ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PGS.TS Đỗ Phúc

GIÁO TRÌNH

KHAI PHÁ DỮ LIỆU
(Data Mining)

NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH - 2016

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PGS.TS Đỗ Phúc

GIÁO TRÌNH

KHAI PHÁ DỮ LIỆU
(Data Mining)

NHÀ XUẤT BẢN ĐẠI HỌC QUỐC GIA
THÀNH PHỐ HỒ CHÍ MINH - 2016

GIÁO TRÌNH
KHAI PHÁ DỮ LIỆU

PGS. TS ĐỖ PHÚC

LỜI NÓI ĐẦU

Khai phá dữ liệu (data mining) nhằm tìm kiếm và nhận dạng các mẫu hữu ích tiềm ẩn trong khối dữ liệu lớn. Hiện nay, công nghệ thông tin (CNTT) được áp dụng rộng rãi trong mọi lĩnh vực của đời sống xã hội, do vậy nhiều cơ sở dữ liệu (CSDL) khổng lồ đã được tạo lập như: các CSDL trong siêu thị, các CSDL lưu trữ các thông tin trên Internet, các CSDL về trình tự sinh học, các CSDL lưu trữ các tín hiệu thu được từ các sensor,... Nhu cầu phát triển các phương pháp cho phép tìm kiếm, nhận dạng các mẫu tiềm ẩn trong dữ liệu là cấp thiết và có nhiều ứng dụng. Ví dụ tìm các quy luật về hành vi mua hàng của khách hàng trong siêu thị, tìm các chuỗi bất thường trong dữ liệu chuỗi thời gian, tìm các quy luật phân lớp khách hàng trong kinh doanh, phân loại xếp hạng doanh nghiệp theo các đặc điểm của doanh nghiệp,... Đây là các bài toán có rất nhiều ứng dụng trong thực tế.

Giáo trình Khai phá dữ liệu tập trung trình bày các nội dung cơ bản liên quan đến khai phá dữ liệu như: tiền xử lý dữ liệu, bài toán tìm tập phổ biến và luật kết hợp, bài toán luật dãy, bài toán tìm luật phân lớp, lý thuyết tập thô và ứng dụng, bài toán phân tích gom cụm, xử lý văn bản. Bên cạnh đó, chúng tôi còn trình bày giải pháp khai phá dữ liệu của SQL Server và phần mềm khai phá dữ liệu Weka nhằm mục đích cụ thể hóa các thực hành khai phá dữ liệu. Sau mỗi chương của giáo trình, chúng tôi chuẩn bị phần bài tập để sinh viên có thể thực hành khai phá dữ liệu.

Mặc dù đã cố gắng trong quá trình biên soạn, nhưng chắc chắn giáo trình còn có những thiếu sót, chúng tôi rất mong nhận được góp ý của người đọc để sách ngày càng hoàn thiện hơn.

Trân trọng

PGS.TS Đỗ Phúc

MỤC LỤC