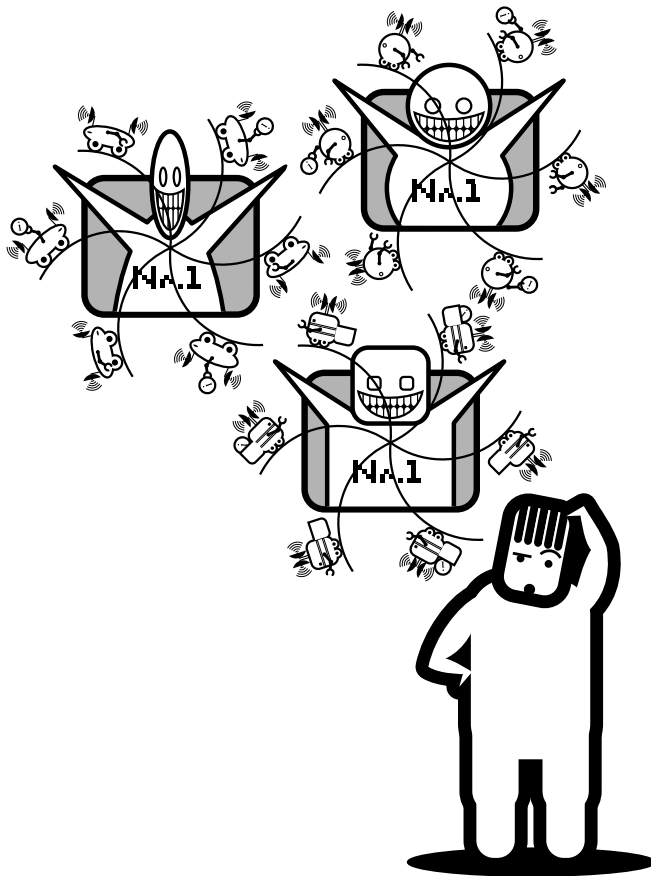


Die Lösung für mein Problem der Internet-Statistiken ist recht einfach. Ich muss versuchen, einen besseren Push-Dienst zu finden, der mehr Flexibilität bietet. Dann kann ich ein Profil definieren, das mein Informationsbedürfnis mit Hilfe von geeigneten Begriffen sehr spezifisch beschreibt. Bei der Zusammenstellung der Suchbegriffe für das Profil achte ich auf dieselben Regeln, die auch für die Suchanfragen bei Suchdiensten gelten.

Wenn ich einen Push-Dienst mit dem einzigen Kanal «Internet» verwende, so erhalte ich dasselbe Spektrum an Dokumenten, wie wenn ich bei OMNISEARCH die Anfrage *internet* durchführe. Besser wäre eine Anfrage mit einer grossen Anzahl von charakteristischen Begriffen. Bei einem geeigneten Push-Dienst kann ich ein entsprechendes Benutzerprofil selber festlegen. Das Profil wird Begriffe wie *internet*, *world wide web*, *statistics*, *demographics*, *growth*, *usage* usw. enthalten.

Kapitel 9

Evaluation von Informationsdiensten



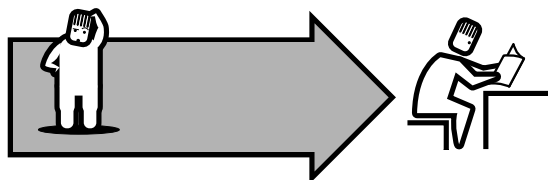


Nach der Lektüre der letzten Kapitel ist mir klar, worauf ich bei der Benutzung der verschiedenen Werkzeuge zum Informationszugriff im Internet achten muss. Ein Problem allerdings bleibt bestehen: Im Internet steht eine riesige Auswahl von Werkzeugen bereit. Welche soll ich wählen?

Der konkrete Umgang mit den Werkzeugen unterscheidet sich grundsätzlich nicht. Trotzdem erhalte ich von allen Seiten – von meinen Freunden und Bekannten – die verschiedensten Ratschläge. Manche Leute schwören auf einen Informationsdienst, den andere verfluchen. Doch all diese Meinungen sind sehr subjektiv gefärbt.

Ich möchte einen Informationsdienst nicht anhand von oberflächlichen, subjektiven Kriterien bewerten. Stattdessen suche ich nach grundlegenden Entscheidungshilfen, die mich bei der Beurteilung unterstützen.

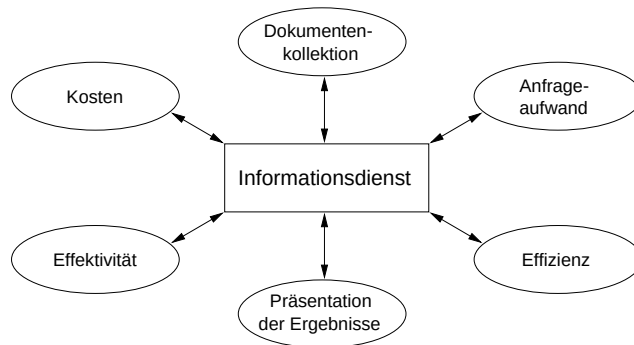
Um dem Anwender weiterzuhelfen, behandeln wir hier einige Ansatzpunkte, wie man Informationsdienste beurteilen kann ...



Kriterien zur Bewertung von Informationsdiensten

Internet-Anwender sind mit einer Fülle von Informationsdiensten konfrontiert. Da stellt sich sofort die Frage: Nach welchen Kriterien können Informationsdienste bewertet und miteinander verglichen werden?

Sechs übergeordnete Kriterienbereiche spielen bei der Evaluation von Informationsdiensten eine Rolle. In manchen Bereichen können weitere, untergeordnete Aspekte unterschieden werden.



Dokumentenkollektion

Das beste Informationssystem nützt wenig, wenn die damit erschlossene Dokumentenkollektion den jeweiligen Ansprüchen nicht genügt. Ein offensichtliches Merkmal von Dokumentenkollektionen ist die *Ausrichtung*: Handelt es sich um eine horizontale oder um eine vertikale Kollektion? Welche Themenbereiche werden abgedeckt?

Der *Abdeckungsgrad* sagt aus, wie viele der insgesamt existierenden Dokumente im jeweiligen Gebiet durch die Dokumentenkollektion erschlossen werden.

Weitere Aspekte beziehen sich auf die einzelnen Dokumente innerhalb der Kollektion: zum Beispiel *inhaltliche Qualität*, *Aktualität* und *Strukturierung*. Die Strukturierung von Dokumenten lässt sich oft für einen effektiveren Informationszugriff ausnützen, weil bestimmte Angaben wie Personennamen oder Adressen eindeutig identifiziert werden können. Die Aktualität der Dokumente ist je nach Anwendung unterschiedlich wichtig.

Anfrageaufwand

Ein Benutzer muss Vorarbeit leisten, um eine Anfrage an einen Informationsdienst zu stellen. Der Anfrageaufwand ist abhängig von verschiedenen Aspekten des Informationssystems. Vorteilhaft ist beispielsweise eine übersichtliche und intuitiv verständliche Benutzerschnittstelle. Bei manchen Systemen dagegen muss man sich als Be-

nutzerin zuerst durch schlecht verfasste Hilfeseiten quälen, bis man in der Lage ist, eine Anfrage zu stellen.

Ebenso spielt der Umfang der Indexierungskomponente eine Rolle. Systeme ohne Buchstabenumwandlung, Wortzerlegung und Wortnormalisierung wälzen Mehrarbeit auf die Benutzer ab. Die Benutzer müssen sich selbst um Flexionen und Umlaute oder Akzente kümmern und diese Problematik in den Suchanfragen berücksichtigen.

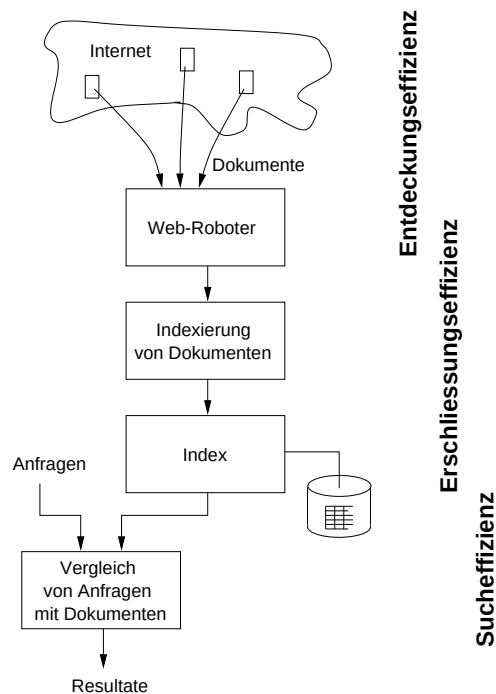
Nicht zuletzt kann der Anfrageaufwand weiter verringert werden, wenn ein System mit hilfreichen Zusatzfunktionen ausgestattet ist. Die automatisierte Relevanzrückkoppelung beispielsweise entlastet die Benutzer zusätzlich.

Effizienz

Die Effizienz – die Schnelligkeit – eines Informationssystems misst sich in in erster Linie an drei Grössen:

- *Sucheffizienz*: Gemeint ist die Antwortzeit des Systems, also die Zeit zwischen dem Erhalt einer Anfrage und der Fertigstellung der Antwort. Die benötigte Zeit für die Verarbeitung von Suchanfragen hängt einerseits von der Anzahl der Suchbegriffe ab, denn für jeden Begriff muss der Index konsultiert und die gefundenen Einträge müssen miteinander kombiniert werden. Andererseits hängt die Antwortzeit auch von der Wahl der Suchbegriffe ab. Für sehr häufige Begriffe existieren im Index sehr lange Listen, die untersucht werden müssen.
- *Entdeckungseffizienz*: Ein neues Dokument entsteht. Wie lange dauert es, bis der Web-Roboter das Dokument gefunden hat? Diese Zeitspanne wirkt sich auch auf die Aktualität der angebotenen Dokumentensammlung aus.
- *Erschliessungseffizienz*: Nach der Entdeckung und dem Bezug müssen neue Dokumente indexiert und in den Index des Systems eingefügt werden. Je nach System sind vielleicht weitere Vorarbeiten nötig, bis Dokumente zum Auffinden bereitstehen. Diese Vorgänge werden unter dem Begriff der Erschliessung zu-

sammengefasst. Die Erschliessungseffizienz gibt Auskunft über die Schnelligkeit, mit der diese Arbeiten durchgeführt werden.



Präsentation der Ergebnisse

Nachdem ein Informationsdienst eine Anfrage bearbeitet und die Antwort zusammengestellt hat, werden die Resultate aufbereitet und meist in Form einer Webseite der Benutzerin zugestellt. Die Präsentation der Ergebnisse kann mehr oder weniger übersichtlich und benutzerfreundlich ausfallen. Manche Benutzerschnittstellen stellen die Suchresultate mittels aufwendiger Grafiken dar. Das macht oft Spass, führt aber nicht unbedingt schneller zum Ziel. Manchmal ist auch die Grösse der Ergebnisse wichtig. Zum Beispiel wird sich ein Privatanwender mit einer langsamen Modemverbindung daran stören, wenn der verwendete Suchdienst zu umfangreiche Ranglisten liefert.

Effektivität

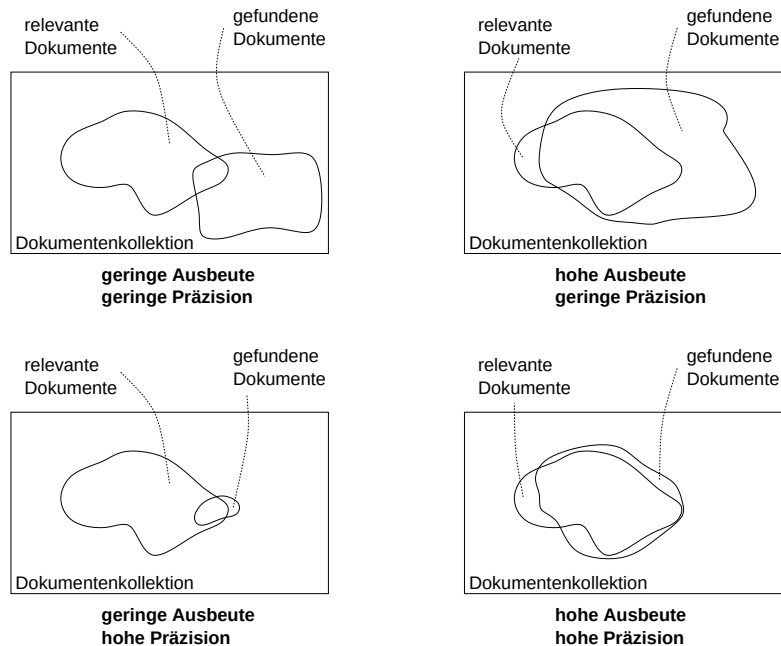
Ein sehr wichtiges Kriterium ist die Sucheﬀektivität. Mit der Eﬀektivität meint man die Fähigkeit eines Informationssystems, die von einer Benutzerin gewünschten Dokumente zu liefern. *Ausbeute* und *Präzision* sind zwei Masse, die sich zur Bewertung der Eﬀektivität eines Informationssystems eignen. Die beiden Masse stehen immer im Zusammenhang mit einer Anfrage sowie der Menge der *gefundenen* Dokumente. Es handelt sich dabei um die bestrangierten Dokumente, die der Benutzer betrachtet. Die Zahl der betrachteten Dokumente hängt vom Informationsbedürfnis ab.

Ziel einer Anfrage ist es, möglichst alle relevanten Dokumente aus der Dokumentenkollektion zu finden. Meistens wird man aber nur einen Teil der gewünschten Dokumente finden. Die Ausbeute gibt an, welchen Anteil der an und für sich in der Dokumentenkollektion vorhandenen relevanten Dokumente man mit einer Anfrage gefunden hat.

Auch wenn eine Anfrage alle relevanten Dokumente findet – also hundertprozentige Ausbeute aufweist –, heisst das noch lange nicht, dass die Benutzerin zufrieden ist. Liefert das Suchsystem neben den relevanten auch eine Unzahl irrelevanter Dokumente, so gehen die relevanten Dokumente möglicherweise unter. Gefragt ist also auch eine hohe Präzision. Mit Präzision bezeichnen wir den Anteil der mit einer Anfrage gefundenen Dokumente, die tatsächlich relevant sind.

Das perfekte System bietet demnach maximale Ausbeute und maximale Präzision gleichzeitig. Das heisst, es liefert alle relevanten Dokumente, die in der Kollektion überhaupt existieren. Zudem taucht kein einziges irrelevantes Dokument in der Antwort auf. Leider bleibt das perfekte System hypothetisch, denn die beiden Grössen sind gegenseitig voneinander abhängig. Steigert man in einem System die Ausbeute, so geht das auf Kosten der Präzision und umgekehrt.

Viele Suchsysteme arbeiten präzisionsorientiert. Solche Systeme liefern nur wenig irrelevante Dokumente und vermitteln dem Benutzer ein Erfolgsgefühl. Da der Benutzer nur über eine lokale Sicht auf die gefundenen Dokumente verfügt, entgehen ihm viele nicht gefundene, aber ebenfalls relevante Dokumente.



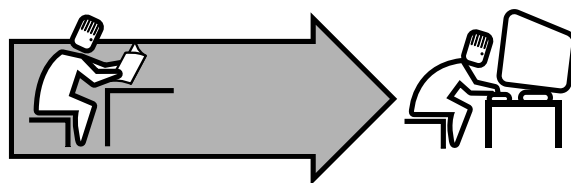
Bei Push-Systemen sind die Masse Ausbeute und Präzision anders definiert. Der Grund: Push-Systeme erzeugen keine Rangliste. Stattdessen entscheiden die Systeme, ob ein Dokument dem Benutzer zugestellt werden soll oder nicht. Deshalb zählt man bei Push-Systemen, wie viele der zugestellten Dokumente relevant und wie viele irrelevant sind. Durch den Vergleich der beiden Zahlen erhält man ein Mass für die Effektivität eines Systems. Ausserdem kann man die Zahlen unterschiedlich gewichten und damit mehr Wert auf Ausbeute oder Präzision legen.

Kosten

Das letzte Kriterium ist weniger wichtig für die Benutzer, aber umso wichtiger für die Betreiberinnen von Informationsdiensten. Systeme unterscheiden sich in Bezug auf den Entwicklungsaufwand und die Unterhaltskosten. Ein manuell betriebener Katalogdienst zum Beispiel verursacht hohe Unterhaltskosten.

Für die Benutzer relevanter in diesem Zusammenhang ist die Frage, wie sich ein Informationsdienst finanziert. Ein öffentlich zugängliches Angebot finanziert sich typischerweise mit Werbung in allen Varianten, mit der die Benutzer konfrontiert werden. Dagegen bleiben Benutzer von eigens eingekauften Systemen in einem Intranet normalerweise von Werbung verschont.

Einige Bemerkungen in Bezug auf die Praxis sind noch nötig, bevor der Internet-Anwender das Kapitel beschliesst ...



Informationsdienste in der Praxis beurteilen

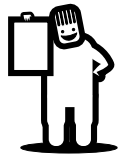
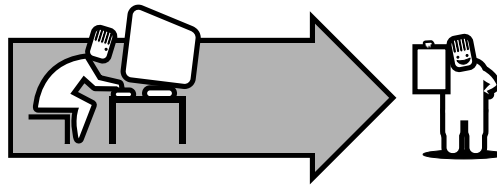
Für eine Benutzerin ist es in der Praxis nicht einfach, einen Informationsdienst objektiv zu bewerten. Der Abdeckungsgrad einer Dokumentensammlung beispielsweise lässt sich lediglich abschätzen. Man muss sich diesbezüglich oft auf die Angaben der Hersteller verlassen. Trotzdem verschafft man sich mit der Zeit ein Gespür für die Qualität der Dokumentensammlung. Auch die inhaltliche Qualität und die Aktualität der angebotenen Dokumente lässt sich einfach erkennen.

Der Anfrageaufwand ist das wohl augenfälligste Kriterium für einen Benutzer. Mittels einiger Beispielanfragen kann recht einfach der Umfang der Indexierung bestimmt werden. In Bezug auf die Effizienz von Informationsdiensten ist in erster Linie die Sucheffizienz von Bedeutung. Sowohl von der Sucheffizienz als auch von der Präsentation der Ergebnisse erhält man bei jeder Suchanfrage einen Eindruck.

Das in den meisten Fällen wichtigste Kriterium ist die Sucheffektivität. Ein Benutzer kann sich bei regelmässiger Anwendung eines Suchdienstes vielleicht einen Eindruck davon verschaffen, welche

Rangierungsprinzipien angewendet werden. Für eine objektive Bewertung der Effektivität sind jedoch aufwendige Experimente nötig. Dabei lässt man verschiedene Suchsysteme mit der gleichen Dokumentenkollektion gegeneinander antreten und vergleicht die Resultate auf eine vorgegebene Serie von Anfragen. Beim «Herumspielen» mit einem System sind hingegen zu viele zufällige Faktoren im Spiel, die eine objektive Bewertung verunmöglichen.

Der Anwender wird nun das Wichtigste nochmals kurz zusammenfassen ...



Was ist für mich wichtig? In erster Linie muss ich die passende Dokumentenkollektion für mein Informationsbedürfnis wählen. Dann steht die Wahl des Informationssystems an: Verwende ich ein Such-, ein Katalog- oder ein Push-System? Anhand dieser Entscheidungen wähle ich einen konkreten Informationsdienst aus der Fülle an Werkzeugen im Internet.

Zudem scheint mir wichtig, dass ich mich bei den tagtäglich verwendeten Werkzeugen auf eine bescheidene Anzahl konzentriere. Dafür versuche ich, diese Dienste immer besser kennen zu lernen, indem ich Beispielsuchanfragen durchführe und die Hilfeseiten lese. Ausserdem stelle ich eine Tabelle mit den wichtigsten Eigenschaften der von mir gewählten Dienste zusammen. Eine solche Tabelle gibt mir Auskunft darüber, wo ich den entsprechenden Informationsdienst im Internet finde. Ausserdem notiere ich darin alles, was ich über die erschlossene Dokumentenkollektion weiss. Weiter sind bei einem Suchdienst einige Angaben zum Umfang der Indexierung wichtig. Wird eine Buchstabenumwandlung durchgeführt? Werden zusammengesetzte Wörter zerlegt und normalisiert oder muss ich als

Benutzer selber darauf achten? Ein weiteres Kriterium sind die Rangierungsprinzipien. Was wird beim Abschätzen der Relevanz eines Dokuments berücksichtigt?

Für die Suchdienste OMNISEARCH und NEWSSEEKER sieht meine Tabelle so aus:

Dienst	OMNISEARCH	NEWSSEEKER	...
URL	http://...	http://...	...
KOLLEKTION	horizontal, Webseiten allg., mehrsprachig	vertikal, News-Artikel, mehrsprachig	...
SYSTEM	Suchsystem	Suchsystem	...
INDEXIERUNG	Wortextraktion	Buchstabenumwandlung, Wortextraktion, Wortzerlegung, Wortnormalisierung	...
STOPPWÖRTER	werden ignoriert	werden ignoriert	...
RANGIERUNG	keine feste Aussage möglich. Sicher Rangierungsprinzipien 1, 2 und 3. Prinzip 5 mittels Phrasensuche.	vermutlich Rangierungsprinzipien 1–6	...
HILFE	http://...	http://...	...

Kapitel 10

Suchtipps



Suchdienste, Katalogdienste, Push-Dienste, Indexierung, Web-Roboter, Rangierungsprinzipien, Profile, Anfragen usw. – die bisherigen Kapitel warteten mit einer Fülle von Informationen und Begriffen für die Leserschaft auf. Funktionsweise und Verwendungszweck der drei wichtigsten Werkzeuge für den Informationszugriff im Internet wurden vermittelt. Der Blick hinter die Kulissen von Informationsdiensten sollte zu einem zielgerichteten Umgang mit diesen Hilfsmitteln führen.

In diesem letzten Kapitel geht es darum, den Notproviant für den Alltag der Informationssuche im Internet zu schnüren. Anhand von zehn Fallbeispielen werden die wichtigsten Merkmale vorgestellt, an die man bei einer Recherche denken sollte. Gleichzeitig bieten die zehn Suchtipps eine Kurzzusammenfassung der vorhergehenden Kapitel.

Alle zehn Suchtipps sind identisch aufgebaut. Zunächst wird ein typisches Anwenderproblem geschildert. Es folgen der Suchtipp sowie etwas Hintergrundinformation zum Thema, bevor die Lösung zum Anwenderproblem präsentiert wird.



Der Frühling steht vor der Tür. Der Garten ruft. Ich möchte mir im Internet ein paar neue Ideen besorgen. Kurzerhand konsultiere ich NEWSSEEKER mit der Anfrage *Gartentipps*. Ich arbeite gerne mit NEWSSEEKER, weil mir der ausführliche Indexierungsprozess viel Arbeit abnimmt. Doch in diesem Fall werde ich arg enttäuscht. Die meisten gefundenen Dokumente handeln von Kindergärten! Irgendwo wird ein neues Kindergartenkonzept eingeführt. An einem anderen Ort gab es einen Brand im Kindergarten. Immerhin in einem Dokument geht es um Gartenzwerge. Aber auch dieses Dokument ist nicht relevant: Einem rückfälligen Gartenzwergrandalierer wird der Prozess gemacht.

Richtige Dokumentensammlung wählen!

Im Allgemeinen ist der Fehler in solchen Fällen nicht beim Suchsystem zu finden. Das Problem liegt bei der Wahl der Dokumentensammlung. Man sollte sich immer bewusst sein, was für eine Kollektion durch den gerade verwendeten Informationsdienst abgedeckt wird. Auch das beste Suchsystem kann ein Informationsbedürfnis nicht befriedigen, wenn es keine relevanten Dokumente in der erschlossenen Kollektion gibt.



Alles klar. Ich darf natürlich nicht denjenigen Suchdienst wählen, der mir am besten gefällt. Wichtiger ist, dass der Dienst eine für mich geeignete Dokumentensammlung anbietet. Offenbar gibt es in der gesamten Kollektion von NEWSSEEKER nicht ein Dokument mit dem Begriff *Gartentipps*. Wegen der Wortzerlegung werden stattdessen Dokumente mit Kindergärten oder Gartenzwerge gemeldet. Ein Suchsystem ohne Wortzerlegung hätte gar nichts gefunden.

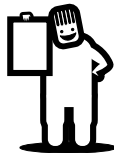
Für mein Problem benütze ich also besser OMNISEARCH. Oder noch besser einen spezifischen Suchdienst im Bereich von Gartenbau und -pflege, falls sich so etwas finden lässt.



Reiseplanung ist angesagt! Meinen nächsten, wohlverdienten Urlaub möchte ich auf Java verbringen. Deshalb suche ich bei OMNISEARCH nach Informationen über diese Insel Indonesiens. Leider scheint das ein völlig hoffnungsloses Unterfangen zu sein. Ich finde Millionen von Dokumenten zu einer Programmiersprache namens Java. Was tun?

Richtiges Werkzeug benützen!

Wir haben drei Grundtypen von Werkzeugen kennen gelernt. Suchdienste liefern auf eine Anfrage eine Rangliste mit möglichst relevanten Dokumenten. Katalogdienste sind hilfreich, wenn man sich einen allgemeinen Überblick über einen Themenbereich verschaffen möchte, und wenn das Informationsbedürfnis mit einer Kategorie des Katalogs übereinstimmt. Push-Dienste versorgen die Benutzer fortlaufend mit den aktuellsten Informationen zu einem bestimmten Thema. Vor jeder Recherche sollte man sich überlegen, welches Werkzeug wohl am ehesten zum Ziel führt.



Dann lasse ich mir das nochmals durch den Kopf gehen. Ich habe nach der Insel Java gesucht und die Programmiersprache gefunden. Ausserdem gibt es auch noch eine Kaffeesorte mit demselben Namen. Drei getrennte Themenbereiche also – ein klarer Fall für einen Katalogdienst. Ein Katalog dürfte genau diese Kategorien anbieten. In meinem Fall heisst die passende Kategorie *Reiseinformationen / Nach Regionen / Asien / Indonesien / Java*. Dort stosse ich auf weitere Unterkategorien zu verschiedenen Aspekten wie Kunst, Unterhaltung, Sport, Tourismus und Transportmöglichkeiten auf Java.

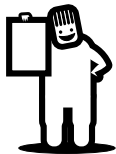


Am 4. Juni 1996 explodierte die Weltraumrakete Ariane 5 etwa 40 Sekunden nach dem Start. Der Schaden belief sich auf rund 500 Millionen Dollar. Ich möchte wissen, was der Grund für die Explosion war.

OMNISEARCH liefert gegen 100 000 Dokumente auf die Anfrage *ariane*. Weit und breit ist keine Spur von der gesuchten Information. Mit dem Suchbegriff *explosion* komme ich ebenfalls nicht zum Ziel. Und auch mit der kombinierten Anfrage *rocket explosion* bleibt der Erfolg aus.

Viele präzise Suchbegriffe verwenden!

Wer Mühe hat, auch nur ein einziges relevantes Dokument zu finden, soll möglichst viele Suchbegriffe verwenden. Man soll aber nicht irgendwelche allgemeinen Begriffe zur Anfrage hinzufügen, sondern eine der folgenden Strategien verfolgen: (1) Möglichst viele Suchbegriffe verwenden, die mit grosser Wahrscheinlichkeit in einem relevanten Dokument vorkommen. (2) Möglichst spezifische Suchbegriffe sowie Phrasen verwenden. Dabei soll bereits das Vorkommen eines einzigen dieser Suchbegriffe die Relevanz des entsprechenden Dokuments gewährleisten.



Ich stelle also eine Liste mit charakteristischen Suchbegriffen zusammen: Es geht um die Rakete Ariane 5, daraus kann ich vorzugsweise eine Phrase bilden. Die Explosion hat im Juni 1996 während des Starts stattgefunden. Folglich lautet die endgültige Anfrage "*ariane 5" rocket june 1996 explosion failure launch*". Diese Anfrage liefert unter den ersten Treffern Bilder vom «teuersten Feuerwerk aller Zeiten» sowie die Antwort zu meinem Informationsbedürfnis, weitere Hintergrundinformationen und sogar den offiziellen Untersuchungsbericht. Offenbar war die Explosion auf einen simplen Rechenfehler zurückzuführen. Eine Zahl wurde zu gross und verursachte den Absturz eines Computersystems.

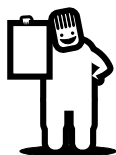


Ich habe die Informationen zu Java gefunden und habe mir einen Eindruck von der Insel verschaffen können. Jetzt brauche ich nur noch ein Flugticket. Ich möchte mit der Fluggesellschaft FreshAir fliegen und hätte am liebsten ein Ticket direkt von der Fluggesellschaft, denn die sind vermutlich die billigsten. Meine Anfrage bei OMNISEARCH lautet *fresh-air ticket zürich zurich java indonesia*. Ich finde zwar die gewünschte Fluggesellschaft, aber leider enthält meine Rangliste viele Einträge von Reisebüros, die ebenfalls die gesuchten Tickets anbieten. Wie verbessere ich eine solche Rangliste?

Irrelevante Dokumente in der Rangliste ignorieren!

Ziel einer Recherche mit Hilfe eines Suchdienstes im Internet ist es nicht, eine perfekte Rangliste zu erhalten. In der Regel tauchen einige irrelevante Dokumente in der Rangliste auf, und meistens kann man diese «schwarzen Schafe» problemlos und innert kürzester Zeit entlarven. Unangenehmer ist die umgekehrte Situation: Es gibt relevante Dokumente in der Kollektion, aber sie tauchen für eine bestimmte Anfrage nicht in der Rangliste auf. Der Benutzer kann gar nicht erkennen, dass er etwas verpasst hat!

Es gilt hier, einen guten Kompromiss im Konflikt Ausbeute gegen Präzision zu finden. Auf der einen Seite sollen möglichst alle relevanten Dokumente gefunden werden. Andererseits soll die Rangliste nicht übermässig mit irrelevanten Einträgen belastet werden.



In meinem Fall ist das Ignorieren der irrelevanten Dokumente denkbar einfach. Ich konzentriere mich ausschliesslich auf Dokumente, die vom Web-Server von FreshAir stammen. Die Dokumente erkenne ich anhand der Adresse, die mit **www.freshair.com** beginnt.

Oder noch besser: Ich befolge den vorhergehenden Suchtipp und füge den Suchbegriff *server:www.freshair.com* zur Anfrage hinzu. Damit mache ich Gebrauch von den Metadaten. In der Rangliste sollten nun die Dokumente vom Server der Fluggesellschaft weit oben erscheinen.



Meine Tochter soll für den Geschichtsunterricht einen Vortrag vorbereiten. Thema: die «Boston Tea Party», als ausgewanderte Amerikaner eine Schiffsladung Tee aus England in den Bostoner Hafen kippten, um gegen die Besteuerung durch das englische Mutterland zu protestieren. Ich helfe meiner Tochter beim Recherchieren. Unsere Anfrage bei OMNI-SEARCH lautet *tea party*, und wir erhalten alles – nur nicht das Gesuchte.

Dem System mitteilen, was man weiss!

Man stelle sich vor: Ein USA-Reisender betritt ein New Yorker Reisebüro mit der Frage: «Car?» Welche Antwort wird er wohl erhalten? Möchte er wissen, wo die Langstreckenbusse von New York City nach Seattle fahren? Versucht er ein Auto zu mieten? Oder möchte er sogar ein Auto kaufen? Oder sucht er nach Mitfahrgelegenheiten für die Reise nach Washington? Vielleicht interessiert er sich auch für ein Automobilmuseum? Man weiss es nicht.

Suchsysteme können keine Gedanken lesen! Je mehr Information die Benutzerin einem Suchdienst zukommen lässt, desto zutreffendere Antworten kann das System liefern.



Ungeschickt von uns, denn eigentlich haben wir oben bereits all unser Hintergrundwissen zum Thema geschildert. Jetzt müssen wir es aber dem Suchdienst auch mitteilen. Zum Beispiel mit der Anfrage *“boston tea party“ tea tax america britain “united kingdom“ “united states“*.

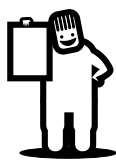


Eine vor allem im Altertum beliebte Strategie, um beispielsweise ein Volk zu unterwerfen, ist unter dem Namen «divide et impera» bekannt. Zu Deutsch: «teile und herrsche». Zuerst teilt man das Volk in einzelne, rivalisierende Stämme. Anschliessend wird jeder Stamm einzeln unterworfen, denn die Widerstandskraft der einzelnen Stämme ist niemals so gross wie diejenige des vereinten Volkes.

So viel zur Bedeutung von «divide et impera». Ich möchte nun aber wissen, wer diesen Namen geprägt hat. OMNISEARCH liefert auf die Anfrage “*divide et impera*“ an erster Stelle ein Dokument mit dem Titel «Schnelle Sortieralgorithmen unter der Lupe». Ich frage mich kurz, was dieses Dokument an erster Stelle zu suchen hat, und überspringe es dann. Doch weiter hinten finde ich auch nichts Relevantes. Enttäuscht gebe ich die Suche auf.

Besträngiertes Dokument vollständig untersuchen!

Auch wenn Titel und Zusammenfassung etwas völlig Unpassendes versprechen, das besträngierte Dokument liegt nicht ohne Grund in der «Pole Position». Bei guten Suchsystemen (und guten Anfragen) enthält das besträngierte Dokument häufig relevante Information, obwohl das nicht auf den ersten Blick ersichtlich ist.



Das möchte ich sehen! Ich schaue das besträngierte Dokument an. Wie erwartet werde ich mit unverständlichen Informatikthemen konfrontiert. Sogar Programmbeispiele kommen vor, die mir rein gar nichts sagen.

Doch tatsächlich – im Zusammenhang mit einem der Algorithmen fällt das Stichwort «divide et impera». Der Algorithmus heisst «Quicksort» und wendet ebenfalls das Prinzip an, indem er das Ausgangsproblem in immer kleinere und besser überschaubare Einzelprobleme zerlegt. Als Nebenbemerkung wird an dieser Stelle ausserdem darauf hingewiesen, dass der Ausdruck «divide et impera» vom römischen Feldherrn Julius Cäsar stammt.

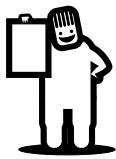


Ich interessiere mich für die internationale Zusammenarbeit im Bereich Umweltschutz. Dazu suche ich nach einer möglichst vollständigen Übersicht über die bestehenden internationalen Abkommen auf diesem Gebiet. Die Anfrage *internationale Zusammenarbeit Umweltschutz Abkommen Verträge* bei NEWSSEEKER liefert bereits einige relevante Dokumente. Wie aber kann ich noch weitere Dokumente zum Thema finden?

Interaktive Techniken anwenden!

Die interaktiven Suchtechniken sind wichtig, wenn man einige relevante Dokumente gefunden hat und zusätzliche Dokumente finden möchte. Interaktive Techniken helfen häufig auch weiter, wenn man beim Recherchieren nicht mehr weiterkommt, weil sich zum Beispiel keine nützlichen Suchbegriffe mehr finden lassen.

Man soll sich unbedingt die schon gefundenen relevanten Dokumente anschauen und wichtige Begriffe identifizieren. Anschliessend lässt sich mit der Suchanfrage spielen – man kann neue Suchbegriffe hinzufügen oder vielleicht andere entfernen. Das lässt sich manuell erledigen. Man kann aber auch das Suchsystem arbeiten lassen. Vorausgesetzt, es existieren Funktionen für die Relevanzrückkoppelung oder für die Suche nach ähnlichen Dokumenten.



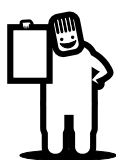
Zum Glück bietet NEWSSEEKER die praktische Relevanzrückkoppelung an. Die kann ich bei meinem Problem anwenden. Dazu markiere ich die gefundenen relevanten Dokumente und lasse anschliessend vom System die Relevanzrückkoppelung durchführen. Ich erhalte dann weitere relevante Dokumente. Meine ursprüngliche Anfrage wird um weitere nützliche Suchbegriffe erweitert. Neu wird nämlich auch nach Begriffen wie *Nachhaltigkeit* und *Klimaveränderung* oder nach Organisationen wie *UNEP* (United Nations Environment Programme) oder *IISD* (International Institute for Sustainable Development) gesucht.



Ist der Papst wohl auch im Web vertreten? Um das zu überprüfen, mache ich mich auf die Suche nach der offiziellen Web-Site des Vatikans. Dazu benutze ich OMNISEARCH mit der Anfrage *homepage "home page" "web site" site vatican pope*. Das Ergebnis ist enttäuschend. Es gibt so viele Webseiten, die den Papst oder den Vatikan erwähnen, dass ich keine Chance habe, ans Ziel zu kommen.

In Teilkollektionen suchen!

Viele Suchsysteme erlauben es, in Teilkollektionen zu suchen. Entweder durch eine vorgegebene Auswahl von Teilkollektionen (zum Beispiel die Kategorien in einem Katalogdienst) oder durch die Anwendung von Boole'schen Funktionen auf Metadaten. Die Suche in Teilkollektionen ist vor allem dann lohnenswert, wenn man mit einer Fülle von irrelevanten Dokumenten überschwemmt wird und sich diese Dokumente eindeutig charakterisieren lassen. Zum Beispiel aufgrund der Herkunft, der Sprache oder eines Datums. Aber es ist Vorsicht geboten. Mit Boole'schen Einschränkungen kann man relevante Dokumente ausschliessen, ohne es zu merken.



Ich könnte es in meinem Fall mit einer Einschränkung der Dokumentensammlung auf die Webseiten innerhalb der Domäne des Vatikanstaates versuchen. Aber wie lautet die Länderkennung des Vatikans? Um das herauszufinden, bemühe ich ebenfalls OMNISEARCH. Die Anfrage *"country codes"* führt rasch zum Ziel – die gesuchte Kennung lautet «va».

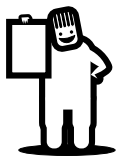
Jetzt kann ich meine Suche entsprechend einschränken. Das mache ich mit der Anfrage *+domain:va vatican homepage "home page" "web site" site pope*. Siehe da! Es gibt tatsächlich einen offiziellen Web-Server, der sich dem Heiligen Stuhl widmet.



Ich möchte mich über die wichtigsten neuen medizinischen Methoden zur Krebsbehandlung informieren. Für die Recherche verwende ich NEWSSEEKER, denn ich gehe davon aus, dass die wichtigeren Methoden auch durch entsprechende Presseartikel einer breiten Öffentlichkeit vorgestellt wurden. Meine Anfrage lautet: *Medizin medizinische Technologie Behandlung Therapie Krebs Leukämie Brustkrebs*. Ich finde einige relevante Artikel, bin aber noch nicht ganz zufrieden. Also versuche ich es mit einer Relevanzrückkoppelung, allerdings ohne Erfolg. Ich denke, es müsste noch mehr Informationen zum Thema geben.

Anfragen in verschiedenen Sprachen formulieren!

Englisch hat im Internet traditionell einen grossen Vorsprung gegenüber anderen Sprachen, weil das Netz im amerikanischen Raum entstanden ist. Doch die übrigen Sprachen holen auf. Wer also mehrere Sprachen beherrscht, sollte seine Sprachkenntnisse auch bei der Informationssuche ausnützen. Andernfalls können Übersetzungshilfen benutzt werden. Hinzu kommt, dass manchmal Informationen über ein spezifisches Thema vor allem in einem bestimmten Sprachraum auftauchen. Offensichtliches Beispiel: Informationen über die deutsche Rechtschreibreform sind sicherlich vorwiegend in deutscher Sprache verfasst.



Für meine Frage ist Englisch sicher einen Versuch wert. Mit der übersetzten Anfrage *medicine medical technologies therapy treatment breast cancer leukemia leukaemia* lassen sich tatsächlich noch ein paar zusätzliche relevante Dokumente aufspüren.



Eine meiner Geschäftsreisen führt mich in die Hauptstadt Japans, Tokio. Ich möchte mich vorher mit den nötigsten Informationen eindecken und suche darum nach einem Stadtplan von Tokio. Am liebsten wäre mir ein kombinierter Stadtplan inklusive Streckennetz der U-Bahn. Ich denke, so etwas sollte sich im Internet einfach finden lassen. Ich wende mich an OMNISEARCH mit der Anfrage *tokio city subway map*. In der Rangliste erhalte ich Dutzende von Stadtplänen von New York, Boston, Chicago, London und vieler anderer Städte. Zu Tokio finde ich nichts.

Rechtschreibung und alternative Schreibweisen beachten!

Falls die Rangliste leer bleibt oder die gemeldeten Resultate keinen Sinn machen, sollte man zuerst die Anfrage sorgfältig überprüfen. Häufig sind Tippfehler oder andere Probleme mit der Rechtschreibung in der Anfrage die Ursache für eigenartige Phänomene. Ausserdem gibt es häufig alternative Schreibweisen für einen Begriff, so zum Beispiel unterscheiden sich viele Ausdrücke im Amerikanischen und im Britischen. Genauso existieren für zahlreiche Begriffe alte und neue deutsche Schreibungen.



Die Rechtschreibung ist auch in meinem Fall das Problem. Unterdessen habe ich nämlich herausgefunden, dass das Wort «Tokio» in Deutsch und Englisch unterschiedlich geschrieben wird. Mit der korrigierten Anfrage – *tokyo city subway map* – finde ich den gesuchten Stadtplan ohne weitere Schwierigkeiten.

Stichwortverzeichnis

A

Akzente, 58, 136
AND, *siehe* Boole'sche Operatoren
AND NOT, *siehe* Boole'sche Operatoren
Anfrage-Dokumentenvergleich, 24, 63
Anfrageaufwand, 135–136, 140
Anfragen, 32, 153
 automatische Modifikation, 101
 manuelle Modifikation, 101
Antwortzeit, 136
Ausbeute, 138–139, 148

B

Benutzerprofile, *siehe* Profile
Benutzerschnittstelle, 23, 137
Boole'sche Operatoren, 81, 89
Boole'sche Suchmethoden, 81–83, 88, 152
 im Dokumentinhalt, 81–83, 91–93
Bring-Prinzip, *siehe* Push-Prinzip
Browser, *siehe* Web-Browser
Buchstabenumwandlung, 52, 136

C

CGI, 18
Common Gateway Interface, *siehe* CGI
Crawler, *siehe* Web-Roboter

D

Daten, 15
Deskriptoren, 84
Dokumente, 20
Dokumentenkollektionen, 19–21, 112, 126, 135, 145
 horizontale, 20
 vertikale, 21

E

E-Mail, 16
Effektivität, *siehe* Sucheffektivität
Effizienz, 136–137
 Entdeckungs-, *siehe* Entdeckungseffizienz
 Erschliessungs-, *siehe* Erschliessungseffizienz
 Such-, *siehe* Sucheffizienz
Entdeckungseffizienz, 136
Erschliessung, 19, 136
Erschliessungseffizienz, 136–137
Extensible Markup Language, *siehe* XML

F

Feedback, 101
Filterung, 127–128
Firewall, *siehe* Firewall-Rechner
Firewall-Rechner, 18
Flexionen, 53, 136

G

Gross- und Kleinschreibung, 53, 58–59

H

Hintergrundwissen, 33
Hol-Prinzip, *siehe* Pull-Prinzip
Hyperlinks, 55, 64

I

Index, 24, 66–69
Indexierung, 23, 51, 56, 136
Information, 15
Information Retrieval, 31
Information-Retrieval-Systeme,
 siehe Suchsysteme
Informationsbedürfnis, 32, 124
Informationsdienste, 19–20
 Zugriff auf, 19
Informationssuche, 31
 Grundproblem, 32
 iterativ, 99–101
Informationssysteme, 19, 21–22
 Komponenten, 23–24
Internet, 16
 Zugriff auf Daten, 17–18
Interpunktionszeichen, 52
Intranet, 18

K

Katalogdienste, 112, 146
 automatische Erstellung,
 113–116
 manuelle Erstellung, 112–113
 Nachteile, 118
 Vorteile, 117
Kataloge, *siehe* Katalogdienste
Katalogsysteme, 21, 111
 Aufbau, 111–112
Kategorien, 22, 111
Kategorienhierarchie, 112
Kategorienprofile, *siehe* Profile
Keyword Spamming,
 siehe Web Spamming
Klassifizierung, 113
Kollektionen, *siehe* Dokumenten-
 kollektionen

Komposita, 53

Kosten, 139–140
 Entwicklungsaufwand, 139
 Unterhaltskosten, 139

L

Links, *siehe* Hyperlinks
Linktiefe, 73

M

Mailing-Listen, 130
Manuelle Anfrageerweiterung, 104
Meta-Tags, 55
Metadaten, 83–85, 114, 152
 nicht normalisierte, 84
 normalisierte, 84, 87
Metadokumente, 83
Metasuchdienste, 74–76
 Arbeitsweise, 74–75
 Nachteile, 75–76
 Vorteile, 75
Modifikationsdatum, 55

N

News, 16, 90, 130
Newsgroups, 130
NewsSeeker, 25
Normalform, 53
NOT, *siehe* Boole'sche Operatoren

O

OmniSearch, 25
OR, *siehe* Boole'sche Operatoren

P

Page Spamming, *siehe* Web Spam-
 ming
Phrasensuche, 93
Platzhalter, *siehe* Wildcards
Präzision, 138–139, 148
Profilanpassung, 129

- Profile, 22, 114–116, 127
- Profilerzeugung, 116, 128
- Pull-Dienste, 130
- Pull-Prinzip, 125
- Pull-Systeme, 22
- Push-Dienste, 125–130, 146
 - Dokumentenbezug, 126
 - Dokumentenfilterung, 127–128
 - Dokumentenzustellung, 128
- Push-Prinzip, 125
- Push-Systeme, 22, 126, 139

Q

- Querverweise, 112

R

- Rangierungsprinzipien, 38–45
- Rangliste, 37, 44, 148
- Relevance Ranking, 36, 85
 - Ablauf, 36–37
- Relevanz, 34
 - Berechnung, 38
 - geschätzte, 35
 - inhaltsunabhängige Bestimmung, 45–46
 - objektive, 35
 - subjektive, 34, 45, 46
- Relevanzrückkoppelung, 101–103, 116, 129, 136, 151
 - Abdriften, 103
 - manuelle, 104–105
- Relevanzwert, 33, 115, 128
- Robot Exclusion, 73

S

- Satzzeichen, *siehe* Interpunktionszeichen
- Schlagwörter, 84
- Server, 16
- Spamdexing, *siehe* Web Spamming
- Spider, *siehe* Web-Roboter
- Sprachidentifikation, 52
- Stoppwörter, 52, 57–58

- Stoppwortelimination, 52, 57
- Stoppwortliste, 57
- Suchanfragen, *siehe* Anfragen
- Suchbegriffe, 38, 51, 69, 136, 147
- Suchdienste, 146
 - Abdeckungsgrad, 72
- Suche nach ähnlichen Dokumenten, 101–102, 151
- Sucheffektivität, 69, 138–140
- Sucheffizienz, 69, 136
- Suchmaschinen, *siehe* Suchsysteme
- Suchsysteme, 21, 24–25
 - Funktionsweise, 69–71

T

- Teildokumentenkollektionen, *siehe* Teilkollektionen
- Teilkollektionen, 85, 152
 - Definition von, 85, 87–91
- Textdokumente, 20

U

- Umlaute, 52, 58, 136
- Uniform Resource Locator, *siehe* URL
- URL, 18, 55

W

- Web Spamming, 46–47
- Web-Browser, 17
- Web-Roboter, 23, 64–66, 114, 126, 136
 - Arbeitsweise, 64
 - Regeln, 66
- Web-Server, 17
- Web-Site, 18
- Webseiten
 - dynamische, 18, 72
 - isolierte, 65, 72
 - lokale, 17
 - statische, 18
- Wildcards, 57
- Wissen, 15

World Wide Web Consortium, 16
Wortextraktion, 52
Wortnormalisierung, 53, 57, 136
 Grundformermittlung, 53
 Wortstammreduktion, 53
Wortzerlegung, 53, 57, 136
WWW, 16

X

XML, 84