Lehrstuhl Praktische Informatik III
Universität Mannheim
Prof. Dr. Guido Moerkotte
Sven Helmer

Juli 2004

Hauptdiplomklausur
Einführung in Information Retrieval
Sommersemester 2004

Name: .....
Vorname: .....
Matrikelnummer: .....
Studienfach: .....

Wichtige Hinweise:

1. Prüfen Sie Ihr Klausurexemplar auf Vollständigkeit (6 Seiten).
2. Es sind keine Hilfsmittel zugelassen.
3. Die Klausur dauert 33 Minuten.
4. Jede Aufgabe ist auf dem zugehörigen Aufgabenblatt (und ggf. auf separaten Lösungsblättern) zu bearbeiten.
5. Vermerken Sie Ihren Namen und Ihre Matrikelnummer auf jedem Aufgaben- (bzw. Lösungsblatt). Blätter ohne Namens- und Matrikelnummerangabe werden nicht bewertet.
6. Das Deckblatt sowie alle Aufgabenblätter (evtl. Lösungsblätter) sind abzugeben.

	maximale Anzahl Punkte	erreichte Anzahl Punkte
Aufgabe 1	3	
Aufgabe 2	4	
Aufgabe 3	6	
Aufgabe 4	5	
Aufgabe 5	4	
Aufgabe 6	5	
Aufgabe 7	6	
	33	

1. (3 Punkte)
   Gegeben sei die folgende Wortmenge: { Kante, Kantor, Kanu, Kapitel, Kaviar, Kern, Kutsche }. Was sind die Nachfolgervielfalten von *Kantor*?

	mögliche Fortsetzungen	Nachfolger-vielfalt
K		
Ka		
Kan		
Kant		
Kanto		
Kantor		

2. (4 Punkte)
   Eine Dokumentsammlung wird mit Hilfe einer invertierten Datei indexiert:

⋮
Kante 2,3,9,11,17,18,21
Kantor 3,12,17
Kanu 15,16
Kapitel 1,2,3,5,9,11,15,18,21,35
Karren 1,2,15,16,18
Kasten 3,5,18,22,25
Kaviar 11,15,16,17
⋮

   Folgende Anfrage (im Booleschen Retrievalmodell) wird an die Dokumentsammlung gestellt: Kante $\wedge$ Kapitel $\wedge$ Kasten. Welche Dokumente werden als Antwort zurückgeliefert? Geben Sie die Zwischenergebnisse nach jedem Zugriff auf eine invertierte Liste an.

   1. Schritt:

   2. Schritt:

   Endergebnis:

3. (6 Punkte)
   In einem IR-System werden folgende Dokumente in einer Rankingliste zurückgeliefert. Markieren Sie die relevanten Dokumente und geben Sie für jeden Schritt die Precision an. (Sie können die Angaben in Brüchen machen, also z.B. $\frac{1}{8}$ statt 12,5%.)

Ranking	Recall	Precision
1. d_{76}	0%	
2. d_{23}	20%	
3. d_{298}	20%	
4. d_{412}	40%	
5. d_{99}	40%	
6. d_{87}	40%	
7. d_{723}	60%	
8. d_{615}	60%	
9. d_{187}	60%	
10. d_{399}	60%	
11. d_{12}	60%	
12. d_{54}	80%	

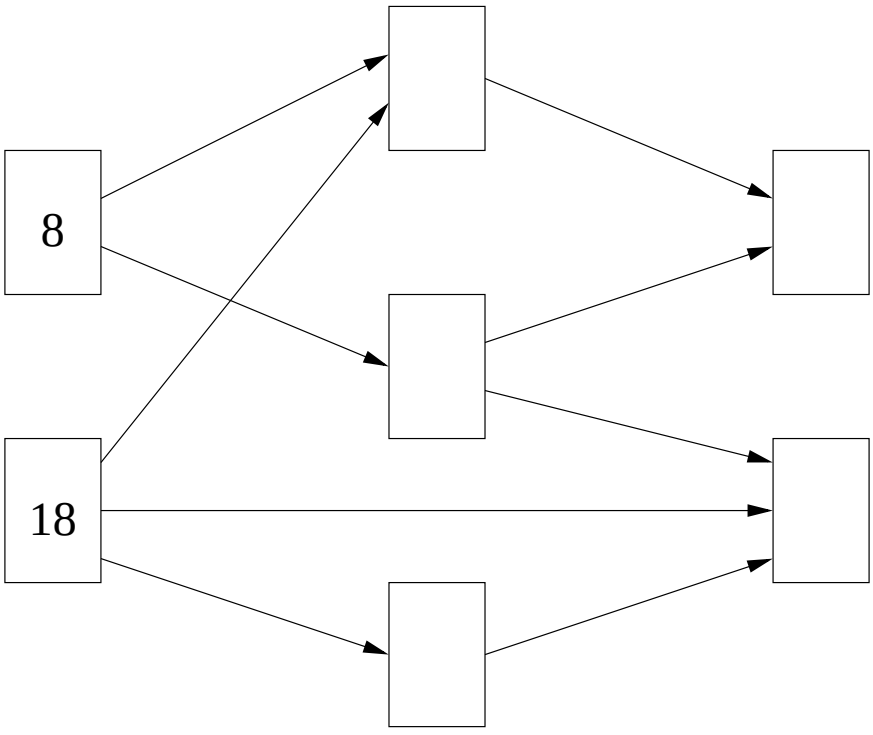
4. (a) (3 Punkte)
   Gegeben sei das Alphabet $\Sigma = \{a, b\}$. Sie bekommen den Kode 0,49 in arithmetischer Kodierung zugeschickt. Dekodieren Sie die ersten drei Zeichen des Textes. (Gehen Sie davon aus, dass das Zeichen a immer durch das niederwertige Intervall repräsentiert wird.)

   (b) (2 Punkte)
   Nennen Sie einen Vorteil und einen Nachteil der arithmetischen Kodierung gegenüber der Huffmankodierung.

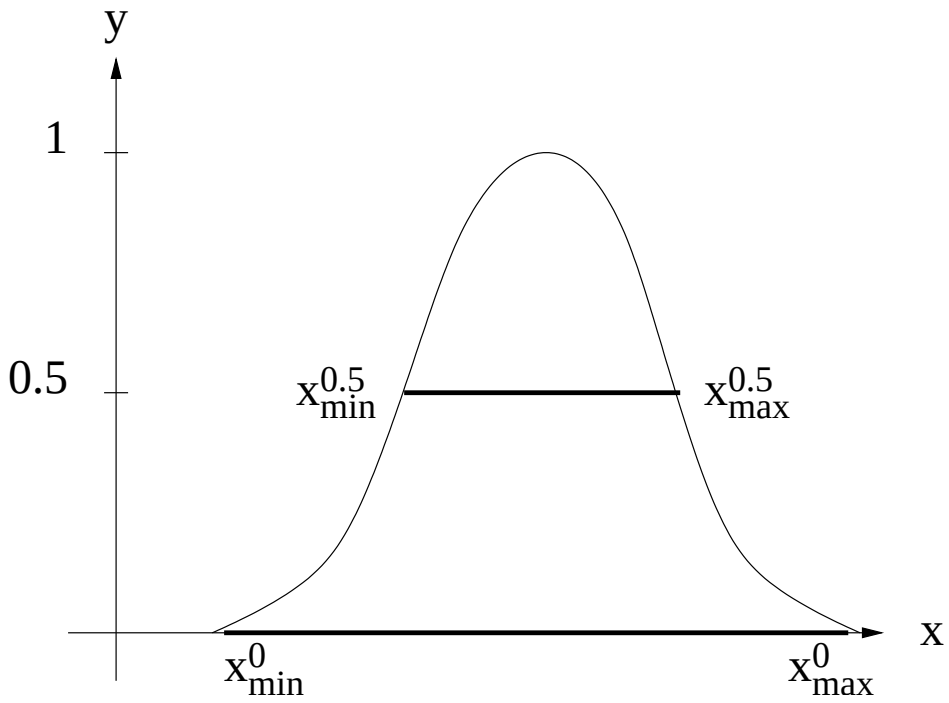
5. (4 Punkte)
   Das Suchmuster `aba` wird mit maximal einer Ersetzung im Text `abcba` gesucht. Geben Sie die Zustände der NEAs R und R^1 beim Durchsuchen des Textes an (die Startzustände sind schon eingetragen). (Hinweis: beim Ersetzungsfall gilt für den Folgezustand von R^1 folgendes: $R^1_{j+1}[i] = R_j[i-1] \vee (R^1_j[i-1] \wedge (T[j+1] = P[i]))$.)

	$R[0]$	$R[1]$	$R[2]$	$R[3]$	$R^1[0]$	$R^1[1]$	$R^1[2]$	$R^1[3]$
Start	1	0	0	0	1	0	0	0
a								
b								
c								
b								
a								

6. (5 Punkte)
   Einige Webseiten sollen mit PageRank bewertet werden. Das Ranking der ersten beiden Seiten ist eingetragen, geben Sie das Ranking der anderen Seiten an. Gehen Sie davon aus, dass $d = 1.0$, d.h. es gibt keine Sprünge.



7. (6 Punkte)
   In einem Multimedia-IR-System sollen Messkurven abgespeichert werden. Die Messkurven haben folgendes Aussehen: zuerst steigt die Kurve monoton bis zu einem Maximalwert (von 1) an, danach fällt sie wieder monoton.



   Zusätzlich zu jeder Kurve werden die Werte x^0_{min}, x^0_{max}, $x^{0.5}_{\text{min}}$ und $x^{0.5}_{\text{max}}$ abgespeichert, die die minimale bzw. maximale x-Koordinate auf der Höhe 0 bzw. 0.5 angeben. In einer Anfrage wird immer ein Punkt mit den Koordinaten (x,y) angegeben und man möchte alle Kurven bei denen der Punkt unterhalb der Kurve liegt. Wie könnte ein „Quick-and-Dirty“-Test für solche Anfragen aussehen?