Evaluation des WWW - Suchdienstes GERHARD unter besonderer Beachtung der automatischen Indexierung

Diplomarbeit

im Fach Inhaltliche Erschließung
Studiengang Informationsmanagement der
Fachhochschule Stuttgart - Hochschule für Bibliotheks- und Informationswesen

Carmen Krüger, Stuttgart

Erstprüfer: Prof. Holger Nohr
Zweitprüfer: Prof. Dr. Wolfgang von Keitz

Angefertigt in der Zeit vom 09. Juli 1999 bis 11. Oktober 1999

Stuttgart, Oktober 1999

Zusammenfassung

Die vorliegende Arbeit beinhaltet eine Beschreibung und Evaluation des WWW - Suchdienstes GERHARD (German Harvest Automated Retrieval and Directory). GERHARD ist ein Such- und Navigationssystem für das deutsche World Wide Web, welches ausschließlich wissenschaftlich relevante Dokumente sammelt, und diese auf der Basis computerlinguistischer und statistischer Methoden automatisch mit Hilfe eines bibliothekarischen Klassifikationssystems klassifiziert.

Schlagwörter:
Suchdienst, World Wide Web, automatische Klassifikation, Universelle Dezimalklassifikation, Information Retrieval

Abstract

The following study is a description and evaluation of the search engine GERHARD (German Harvest Automated Retrieval and Directory). GERHARD is a search and navigation system for the German World Wide Web. GERHARD only collects academically relevant documents, which are automatically classified with computer - linguistics and statistical methods using a library classification scheme.

Keywords:
search engine, World Wide Web, automatic classification, Universal Decimal Classification, information retrieval

Inhaltsverzeichnis

0. Einleitung

Mit dem DFG - Projekt GERHARD ist der Versuch unternommen worden, mit einem auf einem automatischen Klassifizierungsverfahren basierenden World Wide Web - Dienst eine Alternative zu herkömmlichen Methoden der Interneterschließung zu entwickeln. GERHARD ist im deutschsprachigen Raum das einzige Verzeichnis von Internetressourcen, dessen Erstellung und Aktualisierung vollständig automatisch (also maschinell) erfolgt. GERHARD beschränkt sich dabei auf den Nachweis von Dokumenten auf wissenschaftlichen WWW - Servern.

Die Grundidee dabei war, kostenintensive intellektuelle Erschließung und Klassifizierung von Internetseiten durch computerlinguistische und statistische Methoden zu ersetzen, um auf diese Weise die nachgewiesenen Internetressourcen automatisch auf das Vokabular eines bibliothekarischen Klassifikationssystems abzubilden.
GERHARD steht für German Harvest Automated Retrieval and Directory.
Die WWW - Adresse (URL) von GERHARD lautet: http://www.gerhard.de.

Im Rahmen der vorliegenden Diplomarbeit soll eine Beschreibung des Dienstes mit besonderem Schwerpunkt auf dem zugrundeliegenden Indexierungs- bzw. Klassifizierungssystem erfolgen und anschließend mit Hilfe eines kleinen Retrievaltests die Effektivität von GERHARD überprüft werden.

1. Suchdienste im World Wide Web

1.1 Ausgangssituation

Immer mehr Wissen ist in elektronischer Form über Datennetze verfügbar. Das Internet enthält eine ständig wachsende Menge an Dokumenten. Für den Informationssuchenden wird es daher immer schwieriger, die für ihn relevanten Dokumente ausfindig zu machen. Da das Internet nicht zentral verwaltet wird, gibt es kein einheitliches, verbindliches Ordnungsschema, das die Suche nach Informationsressourcen erleichtern würde.

Der Informationssuchende ist daher auf die mittlerweile zahlreich vorhandenen Suchmaschinen - oder allgemeiner: Suchdienste - angewiesen, die versuchen, einen Überblick über die riesige Datenmenge zu behalten.

Im folgenden soll kurz erläutert werden, wie diese Suchdienste arbeiten, und welche Probleme sich durch die ihnen eigenen Erschließungsmethoden ergeben. Anschließend soll dargestellt werden, wie der Suchdienst GERHARD mit den genannten Problemen umgeht.

1.2 Erschließungsmethoden konventioneller Dienste

Die verschiedenen Suchdienste unterscheiden sich durch ihre unterschiedlichen Funktionsweisen. Dabei gibt es unterschiedliche Ansätze in der angewandten Erschließungsmethode. Unterschieden werden dabei grob zwei Arten von Suchdiensten, die kurz vorgestellt werden sollen, um Vergleiche vom Suchdienst GERHARD zu seiner Konkurrenz ziehen zu können.

1.2.1 Verzeichnisbasierte, manuell erstellte Dienste

Verzeichnisbasierte Dienste strukturieren Internetressourcen in nach Themengebieten geordneten „Katalogen“. Dabei wird unterschieden zwischen Verzeichnissen für bestimmte Themengebiete und großen, allgemeinen, die - im Idealfall - alle Themengebiete umfassen. Thematische Verzeichnisse sind hierarchisch aufgebaut. Von Oberkategorien wie etwa Wirtschaft oder Wissenschaft aus kann sich der Benutzer zu Unterkategorien klicken. Dabei bewegt er sich demnach immer vom Allgemeinen zum Speziellen.[1]

„Kataloge eignen sich vor allem, wenn man zu einem gewissen Thema bzw. Sachgebiet einen Einstieg finden will ohne dabei ganz konkrete Informationsprobleme zu haben.

[1] Vgl. Babiak (1997) S. 53 ff

Das Browsen des Benutzers in einem Katalog erlaubt auch Serendipity - Effekte, die beim Einstieg in neue Gebiete durchaus wünschenswert sind und bei der Stichwortsuche eher ausbleiben.“[2]

Beispiele für Verzeichnisdienste (Directory Services) sind Yahoo! oder Web.de. Die Zuordnung der Dokumente zu den einzelnen Kategorien geschieht in den meisten Fällen manuell.
Manche Suchdienste verwenden als Struktur ihrer Verzeichnisse Bibliotheksklassifikationssysteme; auf diesen besonderen Fall wird in Kap. 1.4 näher eingegangen werden.

1.2.2 Roboterbasierte Dienste (Suchmaschinen)

Roboterbasierte Dienste werden oft auch als search engines oder Suchmaschinen bezeichnet. Es handelt sich allgemein um Programme, die WWW - Hypertextstrukturen automatisch verarbeiten können: Ein Suchroboter durchsucht dabei das World Wide Web nach bestimmten Kriterien und indiziert die gefundenen HTML-Seiten im Volltext (oder Teile davon). Diejenigen Begriffe, die später für den Benutzer suchbar sein sollen, werden damit als Indexeinträge in einer Datenbank abgelegt. Der Benutzer, der ein Stichwort in die Suchmaschine eingibt, sucht also zunächst einmal nicht im WWW direkt, sondern in der Datenbank des jeweiligen Anbieters. Die in der Datenbank gespeicherten Stichworte verweisen dann auf die URLs der entsprechenden Dokumente im World Wide Web. Der Benutzer erhält als Ergebnis seiner Suche eine Liste aller Dokumente, in denen sein Suchwort vorkommt. Eine solche Trefferliste ist meist nach Relevanz der gefundenen Dokumente sortiert (sogenanntes Relevance Ranking).
„Roboterbasierte Suchdienste sind vor allem geeignet für spezielle Suchfragen, bei denen spezifische oder wenig gebräuchliche Suchbegriffe verwendet werden können. [...] Sie sind ungeeignet bei zu allgemeinen Suchbegriffen und zu weiten Themen, weil sie dann zu umfangreiche Treffermengen hervorbringen.“[3]

Bekannte roboterbasierte Suchdienste sind Alta Vista, Execite, Lycos und WebCrawler.
Viele Suchdienste bestehen auch aus einer Kombination der beiden beschriebenen Methoden; die nachgewiesenen Seiten sind auf der einen Seite nach Kategorien

[2] Bakavac (1996) S. 197

sortiert, zusätzlich aber auch im Volltext suchbar. Auch GERHARD ist solch ein kombinierter Dienst. Auf die daraus resultierenden Suchmöglichkeiten für den Benutzer wird in Kap. 2.4 noch näher eingegangen werden.

Ferner existieren sogenannte Meta - Suchdienste, die eine Suche in unterschiedlichen Suchmaschinen gleichzeitig ermöglichen. Meta - Suchmaschinen leiten Suchfragen zur parallelen Bearbeitung an verschiedenste andere Suchdienste weiter und ermöglichen somit eine sehr umfassende Suche. Beispiele für Meta - Suchdienste sind MetaGer (http://mserv.rrzn.uni-hannover.de) für den deutschsprachigen Raum und Metacrawler (http://www.metacrawler.com).

1.3 Aktuelle Probleme von Suchdiensten

Das Hauptproblem der bestehenden Suchdienste ist die Quantität der zu erschließenden Dokumente. Es gibt unterschiedliche Ansätze, dieses Problem zu lösen, z.B. die Spezialisierung der Dienste auf bestimmte Fachgebiete oder auch geographische Einschränkungen. Vor allem die manuell erstellten Verzeichnisdienste wie Yahoo! haben mit dem Mengenproblem zu kämpfen: „An Größe weit hinter den robotergenerierten Webindices ... zurückbleibend, schafft es Yahoo! nicht mehr, irgendeine systematische Auswahl zu treffen, und nur 25 bis 30 Prozent der angemeldeten Ressourcen werden überhaupt noch, oft mit erheblicher Verzögerung, aufgenommen.“[4]

Das Mengenproblem wird also von den roboterbasierten Suchmaschinen besser gelöst, wobei sich dieser Vorteil bei der Suche auch als Nachteil herausstellen kann, „denn die große Zahl verfügbarer Internetquellen sorgt bei vielen Suchen für nicht mehr praktikable Ergebnismengen, die nicht selten mehr als 10.000 Nachweise anbieten und dadurch die Trennung von Treffer und Nicht-Treffer in einen wenig erfolgversprechenden intellektuellen Suchprozeß münden lassen.“[5]

Manuell erstellte Verzeichnisse selektieren durch eine Vorabauswahl der Quellen mehr nach Qualität: „Der Zwang zur intellektuellen Auswahl reduziert einerseits die Menge der verfügbaren Quellen drastisch, sorgt aber andererseits dafür, daß die

---

[3] Oehler (1998) Online
[4] Koch (1998) S. 326 - 335
[5] Lepsky (1998) S. 336

nachgewiesenen Quellen um Größenordnungen über der Durchschnittsqualität im Internet liegen.“[6]

Die schwache Netzabdeckung verzeichnisorientierter Dienste muß dennoch eindeutig als Manko erkannt werden: „Grundsätzlich ist ... der starke Auswahlcharakter des Verfahrens ein echter Schwachpunkt in Hinblick auf einen möglichst vollständigen Nachweis.“[7]

Der WWW - Suchdienst GERHARD will im Prinzip die Vorteile von Verzeichnisdiensten und roboterbasierten Suchmaschinen miteinander verbinden und durch die Verwendung automatischer Verfahren größere Dokumentenmengen für das Browsing erschließen als herkömmliche Verzeichnisdienste.[8]

1.4 Klassifikation von Internetressourcen

Herkömmliche WWW-Kataloge sind inzwischen so umfangreich, daß die Navigation in ihnen schwerfällig geworden ist.[9] Bei Yahoo! z.B. entschied man sich von Anfang an, „eine völlig eigene Struktur zu entwickeln, was heute zu etwa dreißigtausend Kategorien mit unzähligen Querverbindungen geführt hat, deren Logik nur schwer zu durchschauen ist.“[10]

Daher soll im folgenden der Einsatz von bibliothekarischen Systematiken als Alternative zu den meist „hausgestrickten“ Systemen herkömmlicher Verzeichnisdienste näher betrachtet werden, da auch GERHARD mit der UDK (Universelle Dezimalklassifikation) auf eine bibliothekarische Universalklassifikation zurückgreift.

Erfahrungen, große Sammlungen von Dokumenten zu strukturieren, um sie auf diese Weise für den Informationssuchenden zugänglich zu machen, haben Bibliotheken bereits seit Jahrzehnten gesammelt. Warum also das Rad neu erfinden?

1.4.1 Allgemeines zur Klassifikation

„Die Klassifikation spiegelt den Zusammenhang und die Gliederung aller Wissensgebiete wider, wobei sie von den einzelnen Wissenschaften ausgeht und diese

---

[6] Ebd.
[7] Ebd.
[8] Vgl. Wätjen (1998b) Online
[9] Vgl. Bekavac (1996) S. 197
[10] Koch (1998) S. 326

dann in immer kleinere und speziellere Begriffe untergliedert. Diese verschiedenen Gruppen und Unterteilungen erhalten jeweils eine bestimmte Notation[11].“[12]

Klassifikationen wurden bisher vor allem in (öffentlichen und wissenschaftlichen) Bibliotheken als Grundlage für den systematischen Katalog und für eine systematische Aufstellung der Bestände verwendet.

Unter Klassifikationen versteht man hierarchisch organisierte Verzeichnisse, die die verschiedenen Wissensgebiete formal strukturieren, um auf diese Weise einen sachgebietsorientierten Zugang zu den nachgewiesenen Ressourcen zu ermöglichen. Klassifikation ist also „eine von mehreren Methoden von Wissensorganisation. Sie ist kein Selbstzweck, sondern eingebunden in die Aufgabenbereiche der Dokumentenverwaltung und des Navigierens, sowie der Suche in Informationssammlungen.“[13]

Es wird unterschieden zwischen Spezialklassifikationen, die sich auf ein begrenztes Sachgebiet beziehen, und allgemeinen, universellen Klassifikationen für alle Wissensgebiete. Universalklassifikationen sind fächerübergreifend. Sie finden Anwendung bei der Klassifikation von großen, umfangreichen Dokumentensammlungen, die keine Spezialisierung aufweisen, sondern das gesamte Spektrum menschlichen Wissens abdecken.[14]

1.4.2 Klassifikationsstrukturen im World Wide Web

Auch das World Wide Web kann im Prinzip als eine solche unspezialisierte Sammlung von Texten bezeichnet werden. Zahlreiche thematische Verzeichnisse von Internetressourcen greifen daher auf Bibliotheksklassifikationen als Browsingstrukturen zurück.

Eine Auflistung von Internetdiensten, die mit Klassifikationssystemen bzw. kontrolliertem Vokabular arbeiten, findet sich im World Wide Web unter der URL: http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm.[15]

Eine nähere Beschreibung der auf Klassifikationssystemen basierenden Dienste und Verzeichnisse findet sich bei Traugott Koch.[16]

---

[11] Unter einer Notation (lat. = Bezeichnung) versteht man allgemein eine Kombination aus Buchstaben oder Zahlen, mit Hilfe derer ein Dokument eindeutig einer bestimmten Gruppe innerhalb eines Wissenschaftssystems zugeordnet werden kann. (Vgl. Rehm (1991) S. 204)

[12] „Wörterbuch der Fachbegriffe“ der Universitätsbibliothek Bielefeld

[13] Koch (1998) S. 326

[14] Vgl. Buchanan (1989) S. 109 ff

[15] McKiernan (1999) Online

[16] Vgl. Koch (1998) S. 326 - 335

Bekanntestes Interneterschließungsprojekt dieser Art ist Scorpion,[17] ein Forschungsprojekt von OCLC (Online Computer Library Center). Scorpion nutzt die Dewey Dezimal Klassifikation für eine automatische Klassifikation von Internetressourcen.

Folgende Gründe sprechen unter anderem für den Einsatz von Klassifikationssystemen zur Strukturierung von Internetressourcen: Zunächst können Klassifikationen die Suchmöglichkeiten erheblich verbessern:
„Das Browsing in Informationssammlungen wird deutlich erleichtert. Gerade unerfahrene Nutzer oder Personen, die mit dem Fachgebiet, seiner Struktur und Terminologie nicht vertraut sind, werden durch ein logisch strukturiertes Angebot besser unterstützt.“[18]
Klassifikationen ermöglichen ferner einen mehrsprachigen Zugriff auf die entsprechenden Ressourcen.[19]

Durch eine OCLC - Studie konnte gezeigt werden, daß universelle Klassifikationssysteme wie z.B. die DDC (Dewey Decimal Classification) oder die LCC (Klassifikation der Library of Congress) in Bezug auf Umfang und Themenabdeckung durchaus mit dem Verzeichnissystem Yahoo! zu konkurrieren in der Lage sind. Dazu wurden die am häufigsten aufgerufenen Yahoo - Kategorien mit den Systemklassen von DDC und LCC verglichen. Es wurde festgestellt, daß beide Systeme populäre Themen ebenso gut abdecken wie Yahoo und daher für eine Klassifikation von Internetressourcen durchaus geeignet sind.[20]

1.4.3 Die Universelle Dezimalklassifikation (UDK)

GERHARD verwendet als Schema für das Verzeichnis eine maschinenlesbare, leicht abgewandelte und modernisierte Version der UDK (Universelle Dezimalklassifikation), die dem Projekt von der ETH - Bibliothek Zürich zur Verfügung gestellt wurde. Warum gerade die UDK?

„Angesichts der zu erwartenden Mengen zu klassifizierender Dokumente kam nur ein mächtiges Klassifikationssystem in Frage, das zudem alle Fachgebiete in einer

[17] Homepage unter http://orc.rsch.oclc.org:6109/ (Stand 10/99)
[18] Koch (1998) S. 327
[19] Koch (1998) S. 328

hierarchischen Struktur abdecken, maschinenlesbar und möglichst deutsch- und englischsprachig vorliegen sollte.“[21]
Diese Kriterien werden von der UDK erfüllt. Die UDK ist mit insgesamt rund 60.000 Einträgen und ca. 500.000 Textzeilen (27MB) sehr umfangreich, deckt alle Wissensgebiete ab und ist zudem mehrsprachig (Deutsch, Englisch und Französisch).

Die UDK ist eine auf dem Dezimalsystem basierende Universalklassifikation. Jeder Systemstelle innerhalb der UDK ist eine numerische Notation zugeordnet. Die Notation 5 entspricht beispielsweise dem Bereich „Mathematik / Naturwissenschaft“, die Notation 53 dem Bereich „Physik“, 536 dem Bereich „Wärmelehre und Thermodynamik“. Spezifischere Themengebiete erhalten auf diese Weise i.d.R. komplexere Notationen. Durch die Struktur der Notation wird die von GERHARD vorgenommene Klassifikation der Dokumente transparent.[22]
Eine Abbildung der Hauptklassen der (abgewandelten) UDK findet sich auf S.16.

Die Züricher Version der UDK:
Grundlage der UDK bildet ein Sachregister, das auf den Zahlen der Dezimalklassifikation beruht: „Das Sachregister der ETH - Bibliothek ist hierarchisch aufgebaut (Oberbegriffe, Unterbegriffe, Querverweise), d.h. die Notationen (DK - Zahlen) sind in einem Netz von Notationen verankert.“[23]
Die ETH - Bibliothek hat mit ihrem Online - System ETHICS das Experiment unternommen, „die Vorteile eines hierarchisch strukturierten Systems (UDK) mit denen eines Schlagwortsystems zu kombinieren. Das bedeutet, daß im ETHICS ... die DK - Zahlen direkt benannt wurden.“[24]
Die einzelnen Notationen erhalten also mehrsprachige Benennungen. Außerdem werden bestimmten Begriffen ihre entsprechenden Synonyme (so vorhanden) zugeordnet.
Beispiel:[25]

Notation DK 581,5

---

[20] Vgl. Vizine-Goetz (1996) Online
[21] Vgl. Wätjen (1998b) Online
[22] Vgl. Wätjen (1998a) S. 14
[23] Schwaninger (1997) S. 81
[24] Loth (1996) S.17
[25] Beispiel entnommen: Schwaninger (1997) S. 89

Deutsche Benennungen:	**GEOBOTANIK**
	Pflanzengeographie
Englische Benennungen:	**GEOBOTANY**
	Plant geography
Französische Benennungen:	**GEOBOTANIQUE**
	Phytogéographie

Der Vorteil ist offensichtlich: Die Benutzer finden die entsprechenden Dokumente zum Thema, „gleich ob sie in deutscher, englischer oder französischer Sprache suchen und gleich mit welchem Synonym sie ihre Recherche durchführen."[26]

Die UDK ist ferner in der Lage, Relationen zwischen Dokumenten darzustellen. Neben den hierarchischen Relationen existieren zwölf weitere Beziehungstypen zwischen Notationen, wie z.B. Querverweise.

[26] Loth (1996) S. 17

2. Das Projekt GERHARD

2.1 Ziele und Partner

GERHARD ist ein von der Deutschen Forschungsgesellschaft (DFG) im Rahmen des Förderprogramms “Elektronische Publikationen im Literatur- und Informationsangebot wissenschaftlicher Bibliotheken"[27] gefördertes Projekt unter der Leitung des Bibliotheks- und Informationssystems (BIS) der Carl von Ossietzky - Universität Oldenburg in Zusammenarbeit mit dem Institut für Semantische Informationsverarbeitung (ISIV) der Universität Osnabrück und dem Oldenburger Forschungs- und Entwicklungsinstitut für Informationswerkzeuge und -systeme (OFFIS). Laufzeit des Projektes war Oktober 1996 bis März 1998.
Das Ziel war die Entwicklung eines Nachweissystems von WWW-Seiten für den deutschen Wissenschaftsbereich.

„Das GERHARD - Angebot richtet sich in erster Linie an eine wissenschaftliche Nutzerschaft, die wissenschaftlich relevante deutsche
Internetressourcen in einer möglichst vollständigen Datenbank gezielt suchen oder darauf systematisch zugreifen will.“[28] Der entstehende Suchdienst sollte eine Integration von Searching (gezielte Suche) und Browsing (thematische Navigation) bieten, also einen „transparenten Übergang von Einzeltreffern zur Klassifikation und verwandten Treffern.“[29]
Auf intellektuelle und manuelle Verfahren sollte beim Aufbau des Dienstes weitgehend verzichtet werden.

[27] Nähere Informationen auf der Homepage der Deutschen Forschungsgemeinschaft (DFG) unter http://www.dfg.de/foerder/formulare/1_51.htm (Stand: Oktober 1999).
[28] Wätjen (1998a) S. 7
[29] Ebd.

2.2 Zuständigkeiten (nach Modulen)

Eine klassische, roboterbasierte Suchmaschine besteht aus drei Hauptmodulen: [30]

- Gatherer
- Datenbanksystem
- Benutzerschnittstelle

Gatherer werden zum Teil auch als Roboter, Searcher oder Agentenprogramme bezeichnet. Sie durchsuchen das Internet nach bestimmten vom Administrator vorher festgelegten Kriterien nach HTML - Seiten und sammeln Informationen über im Internet vorhandene Dokumente, Dateien etc. oder auch die Dokumente und Dateien selbst. Es sind demnach „Programme, die entlang von WWW - Hypertextstrukturen Dokumente automatisch verarbeiten. Dabei wird ein Dokument geladen, verarbeitet und es werden referenzierte Dokumente rekursiv weiterverfolgt.“[31]

Das **Datenbanksystem** verwaltet die vom Gatherer gelieferten Dokumente und ist für die Indexierung der Daten sowie für die Recherchemöglichkeiten verantwortlich.[32]

Die **Benutzerschnittstelle** besteht bei den meisten Suchmaschinen aus einer formularbasierten Eingabemaske. Sie nimmt die Suchanfragen entgegen und gibt daraufhin eine geordnete Trefferliste aus.[33]

Auch bei GERHARD ist diese modulare Aufteilung zu finden. Entsprechend den verschiedenen Modulen von GERHARD (plus Klassifikationskomponente) wurden die Zuständigkeiten der Projektpartner festgelegt: [34]

- BIS Oldenburg: Leitung, Koordination, Gatherer und Prozeßsteuerung
- ISIV Osnabrück: Klassifikationsverfahren
- OFFIS Oldenburg: Datenbank und Benutzerschnittstelle

2.3 Sammelprozesse

GERHARD verwendet Teile des Programmsystems Harvest zum Sammeln der Dokumente. Harvest ist eine Zusammenstellung von Tools, um Dokumente im Internet zu sammeln, zu analysieren, zu organisieren und Informationen daraus auf Anfrage

[30] Vgl. Mönnich (1999) S. 142 f.
[31] Bekavac (1996) S. 197
[32] Vgl. Mönnich (1999) S. 142 f.
[33] Vgl. Mönnich (1999) S. 143

bereitzustellen.[35] Harvest wurde von der Internet Research Task Force Research Group on Resource Discovery (IRTF-RD) zwischen 1995 und 1996 entwickelt und ist heute weltweit im Einsatz.

Harvest wurde aus folgenden Gründen für das Projekt ausgewählt:
Harvest ist gut konfigurierbar hinsichtlich Suchraum, Dokumententyp und Zugriffsweise und zudem kostenlos verfügbar.[36] Harvest beachtet das Robot Exclusion - Protokoll[37] und arbeitet netzschonend, d.h. sammelt Dokumente nur wenn sie neu sind oder nach einer bestimmten Zeit auf Änderungen hin überprüft werden müssen.

Das Gesamtsystem von Harvest besteht aus mehreren unabhängig voneinander arbeitenden Komponenten. Die beiden Hauptkomponenten sind folgende:

- Der **Gatherer**, sammelt Dokumente aus dem World Wide Web, extrahiert aus ihnen die relevanten Daten und schreibt diese in eine Datenbank. Ein SGML - Parser analysiert dabei die Struktur der HTML - Seiten und gibt sie in Form des Austauschformates SOIF aus, um auf diese Weise eine effektive Weiterverarbeitung der Dokumente zu gewährleisten. Das Summary Object Interchange Format (SOIF) wurde entwickelt, um gesammelte Informationen leicht austauschbar zu halten. Das Format wird für die interne Darstellung der extrahierten Informationen aus dem Dokumenten genutzt.
- Der **Broker** empfängt die Daten vom Gatherer. Er eliminiert mehrfache Informationen und übernimmt die eigentliche Indexierung der Daten. Über ein WWW-Formular werden die in den SOIF - Dateien gespeicherten Informationen zugänglich gemacht.

Als Benutzerschnittstelle dient ein World Wide Web Interface.
Insgesamt sind bisher etwa eine Mio. Dokumente in Form von SOIF - Datensätzen in der Konfigurationsdatenbank von GERHARD gespeichert. Dabei findet das Datenbanksystem Oracle Anwendung. Die Datenbank speichert zusätzlich

---

[34] Vgl. Wätjen (1998a) S. 7
[35] Vgl. Frequently Asked Questions (and Answers) about Harvest
[36] Vgl. Wätjen (1998a) S. 10
[37] Die ständigen Zugriffe der Suchmaschinen belasten weltweite WWW-Server. Das Robot Exclusion - Protokoll ist eine Möglichkeit, Server vor Roboterzugriffen zu schützen. Eine einfach strukturierte Textdatei mit dem Namen robots.txt im Wurzelverzeichnis des Servers gibt Auskunft, welche Seiten für Web Roboter nicht zugänglich sind.

Informationen über Filtereinstellungen, Suchtiefe, Aktualität sowie statistische Daten zu den Indexierungsläufen (z.B. Anzahl der Dokumente pro Domain und Zugriffszeiten).[38]

Bisher werden insgesamt 350 Domains erfaßt, wobei sich das Gathering auf deutsche wissenschaftlich relevante Server aus folgenden Gebieten beschränkt:

- Universitäten, Fachhochschulen, sonstigen Hochschulen
- staatlichen und halbstaatlichen wissenschaftlichen Einrichtungen
- wissenschaftlich relevanten Einrichtungen u. Ämtern auf Bundesebene
- Parteien auf Bundesebene.[39]

Die Auswahl der Server konnte nicht vollständig durch automatische Verfahren vonstatten gehen. Der Suchraum wird manuell aktualisiert.
GERHARD sammelt bisher nur HTML - Dateien, eine Erweiterung um Postscript und PDF - Dateien ist jedoch bereits in Planung.
Bisher ist nicht geklärt, ob das Programmsystem Harvest in naher Zukunft weiterentwickelt werden wird. Ferner verbraucht Harvest in hohem Maße Systemressourcen wie z.B. Speicherbedarf. Es ist daher geplant, Harvest durch den im DESIRE - Projekt entwickelten Gatherer COMBINE zu ersetzen.[40]
Derzeit wird noch Harvest verwendet. Die Integration von COMBINE hat sich als sehr aufwendig herausgestellt und konnte daher noch nicht abgeschlossen werden.[41]

2.4 Suchmöglichkeiten

GERHARD ermöglicht sowohl die konventionelle Stichwortsuche (searching), wie man sie von herkömmlichen Suchmaschinen her kennt, als auch die Navigation durch die Verzeichnisstruktur der UDK (browsing). Ferner ermöglicht der Dienst eine integrierte Suche:
„GERHARD bietet für die Benutzer gegenüber den bereits bekannten Suchmaschinen des World - Wide - Web einen echten Mehrwert durch die Verwendung der UDK als Klassifkationsschema. Dabei haben die Benutzer einerseits die Möglichkeit, durch

[38] Vgl. Wätjen (1998a) S. 13
[39] Vgl. Wätjen (1998a) S. 11
[40] Ebd.
[41] Auskunft von B. Diekmann (Projektteilnehmer) 9/99

Navigation (Browsing) in der UDK zu gewünschten Dokumenten zu gelangen, andererseits von bereits gefundenen Dokumenten in die UDK zurückzuspringen, um von dort zu dem gefundenen Dokument verwandte Treffer zu finden.“[42]

Der Benutzer hat die Wahl zwischen folgenden drei Suchmöglichkeiten:

- Navigation im Verzeichnis
- Suche im Verzeichnis
- Suche in den Dokumenten

Links zu den verschiedenen Suchmöglichkeiten befinden sich im Hauptmenü im linken Frame (Rahmen). Auf diese Weise sind sie jederzeit erreichbar, und der Benutzer kann spontan seine Suchstrategie ändern.
Über das Hauptmenü ist ferner eine recht ausführliche, kontextsensitive Hilfefunktion aktivierbar.

2.4.1 Navigation im Verzeichnis

Auf der Einstiegsseite sieht der Benutzer die Hauptklassen der UDK, von denen aus er seine Suche beginnen kann. Die Einstiegsseite enthält folgende Informationen (vgl. Abb.1):

- die Namen der Hauptklassen der (modifizierten) UDK, von denen aus der Benutzer die Navigation beginnen kann
- die Anzahl der Dokumente, die der entsprechenden Hauptklasse zusammen mit ihren Unterklassen zugeordnet sind
- die Anzahl der Dokumente, die nur der entsprechenden Hauptklasse zugeordnet sind

Die folgende Abbildung zeigt die Einstiegsseite für die Navigation im Verzeichnis mit den UDK – Hauptklassen:

Abbildung 1: Einstieg in die Navigation: Die UDK - Hauptklassen

[42] Wätjen (1998a) S. 19

Von den alphabetisch angeordneten Hauptklassen aus kann der Benutzer zu den Unterklassen / Unterbegriffen und von dort aus auf immer speziellere Themengebiete navigieren, um zur gewünschten Information zu gelangen.
Dabei sind immer nur diejenigen UDK - Stellen sichtbar, denen Dokumente zugeordnet sind.
Der Begriff, bei dem sich der Benutzer gerade befindet, ist zentriert und fettgedruckt dargestellt. Oberhalb links eingerückt befinden sich seine Oberbegriffe, unterhalb rechts sind seine Unterbegriffe angeordnet. Ein Begriff kann dabei mehrere Oberbegriffe besitzen.
Im folgenden Beispiel ist „**Studentenschaft, Studierende, Hochschulleben**“ der aktuelle Begriff. „**Hochschulorganisation**“ ist in der Hierarchie höher angesiedelt und stellt den Oberbegriff zum aktuellen Begriff dar. Die Einträge „**Studentische Politik**“, „**Studierende / Mobilität**“, „**Aeltere Studierende**“ usw. bilden die Unterbegriffe des aktuellen Begriffes.

Abbildung 2: Verzeichnisstruktur von GERHARD

Auf diese Weise kann sich der Benutzer dem gesuchten Begriff langsam annähern.
Durch das Anklicken der Zahl neben dem gewünschten Begriff gelangt der Benutzer in eine Übersicht aller Dokumente, die diesem Begriff durch die automatische Klassifikation zugeordnet worden sind: die sogenannte Dokumentübersicht von GERHARD. Sie ist in der folgenden Abbildung zu sehen:

Abbildung 3: Dokumentübersicht

In der Dokumentübersicht werden maximal 25 Dokumente gleichzeitig aufgeführt. Der Benutzer kann sich mit Hilfe von Buttons vor- bzw. zurückbewegen.

Von der Dokumentübersicht aus kann der Benutzer entweder sofort über die URL zum Originaldokument wechseln oder sich zuvor die sog. Ausführliche Anzeige des Dokuments am Bildschirm ausgeben lassen.
Wenn unter den gefundenen Dokumentenverweisen nichts Passendes dabei war, besteht für den Benutzer die Möglichkeit, von der Dokumentübersicht zur

Ausgangsstelle im Verzeichnis zurückzukehren oder zu den Unterklassen der angezeigten Klasse - so vorhanden - zu wechseln.

Die ausführliche Anzeige bezieht sich also jeweils auf eine einzige gefundene Quelle und enthält sowohl extrahierte Informationen aus dem HTML - Dokument wie Titel, URL, Autor / Adresse sowie Inhalte aus dem Dokument selbst, als auch vom System generierte Informationen in Form der dem Dokument zugeordneten Stellen der UDK. Auch von der ausführlichen Anzeige aus kann der Benutzer wieder direkt zum WWW - Originaldokument wechseln. Von den zugeordneten Verzeichniseinträgen in der Ausführlichen Anzeige aus kann er aber auch direkt zu verwandten Themengebieten navigieren und dabei durch Klicken diejenigen zugeordneten Verzeichniseinträge aktivieren, die mit dem gewünschten Dokument ebenfalls verknüpft sein sollen. Auf diese Weise kann eine Suche nach und nach verfeinert werden.

Abbildung 4: Ausführliche Anzeige eines Dokuments in GERHARD

2.4.2. Suche im Verzeichnis

Diese Suchfunktion erlaubt eine Stichwortsuche im UDK - Verzeichnis selbst. Es kann dabei nur nach *einem* Suchwort recherchiert werden; Verknüpfungen von Suchbegriffen sind an dieser Stelle nicht möglich. Das System ignoriert dabei Groß- und Kleinschreibung und konvertiert automatisch Umlaute und Sonderzeichen im eingegebenen Suchbegriff.
Als Ergebnis werden alle UDK - Verzeichniseinträge angezeigt, die diesen Suchbegriff enthalten.
Von hier aus kann der Benutzer zur Navigation in die Verzeichnisstruktur wechseln, indem er den Namen der gewünschten UDK - Klasse anklickt und so zu den Ober - bzw. Unterbegriffen des Eintrags gelangt oder er sieht sich die zugeordneten Dokumente in der Dokumentliste an.

Vorteil der Suche im Verzeichnis:
„Statt des hierarchischen Navigierens können die Klassenbezeichnungen auch direkt recherchiert werden, um mit dem Browsing inmitten der Hierarchie und nicht von ganz

oben zu beginnen oder um interdisziplinäre Themen in verschiedenen Ästen der UDK direkt aufzusuchen.“[43]

2.4.3 Suche in den Dokumenten

Wie konventionelle Suchmaschinen erlaubt auch GERHARD eine Stichwortsuche im Volltext der nachgewiesenen Informationsressourcen. Dabei kann der Benutzer zwischen einer einfachen und einer erweiterten Suche in den Dokumenten wählen.

Einfache Stichwortsuche

In ein einfaches Formular mit nur einem Eingabefeld kann der Benutzer in den ConText - Volltext - Indices[44] nach einem oder mehreren Stichwörtern suchen. Bei Eingabe mehrerer Suchworte besteht die Möglichkeit, logische Verknüpfungen einzusetzen, um die Suchanfrage zu spezifizieren. Ferner besteht die Möglichkeit der Trunkierung der Suchbegriffe mit sog. Wildcards: Dabei steht „%“ für eine beliebig lange Zeichenkette und „_“ für ein beliebiges Zeichen.[45]

Die Ergebnisausgabe zur einfachen Suche ist formal identisch mit der Dokumentliste in Abbildung 3, wenngleich hier die Dokumente auch unterschiedlichen UDK - Stellen zugeordnet sein können.

Erweiterte Suche

Die erweiterte Suche ermöglicht eine feldbezogene Suche nach Titel, Überschrift, Autor / Adresse sowie Stichwörtern im Text. Dabei können - wie bei der einfachen Stichwortsuche - logische Verknüpfungen zwischen den einzelnen Suchbegriffen vorgenommen werden.

Abbildung 5: Erweiterte Suche in GERHARD

Da aber die wenigsten HTML - Dokumente Metadaten beinhalten, können Felder wie Autor oder Adresse in nur ca. 8% der Fälle ausgefüllt werden.

---

[43] vgl. Wätjen (1998b) Online

[44] ConText ist ein Modul innerhalb der Oracle - Datenbank für die Speicherung von und die Suche in Volltexten

[45] Die Stemming - Funktion von Oracle ermöglicht die Reduzierung des Suchwortes auf dessen Grundform. Werden Wildcards benutzt, so wird Stemming in genau dem Wort, in dem sie verwendet werden, abgeschaltet (vgl. auch Online - Hilfe).

2.4.4 Bewertung der Suchmöglichkeiten

Ich würde jedem GERHARD – Benutzer empfehlen, vor einer Suche die wichtigsten Hilfethemen durchzulesen, da sich die Struktur der UDK für Laien meiner Meinung nach nicht auf Anhieb erschließt. Ein rein intuitiv agierender Benutzer käme zwar u.U. auch zu einem Ergebnis, indem er sich z.B. an der Browsingstruktur „entlangklickte“, würde auf diese Weise aber nicht alle Recherchemöglichkeiten nutzen, die der Suchdienst bietet. Die Hilfe ist recht übersichtlich aufgebaut; Suchmöglichkeiten und Fachvokabular werden in leicht verständlicher Form erklärt.
Die Integration von Searching und Browsing ist meiner Meinung nach gut gelöst. Bei der Navigation durch die Verzeichnisstruktur kann die Betrachtung der begrifflichen Umgebung des Suchbegriffs helfen, die Suche zu spezifizieren.
Insgesamt sind die Suchmöglichkeiten des Dienstes meines Erachtens sehr vielfältig.

Positiv anzumerken ist ferner die Tatsache, daß GERHARD eine mehrsprachige Suche erlaubt. Da viele wissenschaftliche Texte in englischer Sprache publiziert werden, ist es ein Vorteil, daß GERHARD gleichzeitig deutsch- und englischsprachige Texte zum Thema findet, ohne daß die Suchformulierung geändert werden muß.

In Bezug auf die Verzeichnisstruktur ist anzumerken, daß die UDK zum Teil nicht fein genug untergliedert ist. In der Konsequenz sind bestimmten Verzeichniseinträgen unübersichtlich viele Dokumente zugeordnet. Beispiele sind die Klassen „Unix“ mit 12.413 oder „Java“ mit 27.403 zugeordneten Dokumenten. Hier wäre eine weitere Spezifizierung der Klassen angebracht, da gerade Informatikthemen im Internet naturgemäß gehäuft anzutreffen sind. Ein Nachweisdienst sollte sich diesen Gegebenheiten anpassen.

3. Die Klassifikationskomponente in GERHARD

Im folgenden Kapitel soll auf die Klassifikationskomponente des Suchdienstes eingegangen werden.
An erster Stelle soll die kurze Definition einiger grundlegender Begriffe stehen. In diesem Zusammenhang folgt eine Übersicht über die unterschiedlichen Ansätze automatischer Indexierungsverfahren. Anschließend wird das automatische Klassifikationssystem von GERHARD beschrieben und einer kritischen Betrachtung unterzogen.

3.1 Begriffsklärungen und Definitionen

3.1.1 Computerlinguistik

Unter Computerlinguistik im weiteren Sinne versteht man „ein zwischen Linguistik und Informatik liegendes interdisziplinäres Forschungsgebiet, das sich mit der maschinellen Verarbeitung natürlicher Sprachen beschäftigt. Computerlinguistik im engeren Sinne ist ein Teilgebiet der modernen Linguistik, das berechenbare Modelle menschlicher Sprache entwirft, implementiert und untersucht.“[46]
Computerlinguistische Verfahren werden heute vermehrt eingesetzt, um große Datenmengen zu strukturieren und - soweit möglich - auch inhaltlich zu erschließen. Das Hauptproblem der maschinellen Sprachverarbeitung liegt dabei in der Komplexität natürlicher Sprache.

3.1.2 Automatische Indexierung[47]

Automatische Indexierung ist eine Form von Inhaltserschließung.

[46] Vgl. Uszkoreit (1998) Online
[47] Die Begriffe Automatische Indexierung und Automatische Klassifizierung werden in der Literatur häufig synonym gebraucht.

Inhaltserschließung von Texten hat zum Ziel, diese wiederfindbar zu machen:
„Der klassische Ausgangspunkt manueller Indexierung - etwa in einem bibliothekarischen Umfeld - ist der, daß eine Dokumentationseinheit ohne Inhaltserschließung zunächst nicht suchbar ist. Entsprechend hoch ist der Stellenwert der manuellen Indexierung, die einen gezielten inhaltsorientierten Zugriff erst ermöglicht.“[48] Manuelle Verfahren aber bedeuten zum Großteil einen hohen Aufwand.
Unter der Bezeichnung „Automatische Indexierungsverfahren“ werden alle Verfahren gefaßt, die eine automatisierte, sachliche Erschließung von Textinhalten zum Ziel haben.
Eine detaillierte Beschreibung der verschiedenen Indexierungsverfahren würde den Rahmen dieser Arbeit eindeutig sprengen; daher werde ich mich im folgenden auf eine kurze Erläuterung der unterschiedlichen Ansätze beschränken.
Es lassen sich grob vier Ansätze unterscheiden:[49]

- Freitextverfahren
- Statistische Verfahren
- Computerlinguistische Verfahren
- Begriffsorientierte Verfahren

Freitextverfahren stellen die einfachste Form automatischer Indexierung dar: Bei diesen Verfahren wird jedes Wort eines Textes - nachdem es mit einer Stopwortliste abgeglichen wurde - im Index abgelegt und damit zum suchbaren Stichwort. Freitextverfahren sind leicht zu implementieren und daher weit verbreitet.
Die Suche für den Benutzer wird allerdings erschwert: Eine reine Stichwortsuche bringt oft nicht das gewünschte Ergebnis oder einfach zu viele Treffer auf eine Suchanfrage.
„Die Schwäche des Ansatzes ist offensichtlich: Den Inhalt eines Textes kann man nur sehr bedingt durch die darin vorkommenden Wörter (u.U. noch deren Reihenfolge) charakterisieren.“[50]
Das Hauptproblem liegt also darin, „daß in keinster Weise der Tatsache Rechnung getragen wird, daß gleiche Inhalte völlig unterschiedlich formuliert sein können. Damit

[48] Knorz (1994) Online
[49] Vgl. Knorz (1994) Online
[50] Reimer (1992) S. 174

ein Benutzer trotzdem möglichst viele der von ihm zu einem bestimmten Thema gesuchten Texte findet, muß er seine Frageformulierung entsprechend ausbauen.“[51]

Statistische Verfahren ziehen Rückschlüsse von der Häufigkeit des Auftretens bestimmter Wörter (bzw. Terme) auf ihre Relevanz für das entsprechende Dokument. Eine einfache Formel für eine solche Termgewichtung wäre beispielsweise folgende:
inverse Dokumenthäufigkeit (t, d) = Auftretenshäufigkeit von t in d /
Dokumenthäufigkeit von d
„Nach diesem Ansatz ist ein Indexterm also je aussagekräftiger für den Inhalt eines Dokuments, je häufiger er in dem Dokument auftritt und je seltener er überhaupt vorkommt.“[52]

Computerlinguistische Verfahren arbeiten mit morphologischen und syntaktischen Methoden, um bei der Indexierung die sprachlichen Gesetzmäßigkeiten der Dokumententexte zu berücksichtigen. Das Ziel dabei ist, verbesserte Retrievalergebnisse gegenüber dem reinen Freitextretrieval zu erzielen. [53]
Computerlinguistische Verfahren basieren auf Regelwerken und / oder Wörterbüchern. Danach wird eine Unterteilung in regelbasierte und wörterbuchbasierte Verfahren vorgenommen.

Regelbasierte Verfahren:

Bei diesen Verfahren werden die zu indexierenden Dokumententexte mit Hilfe von sogenannten Parsern entsprechend einem vorgegebenen Regelwerk analysiert. Das zugrundeliegende Regelwerk besteht zumeist in einer Grammatik.
Auf diese Verfahren soll hier nicht näher eingegangen werden, da sie bei GERHARD keine Verwendung finden.

Wörterbuchbasierte Verfahren:

Hier erfolgt eine Analyse der Texte in Form eines Abgleichs mit einem oder mehreren Wörterbüchern, die beispielsweise Relationen zwischen Wortformen und ihren zugehörigen Grundformen enthalten können.
Es sind weiterhin Wörterbücher denkbar, die auch semantische Beziehungen zwischen den Wörterbucheinträgen enthalten.

---

[51] Reimer (1992) S. 174
[52] Reimer (1992) S. 175
[53] Vgl. Knorz (1994) Online

Wörterbuchbasierte Verfahren sind sehr pflegeintensiv und benötigen oft Updates. Da sich mit ihnen aber jeder Einzelfall regeln läßt, arbeiten sie wesentlich genauer als regelbasierte Verfahren, die sich häufig nur an einer Grammatik orientieren.[54]

Begriffsorientierte Verfahren versuchen die *Bedeutung* von Texten zu ermitteln und sich auf diese Weise der manuellen Indexierung zu nähern. Es sollen also nicht nur Stichworte aus dem Text übernommen werden, sondern auch solche Deskriptoren vergeben, die sich nur aus dem Gesamtzusammenhang erschließen lassen.[55]

Oft integrieren automatische Indexierungssysteme mehrere der vorgestellten Ansätze. Dies ist auch bei GERHARD der Fall: Das automatische Klassifikationsverfahren greift sowohl auf den computerlinguistische als auch auf den statistischen Ansatz zurück.

3.2 Automatische Klassifikation in GERHARD

Die automatische Klassifikation der HTML - Texte basiert auf computerlinguistischen und statistischen Methoden. Ziel dabei ist, „den natürlichsprachlichen Gehalt der gesammelten Internet - Dokumente auf die UDK abzubilden und so eine Klassifikation dieser Dokumente zu erreichen.“[56]
Das Institut für Semantische Informationsverarbeitung (ISIV) der Universität Osnabrück hat eine Methode für den automatischen Klassifikationsprozeß von GERHARD entwickelt, die im folgenden beschrieben und analysiert werden soll.

Der Klassifikationsprozeß besteht aus folgenden Verfahrensschritten:

- Linguistische Aufbereitung der UDK
- Erstellung eines UDK - Lexikons
- Aufbereitung der zu klassifizierenden Dokumente
- Analyse der Notationen

3.2.1 Linguistische Aufbereitung der UDK

Im Sinne des Projektziels wurde die UDK einigen Änderungen unterzogen:

[54] Knorz (1994) Online
[55] Vgl. Knorz (1994) Online
[56] Vgl. Carstensen (1997) Online

„Die UDK - Einträge mußten für eine Abbildbarkeit der Dokumente auf die Einträge aufbereitet werden. Dies geschah durch die Erstellung eines aus der UDK gebildeten Wörterbuches, in dem die natürlichsprachlichen Begriffe ihren entsprechenden Notationen zugeordnet werden.“[57]

Die UDK wurde zu diesem Zweck folgendermaßen aufbereitet:
Es wurde eine Normierung der Umlaute vorgenommen, diakritische Zeichen wurden beseitigt und eine einheitliche Kleinschreibung eingeführt. Verweise, Anmerkungen, Kommentare und Klammerungen wurden aus der UDK entfernt. Ferner mußten Umformungen von einzelnen UDK - Einträgen vorgenommen werden.
Beispiel:

Ein UDK - Eintrag, der in der Form „Übersetzungen / Technische u. naturwissenschaftliche“ vorliegt wird umgewandelt in „Technische u. naturwissenschaftliche Übersetzungen“, da nur die letztere Form in einem natürlichsprachlichen Dokument vorkommen wird. Auf diese Weise wird ein späterer Abgleich im Rahmen einer späteren Textanalyse überhaupt erst ermöglicht. [58]
Die UDK enthält viele komplexe Einträge, wie beisp. Aufzählungen. Es müssen daher die Einzelbegriffe aus Aufzählungen erst selektiert und dann ihren entsprechenden Notationen zugeteilt werden.
Die Aufbereitung der UDK (durch einen Compiler) erfolgte einmal offline und dann nur noch nach eventuellen Updates der UDK neu.

Schließlich wurde die Extraktion von natürlichsprachlichen Begriffen (Phrasen) aus den UDK - Einträgen vorgenommen. Dies geschah durch ein maschinelles Verfahren.

3.2.2 Erstellung eines UDK - Lexikons

Aus den aus der UDK extrahierten Begriffen wurde ein Lexikon gebildet.
Die Einträge des UDK - Lexikons liegen allgemein in folgender Form vor:

NATÜRLICHSPRACHLICHER_SCHLÜSSEL TRENNSYMBOL NOTATION

Der Extrahierte Begriff „Esperanto“ wird demnach folgendermaßen abgelegt werden:

[57] Wätjen (1998a) S. 15

esperanto:=089.2

Jeder Lexikoneintrag ist also einer bestimmten UDK - Notation zugeordnet.
„Die Erstellung des UDK - Lexikons erfolgte unter der Prämisse, daß das Verfahren zur Textanalyse Elemente eines Textdokuments auf die natürlichsprachlichen Lexikoneinträge abbilden kann, die somit als Suchschlüssel für die entsprechenden Notationen dienen.“[59]

Zu diesem Zweck wurden sogenannte Stoppwörter - also für das Retrieval nicht relevante Wörter - aus den Einträgen entfernt, um auf diese Weise Fehlklassifikationen einzuschränken. Dazu gehören nicht - inhaltstragende Wörter wie z.B. bestimmte Wortarten (Konjunktionen, Präpositionen, Artikel) und häufig vorkommende Verben und Hilfsverben sowie Abkürzungen („usw.“, „bzw.“). GERHARD verwendet dabei eine vom Max Planck Institut Nijmegen entwickelte Liste aus deutschen und englischen Wörtern.

Ferner wurden die aus der UDK extrahierten Begriffe einer morphologischen Reduktion unterzogen, so daß im UDK - Lexikon nur unflektierte Stammformen der Begriffe zu finden sind:
„Die morphologische Reduktion besteht zur Zeit in der Trunkierung von Flexionsendungen der in UDK - Einträgen vorkommenden Wörter.“[60]
Durch die Verwendung der Trunkierung werden also alle grammatikalischen Varianten der Begriffe berücksichtigt. Auf diese Weise können alle im Text vorkommenden Wortformen und -endungen durch eine morphologische Analyse anhand der UDK - Einträge identifiziert werden.

Nach diesen Umformungen konnte das UDK - Lexikon implementiert werden. Dies geschah in Form eines sog. Buchstabenbaums, der neben den Zeichen / Buchstaben der gespeicherten Begriffe auch ein Trunkierungssymbol verwaltet (z.B. ‚#‘).[61] Das Trunkierungssymbol kennzeichnet ein variables Wortende.
„Auf diese Weise lassen sich bei der Textanalyse spezifische Wortformen auf die reduzierten Einträge abbilden. Da momentan auch die nicht - trunkierten Wörter der

[58] Vgl. Carstensen (1997) Online
[59] Wätjen (1998a) S. 16
[60] Wätjen (1998a) S. 17

Einträge durch dieses Symbol ergänzt werden, bilden somit Zeichenketten wie die folgende den Input für den Aufbau eines in dieser Art implementierten UDK - Lexikons: **umwelt# frau#:396,5.00.504**“[62]

3.2.3 Aufbereitung der Dokumente

Auch die zu klassifizierenden HTML - Dokumente werden aufbereitet: Zunächst erfolgt auch hier ein Abgleich mit einer Stopwortliste. Der HTML - Text wird in einen ASCII - Text umgewandelt und die Zeichenformate an die des UDK - Lexikons (normierte Umlaute und einheitliche Kleinschreibung) angepaßt. Auf diese Weise soll ein effizientes Nachschlagen der Begriffe im UDK - Lexikon ermöglicht werden. [63]

„Die Analyse eines Dokuments erfolgt danach als sequentielle Abarbeitung seines bereinigten Textes durch iteratives Look - up eines Präfix dieses Textes in dem UDK - Lexikon (plus nachfolgendem Abschneiden). Die lineare Abarbeitung sowie die kompakte Implementierung als Buchstabenbaum sichern dabei eine effiziente Verarbeitung.“[64]

Dabei wird jeweils der längste matchende Präfix gesucht und als Ergebnis geliefert.[65] Die Textanalyse liefert also eine bestimmte Anzahl von Übereinstimmungen und ermittelt auf diese Weise passende Notationen für das Dokument. Diese werden - zusammen mit Angaben über die Häufigkeit des Auftretens der jeweiligen Begriffe - an die statistische Analyse weitergegeben.

3.2.4 Analyse der Notationen

Die ermittelten Notationen werden einer statistischen Analyse unterzogen. Dabei wird die systematische Struktur der UDK - Notationen folgendermaßen für eine Relevanzbewertung der einzelnen zugeordneten Notationen ausgenutzt:

„Je mehr Notationen mit einem gemeinsamen Präfix vorliegen, desto sicherer ist die Zuordnung zu dem entsprechenden Themenbereich in der UDK. Je länger dieser Präfix ist, desto spezifischer ist die inhaltliche Klassifikation. Das Verfahren verrechnet beide Faktoren miteinander und selektiert die relevantesten Notationen.“[66]

---

[61] Vgl. Wätjen (1998b) Online
[62] Carstensen (1997) Online
[63] Vgl. Carstensen (1997) Online
[64] Wätjen (1998b) Online
[65] Vgl. Carstensen (1997) Online
[66] Wätjen (1998b) Online

Das zu klassifizierende Dokument wird dabei zweimal analysiert. Zum einen werden Notationen aufgrund des Titels, zum anderen in Bezug auf das Gesamtdokument vergeben. Dabei erhalten diejenigen Notationen, die aus der Titelanalyse hervorgegangen sind, einen höheren Relevanzwert. Ein berechneter Relevanzfaktor gibt dabei jeweils an, wie exakt die Zuordnung eines Dokuments zu einer UDK - Klasse ist. Es findet also ein Ranking der Dokumente nach Qualität der Zuordnung statt. Dadurch soll erreicht werden, daß die für eine Klasse relevanteren Dokumente vor den weniger relevanten positioniert werden.[67]

Ein Dokument wird auf diese Weise durchschnittlich sechs bis sieben verschiedenen UDK - Klassen zugeordnet.

3.3 Probleme bei der automatischen Klassifikation

Nicht alle Dokumente werden durch die automatische Klassifikation den richtigen UDK - Klassen zugeordnet. Klassifizierungsfehler sind z.B. durch Mehrsprachigkeit und Homonyme bedingt:
„Das Stichwort ‚Windows' führt nicht nur zu Eintragungen in der Informatikklasse ‚Betriebssysteme', sondern auch im Bauingenieurwesen bei ‚Fenstern und Türen'. Problematisch sind auch die Häufung von Homonymen bei kurzen Klassenbezeichnungen wie chemischen Elementen."[68]
Fehlklassifikationen treten zum Teil auch bedingt durch die stark variierende Struktur der HTML - Dokumente auf.[69]

Zusammenfassend kann man sagen, daß eine automatische Klassifikation zwar Zeit spart und eine große Dokumentenmenge verfügbar macht, man erkauft sich aber „die Quantität ein Stück weit auf Kosten der Qualität".[70]

Wie exakt GERHARD im Einzelnen klassifiziert, soll in Kapitel 5 anhand von konkreten Ergebnissen einiger Suchfragen evaluiert werden.

---

[67] Vgl. Jahresbericht OFFIS (1997)
[68] Wätjen (1998a) S. 31
[69] Vgl. Wätjen (1998a) S. 30

4. Retrievaltest (Methodik)

4.1 Bewertungskriterien für Retrievalergebnisse

Im folgenden Kapitel sollen zunächst allgemeine Methoden zur Bewertung von Retrievalsystemen beschrieben werden. Im zweiten Teil des Kapitels werden Vorgehen und Bewertungskriterien für den folgenden Test dargestellt.

4.1.1 Begriffsklärungen

„Gegenstand des **Information Retrieval** ist die Repräsentation, Speicherung und Organisation von Informationen und der Zugriff zu Informationen.“[71]

Mit Hilfe eines sog. **Retrievalsystems** kann auf (meist in elektronischer Form) gespeicherte Dokumente zugegriffen werden. Dies soll in möglichst ökonomischer Weise geschehen: „Der Stellenwert eines Retrievalsystems hängt von der Fähigkeit ab, die gesuchte Information rasch und ohne Ballastinformation nachzuweisen, der Möglichkeit nichtrelevante Informationen auszusondern und der Vielseitigkeit der nutzbaren Retrievalmethoden.“[72]

Auch eine Suchmaschine wie GERHARD ist ein Retrievalsystem.

Einige unterschiedliche Methoden zur Bewertung von Retrievalsystemen, sollen im folgenden kurz erläutert werden.

4.1.2 Allgemeine Bewertungskriterien

Die Bewertung von Retrievalsystemen kann unter unterschiedlichen Aspekten geschehen. Zum einen kann die sog. Systemeffektivität bewertet werden, die vor allem den Nutzen des Systems für den Benutzer, also den Informationssuchenden, im Auge

---

[70] Tröger (1998) Online
[71] Salton / McGill (1983) S. 1
[72] Salton / McGill (1983) S. 168

behält, und zum anderen die sog. Systemeffizienz, welche vor allem Kosten und Nutzen des Systems gegeneinander abwägt.[73]

Die am häufigsten verwendeten Kriterien zur Bewertung von Retrievalsystemen sind die beiden Meßzahlen Recall (Vollständigkeit) und Precision (Präzision).

Als Recall bezeichnet man dabei die Fähigkeit eines Systems, alle relevanten Dokumente nachzuweisen, als Precision hingegen die Fähigkeit, *nur* relevante Dokumente nachzuweisen. [74]

Anders formuliert: Der Recall mißt, welcher Anteil der im System vorhandenen relevanten Dokumente bei einer Suchanfrage gefunden wird, und die Precision gibt an, welcher Anteil der gefundenen Dokumente als relevant einzustufen ist.

Recall = Zahl der nachgewiesenen relevanten Dokumente / Zahl aller relevanten Dokumente der Datenbank

Der Recall ist somit ein Maß für den quantitativen Erfolg einer Recherche.[75]

Der Wertebereich für den Recall liegt zwischen 0 und 1 (0 für das schlechteste, 1 für das bestmögliche Ergebnis).

Precision = Zahl der nachgewiesenen relevanten Dokumente / Zahl aller relevanten Dokumente

Die Precision bestimmt die Genauigkeit bzw. den Ballast einer Recherche.

Auch der Wertebereich der Precision liegt zwischen 0 und 1, wobei auch hier 1 als das beste und 0 als das schlechteste mögliche Ergebnis zu werten ist. Die Precision kann auch als Prozentzahl ausgedrückt werden:

Precision - Wert von 1 = 100% (100 % der gefundenen Dokumente sind relevant)

Precision - Wert von 0 = 0 % (0% der gefundenen Dokumente sind relevant)

Mit ansteigendem Recall fällt typischerweise die Precision - und umgekehrt. Maximaler Recall und maximale Precision sind in der Praxis gleichzeitig daher nie zu erreichen.

Weitere Leistungskriterien können sein: Der Suchaufwand für den Benutzer, die Antwortzeiten des Systems, die Präsentation der Ergebnisse einer Recherche und die Abdeckung der Datenbank. [76]

---

[73] Vgl. Salton / McGill (1983) S.168
[74] Vgl. Salton / McGill (1983) S.172
[75] Vgl. Lepsy (1998) S. 337

4.1.3 Bewertung von Internetsuchdiensten

In bisher durchgeführten Retrievaltests von Suchdiensten im Internet kommen unterschiedliche Bewertungskriterien zur Anwendung. In den meisten Fällen wird aber auf die in Kap. 4.1.2 erläuterten Meßzahlen zurückgegriffen.

Recall - Werte sind dabei jedoch schwer zu ermitteln:
„True recall, the ratio of the total number of relevant elements in the space to the total of relevant results returned by the search, cannot be calculated for Web space, because the total number of relevant links changes quickly and is practically unknowable.“[77]

Für eine Bewertung von Suchdiensten wird daher meist die Precision als Maßzahl herangezogen.[78] Die verschiedenen Test unterscheiden sich in der Anzahl der Dokumente aus der Treffermenge einer bestimmten Suchanfrage, die für die Berechnung der Precision als Grundmenge verwendet wird. Bei den meisten Tests werden die ersten zehn oder zwanzig Treffer einer Suchanfrage (die mit Hilfe eines integrierten Ranking - Algorithmus als die relevantesten ermittelt wurden) für die Ermittlung der Precision herangezogen. Man spricht dabei auch von der „First X Precision“:

„True precision, the ratio of relevant elements returned to the total number of elements returned, is too arduous to calculate, because it would mean examining all of the links returned by a service, which may number in the thousands or millions. The ‚first X precision‘ is designed to reflect the quality of service delivered to the user: how good is the relevancy within the first few pages of results?“[79]

4.1.4 Relevanzkriterien

Die Meßzahlen Recall und Precision basieren auf der Bestimmung der sog. Relevanz von Dokumenten. Wie aber ist der Begriff Relevanz zu definieren?

Zunächst einmal ist zu unterscheiden zwischen der sog. objektiven Relevanz, die eine Beziehung zwischen Dokument und Suchanfrage darstellt, und der sog. subjektiven

---

[76] Vgl. Salton / McGill (1983) S. 172
[77] Leighton / Srivastava (1997) Online
[78] Vgl. Leighton / Srivastava (1997) Online

Relevanz, hinter der eine Beziehung zwischen Dokument und Informationsbedürfnis steht: „Von mehreren gefundenen Dokumenten mit gleichen vom Benutzer gefragten Eigenschaften ... werden einige von ihm trotzdem als nicht relevant eingestuft.“[80]
Im folgenden soll von einem benutzerorientierten Relevanzbegriff ausgegangen werden. Relevanz kann dabei als „Relation zwischen einem Dokument (bzw. Dokumentbeschreibung) und einem Benutzer im Bezug zu seinem Informationsbedürfnis“[81] verstanden werden. Dabei kann von Benutzer zu Benutzer differieren, welche Dokumente er in Bezug auf eine bestimmte Fragestellung als relevant bezeichnet - dies kann z.B. von seinem Vorwissen abhängen.
Ein objektives Relevanz - Kriterium gibt es nicht. In Retrievaltests spielen daher immer auch subjektive Bewertungen eine Rolle.

Für eine Bewertung der Effektivität eines Suchdienstes müssen daher vorab genaue Kriterien für die Relevanzbewertung der gefunden Seiten festgelegt werden: Welche Kriterien muß ein Treffer erfüllen, um von einem Benutzer möglicherweise als relevant bzw. nützlich eingestuft zu werden?
Beachtet werden sollte dabei auch die mögliche Zielgruppe des Dienstes. Welche Anforderungen stellt diese an den Dienst?
Trotz aller Versuche, objektive Relevanzkriterien zu finden, ist eine Beurteilung dessen, was als „relevant“ in Bezug auf eine Suchanfrage bezeichnet wird, sehr subjektiv. Im Idealfall sollten daher die Bewertungen mehrerer Personen in den Test eingehen. Durch die knapp bemessene Zeit für diese Arbeit mußte ich den folgenden Test allein durchführen, daher ist er natürlich keinesfalls als repräsentativ anzusehen.

4.2 Kriterien und Vorgehen beim folgenden Test

Der GERHARD - Projektbericht beinhaltet die folgende Aussage:
„Eine perfekte automatische Klassifizierung ist unmöglich, aber zu 80% richtige automatische Zuordnungen sind effizienter und besser als Hand- und Kopfarbeit.“[82]
Bisher wurde noch kein GERHARD - Retrievaltest durchgeführt, die 80% beruhen auf Schätzungen der Projektleitung.[83]

---

[79] Leighton / Srivastava (1997) Online
[80] Panyr (1983) S.97
[81] Ebd.
[82] Wätjen (1998a) S.31
[83] Auskunft eines Projektteilnehmers (im Gespräch), 9/99

Im folgenden Test soll daher zunächst eine empirische Überprüfung der 80% durchgeführt werden. Ferner soll ein Precision - Wert für die gefundenen Dokumente ermittelt werden und schließlich eine Bewertung der Aktualität des Suchdienstes erfolgen.

Aus Gründen der weiten Verbreitung und der einfachen Interpretierbarkeit bietet es sich an, die Maße Recall und Precision für eine Beurteilung der Ergebnisse des Retrievaltests zu GERHARD heranzuziehen. Aus bereits genanntem Grund ist der Recall jedoch nicht, bzw. nur sehr schwer zu ermitteln. Daher soll sich der folgende Test auf die Bestimmung der Precision einiger Suchergebnisse beschränken.

4.2.1 Bestimmung der Precision

Anhand von insg. 20 Beispielklassen erfolgt eine Relevanzbeurteilung der Dokumente, um auf der Basis dieser Ergebnisse einen Precision - Wert ermitteln zu können:

Stufe 1:

In der ersten Stufe werden die ersten 20 (erreichbaren) Dokumente der Dokumentliste hinsichtlich ihrer Zuordnung zur entsprechenden UDK - Klasse analysiert, um eventuelle Fehlklassifikationen ausfindig zu machen.

Ein Dokument gilt als „richtig klassifiziert“, wenn es thematisch zur zugeordneten UDK - Klasse paßt, unabhängig davon, ob es tatsächlich verwertbare Informationen zum Thema enthält.

Beispiel: Der Prosatext über einen „Hacker“ auf der Homepage eines Studenten würde in diesem Fall in der UDK - Klasse „Hacker“ (als Unterbegriff von „Datensicherheit“) als „richtig klassifiziert“ eingestuft werden. Ob dieses Dokument tatsächlich weiterführende, wissenschaftliche Informationen zum Thema enthält, ist dabei zunächst noch nicht von Interesse.

Die Informationsseite über den Bundestagsabgeordneten Hans - Joachim *Hacker* in derselben Klasse ist dagegen eindeutig das Ergebnis einer Fehlklassifikation. Hier paßt das Dokument zwar formal zur Suchanfrage, da es die gesuchte Zeichenfolge enthält, das ausgewiesene Thema wird im Dokument jedoch nicht behandelt.

Stufe 2:

Die als „richtig klassifiziert“ eingestuften Dokumente werden im zweiten Schritt einer genaueren Analyse unterzogen.

Es erfolgt eine Relevanzbewertung der Dokumente. Dabei soll von einem Benutzer aus einem wissenschaftlichen Umfeld ausgegangen werden, der sich über ein bestimmtes Thema informieren will. In Anlehnung an Gödert soll im folgenden Test ein vergleichsweise „weiches“ Relevanz - Kriterium Anwendung finden:

„Ein Dokument wird als relevant für eine sachliche Frage betrachtet, wenn vermutet werden kann, daß sich ein Nutzer dieses Dokument näher ansehen würde.“[84]

Im folgenden Test gelten als **relevant**:

- Texte, die das entsprechende Thema behandeln, z.B. Vorlesungsskripte oder wissenschaftliche Abhandlungen, FAQs zum Thema etc.
- Literaturhinweise zum Thema, ebenso Sammlungen von Links (Web - Bibliographien) oder Buchbesprechungen
- Seiten, die auf Diplomarbeiten oder Dissertationen zum entsprechenden Thema verweisen
- Homepages von Instituten zum entsprechenden Thema. Hier befinden sich oftmals Publikationen, Verweise auf Publikationen zum entsprechenden Thema, Projektbeschreibungen u.ä.

Als **nicht relevant** gelten alle Dokumente, die keine näheren Informationen zum entsprechende Thema enthalten:

- Vorlesungsverzeichnisse und Informationen zu Lehrveranstaltungen
- Informationen zum Erwerb von Leistungsnachweisen
- Merkblätter für Studierende, Semesterpläne, Informationen zu Klausurthemen
- Ankündigungen von Vorträgen, Veranstaltungsübersichten u.ä.

Als nicht relevant gelten damit alle Dokumente, die für die entsprechenden Institutionen (Universitäten, Fachhochschulen etc.) nur von internem Interesse sind.

[84] Gödert (1997) S. 12

Da sich der Suchdienst GERHARD explizit an eine wissenschaftliche Nutzerschaft richtet, soll die Relevanzbewertung der gefundenen Dokumente ferner unter einem wissenschaftlichem Aspekt geschehen:
Eine Informationsseite über den Kinofilm „The Net“ wird demnach als „nicht relevant“ eingestuft, wenngleich das entsprechende Suchwort (in diesem Fall „Hacker“) in dem Dokument enthalten ist.

4.2.2 Bewertung der Aktualität

Um eine Aussage über die Aktualität des Suchdienstes machen zu können, wird über die Bewertung der aktiven Links hinaus, für jede Klasse die Prozentzahl der inaktiven Links ermittelt.
Inaktive Links, sind zum einen Verweise auf URLs, die sich geändert haben bzw. nicht mehr existieren und auf dem entsprechenden Server nicht mehr gefunden werden können[85] und zum anderen Links, die zu einer internen Fehlermeldung des Suchdienstes führen.
Gründe für interne Fehlermeldungen können z.B. folgende sein: [86]

- Eine Komponenete der Datenbank ist gerade heruntergefahren (z.B. aus Wartungsgründen),
- eine Komponente der Datenbank wird neu konfiguriert,
- alle Webserver sind beschäftigt.

Die ermittelte Prozentzahl der inaktiven Links ist dabei jeweils bezogen auf die Zahl der URLs, die für den Test angeklickt werden mußte, um für eine Bestimmung der Precision an 20 zugängliche Dokumente zu gelangen.

[85] Fehlermeldung 404 (file not found)

5. GERHARD Retrievaltest

5.1 Fragenkatalog

Es wurden insgesamt 20 Klassen für den Retrievaltest ausgewählt, die zum Teil allgemeine, zum Teil aber auch sehr spezielle Themengebiete abdecken. Zehn Klassen stammen aus dem Bereich Informatik und Computerwissenschaften, weitere zehn aus dem Bereich Wirtschaftswissenschaften. Jeweils die ersten 20 (verfügbaren) Dokumente aus der Dokumentliste wurden bewertet. Die Testkollektion umfaßt somit 400 Dokumente.

Folgende Klassen / Unterklassen wurden für den Test untersucht:

1. Wirtschaftsethik, Unternehmensethik
2. Wirtschaftsgeschichte
3. Wirtschaftstheorien
4. Löhne, Gehälter
5. Marktforschung
6. Bruttosozialprodukt
7. Marketing + Marketingstrategie + Absatzwirtschaft
8. Optionen (Börsenwesen)
9. Ablauforganisation
10. Lean Management
11. Computergeschichte, Informatikgeschichte
12. SGML
13. Datenschutz

[86] Auskunft eines Projektmitglieds (10/99)

14. Hypermedia, Hypertext
15. Relationale Datenbanken
16. Kommunikationsprotokolle
17. Unix
18. Computerviren
19. Hacker
20. Pascal (Programmiersprache)

Eine repräsentative Auswahl zu treffen, war angesichts der knapp bemessenen Zeit nicht möglich. Eine solche müßte beispielsweise auch die Tatsache berücksichtigen, daß manche Klassen anfälliger für Fehlklassifikationen sind als andere, um nicht durch eine Häufung solcher Klassen innerhalb der Auswahl ein verfälschtes Ergebnis entstehen zu lassen.
Die Auswahl der zu analysierenden Klassen im folgenden Test erfolgte nach dem Zufallsprinzip.

Die Relevanzbewertung erfolgte nur anhand der HTML - Dokumente selbst, nie allein aufgrund ihrer URL oder Dokumentbeschreibung. Daher konnten nur Dokumente berücksichtigt werden, die zum Zeitpunkt des Tests auch tatsächlich verfügbar waren.

5.2 Ergebnisse

Die Ergebnisse des Tests wurden in zwei Tabellen festgehalten.
Die erste Tabelle beinhaltet die entsprechenden Daten zur Bestimmung der Precision des Suchdienstes sowie Informationen über den Anteil der falsch klassifizierten Dokumente, die zweite Tabelle läßt Rückschlüsse auf die Aktualität von GERHARD zu.

UDK - Klasse	falsche Zuordnung	relevant	nicht relevant	Precision in %
Wirtschafsethik, Unternehmensethik	0	14	6	70%
Wirtschaftsgeschichte	0	12	8	60%
Wirtschaftstheorien	1	12	7	60%
Löhne, Gehälter	15	5	0	25%
Marktforschung	3	9	8	45%
Bruttosozialprodukt	0	20	0	100%
Marketing + Marketingstrategie	0	14	6	70%

Marketingstrategie + Absatzwirtschaft				
Optionen (Börsenwesen)	19	1	0	5,0%
Ablauforganisation	0	16	4	80%
Lean Management	4	7	9	35%
Computergeschichte, Informatikgeschichte	0	20	0	100%
SGML	1	18	1	90%
Datenschutz	3	12	5	60%
Hypermedia, Hypertext	2	11	7	55%
Relationale Datenbanken	2	11	7	55%
Kommunikations-protokolle	0	9	11	45%
Unix	4	11	5	55%
Computerviren	3	14	3	70%
Hacker	6	9	5	45%
Pascal (Programmiersprache)	2	11	7	55%
?	3,25	11,8	4,95	59%

Abbildung 6: Testergebnisse (1)

Von 20 Dokumenten sind im Durchschnitt 3,25 falsch klassifiziert (16,25%). Demnach wurden 83,75% der Dokumente durch die automatische Klassifikation der richtigen UDK - Klasse zugeordnet.

Die geschätzte Größe des Projektteams von 80% kann damit bestätigt werden.

Bei Betrachtung der Tabelle fällt auf, daß die falsch klassifizierten Dokumente gehäuft in einigen wenigen UDK - Klassen zu finden sind.

Die Häufung der Fehlklassifikationen in der Klasse „Optionen (Börsenwesen)“ ist beispielsweise durch eine Verwechslung mit dem Begriff *„Option“* aus dem Bereich der Informatik zu erklären.

Die Klasse „Löhne und Gehälter“ weist zahlreiche Dokumente nach, die die Verbform *„gehalten“* beinhalten - und eine Informationsseite über den CDU Kreisverband der Stadt „*Lohne“* wird in diesem Zusammenhang vermutlich ebenfalls auf wenig Interesse stoßen.

Ursache für die Fehlklassifikationen sind also zum großen Teil Homonyme.

Die durchschnittliche Anzahl der richtig klassifizierten und relevanten Dokumente liegt bei 11,8 Dokumenten (von 20), die Precision für die getesteten Klassen liegt damit bei durchschnittlich 59%.
Die Zahl der relevanten / nicht relevanten Dokumente variiert ebenfalls von Klasse zu Klasse. Viele nachgewiesene Seiten aus dem Informatikbereich konnten nicht als relevant bezeichnet werden, weil sie eher den internen Rechnerbetrieb der jeweiligen Institutionen betrafen und daher nicht von allgemeinem Interesse waren, (z.B. Informationen über Paßwortvergabe, Rechnerlizenzen etc.).
Ferner weist der Suchdienst sehr viele Vorlesungsverzeichnisse u.ä. nach, welche ebenfalls keine Informationen zum Thema beinhalten.

Die zweite Tabelle läßt Rückschlüsse auf die Aktualität des Suchdienstes zu. Als Indikator für die Aktualität des Suchdienstes gilt dabei die Prozentzahl der Links, die sich seit dem letzten Sammellauf des Suchdienstes geändert haben, also nicht mehr aktuell sind und eine Fehlermeldung bringen.

Als inaktive Links gelten ferner URLs, deren Aufrufen eine interne Fehlermeldung des Suchdienstes zur Folge hat. Diese sollen zwar nicht für eine Bewertung der Aktualität herangezogen werden, sie werden aber dennoch mit aufgelistet, da sie nicht weiterführen und damit für einen Benutzer des Suchdienstes nutzlos sind.

UDK - Klasse	**Fehlermeldung 404**	**Request failed (interne Fehlermeldung)**	**Inaktive Links insgesamt**
Wirtschafsethik, Unternehmensethik	29,0%	6,5%	35,5%
Wirtschaftsgeschichte	9,1%	0%	9,1%
Wirtschaftstheorien	41,2%	0%	41,2%
Löhne, Gehälter	28,6%	0%	28,6%
Marktforschung	48,7%	0%	48,7%
Bruttosozialprodukt	28,6%	0%	28,6%
Marketing + Marketingstrategie + Absatzwirtschaft	31,0%	0%	31,0%
Optionen (Börsenwesen)	36,1%	8,3%	44,4%
Ablauforganisation	38,2%	2,9%	41,1%
Lean Management	34,4%	3,1%	37,5%

Computergeschichte, Informatikgeschichte	0,0%	0,0%	0,0%
SGML	12,9%	22,6%	35,5%
Datenschutz	50%	0,0%	50,0%
Hypermedia, Hypertext	33,3%	15,4%	48,7%
Relationale Datenbanken	48,7%	0,0%	48,7%
Kommunikations-protokolle	25,9%	0,0%	25,9%
Unix	13,0%	43,5%	56,5%
Computerviren	35,5%	0,0%	35,5%
Hacker	25,9%	0,0%	25,9%
Pascal (Programmiersprache)	63,5%	4,8%	68,3%
?	31,7%	5,4%	37,1%

Abbildung 7: Testergebnisse (2)

Im Durchschnitt brachten 37,1% der für den Test untersuchten URLs eine Fehlermeldung. 31,7% der Dokumente führte zur Fehlermeldung 404 („File not found“). Knapp jeder dritte aufgerufene Link der Testkollektion war damit veraltet.

5.3 Bewertung

Ein Problem ist meiner Meinung nach die Tatsache, daß GERHARD zu viele irrelevante Seiten nachweist, welche für die jeweiligen wissenschaftlichen Institutionen eher von internem Interesse sind, wie beisp. Informationen zu Lehrveranstaltungen, Termine etc. Zum Teil ist es sehr mühsam, die Spreu vom Weizen zu trennen.

Für die Precision von 59% liegen leider keine Vergleichswerte vor, da bei mir vorliegenden Tests von Suchmaschinen andere Relevanzkriterien angesetzt wurden. Angesichts der Tatsache, daß die Klassifikation zu 100% auf automatischen Verfahren beruht, ist das Ergebnis meines Erachtens recht gut, jedoch ist die Precision von 59% vor allem auch das Ergebnis des sehr weit gefaßten Relevanzkriteriums.

Die Aktualität der Links läßt zum Teil sehr zu wünschen übrig: Viele Links sind veraltet oder bringen Fehlermeldungen. Die Datenbank sollte häufiger aktualisiert, d.h. die vorhandenen Seiten reindexiert werden.

6. Ausblick

Abschließend würde ich sagen, daß GERHARD für eine wissenschaftliche Informationssuche geeignet ist.

In der Tatsache, daß sich der Sammelraum nur auf Deutschland beschränkt, sehe ich allerdings einen Nachteil, da vermutlich kein Benutzer zu einem Thema explizit deutsche Publikationen sucht. Eine Erweiterung des Suchdienstes auf internationale Server wäre sicher ein interessantes Projekt, leider ist bisher nur eine Erweiterung um Server des deutschsprachigen Auslandes in Planung.

Die automatische Klassifikation des Suchdienstes erreicht noch nicht das Niveau intellektueller Verfahren; dennoch verfolgt das Projekt GERHARD meiner Meinung nach mit dem Einsatz eines automatischen Erschließungsverfahrens einen sehr vielversprechenden Ansatz, der Informationsflut im Internet Herr zu werden.

Literaturverzeichnis

(URLs entsprechen dem Stand 11/99)

Babiak, U.: *Effektive Suche im Internet: Suchstrategien, Methoden, Quellen*. Köln (1997).

Becavac, B.: *Suchverfahren und Suchdienste des World Wide Web*. In: Nachrichten für die Dokumentation 47 (1996) 4, S. 195 - 213

Buchanan, B.: *Bibliothekarische Klassifikationslehre*. München (1989).

Carstensen, K.-U.: *Die Klassifizierung von Texten in GERHARD*. Online: http://www.cl-ki.uni-osnabrueck.de/~kai/Klassifikation.html

Frequently Asked Questions (and Answers) about Harvest. Online: http://www.rz.go.dlr.de:8081/www/harvest-1.4.pl2-docs/FAQ.html

Gödert, W.; Lepsy, K.: *Semantische Umfeldsuche im Information Retrieval in Online - Katalogen.* Fachhochschule Köln, Fachbereich Bibliotheks- und Informationswesen (1997). (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; 7). Online: http://www.fbi.fh-koeln.de/fachbereich/papers/index/band7/sem_in.htm

Jahresbericht OFFIS Uni Oldenburg (Oldenburger Forschungs- und Entwicklungsinstitut für Informationswerkzeuge und -systeme) Projekte. (1997), Online: http://www.offis.uni-oldenburg.de/jahresbericht/jb97/p9_3.htm

Knorz, G.: *Automatische Indexierung*. In: Wissensrepräsentation und Information Retrieval. Hrsg.: R.-D. Hennings (u.a.). Potsdam (1994), S. 138-196. Online: http://www.iud.fh-darmstadt.de/iud/wwwmeth/publ/skript/autind94/paper1.htm

Koch, T.: *Nutzung von Klassifikationssystemen zur verbesserten Beschreibung, Organisation und Suche von Internetressourcen*. In: Buch und Bibliothek 50 (1998) 5, S. 326 - 335

Leighton, H.; Srivastava, J.: *Precision among World Wide Web Search Services - (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos*. (6/1997) Online: http://www.winona.msus.edu/library/webind2/webind2.htm

Lepsky, K.: *Im Heuhaufen suchen - und finden. Automatische Erschließung von Internetquellen: Möglichkeiten und Grenzen*. In: Buch und Bibliothek 50 (1998) 5, S. 336 - 340

Loth, K.: *Wissensorganisation durch ein neues Notationssystem - eine konstruktive Kritik der UDK*. In: ABI - Technik 16 (1996) 1, S. 17 - 28

McKiernan, G.: *Beyond Bookmarks: Schemes for Organizing the Web*. (1999). Online: http://www.public.iastate.edu/~CYBERSTACKS/CTW.htm

Mönnich, W.*: Kriterien zur Bewertung und Auswahl von Internetsuchmaschinen*. In: 21. Online - Tagung der DGI. Frankfurt am Main, 18. bis 20. Mai 1999. Proceedings.

Oehler, A.: *Informationssuche im Internet - In welchem Ausmaß entsprechen existierende Suchwerkzeuge für das World Wide Web Anforderungen für die wissenschaftliche Suche.* Berlin (1998) Magisterarbeit im Fach Informationswissenschaft. Online: http://userpage.fu-berlin.de/~angela/mag/mag.htm

Panyr, J.: *Automatische Indexierung und Klassifikaiton*. In: Automatisierung in der Klassifikation (1983), S.90 - 111

Reimer, U.: *Verfahren der automatischen Indexierung. Benötigtes Vorwissen und Ansätze zu seiner automatischen Akquisition: Ein Überblick*. In: Experimentelles und praktisches Information Retrieval: Festschrift für Gerhard Lustig / Hrsg.: R. Kuhlen. Konstanz (1992), S. 171-194

Rehm, M.: *Lexikon Buch - Bibliothek - neue Medien.* München (1991).

Salton, G.; McGill, M. J.: *Information Retrieval - Grundlegendes für Informationswissenschaftler*. Hamburg (1987).

Schwaninger, L.: *Mehrsprachigkeit und Begriffshierarchie bei der Literaturrecherche an der ETH - Bibliothek Zürich*. In: 19. Online - Tagung der DGD. Frankfurt am Main, 14. bis 16. Mai 1997. Proceedings.

Tröger, B*.: „Und wie halten Sie es mit der Internet - Erschließung?“ - Bibliothekarische Gretchenfragen von IBIS bis GERHARD*. In: Bibliotheksdienst 32 (1998) 11, Online: http://www.dbi-berlin.de/dbi_pub/bd_art/98_11_03.htm

Uszkoreit, H.: Vorlesungsskript *„Einführung in die Computerlinguistik“* (1998), Online: http://www.coli.uni-sb.de/~hansu/VLCL1/index.htm

Wätjen, H.-J. et al.: *GERHARD: German Harvest Automated Retrieval and Directory. Bericht zum DFG-Projekt*. Bibliotheks- und Informationssystem der Universität Oldenburg, (1998a), Online: http://www.gerhard.de/info/index_de.html

Wätjen, H.-J.: *GERHARD - Automatisches Sammeln, Klassifizieren und Indexieren von wissenschaftlich relevanten Informationsressourcen im deutschen World Wide Web*. In: B.I.T. online 1 (1998b) 4, Online: http://www.bitonline.de/archive/ausgabe4-98/98120101.html

„Wörterbuch der Fachbegriffe“ der Universitätsbibliothek Bielefeld. Online: http://www.ub.uni-bielefeld.de/help/dictionary.htm

Vizine-Goetz, D.: *Using Library Classification Schemes for Internet Resources* (1996), Online: http://www.oclc.org/oclc/man/colloq/v-g.htm