

ĐỀ CƯƠNG MÔN HỌC

1. THÔNG TIN VỀ MÔN HỌC

1.1 Tên môn học:	KHAI PHÁ DỮ LIỆU
1.2 Tên tiếng Anh:	DATA MINING
1.3 Mã môn học:	ITEC3417
1.4 Khoa phụ trách:	Công nghệ Thông tin
1.5 Số tín chỉ:	04 (04 LT, 0 TH)

2. MÔ TẢ MÔN HỌC

Khai phá dữ liệu là quá trình khám phá các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có. Mục tiêu của khai phá dữ liệu là sử dụng các giải thuật xử lý dữ liệu để biến dữ liệu thô thành dữ liệu có cấu trúc dễ hiểu để sử dụng tiếp. Các giải thuật này được tổng hợp từ nhiều lĩnh vực thuộc khoa học máy tính như trí tuệ nhân tạo, máy học, thống kê và hệ thống cơ sở dữ liệu.

3. MỤC TIÊU MÔN HỌC

3.1. Mục tiêu chung

Môn học cung cấp một cái nhìn tổng quan về khai phá dữ liệu trong việc hỗ trợ ra quyết định điều hành. Các kỹ thuật khai phá dữ liệu được giới thiệu cho việc khám phá các thông tin ẩn từ dữ liệu thu thập được dựa trên các nhu cầu thực tế.

3.2. Mục tiêu cụ thể

3.2.1. Kiến thức

- Hiểu và áp dụng được khái niệm của khai phá dữ liệu, các thuận lợi và giới hạn của nó.
- Hiểu và áp dụng được các giải thuật khai phá dữ liệu phổ biến.
- Hiểu được cách khai thác tri thức, thông tin từ dữ liệu.
- Hiểu và ứng dụng khai phá dữ liệu trong khoa học máy tính cũng như ứng dụng thực tế.
- Hiểu được các hướng nghiên cứu và ứng dụng hiện nay về khai phá dữ liệu.

3.2.2. Kỹ năng

- Nâng cao kỹ năng về làm việc nhóm, kỹ năng tìm kiếm, tổng hợp và trình bày vấn đề, đặc biệt liên quan đến khai phá dữ liệu.
- Nâng cao kỹ năng đọc và hiểu tài liệu tiếng Anh chuyên ngành.

3.2.3. Thái độ

- Tinh thần tự học, tự nghiên cứu
- Không ngừng cập nhật những kết quả nghiên cứu mới lĩnh vực khai phá dữ liệu

4. NỘI DUNG MÔN HỌC

STT	Tên chương	Mục, tiểu mục	Số tiết			Tài liệu tự học
			TC	LT	TH	
1	Tổng quan về khai phá dữ liệu	1.1 Tổng quan khai phá dữ liệu 1.2 Dạng dữ liệu và mẫu cần tập trung khai phá 1.3 Các phương pháp và ứng dụng cho bài toán khai phá dữ liệu 1.4 Các yếu tố ảnh hưởng	9	9	0	[1] Chương 1 [2] Chương 1 [3] Chương 1
2	Tiền xử lý dữ liệu	2.1 Khảo sát dữ liệu 2.2 Thống kê 2.3 Hình tượng hóa 2.4 Đo đặc độ tương đồng 2.5 Làm sạch dữ liệu 2.5.1 Xử lý dữ liệu bị thiếu 2.5.2 Xử lý dữ liệu không đúng và không nhất quán 2.5.3 Chuẩn hóa và mở rộng quy mô 2.6 Tích hợp và thu gọn dữ liệu 2.6.1 Chọn mẫu 2.6.2 Lựa chọn tập con đặc trưng 2.6.3 Giảm chiều với trục xoay 2.6.4 Giảm chiều với chuyển đổi kiểu 2.7 Biến đổi dữ liệu và rời rạc hóa	15	15	0	[1] Chương 2 [2] Chương 6 [2] Chương 7
3	Khai thác các mẫu phổ biến	3.1 Khai thác tập hợp 3.1.1 Tập phổ biến và luật kết hợp 3.1.2 Các giải thuật khai thác tập hợp 3.1.3 Các luật kết hợp tổng quát 3.2 Khai thác sự tuần tự 3.2.1 Trình tự phổ biến 3.2.2 Khai thác các trình tự phổ biến	12	12	0	[1] Chương 4 [1] Chương 5 [2] Chương 8 [2] Chương 10 [2] Chương 11

STT	Tên chương	Mục, tiểu mục	Số tiết			Tài liệu tự học
			TC	LT	TH	
		3.2.3 Khai thác các chuỗi con thông qua các cây hậu tố 3.3 Khai thác các mẫu Graph 3.3.1 Hỗ trợ và đẳng cấu 3.3.2 Khởi tạo các ứng viên 3.3.3 Thuật toán gSpan				[3] Chương 6
4	Phân loại dữ liệu	4.1 Nền tảng toán học 4.2 Cây quyết định 4.2.1 Các cây quyết định 4.2.2 Các tiêu chí để tách 4.2.3 Cắt tỉa cây và tiêu chí dừng 4.3 Thuật toán Bayes 4.4 Phân loại dữ liệu dựa vào các luật 4.4.1 Khởi tạo các luật từ các cây quyết định 4.4.2 Các giải thuật bao phủ tuần tự 4.4.3 Cắt tỉa các luật 4.4.4 Các phân loại kết hợp 4.5 Các kỹ thuật nâng cao độ chính xác.	12	12	0	[1] Chương 10 [2] Chương 18 [2] Chương 19
5	Gom cụm dữ liệu (data clustering)	5.1 Nền tảng toán học 5.2 Phân hoạch k-mean 5.3 Gom cụm dựa trên cấu trúc lồng nhau 5.4 Gom cụm dựa trên lưới 5.5 Đánh giá độ hiệu quả của việc gom cụm	12	12	0	[1] Chương 6 [1] Chương 7 [2] Chương 13 [2] Chương 14 [2] Chương 15 [2] Chương 16 [3] Chương 7

5. TÀI LIỆU THAM KHẢO

5.1. Tài liệu chính

[1] Charu C. Aggarwal, *Data Mining: The Textbook*, Springer, 2015.

[2] Mohammed J. Zaki, Wagner Meira Jr, *Data Mining and Analysis Fundamental Concepts and Algorithms*, Cambridge University Press, 2014.

5.2. Tài liệu tham khảo

[3] Jure Leskovec, Anand Rajaraman, Jeffrey David Ullman, *Mining of Massive Datasets (2nd edition)*, Cambridge University Press, 2014.

6. ĐÁNH GIÁ KẾT QUẢ HỌC TẬP

STT	Hình thức đánh giá	Trọng số
1	Đánh giá trên lớp, chuyên cần	10%
2	Báo cáo tiêu luận (bài tập lớn)	30%
3	Thi cuối kỳ (thi viết trên giấy)	60%

7. KẾ HOẠCH GIẢNG DẠY

STT	Buổi học	Nội dung	Ghi chú
1	Buổi 1	Tổng quan về khai phá dữ liệu (4,5 tiết)	
2	Buổi 2	Tổng quan về khai phá dữ liệu (tt) (4,5 tiết)	
3	Buổi 3	Tìm hiểu và tiền xử lý dữ liệu (4,5 tiết)	
4	Buổi 4	Tìm hiểu và tiền xử lý dữ liệu (tt) (4,5 tiết)	
5	Buổi 5	Tìm hiểu và tiền xử lý dữ liệu (tt) (4,5 tiết)	
6	Buổi 6	Tìm hiểu và tiền xử lý dữ liệu (tt) (1,5 tiết) Luật kết hợp (3 tiết)	
7	Buổi 7	Luật kết hợp (tt) (4,5 tiết)	
8	Buổi 8	Luật kết hợp (tt) (4,5 tiết)	
9	Buổi 9	Phân loại dữ liệu (4,5 tiết)	
10	Buổi 10	Phân loại dữ liệu (tt) (4,5 tiết)	
11	Buổi 11	Phân loại dữ liệu (tt) (3 tiết) Gom cụm dữ liệu (1,5 tiết)	
12	Buổi 12	Gom cụm dữ liệu (tt) (4,5 tiết)	
13	Buổi 13	Gom cụm dữ liệu (tt) (4,5 tiết)	
14	Buổi 14	Gom cụm dữ liệu (tt) (1,5 tiết)	

TRƯỜNG KHOA
(đã ký)
TS. Lê Xuân Trường