TRƯỜNG ĐH NGOẠI NGỮ TIN HỌC TPHCM
KHOA CNTT

ĐỀ 2

Đề thi môn **KHAI KHOÁNG DỮ LIỆU**

Khoá 2006 – Thi lần 1 – Học kỳ I – Ngày thi: 5/1/2010 – Thời gian làm bài: 90'
(Sinh viên không được sử dụng tài liệu)

Câu 1: (4 điểm)

a) So sánh kết quả các phương pháp tìm tập phổ biến về: kết quả, không gian lưu trữ và thời gian khai thác.

b) Cho cơ sở dữ liệu giao tác như sau:

Mã giao dịch	A	B	C	D	E	F
1	0	0	1	0	1	1
2	1	0	1	1	1	0
3	0	0	1	1	0	0
4	0	0	0	1	1	0
5	1	1	1	0	1	0

i) Tìm tất cả các tập phổ biến (FIs) có trong CSDL với minSup = 40% theo phương pháp FP-Tree

ii) Tìm tất cả các tập phổ biến đóng theo thuận toán CHARM với minSup = 40%

Câu 2: (4 điểm) Cho bảng quan sát về thời tiết như sau:

Ngày	**Quang cảnh**	**Nhiệt độ**	**Độ ẩm**	**Gió**	**Chơi Tennis?**
1	Mưa	Lạnh	TB	Thấp	Đi
2	Nắng	TB	Cao	Thấp	Không đi
3	Nắng	Lạnh	TB	Thấp	Đi
4	Âm u	Lạnh	TB	Cao	Đi
5	Mưa	TB	TB	Thấp	Đi
6	Nắng	Nóng	Cao	Cao	Không đi
7	Nắng	TB	TB	Cao	Đi
8	Âm u	TB	Cao	Cao	Đi
9	Âm u	Nóng	TB	Thấp	Đi
10	Mưa	TB	Cao	Cao	Không đi
11	Nắng	Nóng	Cao	Thấp	Không đi
12	Âm u	Nóng	Cao	Thấp	Đi
13	Mưa	Lạnh	TB	Cao	Đi
14	Âm u	TB	Cao	Thấp	Không đi
15	Mưa	TB	Cao	Thấp	Đi

a) Tính Ratio của thuộc tính Quang cảnh và Nhiệt độ dựa trên 10 dòng dữ liệu đầu tiên. Nếu chọn lựa nút làm gốc ta nên lựa thuộc tính nào? Tại sao?

b) Dự đoán mẫu của các lớp 10-15 dựa trên 10 dòng đầu tiên theo phương pháp Bayesian. Dựa vào kết quả này, cho biết độ chính xác phân lớp?

Câu 3: (2 điểm)

Cho 6 điểm A, B, C, D, E, F có toạ độ được cho như sau: A(1,1,1), B(2,2,2), C(4,1,0), D(2,0,-1), E(0,1,3), F(0,0,2). Sử dụng phương pháp phân cấp để gom các điểm trên (sử dụng độ đo khoảng cách Manhattan).