

TUSTEP und das Netz

Eine Beispielanwendung für die Recherche in WWW-basierten Datenbanken

TUSTEP kann schon seit langem mit strukturierten Daten (Datenbanken) umgehen. Es bietet mächtige, vom Benutzer fast beliebig konfigurierbare und trotzdem effiziente Suchmöglichkeiten für Texte aller Art. Es ist außerdem ein ideales Werkzeug zur Wartung und Publikation von Dokumenten mit inhaltlichen Auszeichnungen.

Auf diese Leistungen von TUSTEP kann jetzt auch über das WWW zugegriffen werden.

Anforderungen

Eine Suchmaske sollte es dem Nutzer so einfach wie möglich machen, auf die von ihm benötigten Informationen zuzugreifen. Dieses an sich selbstverständliche Postulat führt zu einer Reihe von Anforderungen, die von vielen Produkten am Markt nicht oder nur schlecht erfüllt werden können:

- Die erzeugte Suchmaske und die Präsentation der Ergebnisse müssen frei gestaltet und so dem Nutzerprofil angepaßt werden können.
- Die Suche an sich muß durch »intelligente Unschärfe« den Erwartungen des Nutzers Rechnung tragen.
- Masken sollten interaktiv vom Benutzer veränderbar sein, um ihm so Einfluß auf die konkrete Ausgestaltung der Suche und die Ergebnispräsentation zu geben. Er darf aber

deshalb vor lauter technischen Details nicht den Überblick verlieren.

Vom Gesichtspunkt des Entwicklers kommen noch weitere Wünsche hinzu:

- Die im Netz präsentierten Daten sollten sich mit minimalem Aufwand auch »zweitverwerten« lassen (Publikation in Buchform, CD-ROM-Versionen, Transfer in andere Datenbankformate . . .).
- Die Programmierung sollte möglichst einfach und schnell gehen.

Zusätzlich sind natürlich Sicherheitsüberlegungen hochrelevant: CGI-Skripte sind potentielle Sicherheitslücken, die es u. U. »Hackern« erlauben können, in das lokale Netzwerk einzudringen und Schaden anzurichten.

Der neue OPAC der ZDV-Bibliothek

Der neue, TUSTEP-basierte OPAC (siehe Abb. 1) hat das Ziel, obigen Anforderungen gerecht zu werden. Die Benennung der Textfelder ist selbsterklärend und entspricht den üblichen Konventionen: *Nachname*, *Vorname* etc. Im *Titel*-Feld können mehrere Stichwörter angegeben werden, die durch ODER verknüpft werden. Sollen Stichwörter zusätzlich oder aber nicht vorkommen, können diese im zweiten Feld angegeben werden. Anstelle eines Stichwortes kann auch eine in Anführungszeichen eingeschlossene Wortfolge angegeben werden. Im Gegensatz zu anderen Suchmaschinen können für *Jahr* auch Zeiträume angegeben werden. Die Möglichkeit, die Suche auf bestimmte Signaturen einzuschränken, ist für die ZDV-Bibliothek wichtig, da sich nur so gezielt auf den weit über Tübingen hinaus bekannten

Sammelschwerpunkt *EDV in den Geisteswissenschaften*, dessen Signaturen mit G beginnen, zuzugreifen läßt. In anderen OPACs verwendete Suchkriterien wie *Verlag* oder *Erscheinungsort* schienen für den konkreten Anwendungszusammenhang verzichtbar und wurden daher weggelassen. Bei Bedarf könnten sie natürlich problemlos berücksichtigt werden.

Der Benutzer kann über *Treffer pro Seite* die Art der Ergebnispräsentation beeinflussen. Insbesondere kann er mit *Alle Treffer* bequem alle Suchergebnisse in einer Datei sammeln und zur weiteren Verwendung abspeichern. Über *Ausgabeformat* kann er noch bestimmen, ob bei unselbständigen Publikationen auch die Quelle mit ausgegeben werden soll (default) oder ob Autor und Titelangabe reichen. Ebenso kann mit *Publikationsart* die Suche auf Monogra-

phien beschränkt werden (normalerweise werden auch Artikel in Sammelbänden o. ä. mit in die Suche einbezogen).

Ein Beispiel für das Ergebnis einer Suchanfrage (hier die Suchanfrage aus Abb. 1) enthält die Abb. 2.

Erlaubt es TUSTEP im Gegensatz zu vielen anderen Produkten bereits, Suchmasken mit relativ geringem Aufwand exakt auf die vorhandenen Bedürfnisse zuzuschneiden, so spielt es seine wirkliche Stärke in der »intelligent unscharfen« Suche aus, die sich hier hinter dem Punkt *Exaktheit* verbirgt. Hier kann der Benutzer zwischen drei Optionen wählen: *Exakte Übereinstimmung* findet nur in der Schreibung identische Ausdrücke, während *Normalisierte Suche* versucht, auf die Erwartungen des Benutzers einzugehen, indem in der Sprache verbreitete äquivalente Schreibweisen berücksichtigt werden. Im konkreten Beispiel findet die Suche nach »Tübinger« korrekt auch die Schreibung »TUEbinger« – allgemeiner gesprochen, es werden Diakritika für die Suche ignoriert und Umlaute sowie ae/oe/ue-Kombinationen auf ihre Grundbuchstaben zurückgeführt. Sowohl in den Daten als auch in der Ausgabe findet sich natürlich die Originalform.

Grundgedanken der Realisierung

Der neue OPAC ist intern ein TUSTEP-Makro, das über ein CGI-Skript aktiviert wird. CGI (*Common Gateway Interface*) ist ein Mechanismus, der es einem irgendwo im WWW laufenden Clienten (normalerweise einem Browser) erlaubt, ein dazu freigegebenes Programm auf dem WWW-Server zu starten, ihm Parameter zu übergeben und die Ausgabe entgegenzunehmen. Damit ist CGI die Grundlage für interaktives Arbeiten im WWW und eine der Ursachen für dessen immensen Erfolg. Für Details sei hier auf zahlreiche Publikationen zu diesem Thema verwiesen (gleich eine Gelegenheit, einmal im neuen OPAC zu stöbern!)

Nur soviel sei gesagt: TUSTEP erzeugt eine vom Entwickler frei zu gestaltende HTML-Suchmaske, in die der Benutzer seine Anfragen eingeben kann. Dem hierzu im Batch-Betrieb laufenden TUSTEP werden diese Daten dann vom Browser per CGI in einer fixen Umgebungsvariablen QUERY_STRING übergeben. Das Makro analysiert deren Inhalt und nutzt die neuen Makro-Funktionen zur Datenbankabfrage (die Inhalt eines BI-Artikels in einer der nächsten Ausgaben sein werden), um passende Einträge zu finden. In Abhängigkeit von

Die gewünschte Unschärfe kann mit minimalem Aufwand den konkreten Anforderungen angepaßt werden und damit z. B. unterschiedliche Schreibungen im englischen und amerikanischen Englisch (*-ize* vs. *-ise* u. ä.) berücksichtigen. Der Entwickler kann hier nahezu unbeschränkt auf Kundenwünsche eingehen.

Zusätzlich zur *Normalisierten Suche* gibt es noch die Option *Fuzzy-Suche*, die in der hier programmierten Form den gesamten Datenbestand nach Einträgen durchsucht, die sich maximal um einen Buchstaben von dem Suchbegriff unterscheiden (z. B. findet die Suche nach »tastep« neben »TUSTEP« auch »tasten«). So können um den Preis höherer Rechenzeit auch dann Publikationen lokalisiert werden, wenn der Benutzer sich unsicher über die korrekte Schreibweise etwa des Autorennamens ist. Diese Funktionalität, die wiederum weitgehend anpaßbar ist (z. B. phonetische Suche, Soundex-Suche ...), ist u. a. für Recherchen in Textkorpora interessant, in denen entweder eine nicht-standardisierte Orthographie verwendet wird oder viele Erfassungsfehler vorliegen (z. B. bei aus OCR resultierenden Texten). Die verschiedenen Methoden lassen sich kombinieren.

Benutzereinstellungen wird dann HTML-Code erzeugt und ausgegeben, über den der Entwickler völlige Kontrolle hat.

Aufgrund der »Sandbox«-Struktur von TUSTEP, die die Kommunikation mit dem dahinter stehenden Betriebssystem nur über genau definierte Schnittstellen ermöglicht, ist es nach menschlichem Ermessen unmöglich, daß ein Eindringling Zugriff auf nicht explizit freigegebene Systemressourcen erhalten kann.

Falls diese knappen Ausführungen mehr Fragen offen lassen als Antworten geben, sei auf den vom 15.-17. 1. 1999 in Baubeuren geplanten Workshop zum Thema WWW – CGI – TUSTEP: *Recherchieren in strukturierten Daten* verwiesen, in dem Ihre Neugier befriedigt wird (TUSTEP-Grundkenntnisse sind nützlich, aber keine Vorbedingung). Ihre Anmeldungen richten Sie bitte an tustep@zdv.uni-tuebingen.de (Tel.: 07071/29-70347 oder 29-70210).

Für Rückfragen hierzu und für Anregungen zum neuen OPAC stehe ich Ihnen gerne zur Verfügung.

Marc Wilhelm Küster
kuester@zdv.uni-tuebingen.de

Abb. 1: Beispiel für eine Suchanfrage

OPAC der ZDV-Bibliothek

Nachname:

Vorname:

Körperschaft:

Titel:

UND ABER NICHT

Jahr: bis

ISBN:

Signatur: (EDV in den Geisteswissenschaften: Signatur G*)

Treffer pro Seite: 1 5 10 20 50 100 Alle Treffer

Exaktheit: Exakte Übereinstimmung Normalisierte Suche
 Fuzzy-Suche (nur Titel und Nachnamen)

Ausgabeformat: Vollständige Einträge Bei Artikeln nur Autor und Titel

Publikationsart: Alle Veröffentlichungen Nur Monographien

Sie können in Ihrer Suche Wildcards benutzen. Die TUSTEP-basierte Suchmaschine akzeptiert "*" (auch am Wortanfang und im Wortinneren) für beliebig viele beliebige Zeichen und "?" für ein beliebiges Zeichen. Weitere Erläuterungen zur Benutzung des OPAC finden Sie in der [Hilfe](#).

Zusätzliche Informationen über die ZDV-Bibliothek finden Sie auf deren [Homepage](#).

Abb. 2: Das dazugehörige Ergebnis

Weiter

Testversion

1 Treffer zu Ihrer Suchanfrage

Treffer 1

Autor Castrillo Benito, Nicolas

Titel Aplicacion del tratamiento de textos a la investigacion programmas TUSTEP (TUEbinger System von TExtverarbeitungs-Programmen). (Traducción del original alemán)

Verlag Valladolid: Universidad 1992

Umfang VI, 474 S.

Serie Lingüística y Filología, Bd. 14

ISBN 84-7762-284-1

Sign. G CAS