# Suchhilfen für das World Wide Web: Funktionsweisen und Metadatenstrukturen

Manfred A. Jeusfeld, Matthias Jarke

## 1 Einleitung

Stellen Sie sich vor, Sie seien Bibliothekar einer Bibliothek mit gut 50 Millionen Nutzern aus über hundert Ländern, die über einen Bestand von einigen Dutzend Millionen Dokumenten unterschiedlichster Form verfügt. Täglich gibt es mehrere Millionen Nachfragen nach Dokumenten, täglich werden tausende Dokumente gelöscht und zehntausende neue Dokumente hinzugefügt [Sher95]. Wie würden Sie einen Katalog dieser Bibliothek aufbauen, der den Nutzern eine verläßliche Hilfe bei der Suche nach Information bietet? Vor dieser Frage standen auch die Entwickler des World Wide Web [Bern94], als sie es vor etwa 7 Jahren starteten. Ihre Antwort lautete damals: es gibt keinen zentralen Katalog, jeder Nutzer ist selbst dafür verantwortlich, sich seinen persönlichen Index interessanter Dokumente aufzubauen. Diese Antwort ist bei der schieren Größe der Bibliothek und deren exponentiellen Wachstum heute nicht mehr zufriedenstellend. Laut [Gray95] verdoppelte sich die Zahl der Informationsanbieter im WWW im Jahre 1993 alle 3 Monate. Selbst im Jahre 1995 lag diese Periode immer noch bei 5 Monaten. Die Suche in einem solch großen, dynamischen und unstrukturierten Informationsmarkt ist zweifellos eine enorme Herausforderung. Wir sprechen hier von *Informationsmarkt* im Vorgriff auf eine sich abzeichnende Entwicklung. Eine große Zahl Informationsanbietern und -nachfragern treffen aufeinander. Suchhilfen sind die Werkzeuge, sich in diesem Markt zurechtzufinden. Dieser Beitrag soll die verschiedenen Suchmethoden im globalen Informationsmarkt vorstellen und einordnen.

## 2 Dokumentzugriff und Suche im World Wide Web

Bevor wir uns den Suchmaschinen zuwenden, ist der Grundmechanismus des World Wide Web zu erklären. Das World Wide Web beruht wesentlich auf der eindeutigen Identifizierung eines Dokuments (*uniform resource locator* [Bern95], im folgenden URL oder Verweis genannt). Ein typischer URL ist wie folgt aufgebaut:

http://www.marketing.www-ag.de/kunde/produkte.html

Der erste Teil (hier: http) spezifiziert das Übertragungsprotokoll, danach folgt der eindeutige Name eines Knotens im weltweiten Rechnernetz Internet (bis zum 3. Schrägstrich) und der Verzeichnispfad zu dem Dokument auf dem spezifizierten Knotenrechner. Jeder gängige WWW-Browser ist in der Lage, bei Angabe des URLs das entsprechende Dokument zuzugreifen. Der eindeutige Rechnername wird im Internet via eines sogenannten Name-Servers auf eine eindeutige Rechnernummer abgebildet. Die Internetnummer dient wie eine Postanschrift der Weiterleitung (engl.: routing) von Datenpaketen im Rechnernetz genutzt wird. Bild 1 zeigt das Prinzip des Internet-Zugriffs unter Berücksichtigung von Proxy-Servern (engl.: Stellvertreter). Solche Rechner werden vielerorts zur Beschleunigung und als Sicherheitsbarriere eingesetzt. Ein Proxy-Server fängt Internet-Zugriffe ab und überprüft, ob das Ergebnis eines Zugriffs bereits in seinem Cache (Zwischenspeicher) abgelegt ist. Wenn ja, wird der Zugriff aus dem Cache bedient. Die Auflösung eines URL-Zugriffs geschieht vereinfacht dargestellt in folgenden Schritten:

1. Zunächst wird der Zugriff an den lokalen Proxy-Server weitergeleitet. Dieser durchsucht seinen Zwischenspeicher, ob das spezifizierte Dokument dort vorhanden ist. Wenn ja, so wird es an den WWW-Browser transferiert und die Anfrage ist beantwortet.
2. Wenn es nicht im ersten Cache vorhanden ist, so leitet der Proxy-Server die Anfrage zu benachbarten Proxy-Servern weiter. Wenn es dort vorhanden ist, so wird es entlang der Aufrufhierarchie (fett gezeichnete Pfeile in Bild 1) an den WWW-Browser zurückgegeben.
3. Falls auch die benachbarten Proxy-Server das Dokument nicht gespeichert haben, so beschafft der letzte Proxy-Server in der Aufrufhierarchie das Dokument im Internet. Dieser Zugriff kann entweder direkt beim spezifizierten Ziel-Server www.marketing.www-ag.de ankommen oder wiederum bei einem Proxy-Server, diesmal dem Stellvertreter für den Ziel-Server.

Manfred A. Jeusfeld, Matthias Jarke,
RWTH Aachen, Lehrstuhl Informatik V,
Ahornstr. 55, D-52056 Aachen,
http://www-i5.informatik.rwth-aachen.de

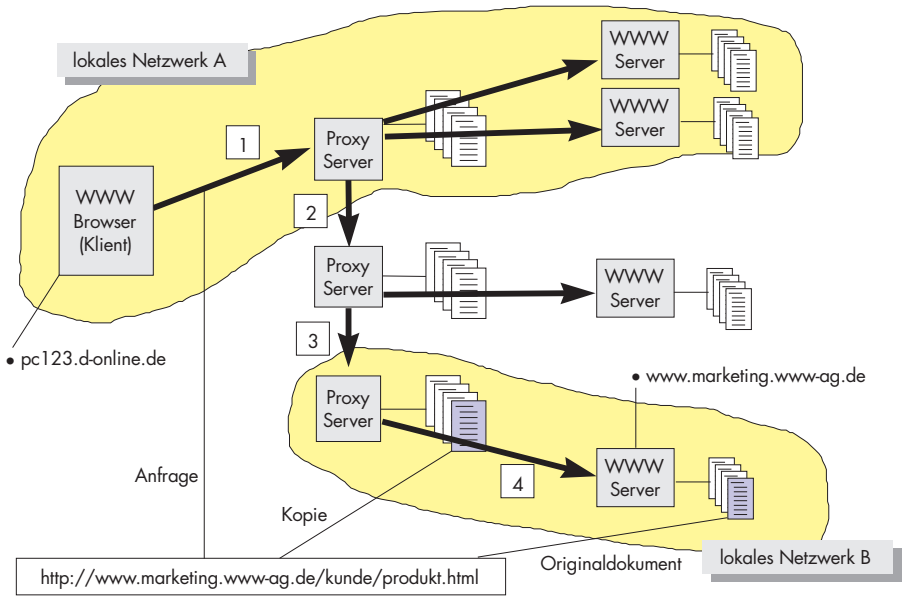


**Bild 1** Prinzip des Dokumentzugriffs mit Proxy-Servern

4. Falls auch dieser das Dokument nicht zwischengespeichert hat, so erfolgt letztendlich der Zugriff auf das Originaldokument. Als Seiteneffekt wird das Dokument im Cache der beteiligten Proxy-Server repliziert. Zusätzlich verwalten manche WWW-Browser zugegriffene Dokumente in einem lokalen Cache.

Vorteil der Proxy-Server ist eine bessere Ausnutzung der begrenzten Übertragungskapazitäten. Erkauft wird dies mit einer gewissen Rate an veralteten Dokumenten in den Zwischenspeichern, die anstatt des aktuellen Originaldokuments an den Benutzer zurückgegeben werden. Dies ist ein grundsätzliches Problem.

Die per URL bezeichneten Dokumente können beliebiger Art sein: Textdokumente, Bilder, Videosequenzen, Töne, Programmcode, Datenbankinhalte, Spreadsheets, Meßreihen und so weiter. Um dem Klienten (hier: WWW-Browser) eine Hilfestellung bei der Einordnung eines Dokuments zu geben, wird vom WWW-Server neben dem Dokument auch seine Typangabe zurückgegeben. Als Format hat sich der sogenannte MIME-Typ durchgesetzt. Zum Beispiel steht die Typangabe ‚text/plain' für Texte. Die Liste erlaubter Typangaben ist nicht beschränkt. Jeder Administrator eines WWW-Servers kann neue Typangaben definieren. Eine Überprüfung, ob ein Dokument seinem Typ entspricht, ist nicht vorgeschrieben. Die am weitesten verbreitete Typangabe ist ‚text/html' für Hypertextdokumente. Solche Dokumente zeichnen sich dadurch aus, daß sie in ihrem Text Verweise (URL) auf andere Dokumente enthalten.

| | generisch | domänenspezifisch |
|---|---|---|
| referenzbasiert | HTML / HTTP-Navigation | Hotlists Themenkataloge |
| wertebasiert | Suchmaschinen im engen Sinne | Informations-Broker |

**Bild 2** Einordnung von Suchhilfen

## Anforderungen an Suchhilfen und Klassifikationskriterien

Nach [BDMS94] unterscheidet sich das WWW von traditionellen Datenbanken vor allem in drei Aspekten:

- einem enormen und ständig wachsendem Datenvolumen
- einem wesentlich breiteren Benutzerspektrum mit sehr heterogenen Rechnerkenntnissen
- einer großen Diversität der verwalteten Dokumente, die wenig Annahmen über eine einheitliche Struktur zuläßt.

Die ursprünglich realisierten Konzepte kamen auch nicht aus dem Datenbankbereich, sondern wurden von Spezialisten für verteilte Betriebssysteme entwickelt [BCLN94].

Die weitere Entwicklung weist aber überraschende Analogien zur Entwicklung der Datenbanksysteme aus Dateisystemen in den 60er und 70er Jahren auf. Diese sind in Bild 2 gegenübergestellt. In der einen Dimension dieser Tabelle werden wie im Datenbankbereich referenzbasierte Ansätze und wertebasierte[1] Ansätze gegenübergestellt. In der anderen kontrastieren wir generische Konzepte und domänenspezifische Sichten. Als Felder der Tabelle ergeben sich damit Navigationstechnologien, benutzerdefinierte Hotlists, Suchmaschinen und Informations-Broker.

Die ersten Ansätze waren wie das hierarchische Datenmodell und das Netzmodell referenzbasiert [Date95]. Wie im vorigen Abschnitt beschrieben, stellt sich das WWW als weltweites Hypertext-Netz dar, in dem die Benutzer unter Nutzung der URL und diverser Dokument-Zwischenspeicher mehr oder weniger effizient navigieren können. Dies ist der vor allem durch den Erfolg von Netscape weitverbreitete Browser-Ansatz. Krasser noch als in anderen Hypertexten stellt sich hier das Problem des „getting lost in [hyper] space" [Conk87]. Schon nach wenigen Hypertext-Links fragt sich der Benutzer: Wo bin ich? Wie bin ich hierher gekommen? Wohin gehe ich sinnvollerweise weiter? Wie kann ich die Ergebnisse erfolgreicher Suchvorgänge für mich und andere persistent verfügbar machen?

Der Aufbau benutzerdefinierter *Hotlists*, Bookmarks o.ä. stellt hier eine erste wichtige Strategie dar, die noch weitgehend innerhalb des Navigationsansatzes verbleibt. Die Hotlists sind im einfachsten Falle selbsterstellte WWW-Seiten, auf denen Verweise gesammelt werden; dies wird mittlerweile von allen Browser-Herstellern routinemäßig angeboten. Weitergehende Ansätze etwa im Groupware-Bereich gestatten es, derartige Hotlists untereinander und zu organisationsspezifischen Themen in Beziehung zu setzen, um so den Zugriff auf Hotlists anderer Gruppenmitglieder zu erleichtern [Hall97].

Unter *Suchmaschinen* im engeren Sinne werden jedoch Systeme verstanden,

die – analog zum Übergang vom netzbasierten zum relationalen Modell – werteorientierte Suche gestatten. Aufgrund der Diversität der zuzugreifenden Dokumente, aber auch aufgrund der geringen Vorkenntnisse der meisten Benutzer ist es bisher jedoch nicht gelungen, ein vergleichbar einfaches Daten- und Abfragemodell zu entwickeln. Vielmehr weisen die einzelnen Suchmaschinen große Unterschiede darin auf, wie sie Metadaten über die Dokumentenbasis erzeugen und welchen Ausgleich zwischen Benutzerfreundlichkeit, Ausführungseffizienz und Ausdrucksfähigkeit sie bei der Ausgestaltung der Abfragesprache selbst verfolgen.

Viele Forscher argumentieren, daß sich generische Suchmaschinen ihren Grenzen bezüglich Informationsqualität und Ressourcenbelastung nähern. Auch ist zunehmender Mißbrauch durch bewußte Fehlklassifikation aus Marketinggründen zu beobachten. Eine Reihe von Arbeiten widmen sich daher dem Entwurf höherwertiger, domänenspezifischer *Informations-Broker*, welche mit mäßig höherem Aufwand als derzeitige Suchmaschinen eine wesentlich verbesserte Informationsqualität für bestimmte Kundenkreise zu erreichen suchen. Hier kann eine Analogie zur Informationsssuche in heterogenen Datenbanken [PaMi96] gesehen werden. Auch Verfahren aus der Künstlichen Intelligenz werden hier erprobt.

### Bewertungskriterien

Wenn wir das WWW als globalen Informationsmarkt betrachten, so sind Suchhilfen aus Anbieter- und aus Kundensicht zu beurteilen. Anbieter sind daran interessiert, den Ort der Verweise so zu bestimmen, daß der Leserkreis optimal erreicht wird. Wegen der Fülle des Angebots und der begrenzten Aufnahmefähigkeit der Leser ist dies ein Problem, das mittels Preisgestaltung handzuhaben ist. Wir befassen uns hier jedoch ausführlicher mit der Kundensicht. Im Information Retrieval [SaBu88; Tenn96] werden Suchhilfen klassisch nach den Kriterien Recall (Wiederfindungsgrad) und Precision (Anteil fehlerhafter Einträge in der Antwort) bewertet. Eine differenzierteres Beurteilungsschema der Informationsqualität an der Sloan School des MIT entwickelt ([WaWa96], Bild 3).

Im folgenden stehen Daten für die Antworten der Suchhilfe auf eine Suchanfrage

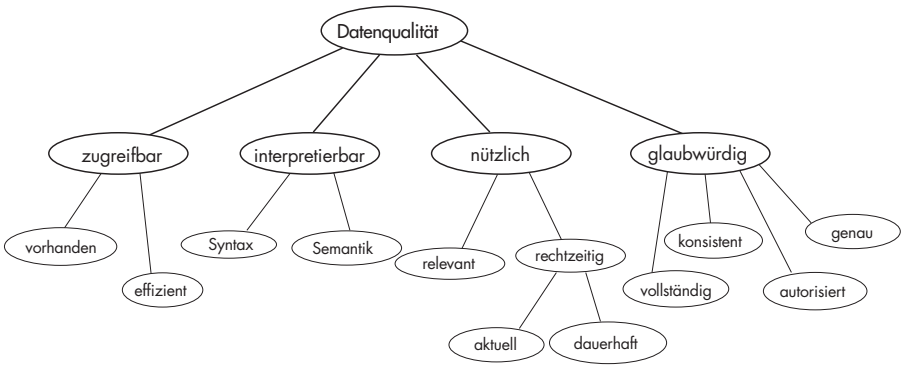


**Bild 3** Ontologie der Datenqualität nach [WaWa96]

und nicht für die Originaldokumente, deren Referenzen in den Antworten auftauchen. Der Aspekt *Zugreifbarkeit* betrifft das Antwortverhalten der Suchhilfe bei Eingabe einer Suchanfrage. Die *Interpretierbarkeit* der Antwort hängt von ihrer gleichförmigen Darstellung (Syntax) und dem Grad der Übereinstimmung über die Semantik von Worten ab. Die *Nützlichkeit* mißt einerseits, ob die Antwort aktuell ist und andererseits, ob sie relevant zur Suchanfrage ist (entspricht der Präzision beim Information Retrieval). Die *Glaubwürdigkeit* einer Antwort hängt unter anderem von der Konsistenz und der Vollständigkeit ab.

Die Zugreifbarkeit wird einerseits durch den Abdeckungsgrad der gewählten Suchhilfe bestimmt, was für vollautomatische und generische Suchmaschinen spricht. Andererseits spielt auch die Effizienz eine Rolle; diese stößt bei sehr großen Nutzerzahlen zunehmend an ihre Grenzen. Hinzu kommt das Problem, daß durch das ansteigende Datenvolumen die Metadaten der Suchmaschinen zunehmend veralten – dazu später mehr.

## 3 Vom Auskundschaften und Auskunftgeben

Kein Mensch kann den Überblick über das gesamte Netz zu einem bestimmte Zeitpunkt haben und über die sich dynamisch wandelnde Struktur behalten. Automatisierte Suchmaschinen sammeln rund um die Uhr Informationen *über* die Dokumente und die aktuelle Struktur des Netzes. Mit Hilfe der gesammelten Metadaten über das Netz können dem Nutzer Verweise auf Dokumente geliefert werden, die z.B. eine bestimmte Kombination von Stichwörtern enthalten. *Metadaten* sind alle Daten über ein Dokument, also sein MIME-Typ, seine Autoren, sein Erstellungsdatum, das Datum der letzten Änderung, die im Dokument enthaltenen Stichworte und so weiter.

Eine *Suchmaschine* besteht aus drei Teilen: 1) ein Metadaten sammelndes Programm, 2) den gesammelten Metadaten und 3) einer Schnittstelle, über die der Benutzer Auskunft von der Suchmaschine verlangt. Das Suchprogramm traversiert das WWW, indem es Verweisen innerhalb der bearbeiteten Dokument auf noch nicht bekannte Dokumente folgt. Dabei baut es eine Datenbank mit Informationen über die Dokumente und die Struktur des Netzes (d.h. die expliziten Verweise zwischen den Dokumenten) auf. Die Suchmaschine wandert nicht selbst durch das Netz, wie seine populären Namen Roboter, Wanderer, Wurm oder Spinne [Kost 96a] nahelegen. Sie „bewegt“ sich nur dadurch fort, indem sie Dokumente von immer wieder anderen Stellen des Netzes liest. Die grobe Vorgehensweise ist bei den automatisierten Suchprogrammen im wesentlichen gleich. Von einer kleinen Menge von Dokumenten ausgehend folgt es schrittweise den Verweisen auf weitere Dokumente. Immer wenn ein neues Dokument besucht wird, werden dessen bibliographische Informationen extrahiert und in die Datenbank der Suchmaschine eingetragen. Zudem werden die enthaltenen neuen Verweise (URL) in die Liste der noch zu verfolgenden Verweise eingefügt. Dieser Prozeß läuft im Prinzip unendlich. Ergebnis ist die *Metadatenbank* der Suchmaschine als *Abbild* der vernetzten

Dokumente im WWW. Für die Auswahl des nächsten zu verfolgenden Verweises aus der Liste sind in den Suchprogrammen unterschiedliche Strategien implementiert, auf die hier jedoch nicht näher eingegangen wird.

## Extraktion von Metadaten

Nachdem die grundsätzliche Vorgehensweise zum Aufsuchen der Dokumente geklärt ist, stellt sich die Frage, welche Information über die Dokumente in der Metadatenbank abzulegen ist. Ziel muß sein, ausreichend Information für die weiter unten beschriebene Auskunftsfunktion aufzubauen. Wir unterscheiden hier automatisierte und manuelle Bestimmung der Metadaten. Das generisches Schema für die Metadaten sieht dabei etwa wie folgt aus:

*(Metakategorie, Metadatum, URL)*

Die Metakategorie klassifiziert das Metadatum und der URL verweist auf das Originaldokument. Ein Dokument kann wie folgt in der Metadatenbank beschrieben sein:

(Autor, Peter Name, http://server.dept.org/doku1.html)
(Erstellt, 1.4.1996, http://server.dept.org/doku1.html)
(Stichwort, Datenbank, http://server.dept.org/doku1.html)

Bei manuellen Methoden extrahiert ein Experte (oft der Autor) die Metadaten aus dem Dokument. Diese Arbeit ist analog zu der Erfassung von Dokumenten in einer Bibliothek: Autor, Titel, Schlüsselwörter etc. werden im Zusammenhang mit der Inventarisierung des Dokuments in den Bibliothekskatalog eingetragen. Die im WWW vielfach verwendete Dokumentbeschreibungssprache HTML bietet hier eine elegantere Möglichkeit: Metadaten werden dem Dokument quasi als Kommentar hinzugefügt. Die Suchmaschine extrahiert dann diese Kommentare und bildet daraus die Einträge in der Metadatenbank. Die Voraussetzung, geeignete Kategorien für die Kommentare vorzugeben, ist bisher nicht erfüllt. Schließlich möchte man nicht für jede Suchmaschine spezielle Kommentare schreiben müssen. Die Suchmaschine ALTAVISTA (www.altavista.com) etwa berücksichtigt folgende Kommentare in zu indizierenden HTML-Dokumenten:

```
<META   NAME=„DESCRIPTON"
        CONTENT=„LEITSEITE DES
        PROJEKTES ALPHA">
<META   NAME=„KEYWORDS"
        CONTENT=„KUNDEN-
        ORIENTIERUNG INFORMA-
        TIONSSYSTEM">
```

Themenkataloge wie YAHOO (www.yahoo.com) bauen ebenfalls auf die Mitwirkung der Autoren. Bei Eintragen in das Verzeichnis muß sowohl Klassifikation in die Hierarchie als auch ein kurzer zusammenfassender Absatz über das Dokument manuell vorgeben werden. Bei der *manuellen Methode* haben also die Autoren der Dokumente einen erheblichen Einfluß darauf, in welcher Weise ihr Dokument in der Metadatenbank der Suchmaschine erscheint. *Automatische Methoden* extrahieren die Metadaten aus dem Inhalt der Dokumente:

1. Hole die Textdatei
2. Extrahiere alle Wörter außer Funktionswörtern wie ‚the' etc.[2]
3. Kreiere die Metadaten, wobei jedes der extrahierten Wörter ein Stichworteintrag ergibt

Das Ergebnis dieser Stichwortextraktion ist insofern präziser als beim manuellen Verfahren, da prinzipiell alle Wortvorkommen berücksichtigt werden. Auf der anderen Seite können automatische Verfahren nur sehr eingeschränkt die gefundenen Stichwörter ihrer Bedeutung nach gewichten. Ein besonderes Problem ist die Anfälligkeit gegenüber Dokumentänderungen. Falls das ganze Dokument Gegenstand der Stichwortextraktion ist, so können selbst kleine Änderungen des Textes die Metadaten ungültig bzw. unvollständig machen. Daher beschränken sich manche Suchmaschinen bei der Stichwortextraktion auf bestimmte Teile des Dokuments, etwa Kapitelüberschriften oder die Einleitung. Wir werden uns mit dem Problem des Veraltens der Metadatenbank weiter unten beschäftigen.

Weiter fortgeschrittene Verfahren erlauben die automatische Erkennung von verwandten Begriffen aus Textdokumenten. Ein bekanntes System ist ConceptSpace [CHOh94; Chen94], das Textdokumente aus einer eng umgrenzten Domäne analysiert. Aus der Vorkommenshäufigkeit eines Wortes wird auf die Wichtigkeit geschlossen. Aus der relativen Häufigkeit des gleichzeitigen Auftretens zweier Wörter in einem Dokument kann auf die semantische Nähe der Wörter geschlossen werden. Die Experimente mit ConceptSpace beschränkten sich zunächst auf wenige tausend Dokumente. Das Verfahren wurde besonders erfolgreich in elektronischen Konferenzräumen eingesetzt, in denen sich eine Gruppe auf eine gemeinsame Terminologie einigen muß. Inzwischen wurde das Verfahren durch Einsatz von Höchstleistungsrechnern für die Indexierung von wissenschaftlichen Artikeln in digitalen Bibliotheken eingesetzt [ScCh96].

Automatische Methoden sind heutzutage auf Dokumente beschränkt, die bereits als Text (vorzugsweise im HTML oder ASCII-Format) vorliegen. In bestimmten Fällen kann eine Textfassung rücktransformiert werden, so etwa aus Postscript-Dokumenten von technischen Berichten. Dies wird in der Suchmaschine WAIKATO (www.cs.waikato.ac.nz/~nzdl) angeboten.

## Die Auskunft bei Suchmaschinen

Der eigentliche Zweck der Suchmaschinen ist natürlich die Suche nach Dokumenten. Idealerweise sollte die Anfrage natürlichsprachlich sein („Welche Dokumente enthalten Vorschläge zur Lösung der momentanen Weltwirtschaftskrise") und die Antwort sollte alle relevanten Dokumente enthalten. Von diesem Ideal ist die Praxis allerdings noch weit entfernt. Typischerweise wird nur die Suche mit Stichwörtern unterstützt. Stichworte können logisch (AND, OR, NOT) sowie bezüglich ihres relativen Vorkommens (NEAR) kombiniert werden. Die Antwort wird allein durch Rückgriff auf die Informationen der Metadatenbank berechnet. Merkt sich ein Suchprogramm z.B. nur bibliographische Informationen wie den Dokumenttitel, so werden auch nur solche Lösungen zurückgegeben, welche die Stichworte im Titel enthalten.

Da über die Struktur der Dokumente und den Kontext der in ihnen vorkommenden Stichwörter im allgemeinen wenig bekannt ist, reicht für viele Recherchen eine solch einfache Stichwortsuche nicht aus. Eine logische Verknüpfung von Stichwortvorkommen ist nur eine schwache Repräsentation des Rechercheziels. Es mag etwa sein, daß ein Dokument zwei von drei der gesuchten Stichwörter enthält, jedoch statt des dritten ein Synonym verwendet. Abhilfe schafft hier die Berechnung einer Rangzahl für die Antwort-

liste. Der grundsätzliche Ablauf der Antwortbestimmung ist nachfolgend wiedergegeben[3]:

1. Erfrage die Suchbegriffe und deren Verknüpfung vom Benutzer
2. Bestimme für jedes in der Metadatenbank beschriebene Dokument die Rangzahl als Maß für die Übereinstimmung mit den Suchbegriffen. Je größer die Übereinstimmung, desto größer die Rangzahl.
3. Ordne die Dokumente absteigend nach ihrer Rangzahl und gebe die Antwortliste aus.

Die Wahl der Funktion in Schritt 2 ist nicht trivial. Das System INQUERY [CCH92] setzt hierzu ein sogenanntes Bayes'sches Netz ein: die Gewichte von Stichwortvorkommen für ein Dokument werden in der Metadatenbank eingetragen. Die Rangzahl wird dann gemäß der logischen Verknüpfung aggregiert. Die Güte des Verfahrens mißt sich an der Rückmeldung (Anzahl der Dokumente in der Antwortliste) und der Präzision (Relevanz der Dokumente in der Antwortliste). Die Gewichte der Stichwortvorkommen leitet sich aus ihrer Häufigkeit ab. Qian und Frei [QF96] nutzen einen Ähnlichkeitsthesaurus, um die Treffsicherheit einer Anfrage zu erhöhen. Der Thesaurus wird wie in [Chen94] automatisch aus einer Dokumentmenge aufgebaut, die einer bestimmten Domäne entstammt.

Idealerweise sollte die Reihenfolge in der Antwort auch der tatsächlichen Wichtigkeit des Dokuments für den Benutzer entsprechen. Die Treffsicherheit kann durch die Erweiterung der Anfrage verbessert werden: der Benutzer soll sein Rechercheziel genau spezifizieren. Allerdings ist diese Strategie allzu naiv: Wüßte der Benutzer genau, nach welcher Information er sucht, so würde er kaum eine Suchmaschine benötigen. Alternativ kann die Suchmaschine ein Benutzerprofil erstellen. Die Informationen aus dem Benutzerprofil werden zur Steuerung der Suche in der Metadatenbank mit herangezogen. Wegen der Kopplung an die Begriffswelt des Benutzers stellen Synonyme, Homonyme und Abkürzungen eine Herausforderung dar. Man stelle sich etwa vor, daß der Benutzer Dokumente mit dem Stichwort 'Expertensystem' sucht und ein Dokument statt dessen durchgängig die Abkürzung XPS verwendet. Abhilfe kann ein Synonymlexikon schaffen. Es kann transparent oder auf Intervention des Benutzers eingeschaltet werden. Man beachte, daß solche Synonymlexika prinzipiell fachspezifisch, im Extremfall sogar personenbezogen sind.

In [Mase97] wird ein Überblick über datenbankbasierte Anfragesprachen für das World Wide Web gegeben. Idee ist hierbei, einen Dialekt der Anfragesprache SQL anstatt der bisher vorherrschenden Stichwortsuchen zu verwenden. Auf diese Weise können präzise Selektionsbedingungen formuliert werden. Bisher konnten sich solche Anfragesprachen aber nicht durchsetzen, da anders als bei Datenbanken fast nichts über die Struktur der in der Metadatenbank indizierten Objekte bekannt ist. Das Problem ist sogar noch schwieriger als die Abfrage heterogener Datenbanken, für das noch keine zufriedenstellende Lösung gefunden wurde. Ursache sind unterschiedliche Interpretationen für Metadaten in den Teildatenbanken bzw. den beteiligten WWW-Servern. Man nehme etwa an, zwei Teildatenbanken offerieren Angaben über den Schadstoffgehalt in Lebensmitteln. In der einen Teildatenbank kann sich die Angabe auf Mittelwerte beziehen, in der anderen auf die Maximalwerte. In einem solchem Fall ist eine Anfrage nach allen Lebensmitteln, deren Schadstoffgehalt über einem Grenzwert liegt, nicht fehlerfrei auf die Teildatenbanken abbildbar.

Die Ausdrucksfähigkeit der Anfragesprache hängt stark von der Realisierung der Metadatenbank ab. Zur Zeit gibt es zwei konkurrierende Ansätze: zentral und dezentral. Zentrale Realisierungen wie ALTAVISTA speichern die gesamte Metadatenbank auf dem physischen Plattenspeicher eines leistungsstarken Rechners ('Supercomputer'). Somit kann die Metadatenbank komplexen Anfragen unterworfen werden. Nachteil dabei ist die Kapazitätsbeschränkung: wenn Engpässe auftreten, etwa in der Antwortzeit, so muß der Zentralrechner ausgebaut werden. Die Skalierbarkeit (Anzahl der Prozessoren, Hauptspeicherkapazität etc.) ist jedoch durch die Architektur des Zentralrechners beschränkt. Dezentrale Realisierungen zerteilen die Metadatenbank in Fragmente die auf eine Vielzahl relativ leistungsschwacher Workstations aufgeteilt werden. Eine Suchanfrage wird dann nach einem bestimmten Algorithmus in Teilanfragen für die Workstations dekomponiert. Die Antworten auf die Teilanfragen werden gemischt und als Gesamtantwort zurückgegeben. Vorteil ist die nahezu beliebige Skalierbarkeit, was jedoch mit deutlichen Beschränkungen in der Ausdrucksstärke der Anfragesprache bezahlt wird. Beispiel einer dezentral realisierten Suchmaschine ist HOTBOT (www.hotbot.com).

## Verfügbare Suchhilfen

Ein Überblick über verfügbare Suchhilfen ist in dem Parallelbeitrag [Teut97] im vorangehenden Heft enthalten. Es gibt inzwischen weit über hundert frei zugreifbare Suchhilfen. Für eine ausführliche Liste sei der Leser auf die Datenbank in [Kost97] verwiesen. Die erfolgreichsten Suchmaschinen [Tenn96] sind mittlerweile aus dem Universitätsbereich in die kommerzielle Nutzung gewandert. Vorsicht ist jedoch bei Angaben über die Indexgröße geboten. Bei ALTAVISTA werden etwa 31 Millionen Dokumente in der Metadatenbank indiziert. Diese Obergrenze im Jahre 1997 erreicht und seitdem verharrt die Datenbank von ALTAVISTA auf dieser Größe. Die Suchmaschine HOTBOT (www.hotbot.com) gibt für ihre Datenbank die Größenordnung von 53 Millionen Einträgen bekannt. Bei LYCOS (www.lycos.com) fanden sich bis vor kurzem Angaben, daß dort 50 Millionen Dokumente indiziert seien. Stichproben ergaben jedoch für LYCOS eine Größe von nur 17 Millionen indizierten Dokumenten. Auch die Angaben über die im Netz befindlichen WWW-Server schwanken stark. ALTAVISTA gibt im Juli eine Zahl von ca. 600.000 an. Exakte Zahlen sind nur schwer zu ermitteln. Ein qualitativer Vergleich der Suchhilfen ist in [SiLi96] enthalten.

Der METACRAWLER (www.metacrawler.com) ist eine *föderierte Suchmaschine* [BDHM95], die keine eigene Metadatenbank unterhält, sondern eine Stichwortanfrage an etwa 7 global indizierende Suchmaschinen (wie ALTAVISTA) weiterleitet und deren Antworten zu einer Gesamtantwort kombiniert. Maßgeblich für die Reihenfolge der Antworten ist das Ergebnis der Teilanfrage an den WEBCRAWLER (www.webcrawler.com), einer sehr präzisen, jedoch unvollständig arbeitenden Suchmaschine.

# 4 Probleme existierender Suchhilfen

Der Grundfunktion einer Suchmaschine ist die präzise Auflistung der unter eine Suchanfrage fallenden Dokumente. In diesem Kapitel wollen wir uns fragen, welche Gegebenheiten diese Grundfunktion gefährden können – und es sind derer viele.

## Veralten der Metadatenbank

Betrachten wir die Metadatenbank der Suchmaschine als Abbild der Wirklichkeit, so kann dieses Abbild schlicht veralten. Die Auskunftsfunktion gibt dann fehlerhafte Antworten bei Suchanfragen. Der Qualität der Suchmaschine nimmt ab. Eine mögliche Lösung des Problems ist, daß die Suchmaschine bereits bekannte Dokumente periodisch wieder aufsucht und bei Änderung die Metadatenbank aktualisiert. Natürlich erlaubt die schiere Größe des WWW nur relativ wenige dieser Auffrischungsbesuche pro Monat. Folgende Überschlagsrechnung macht dies deutlich:

Konservativ geschätzt gibt es 30 Millionen Dokumente im Suchraum. Der Zugriff über das weltweite Internet dauert bei heutiger Infrastruktur im Durchschnitt mindestens 1 Sekunde. Also werden für eine Totalauffrischung über 200 Tage benötigt[4]! Innerhalb dieser 200 Tage können bereits aufgesuchte Dokumente erneut geändert werden, so daß trotz der Totalauffrischung nie ein vollkommen korrektes Abbild entsteht.

Die Unvorhersehbarkeit der Änderungen in Verbindung mit den Transferkosten führt dazu, daß jede reale Metadatenbank in gewissem Umfang inkorrekt ist. Es gibt einige Heuristiken, um die Fehlerquote gering zu halten.

1. Speichere nur wenig Metadaten zu einem Dokument. Dadurch verringert sich die Wahrscheinlichkeit, daß eine Änderung am Originaldokument dessen Metadaten verändert. Im Extremfall reduzieren sich die Metadaten auf die Dokumentadresse. Natürlich verringert dies auch den Nutzen der Metadatenbank.
2. Beachte das Verfallsdatum der Dokumente. Das Verfallsdatum wird dem Dokument vom Autor beigefügt. Nur verfallene Dokumente werden aufgefrischt. Leider benutzen nur wenige Autoren diese Funktion.
3. Leite die Änderungswahrscheinlichkeit aus der Metadatenbank ab. Sind zwei Dokumente per Verweis verbunden, so stehen ihre Änderungswahrscheinlichkeiten in Beziehung. Stichproben auf früherer Aufrischungszugriffe werden zur Schätzung der Änderungswahrscheinlichkeit herangezogen.
4. Beschränke den Suchraum. Spezialisierte Suchmaschinen können in kürzeren Zeitabständen eine Auffrischung vornehmen.
5. Verwalte die Metadaten in einer verteilten Datenbank, wobei die Teildatenbanken dezentral ihre Inhalte auffrischen (siehe HARVEST [BDHM95]).

## Duplikate und Phantomkopien

*Duplikate* sind physische Kopien von Dokumenten. Die Suchprogramme sind im allgemeinen nicht in der Lage, Kopien vom Original zu unterscheiden. Bei einer Anfrage werden daher sowohl Kopien als auch Original als Antwort zurückgegeben, sofern die Suchkriterien zutreffen und sowohl Kopie als auch Original im WWW publiziert sind. Besonders bedenklich sind veraltete Duplikate, die ohne Wissen der Autorin oder des Autors angelegt wurden. Über die Suchmaschinen werden sie für eine allgemeine Leserschaft zugänglich.

*Phantomkopien* entstehen, wenn ein und dasselbe physische Dokument durch mehrere URL-Adressen zugreifbar ist. Dies geschieht beispielsweise durch Definition eines Alias-Namens. Ohne Zusatzinformation haben die Suchmaschinen keine Möglichkeit, solche Mehrfachnennungen zu unterdrücken. Auf den ersten Blick erscheint dies als eher kleines Problem. Man möge sich aber vergegenwärtigen, daß Nutzer der Suchmaschinen die Antworten auf eine Anfrage in eigene Dokumente einbauen. Wird nun eine Phantomkopie durch Löschung des Alias-Namens unzugreifbar, so werden solche Dokumente unnötigerweise inkonsistent.

Prinzipiell sind Duplikate nicht zu verhindern. In vielen Fällen ist eine Duplizierung sogar von der Autorin bzw. dem Autor gewünscht, um den Nutzern einen leichteren Zugriff zu ermöglichen. Ein Beispiel sind die frei verfügbaren Softwarepakete, die über Computernetze verbreitet werden. Ansonsten kann man nur an die Nutzer des WWW appellieren, auf nichtautorisierte Duplikate zu verzichten.

## Autorisierung

Vom Standpunkt des Autors eines Dokumentes ist das WWW ein Medium zur Verbreitung von Information. Wie bei herkömmlichen Publikationen ist eine Zuordnung der verbreiteten Dokumente zu Autoren grundsätzlich wünschenswert und ein wichtiges Metadatum für Suchmaschinen. Mangelhafte Autorisierung äußert sich auf zweifache Weise.

1. Das WWW-Dokument eines Autors ist ursprünglich nur für eine begrenzte Leserschaft gedacht, wird aber tatsächlich über diesen Kreis hinaus gelesen. Abhilfe kann durch Paßwortschutz, Verschlüsselung etc. geschaffen werden. Manche Suchmaschinen können auch durch eine Hinweisdatei [Kost 96b] von der Erfassung bestimmter Dokumente ausgeschlossen werden.
2. Das Dokument ist zwar für die breite Leserschaft gedacht, aber die Autorenschaft ist bewußt verschleiert oder verfälscht. Digitale Unterschriften bieten eine Möglichkeit, diesen Mißbrauch zu erschweren.

## Mehrsprachigkeit

Bisher sind wir implizit davon ausgegangen, daß alle WWW-Dokumente in derselben Sprache erstellt wurden. Dies ist gerade im europäischen Kontext eine fragwürdige Annahme. Aus historischen Gründen herrscht zwar die englische Sprache vor, aber mit der zunehmenden weltweiten Nutzung des Internets nehmen die fremdsprachlichen Dokumente an Bedeutung zu. Für die Suchmaschinen stellt sich nun das Problem, wie man Informationen über Sprachbarrieren hinweg verknüpfen kann. Zentral aus Sicht der Suchmaschinen ist hierbei die Metadatenbank, da die in ihr enthaltenen Schlüsselwörter zunächst nur in der Sprache des jeweils beschriebenen Dokuments vorliegen.

Das Problem der Mehrsprachigkeit ist nicht neu. In Ländern wie Belgien liegen viele staatliche Dokumente in allen Landessprachen (hier: Französisch, Niederländisch und Deutsch) vor. In der Europäischen Union sind Gesetze jeweils in alle Sprachen der Mitgliedstaaten zu übersetzen. Aus Sicht der Suchmaschinen gibt

es drei prinzipielle Möglichkeiten des Umgangs mit der Mehrsprachigkeit:

1. Man einigt sich auf eine Referenzsprache (meist Englisch). Dokumente in anderen Sprachen werden ignoriert bzw. nicht als solche behandelt. In diesem Fall ist an der Metadatenbank nichts zu tun. Allerdings ist das Problem nicht gelöst, sondern nur umgangen.
2. Man erwartet, daß Dokumente auch in ihren Übersetzungen im WWW vorliegen. In diesem Fall baut die Suchmaschine auch einen Index für die übersetzten Dokumente auf. Wiederum ist aus Sicht der Suchmaschine nicht viel zu tun. Die Verantwortung (und Arbeit) für die Übersetzung liegt bei den Autoren.
3. Die Suchmaschine übersetzt ihre Metadaten (Stichwörter, Konzeptnamen) in die jeweiligen Sprachen. Eine einfache Vorgehensweise ist die Übersetzung mittels eines Wörterbuchs. Um präzise Übersetzungen zu gewährleisten, reicht dies kaum. Es muß der Kontext eines Wortes im Dokument bei der Übersetzung berücksichtigt werden.

Die großen amerikanischen Suchhilfen bieten seit einigen Monaten nationale Ableger ihrer Systeme an, etwa das deutsche YAHOO (www.yahoo.de). Eine Spezialisierung auf die jeweilige Landessprache erhöht die Antwortqualität für solche Nutzer, die Dokumente in dieser Landessprache suchen. YAHOO hat als besondere Dienstleistung eine Verzeigerung der Kategorien zwischen den nationalen Ablegern. So kann ein Benutzer von der Kategorie *Unterhaltung/Kino* von YAHOO Deutschland direkt in die Kategorie *Entertainment/Movies* des amerikanischen Ablegers wechseln. Die Domänenbegriffe der Hierarchie liegen also übersetzt vor, nicht jedoch die Inhalte der Dokumente.

## Multimedialität und strukturierte Daten

Trotz der bereits erwähnten Ansätze für Bilddokumente in Lycos sind die Suchmaschinen bisher nur für Textdokumente nutzbar. Wenn das Internet sich weiter in Richtung interaktives Fernsehen entwickelt (siehe [GVU95] als Beleg), so werden nicht-textuelle Dokumente wie Bilder, Videosequenzen und Töne enorm an Bedeutung zunehmen. Eine Suchhilfe, die Benutzer effizient zu solche Informationsquellen leitet, ist das Äquivalent zu einer guten Fernsehzeitschrift. Die Aufgabe der Klassifikation dieser Dokumente ist um Größenordnungen schwieriger als bei Texten. Digitale Repräsentation für Videosequenzen (Stichwort MPEG) erlauben das Verstecken von Metadaten wie Szenentitel in den Bilddaten. Solche Zusatzinformation des Dokumentautors ist das Analogon zu den META-Feldern in HTML (vergleiche Kapitel 3). Formatierte Daten, etwa aus Datenbanken, werden unseres Wissens nicht im nennenswerten Maße durch Suchmaschinen erfaßt.

| Suchhilfen \ Qualität | Zugreifbarkeit | Interpretierbarkeit | Nützlichkeit | Glaubwürdigkeit |
|---|---|---|---|---|
| HTML/HTTP-Navigation | schlecht | schlecht | schlecht | gut |
| Hotlists Themenkataloge | gut | gut | mittel | mittel |
| Suchmaschinen im engeren Sinne | gut | schlecht | mittel | schlecht |
| Informations-Broker | gut(?) | gut(?) | gut(?) | gut(?) |

**Bild 4** Informationsqualität der Suchhilfen

## Versuch einer Gesamtbewertung

Interpretierbarkeit, Relevanz und Glaubwürdigkeit der gefundenen Informationen werden in offenen Systemen weitgehend von der Güte des Auswahlverfahrens bestimmt. In kleinen Gruppen oder themenbezogenen Foren haben hier strukturierte Hotlists mit Rückverfolgbarkeit zu den jeweiligen Gutachtern Vorteile, während alle domänenunabhängigen Verfahren (referenz- oder wertebasiert) hier große Probleme haben: der Recall ist oft sehr gut, aber die Präzision leidet. Im Überwinden dieses Konfliktes zwischen Verfügbarkeit und den anderen Kriterien liegt die zentrale Forschungsaufgabe bei der Entwicklung von Suchhilfen der nächsten Generation.

Im Licht der Bewertungskriterien aus Kapitel 2 ergibt sich für die vier Klassen von Suchhilfen ein gemischtes Bild (vgl. Bild 4). Die reine *Navigation* mit einem Browser erhält erwartungsgemäß die schlechtesten Noten. Die Suchhilfe hier ist das Gedächtnis des Benutzers inklusive der im momentan dargestellten Dokument enthaltenen Verweise. Da allein durch manuelle Navigation ein Überblick über passende Dokumente gewonnen wird, ist der Zugriff auf die Antwort auf eine (hier implizite) Suchanfrage sehr schlecht. Auch für die Interpretation wird keine Hilfe geboten. In der Tat ist der Benutzer mit der nahezu beliebigen Heterogenität der Dokumente konfrontiert. Eine Zuordnung zu Konzepten der Suchanfrage findet mangels expliziter Repräsentation der Suchanfrage nicht statt. Ob ein per Navigation besuchtes Dokument relevant für die Suchanfrage ist und ob man es rechtzeitig findet, ist weitgehend dem Zufall überlassen. Also ist auch die Nützlichkeit negativ zu bewerten. Allein die Glaubwürdigkeit verdient gute Noten, weil mit Ausnahme der in Kapitel 2 beschriebenen Proxy-Server keine Filterung und Verfälschung der Antwort stattfindet.

Im Gegensatz dazu fällt das Urteil für *Hotlists* und *Themenkataloge* eher positiv aus. Die Information über Dokumente ist in einen Index klassifiziert, der nach Ober- und Unterbegriffen aufgeteilt ist. Die Zugreifbarkeit profitiert von dem vor der Suchanfrage erstellten Index. Die Interpretierbarkeit hängt dieser Art der Suchhilfen von der Wahl der Oberbegriffe des Index ab. Sind sie gut, so ist auch die Interpretierbarkeit des Index gut. Die Nützlichkeit und Glaubwürdigkeit ist mit Abstrichen zu versehen. Die etwa in YAHOO übliche Praxis, daß Einträge unter Oberbegriffe von den Autoren der Dokumente vorgenommen werden, führt zu besonders bei kommerziellen Angeboten zu der Neigung, Einträge unter vielbesuchte anstatt unter die relevanten Oberbegriffe einzuordnen.

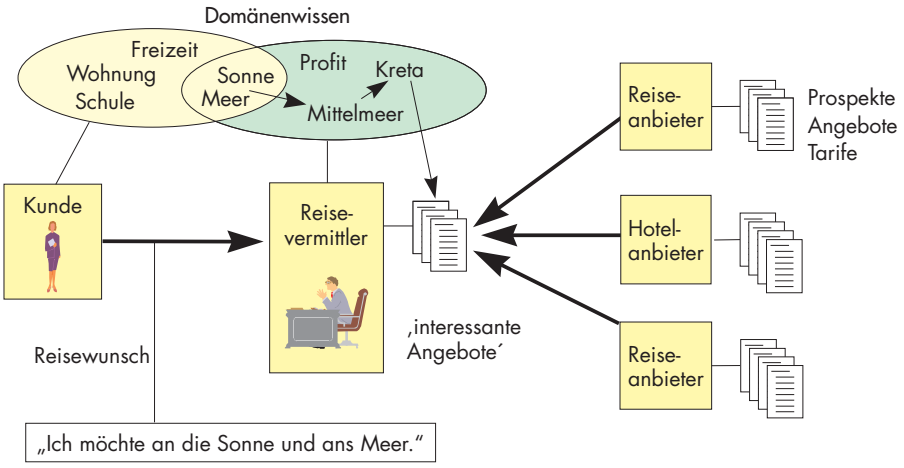


**Bild 5** Informationsvermittlung mit Domänenwissen

Die Bewertung der dritten Klasse von Suchhilfen, den *Suchmaschinen* im engen Sinn, scheint im Widerspruch zu ihrem Erfolg zu stehen. Suchmaschinen wie ALTAVISTA leiden prinzipiell unter der Diversität der indizierten Dokumente. Wenn die zu indizierende Informationsmenge weiter zunimmt, so müssen Interpretierbarkeit, Nützlichkeit und Glaubwürdigkeit der Antworten zwangsläufig leiden. Ihnen fehlt das Domänenwissen, das die globalen Kataloge auszeichnet. Berghel [Berg97] prophezeit sogar den Fehlschlag dieser Art von Suchhilfen.

Als letzte Kategorie werden die *Informations-Broker* genannt. Ihr Ziel ist Kombination der automatischen Indizierung, also der Extraktion von Information aus den Originaldokumenten, mit zusätzlichem Wissen. Diese Vision zukünftiger Suchhilfen ist Gegenstand den nachfolgenden Kapitels.

## 5 Informations-Broker als intelligente Vermittler

Die momentan verfügbaren Suchmaschinen sind geeignet, Textdokumente anhand von Stichwörtern aufzufinden. Ihre genannten Schwächen rühren einerseits von Fehlern bei der Indizierung der Dokumente und anderseits von der ungenauen Spezifikation der Suchanfrage her. Offenbar reicht das in einem Dokument enthaltene Wissen nicht in allen Fällen aus, um es korrekt zu verschlagworten. Zudem ist zu wenig über den Hintergrund des Anfragers bekannt, bzw. dieses Wissen wird nicht berücksichtigt. In diesem Kapitel beschäftigen wir uns daher mit Verfahren, dieses Informationsdefizit zu beheben. Wir schlagen die Hinzunahme von Domänenwissen vor. Ergebnis ist der bereits angesprochene Informations-Broker.

Um diesen Ansatz zu charakterisieren, betrachten wir als Analogie die Situation eines Reisevermittlers (Bild 5). Der signifikante Unterschied ist die Rolle des Domänenwissens (über Reisen) bei der Anfrageformulierung (hier: Reisewunsch) und der Bild auf Reiseangebote durch den Vermittler. Reisevermittler und Kunde kommunizieren über den überlappenden Anteil ihres Domänenwissens. Ist eine Anfrage einmal formuliert, so klassifiziert der Vermittler die darin vorkommenden Begriffe in seine Terminologie.

Beispiele wie die Gelben Seiten von Telefongesellschaften und Themenkataloge wie YAHOO zeigen, daß hierarchisch aufgebaute Terminologien für die Klassifikation besonders geeignet sind. Die Rolle der Terminologie für die Vermittlung ist ihre Verknüpfung mit den Originaldokumenten (hier: Prospekte, Angebote, Tarife etc.). Auch diese werden vom Reisevermittler in die Hierarchie der Begriffe eingeordnet. So sind unter Kreta unter anderem alle Hotels auf Kreta zu finden. Interessante Angebote sind also solche, die sich in die Terminologie des Reisevermittlers klassifizieren lassen. Auf diese Weise wird der Fokus auf die Domäne gesetzt.

Aus dem Beispiel lassen sich die Aufgaben und Teilprobleme eines Informations-Brokers als intelligentem Informationsvermittler ableiten. Die erste Aufgabe ist der *Aufbau des Domänenwissens*. Für diesen nicht-trivialen Prozeß wurden in gewissen Domänen bereits erhebliche Vorarbeiten geleistet. In der Medizin existieren umfangreiche Terminologien, die neben medizin-statistischen Gründen auch zur Informationssuche genutzt werden. Genannt sei hier das System MeSH-System (Medical Subject Headings, www.dimdi.de/klassi/mesh/basmesh.htm), das zur Suche nach medizinischen Publikationen genutzt wird. Ferner sei das UMLS-System [LHM93] genannt, das ein semantisches Netz medizinischer Grundbegriffe bereitstellt, über das auf Spezialterminologien wie MeSH zugegriffen werden kann. Papazoglou und Milliner [PaMi96] nutzen ganz ähnlich eine Begriffshierarchie sogenannter generischer Konzepte zur Anfrageübersetzung für heterogene Informationssysteme. Weitere Beispiele sind in [JePa96] zu finden.

Für die *Einordnung von Informationsangeboten* in das explizit dargestellte Domänenwissen können manuelle und automatisierte Verfahren unterschieden werden. Bei den manuellen Verfahren nimmt ein Gutachter (oder Autor) die Zuordnung des Dokuments in die vorgegebene Terminologie vor. Die Arbeit von Eherer [Eher95] zeigt, daß ein semantisches Begriffsnetz bereits zur Erstellung von Hypertexten Verwendung finden kann. Wird dies mit dem semantischen Netz der Domäne verknüpft, so ist das Dokument klassifiziert. Bei automatisierten Verfahren werden Techniken der Künstlichen Intelligenz erprobt. Ein vielversprechender Ansatz auf Basis sogenannter selbstorganisierender Merkmalskarten (self organizing maps) wurde inzwischen von der Domäne ‚Informatik-Publikationen' [ScCh96] auf das Internet übertragen [CSO96]. Das System erstellt nach Durchsicht der Originaldokumente gemäß ihrer Ähnlichkeit eine Hierarchie von ‚Schwerpunkten'. Diese Schwerpunkte werden dann von einem Experten mit lesbaren Namen versehen. Experimente haben Vorteile dieses Ansatzes gegenüber manuellen Indexen erwiesen. Man beachte, daß die Begriffshierarchie eine Funktion der als Eingabe benutzen Dokumente ist. Eine eingeschränkte Dokumentenmenge liefert mithin eine fokussierte Begriffshierarchie als Domänenwissen.

Die wohl weitestgehende Realisierung eines Informations-Brokers findet sich im System Information Manifold [Kirk96]. Es

kombiniert die Repräsentation des Domänenwissens in einer Begriffstaxonomie mit der Nutzung der klassischen Suchmaschinen als Rohdatenlieferanten (analog zu METACRAWLER, vgl. Kapitel 3). Die Begriffstaxonomie kodiert die Suchpräferenzen eines Benutzers oder einer Benutzergruppe. Sie wird bei der Filterung der Ergebnisse der Rohdatenlieferanten herangezogen. Anfrageergebnisse werden graphisch in die betroffenen Teile der Taxonomie integriert.

## 6 Zusammenfassung

Zusammenfassend bleibt festzuhalten: Der globale Informationsmarkt steht am Anfang seiner Entwicklung und ist zur Zeit weitestgehend unorganisiert. Heutige Suchmaschinen und Themenkataloge sind nur ein erster Schritt, Ordnung in das Informationschaos zu bringen. Die absichtsvoll unkontrollierte Typ- und Sprachvielfalt der Dokumente schafft Teilmärkte, in denen sich spezifische Regeln für die Dokumentbeschreibung und -suche herausbilden müssen. Wir sehen Informations-Broker als die nächste Generation von Suchhilfen für diese Teilmärkte. Sie nutzen Domänenwissen, um den Suchraum einzuschränken.

**Danksagung:** Die Autoren bedanken sich bei Stefanie Kethers und Michael Gebhardt für interessante Hinweise und nützliche Kommentare.

## Anmerkungen

[1] Dies heißt: basierend auf dem Inhalt der Dokument anstatt ihrer URL-Referenz.

[2] Dies führt bei älteren Suchmaschinen zu dem Effekt, daß nicht nach Textstellen wie ‚as you like it' gesucht werden kann. Diese Schwäche wurde aber inzwischen erkannt und behoben.

[3] Zur Beschleunigung der Suche werden invertierte Indexe eingesetzt. Meist wird nicht die ganze Antwortliste auf einmal berechnet, sondern nur seitenweise auf Anforderung durch den Benutzer.

[4] Durch parallelen Einsatz vieler Metadatensammler kann diese Zeit beträchtlich verkürzt werden. Die dadurch entstehende Netzbelastung geht allerdings auf Kosten der Verfügbarkeit der WWW-Server.

## Literatur

[BCLN94] *Berners-Lee, T. et al.:* The World-Wide Web. In: Communications of the ACM 37 (1994) 8, S. 76-82.

[BDHM95] *Bowman, C.M. et al.:* Harvest: A Scalable, Custumizable Discovery and Access System. Technischer Bericht CU-CS-732-94, University of Colorado-Boulder, März 1995.

[BDMS94] *Bowman, C.M. et al.:* Scalable Internet Resource Discovery: Research Problems and Approaches. In: Communications of the ACM 37 (1994) 8, S. 98-107.

[Berg97] *Berghel, H.:* Cyberspace 2000: Dealing with Information Overload. In: Communications of the ACM 40 (1997) 2, S. 19-24.

[Bern95] *Berners-Lee, T.:* Uniform Resource Locators. Http://www.w3.org/hypertext/WWW/Addressing/URL/URL_TOC.html, 1995.

[CCH92] *Callan, J.P.; Croft, W.B.; Harding, S.M.:* The INQUERY Retrieval System. In: Proceedings of the 3rd International Conference on Database and Expert Systems Applications, 1992.

[Chen94] *Chen, H.:* Collaborative Systems: Solving the Vocabulary Problem. In: IEEE Computer 17 (1994) 5.

[CHOH94] *Chen, H. et al.:* Automatic Concept Classification of Text from Electronic Meetings. In: Communications of the ACM 37 (1994) 10.

[Conk87] *Conklin, J.:* Hypertext: An Introduction and Survey. In: IEEE Computer 20 (1987) 9, S. 17-41.

[CSO96] *Chen, H.; Schuffels, C.; Orwig, R.:* Internet Categorization and Search: A Self-Organizing Approach. In: J. Visual Communication and Image Representation 7 (1996) 1, S. 88-102.

[Date95] *Date, C.J.:* Introduction to Database Systems. 6. Aufl., Addison-Wesley, 1995.

[Eher95] *Eherer, S.:* Eine Software-Umgebung für die kooperative Erstellung von Hypertexten. Niemeyer, Tübingen 1995.

[Gray95] *Gray, M.:* Measuring the Growth of the Web. Http://www.netgen.com/info/growth.html, Juni 1995.

[GVU95] *GVU's WWW Surveying Team:* GVU's 4th WWW User Survey. Http://www.Cc.gatech.edu/gvu/user_surveys/survey-10-1995/, Oktober 1995.

[Hall97] *Hallström, M.:* Gruppenindexe. Persönliche Kommunikation mit dem Autor, SISU, Ellectrum Kista, Stockholm, Schweden, Februar 1997.

[JePa96] *Jeusfeld, M.A. ; Papazoglou, M.:* Information Brokering: Design, Search and Transformation. Aachener Informatik-Berichte, Bd. 96-18, 1996.

[Kirk96] *Kirk, T.:* Information Manifold: Knowledge Based Access to Information in the WWW. In: Proceedings Workshop Artificial Intelligence-based Tools to Help W3 Users, 5th International World Wide Web Conference, Paris, Frankreich, Mai 1996.

[Kost96a] *Koster, M.:* World Wide Web Robots, Wanderes, and Spiders. Http://info.Webcrawler.com/mak/projects/robots/robots.html, Februar 1996.

[Kost96b] *Koster, M.:* A Standard for Robot Exclusion. Http://info.webcrawler.com/mak/projects/robots/norobots.html, 1996-03-15.

[Kost97] *Koster, M.:* Database of Web Robots: Overview. Http://info.webcrawler.com/mak/projects/robots/active/html/, WebCrawler Corp, USA, März 1997.

[LHM93] *Lindberg, D.A.B.; Humphreys, B.L.; McCray, A.T.:* The Unified Medical Language System. In: Yearbook of Medical Informatics 1993, S. 41-54.

[Mase97] *Masermann, U.:* Suchdienste im World-Wide Web – Anforderungen und Perspektiven. In: *Jeusfeld, M. (Hrsg.):* Informationsserver für das Internet – Anforderungen, Konzepte, Methoden. Proceedings EMISA-Fachgruppentreffen 1996 [EMISA-FORUM (1997) 1], Gesellschaft für Informatik, 1997.

[PaMi96] *Papazoglou, M.P.; Milliner, S.:* Pro-Active Information Elicitation in Wide-Area Information Networks. In: Proc. Intl. Symposium on Cooperative Database Systems for Advanced Applications, Kyoto, Japan, Dezember 1996.

[QF96] *Qian, Y.; Frei, H.-P.:* Improving the Retrieval Effectiveness by a Similarity Thesaurus. Technischer Bericht 225, ETH Zürich, Institut für Informatik, 1996.

[SaBu88] *Salton, G.; Buckley, C.:* Term weighting approaches in automatic text retrieval. In: Information Processing Management 24 (1988) 5, S. 513-523.

[ScCh96] *Schatz, B.; Chen, H.:* Building Large-Scale Digital Libraries. In: IEEE Computer (1996) Mai, S. 22-26.

[Sher95] *Sher, D.:* Lycos catalogs more than 10 million web sites. Http://www.lycos.com/press/press-release.html, Boston, USA, 1995-10-30.

[SiLi96] *Singh, A.; Lidsky, D.:* All-out Search. In: PC Magazine 15 (1996) 21, http://www8.zdnet.com/pcmag/iu/srchsite/_open.htm, USA, 1996.

[Tenn96] *Tennant, R.:* Effective Web Searching. 5th Intl. World Wide Web Conf., Paris, Mai 1996, Tutorial Notes, O'Reilly & Ass. Inc., USA, 1996.

[Teut97] *Teuteberg, F.:* Effektives Suchen im World Wide Web:Suchdienste und Suchmethoden. In: Wirtschaftsinformatik 39 (1997) 4.

[WaWa96] *Wand, Y.; Wang, R.Y.:* Anchoring Data Quality Dimensions Ontological Foundations. In: Communications of the ACM 39 (1996) 11.