COLLEGE OF ENGINEERING ,PUNE.

(An Autonomous Institute of Government of Maharashtra, Pune - 411005)

End Semester Examination

CT 419 – Information Retreival (Elective-II)

Year : B.Tech
Academic Year : 2012-13
Duration: 3 hr

Branch : Computer Engg.& I.T.
Date : 23 /04/2013
Max. Marks : 50

Instructions:

i) Answer all questions .
ii) Figures to the right indicate full marks .

Q.1 **A)** What defines a data and information in an IR system ? Give six points differentiating data retrieval from information retrieval ? 05

B) What is Inverted indexing ? Explain inverted indexing with map-reduce using the following example ? Design a pseudo-code for inverted indexing using map-reduce ? 05

Doc1 : one fish, two fish
Doc2 : red fish, blue fish
Doc3 : cat in the hat

Q.2 **A)** Cluster the following eight points (with (x, y) representing locations) into three clusters A1(2, 10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9). Initial cluster centers are: A1(2, 10), A4(5, 8) and A7(1, 2). The distance function between two points $a=(x1, y1)$ and $b=(x2, y2)$ is defined as: $\rho(a, b) = |x2 - x1| + |y2 - y1|$. 08

Use k-means algorithm to find the three cluster centers after the second iteration.

B) What is an entropy ? How entropy is related to information gain ? 02

Q.3 **A)** How is the idea of relevance feedback used in information retrieval ? What is the key concept used in relevance feedback ? Explain the Rocchio's relevance feedback algorithm ? 05

B) Mr.X has implemented a relevance feedback web search system, where he is going to do relevance feedback based only on words in the title text returned for a page (for efficiency). The user is going to rank 3 results. The first user, Mr. Y, queries for: banana slug and the top three titles returned are: 05

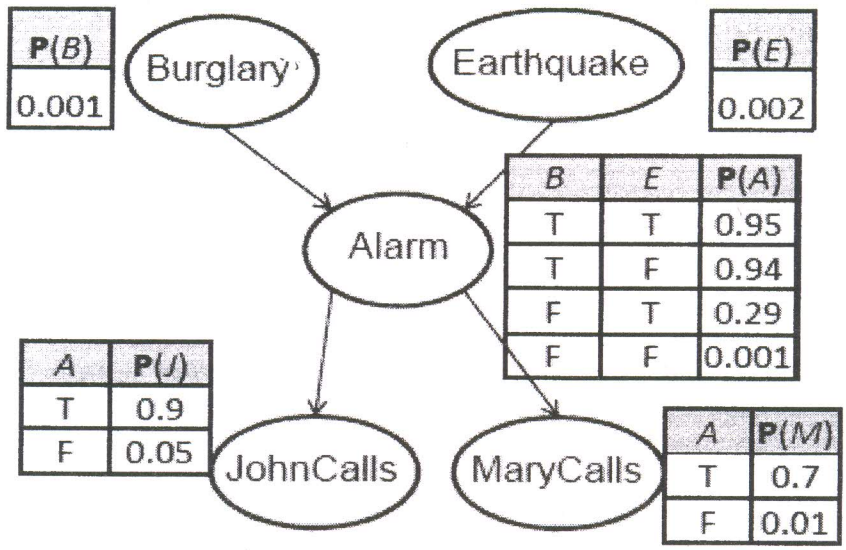
- banana slug Ariolimax columbianus
- Santa Cruz mountains banana slug
- Santa Cruz Campus Mascot

Mr.Y Judges the first two documents relevant, and the third nonrelevant. Assume that Mr.X's search engine uses term frequency but no length normalization nor IDF.Assume that he is using the Rocchio relevance feedback mechanism, with $a = b = g = 1$.

Show the final revised query that would be run. (List the vector elements in

alphabetical order.)

Q.4 **A)** Consider the following Bayesian Network with CPT's as given : 06



P(B)
0.001

P(E)
0.002

B	E	P(A)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J)
T	0.9
F	0.05

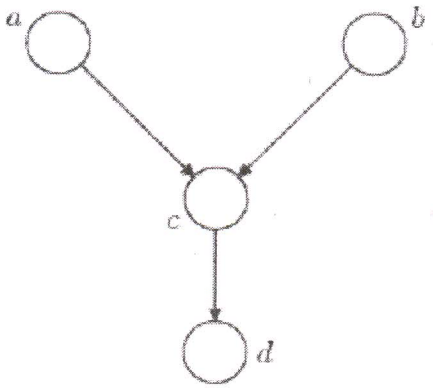
A	P(M)
T	0.7
F	0.01

Estimate the following :
1) What is the probability that Mary calls given that there is a burglary.
2) What is the probability that alarm has gone off given that John called ?
3) What is the probability that alarm rings given that there is a burglary ?

OR

A) Describe MIMD architecture with respect to Parallel IR. How inverted file is used for MIMD. ? 06

B) Consider the BN shown below over 4 RV *a*, *b*, *c*, *d*. Show that RV *a* and *b* are independent when none of the RV are observed. Show also that, in general, *a* and *b* are not independent when *d* is observed 04



Q.5 **A)** Discuss the five key abstractions used to design or model Digital Libraries. 05

B) The figure below shows the output of two information retrieval systems on the same two queries in a competitive evaluation. The top 15 ranks are shown. Crosses correspond to a document which has been judged relevant by a human judge; dashes correspond to irrelevant documents. There are no relevant 05

documents in lower ranks. Explain the following evaluation metrics and give results for query Q1 for both systems.

(a) (i) Precision at rank 10.

(ii) Recall at precision 0.5.

(b) Give the formula for mean average precision (MAP), and illustrate the metric by calculating System 1's MAP.

(c) For each system, draw a precision-recall curve. Explain how you arrived at your result. How could one create more informative curves?

OR

B) Explain Collection Partitioning, source selection and query processing with respect to Distributed IR ? 05

--- Best of Luck ----