Fachbeiträge

Datenbanktechnische Betrachtungen zu Text- und Bibliographiedatenrecherchesystemen unter WWW

Prof. Dr. Gregor Büchel Fachhochschule Köln Fachbereich Nachrichtentechnik Betzdorfer Str.2 50679 Köln Tel.: 0221/8275-2488 (-2493, -2431) *EMAIL: buechel@fh-koeln.de*

Einleitung

Die nachfolgenden Betrachtungen sind entstanden aus Entwicklungsvorhaben, die momentan am Fachbereich Nachrichtentechnik größtenteils durch Diplomarbeiten geplant und realisiert werden, die die Frage der "Anbindung" bestehender oder neu zu entwickelnder Informationssysteme an das World Wide Web (=: WWW) behandeln. Das WWW ist innerhalb des INTERNET ein Dienst, der durch wachsende Benutzerakzeptanz und Zuwachszahlen an Teilnehmern (Informationsanbieter und Informationsnehmer) gekennzeichnet ist. Dieses geht insbesondere für die Fachhochschule Köln aus der Projektzusammenarbeit mit externen Partnern und aus statistischen Aufzeichnungen auf dem für die Domäne *jh-koeln.de* zentralen WWW-Server hervor.

Ein Problem, das sowohl auf Seiten der Informationsanbieter (Server) und erst recht auf Seiten der Informationsnehmer (Client) besteht, ist, daß die allermeisten angebotenen und empfangenen Informationen in herkömmlichen Dateisystemen, die das Betriebssystem des jeweiligen WWW-Server- bzw. -Client-Rechner gerade bieten, abgelegt werden und daß dadurch komfortablere Methoden der Datenhaltung, wie z.B. eine Datenstrukturierung durch ein Data-Dictionary, eine normierte und mächtige Zugriffssprache wie ANSI-SQL und die Möglichkeit des Datenschutzes auf Tabellen- bzw. Feldebene, zunächst einmal nicht gegeben sind, wie es der Standard für relationale Datenbanksysteme (RDBS) auf Host-Systemen oder in lokalen Netzen garantiert. Die Entwicklung von WWW-Anwendungen stellt daher alte Fragen der Datenbanktechnik, aber mit breiterer Wirksamkeit, neu. In Betrachtung, daß das WWW für die Entwicklung von Infor-

mationssystemen auf verteilten Datenbanksystemen interessante Testfalle mit großen Teilnehmerzahlen bereitstellt, sollen Einzelvorschläge für die Verbesserung von Text- und Bibliographiedatenrecherchesysteme bei Neuentwicklungen im WWW diskutiert werden, die, obwohl sie als "theoretische" Vorschläge teilweise alt sind, teilweise in vorliegenden nicht alten Systemen nicht berücksichtigt wurden. Diese Vorschläge werfen im Aspekt auf ihre datenbanktechnische Umsetzung besondere "praktische" Fragen auf, wie sie z.B. in Hinsicht auf die Definition von Stoppwortlisten, auf die Einbeziehung von maschinenlesbarer Literatur, auf die Darstellung von Rechercheergebnissen und auf empirische Ansätze der Thesaurierung von Schlagwortbeständen bestehen.

1 Zum Problem des Durchgriffes von WWW-Anwendungen auf relationale Datenbanksysteme

Dieses Problem stellte sich bisher häufig, weil viele WWW-Server-Rechner nicht über ein Datenbanksystem verfügen. Selbst auf den WWW-Server-Rechnern, die über ein lokales Netz mit einen Datenbank-Server verbunden waren, stellten sich Probleme des Durchgriffes auf die vorhandenen Datenbanktabellen ein, weil viele Datenbankhersteller noch keine Werkzeuge für die Programmierung von Schnittstellen für die Verarbeitung von Suchanfragen, die an den WWW-Server abgesetzt wurden, zur Verfügung stellten. Inzwischen hat hier eine Trendwende eingesetzt und es gibt neben den großen kommerziellen relationalen Datenbankprodukten auch solche, die relevante Teilmengen des ANSI-SQL-Sprachinventars auf UNIX-Maschinen unterstützen und als Shareware im INTERNET erhältlich sind (wie z.B. das Mini SQL, das an der australischen Bond University entwickelt wurde) [Schmelzle/Gast 1996], was insbesondere für Studierende, die zu Hause über einen UNIX-fähigen PC verfügen, als WWW-Clienten arbeiten und im WWW recherchierte Daten strukturiert ablegen möchten, von Interesse sein könnte.

Komfortabel ist die Programmierung eines relationalen Datenbankzugriffes dann, wenn auf einer WWW-Seite im HTML-Kode (HTML:= Hypertext Markup Language) durch SQL-Befehle, die in SGML-Kommentaren (SGML:= Standard Generalized Markup Language) eingebettet werden, lesende, bzw. schreibende Datenbankabfragen abgesetzt werden können. Weiterhin ist es hilfreich, wenn ein solches Schnittstellenwerkzeug als Ergebnis einer SELECT-Anfrage eine Tabelle innerhalb einer HTML-Ergebnisseite generiert, die im Fall einer erfolgreichen Anfrage die Liste aller erfragten Feldelemente wiedergibt und eine Meldung mit der Anzahl gefundener Tabelleneinträge anzeigt. Anband der Programmierung eines Literaturdatenbanksystems für das unter WWW befindliche Kommunikationssystem des Landesarbeitskreises ,,Parallelverarbeitung an Fachhochschu-

len in NRW" wurden Erfahrungen in der Problematik dieser Schnittstellenprogrammierung gesammelt.

2 Ein relationales Datenbanksystem zum Online-Zugriff auf die Katalogbestände der Bibliothek der Fachhochschule Köln unter WWW

Für viele Studierende unserer Fachhochschule, die über einen WINDOW-fähigen PC, der mit einem MODEM oder mit einer ISDN-Karte an ein Telekommunikationsnetz gekoppelt werden kann, verfügen, ist es eine Arbeitserleichterung, wenn sie auch außerhalb der Bibliotheksöffnungszeiten und innerhalb von ungünstigen Nahverkehrsanreisezeiten vom häuslichen Schreibtisch aus lesend auf Katalog-, Bestands- und Leihdaten der FH-Bibliothek zugreifen können. Unsere Bibliothek verfügt bereits über ein lokales Rechner-Netz, innerhalb dessen die Katalogdaten unter einem C-ISAM-Datenbanksystem und die Bestands-/Leihdaten unter einem relationalen Datenbanksystem abfragbar sind. Es ist aktuell ein WWW-Bibliothek-Serversystem realisiert worden, auf das von WWW-Clienten aus, die über einen WWW-Browser verfügen, lesend zugegriffen werden kann [Lier 1996] (s. Anlage 1: Gesamtübersicht). Das aktuelle System enthält im Moment ca. 170000 Katalogisate, auf die via WWW nun zugegriffen werden kann.

Als Recherchemöglichkeiten werden dem WWW-Clienten alle die Suchkriterien angeboten, die den Bibliotheksbenutzern bisher zur Verfügung stehen (Recherche über die Titel der Bibliothek nach Titel, Autor, Verlag, ISBN/ ISSN, Körperschaft, Signatur, Jahr ...; bei positivem Rechercheergebnis sollen die verfügbaren Titel, bzw. die Leihfristen der entliehenen Titel angezeigt werden). Erweiterungen der Suchkriterien werden im Moment von uns geprüft. Das Datenbanksystem des WWW-Bibliothek-Servers ist ein durchgängiges relationales System, das den Suchkornfort des vollen SQL-Sprachurnfangs (incl. Match-Kode-Suche) bietet. Es soll in Hinsicht auf die Katalog-, Bestands- und Leihdaten des bisherigen Bibliothekssystem offline betrieben werden, damit eine wechselseitige Ausfallsicherheit gegeben ist. Der dadurch im Unterschied zum Online-Betrieb zusätzlich entstehende Update-Aufwand ist minimal, da er parallel zu den entsprechenden Backup-Läufen des bisherigen Bibliothekssystem ausgeführt werden kann.

Die Arbeitsfolge der bibliographischen WWW-Recherche gliedert sich in folgende vier Arbeitsschritte. Eine WWW-Recherche kann von der Homepage der Fachhochschule Köln *(http://www.fh-koeln.de)* aus oder direkt gestartet werden: *http://wwwopac.fh-koeln.de:8080/opacsuch.html*

1. Rechercheaufruf (s. Anlage 2: Rechercheaufruf). Hier kann der WWW-Client zwei Suchkategorien aus einer Menge von jeweils acht Kategorien auswählen und zugehörige Suchbegriffe eingeben. Er kann die maximale Trefferanzahl bestimmen und den boolschen Operator der Verknüpfung von zwei Suchkategorien festlegen.
2. Anzeige der Ergebnisliste (s. Anlage 3: Ergebnisliste): Im positiven Fall bekommt der WWW-Client die Liste aller Titel des Kataloges angezeigt, die die Bedingung des Rechercheaufrufes erfüllen. Im negativen Fall erhält der Client eine Meldung. Die Ergebnisliste stellt die gefundenen Katalogisate in Zeilenform dar: Jede Ergebniszeile besteht aus den Informationselementen VERFASSER, KURZTITEL, ERSCHEINUNGSJAHR.
3. Titelanzeige (s. Anlage 4: Titelanzeige): Der WWW-Client kann in der Ergebnisliste Einzeltitel auswählen und per Mouse-Click den vollen Katalogeintrag anfordern (incl. Schlagwortangaben, Signatur et. al.).
4. Exemplaranzeige (s. Anlage 5: Exemplaranzeige): Möchte der WWW-Client ein Exemplar des angezeigten Einzeltitels ausleihen, bekommt er in der Exemplaranzeige die Information, welche Exemplare an welchen Standorten ausleihbar sind, bzw. wie lange die Leihfrist dauert. Die Exemplaranzeige wird künftig die Schnittstelle sein, wenn bestimmten WWW-Clienten (d.h. den Angehörigen der FH-Köln) die Möglichkeit eröffnet wird, via WWW-Reservierungen vorzunehmen. Unter Verwendung vorhandener WWW-Server-Standardsoftware kann in bezug auf WWW-Clienten eine Autorisierungsprüfung vorgenommen werden (z.B. es kann künftig mit dem Bibliotheksbenutzerausweis ein Paßwort vergeben werden).

Als einen weiteren Zusatzdienst für WWW-Clienten könnte in Zukunft ein SQL-Bibliotheksdatenbank-Toolkit angeboten werden, das auf Freeware-Software basiert und das eine gespiegelte Version des Bibliotheksauskunftssystem darstellt, so daß der Client die Möglichkeit hat, Ergebnisse erfolgreicher bibliographischer Recherchen zur weiteren Verarbeitung in ein Client-Datenbanksystem abzulegen.

3 Zum Problem der Definition von Stoppwortlisten

Zur Beschleunigung von Recherche-Läufen und innerhalb des automatischen Indexings [Jones 1991, Salton/McGill 1987] werden Stoppwortlisten hochfrequenter Wörter einer vorliegenden Dokumentensprache verwandt. Die Planung des oben beschriebenen WWW-Bibliotheksauskunftssystem gestattete, die im bisherigen Bibliotheksauskunftssystem (SIKIS) verwendete Stoppwortliste einer

Analyse zu unterziehen und sie während einer Testphase durch eine umfangreichere, nach Wortklassen der deutschen Sprache und nach Frequenzen bestimmte Stoppwortliste zu ersetzen. Letztere Liste ist die mit Wortklassenangaben markierte Stoppwortliste des LEMMA2_C- Verfahrens, einer Portierung von LEMMA2 nach C unter UNIX [Willee 1982, Büchel 1995].

Im folgenden wird ein Vergleich beider Stoppwortlisten auf Wortklassenebene gegeben. In Absicht, das Laufzeitverhalten der Recherche zu verkürzen, benutzt das WWW-Bibliotheksauskunftssystem eine Stoppwortliste mit grammatisch vollständigerer Filterleistung. Z.B. sind die Konjunktionen "deshalb", "entweder", "indem", "nachdem" und "sodaß" nur in der Stoppwortliste des LEMMA2Verfahrens enthalten. Ein Ziel besteht weiterhin darin, aus dieser Liste eine Stoppwortliste zu erzeugen, die hinsichtlich der Wortklassenmenge FUW: ={Interjektionen, Konjunktionen, Numeralia, Präpositionen, Pronomina} und hinsichtlich einer Menge flektierter Formen von Hilfs- und Modalverben komparativ vollständig ist. Zum Vergleich wird folgende Tabelle gegeben:

Stoppwortliste:	Wortformen insgesamt:	Adverbien:	FUW:	Verbformen:	übrige:
SIKlS	191	27	124	35	5
LEMMA2	1538	451	538	119	430

Es sollen Leistungsdaten ermittelt werden, in welchem Maße das Suchverhalten durch verschiedene Varianten der Stoppwortliste beschleunigt wird. Im Rahmen eines weiteren Vorhabens soll untersucht werden, wie für fremdsprachige Buchtitel der Bibliiothek Datenbanktabellen mit Stoppwortlisten in den entsprechenden Sprachen aufgebaut werden können.

4 Einbeziehung von maschinenlesbarer Literatur

Für die nächste Zukunft kann davon ausgegangen werden, daß der Bestand maschinenlesbarer Texte einer Hochschulbibliothek deutlich wachsen wird. Dafür gibt es mindestens drei Quellen: (1) Texte, Textsammlungen und Nachschlagewerke, die auf CD-ROM vorliegen, (2) Fachzeitschriften, die an Abonnenten im INTERNET als Graphikformate versandt werden (electronic publishing), (3) Textkorpora, die im INTERNET als Share- oder Freeware zur Verfügung stehen.

Die nachfolgenden Bemerkungen sind rein datentechnischer Natur, sie werden ohne Betrachtung der in diesem Gebiet wichtigen Copyrightfragen ausgeführt.

Unter dem Stichwort "Elektronisches Bücherregal" wurde ein softwaretechnisches Entwicklungsvorhaben begonnen, um unter Zugriff von einem WWW-Server aus (z. B. vom künftigen WWW-Server der Hochschulbibliothek) auf einem Text-Korpora-Server Volltextrecherchen in maschinenlesbaren Textbeständen ausführen zu können. Im Hinblick auf die genannten Server, die UNIX-Maschinen sind, wird eine Text-Retrieval-Software in C unter UNIX mit WWW-Anbindung entwickelt, die vom Leistungsumfang her die Funktionalität kommerzieller auf PC's realisierter Retrieval-Software nachbildet. Die Software besteht aus einer Index- und einer View-Komponente. In der Index-Komponente werden die Texte eines neu in den Server aufzunehmenden Korpus unter Verwendung einer durch den Benutzer spezifizierbaren Stoppwortliste indiziert. Die View-Komponente enthält die Software für die Volltextrecherche und die Darstellung der Rechercheergebnisse. Im Unterschied zu einer Software, die häufig für Suchläufe auf CD-ROM- Texten unter einem WINDOW-System genutzt wird [Greatest Books CollectionTM (1995), Goethe (1995)] und die die Fundstellen von Suchläufen bloß invers unterlegt, werden hier für jeden Treffer die Konkordanzzeilen in einem gesonderten Datenbestand aufbereitet (z.B. mit Seiten/Zeilennummerierung) und es wird die globale Trefferanzahl angegeben. Die modulare Softwareentwicklung gestattet, andere Formen der Konkordanzen benutzerparametrierbar zu programmieren (2*n-stellige Wortumgebungen, Hauptsatz} Satzgefüge-Konkordanz [Büchel 1992]).

5 Eine Vorstudie zu einer empirischen Thesaurierung von Schlagwortbeständen der Informatik-Literatur

Für das vorhandene Mengengerüst der mit Schlagwortketten versehenen Informatiktitel (das ist eine Teilmenge der nach 1990 erschienenen Informatikliteratur) der relationalen Katalog-Datenbank des WWW-Servers der Bibliothek soll ein Abgleich mit dem ACM-Klassifikationssystem der Informatik-Literatur (ACM:= Association for Computing Machinery), die bereits in einer relationalen Datenbanktabelle vorliegt, durchgeführt werden.

Hierzu soll maschinell eine dynamische Listenstruktur für eine repräsentative Menge von Schlagwörterfolgen, die mit normierten Schlagwortphrasen (=: NSWPH) der ACM-Klassifikation übereinstimmen, aufgebaut werden. In dieser Liste werden die (1 :m)-Relationen KATEGORIE - NSWPH und die (l:n)-Relationen NSWPH - BUCHTITEL aufgenommen. Durch sortierte Verarbeitung dieser dynamischen Listenstruktur wird ein Suchbaum mit Knotenfolge gemäß der

Kategorienordnung des ACM-Klassifikationssystem aufgebaut, dessen Blätter normierte Schlagwortphrasen sind. Jedes Blatt ist dann eingefärbt mit einer kreisförmigen verketteten Liste aller Buchtitel, die die gegebene normierte Schlagwortphrase des Blattes als Substring in der Schlagwortkette des Buchtitels enthalten. Für die Bibliothek unserer Fachhochschule wäre dieser maschinell erzeugte Suchbaum dann ein erster für eine weitere Bearbeitung zur Verfügung stehender Thesaurus, der einen Teilbestand der Informatik-Titel systematisch erschließen soll.

Dieses Verfahren soll mit folgendem Beispiel erläutert werden: Ein Benutzer der FH-Bibliothek interessiert sich für Literatur zum Thema *„Datenbanksystem"*. Das Stichwort "Datenbanksystem" tritt in der ACM-Klassifikation in der KATEGORIE "H.2.4. Datenbanksysteme" auf. Die Verbindung zwischen einer KATEGORIE der ACM-Klassifikation und den normierten Schlagwortphrasen der Katalogisate wird durch die (1:m)-Relation KATEGORIE ~ NSWPH modelliert. Jeder zu dieser Relation zugehörige Tupel wird mittels einer Zeile einer Verweistabelle des RDBMS realisiert, die im Feld KATEGORIE den ACM-SchlüsseI (z.B. "H.2.4") und im Feld NSWPH eine der normierten Schlagwortphrasen des Katalogbestandes enthält, die das Stichwort der gegebenen ACM-KATEGORIE als Hauptschlagwort hat (im augenblicklichen Zustand des Kataloges der FH-Bibliothek sind dieses die NSWPH-Einträge: "Datenbanksystem", "Datenbanksystem, deduktives", „Datenbanksystem, objektorientiertes", "Datenbanksystem, relationales", "Datenbanksystem, verteiltes").

Ein Verfahren zur maschinellen Generierung von normierten Schlagwortphrasen, die durch Zerlegung der Schlagwortketten im Katalogbestand gewonnen werden, ist bereits als prototypisches ESQL/C-Programmsystem für den Katalog der FH-Bibliothek realisiert [Huck 1996]. Die Implementation der (l:m)-Relation KATEGORIE->NSWPH wird im Jahr 1997 vorbereitet. Durch eine weitere Tabelle des RDBMS ist prototypisch die (l:n)-Relation NSWPH->BUCHTITEL realisiert, die zu jeder gegebenen normierten Schlagwortphrase alle Titel angibt, die in ihrer Schlagwortkette eine Belegphrase zur gegebenen normierten Schlagwortphrase enthalten. Für die normierte Schlagwortphrase "Datenbanksystem, relationales" wird ein künftiger WWW-Klient (bezogen auf den heutigen Bestand der FH-Bibliothek) n=33 Titel angezeigt bekommen. Faßt man dieses Verfahrensbeispiel zusammen, so erzeugt das projektierte Programmsystem einen Suchbaum mit dem Aufbau: Stichwort (Wurzel) -> ACM-KATEGORIE (Ast) -> normierte Schlagwortphrasen (Blätter), die mit der Liste der zugehörigen BUCHTITEL gefärbt sind (Zielinformation).

Danksagung: Hinweise, Anregungen und Kenntnisse entstanden aus Diskussionen mit Kollegen, Mitarbeitern und Diplomanden des Fachbereiches. Beson-

ders möchte ich hier die Herren Dipl.-Ing. M. Bank, D. Huck, B. Lier, M. Müller und Herrn U. Villers nennen.

Literatur

Büchel, Gregor (1992): "Maschinelle Erschließung und Repräsentation von Terminologiebeständen am Beispiel von G.W.F. Regel: 'Wissenschaft der Logik'", in: Sprache und Datenverarbeitung 2/92, S.3-14.

Büchel, Gregor (1995): "Können Verben semantische Relationen markieren?", in: Weber, Nico (ed.): "Semantik, Lexikographie und Computeranwendungen", Kolloquium, Bonn 27.-28.01.1995, Tübingen (Niemeyer), erscheint Frühjahr 1996.

Goethe, Johann Wolfgang (1995): "Faust - Der Tragödie erster Teil", Klassiker auf CD ROM, Stuttgart, Berlin (Rec1am, in Zusammenarbeit mit Silver Spring).

Greatest Book Collection TM (1995): "World Library's Greatest Book Collection TM Complete, Unabridged 150 Literary Titles on one CD-ROM", Garden Grove, California, USA (World Library Inc.).

Ruck, Dirk J. (1996): "Entwicklung eines Datenbanksystems zur maschinellen Generierung normierter Schlagwortbestände" (Diplomarbeit am FB Nachrichtentechnik der FR Köln), Köln November 1996.

Jones, Susan (1991): "Text and Context - Document Storage and Processing", London, Berlin, Reidelberg etc. (J. Springer).

Lier, Bemd (1996): ,,Entwicklung eines Datenbankssystems zum Online-Zugriff aus die Katalogbestände der FB-Bibliothek im Internetdienst WWW" (Diplomarbeit am FB Nachrichtentechnik der FR Köln), Köln Oktober 1996.

Salton, Gerard / McGill, Michael (1987): " Information Retrieval - Grundlegendes für Informationswissenschaftler", Ramburg, New York, St. Louis etc. (McGraw Hill).

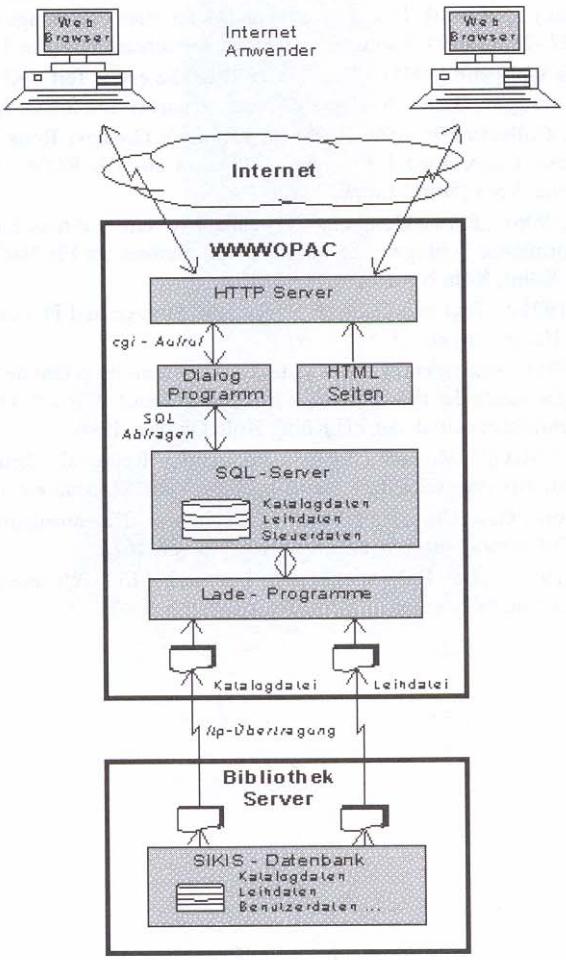
Schmelzle, Oliver / Gast, Christian (1996): ,,Angebunden - Datenbankanwendungen mit dem W3-Gateway für mySQL", iX 2 /1996, S.158-l62.

Willee, Gerd (1982): "Das Programmsystem Lemma2 - Eine Weiterentwicklung von 'Lemma'" in: IKP-Arbeitsberichte Nr.2, Bonn, S. 1-47.

Anlagen

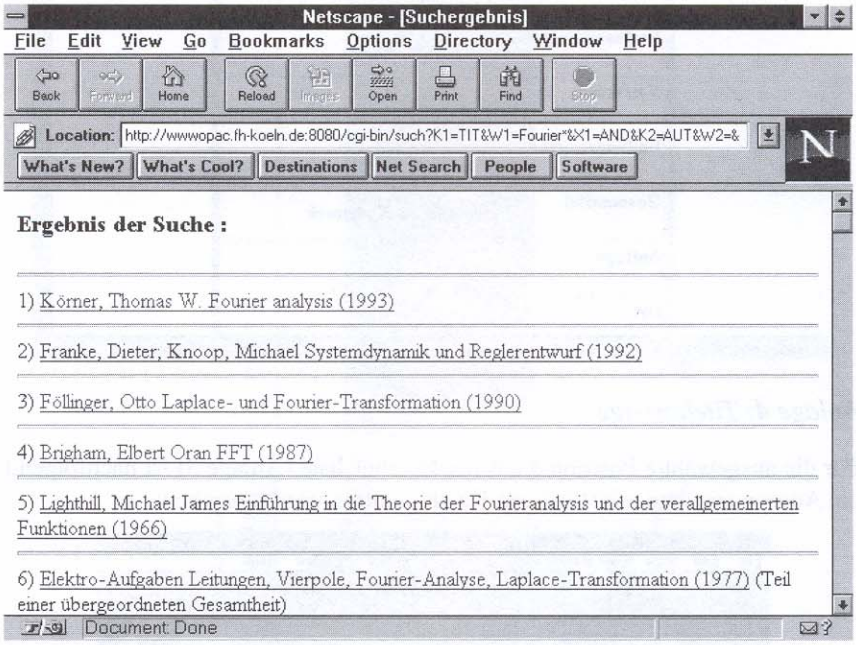
Anlage 1: Gesamtübersicht

Ein Schaubild, das die Rechnerstruktur des in [Lier 1996] entwikkelten WWWOPAC anzeigt, findet man unter: *http://wwwopac.fh-koeln.de:8080/opac_struct.html*



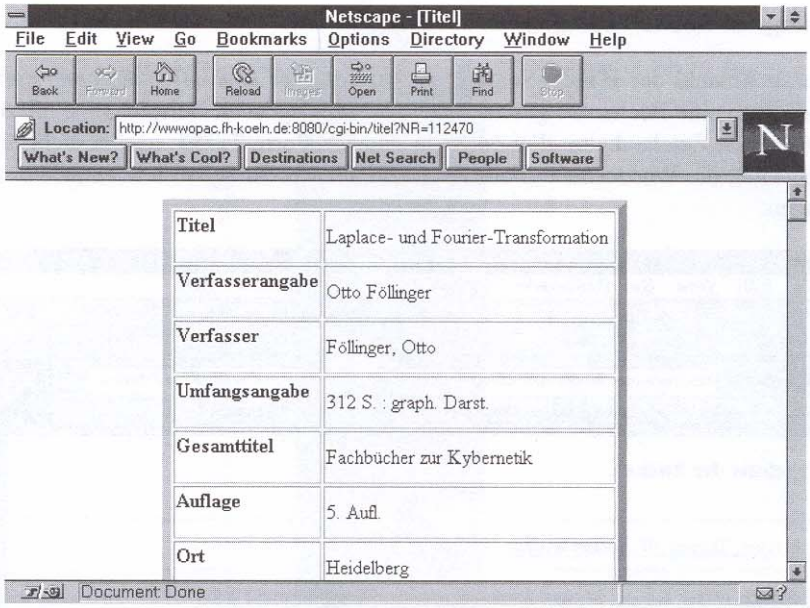
Anlage 2: Rechercheaufruf

Durch Anwahl der HTML-Seite: *http://wwwopac.fh-koeln.de:8080/opac_such.html* kann eine Recherche gestartet werden. Im folgenden Beispiel wird mit der trunkierten Zeichenkette „Fourier*" in der Kategorie TITEL nachgesucht, um z.B. Titel, die Wortformen „Fourierreihe", „Fourier-Analyse" u.ä. enthalten, zu finden:



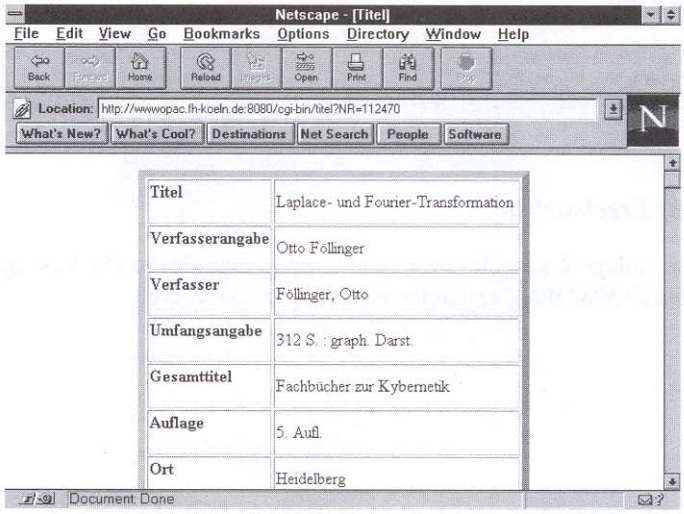
Anlage 3: Ergebnisliste

Für die in Anlage 2 ausgeführte Recherche ist nachfolgend ein Auszug aus der mittels des WWWOPAC erzeugten Ergebnisliste angezeigt:



Anlage 4: Titelanzeige

Für die ausgewählte Position 3 aus der Ergebnisliste (Anlage 3) ist nachfolgend ein Auszug aus dem vollständigen Katalogeintrag wiedergegeben:



Anlage 5: Exemplaranzeige.

Nachfolgend ist die Exemplaranzeige des in Anlage 4 ausgewählten Einzeltitels angezeigt: