

**ĐẠI HỌC THÁI NGUYÊN
KHOA CÔNG NGHỆ THÔNG TIN**

DƯƠNG THỊ NHUNG

**NGHIÊN CỨU PHƯƠNG PHÁP CÂY QUYẾT ĐỊNH
VÀ CÀI ĐẶT MÔ PHÒNG THUẬT TOÁN ID3**

Chuyên ngành: Khoa học máy tính

Mã số: 60.48.01

**LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH
NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS.TS VŨ ĐỨC THI**

Thái Nguyên 2010

LỜI CAM ĐOAN

Tôi xin cam đoan toàn bộ nội dung luận văn này là do tôi tự sưu tầm, tra cứu thông tin trên mạng internet, trong một số sách tham khảo để sắp xếp, hoàn thiện cho phù hợp với nội dung yêu cầu của đề tài.

Đến nay, nội dung luận văn của tôi chưa từng được công bố hay xuất bản dưới bất kỳ hình thức nào. Nếu sai tôi xin chịu hoàn toàn trách nhiệm.

Ngày 10 tháng 10 năm 2010

Tác giả

Dương Thị Nhung

LỜI CẢM ƠN

Để hoàn thành luận văn này tôi đã nhận được sự giúp đỡ tận tình của các thầy cô Khoa Công nghệ thông tin – Đại học Thái Nguyên, các thầy cô viện công nghệ thông tin – viện Khoa học và công nghệ Việt Nam, các bạn bè đồng nghiệp. Đặc biệt là PGS.TS Vũ Đức Thi, người thầy trực tiếp hướng dẫn tôi trong quá trình nghiên cứu và thực hiện luận văn.

Nhân dịp này tôi xin được bày tỏ lời cảm ơn tới tất cả các thầy cô giáo viện Công nghệ thông tin – Viện Khoa học và công nghệ Việt Nam, các thầy cô ở khoa Công nghệ thông tin – Đại học Thái nguyên đã giảng dạy và tạo mọi điều kiện thuận lợi giúp đỡ chúng tôi trong quá trình học tập, nghiên cứu.

Tôi xin trân trọng cảm ơn PGS.TS. Vũ Đức Thi – Viện công nghệ thông tin, người thầy trực tiếp hướng dẫn, đưa ra ý tưởng, định hướng, đóng góp các ý kiến chuyên môn và tận tình giúp đỡ tôi trong suốt quá trình nghiên cứu và thực hiện luận văn này.

Tôi xin cảm ơn các bạn bè đồng nghiệp và gia đình đã giúp đỡ, đóng góp ý kiến và động viên tôi trong suốt quá trình học, quá trình nghiên cứu và hoàn thành luận văn này.

Tác giả

Dương Thị Nhung

MỤC LỤC

LỜI MỞ ĐẦU	1
Phần 1: TỔNG QUAN VỀ PHÁT HIỆN TRI THỨC VÀ	2
KHAI PHÁ DỮ LIỆU	2
1.1. Khái quát chung về phát hiện tri thức và khai phá dữ liệu.....	2
1.2. Quá trình khám phá tri thức.....	3
1.2.1. Hình thành và định nghĩa bài toán	4
1.2.2. Thu thập và tiền xử lý dữ liệu	4
1.2.3. KPDL và rút ra các tri thức	5
1.2.4. Phân tích và kiểm định kết quả	5
1.2.5. Sử dụng các tri thức phát hiện được	5
1.3. Quá trình KPDL	6
1.3.1. Gom dữ liệu (gatherin)	6
1.3.2. Trích lọc dữ liệu (selection)	7
1.3.3. Làm sạch và tiền xử lý dữ liệu (cleansing preprocessing preparation).	7
1.3.4. Chuyển đổi dữ liệu (transformation)	7
1.3.5. Phát hiện và trích mẫu dữ liệu (pattern extraction and discovery)	7
1.3.6. Đánh giá kết quả (evaluation of result)	7
1.4. Chức năng của KPDL	8
1.5. Các kỹ thuật KPDL	8
1.5.1. Phân lớp dữ liệu:	8
1.5.2. Phân cụm dữ liệu:	9
1.5.3. Khai phá luật kết hợp:	9
1.5.4. Hồi quy:	9
1.5.5. Giải thuật di truyền:	9
1.5.6. Mạng nơron:	9
1.5.7. Cây quyết định.	10
1.6. Các dạng dữ liệu có thể khai phá được	10
1.7. Các lĩnh vực liên quan và ứng dụng của KPDL	10

1.7.1. Các lĩnh vực liên quan đến khám phá tri thức và KPDL.....	10
1.7.2. Ứng dụng của KPDL	11
1.8. Các thách thức và hướng phát triển của KPDL.....	12
Phần 2: CÂY QUYẾT ĐỊNH VÀ CÁC THUẬT TOÁN KHAI PHÁ DỮ LIỆU BẰNG CÂY QUYẾT ĐỊNH	13
2.1 Cây quyết định.....	13
2.1.1 Mô tả	13
2.1.2 Định nghĩa cây quyết định	13
2.1.3 Ưu điểm của cây quyết định.....	15
2.1.4 Vấn đề xây dựng cây quyết định	16
2.1.5 Rút ra các luật từ cây quyết định.	17
2.2 Các thuật toán KPDL bằng cây quyết định	18
2.2.1 Thuật toán CLS.....	18
2.2.2. Thuật toán ID3.....	23
2.2.3. Thuật toán C4.5	38
2.2.4. Thuật toán SLIQ.....	52
2.2.5. Cắt tỉa cây quyết định	61
2.2.6. Đánh giá và kết luận về các thuật toán xây dựng cây quyết định.	66
Phần 3: CÀI ĐẶT MÔ PHỎNG THUẬT TOÁN ID3.....	69
3.1. Mô tả bài toán	69
3.2. Màn hình nhập dữ liệu của chương trình.....	69
3.3. Màn hình phân tích dữ liệu đưa ra kết quả của chương trình.....	70
Phần 4: KẾT LUẬN	71
TÀI LIỆU THAM KHẢO	72
Tài Liệu Tiếng Việt.....	72
Tài Liệu Tiếng Anh.....	72
Danh Sách Website.....	72

LỜI MỞ ĐẦU

Trong nhiều năm qua, cùng với sự phát triển của công nghệ thông tin và ứng dụng của công nghệ thông tin trong nhiều lĩnh vực của đời sống xã hội, thì lượng dữ liệu được các cơ quan thu thập và lưu trữ ngày một nhiều lên. Theo thống kê thì trong lượng dữ liệu đó chỉ có khoảng từ 5% đến 10% dữ liệu là luôn được phân tích, số còn lại người ta không biết sẽ phải làm gì với chúng. Tuy nhiên con người vẫn tiếp tục thu thập và lưu trữ dữ liệu vì cho rằng chúng ẩn chứa những giá trị nhất định nào đó, chẳng hạn có thể cung cấp cho họ thông tin đưa ra quyết định chiến lược một cách nhanh chóng trong một lúc nào đó. Chính do lượng dữ liệu được lưu trữ ngày càng nhiều lên đến mức khổng lồ thì các phương pháp quản trị và khai thác dữ liệu truyền thống ngày càng không đáp ứng được nhu cầu của thực tế và đã làm phát triển một khuynh hướng kỹ thuật mới đó là Kỹ thuật khám phá tri thức và KPD L (*Khai phá dữ liệu*).

Kỹ thuật khám phá tri thức và KPD L đã và đang được nghiên cứu, ứng dụng trong nhiều lĩnh vực khác nhau trên thế giới. Tại Việt Nam, kỹ thuật này còn tương đối mới mẻ tuy nhiên cũng đang được nghiên cứu và dần đưa vào một số ứng dụng thực tế; đặc biệt trong lĩnh vực giáo dục đào tạo thì chưa có một ứng dụng thiết thực nào. Hiện nay vấn đề khám phá tri thức và KPD L cũng đang thu hút được sự quan tâm của nhiều người và nhiều công ty phát triển ứng dụng công nghệ thông tin ở nước ta.

Luận văn này bao gồm các nội dung sau:

Chương 1: Tổng quan về phát hiện tri thức và khai phá dữ liệu.

Chương 2: Cây quyết định và các thuật toán khai phá dữ liệu bằng cây quyết định.

Chương 3: Cài đặt mô phỏng thuật toán ID3

Phần 1: TỔNG QUAN VỀ PHÁT HIỆN TRI THỨC VÀ KHAÍ PHÁ DỮ LIỆU

1.1. Khái quát chung về phát hiện tri thức và khai phá dữ liệu.

Trong vài thập kỷ gần đây, khả năng tạo sinh và lưu trữ dữ liệu của con người đã tăng lên nhanh chóng. Lượng dữ liệu lớn được lưu trữ dẫn đến một đòi hỏi cấp bách phải có những kỹ thuật mới, những công cụ tự động mới trợ giúp con người một cách thông minh trong việc chuyển đổi một lượng lớn dữ liệu thành thông tin hữu ích và tri thức. Vì vậy mà kỹ thuật khám phá tri thức (*Knowledge Discovery*) đã ra đời và ngày càng phát triển để đáp ứng nhu cầu của con người trong việc xử lý các kho dữ liệu lớn.

Vậy tri thức ở đây là gì? Thông thường chúng ta coi dữ liệu như là một dãy các bit, các số và các ký hiệu, hoặc các “đối tượng” được gửi cho một chương trình dưới một định dạng nhất định nào đó. Chúng ta sử dụng các bit để đo lường thông tin và xem nó như là dữ liệu đã được lọc bỏ dư thừa, được rút gọn tới mức tối thiểu. Bit được dùng làm đơn vị đặc trưng cho dữ liệu. Chúng ta có thể xem tri thức như là các thông tin tích hợp, bao gồm các sự kiện và các mối quan hệ giữa chúng. Các mối quan hệ này có thể được hiểu, được phát hiện ra, hoặc có thể được học. Nói cách khác, tri thức có thể coi là dữ liệu có độ trừu tượng và tổ chức cao.

Hiện nay khám phá tri thức đang phát triển mạnh mẽ trong nhiều ngành học thuật. Nó được kết hợp cùng với việc quản lý cơ sở dữ liệu, khoa học thống kê, học máy, nghiên cứu mối quan hệ giữa các lĩnh vực nhằm rút ra các tri thức có ích từ tập hợp lớn dữ liệu.

Khám phá tri thức là quá trình nhận biết cái logic, cái mới lạ, những tri thức tiềm tàng hữu ích từ cơ sở dữ liệu, và cuối cùng là việc hiểu được các mẫu các mô hình trong dữ liệu.

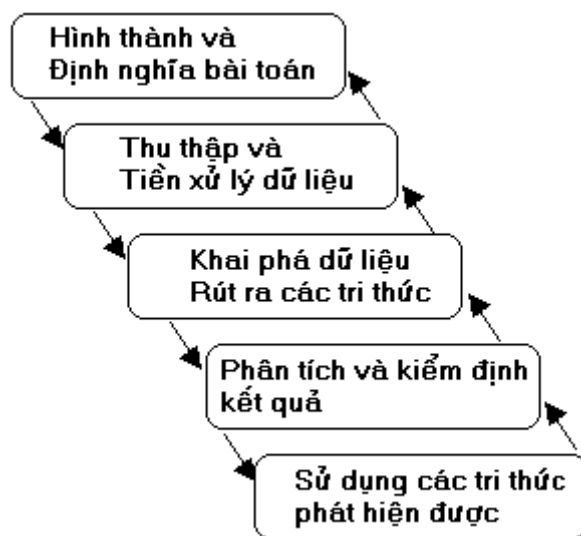
Còn thuật ngữ KPD (Khai phá dữ liệu) ra đời vào những năm cuối của thập kỷ 1980. Có rất nhiều định nghĩa khác nhau về "KPD" đã được đưa ra. Giáo sư Tom

Mitchell đã đưa ra định nghĩa về KPDL như sau: “KPDL là việc sử dụng dữ liệu lịch sử để khám phá những quy tắc và cải thiện những quyết định trong tương lai.”. Với một cách tiếp cận ứng dụng hơn, tiến sĩ Fayyad đã phát biểu: ”KPDL thường được xem là việc khám phá tri thức trong các cơ sở dữ liệu, là một quá trình trích xuất những thông tin ẩn, trước đây chưa biết và có khả năng là hữu ích, dưới dạng các quy luật, ràng buộc, quy tắc trong cơ sở dữ liệu.”. Còn các nhà thống kê thì xem " KPDL như là một quá trình phân tích được thiết kế thăm dò một lượng cực lớn các dữ liệu nhằm phát hiện ra các mẫu thích hợp và/ hoặc các mối quan hệ mang tính hệ thống giữa các biến và sau đó sẽ hợp thức hoá các kết quả tìm được bằng cách áp dụng các mẫu đã phát hiện được cho tập con mới của dữ liệu". Nói chung, KPDL là cốt lõi của quá trình khám phá tri thức. Nó gồm có các giải thuật KPDL chuyên dùng, dưới một số quy định về hiệu quả tính toán chấp nhận được. KPDL là nhằm tìm ra những mẫu mới, mẫu có tính chất không tầm thường, những thông tin tiềm ẩn mang tính dự đoán chưa được biết đến và có khả năng mang lại ích lợi.

Nói tóm lại, mục đích chung của việc khám phá tri thức và KPDL là tìm ra các mẫu được quan tâm nhất và/ hoặc các mô hình tồn tại trong cơ sở dữ liệu, nhưng chúng lại bị che dấu bởi một số lượng lớn dữ liệu.

1.2. Quá trình khám phá tri thức

Quá trình khám phá tri thức được tiến hành qua 5 bước sau:



Hình 1.1. Quá trình khám phá tri thức

1.2.1. Hình thành và định nghĩa bài toán

Đây là bước tìm hiểu lĩnh vực ứng dụng và hình thành bài toán, bước này quyết định cần rút ra những tri thức dạng như thế nào, đồng thời lựa chọn các phương pháp KPDL thích hợp với mục đích ứng dụng và bản chất của dữ liệu.

1.2.2. Thu thập và tiền xử lý dữ liệu

Trong bước này dữ liệu được thu thập ở dạng thô (nguồn dữ liệu thu thập có thể là từ các kho dữ liệu hay nguồn thông tin internet). Trong giai đoạn này dữ liệu cũng được tiền xử lý để biến đổi và cải thiện về chất lượng cho phù hợp với phương pháp KPDL được chọn lựa trong bước một.

Bước này thường chiếm nhiều thời gian nhất trong quá trình khám phá tri thức.

Các công việc tiền xử lý dữ liệu bao gồm :

1. Xử lý dữ liệu bị mất/ thiếu: Các dữ liệu bị thiếu sẽ được thay thế bởi các giá trị thích hợp.
2. Khử sự trùng lặp: các đối tượng dữ liệu trùng lặp sẽ bị loại bỏ. Kỹ thuật này không được sử dụng cho các tác vụ có quan tâm đến phân bố dữ liệu.
3. Giảm nhiễu: dữ liệu nhiễu và các đối tượng tách rời khỏi phân bố chung sẽ bị loại bỏ khỏi tập dữ liệu.

4. Chuẩn hoá: thông thường là chuẩn hoá miền giá trị của dữ liệu cho phù hợp.
5. Rời rạc hoá: chính là việc biến đổi các dữ liệu dạng số về dữ liệu với các giá trị rời rạc.
6. Rút trích và xây dựng đặc trưng mới từ các thuộc tính đã có.
7. Giảm chiều: là loại bỏ bớt các thuộc tính chứa ít thông tin.

1.2.3. KPDL và rút ra các tri thức

Đây là bước quan trọng nhất trong tiến trình khám phá tri thức. Kết quả của bước này là trích ra được các mẫu và/hoặc các mô hình ẩn dưới một khối lượng lớn dữ liệu. Một mô hình có thể là một biểu diễn cấu trúc tổng thể một thành phần của hệ thống hay cả hệ thống trong cơ sở dữ liệu, hoặc miêu tả cách dữ liệu được nảy sinh. Còn một mẫu là một cấu trúc cục bộ có liên quan đến vài biến và vài trường hợp trong cơ sở dữ liệu.

1.2.4. Phân tích và kiểm định kết quả

Bước thứ tư là hiểu các tri thức đã tìm được, đặc biệt là làm sáng tỏ các mô tả và dự đoán. Trong bước này, kết quả tìm được sẽ được biến đổi sang dạng phù hợp với lĩnh vực ứng dụng và dễ hiểu hơn cho người dùng.

1.2.5. Sử dụng các tri thức phát hiện được

Trong bước này, các tri thức khám phá được sẽ được củng cố, kết hợp lại thành một hệ thống, đồng thời giải quyết các xung đột tiềm năng trong các tri thức đó. Các mô hình rút ra được đưa vào những hệ thống thông tin thực tế dưới dạng các môđun hỗ trợ việc đưa ra quyết định.

Các giai đoạn của quá trình khám phá tri thức có mối quan hệ chặt chẽ với nhau trong bối cảnh chung của hệ thống. Các kỹ thuật được sử dụng trong giai đoạn trước có thể ảnh hưởng đến hiệu quả của các giải thuật được sử dụng trong những giai đoạn tiếp theo. Các bước của quá trình khám phá tri thức có thể được lặp đi lặp