The Center for Advanced Computer Studies
University of Southwestern Louisiana

CMPS 561

Final Examination

Date: December 12, 1997 **Instructor:** Dr. Vijay V. Raghavan

Time: 1:30 - 4:00 p.m. **Total Marks:** [100]

Note: Answer in space provided. Use backs of pages for rough work (only).

PART A (30 marks)

There are 7 short questions. You must answer **any 5**. If you answer more than 5, first 5 in order will be graded.

Q1. Grammian Matrix [6]

Q2. IDF weights [6]

Q3. Pseudo-inverse [6]

Q4. Inverted file structure for weighted retrieval [6]

Q5. Heuristic (or confidence) weight [6]

Q6. Jaccard Coefficient [6]

Q7. Learning by sample [6]

PART B (30 marks)

Answer **any 2** of the 3 questions

Q8. (a) In the context of the vector space model (VSM), if you are given matrices A and G_t, show how they should be used in computing $RSVs$ relative to a given query, $\underline{q}$. [6]

(b) Assume $n = 2$. Let $A = \begin{bmatrix} 2.5 & 1.5 \\ 1.2 & 6.5 \end{bmatrix}$ and $G_t = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix}$.

Determine the RSV of the two documents with respect to $\underline{q} = 2\underline{t}_1 + 3\underline{t}_2$. [6]

(c) Many IR practitioners and researchers suggest that if D is a matrix where each row corresponds to a document vector, then its columns can be taken as the vector representation for the various terms.

(i) Restate the above idea using our mathematical notations.
(ii) Do you agree with the statement? Explain why or why not. [3]

Q9. (a) State the main ways in which RUBRIC system differs from the standard Boolean Retrieval model. [3]

(b) Use the following subset of rules from a rule-base, for questions below: [12]

$$device \ \& \ explosion \ => \ bombing \ (0.5)$$
$$grenade \ | \ bomb \ => \ device$$
$$shell \ => \ device \ (0.4)$$

(i) Identify the concepts in the rule-base.

(ii) Determine all minimal sets of term expressions and RSVs with respect to *device*.

(iii) Repeat (ii) with respect to *bombing*.

(iv) Let $d_1 = (\ shell\ )$, $d_2 = (\ explosion\ )$, $d_3 = (\ grenade\ )$. Create an inverted list corresponding to *device* that indicates which documents have non-zero RSVs with respect to *device*.

Q10. We define 3 new operations, $(\cap, \cup, -)$, on ordered sets. They are explained with examples.

Let $X = \{x_1, x_2, x_3, x_4, x_5\}$,
$X_A = \{x_2, x_3, x_5\}$, and
$X_B = \{x_1, x_2, x_5\}$

Then:

$X_A \cap X_B = \{x_2, x_5\}$,
$X_A \cup X_B = \{x_1, x_2, x_3, x_5\}$ and
$X - X_A = \{x_1, x_4\}$

That is, $\cap$, $\cup$ and $-$ are specialized set operations where the order of elements is preserved.

a) Describe how these can be used to implement a retrieval system based on Boolean Retrieval Model that uses inverted file structure. That is, we want to process queries such as $A \wedge B$, $A \vee B$, $\neg A$, etc. [3]

b) Use the table of Q. 12 and construct the inverted lists D_{ti}, for $1 \leq i \leq 4$. [4]

c) Determine RSV_q of documents $\{d_1, d_3, d_5\}$ with respect to the following queries. You must use the inverted lists and operations defined in parts (a) and (b).

(i) $q_1 = \neg t_1 \wedge t_2 \wedge \neg t_3$ [4]

(ii) $q_2 = \neg(t_1 \wedge t_4) \vee (\neg t_2 \wedge t_3)$ [4]

PART C (40 Marks)
ANSWER ANY 2 QUESTIONS

Q11. (a) State and give the meaning of perceptron criterion used in the generalized (for multi-level relevance) perceptron convergence algorithm. [4]

(b) Under what condition(s) will the generalized perceptron convergence algorithm terminate? What kind of retrieval result is guaranteed, if it terminates? [4]

(c) Determine the optimal query term weights, assuming that we have the following documents and user feedback. [12]

	t_1	t_2	t_3	t_4	Feedback
doc. 1	2	1	0	0	P
doc. 5	3	2	1	0	R
doc. 7	0	3	0	3	N
doc. 3	2	1	2	2	N
doc. 6	0	1	2	3	N

Note: 'P' stands for partially relevant

Q12. (a) Under the assumptions of term independence and that the values of any term follow a Bernoulli distribution within the relevant as well as within non-relevant documents, show that the retrieval status value of $d = (w_1, w_2, ..., w_n)$ can be obtained by

$$\sum_{i=1}^{n} w_i \qquad \log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

where p_i is the probability that $w_i = 1$ in relevant documents and q_i is the probability that $w_i = 1$ in non-relevant documents. [8]

(b) Explain the meaning of the above query weight. That is, what does it mean when it is less than, equal to or greater than zero. [3]

(c) Determine the optimal query term weights, assuming that we have the following documents and user feedback. You must use Jeffry's priors. [5]

	t_1	t_2	t_3	t_4	Feedback
doc. 1	1	1	0	0	R
doc. 5	1	1	1	0	R
doc. 7	0	1	0	1	N
doc. 3	1	1	1	1	N
doc. 6	0	1	1	1	N

(d) Determine the R_{norm} of the ranking that results based on the optimal query weights in part (c). [4]

Q.13. (a) Explain in words the meaning of PRECALL. [2]

(b) Provide an informal proof that the value of PRECALL after retrieving NR relevant documents is given by

$$\frac{NR}{NR + j + \frac{s \cdot i}{r}},$$

where

j - # of non-relevant documents in the levels completely retrieved,

s - # of relevant documents actually retrieved from the last level,

r - total # of relevant documents in the last level, and

i - total # of non-relevant documents in the last level. [6]

(c) $\triangle = (- - + \mid - + - - + \mid - - - \mid - + - -)$

For the retrieval output given above, determine

(i) PRECALL at standardized recall values of 0.25, 0.5, 0.75 and 1. [4]

(ii) For another query, retrieval output is:

$$\Delta = (+ + - - \mid - - \mid + - - \mid -)$$

a) Find PRECALL at the same recall values as part (i). [4]

b) What is the averaged PRECALL (precision) values over the 2 queries at the recall values specified? [4]