Informationen im Web anbieten und wiederfinden

Traugott Koch, University Library Lund, NetLab; Florian Seiffert, Hochschulbibliothekszentrum Köln; Hans-Joachim Wätjen, Bibliotheks- und Informationssystem Oldenburg

Abstract:

Die Qualität des Informationsangebotes im Web wird in hohem Maße nicht nur durch die Inhalte, sondern auch durch deren formale Aufbereitung, Strukturierung und Erschließung von den AutorInnen oder den Webmastern selbst bestimmt. Der beste Inhalt nützt niemandem, wenn er über die Suchdienste sogar von den fortgeschrittensten Rechercheuren nicht gefunden werden kann, weil die AutorIn beim Input elementare Fehler gemacht hat. Die Schwierigkeiten der existierenden Such- und Navigationsdienste mit schlechtem Input werden noch verstärkt durch grundsätzliche, immanente Probleme bei der Indexierung von Hypertext.

Anhand von Beispielen und empirischen Auswertungen des umfangreichen Datenmaterials beim HBZ und beim GERHARD-Projekt wird die Qualität des Informationsangebotes von Bibliotheken und wissenschaftlichen Einrichtungen in Deutschland untersucht, um die Art und Häufigkeit typischer Input-Fehler und struktureller Probleme aufzuzeigen.

Die Moderatoren des Tutorials werden den TeilnehmerInnen neue Suchdienste vorstellen und dabei auf die Entwicklung fachspezifischer roboter-basierter Dienste und deren Integration mit intellektuell erstellten Qualitätsdiensten besonders eingehen. Beispielhaft werden Hilfsmittel zur Auswahl der passenden Suchmaschinen und Subject Trees, Recherchestrategien und avancierte Möglichkeiten zur Verbesserung der Suchergebnisse demonstriert.

Die Erzeugung von Metadaten nach dem Dublin Core Set und deren Nutzung in den Such- und Navigationsdiensten eröffnen neue Perspektiven für die Verbesserung des In- und Outputs. Das Ergebnis des Tutorials soll in technischen, praktischen und organisatorischen Vor- und Ratschläge zusammengefaßt werden. Die TeilnehmerInnen sollen sich durch aktive Beiträge am Gelingen des Tutorials beteiligen und prakische Erfahrungen mitbringen.

Was Sie erwartet:

1 Übersicht über die verschiedenen Typen von Suchdiensten

2 Hilfsmittel zur Wahl passender Suchdienste für bestimmte Zwecke

3 Beispiele avancierter Möglichkeiten zur Verbesserung der Suchergebnisse

- 3.1 Konzeptsuche / Erweiterung der Suchanfrage
- 3.2 Kontrolliertes Vokabular
- 3.3 Ranking
- 3.4 Integration von Searching und Browsing

4 Metadaten und Suchdienste

5 Zur Qualität des Inputs im Web

- 5.1 beim Angebot
- 5.2 bei der Abfrage: Folgerungen aus der Auswertestatistik der Bibliothekarischen Suchmaschine des HBZ

1 Übersicht über die verschiedenen Typen von Suchdiensten

von Traugott Koch

- Veränderte Typologie **http://www.ub2.lu.se/tk/websearch_systemat.html#typol**
- Mit neuen/wichtigen Beispielen **http://www.ub2.lu.se/nav_menu.html**
- Focus auf fachspezifische robotbasierte Dienste **http://www.ub2.lu.se/nav_menu.html#rosubj** und Qualitätsdienste **http://www.ub2.lu.se/nav_menu.html#qual**
  (Beispiele: EELS **http://www.ub2.lu.se/eel/**
  und All Engineering mit CrossROADS **http://www.ilrt.bris.ac.uk/roads/cross/**
  und Z39.50 Gateway, SBIG-Robotdienst Integration **http://borg.lub.lu.se/newegwindex.html**)

2 Hilfsmittel zur Wahl passender Suchdienste für bestimmte Zwecke

von Florian Seiffert

Viele Bibliotheken, die in Deutschland im Internet präsent sind, bieten Seiten zum Thema „Suche im Internet“ oder zumindest eine kurze Sammlung von Links auf Suchmaschinen, manchmal auch umfangreiche Linksammlungen anhand derer man sich wunderbar im Cyberspace verirren kann, an. Manchmal werden eigene Seiten gepflegt, oft gibt es Links auf die Seiten anderer Bibliotheken oder Verbünde.

Ein paar Beispiele:

- Suche im Internet an der Bibliothek der Uni Oldenburg
  http://www.bis.uni-oldenburg.de/suche1.html
- Webseiten der Internetquellen: Bibliographische Recherchen, Dokumentlieferdienste, allgemeine und fachliche UB Augsburg
  http://www.bibliothek.uni-augsburg.de/info/quellenfr0.html
- Die Virtuelle Bibliothek der ULB Düsseldorf
  http://www.rz.uni-duesseldorf.de/WWW/ulb/virtbibl.html
- FINT (Fachinformation im INTernet). Ein Service der Fachhochschulbibliotheken in Nordrhein-Westfalen
  http://www.fh-niederrhein.de/bib/fint/index.html
- IBIS (Internet Basiertes Informations System) im HBZ und in der UB Bielefeld
  http://www.hbz-nrw.de/ibis/ibis-hbz.html
- Deutsche Bibliotheken Online im HBZ
  http://platon.hbz-nrw.de/Harvest/brokers/Germlst
- Bibliothekarischer Werkzeugkasten im HBZ
  http://www.hbz-nrw.de/hbz/toolbox

Gibt es verständliche Erläuterungen? Seiten zu Suchstrategien? Ja, es gibt sie.

Ein paar Beispiele:

- Die drei wichtigsten Hilfsmittel zum Thema Suchen !!!
  http://www.hbz-nrw.de/~seiffert/weise-suche.htm
- Browsing and searching, UB Lund
  http://www.ub2.lu.se/nav_menu.html
- Suchstrategien im Internet der Stadtbüchereien Düsseldorf
  http://www.hbz-nrw.de/~untiedt/suchstr.htm
- Suchstrategien im Internet von Michael Otto
  http://ourworld.compuserve.com/homepages/michael_d_otto/

Schauen Sie auch mal auf ausländische Seiten:

- Browsing and searching pages, Lund *(best for experts, very comprehensive)*
  http://www.ub2.lu.se/nav_menu.html
- AskScott *(most interesting concept, not very comprehensive though)*
  http://www.askscott.com/
- Nueva Net Choose the best Search Engine For Your Purpose
  (*quite interesting, comprehensive)*
  http://www.nueva.pvt.k12.ca.us/~debbie/library/research/adviceengine.html
- Windweaver's SearchGuide *(quite interesting)*
  http://www.windweaver.com/searchguide.htm
- Decision Maker, PC Computing 10-17-97
  http://www.zdnet.com/pccomp/features/excl0997/sear/dm.html
- Fondren: Internet Searching QuickGuide
  http://riceinfo.rice.edu/Fondren/Netguides/quickguide.html
- Laura A. Guy Choosing a Search Engine
  http://dpls.dacc.wisc.edu/www_searchers_choice.html
- Features You Should Look for in a Web Search Service
  http://www.library.carleton.edu/websearch/best.html
- Univ. at Albany How to Choose a Search Engine or Research Database
  http://www.albany.edu/library/internet/choose.html
- Univ. at Albany Conducting Research on the Internet
  http://www.albany.edu/library/internet/research.html
- Univ. at Albany Searching the Internet: Recommended Sites and Search Techniques
  http://www.albany.edu/library/internet/search.html
- Ask Jeeves uses search technologies and natural language processing to answer any natural anguage query by directing you to an appropriate Web site.
  http://www.askjeeves.com

3 Beispiele avancierter Möglichkeiten zur Verbesserung der Suchergebnisse

Beispiele aus: "Advanced features of search services" (Lund): source types, fields, citation, filters, query expansion, controlled vocabulary, relevance feedback, sorting/clustering
http://www.ub2.lu.se/tk/advanced_features.html

3.1 Konzeptsuche / Erweiterung der Suchanfrage

von Traugott Koch

AltaVista Refine **http://altavista.digital.com**,
Euroferret **http://www.euroferret.com/**

3.2 Kontrolliertes Vokabular

von Traugott Koch

Beispiel: Medical World Search **http://www.mwsearch.com/**

3.3 Ranking

von Florian Seiffert

Ranking ist das Einstellen der Trefferreihenfolge. Optimieren Sie damit bei großen Treffermengen (die sich nicht immer ganz vermeiden lassen), welche Treffer zuerst angezeigt werden!

Ein Beispiel von Lycos:

- Lycos Power Panel **http://www.lycos.de/homepro.html**
  bei **http://www.lycos.de**
- Lycos Power Panel **http://lycospro.lycos.com/lycospro-powerpanel.html**
  bei **http://www.lycos.com**

3.4 Integration von Searching und Browsing

von Hans-Joachim Wätjen

- Roboter-basierte Suchmaschine und manuelles Verzeichnis als komplementäre, aber nicht integrierte Angebote
- Teilintegration: vom Browsing zur Suche
  - Webcrawler **http://www.webcrawler.com/**
    Während des Browsing kann zur Suche in der roboter-basierten Suchmaschine gewechselt werden.
- Teilintegration: von der Suche zum Browsing
  - Infoseek **http://www.infoseek.com**
    Related Topics aus den Channels werden zu jedem Suchergebnis der Suchmaschine angeboten. Während des Browsing in den Channels kann jederzeit eine Suche über die gesamte Datenbank gestartet werden.
- Weitergehende Integration
  - EELS **http://www.ub2.lu.se/eel/**
    Mit der EI-Klassifikation und dem EI-Thesaurus sind ca. 1.500 Dokumente erschlossen, so daß nach der Suche von Einzeltreffern das Browsing im Klassifikationsbaum oder mit weiteren Deskriptoren ermöglicht werden kann. Während des Browsing in

der Klassifikation und den zugeordneten Dokumenten kann eine Suche in der Klasse einschl. der Unterklassen oder in der All-EELS-Datenbank gestartet werden.

"All EELS" erschließt 120.000 mit einem Roboter gesammelte Dokumente. Browsing ist hier nur nach Domain, Titelalphabet oder in Hotlists möglich.

- GERHARD **http://www.gerhard.de**

  Die automatische Klassifizierung der mit Harvest gesammelten Dokumente (z. Zt. 900.000) nach der Züricher UDK erlaubt ein dreisprachiges Browsing über die 70.000 Klassen. Über die Suche in der Klassifikation können auch relevante Klassen gefunden werden. Von Einzeltreffern ist der Einstieg in die Klassifikation möglich. Eine gezielte Suche nach Dokumenten in Teilen des Klassifikationssystems ist noch nicht möglich.

4 Metadaten und Suchdienste

von Traugott Koch

Metadaten verbessern das Indexieren und die Wiederfindbarkeit der eigenen Informationen. Werkzeuge dafür sind:

- Unser DC Metadata Creator **http://www.lub.lu.se/cgi-bin/nmdc.pl**
- URN Generator **http://www.lub.lu.se/dc/urntest.pl**
- Suchbare Metadaten-Datenbanken
  Schweden **http://gungner.ub2.lu.se/cgi-bin/egwinit.pl/egwirtcl/metatargets.egw?swemeta**
  Dänemark **http://nwi.dtv.dk/cgi-bin/egwzgate/egwirtcl/screen.tcl/name=advanced_metadata&lang=eng&service=uinwi**
- Organisatorische Fragen: Wer macht in Deutschland Was, Wie mit Metadaten?
  - Metadata Server of the SUB Göttingen
    http://www2.sub.uni-goettingen.de
  - Metadatenprojekte und -initiativen an deutschen Bibliotheken
    http://www.dbi-berlin.de/projekte/einzproj/meta/meta03.htm
  - Descriptions of (Dublin Core) Metadata Projects (DC 5)
    http://linnea.helsinki.fi/meta/projects.html
  - Deutsche Uebersetzung (MPI Bildungsforschung, Berlin)
    http://www.mpib-berlin.mpg.de/DOK/metatagd.htm
  - META-LIB, Metadaten-Projekt dt. Bibliotheken (DDB, SUB Göttingen, BSB, DBI)
    http://www.dbi-berlin.de/projekte/einzproj/meta/meta00.htm
  - German Educational Resources Server / Deutscher Bildungs-Server
    http://dbs.schule.de/
  - SSG-Fachinformation (SSG-FI)
    http://www.sub.uni-goettingen.de/ssgfi/
  - IBIS
    http://www.ub.uni-bielefeld.de/ibis.html
  - Der virtuelle Medienserver beim BSZ (Elektronisches-E-Depot) Konstanz
    http://www.swbv.uni-konstanz.de/wwwroot/s71800_d.html
  - Mathematical preprints (Univ. Osnabrueck)
    http://www.mathematik.uni-osnabrueck.de/ak-technik/testinstall.html

  - Math-Net
    http://elib.zib.de/math-net/
- Wie kommen die Ergebnisse in die internationalen Dienste?

Welche Suchdienste nutzen Metadaten und in welcher Weise?

- Webseite **http://www.ub2.lu.se/~traugott/MDsearch.html**
- Beispiele
- Zukunftsperspektiven

5 Zur Qualität des Inputs und der Suche im Web

Grundsätzliche Probleme mit der Indexierung von Hypertext:

- Millionen von AutorInnen produzieren mehrere hundert Millionen Informationsschnipsel verschiedenster Art mit Absprachen (Standards).
- Die Grenzen eines Dokuments verschwimmen.
- Fehler werden toleriert und sogar prämiert (Layoutmißbrauch von HTML).

5.1 beim Angebot

von Hans-Joachim Wätjen

Auswertung des deutschen wissenschaftlich relevanten Web-Angebotes anhand des GERHARD-Projektes **http://www.gerhard.de**, die dokumentiert, wie schlecht das deutsche akademische Web-Angebot ist (Seiten ohne Titel, Metadaten, Fachgebiete mit magerem Angebot, HTML-Benutzung ...).

Von 875.000 gesammelten HTML-Dateien sind

- 8 % ohne Titel
- 32 % ohne Headings (nicht-strukturierte Seiten)
- 25 % ohne irgendeinen Link (isolierte Seiten)
- 0,11 % mit DC-Metatags (947 Seiten, davon 229 Uni Konstanz, 137 DBS, 121 TU München, 105 Uni Halle, 83 Uni Osnabrück, 74 Uni Rostock)
- 92 % ohne Autor- oder Address-Tags (anonyme Seiten)
- 24 % private (~) Seiten (für die Öffentlichkeit gedacht?)

5.2 bei der Abfrage: Folgerungen aus der Auswertestatistik der Bibliothekarischen Suchmaschine des HBZ

von Florian Seiffert

URL der Bibliothekarischen Suchmaschine des HBZ:
http://platon.hbz-nrw.de/Harvest/brokers/Germlst

5.2.1 Kurzer Überblick: Wie kommt die Bibliothekarische Suchmaschine des HBZ an ihre Inhalte?

- Wir benutzen das Harvest Information Discovery and Access System.
  http://harvest.transarc.com/
- Das Konfigurationsfile ("Wo soll die Maschine wie tief suchen?") erzeugen wir aus den Seiten "Deutsche Bibliotheken Online", die im HBZ gepflegt werden.
  http://www.hbz-nrw.de/hbz/germlst/
- Somit enthält die Suchmaschine einen (mehr oder weniger) vollständigen Volltextindex über alle Seiten, die auf deutschen bibliotheksrelevanten Servern aufliegen.

5.2.2 Eine beispielhafte Auswertung des Logfiles der Bibliothekarischen Suchmaschine des HBZ für Dienstag 27. Januar 1998:

Auswertungsstabelle siehe **http://www.hbz-nrw.de/~seiffert/Inet3tab.htm**

5.2.3 Minimalstatistik:

- 233 Abfragen pro Tag insgesamt.
- 149 davon mit keinem Treffer. Das sind 64% !

5.2.4 Folgerungen:

- Viele Personen benutzen die Suchmaschine offenbar, als sei sie das richtige Werkzeug zur Literatursuche. Das ist sie aber nicht! Dafür ist der KVK besser geeignet!
  http://www.ubka.uni-karlsruhe.de/kvk.html

  Aufklärung der Benutzer auf Tagungen **http://www.ub.uni-dortmund.de/lbkon3/tut4.htm**, in Seminaren **http://www.hbz-nrw.de/hbz/fortbildung/3-4-2.htm**, Informationen zu den Suchmaschinen!
- Die Benutzbarkeit und Oberfläche der Suchmaschine ist verbesserungsbedürftig!

  Das HBZ arbeitet daran! Das ist wie immer eine Frage von Prioritäten, Personen, die zuständig sind und die Zeit dafür haben (dürfen), von Chefs, die den Sinn sehen...
  Wir bitten um Geduld.

6 Empfehlungen zur Verbesserung von In- und Output

Was ist bei der Gestaltung von eigenen Web-Angeboten zu beachten, um die Qualität der Suchergebnisse zu verbessern?

- Vermeiden Sie "schlechtes" HTML!
  Siehe: Stefan Karzauinkat: Die goldenen Regeln für schlechtes HTML -
  http://www.karzauninkat.com/Goldhtml/goldhtm1.htm
- Nicht jedes Dokument sollte gleich weltweit publiziert und indexiert werden! Verwenden Sie den Robot Exclusion Standard. Siehe: Martijn Koster: Robot Exclusion Standard -
  http://info.webcrawler.com/mak/projects/robots/norobots.html

Zur Diskussion siehe auch:
http://www.kollar.com/robots.html

Für das Sperren/Indexieren von Einzelseiten bietet HTML4 gute Möglichkeiten. Siehe
http://www.w3.org/TR/REC-html40/appendix/notes.html#recs

- Erschließen Sie Ihre wichtigen und für die Öffentlichkeit bestimmten Dokumente mit Metadaten!
- Versehen Sie Ihr Dokument mit DC Metadaten mit Hilfe eines Creators, umso ausfuehrlicher je wichtiger Ihr Dokument ist.
- Unterschreiten Sie nicht das sinnvolle Minimum der wesentlichsten Metadaten Felder (Title, Creator, Subject, Keywords, Type, Identifier, Language).
- Eine weitere Verbesserung stellt ein Abstract dar (Description), das eventuell auch am Anfang des lesbaren Dokumentes (body text) stehen sollte. In den ersten Zeilen des Dokumentes sollte auf jeden Fall ihr Dokument eindeutig charakterisiert bzw. zusammengefasst sein.
- Halten Sie sich beim Inhalt der Metadatenfelder an die Hinweise der Hilfstexte. Mit abweichenden Inhalten im falschen Feld kann kein Suchdienst oder gar Nutzer etwas anfangen.
- Es empfiehlt sich im Augenblick, neben DC auch Metadaten fuer Alta Vista, Infoseek, Hotbot u.a. anzubieten (keywords, description), wobei fuer Alta Vista 1000 Zeichen nicht ueberschritten werden sollten.
- Formulieren Sie den Titel so, daß er, auch wenn aus seinem Zusammenhang gerissen, das Dokument eindeutig und beschreibt.
- Geben Sie Ihrem Dokument einen authorisierten URN, sobald diese in Deutschland angeboten werden.
- Beeinflussen Sie Dienstanbieter, standardisierte Metadaten korrekt einzusetzen und die Qualität ihrer Dienste damit zu verbessern. Das gilt vor allem fuer Suchdienste.
- Wenn Sie Ihr Dokument in mehreren Sprachen anbieten, nutzen Sie den HTML4 LINK tag, um die Browser darauf hinzuweisen.
  http://www.w3.org/TR/REC-html40/appendix/notes.html#recs