Databases and Information Systems

1. Übungszettel Bioinformatik

Einführung in die Datenbanksysteme
Datenbanken für die Bioinformatik

H. Schweppe

Übungsaufgaben

Aufgabe 1.1

SNPs (Single nucleotide polymorphisms) sind gemeinsame DNA Sequenz Variationen von Individuen. Das SNP Consortium stellt eine Datenbank solcher Variationen bereit (Letztes Release der Daten 2001) Eine Beschreibung der Daten findet man u.a. in der Veröffentlichung
http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=12519964

Siehe auch: http://snp.cshl.org/about/

Ihre Aufgabe ist es, sich einen Überblick über die Art der gespeicherten Daten zu verschaffen. Ermitteln Sie dazu das ER-Schema für diese Datenbank durch "Reengineering". Sie können die Daten in Ihre eigene Datenbank übertragen, wenn Sie wollen (nicht Teil der Aufgabe).

Beschreiben Sie außerdem knapp die typische Art der Verwendung dieser DB.

Aufgabe 1.2

a) Charakterisieren Sie die Unterschiede zwischen Information Retrieval- und (Relationalen) Datenbanksystemen.

b) Lassen sich IR-Systeme mit Hilfe eines RDBS realisieren? Wenn ja: Geben Sie Schemata für den Fall des Boolschen Retrievals und des Vektorraumretrievals an.

c) Geben Sie für das Vektorraumretrieval die SQL-Anfrage an, die eine Liste von Dokumenten (doc_id) mit der zugehörigen Rangfolge (Ranking) für die Suchanfrage ($term_1$, $term_2$, …, $term_k$) produziert.

Aufgabe 1.3

Vektorraum-Modell: Gegeben seien die Terme {a,b,c,d,e} und die Dokumente

d1 abbeaa
d2 bbcacd
d3 bbbeb
d4 abe
d5 babcaab
d6 eceee

(a) Berechnen Sie IDF-Gewichte der Indexterme.
(b) Berechnen Sie die Rangfolgen der folgenden Anfragen zwei Mal:
q1 = a b e
q2 = a b c d e
Benutzen Sie in einem Fall die TF-Gewichte und als Ähnlichkeitsfunktion das Cosinusmaß. In dem anderen Fall benutzen Sie die gleiche Ähnlichkeitsfunktion und IDF*TF-Gewichte für die Index-Terme.