

TỔNG QUAN VỀ PHÁT HIỆN TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

ThS. Trần Hùng Cường, ThS. Ngô Đức Vĩnh

Khoa Công nghệ thông tin

Trường Đại học Công nghiệp Hà Nội

Sự phát triển của công nghệ thông tin và việc ứng dụng công nghệ thông tin ở hầu hết các lĩnh vực trong nhiều năm qua cũng đồng nghĩa với lượng dữ liệu đã được thu thập và lưu trữ ngày càng lớn. Các hệ quản trị cơ sở dữ liệu truyền thống cũng chỉ khai thác được một lượng thông tin nhỏ không còn đáp ứng đầy đủ những yêu cầu, những thách thức mới. Do vậy một khuynh hướng mới được ra đời đó là kỹ thuật phát hiện tri thức trong cơ sở dữ liệu. Xin giới thiệu một cách tổng quan về phát hiện tri thức và khai phá dữ liệu cùng một số kỹ thuật cơ bản để trong khai phá dữ liệu để phát hiện tri thức và một số ứng dụng trong thực tế nhằm hỗ trợ cho tiến trình ra quyết định.

1. TỔ CHỨC VÀ KHAI THÁC CƠ SỞ DỮ LIỆU TRUYỀN THỐNG

Việc sử dụng các kỹ thuật tin học để tổ chức và khai thác các cơ sở dữ liệu (CSDL) đã được phát triển từ những năm 60. Từ đó cho đến nay có nhiều cơ sở dữ liệu đã được tổ chức, phát triển và khai thác ở mọi quy mô và ở khắp các lĩnh vực hoạt động của con người và xã hội. Với sự phát triển mạnh mẽ của công nghệ điện tử tạo ra các bộ nhớ có dung lượng lớn, bộ xử lý tốc độ cao cùng với các hệ thống mạng viễn thông, người ta đã xây dựng các hệ thống thông tin nhằm tự động hóa mọi hoạt động kinh doanh giao dịch của mình. Điều này đã tạo ra một dòng dữ liệu tăng lên không ngừng vì ngay từ các giao dịch đơn giản nhất như kiểm tra sức khỏe, sử dụng thẻ tín dụng,... đều

được ghi vào trong máy tính. Cho đến nay, con số này đã trở nên khổng lồ, bao gồm các cơ sở dữ liệu cực lớn cỡ gigabytes và thậm chí terabytes, chẳng hạn như dữ liệu bán hàng, dữ liệu các tài khoản, các khoản vay, sử dụng vốn,... Nhiều hệ quản trị cơ sở dữ liệu mạnh với các công cụ phong phú và thuận tiện đã giúp con người khai thác có hiệu quả các nguồn tài nguyên dữ liệu. Mô hình cơ sở dữ liệu quan hệ và ngôn ngữ truy vấn chuẩn (SQL) đã có vai trò hết sức quan trọng trong việc tổ chức và khai thác các cơ sở dữ liệu đó.

2. BƯỚC PHÁT TRIỂN CỦA VIỆC TỔ CHỨC VÀ KHAI THÁC CÁC CSDL

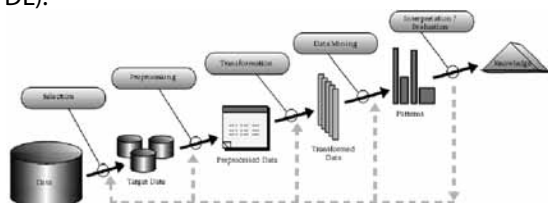
Cùng với việc tăng không ngừng khối lượng dữ liệu, các hệ thống thông tin cũng được chuyên môn hóa, phân hoạch theo các lĩnh vực ứng dụng như sản xuất, tài chính, buôn bán thị trường v.v. Như vậy, bên cạnh chức năng khai thác dữ liệu có tính chất tác nghiệp, sự thành công trong kinh doanh không còn là năng suất của các hệ thống thông tin nữa mà là tính linh hoạt và sẵn sàng đáp lại những yêu cầu trong thực tế, CSDL cần đem lại những "tri thức" hơn là chính những dữ liệu đó. Các quyết định cần phải có càng nhanh càng tốt và phải chính xác dựa trên những dữ liệu sẵn có. Lúc này các mô hình CSDL truyền thống và ngôn ngữ SQL đã cho thấy không có khả năng thực hiện công việc này.

Để lấy được tri thức trong khối dữ liệu khổng lồ này, người ta đã đi tìm những kỹ thuật có khả năng hợp nhất các dữ liệu từ các hệ thống giao dịch khác

nhau, chuyển đổi thành một tập hợp các cơ sở dữ liệu ổn định, có chất lượng, chỉ được sử dụng riêng cho một vài mục đích nào đó. Các kỹ thuật đó được gọi chung là kỹ thuật tạo kho dữ liệu (data warehousing) và môi trường các dữ liệu có được gọi là các kho dữ liệu (data warehouse). Với những thách thức như vậy, các nhà nghiên cứu đã đưa ra một phương pháp mới trên kho dữ liệu đáp ứng cả nhu cầu trong khoa học cũng như trong hoạt động thực tiễn. Đó chính là công nghệ phát hiện tri thức từ cơ sở dữ liệu.

3. QUÁ TRÌNH PHÁT HIỆN TRI THỨC VÀ KHAI PHÁ DỮ LIỆU

Một vấn đề rất quan trọng để dẫn đến thành công là việc biết sử dụng thông tin một cách có hiệu quả. Điều đó có nghĩa là từ các dữ liệu sẵn có phải tìm ra những thông tin tiềm ẩn có giá trị mà trước đó chưa được phát hiện, phải tìm ra những xu hướng phát triển và những yếu tố tác động lên chúng. Thực hiện công việc đó chính là thực hiện quá trình phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database – KDD) mà trong đó kỹ thuật này cho phép ta lấy được các tri thức chính là pha khai phá dữ liệu (KPD).



Hình 1. Quá trình phát hiện tri thức

Trước khi đi vào tìm hiểu các giai đoạn trong KDD ta đưa ra một số ví dụ để phân biệt ba khái niệm: dữ liệu, thông tin và tri thức. Dữ liệu thường được cho bởi các giá trị mô tả các sự kiện, hiện tượng cụ thể. Còn tri thức (knowledge) là gì? Có thể có những định nghĩa rõ ràng để phân biệt các khái niệm dữ liệu, thông tin và tri thức hay không? Tri thức là một khái niệm rất trừu tượng. Do đó, chúng ta sẽ không cố gắng đưa ra một định nghĩa hình thức chính xác ở đây. Thay vào đó, chúng ta hãy cùng nhau cảm nhận khái niệm tri thức bằng cách so sánh nó với hai khái niệm khác là thông tin và dữ liệu.

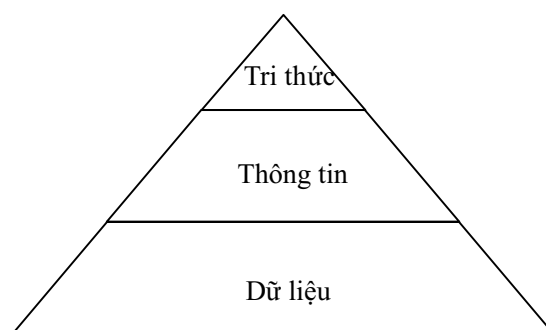
Nhà bác học nổi tiếng Karan Sing đã từng nói rằng “Chúng ta đang ngập chìm trong biển thông tin nhưng lại đang khát tri thức”. Câu nói này làm nổi bật sự khác biệt về lượng lẫn về chất giữa hai khái niệm

thông tin và tri thức. Cũng có thể quan niệm thông tin là quan hệ giữa các dữ liệu. Các dữ liệu được sắp xếp theo một thứ tự hoặc được tập hợp lại theo một quan hệ nào đó sẽ chứa đựng thông tin. Nếu những quan hệ này được chỉ ra một cách rõ ràng thì đó là các tri thức. Chẳng hạn

Trong toán học: Bản thân từng con số riêng lẻ như 1, 1, 3, 5, 2, 8, 13, ... là các dữ liệu. Tuy nhiên, khi đặt chúng lại với nhau theo trật tự 1, 1, 2, 3, 5, 8, 13, 21, 34, ... thì giữa chúng đã bắt đầu có một mối liên hệ. Mối liên hệ này có thể được biểu diễn bằng công thức $Un = Un-1 + Un-2$ nếu $n \geq 3$. Công thức nêu trên chính là tri thức.

Trong vật lý: Từ bảng số liệu về điện trở (R), điện thế (U) và cường độ dòng điện (I) trong một mạch điện. Bản thân những con số trong các cột của bảng trên không có mấy ý nghĩa nếu ta tách rời chúng ta. Nhưng khi đặt kế nhau, chúng đã cho thấy có một sự liên hệ nào đó. Và mối liên hệ này có thể được diễn tả bằng công thức $I = U/R$, công thức này là tri thức.

Như vậy, so với dữ liệu thì tri thức có số lượng ít hơn rất nhiều. Thuật ngữ ít ở đây không chỉ đơn giản là một dấu nhỏ hơn bình thường mà là sự kết tinh hoặc cô đọng lại. Ta hãy hình dung dữ liệu như là những điểm trên mặt phẳng còn tri thức chính là phương trình của đường cong nối tất cả những điểm này lại. Chỉ cần một phương trình đường cong ta có thể biểu diễn được vô số điểm [1].



Hình 2. Mối quan hệ giữa thông tin, dữ liệu và tri thức

Trong hình 1, ta thấy quá trình phát hiện tri thức gồm nhiều giai đoạn. Đầu ra của giai đoạn này là đầu vào của giai đoạn sau. Trong tiến trình này, người ta đặc biệt quan tâm đến pha khai phá dữ liệu (Data mining). Khai phá dữ liệu chính là sử dụng những kỹ thuật, những phương pháp để đưa ra những thông tin có cấu trúc, những tri thức tiềm ẩn trong lượng dữ liệu. Các kỹ thuật phát hiện tri thức được thực hiện

qua nhiều giai đoạn và sử dụng nhiều phương pháp như: phân lớp, gom cụm, phân tích sự tương tự, tổng hợp, phát hiện luật kết hợp và mẫu tuần tự,... Quá trình phát hiện tri thức gồm các bước cơ bản sau:

Chọn lọc dữ liệu (selection): Đây là giai đoạn tập hợp các dữ liệu được khai thác từ một CSDL, một kho dữ liệu, thậm chí từ các nguồn ứng dụng web vào một CSDL riêng. Chúng ta chỉ chọn ra những dữ liệu cần thiết cho các giai đoạn sau. Tuy nhiên, công việc thu gom dữ liệu vào một CSDL lớn thường rất khó khăn vì dữ liệu nằm ở khắp nơi và dạng tạo lập khác nhau.

Tiền xử lý dữ liệu (preprocessing): Phần lớn các CSDL đều ít nhiều mang tính không nhất quán. Vì vậy khi gom dữ liệu rất có thể mắc một số lỗi như dữ liệu không đầy đủ, chặt chẽ và không logic (bị trùng lặp, giá trị bị sai lệch,...). Do đó cần phải được "tiền xử lý" trước khi khai phá dữ liệu nếu không sẽ gây nên những kết quả sai lệch nghiêm trọng.

Chuyển đổi dữ liệu (transformation): Trong giai đoạn này dữ liệu sẽ được chuyển đổi về dạng thuận tiện để tiến hành các thuật toán khám phá dữ liệu.

Khai phá dữ liệu (Data mining): trong giai đoạn này ta sử dụng các kỹ thuật nhằm phát hiện ra các tri thức tiềm ẩn trong dữ liệu. Một số kỹ thuật được sử dụng đó là: phân lớp, gom cụm, luật kết hợp,...

Đánh giá kết quả mẫu: Đây là giai đoạn cuối cùng trong tiến trình KDD. Trong giai đoạn này, các mẫu dữ liệu được chiết xuất bởi các phần mềm khai phá dữ liệu. Không phải bất cứ mẫu nào cũng đều có ích, thậm chí còn bị sai lệch. Chính vì vậy, cần phải xác định và lựa chọn những tiêu chuẩn đánh giá sao cho sẽ chiết xuất ra các tri thức cần thiết.

Nếu phát hiện tri thức là toàn bộ quá trình chiết xuất tri thức từ các CSDL thì KPD L là giai đoạn chủ yếu của quá trình đó. Như trên đã trình bày, trong quá trình phát hiện tri thức, khâu KPD L được thực hiện sau các khâu tinh lọc và tiền xử lý dữ liệu, tức là việc khai phá để tìm ra các mẫu có ý nghĩa được tiến hành trên tập dữ liệu có hy vọng là sẽ thích hợp với nhiệm vụ khai phá đó chứ không phải là khai phá hết dữ liệu với một thời gian đủ dài để lấy được một mẫu không thực sự có ích như khái niệm trong thống kê trước đây. Vì vậy, KPD L thường bao gồm việc thử tìm mô hình phù hợp với tập dữ liệu và tìm kiếm các mẫu từ tập dữ liệu theo mô hình đó. Chẳng hạn ta có mô hình là một luật kết hợp thì mẫu là các yếu tố

tham gia cùng với các độ hỗ trợ (support) và độ tin cậy (confidence) trong các luật tương ứng.

Mặc dù các mẫu có thể được trích lọc từ bất kỳ CSDL nào nhưng chỉ có các mẫu được xem là đáng quan tâm xét theo một phương diện nào đó mới được coi là tri thức. Các mẫu là đáng quan tâm nếu chúng là mới, có lợi,... Một mẫu được xem là mới phụ thuộc vào khung tham chiếu cho trước, có thể đó là phạm vi tri thức của hệ thống hoặc là phạm vi tri thức của người dùng. Ví dụ như việc KPD L có thể tìm ra được luật sau: "Nếu Gây_tai_nạn Thì Tuổi > 16". Đối với hệ thống, tri thức này có thể trước kia chưa biết và rất có ích nhưng đối với một người sử dụng đang thử phân tích các bản ghi về các yêu cầu bảo hiểm thì mẫu này lại không cần thiết và không đáng quan tâm vì không thể hiện được tri thức cần tìm. Ví dụ này cũng cho thấy khái niệm về tính hữu dụng. Tri thức là có ích khi nó có thể giúp đạt được mục đích của hệ thống hay của người sử dụng. Ta có thể coi KPD L giống như một quá trình phát hiện các mẫu mới đáp ứng được các yêu cầu trên, các tương quan mới có ý nghĩa, các xu hướng bằng cách khai thác trong các khối dữ liệu của kho dữ liệu, sử dụng các kỹ thuật và các khái niệm của các lĩnh vực đã được nghiên cứu từ trước như: học máy, nhận dạng, thống kê, hồi quy, xếp loại, phân cụm, các mô hình đồ thị, các mạng Bayes,...

Nếu xét về mặt ý tưởng và mục đích ứng dụng, KPD L là một nhu cầu tất yếu, một sự nhạy cảm đáp lại sự mong mỏi của giới kinh doanh thì về mặt kỹ thuật, đó thực sự là một khó khăn và là cả sự thách thức đối với những nhà khoa học. KPD L được xây dựng dựa trên việc sử dụng các giải thuật mới, được định hướng theo nhu cầu kinh doanh để có thể giải quyết tự động các bài toán kinh doanh bằng các kỹ thuật dễ dùng và có thể hiểu được. Các kỹ thuật đang được nghiên cứu và sử dụng hiện nay bao gồm cây quyết định (CART, CHAID, AID), mạng neuron, phương pháp láng giềng gần nhất, các luật suy diễn, v.v...

4. KHAI PHÁ DỮ LIỆU

Hiện nay trên sách báo, trong các cuộc hội thảo, tiếp thị sản phẩm ứng dụng công nghệ thông tin, người ta nói rất nhiều về KPD L (data mining). Vậy KPD L là gì? Và tại sao lại có nhiều người nói đến vấn đề này trong cả công nghiệp máy tính lẫn trong hoạt động kinh doanh đến như vậy?

4.1. Khai phá dữ liệu là gì?

KPDL là một khái niệm ra đời vào những năm cuối của thập kỷ 80. Nó bao hàm một loạt các kỹ thuật nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn (các kho dữ liệu). Về bản chất, KPDL liên quan đến việc phân tích các dữ liệu và sử dụng các kỹ thuật để tìm ra các mẫu hình có tính chính quy (regularities) trong tập dữ liệu.

Năm 1989, Fayyad, Piatesky-Shapiro và Smyth đã dùng khái niệm Phát hiện tri thức trong cơ sở dữ liệu (Knowledge Discovery in Database – KDD) để chỉ toàn bộ quá trình phát hiện các tri thức có ích từ các tập dữ liệu lớn. Trong đó, KPDL là một bước đặc biệt trong toàn bộ tiến trình, sử dụng các giải thuật đặc biệt để chiết xuất ra các mẫu (pattern) (hay các mô hình) từ dữ liệu.

Khai phá dữ liệu là một tiến trình sử dụng các công cụ phân tích dữ liệu khác nhau để khám phá ra các mẫu dưới nhiều góc độ khác nhau nhằm phát hiện ra các mối quan hệ giữa các dữ kiện, đối tượng bên trong CSDL, kết quả của việc khai phá là xác định các mẫu hay các mô hình đang tồn tại bên trong, nhưng chúng nằm ẩn khuất ở các CSDL. Để từ đó rút trích ra được các mẫu, các mô hình hay các thông tin và tri thức từ các CSDL [4].

4.2. Các bước của quá trình KPDL

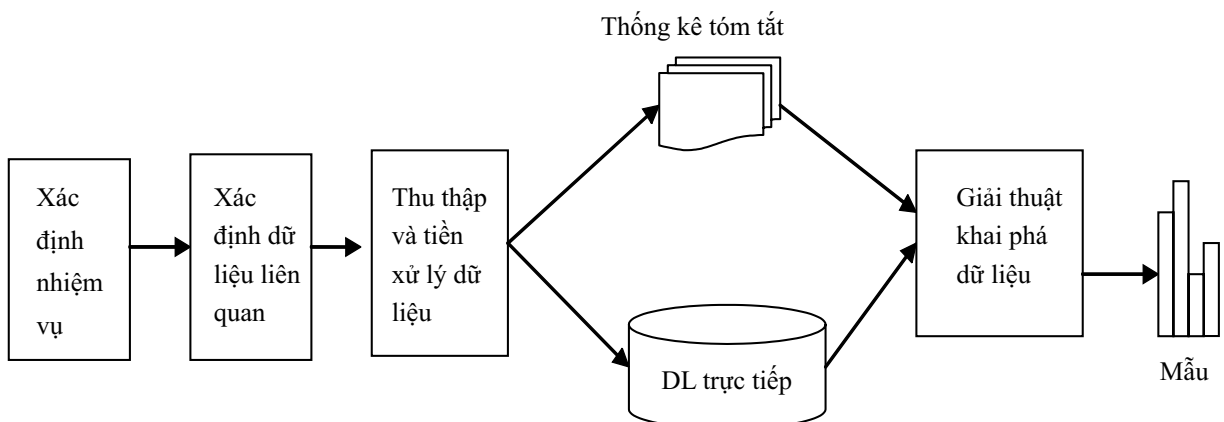
Các giải thuật KPDL thường được mô tả như những chương trình hoạt động trực tiếp trên tập dữ liệu. Với các phương pháp học máy và thống kê trước đây, thường thì bước đầu tiên là các giải thuật nạp toàn bộ tập dữ liệu vào trong bộ nhớ. Khi chuyển sang các ứng dụng công nghiệp liên quan đến việc khai phá các kho dữ liệu lớn, mô hình này không thể đáp ứng được. Không chỉ bởi vì nó không thể nạp hết

dữ liệu vào trong bộ nhớ mà còn vì khó có thể chiết xuất dữ liệu ra các tệp đơn giản để phân tích được.

Quá trình xử lý KPDL bắt đầu bằng cách xác định chính xác vấn đề cần giải quyết. Sau đó sẽ xác định các dữ liệu liên quan dùng để xây dựng giải pháp. Bước tiếp theo là thu thập các dữ liệu có liên quan và xử lý chúng thành dạng sao cho giải thuật KPDL có thể hiểu được. Về lý thuyết thì có vẻ rất đơn giản nhưng khi thực hiện thì đây thực sự là một quá trình rất khó khăn, gặp phải rất nhiều vướng mắc như: các dữ liệu phải được sao ra nhiều bản (nếu được chiết xuất vào các tệp), quản lý tập các tệp dữ liệu, phải lặp đi lặp lại nhiều lần toàn bộ quá trình (nếu mô hình dữ liệu thay đổi),...

Bước tiếp theo là chọn thuật toán KPDL thích hợp và thực hiện việc KPDL để tìm được các mẫu (pattern) có ý nghĩa dưới dạng biểu diễn tương ứng với các ý nghĩa đó (thường được biểu diễn dưới dạng các luật xếp loại, cây quyết định, luật sản xuất, biểu thức hồi quy,...).

Đặc điểm của mẫu phải là mới (ít nhất là đối với hệ thống đó). Độ mới có thể được đo tương ứng với độ thay đổi trong dữ liệu (bằng cách so sánh các giá trị hiện tại với các giá trị trước đó hoặc các giá trị mong muốn), hoặc bằng tri thức (mối liên hệ giữa phương pháp tìm mới và phương pháp cũ như thế nào). Thường thì độ mới của mẫu được đánh giá bằng một hàm logic hoặc một hàm đo độ mới, độ bất ngờ của mẫu. Ngoài ra, mẫu còn phải có khả năng sử dụng tiềm tàng. Các mẫu này sau khi được xử lý và diễn giải phải dẫn đến những hành động có ích nào đó được đánh giá bằng một hàm lợi ích. Mẫu khai thác được phải có giá trị đối với các dữ liệu mới với độ chính xác nào đó.



Hình 3. Quá trình KPDL

Kỹ thuật KPDL thực chất là phương pháp không hoàn toàn mới. Nó là sự kế thừa, kết hợp và mở rộng của các kỹ thuật cơ bản đã được nghiên cứu từ trước như máy học, nhận dạng, thống kê (hồi quy, xếp loại, phân cụm), các mô hình đồ thị, các mạng Bayes, trí tuệ nhân tạo, thu thập tri thức hệ chuyên gia, v.v... Tuy nhiên, với sự kết hợp tài tình của KPDL, kỹ thuật này có ưu thế hơn hẳn các phương pháp trước đó, đem lại nhiều triển vọng trong việc ứng dụng phát triển nghiên cứu khoa học.

4.3. Khái quát về các kỹ thuật KPDL

4.3.1. Khai thác tập phổ biến và luật kết hợp

Đây là tiến trình khám phá các tập giá trị thuộc tính xuất hiện phổ biến trong các đối tượng dữ liệu. Từ tập phổ biến có thể tạo ra các luật kết hợp giữa các giá trị thuộc tính nhằm phản ánh khả năng xuất hiện đồng thời các giá trị thuộc tính trong tập các đối tượng. Một luật kết hợp $X \rightarrow Y$ phản ánh sự xuất hiện của tập X dẫn đến sự xuất hiện đồng thời của tập Y . Chẳng hạn phân tích CSDL bán hàng nhận được thông tin về những khách hàng mua máy tính có khuynh hướng mua phần mềm quản lý nhân sự trong cùng lần mua được miêu tả bằng luật kết hợp như sau:

"Máy tính P Phần mềm quản lý nhân sự"

Luật kết hợp giúp các nhà hoạch định hiểu rõ xu thế bán hàng, tâm lý khách hàng, ... từ đó đưa ra chiến lược bố trí mặt hàng, kinh doanh, tiếp thị,...

4.3.2. Phân lớp dữ liệu

Phân lớp (Classification): Là tiến trình khám phá các luật phân loại hay đặc trưng cho các tập dữ liệu đã được xếp lớp. Tập dữ liệu học bao gồm tập đối tượng đã được xác định lớp sẽ được dùng để tạo mô hình phân lớp dựa trên đặc trưng của đối tượng trong tập dữ liệu học. Các luật phân lớp được sử dụng để xây dựng các bộ phân lớp dữ liệu. Phân lớp dữ liệu có vai trò quan trọng trong tiến trình dự báo các khuynh hướng quy luật phát triển. Áp dụng vào tiến trình phân lớp dữ liệu khách hàng trong CSDL có thể xây dựng các luật phân lớp khách hàng [1]. Một số kỹ thuật thường được sử dụng trong phân lớp:

- Cây quyết định (Decision tree): Cấu trúc dạng hình cây là biểu thị cho các quyết định. Các quyết định này sinh ra các quy tắc để phân lớp và dự đoán (dự báo) tập dữ liệu mới chưa được phân lớp. Tri thức được rút ra trong kỹ thuật này thường được mô tả dưới dạng tường minh, đơn giản, trực quan, dễ hiểu đối với người sử dụng. Tuy vậy, nó cũng đòi hỏi một

không gian nhất định để mô tả tri thức trong phạm vi mà con người có thể hiểu được [3].

- Mạng Nơron (Neural Network): Đây là một trong những kỹ thuật được ứng dụng rất phổ biến hiện nay vì kỹ thuật này bắt chước khả năng tìm kiếm mẫu của bộ não con người. Việc huấn luyện theo phương pháp này được bắt đầu bằng việc cho vào một tập dữ liệu (gọi là tập dữ liệu huấn luyện) mạng sẽ tự động điều chỉnh (học) qua từng lớp trong mạng và cho ra kết quả, quá trình huấn luyện được lặp đi lặp lại nhiều lần. Sau khi mạng học thành công thì nó được xem là một chuyên gia trong lĩnh vực đó [3].

4.3.3. Gom cụm

Gom cụm (clustering): Là tiến trình nhận diện các cụm tiềm ẩn trong tập các đối tượng chưa được xếp lớp. Tiến trình phân cụm dựa trên mức độ tương tự giữa các đối tượng. Các đối tượng được gom cụm sao cho mức độ tương tự giữa các đối tượng trong cùng một cụm là cực đại và mức độ tương tự giữa các đối tượng nằm trong các cụm khác nhau là cực tiểu. Các cụm được đặc trưng bằng các tính chất chung của tất cả các đối tượng trong cụm. Do vậy, khảo sát các cụm sẽ giúp khái quát, tổng kết nhanh chóng nội dung của khối dữ liệu lớn [2]. Một số kỹ thuật dùng trong gom cụm: phương pháp phân cấp và thuật giải k-means.

- Trong phương pháp phân cấp gồm hai giai đoạn: ở giai đoạn gộp mỗi đối tượng thuộc về một lớp, dưới một ngưỡng nào đó chúng được gom lại và tiếp tục như vậy cho đến khi tất cả các đối tượng thuộc về cùng một cụm; còn ở giai đoạn tách sẽ tiến hành ngược lại so với giai đoạn gộp. Trong phương pháp này, để đưa ra được ngưỡng người ta dựa vào liên kết đơn (khoảng cách gần nhất) và liên kết đôi (khoảng cách xa nhất) giữa hai cụm [3].

- Trong thuật giải k-means: khi tiến hành gom cụm ta cần phải biết trước số cụm. Với mỗi cụm như vậy ta chọn phần tử trọng tâm, tiếp đến tính khoảng cách từ mỗi đối tượng đến phần tử trọng tâm ở tất cả các cụm và gán nó vào cụm có khoảng cách nhỏ nhất. Quá trình này tiếp tục như vậy cho đến khi các cụm không còn biến đổi hoặc thoả điều kiện đặt ra.

5. ỨNG DỤNG CỦA KPDL

Mặc dù còn rất nhiều vấn đề mà KPDL cần phải tiếp tục nghiên cứu để giải quyết nhưng tiềm năng của nó đã được khẳng định bằng sự ra đời của rất nhiều ứng dụng. Các ứng dụng của KPDL trong khoa

học cũng được phát triển. Các công ty phần mềm lớn trên thế giới cũng rất quan tâm và chú trọng tới việc nghiên cứu và phát triển kỹ thuật khai phá dữ liệu: Oracle tích hợp các công cụ khai phá dữ liệu vào bộ Oracle9i, IBM đã đi tiên phong trong việc phát triển các ứng dụng khai phá dữ liệu với các ứng dụng như Intelligence Miner, ...Ta có thể đưa ra một số ứng dụng trong các lĩnh vực như:

Ngân hàng: Xây dựng mô hình dự báo rủi ro tín dụng; tìm kiếm tri thức, quy luật của thị trường chứng khoán và đầu tư bất động sản,...

Thương mại điện tử: Công cụ tìm hiểu, định hướng, thúc đẩy, giao tiếp với khách hàng; phân tích khách hàng duyệt web; Phân tích hành vi mua sắm trên mạng và cho biết thông tin tiếp thị phù hợp với loại khách hàng.

Thiên văn học: Hệ thống SKICAT do JPL/Caltech phát triển được sử dụng cho các nhà thiên văn để tự động xác định các vì sao và các dải thiên hà trong một bản khảo sát lớn để có thể phân tích và phân loại (Fayyad, Djorgovski, & Weir).

Sinh học phân tử: Hệ thống tìm kiếm các mẫu trong cấu trúc phân tử (Conklin, Fortier, và Glasgow 1993) và trong các dữ liệu gen (Holder, Cook, và Djoko 1994).

Mô hình hóa những thay đổi thời tiết: các mẫu không thời gian như lốc, gió xoáy được tự động tìm thấy trong các tập lớn dữ liệu mô phỏng và quan sát được (Stolorz et al. 1994).

6. NHỮNG THÁCH THỨC TRONG ỨNG DỤNG VÀ NGHIÊN CỨU KỸ THUẬT KPD L

Khối lượng dữ liệu lớn và từ nhiều nguồn khác nhau: CSDL, internet, các loại thiết bị thu nhận tín hiệu, các loại thiết bị nhận dạng, các loại thiết bị lưu trữ như băng từ, CD, ... Số mẫu tin và số các thuộc tính quá lớn làm cho độ phức tạp và thời gian giải quyết bài toán tăng lên rất nhanh.

Mô hình hay tri thức phát hiện được bị thay đổi theo thời gian tức là mô hình hay tri thức đó phụ thuộc vào thời điểm quan sát, lấy mẫu, thời điểm khai phá, kết quả đạt được sau khai phá cũng gây không ít khó khăn cho khai phá dữ liệu.

Dữ liệu bị ảnh hưởng, bị nhiễu bởi tác động của môi trường bên ngoài, hay bộ dữ liệu không hoàn chỉnh làm cho dữ liệu không phản ánh trung thực, chính xác của các quy luật, tri thức mà ta tìm được.

Các thuộc tính không phù hợp, các bộ giá trị

không đầy đủ, bị thiếu giá trị trong các miền thuộc tính sẽ ảnh hưởng rất lớn đến việc khai phá dữ liệu. Chẳng hạn như khi khai phá dữ liệu, các hệ thống tương tác với nhau, phụ thuộc nhau, chỉ cần thiếu một vài giá trị nào đó sẽ dẫn đến các mẫu thuẫn, không chính xác, không đầy đủ.

Từ những vấn đề đặt ra ở trên nên tốc độ xử lý cần quan tâm trước nhất. Có hai phương hướng để giải quyết vấn đề này là nâng cao năng lực của phần cứng và cải tiến phần mềm. Tuy nhiên khi cải thiện năng lực của máy tính thì dữ liệu cung tăng không ngừng, thậm chí còn tăng nhanh hơn gấp nhiều lần. Do vậy việc nghiên cứu để xuất các thuật toán hiệu quả có khả năng làm việc trên khối lượng dữ liệu lớn, và có độ phức tạp tính toán thấp là một hướng nghiên cứu đầy tiềm năng. Từ nhu cầu thực tế trên, gần đây đã xuất hiện nhiều ngành khoa học công nghệ hỗ trợ KPD L như tính toán song song, máy tính lượng tử, công nghệ nano, phát triển thuật toán, ...

7. KẾT LUẬN

Chúng ta vừa tìm hiểu những nét khái quát nhất về tiến trình phát hiện tri thức trong cơ sở dữ liệu. Tiến trình này bao gồm một dãy các thao tác kế tiếp nhau, đầu ra của giai đoạn trước là đầu vào của giai đoạn sau. Trong đó KPD L chỉ là một giai đoạn trong KDD và người ta quan tâm nhiều đến giai đoạn này. Vì các kỹ thuật khác nhau nên yêu cầu số lượng và các dạng thông tin rất khác nhau do đó chúng thường ảnh hưởng đến việc thiết kế và chọn giải thuật KPD L khác nhau.

TÀI LIỆU THAM KHẢO

- [1] Hoàng Kiếm; Giáo trình nhập môn trí tuệ nhân tạo; NXB ĐHQG TPHCM 2000.
- [2] Đỗ Phúc; Giáo trình khai thác dữ liệu; NXB ĐHQG TPHCM 2008.
- [3] Mehmed Kantardzic; Data mining concepts, models, methods, and algorithms; John Wiley & Sons, 2003.
- [4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth; From data mining to knowledge discovery in databases;