Einführung in IR - Winter 2006/07
Dipl.-Inform. Sascha Kriewel, LF 137
sascha.kriewel@uni-due.de

Übungsblatt 6 — Freitextsuche keine Abgabe

Aufgabe 12: Informatischer Ansatz

(a) Welche spezifischen Probleme gibt es bei der Freitext-Suche im Gegensatz zur Benutzung von Dokumentationssprachen? Wie versucht der informatische Ansatz diesen zu begegnen, und welche Probleme löst dieser Ansatz nicht?

(b) Die INSPEC-Datenbank ist eine bibliographische Datenbank zu internationaler Fachliteratur der Ingenieurwissenschaften. Die Datenbank HealthStar enthält Literaturnachweise zu Themen der Gesundheit und Medizin. Von Rechnern im Universitätsnetz kann über die Universitätsbibliothek auf beide Datenbanken zugegriffen werden:

`http://thetis.hbz-nrw.de/` (Login als Gast über Standort Universitätsbibliothek Duisburg-Essen)

Finde heraus, welche Suchmöglichkeiten und insbesondere welche der in Vorlesung behandelten Operationen bei der Freitextsuche die Datenbanken zulassen.

Alternativ kannst Du auch andere Dir bekannte Literaturdatenbanken (oder andere Datenbanken aus dem Angebot der Digital Bibliothek NRW) betrachten, die Trunkierung, Maskierung und/oder Kontextoperationen bei der Suche ermöglichen.

Aufgabe 13: Computerlinguistischer Ansatz

Führe an dem folgenden Abstrakt die Schritte zur Freitext-Indexierung durch. Die Aufgabe kann auch in Form eines selbstgeschriebenen Programms gelöst werden.

(a) Lege Stopwörter fest und eliminiere sie aus dem Text.

(b) Bestimme die Grundform (oder wahlweise die Stammform) der verbleibenden Terme.

(c) Zähle die Vorkommen der Grundformen (bzw. der Stammformen) und gebe abschließend die Repräsentation des Abstrakts an.

(d) Überlege Dir eine sinnvolle Zerlegung der Komposita im Beispieltext.

Wesentliche Eigenschaften von Datenbankmanagementsystemen (DBMS) wie Datensicherheit, Nebenläufigkeit, Datenschutz und Integrität können dadurch

auch für Information-Retrieval(IR)-Systeme ohne erneuten Entwicklungsaufwand genutzt werden. Durch die zunehmende Verbreitung von IR-Systemen insbesondere auch in Anwendungen mit häufigen Änderungen des Datenbestandes (z.B. Büroinformationssysteme) werden gerade diese Eigenschaften zunehmend wichtiger.

Viele Faktendatenbanken enthalten heute auch textuelle Attribute, für die gängige DBMS (abgesehen von der Speicherung solcher Attribute als „long fields" oder „binary large objects") keine adäquate Unterstützung anbieten. Dies betrifft insbesondere den Aspekt der Anfragesprache, wo bestenfalls substring-Prädikate angeboten werden. Für textuelle Attribute in DBMS sollten die aus IR-Systemen bekannten Funktionen zur Verfügung stehen.