ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CNTT VÀ TRUYỀN THÔNG

NGUYỄN VĂN SỰ

KHAI PHÁ DỮ LIỆU BẰNG CÂY QUYẾT ĐỊNH VÀ ỨNG DỤNG TRONG HỆ HỖ TRỢ QUYẾT ĐỊNH

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Thái Nguyên - 2012

MỤC LỤC

LỜI NÓI ĐẦU

Song song với sự phát triển không ngừng của ngành Công nghệ thông tin nói chung và các lĩnh vực ngành công nghệ phần mềm nói riêng, hệ thống các kho dữ liệu phục vụ trong công tác hỗ trợ ra quyết định và việc phân loại các thông tin cũng như nhu cầu lưu trữ thông tin ngày càng cần thiết. Bên cạnh đó việc tin học hóa trong các công tác quản lý cũng như nhiều lĩnh vực, hoạt động khác đã tạo ra cho nhân loại một thư viện dữ liệu khổng lồ, sẵn sàng phục vụ bất cứ ai quan tâm. Đối với chúng ta nó là một trong những nguồn tài nguyên thông tin vô cùng giá trị, việc tận dụng kho dữ liệu này để làm cơ sở cho việc hỗ trợ ra quyết định trong công tác quản lý mang lại hiệu quả đáng kể. Nhưng vấn đề là chúng ta cần phải phân loại nguồn tài nguyên đó như thế nào để sử dụng có hiệu quả nhất trong từng lĩnh vực cụ thể. Để tìm được đúng thông tin cần tìm trong nguồn tài nguyên khổng lồ này làm cơ sở hỗ trợ ra quyết định trong mọi hoạt động là một thách thức lớn.

Chính vì vậy mà hiện nay các phương pháp quản trị và khai thác cơ sở dữ liệu truyền thống ngày càng không đáp ứng được nhu cầu thực tế, từ những hiện trạng cũng như thách thức này đã làm phát triển một khuynh hướng kỹ thuật mới nhằm giải quyết bài toán này, đó là Kỹ thuật phát hiện tri thức và khai phá dữ liệu (Knowledge Discovery and Data Mining).

Kỹ thuật phát hiện tri thức và khai phá dữ liệu đã và đang được nghiên cứu, ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau ở các nước trên thế giới, tại Việt Nam kỹ thuật này cũng đã và đang được nghiên cứu và dần đưa vào ứng dụng. Bước quan trọng nhất của quá trình này là khai phá dữ liệu (Data Mining), giúp người sử dụng thu được những tri thức hữu ích từ những cơ sở dữ liệu (CSDL) hoặc các nguồn dữ liệu khổng lồ khác để từ đó làm cơ sở ra quyết định xử lý đối với dữ liệu thu được. Rất nhiều tổ chức trên thế giới đã ứng dụng kĩ thuật khai phá dữ liệu vào công tác quản lý đã thu được những lợi ích to lớn. Để làm được điều đó, sự phát triển của các mô hình toán học và các giải thuật hiệu quả chính là chìa

khoá quan trọng. Vì vậy, trong luận văn này tác giả sẽ đề cập tới một số kỹ thuật Khai phá dữ liệu bằng cây quyết định và ứng dụng nó trong hệ hỗ trợ quyết định.

Luận văn gồm 3 chương với nội dung sau:

Chương 1: Trình bày tổng quan về khai phá dữ liệu, các khái niệm cơ bản, các bước thực hiện, các chức năng, kỹ thuật khai phá dữ liệu, và ứng dụng của khai phá dữ liệu trong các lĩnh vực.

Chương 2: Trình bày các khái niệm về cây quyết định, các kiểu cây quyết định và các kỹ thuật khai phá dữ liệu bằng cây quyết định, kỹ thuật cắt tỉa cây quyết định.

Chương 3: Trình bày bài toán ra quyết định khen thưởng trong ngành giáo dục, các quy trình xét duyệt và ra quyết định khen thưởng, xác định yêu cầu bài toán, lựa chọn thuật toán để cài đặt xây dựng công cụ hỗ trợ ra quyết định khen thưởng trong công tác quản lý thi đua khen thưởng của Bộ Giáo dục và Đào tạo.

Chương 1
KHAI PHÁ DỮ LIỆU

1.1. Khám phá tri thức và khai phá dữ liệu

Khám phá tri thức (*Knowledge Discovery*) trong các cơ sở dữ liệu, kho dữ liệu là một quy trình gồm nhiều công đoạn để nhận biết các mẫu hoặc các mô hình trong dữ liệu với các tính năng: hợp thức, mới, khả ích, và có thể hiểu được [18].

Khai phá dữ liệu là việc sử dụng dữ liệu lịch sử để khám phá những qui tắc và cải thiện những quyết định trong tương lai.

Khai phá dữ liệu 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. Mục đích của khai phá dữ liệu là:

- Rút trích thông tin hữu ích, chưa biết, các mẫu hoặc các mô hình tiềm ẩn trong khối dữ liệu lớn dưới dạng các quy luật, ràng buột, quy tắc trong cơ sở dữ liệu.
- Phân tích dữ liệu bán tự động.
- Giải thích dữ liệu trên các tập dữ liệu lớn.

Khai phá dữ liệu là một bước trong quy trình khám phá tri thức để hỗ trợ ra quyết định, dự báo và khái quát dữ liệu.

1.2. Tại sao phải khai phá dữ liệu

Ước tính cứ mỗi năm lượng thông tin trên thế giới lại tăng lên khoảng 2 lần. Chính vì vậy, hiện nay dữ liệu mà con người thu thập và lưu trữ trong các kho dữ liệu là vô cùng lớn, thậm chí lớn đến mức vượt quá khả năng kiểm soát,… Cũng bởi lý do này các nhà khoa học đã đề cập đến việc tổ chức lại dữ liệu sao cho hiệu

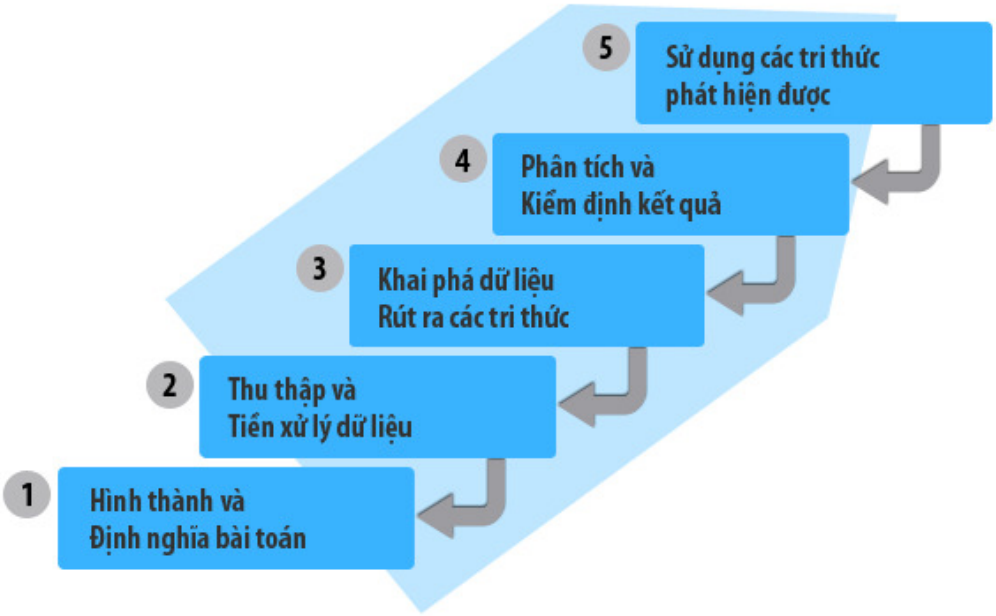
quả, đáp ứng được yêu cầu chất lượng ngày càng cao nhằm hỗ trợ những nhà quản lý ra quyết định trong các tổ chức quản lý tài chính, thương mại, khoa học,...

Với lượng dữ liệu tăng nhanh hàng năm, rõ ràng các phương pháp thủ công truyền thống áp dụng để phân tích dữ liệu sẽ không hiệu quả, tốn kém và dễ dẫn đến những sai lệch. Do đó, để có thể sử dụng hiệu quả hơn nữa các cơ sở dữ liệu lớn thì nhất thiết cần phải có những kỹ thuật mới, và kỹ thuật khai phá dữ liệu đã được các nhà khoa học đề cập tới.

Khai phá dữ liệu là một lĩnh vực khoa học nhằm tự động hóa khai thác những thông tin, tri thức hữu ích, tiềm ẩn trong các CSDL cho các tổ chức, doanh nghiệp,... Các kết quả nghiên cứu cùng với những ứng dụng thành công trong khai phá dữ liệu, khám phá tri thức cho thấy khai phá dữ liệu là một lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích, đồng thời có ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống. Hiện nay, khai phá dữ liệu được ứng dụng rộng rãi trong các lĩnh vực như: Phân tích dữ liệu hỗ trợ ra quyết định, điều trị y học, tin-sinh học, thương mại, tài chính, bảo hiểm, text mining, web mining,...

1.3. 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 [18]

- Bước 1: Hình thành và định nghĩa bài toán

Đây là bước tìm hiểu lĩnh vực ứng dụng và hình thành bài toán, bước này sẽ quyết định cho việc rút ra những tri thức hữu ích, đồng thời lựa chọn các phương pháp khai phá dữ liệu thích hợp với mục đích của ứng dụng và bản chất của dữ liệu.

- Bước 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 khác từ internet). Trong giai đoạn này dữ liệu cũng được tiền xử lý để biến đổi và cải thiện chất lượng dữ liệu cho phù hợp với phương pháp khai phá dữ liệu được chọn lựa trong bước trên.

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

Các giải thuật tiền xử lý dữ liệu bao gồm:

1. Xử lý dữ liệu bị mất/thiếu: các dạng 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 đi.
3. Giảm nhiễu: nhiễu và các đối tượng tách rời khỏi phân bố chung sẽ bị loại đi khỏi dữ liệu.
4. Chuẩn hoá: miền giá trị của dữ liệu sẽ được chuẩn hoá.
5. Rời rạc hoá: các dạng dữ liệu số sẽ được biến đổi ra các giá trị rời rạc.
6. Rút trích và xây dựng đặc trưng mới từ các thuộc tính đã có.
7. Giảm chiều: các thuộc tính chứa ít thông tin sẽ được loại bỏ bớt.

- Bước 3: Khai phá dữ liệu 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 các 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, hay 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.

- Bước 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.

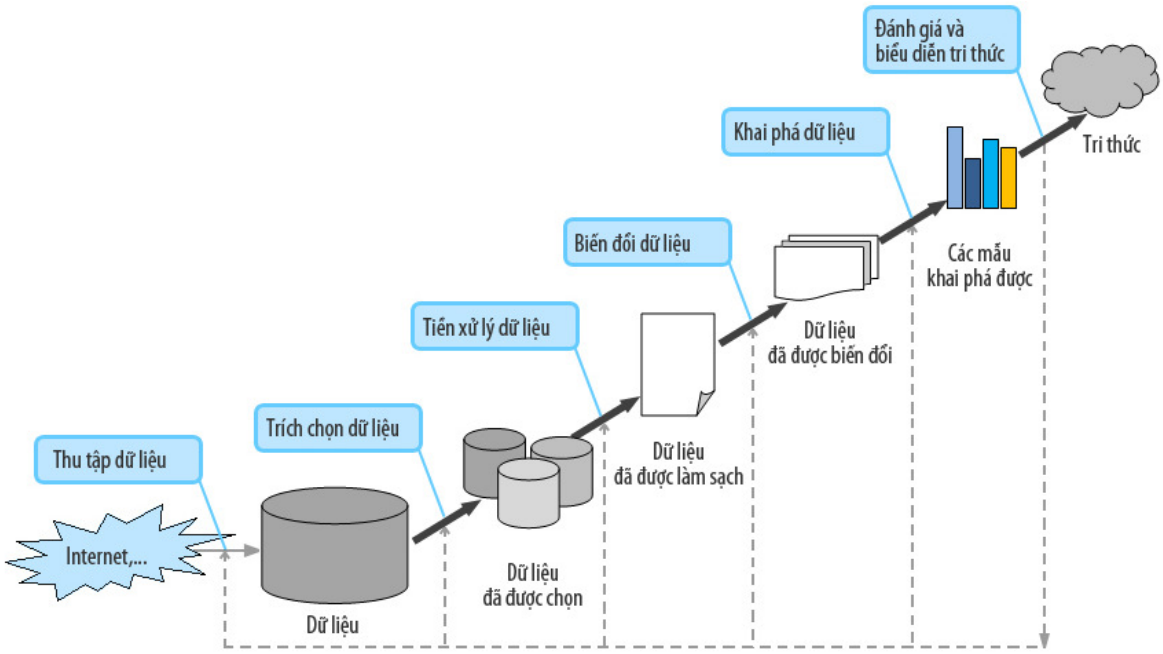
- Bước 5: Sử dụng các tri thức phát hiện được

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 module 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 các giai đoạn tiếp theo. Các bước của quá trình khám phá tri thức có thể được lặp đi lặp lại một số lần, kết quả thu được có thể được lấy trung bình trên tất cả các lần thực hiện.

1.4. Trình tự thực hiện trong quá trình khai phá dữ liệu

Khai phá dữ liệu là hoạt động trọng tâm của quá trình khám phá tri thức. Thuật ngữ khai phá dữ liệu còn được một số nhà khoa học gọi là phát hiện tri thức trong cơ sở dữ liệu (Knowledge discovery in database) (theo Fayyad Smyth và Piatestky-Shapiro 1989). Quá trình này gồm có 6 bước [1]:



Hình 1.2. Quá trình khai phá dữ liệu

Quá trình khai phá dữ liệu bắt đầu với kho dữ liệu thô và kết thúc với tri thức được chiết xuất ra. Nội dung của quá trình như sau:

- Gom dữ liệu (gatherin)

Tập hợp dữ liệu là bước đầu tiên trong khai phá dữ liệu. Bước này lấy dữ liệu từ trong một cơ sở dữ liệu, một kho dữ liệu, thậm chí dữ liệu từ những nguồn cung ứng web.

- Trích lọc dữ liệu (selection)

Ở giai đoạn này dữ liệu được lựa chọn và phân chia theo một số tiêu chuẩn nào đó.

- Làm sạch và tiền xử lý dữ liệu (cleansing preprocessing)

Giai đoạn thứ ba này là giai đoạn thường bị bỏ quên, nhưng thực tế nó là một bước rất quan trọng trong quá trình khai phá dữ liệu. Một số lỗi thường mắc phải trong khi gom dữ liệu là dữ liệu không đầy đủ hoặc không thống nhất, thiếu chặt chẽ, vô nghĩa (ví dụ như: con người có chiều cao = 4 mét → điều này là vô lý), do vậy ở giai đoạn thứ ba này nhằm xử lý các dữ liệu như trên (dữ liệu vô nghĩa, dữ