TS. LÊ VĂN PHÙNG - ThS. QUÁCH XUÂN TRƯỞNG

KHAI PHÁ Dữ liệu

Data Mining

NHÀ XUẤT BẢN THÔNG TIN VÀ TRUYỀN THÔNG

TS. LÊ VĂN PHÙNG - ThS. QUÁCH XUÂN TRƯỞNG

KHAI PHÁ Dữ liệu

Data Mining

NHÀ XUẤT BẢN THÔNG TIN VÀ TRUYỀN THÔNG

Mã số: HT 07 HM 12

LỜI NÓI ĐẦU

Cùng với sự phát triển như vũ bão của công nghệ thông tin, lượng thông tin của nhân loại được lưu trữ trên các thiết bị điện tử ngày một tăng. Nguồn dữ liệu khổng lồ ấy được tích lũy với tốc độ bùng nổ từ rất nhiều lĩnh vực: khoa học, kinh doanh, giao dịch, thương mại, chứng khoán,... Vậy chúng ta có thể khai thác được gì từ những "núi" dữ liệu tưởng chừng như "bỏ đi" ấy không?

Khai phá dữ liệu (Data Mining – DM) ra đời phần nào đó đã giải quyết hữu hiệu cho câu hỏi đặt ra ở trên. Và thế nào là khai phá dữ liệu? Khai phá dữ liệu là một quá trình khám phá, chắt lọc các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có của một công ty, đơn vị, tổ chức nào đó, từ đó giúp cho chúng ta có được quyết định sáng suốt.

Với mục đích cung cấp cho bạn đọc những kiến thức cơ bản về ***khai phá dữ liệu***, giai đoạn quan trọng có thể nói là bậc nhất trong chặng đường đi tìm tri thức trong các kho dữ liệu đồ sộ, Nhà xuất bản Thông tin và Truyền thông xuất bản cuốn sách "***Khai phá dữ liệu***" của TS. Lê Văn Phùng và ThS. Quách Xuân Trường, hiện đang công tác tại Trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên biên soạn giới thiệu với bạn đọc.

Cuốn sách gồm 11 chương chia làm ba phần giới thiệu các khái niệm phổ biến nhất liên quan đến quá trình phát hiện tri thức từ dữ liệu, các phương pháp khai phá các mẫu dữ liệu "hấp dẫn" ẩn chứa trong các tập dữ liệu lớn, một số thuật toán điển hình trong khai phá dữ liệu.

Phần 1 gồm 2 chương, trình bày các khái niệm cơ bản về khai phá dữ liệu như quá trình phát hiện tri thức từ dữ liệu và khai phá dữ liệu.

Phần 2 gồm 6 chương trình bày một số phương pháp khai phá các mẫu dữ liệu hấp dẫn ẩn chứa trong những tập dữ liệu lớn: Phương pháp cây quyết định; Phương pháp phân loại và hồi quy; Phương pháp phân cụm; Phương pháp kết hợp; Phương pháp giải thuật di truyền và Phương pháp mạng Nơ-ron.

Phần 3 gồm 3 chương giới thiệu một số thuật toán điển hình trong khai phá dữ liệu bằng các phương pháp phân cụm dữ liệu và bằng luật kết hợp.

Khai phá dữ liệu là một hướng tiếp cận mới tuy nhiên đã thu hút được rất nhiều sự quan tâm của các nhà nghiên cứu và phát triển nhờ vào những ứng dụng thực tiễn của chúng như: Phân tích dữ liệu và hỗ trợ ra quyết định; Điều trị y học; Tin sinh học; Tài chính và thị trường chứng khoán; Quản lý quan hệ khách hàng, Chăm sóc sức khỏe,...

Hy vọng cuốn sách sẽ thực sự hữu ích đối với các sinh viên, cử nhân, kỹ sư, giáo viên giảng dạy, cán bộ nghiên cứu chuyên ngành công nghệ thông tin trong công việc nghiệp vụ của mình. Cuốn sách cũng là tài liệu tham khảo bổ ích cho tất cả các bạn đọc yêu công nghệ thông tin và khao khát tìm tri thức trong các kho dữ liệu.

Nhà xuất bản Thông tin và Truyền thông xin trân trọng giới thiệu cùng bạn đọc và rất mong nhận được nhiều ý kiến đóng góp của quý vị. Mọi đóng góp của quý vị xin gửi về Nhà xuất bản Thông tin và Truyền thông - số 9, ngõ 90, phố Ngụy Như Kon Tum, quận Thanh Xuân, Hà Nội hoặc gửi trực tiếp cho tác giả theo địa chỉ lvphung@ioit.ac.vn.

NXB THÔNG TIN VÀ TRUYỀN THÔNG

TỪ VIẾT TẮT

1. Tiếng Anh

CLS	Concept Learning System	Thuật toán CLS
CURE	Clustering Using Represen tatives	Thuật toán CURE
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	Thuật toán DBSCAN
DENCLUE	Density - CLUstring Ering	Thuật toán DENCLUE
DM	Data Mining	Khai phá dữ liệu
EM	Expectation Maximization	Thuật toán EM
GA	Genetic Algorithm	Giải thuật di truyền
ID3	Interactive Dichotomizen 3	Thuật toán ID3
KDD	Knowledge Discovery from Data	Phát hiện tri thức từ dữ liệu
OLAP	On-Line Analytical Processing	Xử lý phân tích trực tuyến
PAM	Partitioning Around Medoids	Thuật toán PAM
SLIQ	Supervised Learning In Quest	Thuật toán phân lớp leo thang nhanh

2. Tiếng Việt

CNTT	Công nghệ thông tin
CSDL	Cơ sở dữ liệu
NSD	Người sử dụng
NST	Nhiễm sắc thể
PCDL	Phân cụm dữ liệu

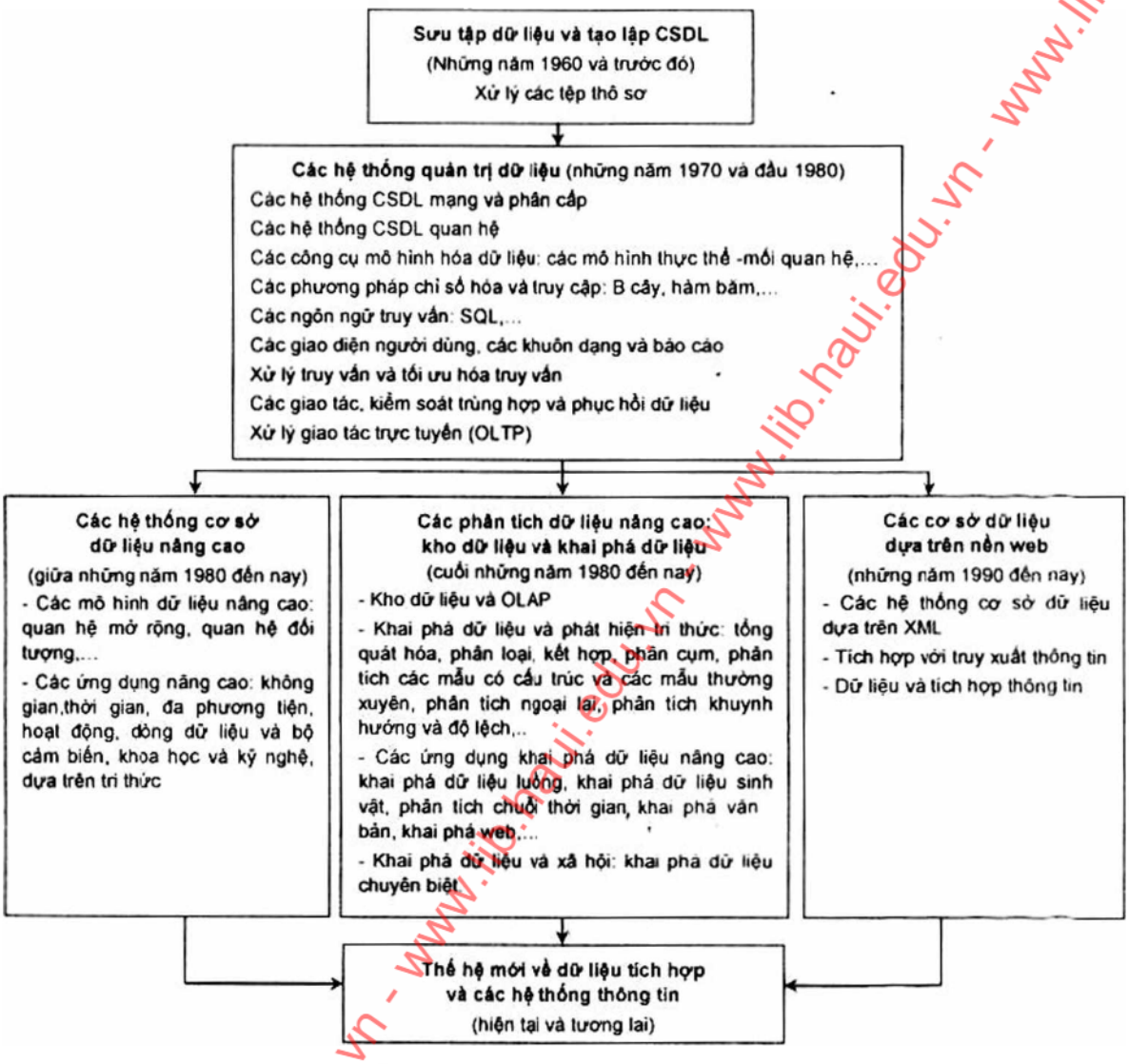
MỞ ĐẦU

Khai phá dữ liệu (Data Mining - DM) và *phát hiện tri thức từ dữ liệu (Knowledge Discovery from Data - KDD)* là một lĩnh vực non trẻ nhưng đầy hứa hẹn. Thông tin và tri thức đã khai phá được có thể sử dụng trong nhiều lĩnh vực từ phân tích thị trường, phát hiện gian lận, giữ chân khách hàng đến điều khiển sản xuất và nghiên cứu khoa học. DM có thể được xem như một kết quả tiến hóa tự nhiên của công nghệ thông tin (CNTT).

Quá trình phát triển hệ thống dữ liệu đã chứng kiến một nhánh tiến hóa thông qua sự phát triển các chức năng sau [28] (hình 1):

Từ những năm 1990, với sự phát triển mạnh mẽ của một loạt các công nghệ (vi xử lý, lưu trữ, truyền thông, thông tin), khối lượng dữ liệu tích lũy được đã tăng nhanh và dẫn đến bùng nổ dữ liệu trong nhiều lĩnh vực đời sống, xã hội, khoa học như thiên văn, hóa học, bảo mật, truyền thông, thương mại, dữ liệu Web, an ninh quốc phòng. Riêng Google tiếp nhận hơn 4 tỷ yêu cầu tìm kiếm mỗi ngày, lưu trữ hàng trăm terabytes dữ liệu, AT&T tiếp nhận 275 triệu cuộc gọi mỗi ngày, France Telecom có 30 terabytes thông tin về khách hàng, Walmart có 20 triệu giao dịch mỗi ngày, Europe`s Very Long Baseline Interforometry (VLBI) có 16 kính thiên văn, mỗi kính thu được 1 gigabits/giây dữ liệu, Cơ quan an ninh quốc phòng Hoa Kỳ (NSA) có trong tay hàng triệu văn bản về khủng bố, El nino cũng lưu trữ vài trăm gigabytes, Internet archive, www.archive.org cũng lưu trữ khoảng 300 terabytes. Người ta dự tính dữ liệu trên toàn cầu sẽ tăng gấp đôi trong vòng 9 tháng.

Sự phong phú về dữ liệu đồ sộ cùng với những nhu cầu về các công cụ phân tích dữ liệu mạnh đã nói lên rằng tình trạng giàu dữ liệu nhưng đói về thông tin như nhà bác học nổi tiếng Karan Sing đã từng nói rằng "*Chúng ta đang ngập chìm trong biển thông tin nhưng lại đang khát tri thức*" [61].



Hình 1. Sự tiến hóa của công nghệ hệ thống cơ sở dữ liệu

Do tăng trưởng nhanh, khối lượng cực lớn của dữ liệu được sưu tập và lưu giữ trong những kho chứa dữ liệu khổng lồ cũng như trên Internet đã vượt quá khả năng hấp thụ của con người nếu không có những công cụ mạnh. Kết quả là các dữ liệu đã được sưu tập trong những kho chứa khổng lồ đó đã trở thành "mồ chôn" dữ liệu. Do đó, những quyết định quan trọng thường không dựa vào những dữ liệu giàu thông tin trong kho chứa mà lại dựa vào quyết định trực giác của người thực hiện vì đơn giản rằng người ra quyết định không có công cụ nào chiết xuất được tri thức có giá trị được nhúng trong bể lớn dữ liệu đó. Hơn thế, các công nghệ kiểu hệ chuyên gia, công nghệ điển hình dựa vào những người sử dụng (NSD) hoặc các chuyên gia lĩnh vực,