**ĐẠI HỌC ĐÀ NẴNG**

**TRƯỜNG CĐ CÔNG NGHỆ THÔNG TIN**

# BÁO CÁO TỔNG KẾT
# ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ
# CẤP CƠ SỞ

# NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DẠNG LƯỚI TRONG LĨNH VỰC TÀI CHÍNH

**Mã số: T2016-07-07**

**Chủ nhiệm đề tài: Th.s Trần Thu Thủy**

**Đà Nẵng, 12/2016**

ĐẠI HỌC ĐÀ NẴNG

**TRƯỜNG CĐ CÔNG NGHỆ THÔNG TIN**

# BÁO CÁO TỔNG KẾT
# ĐỀ TÀI KHOA HỌC VÀ CÔNG NGHỆ
# CẤP CƠ SỞ

## NGHIÊN CỨU ỨNG DỤNG KỸ THUẬT KHAI PHÁ DỮ LIỆU DẠNG LƯỚI TRONG LĨNH VỰC TÀI CHÍNH

**Mã số: T2016-07-07**

**Chủ nhiệm đề tài:** **Th.s Trần Thu Thủy**

**Xác nhận của cơ quan chủ trì đề tài** **Chủ nhiệm đề tài**

**Đà Nẵng,** ***12/2016***

# MỤC LỤC

# DANH MỤC TỪ VIẾT TẮT

| STT | Cụm từ | Viết tắc |
|---|---|---|
| 1. | Management Information System | MIS |
| 2. | Phân cụm dữ liệu | PCDL |

# DANH MỤC HÌNH ẢNH

ĐẠI HỌC ĐÀ NẴNG
**TRƯỜNG CĐ CÔNG NGHỆ THÔNG TIN**

**CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM**
**Độc lập – Tự do – Hạnh phúc**

# THÔNG TIN KẾT QUẢ NGHIÊN CỨU

## 1. Thông tin chung:

- Tên đề tài: ***"Nghiên cứu ứng dụng kĩ thuật khai phá dữ liệu dạng lưới trong lĩnh vực tài chính"***

- Mã số: T2016- 07-07

- Chủ nhiệm: Trần Thu Thủy

- Thành viên tham gia:

- Cơ quan chủ trì: Trường Cao đẳng Công nghệ Thông tin

- Thời gian thực hiện: **Từ 4/2016→ 12/2016**

## 2. Mục tiêu:

Mục tiêu chính của đề tài nhằm tìm hiểu các kĩ thuật khai phá dữ liệu, cụ thể là kĩ thuật khai phá dữ liệu dạng lưới, và từ đó nghiên cứu những ứng dụng của kĩ thuật này trong lĩnh vực tài chính.

## 3. Tính mới và sáng tạo:

Đề tài này nghiên cứu những điểm mạnh và những tiềm năng của kĩ thuật khai phá dữ liệu dạng lưới vào lĩnh vực tài chính, một lĩnh vực đang rất cần có sự can thiệp của khoa học công nghệ khai phá dữ liệu để quản lý tốt hơn những dữ liệu của mình đồng thời tạo cơ sở nền tảng để phát triển kinh doanh tốt hơn.

## 4. Tóm tắt kết quả nghiên cứu:

Kỹ thuật khai phá dữ liệu giúp ngân hàng phân tích và nhận định được đâu là các khách hàng trung thành và đâu là các khách hàng có xu hương chuyển sang ngân hàng khác với mong muốn dịch vụ tốt hơn. Nếu khách hàng chuyển từ ngân hàng của mình sang ngân hàng khác, lý do cho việc chuyển như vậy và giao dịch cuối cùng được thực hiện trước khi chuyển có thể được biết đó sẽ giúp các ngân hàng hoạt động tốt hơn và giữ chân khách hàng của mình.

## 5. Hiệu quả, phương thức chuyển giao kết quả nghiên cứu và khả năng áp dụng:

Đưa ra những tổng kết về kết quả kinh doanh, con số phát triển của các tổ chức có sử dụng kĩ thuật khai phá dữ liệu dạng lưới này vào trong quá trình kinh doanh của mình. Ứng dụng những công nghệ hiệu quả vào trong quá trình phát triển hệ thống

*Đà Nẵng, ngày 15 tháng 12 năm 2016*

**Cơ quan chủ trì** **Chủ nhiệm đề tài**

# MỞ ĐẦU

## 1. TÍNH CẤP THIẾT

Trong ngành công nghiệp dịch vụ tài chính trên toàn thế giới, phương thức liên lạc truyền thống của khách hàng mặt đối mặt (face-to-face) đang được thay thế bằng phương thức điện tử để giảm thời gian và chi phí xử lý các áp dụng cho sản phẩm khác nhau và cuối cùng là cải thiện hiệu quả của việc sử dụng tài chính. Tin học hoá quá trình hoạt động tài chính, sử dụng internet và phần mềm tự động hoàn toàn có thể làm thay đổi các khái niệm cơ bản của kinh doanh và cách hoạt động kinh doanh đang được thực hiện. Hiển nhiên, lĩnh vực ngân hàng không phải là một ngoại lệ. Kể từ những năm 1990 toàn bộ khái niệm ngân hàng đã được chuyển sang cơ sở dữ liệu tập trung, giao dịch trực tuyến và máy ATM được thực hiện trên thế giới, đã làm cho hệ thống ngân hàng mặt mạnh mẽ hơn về mặt kỹ thuật và định hướng khách hàng tốt hơn. Dữ liệu có thể là một trong những nguồn tài nguyên có giá trị nhất của bất kỳ ngân hàng nào, tuy nhiên nó chỉ thực sự có giá trị khi nó biết cách tiếp cận với thông tin có giá trị ẩn chứa trong dữ liệu thô. Khai phá dữ liệu cho phép triết suất các thông tin từ các dữ liệu lịch sử, và dự đoán kết quả các tình huống trong tương lai. Nó giúp cho việc tối ưu hóa các quyết định kinh doanh, tăng giá trị của từng khách hàng và thông tin kết nối, đồng thời cải thiện sự hài lòng của khách hàng.

Số lượng dữ liệu được thu thập bởi các ngân hàng đã tăng nhanh chóng trong những năm gần đây. Với những kỹ thuật phân tích số liệu thống kê hiện khó có thể quản lý tốt với khối lượng lớn dữ liệu hiện có như hiện tại. Sự tăng trưởng bùng nổ này đã dẫn đến sự cần thiết của kỹ thuật phân tích dữ liệu mới và các công cụ mới để tìm ra các thông tin thực sự có ích ẩn chứa trong dữ liệu này. Ngân hàng là lĩnh vực mà tại đây một lượng lớn dữ liệu được thu thập. Dữ liệu này có thể được tạo ra từ các giao dịch của các tài khoản ngân hàng, hồ sơ vay vốn, trả nợ, thẻ tín dụng, v.v... Người ta cho rằng thông tin có giá trị về các hồ sơ tài chính của khách hàng được ẩn chứa trong các cơ sở dữ liệu hoạt động lớn và các thông tin này có thể được sử dụng để cải thiện hiệu suất kinh doanh của các ngân hàng. Tại thời điểm ban đầu tại các trung tâm tin

học đầu mối của các ngân hàng, nhiều gói phần mềm đang được sử dụng cho các giao dịch hàng ngày. Từ đó, nếu như thiết kế mới một Hệ thống thông tin (MIS: Management Information System) mới hoặc cơ cấu lại những cơ sở hạ tầng hiện sẽ khó thể thực hiện được bởi không chỉ đơn giản là cần thay thế các gói phầm mềm tại các trung tâm tin học đó. Giải pháp cho vấn đề này là để thực hiện các khái niệm về kho dữ liệu và khai phá dữ liệu (Data Warehouse and Data Mining).

## 2. MỤC TIÊU NGHIÊN CỨU

- Nghiên cứu tổng quan kiến trúc lưới.
- Nghiên cứu các kĩ thuật khai phá kiến trúc lưới.
- Ứng dụng kĩ thuật khai phá kiến trúc lưới trong lĩnh vực tài chính.

## 3. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

3.1. Đối tượng nghiên cứu

- Kiến trúc lưới.
- Các kĩ thuật khai phá kiến trúc lưới.

3.2. Phạm vi nghiên cứu

Tập trung nghiên cứu khai phá dữ liệu trong mô hình kiến trúc lưới ứng dụng trong lĩnh vực tài chính.

## 4. BỐ CỤC ĐỀ TÀI

Ngoài lời mở đầu và kết luận, đề tài gồm 3 chương:

**Chương 1:** Tổng quan cơ sở lý thuyết của nghiên cứu.

Giới thiệu tổng quan về khai phá dữ liệu, trích chọn thông tin, về kho ngữ liệu, về các công trình nghiên cứu cùng lĩnh vực này đã được công bố.

**Chương 2:** Khai phá dữ liệu phân cụm dựa trên mô hình lưới

Chương này giới thiệu các phương pháp tiếp cận cùng với những ưu và nhược điểm của chúng, từ đó đưa ra giải pháp cho bài toán đang nghiên cứu.

**Chương 3:** Ứng dụng mô hình phân cụm dựa trên lưới vào lĩnh vực tài chính

Chương này giới thiệu về những ứng dụng hiện nay của kĩ thuật PCDL dựa trên lưới trong lĩnh vực tài chính, cụ thể là lĩnh vực tài chính ngân hàng.

# CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

## 1.1 KHAI PHÁ DỮ LIỆU.

Khai phá dữ liệu đề cập đến tri thức chiết xuất từ một lượng lớn dữ liệu. Dữ liệu có thể được dữ liệu không gian, dữ liệu đa phương tiện, dữ liệu chuỗi thời gian, dữ liệu văn bản và dữ liệu web. Khai phá dữ liệu là quá trình khai thác các thông tin hữu ích, thú vị, đặc biệt, tiềm ẩn, chưa được biết và có khả năng hữu ích và tri thức từ một lượng lớn dữ liệu. Nó là tập hợp các hoạt động được sử dụng để tìm kiếm, các thông tin tiềm ẩn hoặc là các không tin không mong đợi trong dữ liệu hoặc hình thức thể hiện khác thường trong dữ liệu. Sử dụng thông tin trong kho dữ liệu, khai phá dữ liệu thường có thể cung cấp các câu trả lời cho các câu hỏi về một tổ chức có một quyết định trước đây không thông qua việc hỏi và khảo sát:

Những sản phẩm nào nên được cất nhắc cho khách hàng đặc biệt? - Mục tiêu quảng bá và bán sản phẩm.

Xác suất mà một khách hàng nhất định sẽ để lại cho một đối thủ cạnh tranh là gì? - Quản lý quan hệ khách hàng

Chẩn đoán thích hợp cho bệnh nhân này này là gì? - Sinh học y tế;

- Khả năng một khách hàng nào đó mặc định hoặc sẽ trả lại một khoản vay là gì? - Ngân hàng.

- Những sản phẩm nào được mua nhiều nhất cùng với nhau? - Phân tích thị trường Giỏ hàng.

Làm thế nào để xác định người gian lận trong ngành công nghiệp viễn thông? - Mô hình phân tích gian lận

Các loại câu hỏi này có thể được trả lời một cách nhanh chóng và dễ dàng nếu các thông tin ẩn trong những lượng lớn dữ liệu trong cơ sở dữ liệu có thể được xác định và sử dụng.

Khai thác dữ liệu thường được coi như là "thông minh phân tích". Một số xu hướng gần đây đã gia tăng sự quan tâm trong lĩnh vự khai phá dữ liệu, chủ yếu là việc giảm chi phí lưu trữ dữ liệu và sự dễ dàng ngày càng tăng của việc thu thập dữ liệu. Với khả năng lưu trữ dữ liệu lớn hơn và chi phí giảm, khai phá dữ liệu đã cung cấp cho các tổ chức một phương thức mới để trong quá trình kinh doanh. Khai phá dữ liệu có thể giúp cho các tổ chức hiểu rõ hơn về tình hình kinh doanh của họ, từ đó họ có thể phục vụ tốt hơn khách hàng của họ, và tăng hiệu quả của tổ chức trong thời gian dài.

Ngày nay, các ngân hàng đã nhận ra những lợi thế khác nhau của việc khai phá dữ liệu. Nó là một công cụ có giá trị mà ngành ngân hàng có thể xác định thông tin hữu ích từ số lượng lớn dữ liệu mà họ thu thập được. Điều này có thể giúp các ngân hàng để đạt được một lợi thế trội hơn so với đối thủ cạnh tranh của họ. Hơn nữa, khai phá dữ liệu có thể giúp các ngân hàng trong việc hiểu rõ hơn về các khối lượng lớn các dữ liệu thu thập bởi các các hệ thống CRM (Customer Relationship Management).

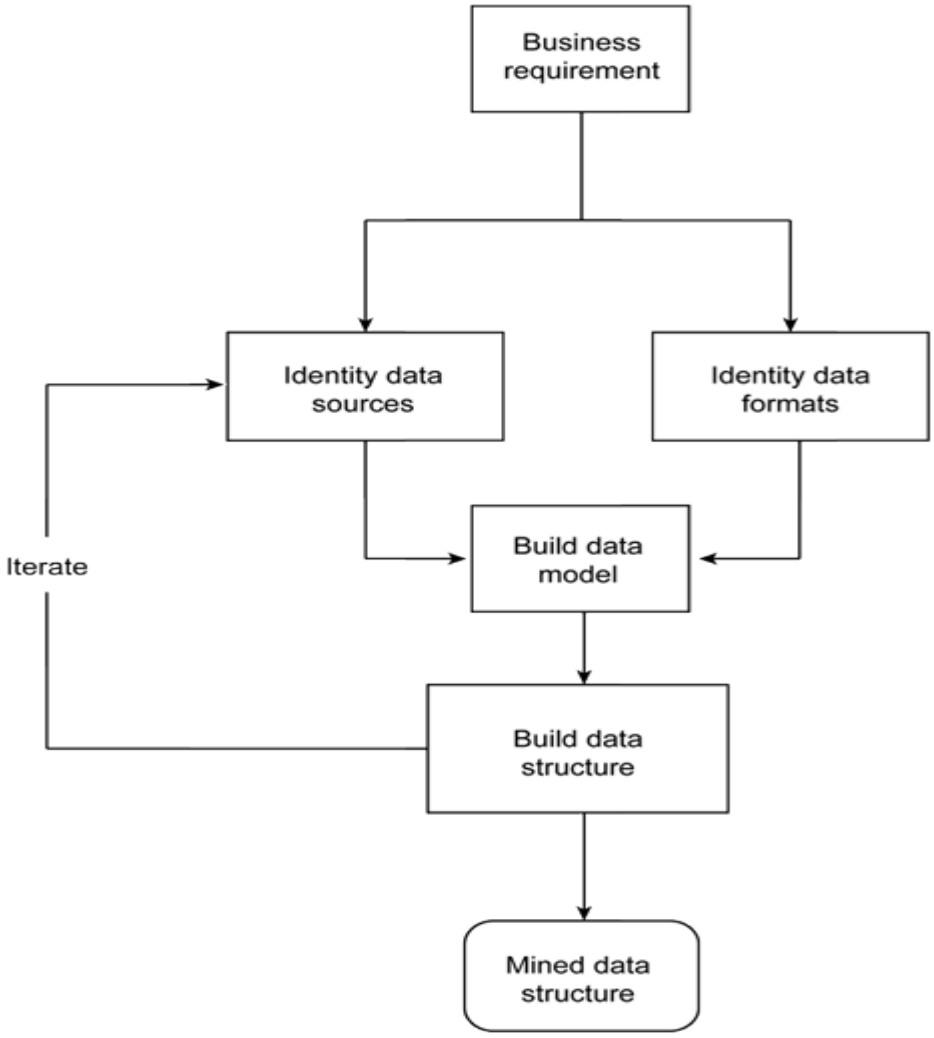
## 1.2 CÁC THUẬT TOÁN VÀ KỸ THUẬT KHAI PHÁ DỮ LIỆU.

Hiện nay có nhiều kỹ thuật khai phá dữ liệu, mô hình truy vấn, mô hình xử lý và thu thập dữ liệu khác nhau. Vậy bạn sẽ sử dụng một kỹ thuật nào để khai phá dữ liệu của mình và bạn có thể sử dụng kỹ thuật nào để kết hợp với phần mềm và cơ sở hạ tầng hiện có của mình? Hãy xem xét các kỹ thuật và các giải pháp khai phá dữ liệu và phân tích khác nhau và tìm hiểu cách xây dựng chúng nhờ sử dụng phần mềm và các bản cài đặt hiện có. Hãy khám phá các công cụ khai phá dữ liệu khác nhau có sẵn và tìm hiểu cách xác định xem kích thước và độ phức tạp của những thông tin của bạn có thể dẫn đến những khó khăn rắc rối về xử lý và lưu trữ không và cần phải làm gì.

Khai phá dữ liệu là một quá trình

Về cơ bản, khai phá dữ liệu là về xử lý dữ liệu và nhận biết các mẫu và các xu hướng trong thông tin đó để bạn có thể quyết định hoặc đánh giá. Các nguyên tắc khai phá dữ liệu đã được dùng nhiều năm rồi, nhưng với sự ra đời của big data (dữ liệu lớn), nó lại càng phổ biến hơn.

Big data gây ra một sự bùng nổ về sử dụng nhiều kỹ thuật khai phá dữ liệu hơn, một phần vì kích thước thông tin lớn hơn rất nhiều và vì thông tin có xu hướng đa dạng và mở rộng hơn về chính bản chất và nội dung của nó. Với các tập hợp dữ liệu lớn, để nhận được số liệu thống kê tương đối đơn giản và dễ dàng trong hệ thống vẫn chưa đủ. Với 30 hoặc 40 triệu bản ghi thông tin khách hàng chi tiết, việc biết rằng 2 triệu khách hàng trong số đó sống tại một địa điểm vẫn chưa đủ. Bạn muốn biết liệu 2 triệu khách hàng đó có thuộc về một nhóm tuổi cụ thể không và bạn cũng muốn biết thu nhập trung bình của họ để bạn có thể tập trung vào các nhu cầu của khách hàng của mình tốt hơn.




***Hình 1.1: Phác thảo quá trình***

Những nhu cầu hướng kinh doanh này đã thay đổi cách lấy ra và thống kê dữ liệu đơn giản sang việc khai phá dữ liệu phức tạp hơn. Vấn đề kinh doanh hướng tới việc xem xét dữ liệu để giúp xây dựng một mô hình để mô tả các thông tin mà cuối cuộc sẽ dẫn đến việc tạo ra báo cáo kết quả. Hình 1 phác thảo quá trình này.

Quá trình phân tích dữ liệu, khám phá dữ liệu và xây dựng mô hình dữ liệu thường lặp lại khi bạn tập trung vào và nhận ra các thông tin khác nhau để bạn có thể trích ra. Bạn cũng phải hiểu cách thiết lập quan hệ, ánh xạ, kết hợp và phân cụm thông tin đó với dữ liệu khác để tạo ra kết quả. Quá trình nhận ra dữ liệu nguồn và các định dạng nguồn, rồi ánh xạ thông tin đó tới kết quả đã cho của chúng tôi có thể thay đổi sau khi bạn phát hiện ra các yếu tố và các khía cạnh khác nhau của dữ liệu.

### 1.2.1. Phân loại.

Phân loại là phương pháp khai phá dữ liệu được áp dụng phổ biến nhất hiện nay. Trong đó sử dụng một tập hợp các ví dụ chưa được phân loại để phát triển một mô hình mà có thể phân loại được. Về cơ bản phân loại được sử dụng để phân loại từng hạng mục trong một tập hợp các dữ liệu vào một trong những tập được xác định trước các lớp hoặc nhóm. Phương pháp phân loại sử dụng các kỹ thuật toán học như cây quyết định, quy hoạch tuyến tính, mạng Neutral và thống kê. Trong việc phân loại, chúng ta tạo ra cho phần mềm có thể hiểu được cách phân loại các thành phần dữ liệu thành các nhóm.

Phát hiện gian lận và rủi ro tín dụng đặc biệt thích hợp với loại hình phân tích này. Phương pháp này thường được sử dụng các thuật toán phân cây quyết định hoặc mạng Neutral. Dữ liệu được phân tích bởi thuật toán phân loại, và được thử nghiệm được sử dụng để ước tính độ chính xác của các quy tắc phân loại. Nếu độ chính xác có thể chấp nhận bởi các quy tắc thì có thể được áp dụng cho các mẫu dữ liệu mới. Đối với một ứng dụng phát hiện gian lận, dữ liệu đầu vào gồm toàn bộ hai tập các bản ghi giả và bản ghi thật các hoạt động. Các thuật toán phân loại sử dụng các dữ liệu chưa được phân loại đó để xác định tập hợp các thông số cần thiết cho những điều chỉnh

thích hợp. Sau đó các thuật toán mã hóa các thông số và chuyển chúng thành một mô hình và được gọi là sự phân loại. Có các loại mô hình phân loại cơ bản sau:

- Phân loại theo cây quyết định.
- Phân loại Bayesian.
- Mạng Neural.

- Phân loại dựa trên sự kết hợp.

### 1.2.2. Luật kết hợp

Luật kết hợp là một trong những kỹ thuật khai thác dữ liệu nổi tiếng nhất. Trong luật kết hợp, một mô hình được phát hiện dựa trên mối quan hệ của một mặt hàng cụ thể đối với các mặt hàng khác trong cùng một giao dịch. Sự kết hợp và tương quan thường được áp dụng trên các bộ dữ liệu lớn. Việc phân loại và tìm kiếm này giúp các doanh nghiệp đưa ra quyết định nào đó, chẳng hạn như thiết kế danh sách danh mục hàng, phân tích hành vi mua sắm của khách hàng. Ví dụ, các kỹ thuật kết hợp được sử dụng trên thị trường phân tích giỏ hàng để xác định những sản phẩm mà khách hàng thường xuyên mua cùng với nhau. Dựa trên dữ liệu này doanh nghiệp có thể có chiến dịch kinh doanh tương ứng để bán sản phẩm với mục đích làm tăng loại nhuận. Các loại luật kết hợp khác nhau bao gồm:

- Sử dụng nhiều luật Kết hợp đồng thời.
- Luật kết hợp đa chiều.
- Luật kết hợp đánh giá.
- Luật kết hợp trực tiếp.
- Luật kết hợp gián tiếp.

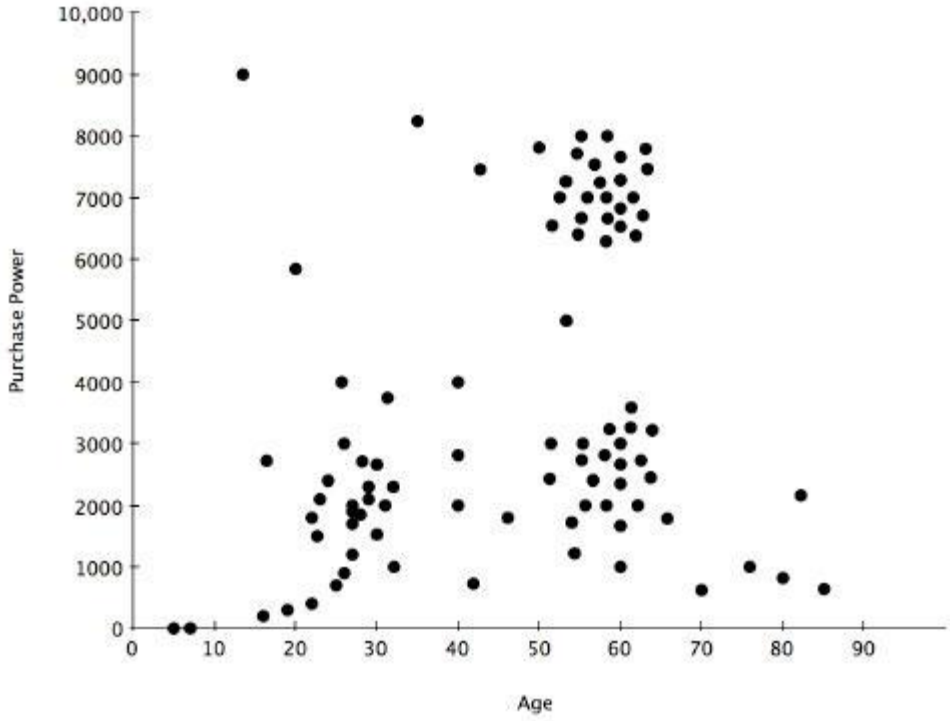
### 1.2.3 Việc phân cụm (Clustering)

Sự phân nhóm là một kỹ thuật khai thác dữ liệu mà làm cho các nhóm có ý nghĩa và hữu ích của các đối tượng có đặc trưng tương tự nhau khi sử dụng các kỹ thuật tự động. Kỹ thuật phân nhóm cũng xác định các lớp và đặt các đối tượng vào

trong trong đó, trong khi đối với kỹ thuật phân loại thì các đối tượng đươc gắn vào các lớp mà chưa được định chưa được định. Phương pháp phân loại cũng có thể được sử dụng cho các phương tiện hiệu quả của các nhóm phân biệt hoặc các lớp đối tượng. Tuy nhiên, nó trở nên khá tốn kém nên có thể phân nhóm thường được sử dụng như phương pháp tiền xử lý trong quá trình khai phá dữ liệu.

Ví dụ: Khách hàng, ở các điểm địa lý khác nhau, với mục đích khác nhau, và với các đặc điểm về công việc khác nhau, họ sẽ có những yêu cầu khác nhau đối với dịch vụ ngân hàng. Tuy nhiên họ vẫn phải yêu cầu và được đảm bảo về sự an toàn ví dụ như là họ không thể chịu chấp nhận rủi ro.Với cùng một bộ dịch vụ áp dụng cho các đối tượng, chúng ta có thể thay đổi một số các chính sách, các ưu đãi để có thể áp dụng cho các đối tượng khách hàng ở khu vực đô thi. Những thông tin này sẽ giúp cho việc tổ chức trong hoạt động bán chéo các sản phẩm của họ. Các đơn vị dịch vụ khách hàng đại diện cho các ngân ngân hàng có thể được trang bị với hồ sơ khách hàng được làm phong phú hơn bằng cách khai phá dữ liệu để giúp họ xác định được sản phẩm và dịch vụ phù hợp nhất với người yêu cầu. Kỹ thuật này sẽ giúp việc quản lý trong việc tìm kiếm các giải pháp của 80/20 cơ bản của việc tiếp thị. Trong đó nói rằng: Hai mươi phần trăm của khách hàng của bạn sẽ cung cấp cho bạn 80 phần trăm lợi nhuận của bạn, tuy nhiên, vấn đề là xác định 20% đó và các kỹ thuật phân nhóm như thế nào.

Trong ví dụ được mô phỏng ở hình dưới đây, chúng ta có thể nhận ra hai cụm, một cụm xung quanh nhóm 2.000 Đô la Mỹ/ 20-30 tuổi và một cụm ở nhóm 7.000-8.000 Đô la Mỹ/ 50-65 tuổi. Trong trường hợp này, chúng tôi đã giả thuyết hai cụm và đã chứng minh giả thuyết của chúng tôi bằng một đồ thị đơn giản mà chúng tôi có thể tạo ra bằng cách sử dụng bất kỳ phần mềm đồ họa thích hợp nào để có được cái nhìn nhanh chóng. Các quyết định phức tạp hơn cần phải có một gói phần mềm phân tích đầy đủ, đặc biệt là nếu bạn muốn các quyết định tự động dựa vào thông tin lân cận gần nhất.




***Hình 1.2: Phác thảo việc phân cụm***

Việc vẽ đồ thị phân cụm theo cách này là một ví dụ đơn giản về cái gọi là nhận ra sự lân cận gần nhất. Bạn có thể nhận ra các khách hàng riêng lẻ bằng sự gần gũi theo nghĩa đen của họ với nhau trên đồ thị. Có nhiều khả năng là các khách hàng trong cùng một cụm cũng dùng chung các thuộc tính khác và bạn có thể sử dụng sự mong đợi đó để giúp hướng dẫn, phân loại và nếu không thì phân tích những người khác trong tập hợp dữ liệu của bạn.

Bạn cũng có thể áp dụng việc phân cụm theo quan điểm ngược lại; dựa vào một số thuộc tính đầu vào, bạn có thể nhận ra các tạo phẩm khác nhau. Ví dụ, một nghiên cứu gần đây về các số PIN 4-chữ số đã tìm ra các cụm giữa các chữ số trong phạm vi 1-12 và 1-31 cho các cặp đầu tiên và thứ hai. Bằng cách vẽ các cặp này, bạn có thể nhận ra và xác định các cụm liên quan đến ngày tháng (các ngày sinh nhật, các ngày kỷ niệm).

Kỹ thuật phân cụm gồm có:

- Phương pháp phân vùng
- phương pháp phân cấp
- Phương pháp dựa trên mật độ
- Phương pháp dựa trên lưới
- Phương pháp dựa trên mô hình

### 1.2.4 Dự đoán

Dự báo là một chủ đề rộng và đi từ dự báo về lỗi của các thành phần hay máy móc đến việc nhận ra sự gian lận và thậm chí là cả dự báo về lợi nhuận của công ty nữa. Được sử dụng kết hợp với các kỹ thuật khai phá dữ liệu khác, dự báo gồm có việc phân tích các xu hướng, phân loại, so khớp mẫu và mối quan hệ. Bằng cách phân tích các sự kiện hoặc các cá thể trong quá khứ, bạn có thể đưa ra một dự báo về một sự kiện.

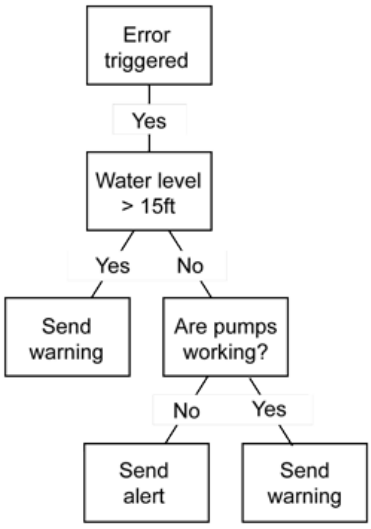
Khi sử dụng quyền hạn thẻ tín dụng, chẳng hạn, bạn có thể kết hợp phân tích cây quyết định của các giao dịch riêng lẻ trong quá khứ với việc phân loại và các sự so khớp mẫu lịch sử để nhận biết liệu một giao dịch có gian lận hay không. Rất có thể là việc thực hiện một sự so khớp giữa việc mua vé các chuyến bay đến Mỹ và các giao dịch tại Mỹ cho thấy giao dịch này hợp lệ.

#### 1.2.4.1 Các mẫu tuần tự

Thường được sử dụng trên các dữ liệu dài hạn, các mẫu tuần tự là một phương pháp có ích để nhận biết các xu hướng hay các sự xuất hiện thường xuyên của các sự kiện tương tự. Ví dụ, với dữ liệu khách hàng, bạn có thể nhận ra rằng các khách hàng cùng nhau mua một bộ sưu tập riêng lẻ về các sản phẩm tại nhiều thời điểm khác nhau trong năm. Trong một ứng dụng giỏ hàng, bạn có thể sử dụng thông tin này để tự động đề xuất rằng một số mặt hàng nào đó được thêm vào một giỏ hàng dựa trên tần suất và lịch sử mua hàng trong quá khứ của các khách hàng.

### 1.2.4.2 Các cây quyết định

Liên quan đến hầu hết các kỹ thuật khác (chủ yếu là phân loại và dự báo), cây quyết định có thể được sử dụng hoặc như là một phần trong các tiêu chí lựa chọn hoặc để hỗ trợ việc sử dụng và lựa chọn dữ liệu cụ thể bên trong cấu trúc tổng thể. Trong cây quyết định, bạn bắt đầu bằng một câu hỏi đơn giản có hai câu trả lời (hoặc đôi khi có nhiều câu trả lời hơn). Mỗi câu trả lời lại dẫn đến thêm một câu hỏi nữa để giúp phân loại hay nhận biết dữ liệu sao cho có thể phân loại dữ liệu hoặc sao cho có thể thực hiện dự báo trên cơ sở mỗi câu trả lời.




***Hình 1.3: Cây quyết định***

Hình 1.3 cho thấy một ví dụ trong đó bạn có thể phân loại một điều kiện lỗi gửi đến.

Các cây quyết định thường được sử dụng cùng với các hệ thống phân loại liên quan đến thông tin có kiểu thuộc tính và với các hệ thống dự báo, nơi các dự báo khác nhau có thể dựa trên kinh nghiệm lịch sử trong quá khứ để giúp hướng dẫn cấu trúc của cây quyết định và kết quả đầu ra.

### 1.2.4.3 Các tổ hợp

Trong thực tế, thật hiếm khi bạn sẽ sử dụng một kỹ thuật trong số những kỹ thuật riêng biệt này. Việc phân loại và phân cụm là những kỹ thuật giống nhau. Nhờ sử dụng việc phân cụm để nhận ra các thông tin lân cận gần nhất, bạn có thể tiếp tục tinh chỉnh việc phân loại của mình. Thông thường, chúng tôi sử dụng các cây quyết định để giúp xây dựng và nhận ra các loại mà chúng tôi có thể theo dõi chúng trong một thời gian dài để nhận biết các trình tự và các mẫu.

Các loại phương pháp dự báo:

- Mô hình dự báo hồi qui tuyến tính
- Mô hình dự báo hồi qui nhiều biến tuyến tính
- Mô hình dự báo hồi qui phi tuyến
- Mô hình dự báo hồi qui nhiều biến phi tính

### 1.2.4.4 Xử lý (bộ nhớ) dài hạn

Trong tất cả các phương pháp cốt lõi, thường có lý do để ghi lại thông tin và tìm hiểu từ thông tin. Trong một số kỹ thuật, việc này hoàn toàn rõ ràng. Ví dụ, với việc tìm hiểu các mẫu tuần tự và dự báo, bạn xem xét lại dữ liệu từ nhiều nguồn và nhiều cá thể thông tin để xây dựng một mẫu.

Trong một số kỹ thuật khác, quá trình này có thể rõ ràng hơn. Các cây quyết định ít khi được xây dựng một lần và không bao giờ được coi nhẹ. Khi nhận biết thông tin mới, các sự kiện và các điểm dữ liệu, có thể cần xây dựng thêm các nhánh hoặc thậm chí toàn bộ các cây mới, để đương đầu với các thông tin bổ sung.

Bạn có thể tự động hoá một số bước của quá trình này. Ví dụ, việc xây dựng một mô hình dự báo để nhận biết sự gian lận thẻ tín dụng là xây dựng các xác suất để bạn có thể sử dụng cho giao dịch hiện tại và sau đó cập nhật mô hình đó với các giao dịch mới (đã được phê duyệt). Rồi thông tin này được ghi lại sao cho có thể đưa ra quyết định một cách nhanh chóng trong lần tới.

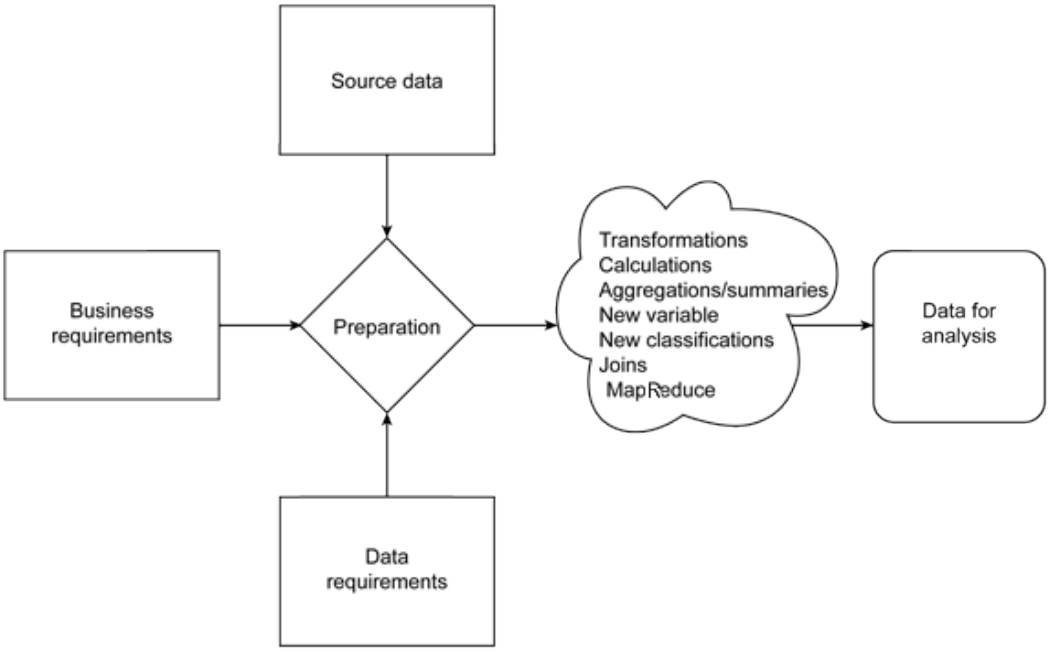
### 1.2.4.5 Chuẩn bị và triển khai dữ liệu

Bản thân việc khai phá dữ liệu dựa vào việc xây dựng một mô hình và cấu trúc dữ liệu phù hợp để có thể sử dụng mô hình và cấu trúc đó để xử lý, nhận biết và xây dựng thông tin mà bạn cần. Bất kể dạng và cấu trúc nguồn dữ liệu, hãy cấu trúc và tổ chức thông tin theo một định dạng để cho phép việc khai phá dữ liệu diễn ra theo một mô hình càng hiệu quả càng tốt.

Hãy xem xét tổ hợp các yêu cầu kinh doanh để khai phá dữ liệu, nhận ra các biến hiện có (khách hàng, các giá trị, quốc gia) và yêu cầu để tạo ra các biến mới để bạn có thể sử dụng chúng để phân tích dữ liệu trong bước chuẩn bị.

Bạn có thể tạo nên các biến phân tích của dữ liệu từ nhiều nguồn khác nhau cho một cấu trúc có thể nhận biết được duy nhất (ví dụ, bạn có thể tạo ra một lớp của một cấp cụ thể và tuổi của khách hàng hoặc một kiểu lỗi cụ thể).

Tùy thuộc vào nguồn dữ liệu của bạn, cách bạn xây dựng và chuyển dịch thông tin này là một bước quan trọng, bất kể bạn sử dụng kỹ thuật nào để cuối cùng phân tích dữ liệu. Bước này cũng dẫn đến một quá trình phức tạp trong việc nhận biết, tổng hợp, đơn giản hóa hoặc mở rộng thông tin cho phù hợp với dữ liệu đầu vào của bạn (xem Hình 1.4).




***Hình 1.4: Chuẩn bị dữ liệu***

Chuẩn bị dữ liệu

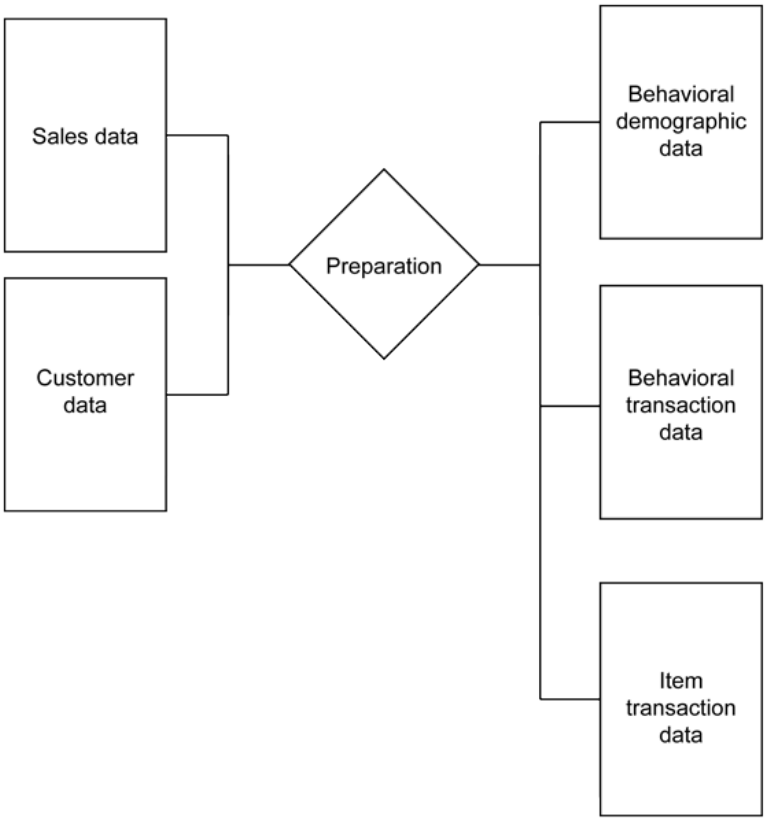
Dữ liệu nguồn, vị trí và cơ sở dữ liệu của bạn ảnh hưởng đến cách bạn xử lý và tổng hợp thông tin đó.

### 1.2.4.6 Việc xây dựng trên SQL

Việc xây dựng trên một cơ sở dữ liệu SQL thường là dễ dàng nhất trong tất cả các cách tiếp cận. Bạn hiểu rõ SQL (và cả cấu trúc bảng bên dưới mà chúng ngụ ý), nhưng bạn không thể hoàn toàn bỏ qua cấu trúc và định dạng thông tin. Ví dụ, khi xem xét hành vi của người dùng trong dữ liệu kinh doanh, bạn có thể sử dụng hai định dạng chính trong mô hình dữ liệu SQL (và việc khai phá dữ liệu nói chung): định dạng giao dịch và định dạng nhân khẩu học-hành vi.

Khi bạn sử dụng InfoSphere Warehouse, việc tạo ra một mô hình nhân khẩu học-hành vi cho các mục đích về khai phá dữ liệu khách hàng để hiểu việc mua và các mẫu mua sắm bao gồm việc lấy dữ liệu SQL nguồn của bạn dựa trên các thông tin giao dịch và các tham số đã biết của các khách hàng của bạn và xây dựng lại thông tin đó

thành một cấu trúc bảng định sẵn. Sau đó InfoSphere Warehouse có thể sử dụng thông tin này cho việc phân cụm và khai phá dữ liệu phân loại để thu được thông tin bạn cần. Dữ liệu nhân khẩu học của khách hàng và dữ liệu giao dịch doanh thu có thể được kết hợp lại và khôi phục lại vào một định dạng để cho phép phân tích dữ liệu cụ thể, như hiển thị trong Hình 1.5.




***Hình 1.5: Định dạng cho việc phân tích dữ liệu cụ thể***

Định dạng cho việc phân tích dữ liệu cụ thể

Ví dụ, với dữ liệu kinh doanh, bạn có thể muốn nhận ra các xu hướng kinh doanh các mặt hàng riêng lẻ. Bạn có thể chuyển đổi các dữ liệu doanh thu thô của các mặt hàng riêng lẻ thành thông tin giao dịch để ánh xạ mã định danh (ID) của khách hàng, dữ liệu giao dịch và mã định danh sản phẩm. Nhờ sử dụng thông tin này, thật dễ nhận ra các trình tự và các mối quan hệ với các sản phẩm riêng lẻ của các khách hàng

riêng lẻ theo thời gian. Điều đó cho phép InfoSphere Warehouse tính toán thông tin liên tục, chẳng hạn như khi một khách hàng rất có thể lại mua sản phẩm đó.

Bạn có thể xây dựng các điểm phân tích dữ liệu mới từ dữ liệu nguồn. Ví dụ, bạn có thể muốn mở rộng (hoặc tinh chỉnh) thông tin sản phẩm của mình bằng cách sắp đặt hay phân loại các sản phẩm riêng lẻ vào các nhóm lớn hơn và sau đó phân tích dữ liệu dựa trên các nhóm này thay cho việc phân tích một sản phẩm riêng lẻ.

#### 1.2.4.7 Các cơ sở dữ liệu tài liệu và MapReduce

Người ta thiết kế MapReduce để xử lý nhiều cơ sở dữ liệu tài liệu hiện đại và NoSQL, như Hadoop, để đối phó với các tập hợp dữ liệu rất lớn và thông tin không phải lúc nào cũng theo định dạng bảng. Khi bạn làm việc với phần mềm khai phá dữ liệu, khái niệm này có thể vừa có ích và vừa có vấn đề.

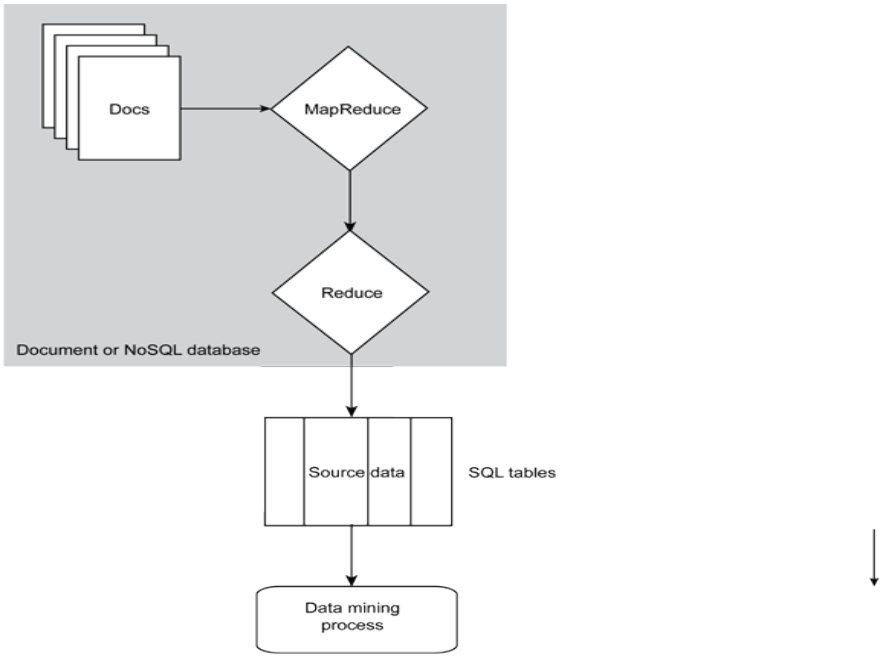
Vấn đề chính với dữ liệu dựa trên tài liệu là ở chỗ định dạng không có cấu trúc có thể cần phải xử lý nhiều hơn là bạn mong đợi để có được thông tin bạn cần. Nhiều bản ghi khác nhau có thể chứa dữ liệu giống nhau. Việc thu thập và phối hợp thông tin này để xử lý nó một cách dễ dàng hơn dựa vào việc chuẩn bị và các giai đoạn của MapReduce.

Trong một hệ thống dựa trên MapReduce, vai trò của bước ánh xạ là lấy dữ liệu nguồn và chuẩn hóa thông tin đó thành một dạng chuẩn của đầu ra. Bước này có thể là một quá trình tương đối đơn giản (nhận biết các trường hoặc các điểm dữ liệu chính) hoặc có thể là một quá trình phức tạp hơn (phân tích cú pháp và xử lý thông tin để tạo ra dữ liệu mẫu). Quá trình ánh xạ tạo ra định dạng chuẩn hóa để bạn có thể sử dụng định dạng đó làm định dạng cơ sở của mình.

Sự rút gọn là việc tóm tắt hoặc định lượng thông tin và sau đó xuất ra thông tin đó dưới dạng một cấu trúc chuẩn hóa, dựa trên các tổng số, các tổng, số liệu thống kê hay phân tích khác mà bạn đã chọn để xuất ra.

Việc truy vấn dữ liệu này thường rất phức tạp, ngay cả khi bạn sử dụng các công cụ được thiết kế để làm việc này. Trong một bài tập khai phá dữ liệu, cách tiếp cận lý tưởng là sử dụng giai đoạn khai phá dữ liệu của MapReduce làm một phần trong bài tập chuẩn bị dữ liệu của bạn.

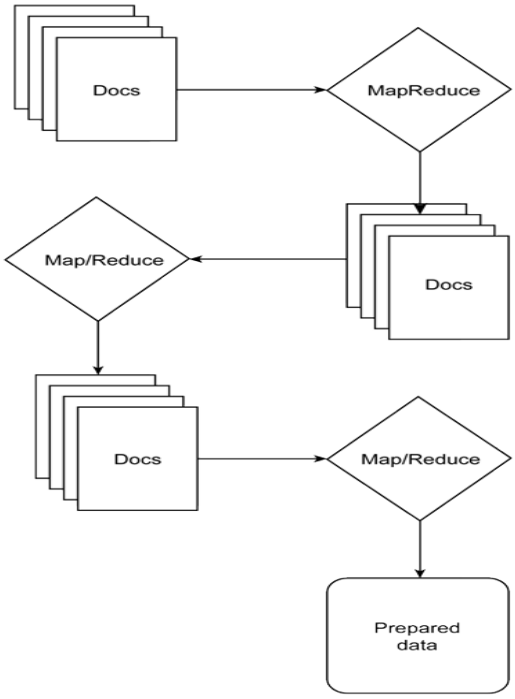
Ví dụ, nếu bạn đang xây dựng một bài tập khai phá dữ liệu để kết hợp hoặc phân cụm, giai đoạn đầu tiên tốt nhất là xây dựng một mô hình thống kê thích hợp để bạn có thể sử dụng mô hình này để nhận biết và trích ra thông tin cần thiết. Việc sử dụng giai đoạn MapReduce để trích ra và tính toán thông tin thống kê đó rồi nhập nó vào phần còn lại của quá trình khai phá dữ liệu, dẫn đến một cấu trúc như thể hiện trong Hình 1.6.




***Hình 1.6: Cấu trúc KPDL***

Trong ví dụ trước, chúng tôi đã thực hiện việc xử lý (trong trường hợp này là MapReduce) của dữ liệu nguồn trong một cơ sở dữ liệu tài liệu và đã chuyển dịch nó sang một định dạng bảng trong một cơ sở dữ liệu SQL cho các mục đích khai phá dữ liệu.

Làm việc với thông tin phức tạp hay thậm chí chưa định dạng này có thể cần phải chuẩn bị và xử lý thông tin còn phức tạp hơn. Có một số kiểu và cấu trúc dữ liệu phức tạp không thể được xử lý và được chuẩn bị chỉ trong một bước thành kết quả đầu ra mà bạn cần. Ở đây bạn có thể nối chuỗi đầu ra của MapReduce của bạn hoặc để ánh xạ và tạo ra cấu trúc dữ liệu mà bạn cần theo tuần tự, như trong Hình 1.7, hoặc riêng lẻ để tạo ra nhiều bảng dữ liệu đầu ra.




***Hình 1.7: Nối chuỗi đầu ra của MapReduce của bạn theo tuần tự***

Ví dụ, việc lấy thông tin ghi nhật ký thô từ một cơ sở dữ liệu tài liệu và chạy MapReduce để tạo ra một khung nhìn tóm tắt các thông tin đó theo ngày có thể được thực hiện chỉ một lần duy nhất. Việc tạo lại thông tin và kết hợp đầu ra đó với một ma trận quyết định (được mã hóa trong giai đoạn MapReduce thứ hai) và sau đó tiếp tục đơn giản hóa thành một cấu trúc tuần tự, là một ví dụ hay của quá trình nối chuỗi này. Chúng tôi cần có toàn bộ dữ liệu đã thiết lập trong giai đoạn MapReduce để hỗ trợ dữ liệu của bước riêng này.

Bất kể dữ liệu nguồn của bạn, có nhiều công cụ có thể sử dụng tệp phẳng, CSV hoặc các nguồn dữ liệu khác. Ví dụ, InfoSphere Warehouse có thể phân tích cú pháp các tệp phẳng ngoài một liên kết trực tiếp đến một kho dữ liệu DB2.

## 1.3 KẾT LUẬN

Khai thác dữ liệu là một công cụ được sử dụng để trích xuất thông tin quan trọng từ dữ liệu hiện có và cho phép ra quyết định kinh doanh tốt hơn trong các ngành công nghiệp ngân hàng và bán lẻ. Họ sử dụng kho dữ liệu để kết hợp các dữ liệu khác nhau từ cơ sở dữ liệu thành một định dạng có thể chấp nhận từ đó có thể tiến hành việc khai phá dữ liệu. Dữ liệu sau đó được phân tích và các thông tin thu được sẽ hỗ trợ cho tổ chức ra quyết định. Kỹ thuật khai phá dữ liệu có thể rất hữu ích cho các ngân hàng thực hiện quá trình kinh doanh tốt hơn, thu hút được khách hàng mới, phát hiện gian lận, cung cấp sản phẩm dựa trên phân tích của các khách hàng để duy trì tốt hơn mối quan hệ của khách hàng. Những ngân hàng này đã nhận ra sự hữu ích của khai phá dữ liệu và đang trong quá trình xây dựng một môi trường khai phá dữ liệu sẽ có được lợi ích to lớn cho quá trình ra quyết định của họ và chiếm được lợi thế cạnh tranh đáng kể trong tương lai.

# CHƯƠNG 2: PHÂN CỤM DỮ LIỆU VÀ PHƯƠNG PHÁP PHÂN CỤM DỰA TRÊN LƯỚI

## 2.1 KHÁI NIỆM CHUNG

Khai phá dữ liệu (Datamining) là quá trình trích xuất các thông tin có giá trị tiềm ẩn bên trong tập dữ liệu lớn được lưu trữ trong các cơ sở dữ liệu, kho dữ liệu. Người ta định nghĩa: "Phân cụm dữ liệu là một kĩ thuật trong Data Minhing, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, quan trọng trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định"

Như vậy, phân cụm dữ liệu là quá trình chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các phần tử trong một cụm "tương tự" với nhau và các phần tử trong các cụm khác nhau sẽ "phi tương tự". Số các cụm dữ liệu được phân ở đây có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định.

## 2.2 BÀI TOÁN PHÂN CỤM TRÊN LƯỚI

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh giá hiệu quả của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, "natural clusters", "useful" clusters, outlier detection

*Kỹ thuật phân cụm có thể áp dụng trong rất nhiều lĩnh vực như:*

- Marketing: Xác định các nhóm khách hàng (khách hàng tiềm năng, khách hàng giá trị, phân loại và dự đoán hành vi khách hàng ,…) sử dụng sản phẩm hay dịch vụ của công ty để giúp công ty có chiến lược kinh doanh hiệu quả hơn;
- Biology: Phận nhóm động vật và thực vật dựa vào các thuộc tính của chúng;
- Libraries: Theo dõi độc giả, sách, dự đoán nhu cầu của độc giả…;
- Insurance, Finance: Phân nhóm các đối tượng sử dụng bảo hiểm và các dịch vụ tài chính, dự đoán xu hướng (trend) của khách hàng, phát hiện gian lận tài chính (identifying frauds);
- WWW: Phân loại tài liệu (document classification); phân loại người dùng web (clustering weblog); …

## 2.3 CÁC PHƯƠNG PHÁP PHÂN CỤM

### 2.3.1 Phương pháp phân cụm phân hoạch

Ý tưởng chính của kỹ thuật này là phân một tập dữ liệu có n phần tử cho trước thành k nhóm dữ liệu sao cho mỗi phần tử dữ liệu chỉ thuộc về một nhóm dữ liệu và mỗi nhóm dữ liệu có tối thiểu ít nhất một phần tử dữ liệu. Các thuật toán phân hoạch có độ phức tạp rất lớn khi xác định nghiệm tối ưu toàn cục cho vấn đề PCDL, vì nó phải tìm kiếm tất cả các cách phân hoạch có thể được. Chính vì vậy, trên thực tế người ta thường đi tìm giải pháp tối ưu cục bộ cho vấn đề này bằng cách sử dụng một hàm tiêu chuẩn để đánh giá chất lượng của các cụm cũng như để hướng dẫn cho quá trình tìm kiếm phân hoạch dữ liệu. Với chiến lược này, thông thường người ta bắt đầu khởi tạo một phân hoạch ban đầu cho tập dữ liệu theo phép ngẫu nhiên hoặc theo heuristic và liên tục tinh chỉnh nó cho đến khi thu được một phân hoạch mong muốn, thoả mãn các điều kiện ràng buộc cho trước. Các thuật toán phân cụm phân hoạch cố gắng cải tiến tiêu chuẩn phân cụm bằng cách tính các giá trị đo độ tương tự giữa các đối tượng dữ liệu và sắp xếp các giá trị này, sau đó thuật toán lựa chọn một giá trị trong dãy sắp xếp sao cho hàm tiêu chuẩn đạt giá trị tối thiểu. Như vậy, ý tưởng chính của thuật toán phân cụm phân hoạch tối ưu cục bộ là sử dụng chiến lược ăn tham để tìm kiếm nghiệm.

Lớp các thuật toán phân cụm phân hoạch bao gồm các thuật toán đề xuất đầu tiên trong lĩnh vực KPDL cũng là các thuật toán được áp dụng nhiều trong thực tế như k-means, PAM, CLARA, CLARANS.

### 2.3.2 Phương pháp phân cụm phân cấp

Phân cụm phân cấp sắp xếp một tập dữ liệu đã cho thành một cấu trúc có dạng hình cây, cây phân cấp này được xây dựng theo kỹ thuật đệ quy. Cây phân cụm có thể được xây dựng theo hai phương pháp tổng quát: phương pháp "trên xuống" (Top down) và phương pháp "dưới lên" (Bottom up).

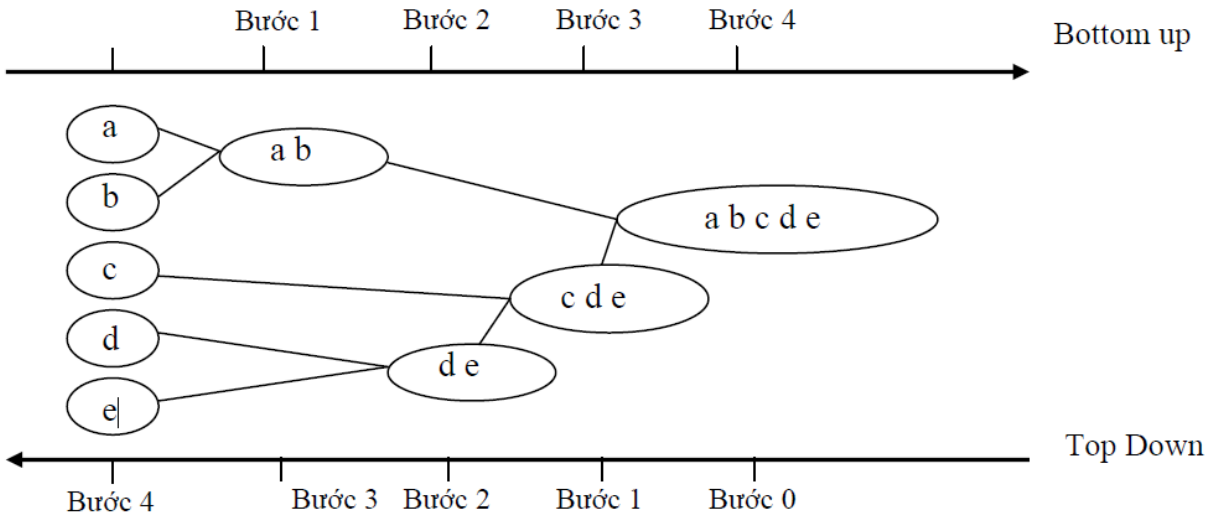
**Phương pháp Bottom up**:

Phương pháp này bắt đầu với mỗi đối tượng được khởi tạo tương ứng với các cụm riêng biệt, sau đó tiến hành nhóm các đối tượng theo một độ đo tương tự (như khoảng cách giữa hai trung tâm của hai nhóm), quá trình này được thực hiện cho đến khi tất cả các nhóm được hòa nhập vào một nhóm (mức cao nhất của cây phân cấp) hoặc cho đến khi các điều kiện kết thúc thoả mãn. Như vậy, cách tiếp cận này sử dụng chiến lược ăn tham trong quá trình phân cụm.

**Phương pháp Top Down**:

Bắt đầu với trạng thái là tất cả các đối tượng được xếp trong cùng một cụm. Mỗi vòng lặp thành công, một cụm được tách thành các cụm nhỏ hơn theo giá trị của một phép đo độ tương tự nào đó cho đến khi mỗi đối tượng là một cụm hoặc cho đến khi điều kiện dừng thoả mãn. Cách tiếp cận này sử dụng chiến lược chia để trị trong quá trình phân cụm.

Sau đây là minh họa chiến lược phân cụm phân cấp bottom up và Top down.



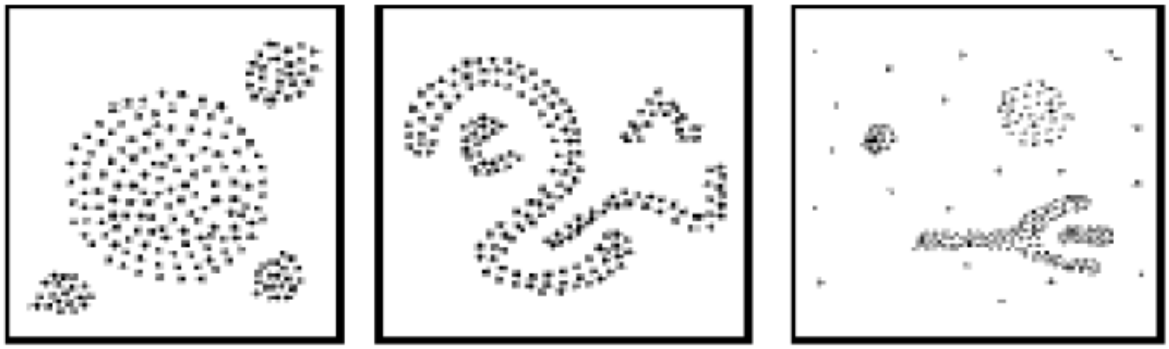

***Hình 2.1. Các chiến lược phân cụm phân cấp***

Trong thực tế áp dụng, có nhiều trường hợp người ta kết hợp cả hai phương pháp phân cụm phân hoạch và phương phân cụm phân cấp, nghĩa là kết quả thu được của phương pháp phân cấp có thể cải tiến thông quan bước phân cụm phân hoạch. Phân cụm phân hoạch và phân cụm phân cấp là hai phương pháp PCDL cổ điển, hiện nay đã có nhiều thuật toán cải tiến dựa trên hai phương pháp này đã được áp dụng phổ biến trong KPDL. Một số thuật toán phân cụm phân cấp điển hình như CURE, BIRCH, Chemeleon, AGNES, DIANA ,...

### 2.3.3 Phương pháp phân cụm dựa trên mật độ

Phương pháp này nhóm các đối tượng theo hàm mật độ xác định. Mật độ được định nghĩa như là số các đối tượng lân cận của một đối tượng dữ liệu theo một ngưỡng nào đó. Trong cách tiếp cận này, khi một cụm dữ liệu đã xác định thì nó tiếp tục được phát triển thêm các đối tượng dữ liệu mới miễn là số các đối tượng lân cận của các đối tượng này phải lớn hơn một ngưỡng đã được xác định trước. Phương pháp phân cụm dựa vào mật độ của các đối tượng để xác định các cụm dữ liệu và có thể phát hiện ra các cụm dữ liệu với hình thù bất kỳ. Tuy vậy, việc xác định các tham số mật độ của thuật toán rất khó khăn, trong khi các tham số này lại có tác động rất lớn đến kết quả PCDL. Hình minh hoạ về các cụm dữ liệu với các hình thù khác nhau dựa trên mật độ được khám phá từ 3 CSDL

khác nhau:



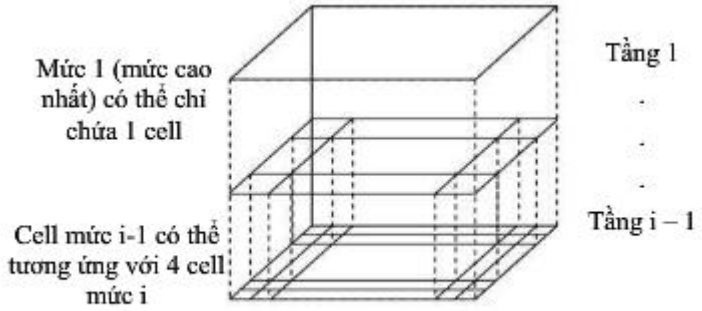

***Hình 2.2: Một số hình dạng khám phá bởi phân cụm dưa trên mật độ***

Các cụm có thể được xem như các vùng có mật độ cao, được tách ra bởi các vùng không có hoặc ít mật độ. Khái niệm mật độ ở đây được xem như là các số các đối tượng láng giềng. Một số thuật toán PCDL dựa trên mật độ điển hình như [2][3]: DBSCAN, OPTICS, DENCLUE, SNN.

### 2.3.4 Phương pháp phân cụm dựa trên lưới

Kỹ thuật phân cụm dựa trên mật độ không thích hợp với dữ liệu nhiều chiều, để giải quyết cho đòi hỏi này, người ta đã sử dụng phương pháp phân cụm dưa trên lưới. Đây là phương pháp dựa trên cấu trúc dữ liệu lưới để PCDL, phương pháp này chủ yếu tập trung áp dụng cho lướp dữ liệu không gian, Thí dụ như dữ liệu được biểu diễn dưới dạng cấu trúc hình học của đối tương trọng không gian cùng với các quan hệ, các thuộc tính, các hoạt động của chúng. Mục tiêu của phương pháp này là lượng hóa tập dữ liệu thành các ô, các ô này tạo thành cấu trúc dữ liệu lưới, sau đó các thao tác PCDL làm việc với các đối tượng trong từng ô này. CÁc tiếp cận dựa trên lưới này không di chuyển các đối tượng trong các ô mà xây dựng nhiều mức phân ấp của nhóm các đối tượng trong 1 ô. Trong ngữ cảnh này, phương pháp này gần giống với phương pháp phân cụm phân cấp nhưng chỉ có điều chúng không trộn các ô. Do vậy các cumgj không dựa trên độ đo khoảng cachs (hay còn gọi là độ đo tương tự đối với các dữ liệu không gian) mà nó được quyết định bởi một tham số xác định trước. Ưu điểm của

phương pháp PCDL dựa trên lưới là thời gian xử lý nhanh và độc lập với số đối tượng dữ liệu trong tập dữ liệu ban đầu, thay vào đó là chúng phụ thuộc vào số ô trong mỗi chiều của không gian lưới. Các thuật toán PCDL dựa trên lưới có thể là String, Wave Cluster..




***Hình 2.3: Phân cụm dựa trên lưới***

### 2.3.5 Phương pháp phân cụm dựa trên mô hình

Phương pháp này cố gắng khám phá các phép xấu củ tốt của các tham số mô hình sao cho khớp với dữ liệu một cách tốt nhất. Chúng có thể dử dụng chiến lược phân cụm phân hoạch hoặc chiến lược phân cụm phân cấp, dựa trên cấu trúc hoặc mô hình mà chúng giả định về tập dữ liệu và cách mà chúng giả định về tập dữ liệu và cách mà chúng ta chỉnh các mô hình này để nhận dạng ra các phân hoạch.

Phương pháp PCDL dựa trên mô hình cố gắng khớp giữa dữ liệu với mô hình toán học, nó dựa trên giả định rằng dữ liệu được tạp ra bằng hỗn hợp phân phối xác suất cơ bản. Các thuật toán phân cụm dựa trên mô hình có hay cách tiếp cận chính: mô hình thống kê và mạng Nơron. Phương pháp này gần giống với phương pháp dựa trên mật độ, bở vù chúng phát triển các cumgj riêng việt nhằm cải tiến các mô hình đã được xác định trước đó, nhưng khôi khi nó không bắt đầu với một số cụm cố định và không sử dụng cùng một khái niệm mật độ cho các cụm

# CHƯƠNG 3: ỨNG DỤNG KĨ THUẬT PHÂN CỤM DỰA TRÊN LƯỚI TRONG LĨNH VỰC TÀI CHÍNH VÀ BÁO CÁO KẾT QUẢ NGHIÊN CỨU

Ngành công nghiệp ngân hàng trên toàn thế giới đã trải qua những thay đổi to lớn trong cách thức kinh doanh. Với việc thực hiện các nhiệm vụ kinh doanh đặc thù của mình trong thời gian gần đây, các ngân hàng đã có sự thay đổi lớn đó là chấp nhận và áp dụng công nghệ thông vào việc kinh doanh của mình. Như một kết quả hiển nhiên, việc thực hiện giao dịch đã trở nên dễ dàng và đồng thời khối lượng dữ liệu từ các giao dịch này đã tăng lên đáng kể. Nó là vượt quá khả năng của con người để phân tích số lượng dữ liệu thô khổng lồ này và chuyển đổi nó thành tri thức hữu ích cho các tổ chức. Khai phá dữ liệu có thể giúp đỡ bằng cách đóng góp trong việc giải quyết các vấn đề kinh doanh bằng cách tìm kiếm các mẫu, các mối kết hợp và các mối tương quan được ẩn chứa trong các thông tin liên quan tới việc kinh doanh được lưu trữ trong cơ sở dữ liệu. Bằng cách sử dụng khai phá dữ liệu để phân tích mô hình và xu hướng này, những người thực hiện công việc kinh doanh trong các ngân hàng có thể dự đoán với độ chính xác tăng lên, khách hàng sẽ phản ứng thế nào với việc điều chỉnh lãi suất, trong đó khách hàng nào sẽ có khả năng chấp nhận sự chào hàng sản phẩm mới, khách hàng nào sẽ có nguy cơ rủi ro cao hơn mặc định trên một khoản vay, và làm thế nào để mối quan hệ khách hàng ngày càng có lợi hơn.

Ngành ngân hàng nhận ra tầm quan trọng của thông tin mà mỗi ngân hàng có về khách hàng của mình một cách rộng rãi. Chắc chắn rằng, họ có một biển thông tin khách hàng, bao gồm nhân khẩu học của khách hàng, dữ liệu giao dịch, creditcards sử dụng mô hình, và nhiều thuộc tính khác nữa. Khi ngành ngân hàng là một bộ phận trong ngành công nghiệp dịch vụ, thì nhiệm vụ duy trì các mối quan hệ khách hàng (CRM: Customer relationship Management) một cách có hiệu quả là một vấn đề quan trọng. Để làm được điều này, các ngân hàng cần phải đầu tư các nguồn lực để hiểu rõ hơn về khách hàng hiện tại và tiềm năng của họ. Bằng cách sử dụng các công cụ khai phá dữ liệu phù hợp, sau đó có thể cung cấp các sản phẩm và dịch vụ thích hợp cho khách hàng của họ.

Có rất nhiều lĩnh vực, trong đó khai phá dữ liệu có thể được ứng dụng trong ngành công nghiệp ngân hàng, trong đó bao gồm việc phân khúc khách hàng và phân chia lợi nhuận, chấm điểm và phê duyệt tín dụng, dự đoán thanh toán mặc định, quảng bá sản phẩm, phát hiện các giao dịch gian lận, quản lý tiền mặt và các hoạt động dự báo, tối ưu hóa danh mục đầu tư chứng khoán và xếp hạng đầu tư. Bằng cách phân tích các dữ liệu trong quá khứ, khai phá dữ liệu có thể giúp các ngân hàng dự đoán số lượng khách hàng có khả năng thay đổi thẻ tín dụng của họ, từ đó họ có thể lập kế hoạch và triển khai ưu đãi đặc biệt khác nhau để giữ lại những khách hàng của mình. Sau đây là một số ví dụ về phương thức mà ngành ngân hàng đã sử dụng có hiệu quả kỹ thuật khai phá dữ liệu trong các lĩnh vực này.

## 3.1. MARKETING

Một trong những lĩnh vực được ứng dụng rộng rãi nhất cho ngành ngân hàng của kỹ thuật khai phá dữ liệu đó là lĩnh vực quảng bá sản phẩm. Bộ phận tiếp thị và bán hàng của các Ngân hàng có thể sử dụng kỹ thuật khai phá dữ liệu để phân tích cơ sở dữ liệu về khách hàng. Khai phá dữ liệu thực hiện các phân tích khác nhau trên bộ dữ liệu thu thập được để xác định hành vi của người tiêu dùng với sự tham khảo sản phẩm, giá và kênh phân phối. Với sự phản hồi của khách hàng đối với các sản phẩm hiện có và các sản phẩm mới, các ngân hàng sẽ có các chiến lược quảng bá sản phẩm, nâng cao chất lượng sản phẩm và dịch vụ và đạt được lợi thế cạnh tranh. Phân tích ngân hàng cũng có thể phân tích các xu hướng trong quá khứ, xác định nhu cầu hiện tại và dự báo hành vi khách hàng các sản phẩm và dịch vụ khác nhau để thu các cơ hội kinh doanh hơn và dự đoán mô hình hành vi. Kỹ thuật khai thác dữ liệu cũng giúp xác định khách hàng nào sẽ mang lại lợi nhuận và khách hàng nào không mang lại lợi nhuận. Các kỹ thuật khai phá dữ liệu có thể được sử dụng để xác định phản ánh của khách hàng như thế nào khi ngân hàng thực hiện điều chỉnh lãi suất.

## 3.2 QUẢN LÝ RỦI RO

Khai phá dữ liệu được sử dụng rộng rãi để quản lý rủi ro trong ngành công nghiệp ngân hàng. Giám đốc điều hành ngân hàng cần phải biết rằng các khách hàng

mà họ đang có liệu đáng tin cậy hay không. Khi cung cấp thẻ tín dụng mới cho khách hàng, mở rộng số lượng khách hàng hiện tại của tín dụng và phê duyệt các khoản vay, họ có thể ra các mang lại sự quyết định rủi ro cho các ngân hàng nếu họ không biết bất cứ điều gì về khách hàng của họ.

Ngân hàng tiến hành quá trình cho các khách hàng của mình vay vốn bằng cách kiểm tra các chi tiết khác nhau liên quan đến việc cho vay như số tiền vay, lãi suất cho vay, kỳ hạn trả nợ, loại tài sản thế chấp, tình hình nhân sự, thu nhập và lịch sử tín dụng của họ. Khách hàng dài hạn với ngân hàng, với các nhóm thu nhập cao có thể nhận được các khoản vay rất dễ dàng. Mặc dù, các ngân hàng đã thận trọng trong khi cung cấp vốn vay cho khách hàng, nhưng vẫn có những khoản nợ mặc định của khách hàng. Kỹ thuật khai phá dữ liệu giúp phân biệt người trả nợ kịp thời với những người không có khả năng trả nợ kịp thời.

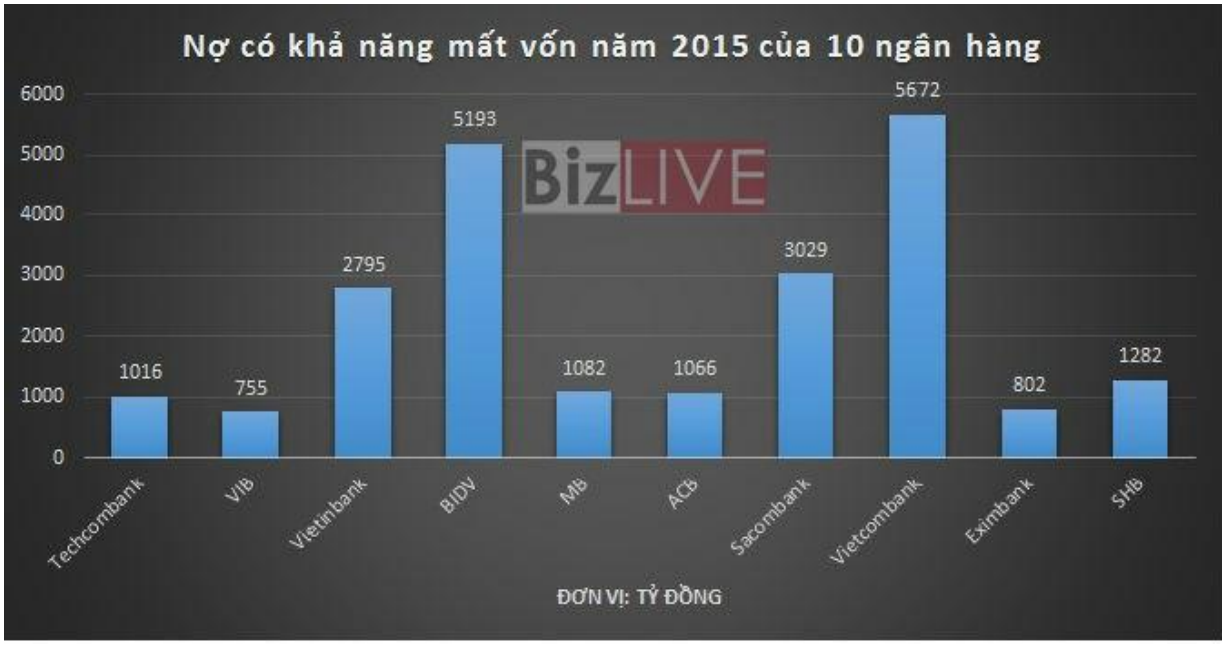
Trên thực tế điểm tín dụng là một trong những công cụ quản lý rủi ro tài chính đầu tiên được phát triển. Điểm tín dụng có thể có giá trị cho người cho vay trong ngành ngân hàng khi đưa ra những quyết định cho vay. Khai phá dữ liệu cũng có thể tìm ra được hành vi tín dụng của từng khách hàng cá nhân với các khoản vay trả góp, thế chấp, tín dụng, bằng việc sử dụng các đặc điểm như lịch sử tín dụng, thời gian làm việc và thời gian cư trú, như vậy đã cho phép một người cho vay đánh giá khách hàng và quyết định khách hàng đó có là một ứng cử viên tốt cho một khoản vay, hoặc nếu có nguy rủi ro nào tiềm ẩn. Khi biết được những gì là cơ hội sẵn có của một khách hàng, tức là khi đó các ngân hàng đang ở trong một vị trí tốt hơn để giảm thiểu rủi ro.

Tính đến thời điểm giữa năm 2016, theo báo cáo tài chính của các ngân hàng gửi cho ngân hàng nhà nước thì tăng trưởng tín dụng toàn ngành 6 tháng đầu năm nay đạt 8,16%, cao hơn mức 7,86% cùng kỳ năm 2015. Đây cũng là mức khá cao so với cùng kỳ những năm gần đây. Dù hoạt động ngân hàng khởi sắc, nhưng gam màu tối trong bức tranh lợi nhuận ngân hàng năm nay chính là trích lập dự phòng rủi ro. Trích lập dự phòng rủi ro tăng vọt đã bào mòn lợi nhuận của nhiều ông lớn ngân hàng.

Điển hình là Ngân hàng TMCP Xuất Nhập Khẩu (Eximbank). Quý IV/2015, lợi nhuận trước thuế Eximbank bị lỗ 588 tỷ đồng, sau thuế lỗ 463 tỷ đồng. Nguyên nhân khiến Eximbank lỗ trong quý này là do ngân hàng đã trích lập dự phòng rủi ro lên tới 935 tỷ đồng.

Sacombank cũng bị lỗ trong quý IV/2015 do dự phòng rủi ro quý cuối năm tăng vọt, từ mức 187 tỷ đồng cùng kỳ năm 2014 lên tới 1.125 tỷ đồng. Lũy kế cả năm trích lập dự phòng cũng tăng gấp hơn 2 lần khiến cho lợi nhuận của ngân hàng bị ảnh hưởng nghiêm trọng. Điều đó khiến cho quý IV/2015, Sacombank lỗ trước thuế 738 tỷ đồng và lỗ sau thuế 583 tỷ đồng.

Mặc dù lợi nhuận vẫn cao nhưng Vietcombank cũng là ngân hàng có mức trích lập dự phòng rủi ro cao. Năm 2015, trích lập dự phòng rủi ro của Vietcombank cũng tăng lên 8.609 tỷ đồng, tăng 21,5% so với năm 2014. Mức trích lập này cũng đã ảnh hưởng tới lợi nhuận của ông lớn này.




***Hình 3.1: Nợ có khả năng mấy vốn của năm 2015***

BIDVcũng bị sụt giảm mạnh do trích lập cao quý IV, BIDV phải trích lập dự phòng rủi ro tín dụng là 1.842 tỷ đồng, lũy kế cả năm là 5.802 tỷ đồng. Điều đó khiến cho lợi nhuận trước thuế của ngân hàng này trong năm 2015 chỉ đạt 7.944 tỷ đồng, trong khi lợi nhuận thuần từ hoạt động kinh doanh trước chi phí dự phòng rủi ro tín dụng cả năm là 13.746 tỷ đồng. Một trong những nguyên nhân dẫn đến nợ xấu tăng là do tín dụng của ngân hàng tăng trong 2 quý đầu năm nay, trong khi đó tác động của kinh tế thế giới phục hồi chậm và kinh tế trong nước tăng trưởng chậm lại làm cho DN gặp khó khăn trong sản xuất, kinh doanh, hàng tồn kho tăng cao, không còn khả năng trả nợ ngân hàng.

## 3.3 PHÁT HIỆN GIAN LẬN

Một lĩnh vực khác trong khai phá dữ liệu có thể được sử dụng trong ngành công nghiệp ngân hàng là việc phát hiện gian lận. Có thể phát hiện các hành động gian lận là một mối quan tâm ngày càng tăng cho nhiều doanh nghiệp, và với sự giúp đỡ của kỹ thuật khai phá dữ liệu các hành động gian lận ngày càng được phát hiện nhiều hơn. Có hai phương pháp tiếp cận phổ biến đã được phát triển bởi tổ chức tài chính để phát hiện các mô hình gian lận. Phương pháp tiếp cận thứ nhất, một ngân hàng cần phải sử dụng đến kho dữ liệu của bên thứ ba và sử dụng các kỹ thuật khai phá dữ liệu để xác định mô hình gian lận. Sau đó, các ngân hàng có thể tham chiếu chéo những mẫu với cơ sở dữ liệu riêng của mình. Phương pháp thứ hai, gian lận được nhận dạng mẫu dựa trên các mẫu thông tin nội bộ riêng của mình mà không phải nhờ vào bên thứ ba. Tuy nhiên, trên thực tế hầu hết các ngân hàng đang sử dụng kết hợp cả hai phương pháp tiếp cận trên.

## 3.4 QUẢN TRỊ QUAN HỆ KHÁCH HÀNG

Trong thời đại cạnh tranh khốc liệt ngày nay nói chung, đặc biệt là trong ngành ngân hàng thì khách hàng được coi là thượng đế. Khai phá dữ liệu là rất hữu ích trong tất cả ba giai đoạn trong một chu kỳ mối quan hệ khách hàng: Tìm kiếm khách hàng, tăng giá trị của khách hàng và duy trì khách hàng. Tìm kiếm khách hàng, chăm sóc và

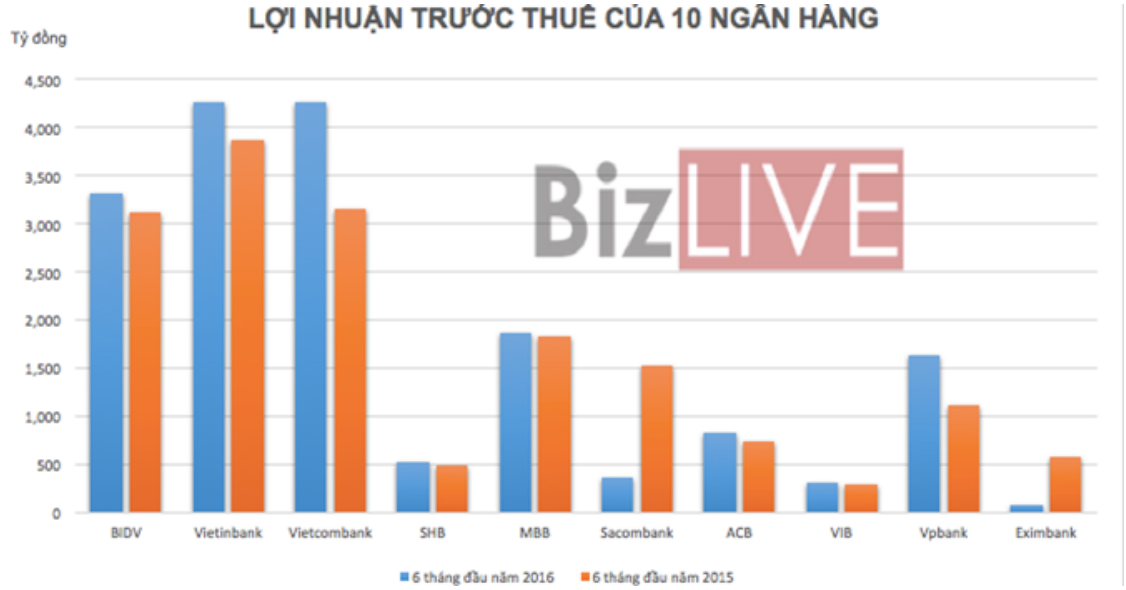
duy trì những khách hàng đã có là mối quan tâm rất quan trọng đối với một lĩnh vực kinh doanh nào, đặc biệt là lĩnh vực ngân hàng.

Ngày nay, khách hàng có nhiều sự lựa chọn bởi nhiều loại sản phẩm và dịch vụ được cung cấp bởi các ngân hàng khác nhau. Do đó, các ngân hàng phải đáp ứng nhu cầu của khách hàng bằng cách cung cấp các sản phẩm và dịch vụ mà họ ưa thích. Điều này sẽ dẫn đến sự trung thành của khách hàng và khả năng giữ khách hàng của các ngân hàng. Kỹ thuật khai phá dữ liệu giúp ngân hàng phân tích và nhận định được đâu là các khách hàng trung thành và đâu là các khách hàng có xu hương chuyển sang ngân hàng khác với mong muốn dịch vụ tốt hơn. Nếu khách hàng chuyển từ ngân hàng của mình sang ngân hàng khác, lý do cho việc chuyển như vậy và giao dịch cuối cùng được thực hiện trước khi chuyển có thể được biết đó sẽ giúp các ngân hàng hoạt động tốt hơn và giữ chân khách hàng của mình.

## 3.5 ĐÁNH GIÁ KẾT QUẢ NGHIÊN CỨU

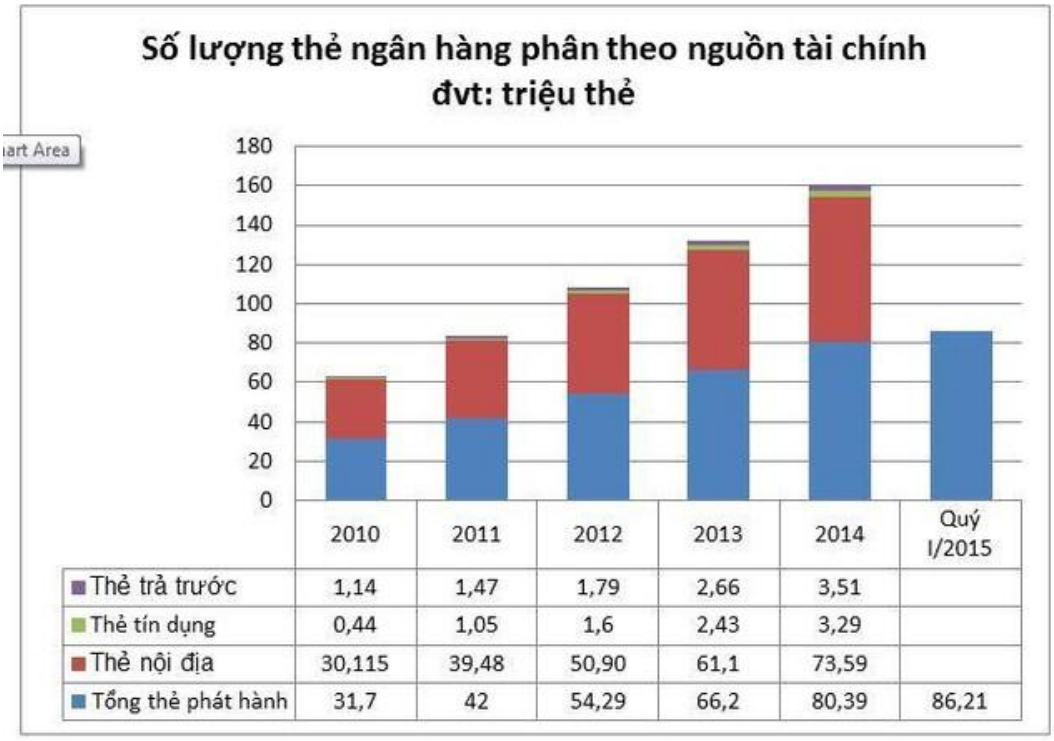
### 3.5.1 Nghiên cứu tập trung ứng dụng vào lĩnh vực quảng bá và bán sản phẩm trong hệ thống ngân hàng Việt Nam

Những năm vừa qua, mạng lưới phục vụ của ngành ngân hàng Việt Nam liên tục tăng trưởng để đáp ứng nhu cầu của người dân. Theo báo cáo của Hiệp hội Thẻ ngân hàng Việt Nam, so với năm 2006, tính đến hết năm 2015, số thẻ phát hành đã tăng từ 5 triệu lên 32 triệu thẻ. Số máy POS tăng gấp 5 lần - lên gần 52.000 POS, số máy ATM cũng tăng hơn 4 lần – đạt gần 12.000 ATM và rất nhiều phương tiện thanh toán qua internet đang được phổ cập.




***Hình 3.2: Lợi nhuận trước thuế của các ngân hàng năm 2015-2016***

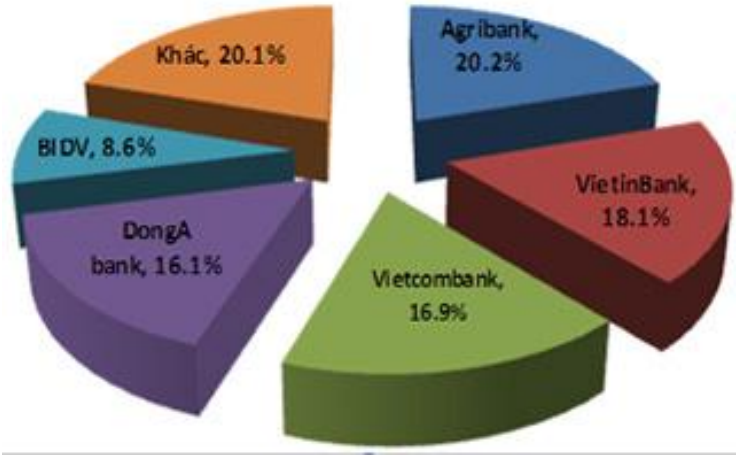
Với việc ứng dụng khai phá dữ liệu dạng lưới vào việc phát triển các hệ thống liên ngân hàng mà tính tới thời điểm hiện nay, hệ thống ngân hàng trong nước và quốc tế đã có những bước tiến đáng kinh ngạc. Ngoài mạng lưới hàng ngàn chi nhánh/phòng giao dịch của các ngân hàng trên cả nước, chính sự hỗ trợ tích cực của công nghệ đã đưa tiện ích ngân hàng đến với hàng chục triệu người tiêu dùng Việt Nam.




***Hình 3.3: Tổng kết doanh số phát triển thẻ tính đến 2015***

Trước đây, việc quản lý khách hàng rất rải rác và vô cùng bất tiện cho khách hàng. Tiền gửi ở đâu, phải đến đó, không thể rút ở điểm giao dịch khác, mặc dù các điểm này đều trong cùng hệ thống một ngân hàng. Thậm chí khách hàng muốn giao dịch ở bao nhiều điểm thì phải mở bấy nhiêu tài khoản. Với sự ra đời của hệ thống liên ngân hàng, và mỗi ngân hàng trang bị core banking hiện đại, khách hàng chỉ cần có một mã duy nhất ở ngân hàng là có thể giao dịch với rất nhiều sản phẩm, và ở bất cứ điểm giao dịch trong cùng hoặc không trong cùng một hệ thống.

Trước năm 2014, có 2 tổ chức lớn đảm nhiệm xây dựng hệ thống chuyển mạch tài chính quốc gia nhằm kết nối các hệ thống thanh toán thẻ nói chung, hệ thống ATM/POS nói riêng của các ngân hàng Việt Nam.




***Hình 3.4: Biểu đồ phân chia thị phần thẻ tính đến 2015***

Việc kết nối này sẽ tạo điều kiện cho các ngân hàng thành viên có khả năng mở rộng mạng lưới dịch vụ của mình với đầu tư hợp lý, tránh được việc đầu tư trùng lặp của các ngân hàng cho hệ thống cơ sở hạ tầng và mạng lưới các thiết bị đầu cuối trên phạm vi toàn quốc, đó chính là SmartLink và Banknetvn. Trong đó có sự phân chia về thị phần phát triển thẻ của các ngân hàng.

Tuy nhiên, hướng tới mục đích tạo sự đồng nhất trong quản lý và phục vụ tốt nhất cho khách hàng, đến 25/12/2014 Công ty Cổ phần Chuyển mạch Tài chính Quốc gia Việt Nam (Banknetvn) và Công ty Cổ phần Dịch vụ Thẻ Smartlink (Smartlink)đồng tổ chức Lễ ký kết Hợp đồng sáp nhập dưới sự chứng kiến của Thống đốc Ngân hàng Nhà nước, Ban chỉ đạo tái cấu trúc Banknetvn, Hội đồng Quản trị, Ban Kiểm Soát và Ban Điều Hành của hai công ty cùng đại diện của 4 ngân hàng TMCP Nhà nước và một số ngân hàng TMCP. Việc liên kết và xây dựng một hệ thống liên ngân hàng đã đẩy mạnh doanh số bán hàng, cũng như là quản lý khách hàng tốt hơn, cụ thể:

**Đối với Ngân hàng:**

- Tiết kiệm thời gian và chi phí triển khai do không phải chỉnh sửa về kỹ thuật (Banknetvn sẽ chịu trách nhiệm chỉnh sửa mọi khác biệt giữa Ngân hàng và các tổ chức chuyển mạch nước ngoài).

- Mang lại doanh thu và nguồn ngoại tệ cho Ngân hàng.
- Góp phần quảng bá thương hiệu của Ngân hàng vượt ra khỏi lãnh thổ Việt Nam, vươn tới tầm khu vực và thế giới.
- Gia tăng tiện ích và đáp ứng nhu cầu khách hàng khi công tác, du lịch nước ngoài.
- Mức phí hấp dẫn hơn so với mức phí mà các tổ chức thẻ và thanh toán quốc tế đang áp dụng.

**Đối với khách hàng:**

- Có thể sử dụng thẻ ghi nợ nội địa thuận tiện khi đi công tác, học tập, du lịch ở nước ngoài thay vì thẻ tín dụng quốc tế như trước kia.
- Chi phí thấp hơn so với sử dụng thẻ tín dụng quốc tế.
- Đảm bảo an toàn vì không phải mang một lượng tiền mặt lớn khi ra nước ngoài.
- Thêm một kênh để gửi tiền từ nước ngoài về Việt Nam và ngược lại.
- Khách du lịch và người nước ngoài tại VN cũng thuận tiện và đỡ tốn chi phí, thời gian hơn trong việc sử dụng dịch vụ thẻ tại VN.

## KẾT LUẬN

Khai phá dữ liệu là một công cụ được sử dụng để trích xuất thông tin quan trọng từ dữ liệu hiện có và cho phép ra quyết định kinh doanh tốt hơn trong các ngành công nghiệp ngân hàng và bán lẻ. Họ sử dụng kho dữ liệu để kết hợp các dữ liệu khác nhau từ cơ sở dữ liệu thành một định dạng có thể chấp nhận từ đó có thể tiến hành việc khai phá dữ liệu. Dữ liệu sau đó được phân tích và các thông tin thu được sẽ hỗ trợ cho tổ chức ra quyết định. Kỹ thuật khai phá dữ liệu rất hữu ích cho các ngân hàng thực hiện quá trình kinh doanh tốt hơn, thu hút được khách hàng mới, phát hiện gian lận, cung cấp sản phẩm dựa trên phân tích của các khách hàng để duy trì tốt hơn mối quan hệ của khách hàng. Những ngân hàng này đã nhận ra sự hữu ích của khai phá dữ liệu và đang trong quá trình xây dựng một môi trường khai phá dữ liệu sẽ có được lợi ích to lớn cho quá trình ra quyết định của họ và chiếm được lợi thế cạnh tranh đáng kể trong tương lai.

*Những kết quả đạt được của nghiên cứu:*

- Trình bày khái quát về các kĩ thuật khai phá dữ liệu
- Nêu lên các phương pháp khai phá dữ liệu đặc biệt là phương pháp khai phá dữ liệu dạng lưới, những ứng dụng hiện nay của kĩ thuật này vào hệ thống tài chính
- Trình bày các ứng dụng cụ thể của kĩ thuật khai phá dữ liệu dạng lưới vào hệ thống quản lý ngân hàng hiện nay, nêu lên những tiến bộ vượt bậc của ngành ngân hàng khi áp dụng kĩ thuật công nghệ vào quá trình xây dựng và kinh doanh.

*Bên cạnh những kết quả đạt được, dù đã rất cố gắng nhưng do sự hữu hạn về thời gian và kiến thức, nghiên cứu vẫn còn một số hạn chế:*

Chưa tìm hiểu sâu hơn được cách thức triển khai hệ thống dịch vụ sử dụng kĩ thuật khai phá dữ liệu dạng lưới của từng đơn vị tài chính ngân hàng, lý do liên quan đến vấn đề bảo mật và cạnh tranh công nghệ sử dụng của từng đơn vị kinh doanh đó. Bởi một khi các ngân hàng tham gia vào liên minh do hệ thống Ngân hàng Nhà nước quản lý đều trang bị cho mình một một hệ thống các phân hệ nghiệp vụ cơ bản của ngân hàng như tiền gửi, tiền vay, khách hàng. Thông qua đó, ngân hàng phát triển thêm nhiều dịch vụ, sản phẩm và quản lý nội bộ chặt chẽ, hiệu quả hơn.

## TÀI LIỆU THAM KHẢO


[1] Mario Cannataro, Domenico Talia, Paolo Trunfio, Distributed data mining on the grid, Future Generation Computer Systems 18 (8) (2002) 1101–1112.

[2] Ian Foster, Carl Kesselman, Jeffrey M. Nick, Steve Tuecke, The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration, Tech. Report Globus Project, 2002.

[3] Ian Foster, the anatomy of the Grid: Enabling scalable virtual organizations, Lecture Notes in Computer Science 2150 (2001).

[4] Alberto S´anchez, Jos´e M. Pe˜na S´anchez, Mar´ıa S. P´erez, Victor Robles, Pilar Herrero, improving distributed data mining techniques by means of a grid infrastructure, in: OTM Workshops, in: LNCS, vol. 3292, 2004, pp. 111–122.

[5] William Allcock, Joe Bester, John Bresnahan, Ann Chervenak, Lee Liming, Steve Tuecke, GridFTP: Protocol extensions to FTP for the Grid, Global Grid Forum Draft, 2001.

[6] Giovanni Aloisio, Massimo Cafaro, Italo Epicoco, Early experiences with the gridftp protocol using the grb-gsiftp library, Future Generation Computer Systems 18 (8) (2002) 1053–1059.

[7] Ian H. Witten, Eibe Frank, Data Mining: Practical machine learning tools with Java implementations, Morgan Kaufmann, San Francisco, 2000.

[8] Agrawal R. & Srikant, R. (1994, September). Fast Algorithms for Mining Association Rules. In Proceedings of the 20th International Conference on Very Large Databases (VLDB'94), Santiago, Chile, 487-499.

[9] Bauer, E., & Kohavi, R. (1999). An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. Machine Learning, 36 (1-2), 105-139.

[10] Dietterich, 2000; Opitz & Maclin, 1999; Bauer & Kohavi, 1999; Merz & Pazzani, 1999