Wirtschafts-informatik

Grundstudium

Suchmaschinen und wie sie genutzt werden

Prof. Dr. Wolfgang G. Stock / Dr. Dirk Lewandowski, Düsseldorf

Suchmaschinen sind aus dem Internet nicht mehr wegzudenken, täglich werden Millionen Anfragen gestellt. Welche Themen werden von den Internet-Nutzern nachgefragt, und wie werden die unterschiedlichen Anfragen von den Suchmaschinen beantwortet?

1. Suchmaschinen im World Wide Web und ihre Nutzer

Bereiche des WWW

Im World Wide Web (WWW) liegen Milliarden von Dokumenten in digitaler Form bereit. Ihr Spektrum reicht von Blogs, in denen Privatleute ihre Ansicht zu allen denkbaren Themen mitteilen, über Produktbeschreibungen kommerzieller Anbieter und Selbstdarstellungen von Unternehmen bis zu wissenschaftlichen Veröffentlichungen. Diese digitalen Dokumente lassen sich in zwei Kategorien einteilen (vgl. Stock/Stock 2004, S. 3 ff.):

- **Oberflächen-Web:** Digitale Dokumente im Web (d.h. fest verlinkt, wobei keine Kosten für den Content entstehen),
- **Deep Web:** Digitale Dokumente in Informationssammlungen (i.d.R. Datenbanken), deren Einstiegsseiten via WWW erreichbar sind.

Suchwerkzeuge

Die meisten Informationen im Oberflächen-Web können von Suchwerkzeugen aufgefunden und ausgewertet werden. Informationen mit speziellen Datenformaten werden nicht immer erfasst. Nicht direkt erfassbar sind Dokumente, die nicht verlinkt sind. Probleme können bei der Verwendung von Frames auftauchen. **Spam-Seiten**, also unerwünschte Werbung oder Dokumente mit irreführenden Erläuterungen oder falschen Angaben, werden von den Suchwerkzeugen nach Möglichkeit nicht in deren Datenbasis aufgenommen.

Es lassen sich drei Arten von Suchwerkzeugen unterscheiden:

- **algorithmisch arbeitende Suchmaschinen** (mit eigener Datenbasis und Retrievalsystem) wie Google, Yahoo, MSN und Ask.com,
- **intellektuell erstellte Webkataloge** wie das Open Directory Project (eingesetzt vom Google Verzeichnis) oder der Yahoo-Katalog,
- **Meta-Suchmaschinen**, die zwar über ein Retrieval-System, aber über keine eigene Datenbasis verfügen und den Content von fremden Suchwerkzeugen beziehen (z.B. Ixquick, Kartoo).

Suchmaschinen verwenden als Bezugseinheit eine Web-Seite, während Webkataloge in der Regel nur die Einstiegseite einer Website berücksichtigen. Dementsprechend ist die Datenbasis bei Suchmaschinen weitaus größer als bei Webkatalogen. Eine Sonderform für Sucheinstiege in das WWW sind Portale. Sie enthalten neben einem Retrieval-System weitere Funktionen zum Browsen, Personalisierungsoptionen sowie Kommunikations- und Kollaborationsmöglichkeiten.

Invisible Web

Die zweite Hauptklasse digitaler Online-Informationen liegt nicht im Web, sondern ist lediglich über das Web erreichbar. Die Einstiegseiten solcher Systeme lassen sich auch mit den Suchmaschinen und Webkatalogen finden, nicht aber die gespeicherten Datensätze. Diese Datenbanken bilden das „Invisible Web" oder „Deep Web". Einige Datenbanken werden professionell und kommerziell betrieben und sind kostenpflichtig. Deep-Web-Angebote in den Wirtschaftswissenschaften sind beispielsweise die WISO-Datenbanken, EconLit und ABI-INFORM.

Frage 1: Welche Arten von Suchwerkzeugen gibt es im WWW?

Für die Deep-Web-Datenbanken, die Suchwerkzeuge im Oberflächen-Web und die entsprechenden Werkzeuge in internen Netzen ist das **Information Retrieval** (IR) zuständig (vgl. Stock 2006), beschränkt man sich auf das WWW, ist es das **Web Information Retrieval** (vgl. Lewandowski 2005).

Drei Nutzergruppen

Je nachdem, weshalb Informationen nachgefragt werden, und entsprechend der Recherchestrategien und Anfragen, lassen sich drei Arten von Nutzern unterscheiden:

1. **Information Professional:** Er ist Experte im Suchen und Finden von Dokumenten (beruflicher Hintergrund: z.B. Informationswissenschaftler oder Informationswirt), kennt den Inhalt der Datenbanken, entwickelt Recherchestrategien und arbeitet mit professionell ausgearbeiteten Anfragen. Er ist vor allem am Deep Web interessiert. Ergänzend nutzt er Quellen im Oberflächen-Web. Er arbeitet überwiegend im Auftrag Dritter (z.B. Neuheitsrecherche vor Beginn eines Forschungsprojektes oder Erstellung eines Firmendossiers vor einer geplanten Unternehmensübernahme).
2. **Professioneller Endnutzer:** Er ist Fachexperte (z.B. Wirtschaftswissenschaftler) und recherchiert, da er Bedarf an Fachinformationen hat. Die einschlägigen fachspezifischen Datenbanken (z.B. WISO und EconLit) und Suchmaschinen (in Deutschland meist Google) sind ihm bekannt, er entwickelt jedoch kaum Recherchestrategien und professionell ausgearbeitete Anfragen. Er arbeitet einfachen Informationsbedarf, der im Berufsalltag anfällt, ad hoc ab.
3. **Laiennutzer:** Er ist der typische „Googler" und recherchiert aus privaten und gelegentlich auch aus beruflichen Gründen ausschließlich im Oberflächen-Web. Systematische Recherchestrategien sind ihm meist unbekannt. Seine Anfragen haben oft beschränkten Umfang.

Funktionsumfang der Suchmaschinen

Suchwerkzeuge im WWW orientieren sich an den Laiennutzern. Sie bieten den professionellen Endnutzern einige, den Information Professionals hingegen kaum zufriedenstellende Recherchemöglichkeiten. Die derzeitigen Suchwerkzeuge arbeiten im Sinne **Boolescher Systeme**, d.h. sie verwenden die Operatoren UND, ODER bzw. NICHT. Gibt ein User keinen Funktor an, wird heutzutage als Boolesches UND interpretiert, während die ursprünglichen Suchwerkzeuge bis ca. zum Jahr 2000 mit ODER gearbeitet haben. Die Suchmaschinen bieten allerdings nicht die volle Boolesche Funktionalität an. So lässt Google keine Klammerung zu und kennt auch keinerlei Abstandsoperatoren außer der direkten Nachbarschaft (Phrase). Trunkierung durch Jokerzeichen ist ebenfalls nicht möglich. Möchte man beispielsweise Literatur über „Angebote" recherchieren, so sind alle grammatischen Varianten mit ODER zu verknüpfen: Angebot OR Angebots OR Angebote OR Angeboten (zum Vergleich mit Trunkierung: Angebot*). Die Klammerung erleichtert verschachtelte Anfragen wie: Gesucht sind Webseiten zu Arbeit und Freizeit in Holland und Belgien, aber nicht solche, in denen Holland und Belgien gemeinsam vorkommen. Bei Google lautet die Formulierung: Arbeit Freizeit Holland OR Belgien -Holland OR -Belgien, die jedoch kaum ein Laiennutzer kennt.

Doch wie arbeiten die Nutzer? Wie gehen sie mit den Suchwerkzeugen um? Damit befasst sich die **Nutzerforschung**, die ihr Verhalten analysiert und zwar:

- durch **Beobachtung** einzelner Nutzer in Laborsituationen,
- durch deren **Befragung** und
- durch die Analyse der **Logfiles** von Suchwerkzeugen.

Frage 2: Welche Nutzergruppen können unterschieden werden?

2. Suchstrategien der Internet-Nutzer

Während die Nutzer von Fachdatenbanken geschult und in der Regel mit den speziellen Abfragesprachen dieser Systeme vertraut sind, sind die Anfragen an Web-Suchmaschinen sehr einfach.

Verwendung von Verknüpfungen

Boolesche Operatoren werden nur bei etwa jeder zehnten Anfrage verwandt (vgl. Spink/Jansen 2004, S. 184), während etwa 20 Prozent der Nutzer angeben, diese öfter zu verwenden (vgl. Machill et al. 2003, S. 167). Eine Untersuchung aus dem Jahr 2000 (vgl. Spink et al. 2000) fand heraus, dass etwa die Hälfte der Booleschen Anfragen Fehler enthalten. Bei den von den Nutzern an Stelle der Booleschen Operatoren bevorzugten Plus- und Minuszeichen (die manchmal dieselben Funktionen ausdrücken) lag die Fehlerquote sogar bei zwei Dritteln. Der Anteil der Anfragen mit Booleschen Operatoren erscheint sehr gering. Allerdings ist die Eingabe dieser Verknüpfungen bei Suchmaschinen

— im Gegensatz zu anderen Recherchesystemen — nicht zwingend notwendig. So sind einfache Anfragen ohne die Eingabe von Operatoren möglich.

Nutzung der Profisuche

Während die Booleschen Operatoren laut der Befragung von Machill et al. (2003) nur etwa der Hälfte der Nutzer bekannt sind, erreichen die **erweiterten Suchformulare** („Profisuche") mit 59 Prozent eine etwas höhere Bekanntheit. Allerdings zeigt sich, dass sie noch seltener genutzt werden als die Operatoren. Nur 14 Prozent der Nutzer geben an, die erweiterte Suche öfter zu verwenden (vgl. Machill et al. 2003, S. 168). In der Laboruntersuchung lag deren Nutzung noch einmal deutlich unter diesem Wert.

Kurze Suchanfragen

Die Nutzung der Operatoren hat sich im Laufe der letzten Jahre nicht verändert (vgl. Spink/Jansen 2004, S. 79). Allerdings nimmt die **Länge der Suchanfragen** langsam zu und liegt mittlerweile bei durchschnittlich etwa 2,6 Termen je Anfrage. Dies sagt jedoch wenig über die Komplexität der Anfragen aus, da sich diese eher in der Verwendung von Operatoren oder anderen Möglichkeiten zur Rechercheeinschränkung zeigen würde. Stattdessen orientiert sich die Entwicklung bzw. Verbesserung der Suchmaschinen am geringen Kenntnisstand der Nutzer, was nicht zu größeren Veränderungen des Nutzerverhaltens führen dürfte.

Auswertung der Treffer

Wie Untersuchungen ergeben, orientieren sich etwa 80 Prozent der Nutzer an den ersten zehn Treffern der Ergebnisliste, also meist nur an der ersten Seite der Trefferliste (vgl. Spink/Jansen 2004). Zudem hat die **Zahl der beachteten Ergebnisseiten** abgenommen, was auch darauf zurückzuführen sein könnte, dass die Anfragen von den Suchmaschinen mittlerweile besser beantwortet werden. In erster Linie werden die ersten Suchergebnisse angeklickt, die ohne Scrollen sichtbar sind.

Im Durchschnitt werden nur etwa fünf Dokumente gesichtet (vgl. Spink/Jansen 2004, S. 101), wobei jedes Dokument nur kurz daraufhin überprüft wird, ob es die gewünschte Information enthält. Wird ein Dokument gefunden, das die Informationswünsche befriedigen kann, wird die Recherche in der Regel beendet. Die Suchdauer einschließlich der Sichtung der Dokumente dauert meist nur etwa 15 Minuten (vgl. Spink/Jansen 2004, S. 101).

Ähnliche Dokumente finden

Einige Suchwerkzeuge bieten die Option, Treffer als „Musterdokumente" auszuwählen und weitere ähnliche Dokumente zu finden. Diese „More like this!"-Suche wird auch **Relevance Feedback** genannt, da der Nutzer Relevanzinformationen zu einer Webseite besitzt und sie für eine iterative Suchstrategie nutzt. Eine Erhebung bei Excite ergab, dass bei weniger als fünf Prozent aller Anfragen die Option „More like this!" verwandt wird (vgl. Spink et al. 2000, S. 326).

> **Frage 3:** Wie lässt sich das Rechercheverhalten der Suchmaschinen-Nutzer charakterisieren?

3. Anfragetypen

Zwei Fragetypen

Nutzer haben einen unterschiedlichen Informationsbedarf, wie sich anhand dieser Beispiele zeigen lässt:

- Fragetyp A
  - Wie heißt die Hauptstadt von Nordrhein-Westfalen?
  - Wie lautet die Web-Adresse der Düsseldorfer Universität?
  - Wie viel kostet eine Walther PPK?
- Fragetyp B
  - Wie kann der Homunculus in Goethes Faust interpretiert werden?
  - Welchen Zusammenhang gibt es zwischen Dienstleistungsmarketing und Qualitätsmanagement?
  - Wie wurde das Unternehmen Boll & Kirch in Kerpen in den letzten Jahren von Analysten bewertet?

Konkreter und problemorientierter Informationsbedarf

Fragetyp A zielt auf Fakteninformation ab, es handelt sich um einen konkreten Informationsbedarf. Frants, Shapiro und Voiskunskii (1997) bezeichnen dies als **Concrete Information Need (CIN)**. Hingegen lässt sich der in Fragetyp B zum Ausdruck kommende Informationsbedarf nicht durch ein Faktum, sondern nur durch eine mehr oder minder große Textsammlung befriedigen. Frants et al. sprechen hier von **Problem Oriented Information Need** (POIN). CIN und POIN lassen sich anhand einiger Charakteristika vergleichen (s. Abb.).

CIN	POIN
1. Thematische Grenzen sind klar abgesteckt.	1. Thematische Grenzen sind *nicht* exakt bestimmbar.
2. Die Suchfrageformulierung ist durch exakte Terme ausdrückbar.	2. Die Suchfrageformulierung lässt *mehrere* terminologische Varianten zu.
3. Eine Fakteninformation reicht i.d.R. aus, um den Bedarf zu decken.	3. In der Regel müssen *diverse* Dokumente beschafft werden. Ob der Informationsbedarf damit abschließend gedeckt ist, bleibt offen.
4. Mit der Übermittlung der Fakteninformation ist das Informationsproblem erledigt.	4. Mit der Übermittlung der Literaturinformation wird ggf. das Informationsproblem modifiziert oder ein neuer Bedarf entdeckt.

Abb.: Concrete Information Need und Problem Oriented Information Need (Frants/Shapiro/Voiskunskii 1997, S. 38)

Ein Treffer vs. Treffermenge

Ob eine Information ein CIN befriedigt, d.h. relevant ist, lässt sich exakt bestimmen: Die Frage wird beantwortet oder nicht. Anders beim POIN. Die gesammelten Texte können mehr oder weniger Aspekte der Frage beantworten, die Relevanz kann also vage sein.

Frage 4: Wie unterscheiden sich konkreter und problemorientierter Informationsbedarf?

Anfragetypen nach Broder

Broder (2002, S. 5 f.) unterscheidet bei Web-Suchmaschinen drei Typen der **Suche**: navigations-, informations- und transaktionsorientierte Anfragen.

Mit **navigationsorientierten Anfragen** soll eine Seite (wieder)gefunden werden, die dem Benutzer bereits bekannt ist oder von der er annimmt, dass sie existiert. Beispiele sind die Suche nach Homepages von Unternehmen („DaimlerChrysler") oder nach Personen („Angela Merkel"). Solche Anfragen haben in der Regel ein richtiges Ergebnis. Das Informationsbedürfnis ist befriedigt, sobald die gewünschte Seite gefunden wird.

Bei **informationsorientierten Anfragen** ist das Informationsbedürfnis meist nicht durch ein einziges Dokument zu befriedigen (POIN). Der Nutzer möchte sich stattdessen über ein Thema informieren und liest deshalb mehrere Dokumente. Informationsorientierte Anfragen zielen auf jeden Fall auf statische Dokumente, nach dem Aufruf des Dokuments ist also keine weitere Interaktion auf der Website nötig, um an die gewünschten Informationen zu gelangen.

Mit **transaktionsorientierten Anfragen** wird eine Website gesucht, auf der anschließend eine Transaktion stattfindet, etwa der Kauf eines Produkts oder der Download einer Datei.

Informationsorientierte Anfragen entspringen in der Regel einem problemorientierten Informationsbedarf, die beiden anderen Suchtypen eher einem konkreten Informationsbedarf. Bei Navigationsfragen reicht beispielsweise ein Treffer zur Befriedigung des Informationsbedarfs.

Untersuchungen zu den Anfragetypen

Broder untersucht den Anteil der verschiedenen Anfragetypen anhand einer Nutzerbefragung und einer Logfile-Auswertung von 400 Suchanfragen. Beide Untersuchungen beziehen sich auf die Suchmaschine Alta Vista. Die navigationsorientierten Anfragen machen 20 bis 24,5 Prozent, die informationsorientierten 39 bis 48 Prozent und die transaktionsorientierten 22 bis 36 Prozent aus.

Um herauszufinden, ob diese Ergebnisse noch aktuell sind und auch für deutsche Suchmaschinennutzer gelten, wurde eine ähnlich angelegte Untersuchung durchgeführt (vgl. Lewandowski 2006). Dabei wurden 1.500 Anfragen der Suchmaschinen Fireball, Metager und Seekport verwandt.

Hoher Anteil an informationsorientierten und navigationsorientierten Anfragen

Die größte Gruppe bilden die informationsorientierten Anfragen, die über alle Suchmaschinen verteilt 45 Prozent aller Anfragen ausmachen, wobei die Werte je nach Suchmaschine zwischen 42 und 47 Prozent liegen. In einem ähnlichen Rahmen bewegen sich die Werte der navigationsorientierten Anfragen, die im Durchschnitt 40 Prozent erreichen. Größere Abweichungen finden sich bei den transaktionsorientierten Anfragen. Diese liegen zwischen 11 und 18 Prozent, mit einem Durchschnitt von 15 Prozent.

Alle Anfragetypen sind von Bedeutung

Vergleicht man die Daten mit der Untersuchung von Broder (2002), zeigen sich deutliche Abweichungen. Während die informationsorientierten Anfragen in dem von Broder ermittelten Rahmen liegen, haben die navigationsorientierten Anfragen einen höheren Anteil. Hingegen machen die transaktionsorientierten Anfragen einen deutlich geringeren Anteil

aus. Die Gründe dafür lassen sich nicht klar bestimmen. Sie können sich aus dem zeitlichen Abstand zwischen den Untersuchungen ergeben, aber auch aus dem spezifischen Verhalten deutscher Nutzer.

Frage 5: Welche Anfragetypen werden im Web unterschieden?

4. Nach welchen Themen wird gesucht?

Erhebung

Neben den Anfragetypen ist auch von Interesse, zu welchen Themen die Nutzer Informationen suchen. Dies lässt sich auf verschiedene Art ermitteln. So kann man die Nutzer direkt befragen, was jedoch zu ungenauen Aussagen führt, wenn es um die Verteilung der Themen geht. Zum einen lässt sich dadurch kaum das Anfragevolumen ermitteln, zum anderen werden die Nutzer häufig nicht wahrheitsgemäß antworten, etwa wenn es um die Themenfelder Sex und Pornografie geht. Ähnliche Probleme ergeben sich bei der Beobachtung der Nutzer in Laboruntersuchungen.

Eine weitere Möglichkeit ist die **Auswertung der Logfiles**. Sie bieten den Vorteil, dass sich mit ihrer Hilfe tatsächliche Suchanfragen ermitteln lassen, ohne dass der Nutzer durch die Untersuchung beeinflusst wird. Allerdings lassen sich auf diese Weise nur Anfragen ermitteln. Direkte Rückschlüsse auf das jeweilige Informationsbedürfnis sind nicht möglich. Bei einer solchen qualitativen Untersuchung besteht die Gefahr, dass die Zuordnung der **Anfragen zu einzelnen Themen** von persönlichen Annahmen beeinflusst wird. Dieses Problem lässt sich abmildern, indem jede Zuordnung von wenigstens zwei Personen vorgenommen wird und diejenigen Fälle, die zu unterschiedlichen Zuordnungen führen, in einer anschließenden Diskussion geklärt werden.

Klassifikation der Themen

Spink und Jansen (2001) nehmen eine grobe Klassifizierung der Themen vor:

- Personen, Orte und Dinge,
- Computer und Internet,
- Handel, Reise, Arbeit und Wirtschaft
- Unterhaltung und Freizeit,
- Gesundheit und (Natur-)Wissenschaft,
- Sex und Pornografie,
- Regierung und Verwaltung,
- Bildung und Geisteswissenschaften,
- Gesellschaft, Kultur, Ethnizität und Religion,
- Kunst,
- nicht bestimmbar und sonstige.

Diese Klassifizierung wurde über mehrere Jahre bei Daten aus Logfiles verschiedener Suchmaschinen angewandt (Spink und Jansen 2004). In der Regel wurden etwa 2.500 Suchanfragen ausgewertet.

Untersuchungsergebnisse

Während des Untersuchungszeitraums (1997 bis 2002) wurde eine **Verschiebung von Entertainment-Inhalten hin zu wirtschaftsorientierten Anfragen** festgestellt. Die am häufigsten nachgefragten Themen der neueren Untersuchungen sind Personen, Orte und Dinge (2002: 41 bis 49 Prozent), Computer und Internet (2002: 12 bis 16 Prozent) sowie Wirtschaft und Arbeit (2002: ca. 12 Prozent).

Im Jahr 2005 wurden von den deutschen Suchmaschinen-Nutzern vor allem Themen aus der Wirtschafts- und Arbeitswelt nachgefragt (29 Prozent), gefolgt von Personen und Orten (12,8 Prozent; vgl. Lewandowski 2006). Die Themen Computer und Internet, Unterhaltung und Freizeit sowie Gesundheit und Wissenschaft liegen jeweils bei 7 bis 8 Prozent, die restlichen Themen zwischen einem und etwas mehr als 4 Prozent.

Differenzen bei den Untersuchungsergebnissen

Die Verteilung unterscheidet sich deutlich von Vergleichsuntersuchungen, in denen mit großem Abstand der Themenbereich Personen, Orte und Dinge führt (41 bzw. 49 Prozent), gefolgt von Computer und Internet sowie Wirtschaft und Arbeit. Die Gründe für die Unterschiede lassen sich nicht verlässlich bestimmen, sie können in unterschiedlichem Nutzerverhalten oder im zeitlichen Abstand liegen. Jedenfalls wird der hohe Anteil der wirtschaftsorientierten Anfragen der deutschen Nutzer deutlich.

Literaturempfehlungen:

Broder, A.: A Taxonomy of Web Search. In: SIGIR Forum, Vol. 36 (2002), S. 3 - 10.

Frants, V.L./Shapiro, L./Voiskunskii, V.G.: Automated Information Retrieval. Theory and Methods. San Diego 1997.

Lewandowski, D.: Web Information Retrieval: Technologien zur Informationssuche im Internet. Frankfurt a.M. 2005.
Lewandowski, D.: Themen und Typen der Suchanfragen an deutsche Web-Suchmaschinen. In: Lehner, F./Nösekabel, H./Kleinschmidt, P. (Hrsg.): Multikonferenz Wirtschaftsinformatik 2006 (MKWI ,06), Bd. 2. Berlin 2006, S. 33 - 43.
Machill, M./Neuberger, C./Schweiger, W./Wirth, W.: Wegweiser im Netz: Qualität und Nutzung von Suchmaschinen. In: Machill, M./Welp, C. (Hrsg.): Wegweiser im Netz. Gütersloh 2003.
Spink, A./Jansen, B.J.: Web Search: Public Searching of the Web. Dordrecht 2004.
Spink, A./Wolfram, D./Jansen, B.J./Saracevic, T.: Searching the Web: The Public and Their Queries. In: Journal of the American Society for Information Science and Technology, Vol. 53 (2001), S. 226 - 234.
Spink, A./Jansen, B.J./Ozmultu, H.C: Use of Query Reformulation and Relevance Feedback by Excite Users. In: Internet Research: Electronic Networking Applications and Policy, Vol. 10 (2000), S. 317 - 328.
Stock, M./Stock, W.G.: Recherchieren im Internet. Renningen 2004.
Stock, W.G.: Information Retrieval. München 2006.

Die Fragen werden im WISU-Repetitorium beantwortet.