

Kapitel II.3: Verteilte Informationssysteme und Suchhilfen¹

Dr. Frank Krüger

Fachbereich 23, Johannes Gutenberg-Universität Mainz

e-mail: krueger@acm.org

Gliederung

- 1 [Weitere verteilte Informationssysteme](#)
 - [Usenet News](#)
- 2 [Suchhilfen im Internet](#)
 - 2.1 [Gezielte Stichwortsuche](#)
 - 2.2 [Systematische Suche](#)

Anhang G: [Einige wichtige Adressen](#) im Internet (insbesondere WWW)

A. Verteilte Informationssysteme

1. Überblick

Lange vor bzw. parallel zum World-Wide Web haben sich im Internet eine Reihe weiterer verteilter Informationssysteme etabliert. Verteilt bedeutet dabei, daß - wie im WWW i.e.S., also den HTTP-konformen Servern - das Informationsangebot nicht zentral auf einem Rechner an einem Ort zur Verfügung gestellt wird, sondern ein eigenes Protokoll (siehe dazu [Abbildung 2 in Kapitel 1.1](#)) den offenen Datenaustausch ermöglicht.

Einige Charakteristika und Eigenschaften der wichtigsten dieser Systeme sind in der folgenden Tabelle gegenübergestellt:

Dienst	Usenet News	Gopher	Wide Area Information System (WAIS)
Grundprinzip	streng hierarchische systematische Gliederung der Gruppen, einzelne Beiträge als "Threads" verknüpft	indiv. hierarchische Gliederung in Menüstruktur ("Ordner" u. "Datei") entsprechend Dateisystem	Datenbank-Abfragestruktur (Stichwortsuche) mit "Meta"-Datenbank zur Vorselektion
Ursprung	in den frühesten Anfängen elektron. Kommunikation, entsprechend Bulletin Boards	Univ. of Minnesota und TU Clausthal (zentrale Home-Gopher), als	Standard Z53.69 (geschaffen durch kommerzielles Firmenkonsortium)
Hauptzweck	offene, meist sehr lebhaft Diskussions- ältere Beiträge werden in Dateien archiviert	Campusinformationssystem, ohne großen Verwaltungsaufwand ("Overhead")	gezielte Suche u.a. mit statistischen Methoden nach Auftreten von Suchbegriffen in sehr großen Textmengen
Suchhilfen	DEJA-News, integriert in ALTAVISTA	VERONICA bzw. JUGHEAD (ebenfalls verteilt, z.B. Uni Köln)	Zentrales Verzeichnis (www.wais.com , Bibliotheken: www.loc.gov)
Verbreitung	weltweit ca. 14.000 Gruppen	v.a. USA, aber immer öfter in das W3 integriert bzw. davon abgelöst	relativ wenig Server, aber jeweils mit großen Textsammlungen (z.B. Doku's des DOE, Kongreßtexte)

Diese Systeme sind in das WWW im weiteren Sinne (also unabhängig von den verwendeten Protokollen) insofern integriert, als die gängigen graphischen [Browser](#) (allen voran NETSCAPE,

¹: Eine elektronische Version dieser Unterlagen findet sich im WWW unter <http://www.fask.uni-mainz.de/cafl/kurse/komm/komm-23.html> (Kleinschreibung beachten!)

LYNX dagegen nur zum Teil) den Zugriff auf diese Systeme ermöglichen (Vergleiche die Erläuterungen zu den Dateitypen, [Kapitel I.4](#)).

Dies gilt besonders für den sog. **GopherSpace** (also die Gesamtheit des über Gopher-Server erreichbaren Informationsangebotes). Dieses Angebot unterscheidet sich weniger durch die Zugriffsstruktur - stark strukturierte HTML-Übersichtsseiten sehen den Gopher-Menüs sehr ähnlich -, sondern vielmehr durch die primitivere Darstellung der Information selbst, nämlich als reiner ASCII-Text ohne jede Formatierung. Dies bewirkte aber auch den großen Erfolg in der Anfangszeit (vor dem Erfolg des WWW), da Gopher-Server mit wesentlich weniger Aufwand einzurichten und zu betreiben waren. Andererseits bedingt gerade dies jetzt auch den starken Rückgang von Gopher-Angeboten, da immer mehr Dokumente in HTML-Seiten für das World-Wide Web umgewandelt werden.

Dagegen haben sich **WAIS-Server** nie auf breiter Basis durchgesetzt, obwohl sie einige neuartige Suchansätze bieten. Dies kann an der (zu frühen) Kommerzialisierung der Suchtechniken und -systeme liegen. Gegenüber der herkömmlichen Boole'schen Suche mit einfachen logischen Mengenoperatoren bietet WAIS etwa statistische Suchverfahren und die Möglichkeit einer schrittweisen Verfeinerung der Ergebnisse durch das sog. Relevance-Feedback, bei dem der Benutzer angibt, welche Suchergebnisse er für besonders relevant hält. Auf der Basis der Ähnlichkeit zu diesen Dokumenten wird dann eine erneute, verfeinerte Suche durchgeführt. Allerdings sind ähnliche Suchverfahren auch bei den nachfolgend beschriebenen großen Suchsystemen im Einsatz - wenn auch nicht als verteilte Datenbank. Die Integration in das World-Wide Web bzw. Browser erfolgt hier durch sog. **Gateways**, d.h. über ein HTML-Formular wird die Suchfrage eingegeben und mittels eines sog. **cgi**-Skriptes (meist kleine Programme in einer relativ einfachen Skriptsprache wie Perl, aber auch C++) in eine entsprechende WAIS-Abfrage umgewandelt. Die Ergebnisse werden entsprechend der Ausgabeformate in eine HTML-Darstellung konvertiert.

Auf die dritte Form der verteilten Informationssysteme, die Usenet News, soll wegen der großen Bedeutung und dem sehr eigenen Charakter im folgenden näher eingegangen werden.

2. Usenet News

Die Usenet News stellen als eine Art "elektronische Notizbretter" oder Diskussionforen eine Zwischenform zwischen der hauptsächlich der persönlichen Kommunikation dienenden E-Mail und den hier beschriebenen "Mensch-Maschine"-Informationssystemen dar, wo der Benutzer mit einer Maschine kommuniziert. Es unterscheidet sich von den anderen Informationssystemen auch durch seine extreme Kurzlebigkeit, da die einzelnen Diskussionsbeiträge oft nur wenige Wochen auf den einzelnen "Notizbrettern", den Newsgroups, aktiv ("ausgehängt") sind. So kann es passieren, dass ein Beitrag innerhalb eines "Diskussionsfadens" ("Thread") auf einen früheren Beitrag antwortet, der schon gar nicht mehr abgerufen werden kann. Allerdings werden von vielen Newsgroups die Beiträge archiviert und sind dann über andere Kommunikationsdienste (z.B. Volltextsuche **innerhalb** eines Webservers, vgl. [Abschnitt 3](#)) nutzbar.

Ein weiteres Charakteristikum ist die strenge Ethik, sog. **Netiquette**, die z.B. kommerzielle Beiträge strikt ablehnt. Diese können deshalb auch entsprechend sanktioniert werden.

Im groben sind die Newsgruppen, wie in der Tabelle oben angedeutet, streng hierarchisch organisiert. Ausgehend von wenigen Hauptkategorien (z.Zt. etwa 40) wird jede Gruppe so weit aufgeteilt, wie es sich als sinnvoll erweist (und genug Interesse an Spezialthemen besteht).

Neben den inhaltlichen Hauptkategorien:

- *alt* = unterschiedliche, oft technische Themen, können solche Kategorien auch einzelnen Ländern, z.B.
- *comp* = Computer, • *de* = Deutschland, *it* = Italien
- *misc* = diverses, oder spezifischen Organisationen/ Vereinigungen zugeteilt sein:
- *rec* = Freizeit, • *bionet* = biowissenschaftliche Themen
- *sci* = Wissenschaft und Forschung, • *fidonet* = Das gleichnamige Mailboxsystem
- *soc* = soziales und politisches Leben

Die Untergliederung fängt aber hier erst an und kann sich bis auf sehr spezielle Themen fortsetzen, z.B.

- *comp.sys.windows.networks* = Netzwerkfragen rund um Windows oder
- *soc.culture.italian* = Italienische Kultur (bzw. Politik)

Technisch ist das Usenet als größtes Forum von Newsgroups - daneben bieten kommerzielle oder kleine gemeinnützige Systembetreiber ihre eigenen, internen "Bulletin Boards" mit mehr oder weniger der gleichen Technik an) - als eigenes Netz von spezifischen Servern organisiert. Jeder dieser Server speichert lokal eine genau definierte Auswahl aller verfügbaren Newsgroups und tauscht neue Diskussionsbeiträge mit den anderen Servern aus. Dies führt dazu, dass z.B. in der "geographischen" Hierarchie die italienischen Gruppen (it..) von Deutschland aus normalerweise nicht zugreifbar sind, da sie auf deutschen Servern nicht gespeichert werden. Dies betrifft allerdings nicht die schon angesprochenen Archive solcher Gruppen, die über andere Methoden bzw. Protokolle zur Verfügung gestellt werden.

Da viele Gruppen inhaltlich identisch mit den entsprechenden Mailinglisten sind oder genau die gleichen Themen behandeln, kann auf viele dieser Beiträge auch zugegriffen werden, ohne einen sog. **Newsreader** (den Client zum Lesen von News) zu verwenden. Allerdings ist mit der Integration der **News in Netscape**, eine dem Web-User relativ vertraute Zugriffsmöglichkeit auf die News geschaffen worden, die auch eine weitgehende Integration erlaubt (so kann z.B. auf einer HTML-Seite für aktuelle Information auf die News verwiesen werden und - was wesentlich häufiger vorkommt - in Beiträgen zu den News auf HTML-/WWW-Seiten, die vom System auch als solche erkannt und damit aktiviert werden können.

Außerdem bietet z.B. das **Stanford Information Filtering System (SIFT, sift.stanford.edu)** die Möglichkeit, eine oder mehrere Gruppen nach bestimmte Stichwörtern zu durchsuchen. Denn trotz der potentiell sehr tiefen Verschachtelung und damit Spezialisierung von solchen

Gruppen, erreichen viele eine sehr große Breite der diskutierten Themen und ein sehr großes Volumen an Beiträgen.

Abschließend soll darauf hingewiesen werden, daß die bei der E-Mail aufgeführten [Netiquette-Regeln, Abkürzungen und Symbole](#) in besonderem Maße auch für die Usenet News gelten und noch durch weitere Regeln ergänzt werden:

- Beiträge sind möglichst kurz zu halten, d.h. es soll nicht endlos aus anderen Beiträgen zitiert und auch keine seitenlange "signature" eingefügt werden.
- Um Beiträge angemessen zu halten, soll eine Diskussion erstmal einige Zeit passiv verfolgt werden, bevor selbst aktiv mitdiskutiert wird. Auf diese Weise lernt man den Themenbereich ("Scope") der Gruppe kennen. Anleitungen und sog. FAQ-Dateien (Frequently Asked Questions) beantworten die wichtigsten Fragen von allgemeinem Interesse und soll unnötige Wiederholungen vermeiden.

B. Suche im Internet

Über das Internet können vielerlei - auch kommerzielle - Informationssysteme zu allen möglichen Themen, Informationstypen und Inhalten abgefragt werden.

Diese Werkzeuge zur Informationssuche lassen sich grob nach folgenden Kriterien aufteilen:

	Referenzinformation, d.h. es wird auf (meist"traditionell" gedruckte) Information verwiesen: Bücher und Zeitschriften(-artikel)	Elektronische Primärinformation, d.h. die volle Information (meist Texte) ist elektronisch verfügbar
Stichwort	meist kommerziell: FIRSTSEARCH, STN, DIALOG	Lycos , Excite , AltaVista Infoseek (auch News)
Mischform		Yahoo , Web.de
Systematisch		
Sachthemen	Bibliothekskataloge , Buchhandels-/ Verlagslisten	Virtual Library (http://www.w3.org/vl) E-journals u. Newsgroups zu best. Themen
Geographisch	Adressensammlungen von Firmen und Institutionen (z.B. Yellow Pages /Gelbe Seiten)	Virtual Tourist und Citynet (Städte und Länder), länderspezifische Listen
andere Informationsarten		
Personen	Listen von (e-mail) Adressen (Four11)	Personenbezogene Homepages (z.B. von Frank Krüger)
Newsgroups, Mailinglisten		DejaNews , Reference (150 Tsd. Diskussionsgruppen)
Programme und andere Dateien	Kataloge bzw. Informationsangebote einschlägiger Versandhäuser (z.B. Softline) und Softwareherstellern (z.B. Microsoft)	Shareware-Sammlungen, z.B. FILEZ oder TUCOWS

Davon zu unterscheiden sind Suchmöglichkeiten, die sich lediglich auf einen bestimmten Server bzw. eine Subdomain (z.B. Universität Mainz) beziehen. Unter diesen "**lokalen**" Suchsysteme hat sich in letzter Zeit das Indexierungs- und Suchsystem "**HARVEST**" neben **GLIMPSE** und einigen anderen besonders hervorgetan.

Hinweis: In diesem Kapitel wird hauptsächlich die grobe Struktur der Informationsquellen im Internet vorgestellt. Eine Sammlung konkreter und (möglichst) aktueller (!) Verweise findet sich in [Anhang G!](#)

1. Gezielte Stichwortsuche

a) Suchroboter

Die als **Roboter** oder **Würmer** (*Web Worm*) bezeichneten Suchsysteme bieten die umfassendsten - wenn auch bei weitem nicht vollständigen - Suchmöglichkeiten im WWW, da sie die Informationen über vorhandene Seiten und Informationsangebote weitgehend automatisch sammeln. Dazu werden regelmäßig alle bekannten Server abgesucht, die dort vorhandenen Seiten eingelese und indexiert. Außerdem wird den Verweisen auf diesen Seiten nachgegangen, so daß auch die neu hinzugekommenen Server und Seiten erfaßt werden können.

Bei der **Suche** mit diesen Systemen werden konkrete Wörter eingegeben, die im Titel oder Text der WWW-Seiten vorkommen sollen. In einigen System (etwa Lycos) wird nicht nur exakt der eingegebene Begriff gesucht, sondern auch alle Varianten und ähnlichen Wörter mit berücksichtigt. Die Ergebnisse werden dann in einer mehr oder weniger sinnvollen Reihenfolge (oder das, was der Rechner dafür hält) sortiert. Weitere Optionen lassen zu, die Suchwörter mit sog. Operatoren (v.a. logisches UND, ODER, NICHT) zu verknüpfen oder die Ergebnisse auf eine bestimmte Höchstmenge zu beschränken.

Der **Nachteil** dieser Roboter ist, daß sie sich - wie das WWW insgesamt - zum großen Teil auf den nordamerikanischen Raum konzentrieren, so daß oft z.B. europäische Verweise entweder gar nicht erst auftauchen oder veraltet sind. Allerdings ist das WWW insgesamt sehr stark von der englischen Sprache geprägt, so daß auch die Suchbegriffe zunächst in englisch eingegeben werden sollten - es sei denn, es werden explizit Dokumente in einer anderen Sprache gesucht. Mittlerweile stehen u.a. auch auf die Bundesrepublik bzw. die deutschsprachigen Ländern spezialisierte Suchsysteme zur Verfügung.

Diese Suchsysteme sind auf alle Fälle als erster Einstieg besonders dann sehr nützlich, wenn man nur eine ungefähre Vorstellung von seiner Suche hat. Anhand der Ergebnisse können dann die Suchbegriffe präzisiert werden oder man gelangt recht schnell zu systematischen Sammlungen und Zusammenstellungen des betreffenden Themas.

Ein besonderer Fall ist das Suchsystem **Yahoo**, daß einerseits zwar seine Informationen automatisch gewinnt, aber andererseits eine systematische Gliederung anbietet, nach der man seine Suche eingrenzen kann.

So kann man etwa nicht nur nach den Begriffen "*Umweltschutz*" oder "*Sport*" suchen, sondern diese Themen auch auf bestimmte Regionen ("*Portugal*") oder Sachgebiete ("*Schiffahrt*") eingrenzen. Außerdem wird Yahoo inzwischen in an bestimmte Länder bzw. Sprachregionen und -kulturen angepaßten Versionen angeboten. So finden sich bei jeder sprachspezifischen Version von YAHOO (also <http://www.yahoo.de> für den deutschsprachigen Raum, www.yahoo.fr für Frankreich/französisch, www.yahoo.it für Italien, www.yahoo.co.jp für japanisch) auch zahlreiche Hinweise auf andere Suchsysteme, die das Informationsangebot in der jeweiligen Sprache systematisch klassifizieren bzw. mit Stichwörtern indexieren.

b) Kommerzielle Datenbanken im Internet

Stellvertretend für diese Gruppe soll der us-amerikanische Dienst FIRSTSEARCH genannt werden, der von der UB Mainz abonniert wurde und damit auch dem Fachbereich ohne weitere nutzungsabhängige Kosten zur Verfügung steht.

FIRSTSEARCH bietet die Möglichkeit der Suche in zahlreichen, meist populärwissenschaftlichen, Datenbasen, den elektronischen Gegenstücken zu Bibliographien und Referatezeitschriften. So werden etwa in einer Datenbasis 14 Tsd. (Mai 1997) Zeitschriftenartikel nachgewiesen, also mit Titeln, Schlagwörtern und Quellenangabe erfaßt. Der volle Text dieser Artikel ist jedoch in den seltensten Fällen elektronisch abrufbar. Allerdings stehen für viele eher wissenschaftliche Fachartikel (z.B. medizinische Aufsätze in der Datenbasis Medline) Kurzfassungen, sog. Abstracts, zur Verfügung, die unter Umständen (etwa als Quelle für einschlägige Terminologie) schon ausreichen.

Die Nutzung dieses Dienstes ist an eine Campus- oder Großkundenlizenz gebunden, wie sie etwa die Universität bzw. UB Mainz abgeschlossen hat. Nähere Information bzw. der Zugang zu FIRSTSEARCH ist über den [Server der UB Mainz](#) verfügbar

2. Systematische Suche

Im Unterschied zu den Systemen, die eine gezielte Stichwortsuche anbieten, erfassen die Listen und Server in dieser Rubrik das Informationsangebot nur selektiv. Aber gerade darin kann ein großer Vorteil in dem Informationsmeer "INTERNET" liegen, wenn nämlich die Angebote redaktionell überprüft und gefiltert sind.

So werden für Deutschland verschiedene manuelle Listen geführt (etwa <http://web.de> oder <http://www.leo.org>), die (alle) Server erfassen und nach bestimmten Kriterien klassifizieren. Allerdings sind die Kriterien für die Auswahl und Klassifizierung nicht immer nachvollziehbar oder auch in kommerziellen Interessen (Werbung!) begründet. Oder die Listen bzw. Server verfolgen einen passiven Ansatz, d.h. alle Anbieter von Servern (sog. Webmaster) oder einzelnen Seiten müssen selbst ihr Angebot an diese Listen melden.

a) Sachthemen

Eine Aufteilung nach Sachthemen ist sicherlich die "klassische" Form der Systematik.

Hier läßt sich deutlich nach Angeboten mit Referenz- vs. Primärinformation unterscheiden:

- Nach Sachgebieten geordnete **Referenzinformationen** sind v.a. in Buchhandels- und Verlagslisten zu finden. Elektronische Bibliothekskataloge (sog. OPACs) lassen sich ebenfalls nach Sachthemen befragen und geben je nach Größe (z.B. Library of Congress für die USA) einen recht umfassenden Überblick zu entsprechenden Büchern.
- Speziell zu Büchern oder buchähnlichen Texten enthalten die sog. **text depositories** (z.B. [Project Gutenberg](#)) die kompletten Volltexte. Dabei kann es sich um reine

ASCII-Texte (d.h. ohne Querverweise oder sonstige Formatierung wie Fettdruck etc.) handeln oder um speziell für das WWW aufbereitete Hypertexte mit Querverweisen (z.B. juristische Texte wie Verfassungen und völkerrechtliche Verträge)

- Die sog. **Virtual Library** ist dagegen ein von dem W3-Konsortium geführtes Gemeinschaftsprojekt vieler Institutionen weltweit, in dem basierend auf einer allgemeinen Klassifikation Nachweise zu den einzelnen Sachgebieten (inklusive Sprachen und Kulturen) dezentral an unterschiedlichen Institutionen gesammelt wird.
- An dieser Stelle sollen auch die **Usenet Newsgroups** genannt werden, die zwar nicht zum WWW im eigentlichen Sinne gehören, oft aber auch eine gute Informationsquelle darstellen. Auf sie kann entweder über die systematische Hierarchie oder mit einigen Suchsystemen (z.B. Altavista) zugegriffen werden.

b) Geographisch

Von den allgemeinen, systematischen Sammlungen zu einem bestimmten Land bzw. Kulturkreis sind die Angebote detaillierter geographischer Information zu unterscheiden (mit vielen Überschneidungen).

Ein solches weltweites geographische Zentralverzeichnis stellt das Angebot "**Virtual Tourist**" (www.vtourist.com) dar, wo v.a. auch Städte- und Länderinformation angeboten werden. Speziell für Hochschulen und wissenschaftliche Einrichtungen gibt es darüberhinaus in vielen Ländern eigene Listen der "akademischen" Server, die auch heute noch in der Mehrzahl sind.

In dieser Rubrik überwiegen jedenfalls die elektronischen "Primärinformationen" deutlich bzw. sind kaum von den "Referenzinformationen" wie Telefonnummern (Auskunft der Telekom), Postleitzahlen und Ansprechpartner für weiterführende Informationen zu trennen.

c) Suche nach Personen und Institutionen

Neben den oben erwähnten Suchrobotern, die eine Suche nach Personen und Institutionen als Stichwörter ermöglichen, werden auch immer mehr spezielle Informationsdienste insbesondere für Firmen gegründet. Solche Systeme sind aber auch wiederum meist geographisch begrenzt, wie etwa www.yip.com als elektronisches Gegenstück zu den Gelben Seiten für USA/Canada.

Somit macht vor allem die Suche nach einzelnen Personen große Schwierigkeiten. Dies betrifft auch die Suchmöglichkeiten außerhalb des WWW (z.B. netfind), da z.B. aus datenschutzrechtlichen Gründen kaum deutsche Adressen verzeichnet sind. Andere Systeme durchforsten für ihre Adreßverzeichnisse v.a. Newsgroups, um die dort enthaltenen E-Mail-Adressen und zugehörigen Personenangaben der Diskussionsteilnehmer zu sammeln. Eine vielversprechendere Methode ist, auf dem Server der jeweiligen Firma oder Institution, bei der eine Person arbeitet, eine Stichwortsuche durchzuführen. Dies bieten zwar noch nicht alle Einrichtungen an, aber v.a. bei der Suche nach Mitarbeitern an Hochschulen oder wissenschaftlichen Forschungseinrichtungen kommt man damit meist zu der gewünschten Adresse, Telefonnummer oder E-Mail Adresse.