

Information Retrieval
Prof. Dr. N. Fuhr

Sascha Kriewel
sascha.kriewel@uni-duisburg.de

Abgabe bis 19. Mai 2004

Übungsblatt Nr. 3

Aufgabe 1: Freitext-Indexierung: Informatischer Ansatz

Die INSPEC-Datenbank ist eine bibliographische Datenbank zu internationaler Fachliteratur der Bereiche Physik, Elektrotechnik, Elektronik, Computer- und Informationstechnik. Von Rechnern im Universitätsnetz kann über die Universitätsbibliothek auf die INSPEC-Datenbank zugegriffen werden:

<http://www.ub.uni-duisburg-essen.de/research/daba/daba.shtml>

Finde heraus, welche Möglichkeiten zur Suche die INSPEC-Datenbank zulässt, insb. welche Operatoren zur Suche in den Feldern erlaubt sind.

10 Punkte

Aufgabe 2: Freitext-Indexierung: Computerlinguistischer Ansatz

Führe an dem folgenden Abstrakt die Schritte zur Freitext-Indexierung durch:

- (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.

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.

10 Punkte

Aufgabe 3: Äquivalenz von Frageformulierungen

Eine positive Eigenschaft von Fuzzy-Retrieval ist die Tatsache, dass äquivalente boolesche Frageformulierungen für ein Dokument meistens identische Relevanzwerte liefern. Diese Aussage soll näher untersucht werden:

- (a) Welche Axiome der booleschen Algebra sind hierzu zu betrachten?
- (b) Untersuche für jedes dieser Axiome, ob jeweils gleiche Relevanzwerte berechnet werden.

10 Punkte