Wer suchet - der findet ?

oder

Die Überwindung der sprachlichen Grenzen bei der Suche in Volltextdatenbanken

DOKUMENT 98

(Franz Reinisch)

Wer suchet - der findet ?

oder

Die Überwindung der sprachlichen Grenzen bei der Suche in Volltextdatenbanken

Das Volumen an "abfragbaren" Texten ist in den letzten Jahren insbesondere durch den Einsatz von Volltextdatenbanken (und dem Internet-Boom) enorm angewachsen.
"Alles verfügbare Wissen (soweit es in maschinenlesbarer Form vorliegt wird, so gut es geht, in Datenbanken abgelegt. Der Nutzen aus solchen Volltextrecherchen (oder Recherchen an sich) ist aber in den letzten Jahren nur unwesentlich gestiegen.

Was sind die Ursachen dieses **Wissensstaus**??

- 80% aller Suchanfragen sind Einwortsuchen
  Logik, Abstandsoperatoren, Wichtung nützen hierbei nichts
- Suchmaschinen haben keine Kenntnis des Sprachumfeldes, gültiger Abkürzungen, Fachbegriffe usw.
- Suchmaschinen haben kein "Verständnis" für den Text

Die Zeit vor den Volltextdatenbanken:

Der frühe Einsatz von Bibliotheksdatenbanken war geprägt von Speicherknappheit und Leistungsschwäche. Die Suche nach Datenbankeinträgen wurde lediglich durch die Eingabe von Beschreibungswörtern (Deskriptoren) unterstützt. Eine Suche im Volltext war daher zur damaligen Zeit noch undenkbar. Der Online-Zugriff auf große Datenmengen war außerdem nur den professionellen Rechercheuren und Bibliothekaren vergönnt. Sie konnten aber bereits früher durch ausgeklügelte Abfragen exakte Antworten aus den gut gepflegten Datenbanken erwarten und mußten für diese Ergebnisse auch entsprechend viel Geld bezahlen.
Diese Datenbanken waren intellektuell beschlagwortet (indexiert). Das heißt, es hat sich ein intelligenter Mensch daran gemacht, den Dokumenteninhalt zu lesen, zu verstehen und mit eigenen Beschreibungsworten (Deskriptoren) zu versehen. Diese Deskriptoren wurden in einen Index aufgenommen und haben extrem schnelle Suchergebnisse geliefert.

Dadurch konnte der Text aus dem Grundgesetz der Bundesrepublik Deutschland : "*Die Würde des Menschen ist unantastbar"* auch durch das Suchwort **"Menschenrecht"** gefunden werden, wenn der zuständige Indexierer (Mensch) sich gedacht hat, das Grundgesetz habe mit **Menschenrecht** zu tun.

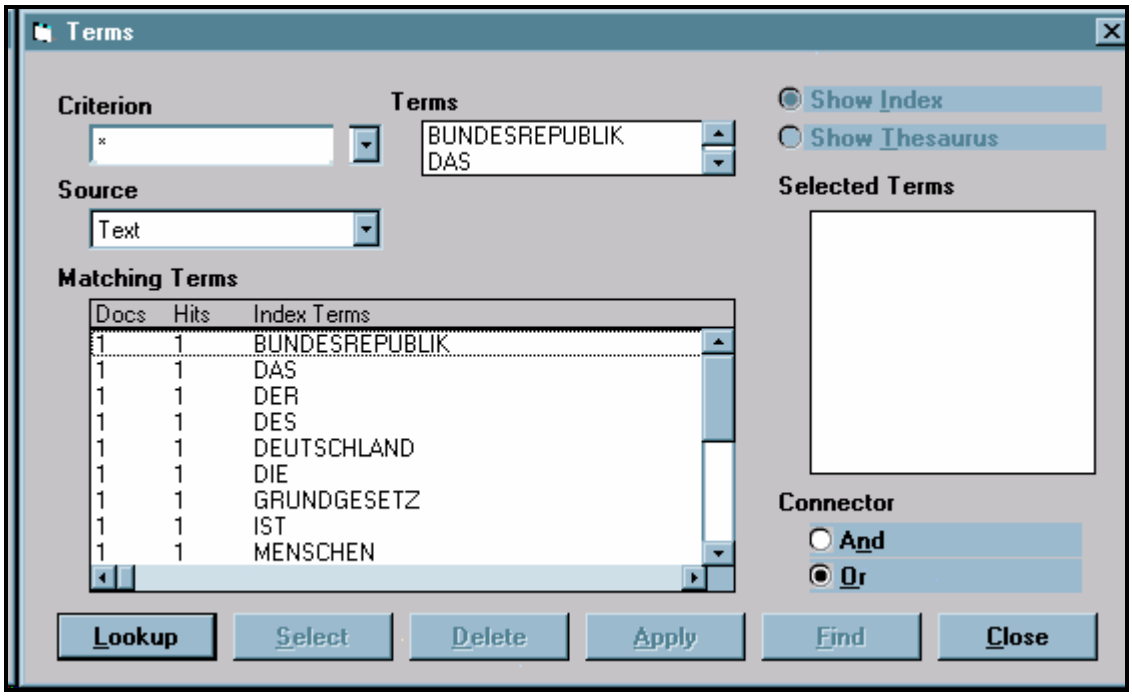
Die **intellektuellen Beschlagwortung** (durch Menschen) nach Thesaurusvorgaben, wie sie in hochqualitativen Bibliotheksdatenbanken und Archiven auch heute noch betrieben wird, **ist aber praktisch unbezahlbar** und bei den enormen Datenmengen auch kaum mehr leistbar.

Die Zeit der Volltextdatenbanken:

Die technische Weiterentwicklung der Datenbanktechnologie sowie das günstige Preis-Leistungsverhältnis des Speicherplatzes haben zum verstärkten Einsatz von Volltextdatenbanken geführt.
Der Wunsch, durch Volltextindexierung auf die Normierung und Klassifizierung von Deskriptoren bzw. den Einsatz von Thesauri verzichten zu können, hat sich leider als grobe Fehleinschätzung erwiesen.

Bei der **maschinellen Indexerschließung** werden alle Wörter (ggf. ausgenommen Stopworte) im Dokument indexiert. Bei wirklich großen Datenmengen erhält man daher meist zu viele Treffer, oder schließt zu viele Dokumente aus. **Das Suchergebnis ist unbestimmt**. Der Benutzer weiß nie, warum bestimmte Treffermengen zustande gekommen sind.

"das Grundgesetz der Bundesrepublik Deutschland"
Die Würde des Menschen ist unantastbar.



(Abb. 1: vollständiger, nicht aufgearbeiteter Index, Quelle: BASISdesktop)

Dieser Text ist natürlich in heutigen Volltextdatenbanken durch Eingabe der Suchwörter: "Die *Würde des Menschen"* auffindbar. Auch Schreibungenauigkeiten lassen sich durch diverse Mechanismen wie Fuzzy-Suche oder Pattern-Recognition überbrücken.

Um den Satz aber mit dem Suchwort: **"Menschenrecht"** zu finden bedarf es linguistischer Methoden.

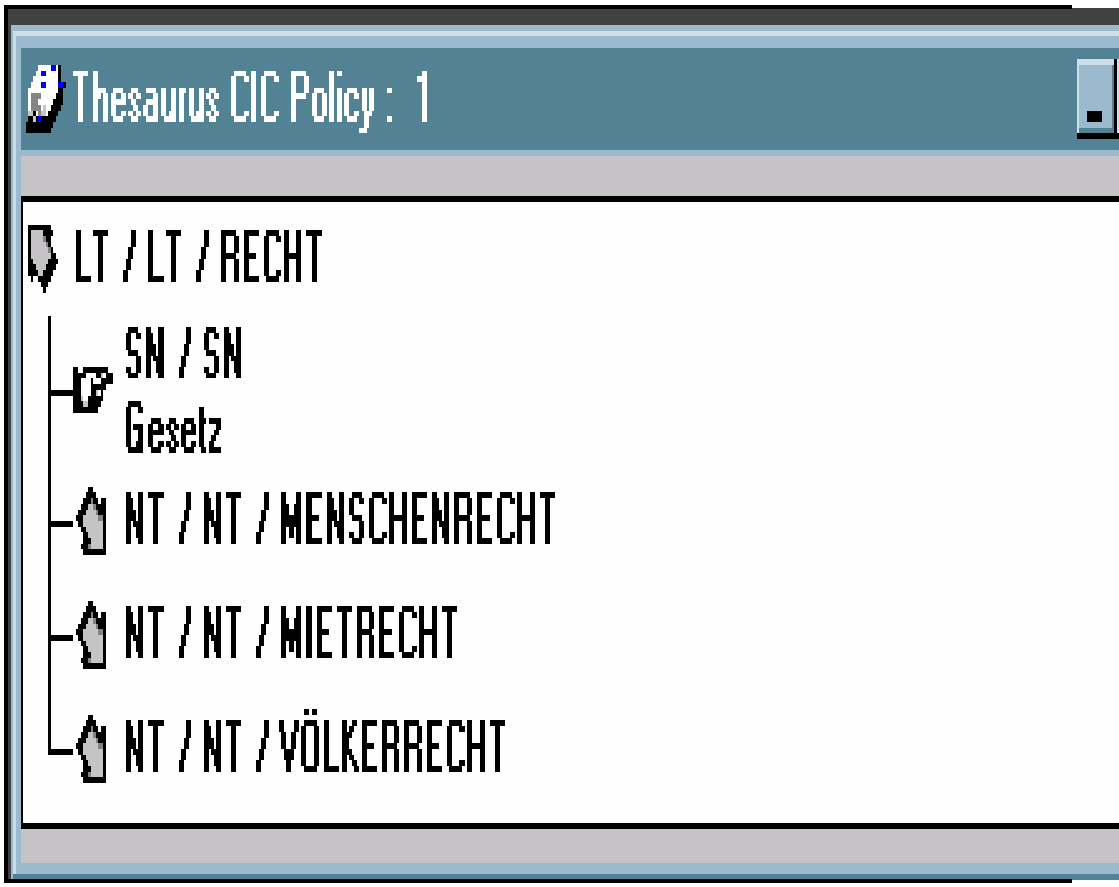
Was kann die deutsche Sprache zusätzlich noch an Schwierigkeiten einbringen?

- Umlautproblematik, deutscher Zeichensatz, vor allem bei (alten) 7-bit ASCII Anwendungen
- Das Verhältnis Grundform/Vollform ist in der deutschen Sprache sehr vielfältig. Ein dt. Adjektiv kann bis zu 50 verschiedene Vollformen haben.
- Zusammengesetzte Wörter — Gastgewerbekonzessionsprüfung; nicht aber: Frühstück
- transitive Verbindungen — Haus- und Hoflieferanten
- Groß- und Kleinschreibung — der DER (deutsches Eisenbahnreisebüro)
- Schreibvarianten — Foto, Photo, Philosoph, Filosof
- Sprachumfeld — Techniker haben eindeutige Wortverwendung; Mediziner/Literaten haben ändernde Wortverwendung

Die drei syntaktischen Wortkategorien: Nomen, Verb und Adjektiv haben sehr komplexe Morphologien. Grundformen sind nicht nur in flektierten Vollformen sondern auch im Innern deutscher Komposita versteckt und damit für die Volltextrecherche recht unzugänglich.

Zur Verbesserung der Suchergebnisse wendet man unterschiedliche Methoden an:

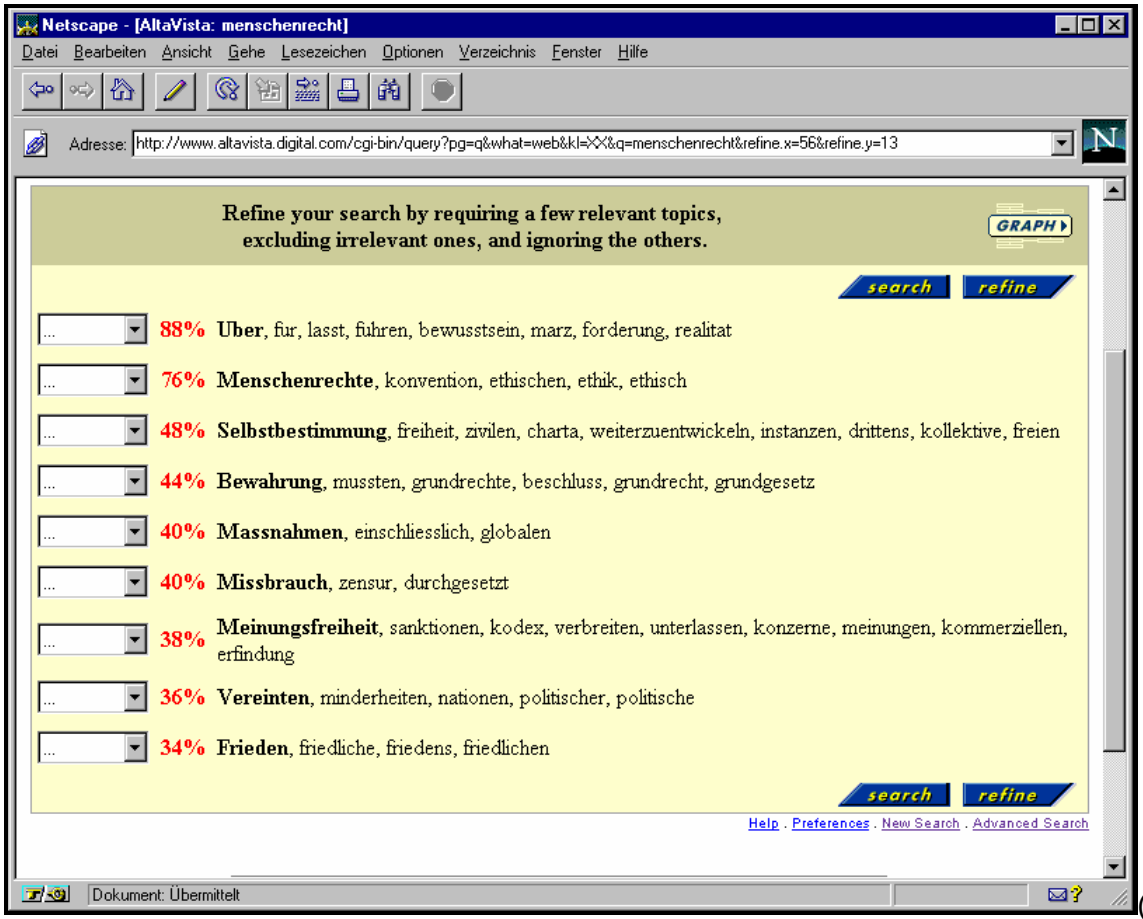
- Thesaurus: aufbauend auf einen strukturierten Wortschatz, z.B.:



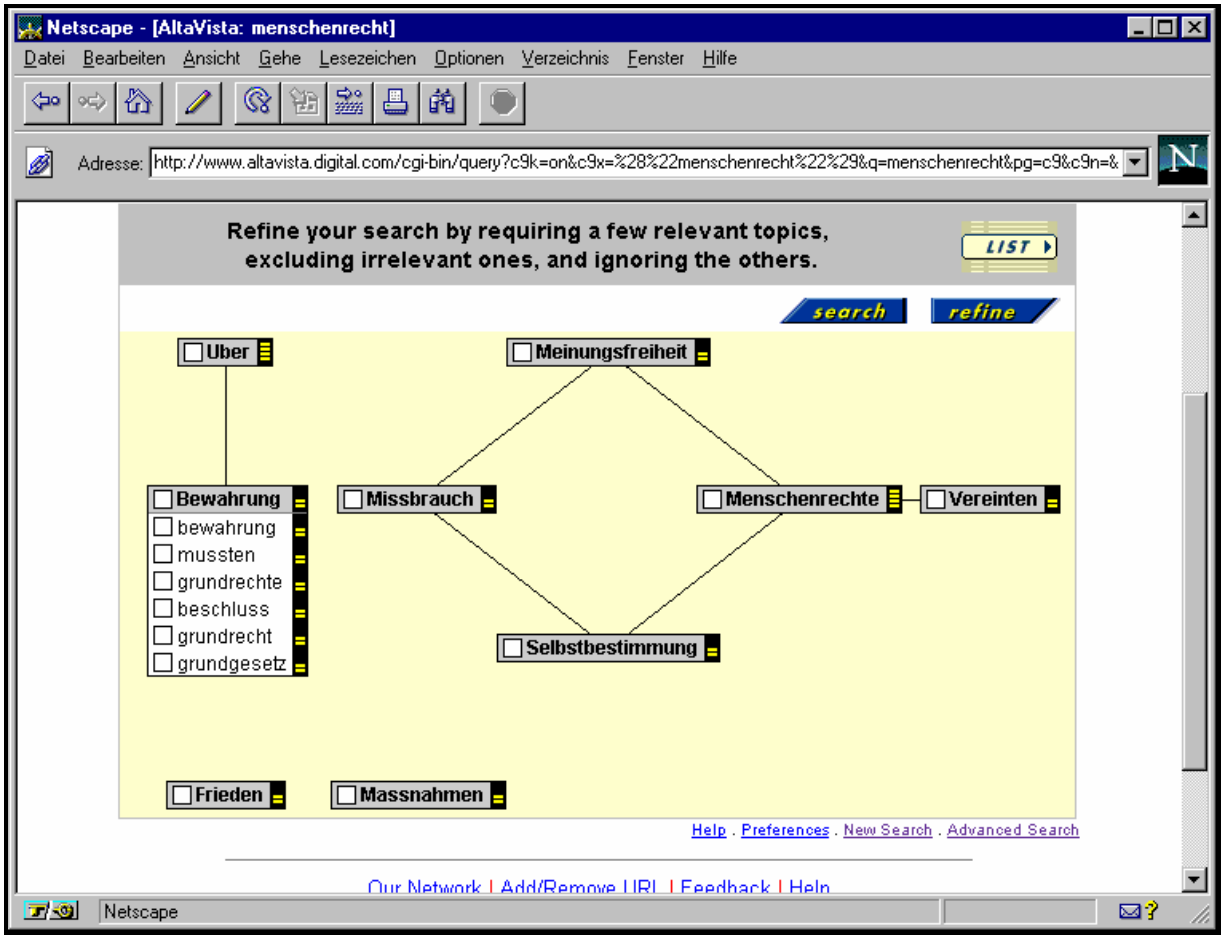
(Abb. 2: Thesaurus Manager, Quelle BASISthesaurus)

Für Volltextdatenbanken müßte ein Thesaurus die GESAMTE deutsche Sprache abbilden.

- Statistik: Überprüfung von Worthäufigkeit und Wortumgebung, Begriffsräume (überschreiten der Unsinns-Schwelle)

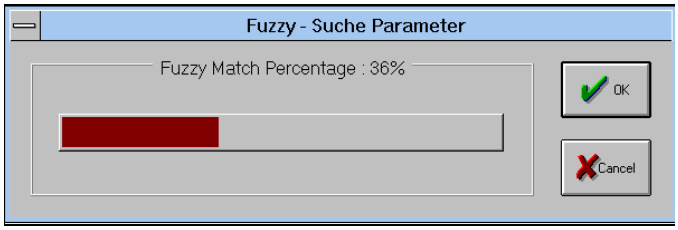


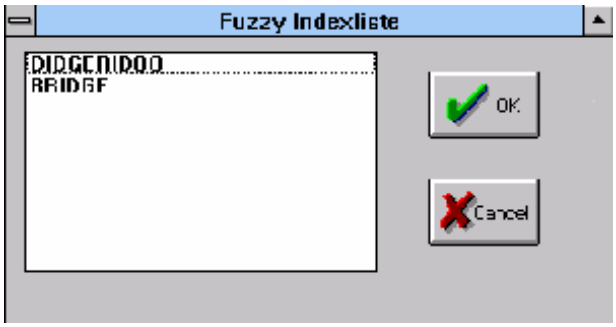
(Abb. 3a: Wortumgebung und Begriffsraum, Quelle: Altavista MAP)



(Abb. 3b: Begriffsraum, Quelle: Altavista MAP)

- Fuzzy Suche, Pattern-Recognition





(Abb. 4: Fuzzy Parameter Quelle: TRIPvbx)

Ein Lösungsweg:

Ein möglicher Ansatz zur Überwindung dieser Probleme ist der Einsatz linguistischer Softwarekomponenten.

Vor einigen Jahren schon wurde im Zusammenhang mit dem Datenbanksystem GOLEM die linguistische Softwarekomponente PASSAT eingesetzt. Ein für die damaligen technischen und linguistischen Bedingungen gelungener Versuch.

Auf Basis einer manuell erstellten Vergleichswortliste (VWL) wurden die Wörter eines Dokumentes, die in dieser Vergleichswortliste vorhanden sind oder sich auf diese reduzieren lassen, in den Index aufgenommen. Weiters wurden Komposita in ihre linken und rechten Teile zerlegt.

Die ABC System GmbH Bad Camberg hat 1996 einen Datenbankmodul entwickelt, der als Grundanforderung "die Leistungsfähigkeiten der PASSAT-Lösung" hatte und nun als Grundlage für weitere linguistische Lösungen genutzt wird.

Das Lexikon CISLEX:

Als zentrale Komponente für die Indexierung dient ein umfangreiches Lexikonsystem (CISLEX) der Universität München (Centrum für Informations- und Sprachverarbeitung Prof. Dr. Franz Guenthner).

Inhalt des Lexikons:

- Einfache Formen
  Borte;fem;NS0;NP4
  leiden,.VST4#<leid,leid,litt,litt,litt>

- Vollformen (%Wortform,Grundform.Kode:merkmalsbündel)
  %zwinget,zwingen.VST1:2mGc
  %zwangt,zwingen.VST1:2mVi
  %zwang,zwingen.VST1:1eVi:3eVi
  %zwingest,zwingen.VST1:2eGc

- Komplexe Formen

- deutsche Eigennamen, UNO-Länderverzeichnis
  sedlmeyer:sedlmeyer.EN:X

- Sonderformen
  do328,Do328.AK:X
  dornier328,Dornier328.EN:X

- Fachwörter
- Phonetisiertes CISLEX (in Vorbereitung)

Vollständigkeit des Lexikons:
Am Beispiel einer Auswertung eines Textkörpers ergibt sich folgende Verteilung von nicht erkannten Wörtern:

Vornamen:	5%
Nachmanen:	20%
geograph. Namen:	32%
Firmennamen:	1%
sonst. Namen:	10%
fremdsprach. Wörter:	16%

Tippfehler:	6%
Abkürzungen:	4%
Dialekt Wörter:	4%
einfache oder komplexe Formen	2,3%

Fast dreiviertel der nicht erkannten Formen stammen aus dem Eigennamenbereich.

Weiteres Beispiel: Auswertung des Altavista-Index mit ca. 50% Eigennamen Anteil!!

Zielsetzung des Lexikon:
Letztendlich ist eine semantische Klassifikation zur Indexierung von Objekten mit Eigenschaften und Beziehungen zu anderen Objekten im Dokument das Ziel.

Die Performanz des CISLEX ist trotz des enormen Umfanges mit 500.000 Wörtern/Sek. extrem hoch.

Die Vorgehensweise bei der "Normalisierung" der Wortformen im Morph-Server:

1. Lematisierung	Waldbrände wird Waldbrand
2. Kompositazerlegung:	Waldbrand wird Wald + Brand
3. Vollformerzeugung	aus dem Lema Wald + Brand wird Waldbrand, Waldbrände, Waldbrandes usw.

4. Umlautauflösung
5. Auflösung von Mehrwortbegriffen
6. Auflösung transitiver Verbindungen
7. Zusammenführung von getrennten Wörtern (harte Silbentrennung)

Passiver linguistischer Ansatz:

Einige Volltextdatenbanken und auch File-Index Systeme verwenden linguistische Komponenten, um einen unbearbeiteten Index zu durchsuchen.

• Indexierung sämtlicher Worte	die Würde des Menschen
• Eingabe Suchwortes	Mensch
• Analyse des Suchwortes	
• Erweitern der Suche um Flexionen (meist nach Regeln)	Mensch, Menschen

Diese "Schrottschußvariante führt naturgemäß zu unvorhersehbaren Ergebnissen.

Aktiver linguistischer Ansatz:

• Analyse jedes einzelnen Wortes	die Würde des Menschen
• Indexierung der analysierten Worte	Würde, Mensch
• Eingabe Suchwortes	Mensch
• Analyse des Suchwortes	
• Erweitern der Suche um Flexionen (mit Hilfe des CISLEX)	Mensch, Menschen

Zusätzlicher Ansatz:

- Zuhilfenahme statistischer Komponenten

- Zuhilfenahme von Fuzzy-Logic
- Zuhilfenahme von Pattern-Recognition Methoden

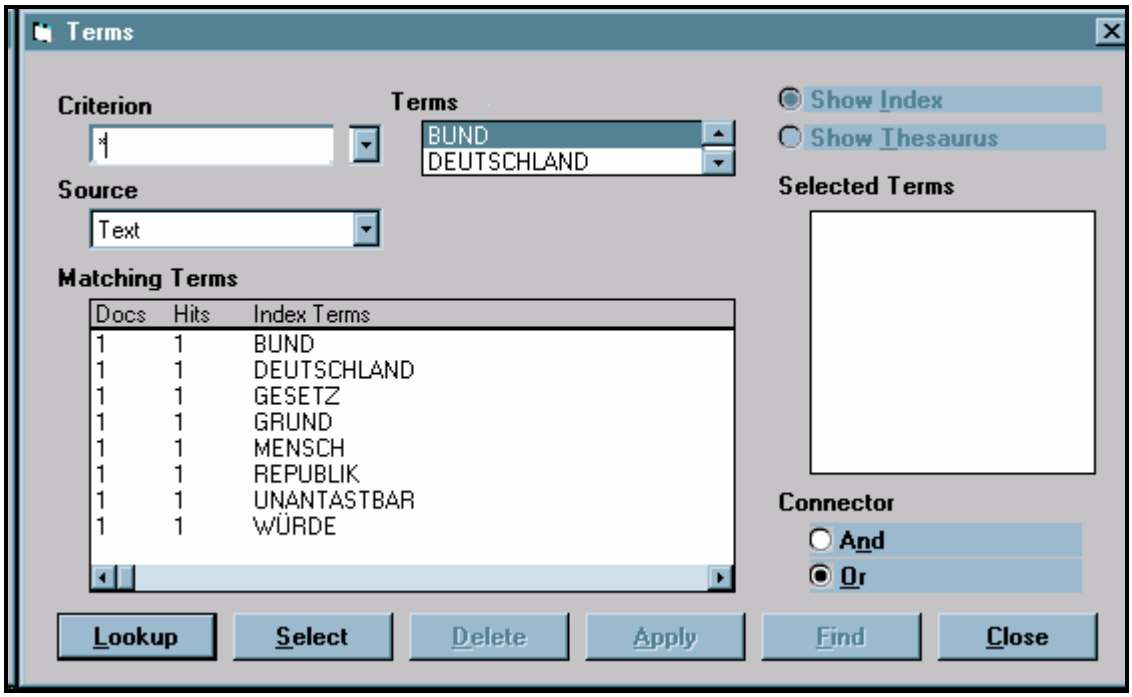
Fuzzy Suche / Pattern-Recognition sind RETRIEVAL orientiert
Linguistische Methoden sind Inhaltsorientiert. (Ziel: Semantische Suche)

Bei der "Aktiven Variante" kann durch den Einsatz linguistischer Komponenten bereits frühzeitig bei der Datenanalyse ein Versagen der Methode erkannt werden.

Rechtschreibfehlerkorrekturen bzw. der Aufbau von Fremdwörterlexika kann in den Workflow fest eingebaut werden.

Statistische Module können als Hilfsmittel zum Aufbau den notwendigen Wortmaterials für den Thesaurus herangezogen werden (WordNet, semantisches Netzwerke).

"das Grundgesetz der Bundesrepublik Deutschland"
Die Würde des Menschen ist unantastbar.



(Abb. 5: überarbeiteter Index, Quelle: BASISdesktop)

Ziele des Linguistik-Modules Morph-Server zusammenfassend:

1. Normierter Index für optimierten Thesauruseinsatz

2. Nutzung Semantischer Klassifizierung im Lexikon
   Indexierung der ausgedrückten Gedanken und größerer Einheiten wie z.B. Phrasen

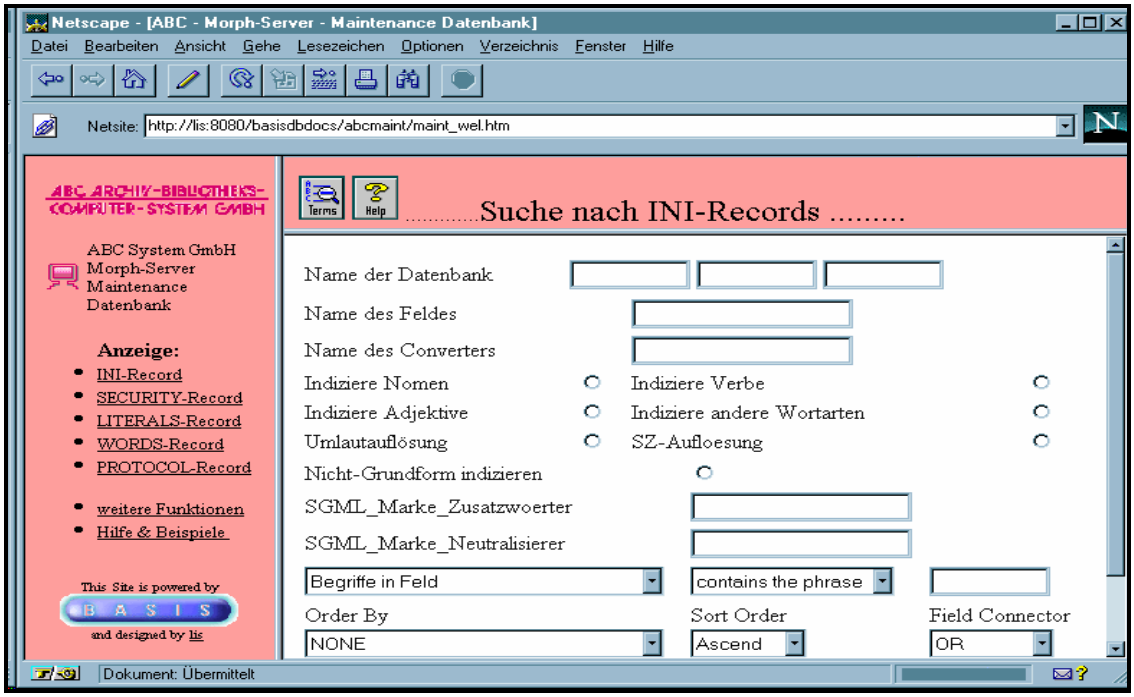
3. SGML-Tagging für zusätzliche kontextbezogene "händische Beschlagwortung" im Textfluß "*Die Würde des Menschen ist unantastbar* " <XABCTERM> MENSCHENRECHT</XABCTERM> Das Treffer-Highlightning kann direkt auf bzw. hinter den versteckten Deskriptor durchgeführt werden.

Sonderfunktionen des Morph-Servers:

Indexierung bestimmter Worttypen, z.B. nur Verben
Inhaltliche Kategorisierung im Dokument durch Zuordnung von Fachbereichen im Lexikon
Bedienung über WEB-Oberfläche
Pflege von eigenen Wörterbüchern (auch für Office-Produkte)

Umsetzung in Produkten:

Morph-Server



(Abb. 6: Parametereinstellung im Morph-Server, Quelle: ABC Morph-Server)

Für den Verband deutscher Rentenversicherer (VDR) wurde zur Ablösung des bestehenden PASSAT Systems ein derartiger Morph-Server entwickelt.

Eine weitere Installation dieses linguistischen Tools befindet sich im Nordrhein Westfälischen Landtag (NRW) im Einsatz.

Produktunabhängigkeit:
Der Morph-Server als linguistische Zusatzkomponente ist in der Programmiersprache C hardwareneutral entwickelt worden. Grundsätzlich ist dieser Modul auf verschiedenste Volltextdatenbanken und File-Indexer anwendbar, soweit diese einen Eingriff in die Indexerstellung erlauben.

Weiter Produktansätze:

Übersetzung mit dem CISLEX

On-the-Fly Übersetzung von HTML-Pages
Übersetzungstools von Systran

Vortrag:

Franz Reinisch

Maschinenbau-Betriebstechnik HTL Graz, 1978
AVL Graz, Betriebswirtsch. Assistent der Geschäftsführung, Projekt Controlling, Organisation Textsysteme, Produktionsplanung und Steuerung
PS Bremen, Projektleitung Produktionsplanungssysteme, Schiffbau
Scientific Control Systems Stuttgart, Bereichsleitung Produktionsplanungssysteme, QS-Systeme. Maintenance, Dokumentationssysteme
strässle Stuttgart, Task Force Krisen-Projektmanagement
PSI Berlin/Baden, Geschäftsführer, Produktionsplanungssysteme, Dokumentenmanagemensysteme, Maintenance Systems
LIS Steinbach, Luftfahrt-Informatik-Service OEG, Geschäftsführer, Dokumentenmanagemensysteme, Aircraft Maintenance Systems, SGML-Systeme, Radar Display Systems.
Vertretung der **ABC Bad Camberg**, BASISplus Master Reseller für Österreich, Moprh-Server und linguistische Dienstleitungen, HTML, SGML-Datenbanken.
Allgemein beeideter gerichtlicher Sachverständiger, mittlere Datentechnik

---

Weitere Informationen:

ABC System GmbH
Quellenweg 7
65520 Bad Camberg

Projekte und Vertrieb Österreich
7441 Steinbach 49
Tel.: 02616 - 4102

http://www.lis-oeg.com
eMail:reinisch@lis-oeg.com