ZACK - eine Metasuchmaschine fuer Bibliotheken

Allegro in Kunst- und Museumsbibliotheken
Herzog-August-Bibliothek, Wolfenbuettel
7. Dezember 2000

by Wolfram Schneider
wolfram@schneider.org
http://wolfram.schneider.org

Zusammenfassung

- Zack ist eine Suchmaschine fuer Bibliotheksdatenbanken, die ueber das Z39.50 Protokoll ansprechbar sind.

- Das Ergebnis sind strukturierte Daten (MAB2), die in das eigene Bibliotheksystem uebernommen werden koennen.

- Bei der verteilten Suche wird gleichzeitig in mehreren Datenbanken gesucht. Dubletten werden als solche erkannt.

Inhalt

- Einfuehrung
  - Geschichte, Zielgruppe
  - Z39.50
  - Struktuierte Daten
  - verteilte Suche
- Detail
  - Normierung
  - Dublettenerkennung
  - parallele Suche
  - Fazit, Grenzen des Systems
- Vorfuehrung
  - Import von strukturierten Daten
  - Verteilte Suche

Einleitung

- Geschichte
  - Projekt KOBV am ZIB, von 1997 bis 2000
  - kooperativer, dezentraler Bibliotheksverbund
  - Bedarf an Dienstleistungen sofort

- Aufgabe
  - Web-Interface zur Datenbank ILTIS der Deutschen Bibliothek fuer Copy-Cataloging in Cottbus und Frankfurt/O.
  - Fremddaten (DNB)

- Zielgruppe
  - Katalogisierer, Bibliothekare, Retrokatalogisierung
  - Virtueller Verbund

Z39.50 Allgemein

- Kommunikationsprotokoll, ISO23950: Information Retrieval
- Client/Server System
- Datenbank-Esperanto: mit jedem Client einen Dialog mit jeder
  Datenbank fuehren koennen
- unabhaengig von Datenbank, lokale Anfragesyntax, Hersteller
- seit 1984 entwickelt, Library of Congress

Z39.50 Detail

- verbindungsorientiert (session/Sitzung)

- wichtigste Services
  - init Verbindungsaufbau
  - search Suche
  - present Ausgabe der Datensaetze
  - close Verbindungsabbau

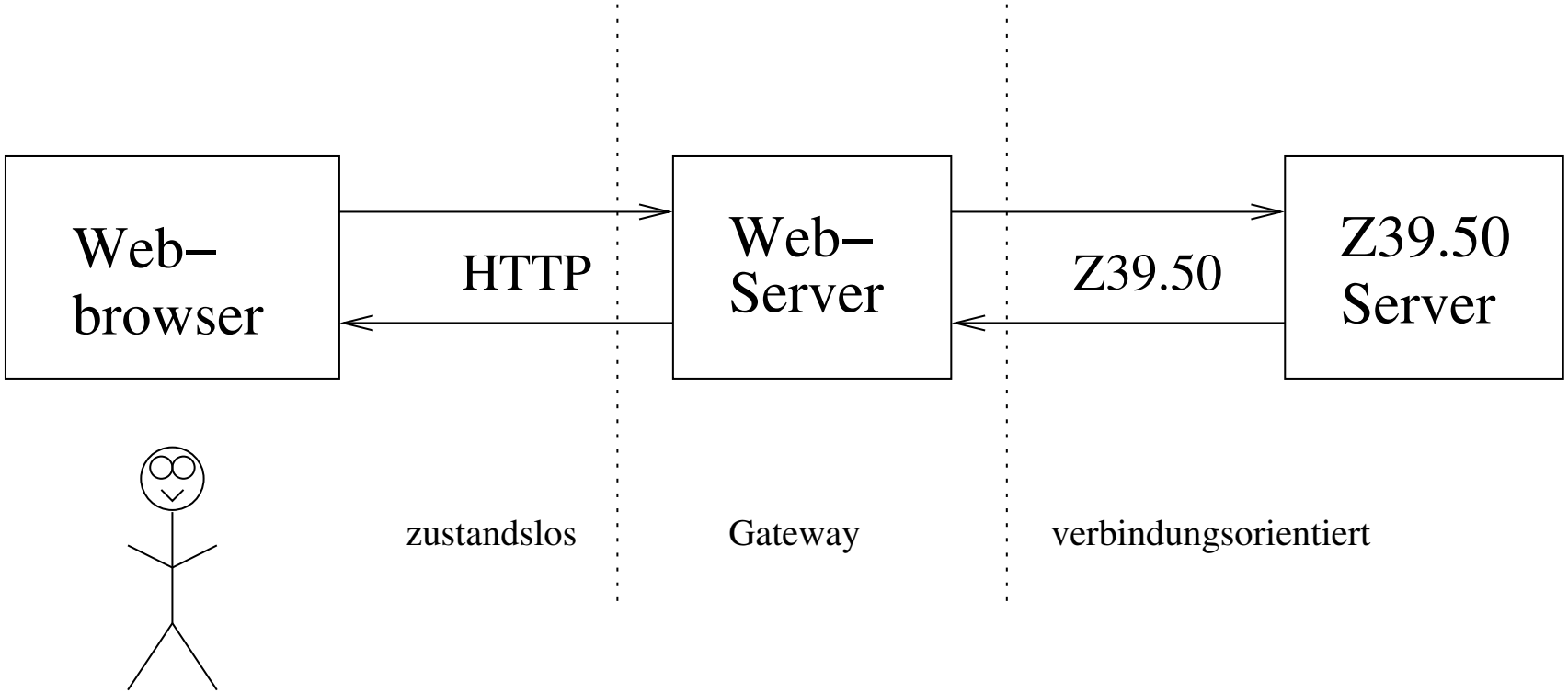
- weitere Services
  - scan Registersuche
  - sort Sortieroptionen fuer Ergebnisse
  - explain (Target profile)

WWW-Z39.50 Gateway

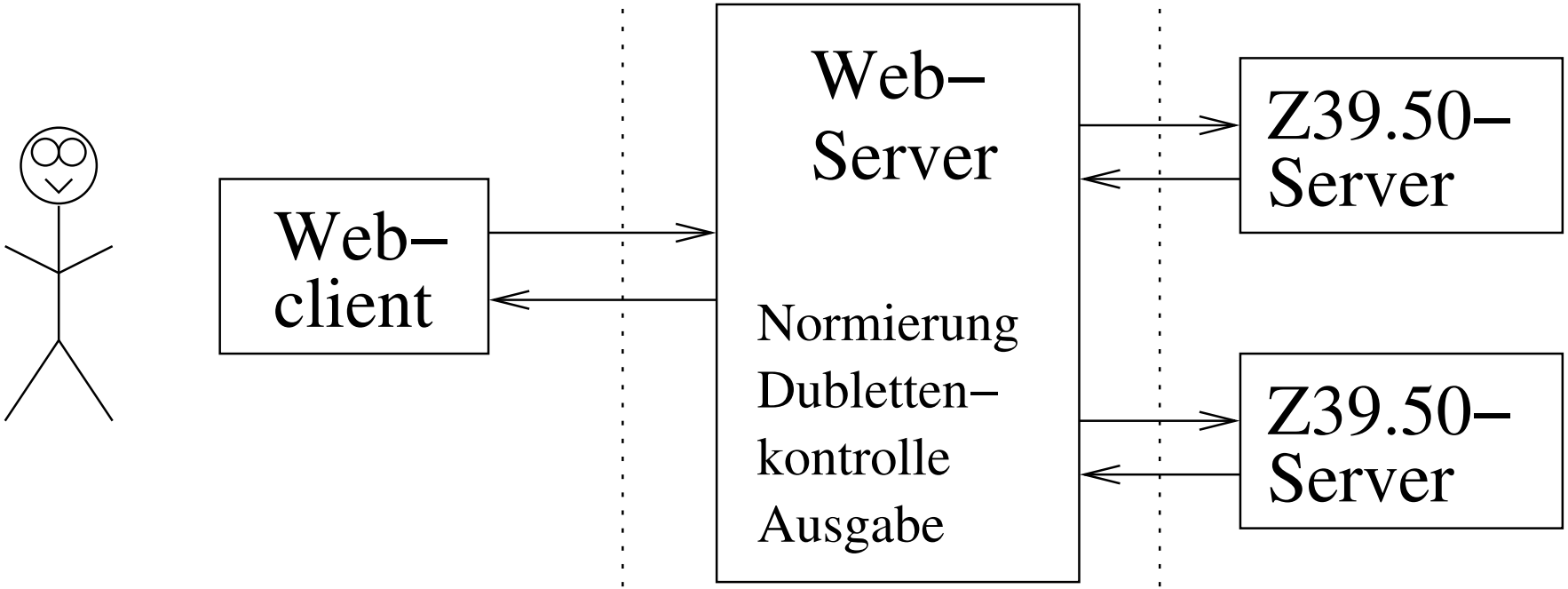
Vorteile von Z39.50 mit Vorteilen des Web verbinden

- Vorteile fuer die Benutzer
  - Web-Browser weit verbreitet
  - gewohnte Software, einfach zu bedienen
  - keine zusaetzlichen Kosten fuer Z39.50 Clients
  - keine Installation und Wartung von Z39.50 Clients
- Nachteile
  - eingeschraenkte Funktionalitaet
  - Performance

WWW-Z39.50 Gateway Modell



WWW-Z39.50 Gateway Verteilte Suche



Parallele Suche

- gleichzeitig Verbindungen aufbauen
- gleichzeitig suchen
- gleichzeitig Lesen
- max. 5-10 Datensaetze pro Sekunde laden
- z.Z. bis zu 16 dt. Datenbanken, insgesamt in mehr als 50 Mio Datensaetzen

Bibliotheksverbund Bayern, KOBV Berlin-Brandenburg, Gemeinsamer Bibliotheksverbund, TU Braunschweig, FH Potsdam, Uni Potsdam, Uni Duesseldorf, FH Brandenburg, Oeffentl. Bibl. Berlins, EUV Frankfurt/O, BTU Cottbus, Max-Planck-Inst. f. Bildungsforschung, Oeffentlichen Bibliotheken im Land Brandenburg (VOEB), Die Deutsche Bibliothek, Suedwestdeutscher Bibliotheksverbund, Nordrhein-westfaelischer Bibliotheksverbund (HBZ)

Normierung

was ist das: unterschiedliche Schreibweisen vereinheitlichen, die keine inhaltliche Bedeutung haben

warum notwendig

- Daten unterschiedlicher Herkunft
- Zeichensaetze
- Tippfehler
- unterschiedliche Katalogisierungsregeln

Beispiele

- Autor: "Dalitz, Wolfgang" -> "dalitz, wolfgang"
- ISBN: "ISBN 3-928861-23-9" -> "3928861239"

Dublettenerkennung

- Ziel: gleiche oder aehnliche Datensaetze erkennen
- nicht besser als ein Mensch
- Expertensystem (MYCIN)

- gleich: gleiche Attribute
- aehnlich: kleinere Abweichungen ignorieren
  - Tippfehler
  - +/- Seitenzahlen

Dublettenerkennung Expertensystem

- gewichteter Vergleich der Attribute
  - hohes Gewicht fuer Titel, Autor, ISBN
  - geringes Gewicht fuer Verlag, Jahr etc.
- Argumente, die fuer eine Dublette sprechen
- Argumente, die gegen eine Dublette sprechen
- positiver/negativer Schwellwert

Dublettenerkennung - Aufwand

- Aufwand: n*(n-1)/2
  - 10 Datensaetze -> 45 Vergleiche
  - 30 Datensaetze -> 435 Vergleiche

- Performance
  - Optimierung mit temporaerem Index
  - Cluster von aehnlichen Datensaetzen

- Grenzen
  - wieviel Toleranz erlaubt

Probleme - Technik

Probleme laufenden Betrieb - Technik

- instabile Z39.50 Server
- Betasoftware Z39.50 Server
- falsch konfigurierte Z39.50 Server
- keine Dokumentation
- nur wenige Server zum Testen
- hoher Wartungsaufwand

Probleme - Semantik Anfragen

Semantik von Z39.50 Anfragen

- Autorsuche
  - "wolfgang dalitz"
  - "dalitz, wolfgang"
  - "dalitz,wolfgang"
- ISBN Nummer
  - ISBN Nummer mit oder ohne Bindestriche (Normierung)

Probleme - Semantik Indexierung

Semantik Indexierung der lokalen Datenbanken

- Wortindex
- Wortgruppen (Phrasen)

- Welche Attribute fuer Indexierung
  - Schlagwoerter im Titelindex
  - Autor: Buecher von oder auch ueber einen Autor
- begrenzte Zahl von Indexen
- Norm- und Titeldaten in einem Index

Probleme - Austauschformate

- Z39.50 orientiert sich an USMARC
- MAB2 kaum unterstuetzt und dokumentiert, z.B Kurzformat
- MAB2 Hierarchien sind in Z39.50 nicht vorgesehen
- MAB2 wird unterschiedlich interpretiert
  - Beispiel: zwei ISBN Nummern - zwei Felder?
- Regelwerke
  - RAK, RAK-WB

Nutzer des Systems

fuer die taegliche Arbeit beim Katalogisieren

- BTU Cottbus, EUV Frankfurt/Oder, Uni Potsdam, FU Berlin, Uni Bonn, RWTH Aachen, ZIKG Muenchen
- (werk-)taeglich ca. 40 Nutzer und 1500-2000 Suchanfragen

Einsatz der Software bzw. Teilen der Software

- ZIB/KOBV Zentrale
- TU Braunschweig
- Hochschulbibliothekszentrum NRW (HBZ)
- Universitaet des Saarlandes (SWB)
- Suedwestdeutscher Bibliotheksverbund (SWB)

Wunschliste Ausgabe

- Benutzerfreundlicheres Design
- Text-Ausgabe der Treffer besser formatieren
- Super-Merged-Record: Datensaetze mischen (bessere Schlagwoerter)
- XML Export der Datensaetze
- Zeichensatz: latin2, Unicode

Wunschliste Performance

- Dublettenkontrolle im Hintergrund mit unvollstaendigen Daten
- parallele Dublettenkontrolle auf Multiprozessorsystemen
- Dublettenkontrolle mit unterschiedlichen Formaten (USMARC + MAB2)
- Aufnahme weiterer Z39.50 Server
- stehende Verbindungen zu Z39.50 Servern

Fazit - was geht

Fazit verteilte Suche mit Dublettenkontrolle

- Verteilte Suche ist moeglich
- Antwortzeiten sind akzeptabel (<10 Sekunden)
- bessere Ergebnisse als bei der Suche in nur einer Datenbank
- Kurztrefferliste ist kuerzer (30-50%) und uebersichtlicher

Fazit - Grenzen des Systems

inhaltlich:

- nur so gut wie die Bibliotheksdatenbanken
- nur so gut wie die Z39.50 Server
- MAB2 Hierarchien nur eingeschraenkt

technisch:

- max. 50 Datenbanken gleichzeitig
- max. 500 Datensaetze bei Dublettenkontrolle
- nur so schnell wie der langsamste Z39.50 Server
- geschaetzt 10.000 Anfragen/Tag

Software

CGI-Scripte, Module:

- Perl5

Z39.50:

- YAZ Toolkit

Sonstiges

- CVS, recode, Apache, LaTeX, Magicpoint, xfig, xv

Betriebssystem & Hardware:

- Solaris, FreeBSD
- mind. 32MB RAM, 100MHz CPU

Diplomarbeit

Wolfram Schneider: Ein verteiltes Bibliotheks-Informationssystem auf Basis des Z39.50 Protokolls. Juli 1999.

Eingereicht als Diplomarbeit an der TU Berlin,
Fachbereich Informatik.

Betreuer:

- Josef Willenborg (ZIB)

Gutachter:

- Prof. Dr. Erhard Konrad (TU Berlin)
- Prof. Dr. Martin Groetschel (TU Berlin/ZIB)

Literatur

- Frei zugaengliche Version von Zack, TU Braunschweig, http://www.biblio.tu-bs.de/zack/

- Diplomarbeit in HTML, PostScript, PDF, ASCII, PalmDoc http://wolfram.schneider.org/lv/diplom/

- Diplomarbeit auf Papier: ZIB Preprint SC 99-21. ISSN 0933-7911.

Copyright (c) Dec 2000 Wolfram Schneider.
All rights reserved.