Trường Đại học Bách Khoa - ĐHQG Tp.HCM
Khoa: Khoa Khoa học và Kỹ thuật Máy tính
Khoa/Bộ môn quản lý MH: Hệ thống Thông tin

Tp.HCM, ngày ..... tháng ..... năm .....

Đề cương môn học Sau đại học

KHAI PHÁ DỮ LIỆU (DATA MINING)

Mã số MH: CO5131

Số tín chỉ:	**Tc (LT.BT&TH.Tự Học): 3**						**ECTS: 6**			
Số tiết -Tổng:	**72**	LT:	**30**	BT:	**6**	TH:	**0**	ĐA:	BTL/TL:	**36**
Đánh giá:	Thuyết trình:		20%	Thuyết trình các nội dung môn học						
	Tiểu luận:		30%	Làm việc theo nhóm						
	Thi cuối kỳ:		50%	Trắc nghiệm + bài tập						
- Môn tiên quyết:										
- Môn học trước:										
- Môn song hành:										
- CTĐT ngành (Mã ngành):	Khoa Học Máy Tính (8480101)									
- Ghi chú khác:										

1. <u>Mục tiêu môn học:</u>

Môn học này nhằm đào tạo sinh viên kiến thức và kỹ năng liên quan đến khai phá dữ liệu (KPDL) như: các kỹ thuật thu thập, tiền xử lý dữ liệu, sự liên hệ giữa KPDL với các lĩnh vực nghiên cứu khác trong khoa học máy tính, các kỹ thuật và công cụ chính trong khai phá dữ liệu, khả năng ứng dụng, các thách thức và hướng phát triển của lĩnh vực này trong tương lai.

Sau khi hoàn thành môn học này, học viên có thể phân tích, chọn lựa/đề xuất, thiết kế phương pháp/kỹ thuật khai phá dữ liệu cho một bài toán đặt ra trong thực tiễn và hiện thực, đánh giá chúng.

Aims:

This course provides students knowledge and skills in data mining including data gathering and pre-processing techniques, relations between data mining and other research fields in computer science, major mechanisms and tools for data mining, real-world applications, research challenges as well as trends of this research field.

After completing this course, students can analyze, select, design data mining methods/techniques for a practical problem and implement, evaluate them.

2. Nội dung tóm tắt môn học:

Trong môn học này, tổng quan về khai phá dữ liệu trong việc hỗ trợ ra quyết định điều hành và mang tính chiến lược sẽ được trình bày. Các kỹ thuật khai phá dữ liệu được giới thiệu cho việc khám phá các thông tin ẩn từ dữ liệu thu thập được. Bên cạnh đó, môn học này cũng giới thiệu cách phân tích các nhu cầu kinh doanh cho việc khám phá tri thức để tạo nên lợi thế cạnh tranh và cách áp dụng các công nghệ khai phá dữ liệu một cách thích hợp để nhận dạng giá trị kinh doanh thực sự của các công nghệ khai phá dữ liệu này. Các đề tài cụ thể thuộc về môn học này bao gồm: các phương pháp và quá trình khai phá dữ liệu, các công nghệ khai phá dữ liệu, các ứng dụng và nghiên cứu cụ thể của khai phá dữ liệu.

Chi tiết, các nội dung sau sẽ được giảng dạy trong môn học:

- Tổng quan về các quá trình khám phá tri thức, khai phá dữ liệu, và quá trình tiền xử lý dữ liệu

- Phân tích về sự liên hệ giữa lĩnh vực khai phá dữ liệu và các lĩnh vực nghiên cứu khác trong khoa học máy tính

- Trình bày các giải thuật và kỹ thuật khai phá dữ liệu chính gồm: hồi qui dữ liệu, phân loại dữ liệu, gom cụm dữ liệu, và phân tích kết hợp – tương quan

- Nghiên cứu, thảo luận về các vấn đề hiện đại trong khai phá dữ liệu như khai phá dữ liệu lớn, phi cấu trúc, các thách thức và xu hướng phát triển của lĩnh vực này

- Tạo khả năng cho sinh viên phát triển và tận dụng các giải thuật và kỹ thuật khai phá dữ liệu cho các ứng dụng và loại dữ liệu khác nhau.

Course outline:

This course gives an overall view about data mining in operational and strategic decision making support. Data mining techniques are introduced for hiden information discovery from acquired data. Besides, this course also discusses analytic methods on business requirements for knowledge discovery in order to identify real business value of data mining technology. Specific topics in this course are listed as follows: data mining techniques and processes, data mining technology, particular applications and studies about data mining.

In details, following topics will be presented in the module:

- Introduction to knowledge discovery process, data mining process, and data preprocessing

- Analyze the relation between data mining and other prinicples in computer science

- Major data mining algorithms and techniques for regression, classification, clustering, association – correlation analysis

- Investigate and discuss on modern issues in data mining such as mining from massive datasets, trends in data mining,...

- Enable students to develop and utilize data mining algorithms and techniques for many different applications and kinds of data

3. Tài liệu học tập:

Giáo trình/Textbook

[1] Jiawei Han, Micheline Kamber, Jian Pei, "Data Mining: Concepts and Techniques", Third Edition, Morgan Kaufmann Publishers, 2012.

[2] Trần Minh Quang, "Khai Phá Dữ Liệu và Kỹ Thuật Phân Lớp", NXB Đại Học Quốc Gia TP. HCM, 2020.

Sách tham khảo/References

[3] J. Leskove, A. Rajaraman, J. D. Ullman,"Mining of messive data", 2nd Edition, Cambridge University Press, 2014.

[4] David L. Olson, Dursun Delen, "Advanced Data Mining Techniques", Springer-Verlag, 2008.

[5] Graham J. Williams, Simeon J. Simoff, "Data Mining: Theory, Methodology, Techniques, and Applications", Springer-Verlag, 2006.

[6] Hillol Kargupta, Jiawei Han, Philip S. Yu, Rajeev Motwani, and Vipin Kumar, "Next Generation of Data Mining", Taylor & Francis Group, LLC, 2009.

[7] Daniel T. Larose, "Data mining methods and models", John Wiley & Sons, Inc, 2006.

[8] Ian H.Witten, Frank Eibe, Mark A. Hall, "Data mining : practical machine learning tools and techniques", Third Edition, Elsevier Inc, 2011.

[9] Florent Messeglia, Pascal Poncelet & Maguelonne Teisseire, "Successes and new directions in data mining", IGI Global, 2008.

[10] Robert Nisbet, John Elder, Gary Miner, "Handbook of Statistical Analysis and Data Mining Applications", Elsevier Inc, 2009.

4. Các hiểu biết, các kỹ năng cần đạt được sau khi học môn học:

STT	Chuẩn đầu ra môn học (CĐRMH)	Công cụ đánh giá CĐRMH	Đóng góp CĐR Chương trình (CĐRCT)		
			Ứng dụng	Nghiên cứu	
CĐRMH.1	Nắm được ý nghĩa của khai phá dữ lệu, sự liên hệ với các lĩnh vực khác trong khoa học máy tính	Thuyết trình	a		1.3
CĐRMH.2	Trình bày được các giải thuật và kỹ thuật khai phá dữ liệu chính gồm: hồi qui dữ liệu, phân loại dữ liệu, gom cụm dữ liệu, và phân tích kết hợp – tương quan	Tiểu luận, Thi cuối kỳ	e		2.1, 2.3
CĐRMH.3	Có khả năng thảo luận về các vấn đề hiện đại trong khai phá dữ liệu như khai phá dữ liệu lớn, phi cấu trúc, các thách thức và xu hướng phát triển của lĩnh vực này	Thuyết trình, Tiểu luận	d, f		2.2, 2.3

STT	Chuẩn đầu ra môn học (CĐRMH)	Công cụ đánh giá CĐRMH	Đóng góp CĐR Chương trình (CĐRCT)		
			Ứng dụng	Nghiên cứu	
CĐRMH.4	Có thể phát triển và tận dụng các giải thuật và kỹ thuật khai phá dữ liệu cho các ứng dụng và loại dữ liệu khác nhau.	Thuyết trình, Tiểu luận	a, g		2.1, 2.3

Learning outcomes:

No.	Course learning outcomes (CLO)	CLO assessment	Matching with PLO		
			Coursework	Research	
L.O.1	Understand the meaning of data mining and relations between this fields to others in computer science	Chapter presentation	a		1.3
L.O.2	Be able to explain major data mining algorithms and techniques for regression, classification, clustering, association – correlation analysis	Project, Final exams	e		2.1, 2.3
L.O.3	Be able to discuss on modern issues in data mining such as mining from massive datasets, trends in data mining,...	Chapter presentation, Project	d, f		2.2, 2.3
L.O.4	Can develop and utilize data mining algorithms and techniques for many different applications and kinds of data	Chapter presentation, Project	a, g		2.1, 2.3

Bảng ánh xạ chuẩn đầu ra môn học và chuẩn đầu ra chương trình ứng dụng:

	Chuẩn đầu ra của chương trình (CĐRCT)										
Chuẩn đầu ra môn học (CĐRMH)	a	b	c	d	e	f	g	h	i	j	k
CĐRMH.1	✓										
CĐRMH.2					✓						
CĐRMH.3				✓		✓					
CĐRMH.4	✓						✓				

Bảng ánh xạ chuẩn đầu ra môn học và chuẩn đầu ra chương trình nghiên cứu:

	Chuẩn đầu ra của chương trình (CĐRCT)										
Chuẩn đầu ra môn học (CĐRMH)	a	b	c	d	e	f	g	h	i	j	k
CĐRMH.1											
CĐRMH.2											
CĐRMH.3											
CĐRMH.4											

5. Hướng dẫn cách học - chi tiết cách đánh giá môn học:

- Đọc hiểu, phân tích các kỹ thuật về khai phá dữ liệu trong text book và trên Internet

- Khảo sát các ứng dụng thực tế mà KPDL có thể được vận dụng

- Tìm cách áp dụng các kỹ thuật KPDL vào một vài bài toán thực tế cụ thể

- Thực hành lập trình hoặc sử dụng tốt các công cụ KPDL (ví dụ Python, Weka, Orange...)

- Mở rộng các vấn đề nghiên cứu trong KPDL

Sinh viên cần đọc sách giáo trình và làm bài tập đầy đủ.

Sinh viên cần thực hành các công cụ khai phá dữ liệu từ MS SQL Server, Oracle, hoặc bất kỳ công cụ khai phá mã nguồn mở khác chẳng hạn như Weka.

Về thực hiện báo cáo tiểu luận: nhóm 1-2 người, bắt đầu thực hiện từ tuần 1 đến tuần 15, nộp đề cương của báo cáo tiểu luận vào tuần 4 (không bắt buộc), nộp báo cáo tiểu luận vào tuần 15 và xung phong trình bày báo cáo tiểu luận vào tuần 15.

Cách đánh giá: 20%*điểm kiểm tra trên lớp + 30%*điểm báo cáo tiểu luận + 50%*điểm thi cuối kỳ từ 5 trở lên (>= 5.0) mới tính là đạt cả môn học.

Learning strategies & Assessment Scheme:

- Comprehensive reading and analyze data mining mechanisms from text book and Internet

- Investigate real-world applications where data mining can be applied

- Find ways to applied data mining to the reality

- Practice with programming and tools (e.g., Python, Weka, Orange,...) for data mining

- Discuss on research issues and development trends in this field

Students need to read textbooks and finish all assignments.

Students need to do practice with data mining tasks from MS SQL Server, Oracle, or any other open source tools such as Weka.

As for seminar report, each group includes one-two students, starts working from week 1 to week 15, optionally submits the description of a seminar report in week 4, submits the final version of a seminar report in week 15, and optionally presents a seminar report in week 15.

Assessment scheme:

20% * in-class exam score + 30% * Report score + 50% * Final exam score >= 5.0 in order to pass this course.

6. Nội dung chi tiết:

Tuần/ Buổi	Chủ đề (chương)	Nội dung	Chuẩn đầu ra môn học	Tài liệu
1	**Chương 1: Tổng quan về khai phá dữ liệu**	1.1.Quá trình khám phá tri thức1.2.Các khái niệm1.3.Ý nghĩa và vai trò của khai phá dữ liệu1.4.Ứng dụng của khai phá dữ liệu1.5.Tóm tắt	CĐRMH.1	[1, 2, 7, 9]
2, 3	**Chương 2: Các vấn đề tiền xử lý dữ liệu**	2.1. Tổng quan về giai đoạn tiền xử lý dữ liệu2.2. Tóm tắt hoá dữ liệu2.3. Làm sạch dữ liệu2.4. Tích hợp dữ liệu2.5. Biến đổi dữ liệu2.6. Thu giảm dữ liệu2.7. Rời rạc hóa dữ liệu2.8. Tạo cây phân cấp ý niệm2.9. Biểu diễn dữ liệu 2.10. Tóm tắt	CĐRMH.1	[1]

Tuần/ Buổi	Chủ đề (chương)	Nội dung	Chuẩn đầu ra môn học	Tài liệu
4, 5	**Chương 3: Hồi qui dữ liệu**	3.1. Tổng quan về hồi qui3.2. Hồi qui tuyến tính3.3. Hồi qui phi tuyến3.4. Ứng dụng3.5. Các vấn đề với hồi qui3.6. Tóm tắt	CĐRMH.2, CĐRMH.3	[1-7]
6, 7	**Chương 4: Phân loại dữ liệu**	4.1. Tổng quan về phân loại dữ liệu4.2. Phân loại dữ liệu với cây quyết định4.3. Phân loại dữ liệu với mạng Bayesian4.4. Phân loại dữ liệu với mạng Neural4.5. Các phương pháp phân loại dữ liệu khác4.6. Tóm tắt	CĐRMH.2, CĐRMH.3	[1-7]
8, 9	**Chương 5: Gom cụm dữ liệu**	5.1. Tổng quan về gom cụm dữ liệu5.2. Gom cụm dữ liệu bằng phân hoạch5.3. Gom cụm dữ liệu bằng phân cấp5.4. Gom cụm dữ liệu dựa trên mật độ5.5. Gom cụm dữ liệu dựa trên mô hình5.6. Các phương pháp gom cụm dữ liệu khác5.7. Tóm tắt	CĐRMH.2, CĐRMH.3	[1-7]
9, 10	**Chương 6: Luật kết hợp**	6.1. Tổng quan về luật kết hợp6.2. Biểu diễn luật kết hợp6.3. Khám phá các mẫu thường xuyên6.4. Khám phá các kết hợp với giải thuật Apriori và các biến thể của giải thuật Apriori6.5. Khám phá các kết hợp dựa trên ràng buộc6.6. Phân tích tương quan6.7. Tóm tắt	CĐRMH.2, CĐRMH.3	[1-7]
11-12	Bài tập tổng hợp tại lớp	- Bài tập tổng hợp- Báo cáo tiểu luận	CĐRMH.2, CĐRMH.3, CĐRMH.4	Tất cả
(Không lên lớp)	**TIỂU LUẬN (36 tiết): Sinh viên thực hiện và báo cáo tiểu luận với GV. GV sẽ không lên lớp cho các buổi tiểu luận. Thay vào đó, HV báo cáo với GV tại phòng Lab của GV.**	HV thảo luận với GV trong việc chọn đề tài, mục tiêu và cách tiếp cậnHV báo cáo tiến độ làm việc với GV và tiếp nhận góp ý, định hướng của GVHV báo cáo kết quả tiểu luận và GV đánh giá (tuần 11 - 12)	CĐRMH.1, CĐRMH.2, CĐRMH.3, CĐRMH.4	[1 - 9]

7. <u>Giảng viên tham gia giảng dạy:</u>

CBGD chính:	PGS.TS Trần Minh Quang

CBGD tham gia:

TS. Lê Thanh Vân

PGS.TS Lê Hồng Trang

PGS.TS Võ Thị Ngọc Châu

XÁC NHẬN CỦA HỘI ĐỒNG XÂY DỰNG CHƯƠNG TRÌNH ĐÀO TẠO VÀ KHOA

Tp. Hồ Chí Minh, ngày ..... tháng ..... năm .....

GIẢNG VIÊN LẬP ĐỀ CƯƠNG

TS. Phạm Hoàng Anh