

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
ĐẠI HỌC ĐÀ NẴNG**

ĐỖ QUANG BẢN

**ỨNG DỤNG KHAI PHÁ DỮ LIỆU XÂY DỰNG
HỆ THỐNG TRỢ GIÚP KINH DOANH DU LỊCH**

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại
ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. PHAN HUY KHÁNH**

Phản biện 1: **TS. NGUYỄN THANH BÌNH**

Phản biện 2: **NGUYỄN MẠU HÂN**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 15 tháng 10 năm 2011

** Có thể tìm hiểu luận văn tại:*

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng
- Trung tâm Học liệu, Đại học Đà Nẵng.

MỞ ĐẦU

1. Lý do chọn đề tài

Khai phá tri thức trong Cơ sở dữ liệu (CSDL - Knowledge Discovery in Databases) đang là một xu hướng quan trọng của nền Công nghệ thông tin (CNTT) thế giới. Nó có khả năng ứng dụng vào rất nhiều lớp bài toán thực tế khác nhau. Bước quan trọng nhất của quá trình này là khai phá dữ liệu, giúp người sử dụng thu được những tri thức hữu ích từ những CSDL hoặc các nguồn dữ liệu khổng lồ khác. Rất nhiều doanh nghiệp và tổ chức trên thế giới đã ứng dụng kỹ thuật khai phá dữ liệu vào hoạt động sản xuất kinh doanh của mình và đã thu được những lợi ích to lớn. Trong lĩnh vực du lịch, áp dụng khai phá dữ liệu để phân tích, dự đoán trong kinh doanh du lịch, nếu có được những phân tích dự đoán về thông tin khách hàng thì sẽ có những chiến lược tốt nhất để đầu tư cơ sở vật chất, các dịch vụ đáp ứng được nhu cầu của khách hàng, vì Du lịch là ngành có số lượng khách hàng rất lớn và có lượng dữ liệu cần lưu trữ khổng lồ. Bên cạnh đó, thị trường Du lịch luôn luôn biến động và đầy cạnh tranh. Xuất phát từ lý do đó tôi đã thực hiện đề tài:

"Ứng dụng khai phá dữ liệu xây dựng hệ thống trợ giúp kinh doanh du lịch" để làm luận văn tốt nghiệp thạc sĩ ngành công nghệ thông tin.

Công Ty Cổ Phần Du Lịch Quảng Ngãi là một đơn vị kinh doanh trong lĩnh vực dịch vụ du lịch. Chính vì vậy việc nghiên cứu và áp dụng kỹ thuật khai phá dữ liệu vào định hướng phát triển là cần thiết.

2. Mục đích và ý nghĩa của đề tài

Nếu biết trước các thông tin về tình hình sử dụng các loại hình dịch vụ du lịch của khách hàng sẽ giúp Công ty giữ được các khách hàng cũ và phát triển nhiều khách hàng mới cũng như mở rộng đầu tư thêm các loại hình dịch vụ du lịch khác giúp cạnh tranh hiệu quả trong hiện tại và tương

lai. Do đó vấn đề ứng dụng kỹ thuật khai phá dữ liệu các thông tin về tình hình sử dụng các loại hình dịch vụ của khách hàng là rất quan trọng và đây cũng chính là mục đích chính của luận văn này.

Hệ thống sau khi xây dựng có các chức năng chính sau:

Phân tích và dự đoán tình hình sử dụng các loại hình dịch vụ du lịch của khách theo các tiêu chí như quốc tịch, độ tuổi, giới tính, khách đoàn, khách lẻ, các dịch vụ khách sử dụng... Để từ đó có những chiến lược quản bá khai phá các loại hình dịch vụ du lịch hoặc đầu tư thêm các loại hình dịch vụ du lịch: Hướng giải quyết nhiệm vụ này là sử dụng phương pháp khai phá luật kết hợp, từ dữ liệu đã có chúng ta sẽ xây dựng nên một kho dữ liệu, áp dụng các thuật toán khai phá luật kết hợp để chiết xuất ra những tập luật có ích. Từ kết quả đó sẽ có chiến lược đầu tư thêm hay mở rộng các loại hình dịch vụ du lịch hay cần nhắm tới những thị trường mục tiêu nào để khai phá...

Vì vậy, quá trình xây dựng mô hình sẽ diễn ra trong thời gian khá lâu. Mặt khác, chu kỳ cập nhật của dữ liệu là khá dài. Để giảm thời gian chờ đợi của người sử dụng, chúng ta sẽ tách riêng bước xây dựng mô hình và bước áp dụng mô hình. Mô hình sau khi xây dựng sẽ được lưu lại để sử dụng khi người dùng muốn dự đoán số liệu mới. Mô hình sẽ được định kỳ xây dựng lại hoặc được xây dựng lại khi người dùng thấy hệ thống dự đoán không còn chính xác nữa.

* Ý nghĩa khoa học và thực tiễn của đề tài

Hòa nhập theo sự phát triển của ngành du lịch ở địa phương và trong cả nước, hiện nay hầu hết các Công ty du lịch trên cả nước phần lớn chưa xây dựng và triển khai kỹ thuật khai phá dữ liệu trong kinh doanh phát triển du lịch, các Công ty du lịch đều sử dụng phần mềm về quản lý du lịch không có chức năng trợ giúp đưa ra những chiến lược kinh doanh cũng như ra quyết định hoặc nếu có đi chăng nữa thì chỉ phân tích ở mức

độ thủ công độ chính xác chưa cao.

Vì lẽ đó đề tài sẽ xây dựng những chức năng, những phân hệ cơ bản mục đích trợ giúp chiến lược kinh doanh phát triển du lịch cho Công ty góp phần khai phá tốt các loại hình dịch vụ du lịch nhằm mục đích cuối cùng là tăng doanh thu, tăng khả năng cạnh tranh của Công ty.

3. Mục tiêu và nhiệm vụ

- Tìm hiểu tri thức chuyên gia trong lĩnh vực hoạch định chiến lược kinh doanh phát triển du lịch, đặt biệt là vấn đề ra quyết định đầu tư các loại hình dịch vụ du lịch tại Công ty hay khai thác các thị trường khác tiềm năng.

- Thu thập và phân tích dữ liệu về tình hình sử dụng các loại hình dịch vụ du lịch của khách lưu trú tại công ty.

- Xây dựng kho dữ liệu về tình hình sử dụng các loại hình dịch vụ du lịch của khách hàng

- Xây dựng thuật toán chuyên cơ sở dữ liệu quan hệ sang cơ sở dữ liệu giao dịch

- Tìm hiểu cơ sở lý thuyết khai phá luật kết hợp

- Xây dựng mô hình và ứng dụng các kỹ thuật của khai phá dữ liệu như: Khai phá luật kết hợp.

4. Bố cục của luận văn

Bố cục của luận văn như sau:

Mở đầu: Trình bày lý do chọn đề tài, mục đích và ý nghĩa của đề tài và mục tiêu nhiệm vụ.

Chương 1: Tổng quan kỹ thuật khai phá dữ liệu và trình bày vấn đề lý thuyết về luật kết hợp, các phương pháp và chương này tập trung trình bày khái quát về khai phá dữ liệu đặt biệt là các phương pháp khai phá dữ liệu cũng như các lĩnh vực ứng dụng.

Chương 2: Khảo sát và phân tích bài toán thực tế, trong chương này

tập trung giới thiệu bài toán, nêu lên bài toán cụ thể và xây dựng mô hình tổng quát cho bài toán, giải pháp giải quyết bài toán.

Chương 3: Xây dựng hệ thống trợ giúp chiến lược kinh doanh phát triển các loại hình dịch vụ du lịch tại Công Ty Cổ phần Du lịch Quảng Ngãi và các kết quả đạt được. Trong chương này tập trung xây dựng kho dữ liệu, xây dựng chương trình, chạy thử và đánh giá kết quả.

Kết Luận: Kết quả đạt được và đề xuất hướng nghiên cứu tiếp theo.

CHƯƠNG 1: TỔNG QUAN VỀ KHAI PHÁ DỮ LIỆU

1.1 Một số kiến thức về(Khai phá dữ liệu

Khai phá dữ liệu (Data mining) 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, khai phá dữ liệu 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.

*** Mục tiêu của khai phá dữ liệu:**

Cuộc cách mạng của kỹ thuật số cho phép số hoá thông tin dễ dàng và chi phí lưu trữ thấp. Với sự phát triển của phần mềm và phần cứng máy tính và trang bị nhanh hệ thống máy tính trong kinh doanh.

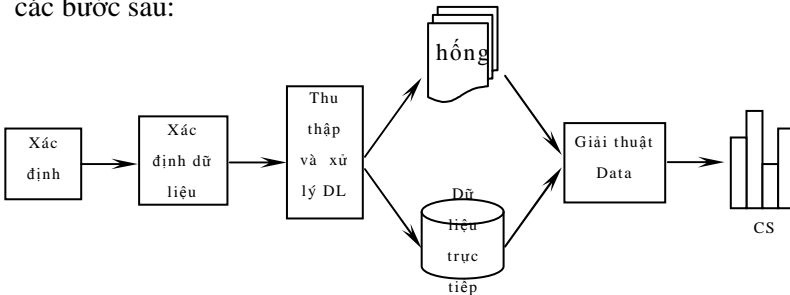
Số lượng khổng lồ của dữ liệu được tập trung và lưu trữ trong cơ sở dữ liệu trên các thiết bị điện tử như: Đĩa cứng, băng từ, đĩa quang, CD-ROM,.... Tốc độ tăng dữ liệu quá lớn. Từ đó dẫn đến kết quả là sự pha trộn của kỹ thuật thống kê và các công cụ quản trị dữ liệu không thể phân tích đầy đủ dữ liệu rộng lớn được nữa.

Dữ liệu của chúng ta sau khi xử lý trực tuyến phục vụ cho một mục đích nào đó được lưu lại trong kho dữ liệu và theo ngày tháng khối lượng

dữ liệu được lưu trữ ngày càng lớn. Trong khối lượng dữ liệu to lớn này còn rất nhiều thông tin có ích mang tính tổng quát, thông tin có tính qui luật vẫn đang còn tiềm ẩn mà chúng ta chưa biết. Các công cụ xử lý phân tích trực tuyến (On-Line Analytical Processing - OLAP) là cần thiết để phân tích dữ liệu, nhưng chưa đủ để rút thông tin từ một khối lượng dữ liệu khổng lồ như vậy. Từ khối lượng dữ liệu rất lớn thì cần phải có những công cụ tự động rút các thông tin và kiến thức có ích. Một hướng tiếp cận mới có khả năng giúp các công ty khai thác các thông tin có nhiều ý nghĩa từ các tập dữ liệu lớn (databases, data warehouses, data repositories) đó là khai phá dữ liệu (Data Mining).

1.1.1 Quá trình khai phá dữ liệu

Có Thể mô hình hóa các bước trong tiến trình khai phá dữ liệu theo các bước sau:



Hình 1.1: Sơ đồ mô tả quá trình khai phá dữ liệu

1.1.2 Kiểu Dữ liệu khai phá

Khai phá dữ liệu có khả năng chấp nhận một số kiểu dữ liệu khác nhau điển hình như sau:

- Cơ sở dữ liệu quan hệ (relational databases):
- Cơ sở dữ liệu đa chiều (multidimension structures, data warehouses, data mart):
- Cơ sở dữ liệu giao tác (transactional databases):

- Cơ sở dữ liệu quan hệ - hướng đối tượng (object relational databases):

1.2 Phương pháp khai phá dữ liệu:

Quá trình khai phá dữ liệu là quá trình phát hiện mẫu trong đó giải thuật khai phá dữ liệu tìm kiếm các mẫu đáng quan tâm theo dạng xác định như các luật, cây phân lớp, hồi quy, phân nhóm,...

1.3 Cơ sở lý thuyết khai phá luật kết hợp

1.3.1 Khai phá luật kết hợp

Trong hoạt động sản xuất kinh doanh, ví dụ: Kinh doanh các loại hình dịch vụ du lịch, các nhà quản lý rất thích có được các thông tin mang tính thống kê như: “90% khách hàng là nữ tuổi từ 25-50 và quốc tịch Pháp lưu trú tại khách sạn thì sử dụng dịch vụ spa, 70% khách hàng là nam tuổi từ 20-35 quốc tịch Mỹ lưu trú tại khách sạn thì sử dụng dịch vụ vui chơi trên biển hay 50% khách lưu trú theo đoàn thì đặt tour đi Sa Huỳnh . Những thông tin như vậy rất hữu ích trong việc định hướng kinh doanh. Vậy vấn đề đặt ra là liệu có tìm được các luật như vậy bằng các công cụ khai phá dữ liệu hay không? Đó chính là nhiệm vụ khai phá luật kết hợp.

1.3.2 Cơ sở dữ liệu giao dịch

1.3.2.1 Khái niệm

1.3.2.2 Giải thuật chuyển đổi cơ sở dữ liệu

Để đơn giản hơn cho các giải thuật khai phá luật kết hợp chúng ta có thể xây dựng giải thuật cho phép chuyển đổi từ một cơ sở dữ liệu dạng quan hệ truyền thống sang cơ sở dữ liệu giao dịch để trợ giúp cho quá trình khai phá dữ liệu tình hình sử dụng các loại hình dịch vụ bằng luật kết hợp. Giải thuật tựa ngôn ngữ được minh họa như sau:

Giải thuật 3-1: ConvertDbRelationToDbTransaction

Begin

Input: D – data set, là cơ sở dữ liệu quan hệ mẫu

Output: D* - data set, là cơ sở dữ liệu giao dịch

foreach (record R into D)

begin

T = { } - là chuẩn dữ liệu giao dịch

foreach (item I into R)

T = T + {I}

D* = D* + {T} – đưa chuẩn dữ liệu giao dịch vào D*

end

End

1.3.3 Một số hướng tiếp cận trong khai phá luật kết hợp

Lĩnh vực khai phá luật kết hợp cho đến nay đã được nghiên cứu và phát triển theo nhiều hướng khác nhau. Có những đề xuất nhằm cải tiến tốc độ thuật toán, có những đề xuất nhằm tìm kiếm luật có ý nghĩa hơn... và có một số hướng chính như sau.

- Luật kết hợp nhị phân
- Luật kết hợp có thuộc tính số và thuộc tính hạng mục:
- Luật kết hợp tiếp cận theo hướng tập thô:
- Luật kết hợp nhiều mức:
- Luật kết hợp mờ:
- Luật kết hợp với thuộc tính được đánh trọng số:
- Luật kết hợp song song:

Ngoài ra, còn có một số hướng nghiên cứu khác về khai phá luật kết hợp như: Khai phá luật kết hợp trực tuyến, khai phá luật kết hợp được kết nối trực tuyến đến các kho dữ liệu đa chiều thông qua công nghệ OLAP, MOLAP, ROLAP, ADO.

1.3.4 Luật kết hợp

1.3.4.1 Khái niệm

1.3.4.2 Một số tính chất liên quan đến các hạng mục phổ biến

Với tập mục phổ biến, có 3 tính chất sau:

Tính chất 1 (Độ hỗ trợ của tập con):

Tính chất 2: Một tập chứa một tập không phổ biến thì cũng là tập không phổ biến.

Tính chất 3: Các tập con của tập phổ biến cũng là tập phổ biến

1.3.5 Phát hiện luật kết hợp trên hệ thông tin nhị phân

1.3.5.1 Các định nghĩa về hệ thông tin nhị phân

1.3.5.2 Thuật toán phát hiện tập chỉ mục và luật kết hợp nhị phân

1.3.6 Một số thuật toán phát hiện luật kết hợp

1.3.6.1 Thuật toán Apriori

Ý tưởng thuật toán Apriori

Apriori là một thuật giải được do Rakesh Agrawal, Tomasz Imielinski, Arun Swami đề xuất lần đầu vào năm 1993. Thuật toán tìm giao dịch t có độ hỗ trợ và độ tin cậy thoả mãn lớn hơn một giá trị ngưỡng nào đó.

Thuật toán được chia bớt những tập ứng cử viên có tập con không phổ biến trước khi tính độ hỗ trợ.

Thuật toán Apriori tính tất cả các tập ứng cử của tập k trong một lần duyệt CSDL. Apriori dựa vào cấu trúc cây băm. Tìm kiếm đi xuống trên cấu trúc cây mỗi khi ta chạm lá, ta tìm được một tập ứng cử viên có tiền tố chung được bao gồm trong giao dịch. Sau đó các tập ứng cử này được tìm

trong giao dịch đã được ánh xạ trước đó. Trong trường hợp tìm thấy biên đếm được tăng lên 1.

Ký hiệu: Giả sử các mục trong mỗi giao dịch được lưu giữ theo trật tự từ điển. Gọi số các mục trong một tập mục là kích thước của nó và gọi tập mục có kích thước k là tập k -mục (tập k mục). Các mục trong mỗi tập mục cũng được giữ ở trật tự từ điển. Ta sử dụng các ký hiệu sau:

L_k: Tập các tập k -mục phổ biến (với độ hỗ trợ cực tiểu minsup nào đó)

C_k : Tập các tập k -mục ứng cử (các tập mục phổ biến tiềm năng)

Thuật toán Apriori

Input: CSDL D , minsup.

Output: Tập các tập mục phổ biến.

1. $L_1 = \{\text{Các 1 - itemset phổ biến}\};$
2. $k=2;$
3. While($L_{k-1} \neq \emptyset$)
4. { $C_k = \text{apriori_gen}(L_{k-1}, \text{minsup});$ // các ứng cử mới theo chương trình con ở dưới đây.
5. for(\forall giao dịch $t \in D$)
6. { $C_t = \text{Subset}(C_k, t);$ // ứng cử viên được chứa trong t
7. for (\forall ứng cử $c \in C_t$)
8. $c.\text{count} ++;$
10. }
11. $L_k = \{ c \in C_k \mid c.\text{count} \geq \text{minsup} \}$
12. $k++;$
13. }
14. Return $L = \cup_k L_k$;

```

// sinh ứng cử viên mới (**)
Void apriori_gen(Lk-1, minsup )
1. {      for ( $\forall$  itemset l1 $\in$  Lk-1)
2.      for ( $\forall$  itemset l2 $\in$  Lk-1)
3.      if((L1(1)== L2(1)&&L1(2) == L2(2)&&...&& L1(k-2) == L2(k-
2)) &&L1(k-1) == L2(k-1))
4.          {      c= L1 kết nối L2;
5.          if( has_infrequent_subset(c, Lk-1)) delete
c;
6.          else add c to Ck;
7.          }
8.      return Ck
9.}

Boolean has_infrequent_subset(c,Lk-1)
1.{      for ( $\forall$  (k-1)-subset s $\in$  c)
2.      if(s  $\notin$  Lk-1) return TRUE;
3.      else return FALSE ;
4.}

```

Giải thích: Lần duyệt đầu tiên, sẽ tính số lần xuất hiện của mỗi mục để xác định các 1- itemset phổ biến. Lần duyệt thứ k ($k \geq 2$) sẽ bao gồm 2 giai đoạn:

Tập phổ biến Lk-1 đã tìm thấy ở lần duyệt thứ k-1 được sử dụng để sinh ra các tập ứng cử viên Ck bằng việc sử dụng hàm Apriori_gen.

Dựa vào CSDL, tính độ hỗ trợ của các ứng cử viên trong Ck. Các ứng cử viên trong Ck mà được chứa trong giao dịch t có thể được xác định một cách hiệu quả bằng việc sử dụng cây băm được mô tả như sau:

Trong giai đoạn 2 (giai đoạn sửa, tĩa): xóa bỏ các tập $c \in C_k$ sao cho một vài $(k-1)$ – tập con của c không nằm trong L_{k-1} . Thủ tục này là đầy đủ bởi đối với bất kì tập nào L_k với độ hỗ trợ tối thiểu thì các tập con kích cỡ $(k-1)$ cũng có độ hỗ trợ tối thiểu, do đó nếu ta mở rộng mỗi tập trong L_{k-1} với tất cả các tập mục có thể và sau đó xóa tất cả các tập mà $(k-1)$ – tập con của nó không nằm trong L_{k-1} , ta sẽ nhận được tập các tập trong L_k .

Việc kết nối là tương đương với việc mở rộng L_{k-1} với mỗi mục nằm trong CSDL và sau đó xóa bỏ các tập này mà đối với nó $(k-1)$ – itemset nhận được bằng việc xóa đi mục thứ $(k-1)$ không nằm trong L_{k-1} . Ở giai đoạn này $C_k \supseteq L_k$. Với lập luận như vậy, giai đoạn tĩa là giai đoạn người ta xóa khỏi C_k tất cả các tập mà các $(k-1)$ tập con của nó không nằm trong L_{k-1} , cũng không xóa bất kỳ một tập nào có thể nằm trong L_k .

Nhận xét: Thuật toán Apriori với n là độ dài lớn nhất của tập được sinh ra. Vậy thì thuật toán sẽ thực hiện duyệt toàn bộ các giao tác $n+1$ lần. Như vậy, nếu bỏ qua thời gian so sánh tìm sự xuất hiện của một mẫu trong một giao tác thì độ phức tạp của thuật toán Apriori là $O(A) > O(n \cdot L)$ trong đó L là kích thước CSDL còn n là độ dài cần đạt được của các mẫu.

Ngoài ra, nếu độ hỗ trợ tối thiểu minsup bị thay đổi thì thuật toán sẽ phải thực hiện lại từ đầu, điều này sẽ rất mất thời gian. Thuật toán Apriori được xây dựng nhằm phát hiện các luật kết hợp giữa các đối tượng với độ hỗ trợ và độ tin cậy tối thiểu.

* Sinh các luật kết hợp từ tập mục phổ biến

1.4 Ứng dụng của khai phá dữ liệu

- Thông tin thương mại:

+ Phân tích dữ liệu marketing, khách hàng

+ Phân tích đầu tư

+ Phê duyệt cho vay vốn

- + Phát hiện gian lận
- Thông tin kỹ thuật
- + Điều khiển và lập lịch trình
- + Quản trị mạng
- + Phân tích các kết quả thí nghiệm..
- Thông tin khoa học
- Thông tin cá nhân...

1.5 Phân loại dữ liệu khai phá

Khai phá dữ liệu dựa trên các tiêu chí khác nhau.

Phân loại dựa trên kiểu dữ liệu được khai phá: Cơ sở dữ liệu quan hệ, kho dữ liệu, cơ sở dữ liệu giao tác, cơ sở dữ liệu hướng đối tượng, cơ sở dữ liệu không gian, cơ sở dữ liệu đa phương tiện, cơ sở dữ liệu văn bản và www.v.v.

Phân loại dựa trên dạng tri thức được khám phá.

Phân loại dựa trên lĩnh vực được áp dụng: Thương mại, viễn thông, tài chính, y học, web mining, v. v.

Phân loại dựa trên kỹ thuật được áp dụng:

1.6 Những vấn đề quan tâm trong hệ thống khai phá dữ liệu

1.7 Kết luận

Như vậy, nhìn vào các phương pháp giới thiệu ở trên, chúng ta thấy có rất nhiều các phương pháp khai phá dữ liệu. Mỗi phương pháp có những đặc điểm riêng phù hợp với một lớp các bài toán với các dạng dữ liệu và miền dữ liệu nhất định. Giả sử đối với bài toán dự đoán theo thời gian, trước kia người ta thường đặt nhiệm vụ cho việc khai phá các mẫu dạng này là hồi quy dự đoán hoặc các mô hình hồi quy tự động dựa trên

thống kê,... Mới đây, các mô hình khác như các hàm phi tuyến, phương pháp dựa trên mẫu, mạng neuron đã được áp dụng để giải loại bài toán này. Mặc dù nhìn bề ngoài ta thấy có rất nhiều các phương pháp và ứng dụng khai phá dữ liệu nhưng cũng không có gì là lạ khi nhận thấy chúng có một số thành phần chung. Hiểu quá trình khai phá dữ liệu và suy diễn được mô hình dựa trên những thành phần này là ta đã thực hiện được nhiệm vụ của khai phá dữ liệu.

Trong quá trình thực hiện luận văn tôi có tham khảo một số tài liệu liên quan nhưng vì trình độ và thời gian còn hạn chế nên nội dung của luận văn này liên quan chủ yếu đến hướng chính là khai phá luật kết hợp để áp dụng khai phá dữ liệu khách hàng về tình hình sử dụng các loại hình dịch vụ du lịch tại Công Ty Cổ Phần Du Lịch Quảng Ngãi.

CHƯƠNG 2 PHÂN TÍCH BÀI TOÁN

2.1 Giới thiệu Công Ty Cổ Phần Du Lịch Quảng Ngãi

Công Ty Cổ Phần Du Lịch Quảng Ngãi có chức năng kinh doanh các loại hình dịch vụ du lịch, Công ty có một trung tâm lữ hành gồm 2 khách sạn 3 sao với hơn 100 phòng và Khu Du lịch Sa Huỳnh quy mô giai đoạn 1 là 52 ha, cùng với các loại hình dịch vụ du lịch rất phong phú và đa dạng, đóng trên địa bàn thành phố Quảng Ngãi, trong những năm gần đây tình hình phát triển du lịch trên địa bàn có nhiều khởi sắc, số lượng khách quốc tế và trong nước đến với Quảng Ngãi ngày càng tăng, theo đó rất nhiều khách sạn nhà hàng được xây dựng lên cùng với các loại hình dịch vụ ngày càng phong phú đa dạng và chất lượng cũng ngày càng đòi hỏi cao hơn để đáp ứng nhu cầu phục vụ du khách. Vấn đề cạnh tranh cũng ngày càng gay gắt và quyết liệt hơn giữa các Công ty du lịch trên địa bàn.

Làm thế nào để mang lại hiệu quả kinh doanh tốt, cần biết Công ty mình đang đứng ở đâu? Công ty mình đang có những loại hình dịch vụ nào? Cần mở rộng phát huy những thế mạnh nào, đầu tư thêm những loại hình dịch vụ nào trong tương lai, để có một quyết định đầu tư mang tính chiến lược góp phần tăng doanh thu cho Công ty.

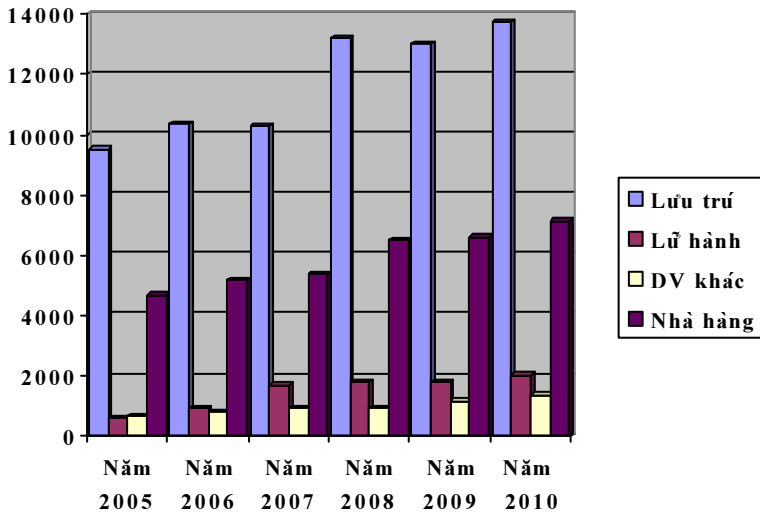
Điều này đòi hỏi Lãnh đạo Công ty phải có một khối lượng tri thức nhất định về tình hình kinh doanh các loại hình dịch vụ du lịch trong quá khứ và hiện tại của Công ty mình đang hoạt động để phân tích và đưa ra những quyết định đúng mang tính chiến lược trong tương lai để đạt mục đích làm tăng doanh thu.

2.2 Hoạt động kinh doanh:

Bảng kế hoạch kinh doanh hằng năm Công ty đề ra từ 2005-2010 (ĐVT: Triệu đồng)

Bảng 2.1 Kế hoạch kinh doanh đặt ra qua các năm từ năm 2005-2010

	Năm 2005	Năm 2006	Năm 2007	Năm 2008	Năm 2009	Năm 2010
Lưu trú	8.350	9.250	9.550	11.800	11.850	11.950
Lữ hành	535	820	1.530	1.660	1.680	1.750
DV khác	596	702	780	918	1.044	1.200
Nhà hàng	4.170	4.630	4.790	5.900	6.000	6.200

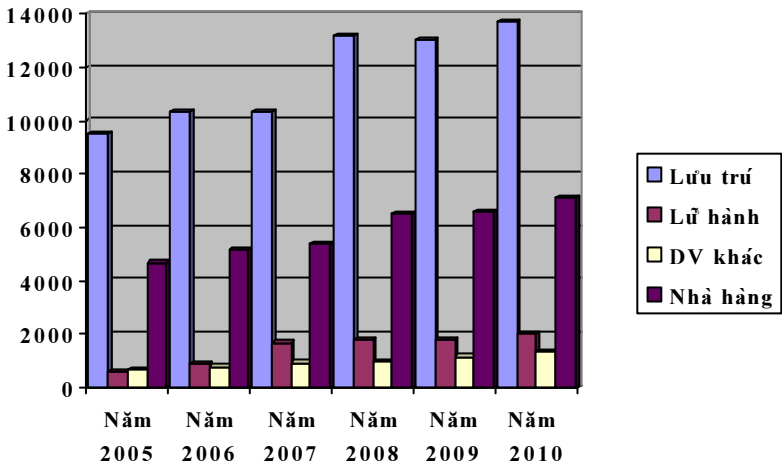


Hình 2.1: Biểu đồ kế hoạch doanh thu

*Bảng kế hoạch kinh doanh đã thực hiện từ 2005-2010 (ĐVT:Triệu đồng)

Bảng 2.2 Kế hoạch kinh doanh đã thực hiện từ năm 2005-2010

	Năm 2005	Năm 2006	Năm 2007	Năm 2008	Năm 2009	Năm 2010
Lưu trữ	9.519	10.360	10.314	13.216	13.035	13.740
Lữ hành	598	902	1.680	1.780	1.780	2.012
DV khác	667	770	923	934	1.145	1.347
Nhà hàng	4.670	5.185	5.364	6.490	6.600	7.130



Hình 2.2: Biểu đồ doanh thu đã thực hiện qua các năm từ năm 2005- 2010

DV khác: spa, vui chơi trên biển, tổng đài taxi, tour, internet...

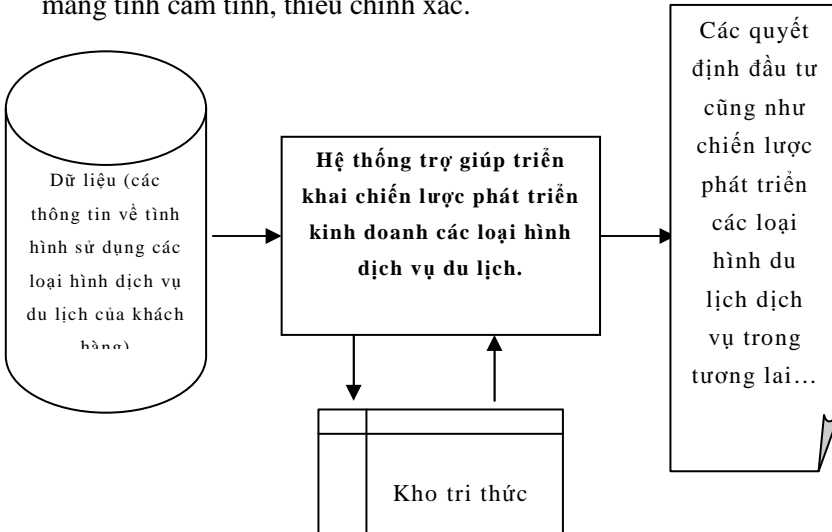
Nhìn vào số liệu trên ta thấy tình hình tăng trưởng trong kinh doanh phát triển du lịch qua các năm tại công ty từ 2005 đến 2010. Mặt khác ta thấy ở biểu đồ tăng trưởng từ năm 2008- năm 2010 tỉ lệ tăng khá cao về doanh thu, ở thời điểm đó do Quảng Ngãi có Khu Kinh tế Dung Quất phát triển rất mạnh đặt biệt đưa vào vận hành Nhà máy Lọc dầu số 1 Việt Nam, trong đó số lượng khách sạn ở Quảng Ngãi chưa nhiều, tính cạnh tranh giữa các khách sạn chưa cao nên cũng có phần lợi thế về nguồn khách.

Để xây dựng được kế hoạch cũng như chiến lược kinh doanh phòng kế hoạch cần phải phân tích tình hình cụ thể cho từng mảng dịch vụ sự biến động của thị trường...

Trong khai phá các loại hình dịch vụ ta chú trọng đến mấy vấn đề (tiêu chí) sau:

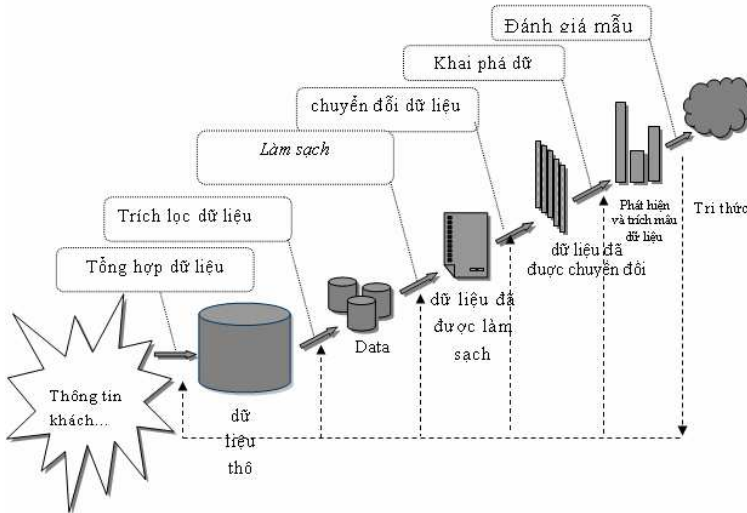
Theo quốc tịch, độ tuổi, giới tính, sở thích, khách theo đoàn, Công ty
lữ hành, loại dịch vụ...

Trong hiện tại vấn đề lập kế hoạch kinh doanh và chiến lược để thực
hiện kế hoạch được phòng kế hoạch lập ra theo phương pháp thủ công dựa
theo phương pháp thống kê là chủ yếu mất khá nhiều thời gian, đôi khi
mang tính cảm tính, thiếu chính xác.



Hình 2.3: Mô hình hệ thống trợ giúp chiến lược kinh doanh

2.3 Giải pháp ứng dụng khai phá dữ liệu



Hình 2.4 Mô hình xây dựng hệ thống

2.3.1 Gom dữ liệu

2.3.2 Trích lọc dữ liệu

2.3.3 Làm sạch, tiền xử lý và chuẩn bị trước dữ liệu

2.3.4 Chuyển đổi dữ liệu

2.3.5 Phát hiện và trích mẫu dữ liệu

2.3.6 Đánh giá kết quả mẫu

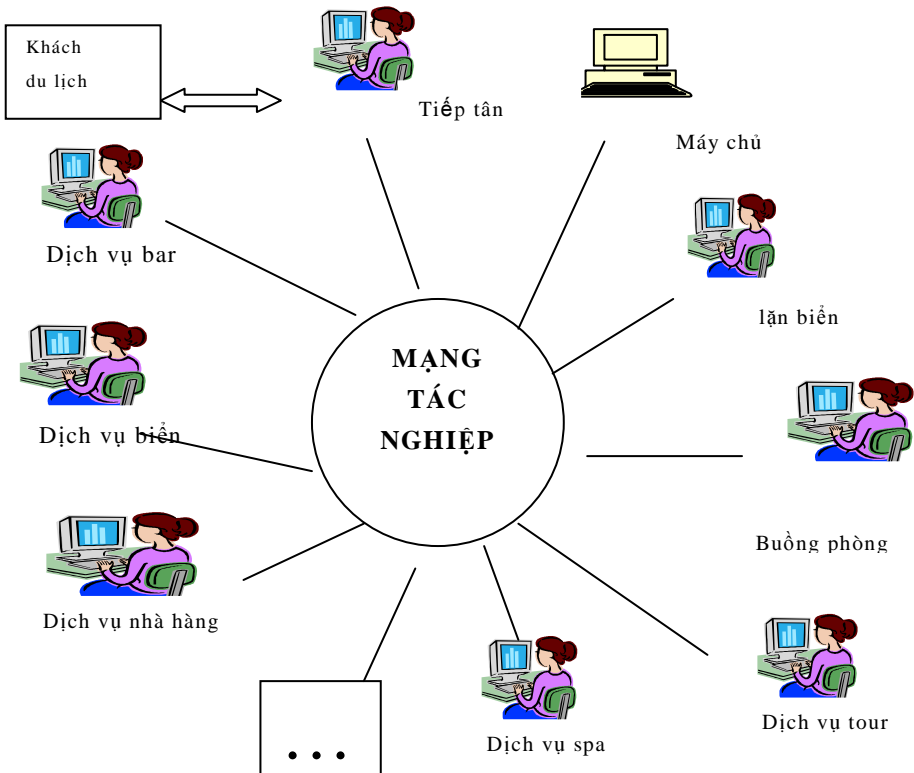
2.4 Tóm tắt chương 2

Trong chương này giới thiệu tổng quan về Công Ty Cổ Phần Du Lịch Quảng Ngãi, nêu lên bài toán cụ thể, cách giải quyết theo cách truyền thống, chỉ ra vấn đề cần giải quyết của hệ thống mới, lập ra mô hình bài toán thực tế, phân tích bài toán và đưa ra giải pháp qui trình giải quyết bài toán.

CHƯƠNG 3 XÂY DỰNG HỆ THỐNG TRỢ GIÚP KINH DOANH DU LỊCH

3.1 Nguồn dữ liệu về tình hình sử dụng các loại hình dịch vụ tại khách sạn

Khi nhận yêu cầu đặt phòng đến khi bộ phận lễ tân checkin và checkout khách thông tin của khách được lưu trong cơ sở dữ liệu hệ chương trình quản lý khách hiện đang sử dụng tại khách sạn:



Hình 3.1: Hệ thống quản lý khách lưu trú tại khách sạn

Tất cả những thông tin về khách du lịch khi vào khách sạn được lưu trữ trong hệ thống tập tin này trong hệ chương trình quản lý khách hiện đang hoạt động tại khách sạn, mỗi tập tin chứa thông tin của khách lưu trú theo từng tháng.

Tất cả những thông tin về khách du lịch khi ra khỏi khách sạn được lưu, được trữ trong hệ thống tập tin sau trong hệ chương trình quản lý khách hiện đang hoạt động tại khách sạn, mỗi tập tin chứa thông tin của khách lưu trú theo từng tháng về tình hình sử dụng các loại hình dịch vụ du lịch khi lưu trú tại khách sạn. Căn cứ vào những thông tin này bộ phận thu ngân tại quầy tiếp tân sẽ lên hoá đơn thanh toán cho khách”

Tuy nhiên tất cả những thông tin này phải được làm sạch loại bỏ những thông tin dư thừa cho việc khai phá dữ liệu, dữ liệu sẽ được tổ chức lại và khai phá theo quý và theo năm từ đó áp dụng thuật toán khai phá luật kết hợp để đưa ra những luật và chọn các luật tốt nhất để trợ giúp cho việc xây dựng chiến lược kinh doanh các loại hình dịch vụ du lịch.

Những thông tin sau là cần thiết cho việc khai phá dữ liệu

MS:Mã khách, TEN:Tên khách, TUOI: Tuổi khách GT: Giới tính, NDEN: Ngày đến, NĐI: Ngày đi, DICHVU:Các dịch vụ đã sử dụng trong thời gian lưu trú

Các dịch vụ như:

TOUR: Đặt tour đi Lý Sơn, Cù Lao Chàm, Sa Huỳnh...

NH: Ăn tại nhà hàng, Bar: uống tại quầy bar, DVBIEN: Các dịch vụ vui chơi giải trí trên biển, SPA: Các dịch vụ thư giãn chăm sóc sắc đẹp, các dịch vụ khác...

Bộ phận tiếp tân:

Tiếp nhận khách vào và cho khách nhận phòng, thông tin về khách được lưu trong cơ sở dữ liệu của hệ chương trình quản lý khách sạn:

Bộ phận thu ngân tại các điểm dịch vụ:

Nhập vào cơ sở dữ liệu về tình hình sử dụng các dịch vụ của khách khi khách sử dụng, tất cả các thông tin được lưu trữ trên server

Bộ phận thu ngân tại quầy tiếp tân:

Khi khách checkout bộ phận này có nhiệm vụ kiểm tra tất cả các dịch vụ khách đã dùng trong thời gian lưu trú tại khách sạn trong cơ sở dữ liệu Dichvu và lên hoá đơn thanh toán.

Với cách quản lý khách lưu trú tại khách sạn như trên, khách sạn hoàn toàn có các tổng kết đánh giá về các thông tin như: Tình hình sử dụng các loại hình dịch vụ (theo quốc tịch, độ tuổi, giới tính...). Từ đó có thể có các chiến lược kinh doanh cho phù hợp.

Bài toán phát hiện luật kết hợp trong kinh doanh các loại hình dịch vụ với số lượng khách du lịch lưu trú tại khách sạn ngày càng tăng, bình quân mỗi năm hơn 80% công suất phòng, dữ liệu cần xử lý tương đối lớn.

Từ các dữ liệu thu được, các nhà quản lý rất cần có các thông tin thống kê, dự đoán sau:

-Tình hình sử dụng dịch vụ theo quốc tịch của khách lưu trú, ví dụ: Khách Mỹ thích dùng loại dịch vụ nào...

-Tình hình sử dụng dịch vụ theo độ tuổi, ví dụ: 70% khách từ 25-50 tuổi là quốc tịch Nhật thì sử dụng dịch vụ spa.

-Tình hình khách lưu trú theo mùa, ví dụ: Vào mùa nắng khách nội địa tăng cao, vào mùa mưa khách quốc tế tăng cao...

Sự kết hợp của các dịch vụ khác nhau trong cùng một lần lưu trú tại khách sạn v.v...

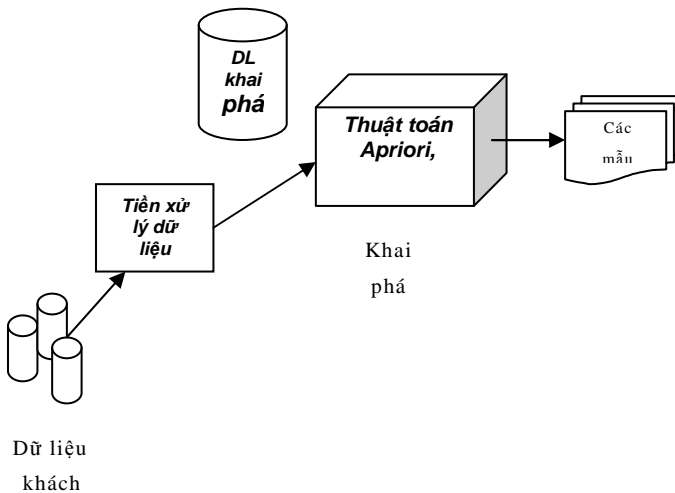
Trên cơ sở đó, các nhà quản lý khách sạn sẽ có phương án, kế hoạch như:

- Đẩy mạnh đầu tư mở rộng nâng cấp các dịch vụ có xu hướng phát triển mạnh

- Lên kế hoạch làm việc với các hãng lữ hành quốc tế tranh thủ nguồn khách của các quốc gia có xu hướng đến khách sạn tăng cao...

3.2 Phân tích chương trình

Quá trình phân tích và xây dựng hệ thống thông tin cho chương trình qua nhiều giai đoạn, tuy nhiên trong giới hạn của phần tóm tắt, tôi minh họa sơ đồ mô hình hóa tiến trình phân tích và thiết kế chương trình như sau:



Hình 3.3. Quá trình thực hiện ứng dụng khai phá luật kết hợp

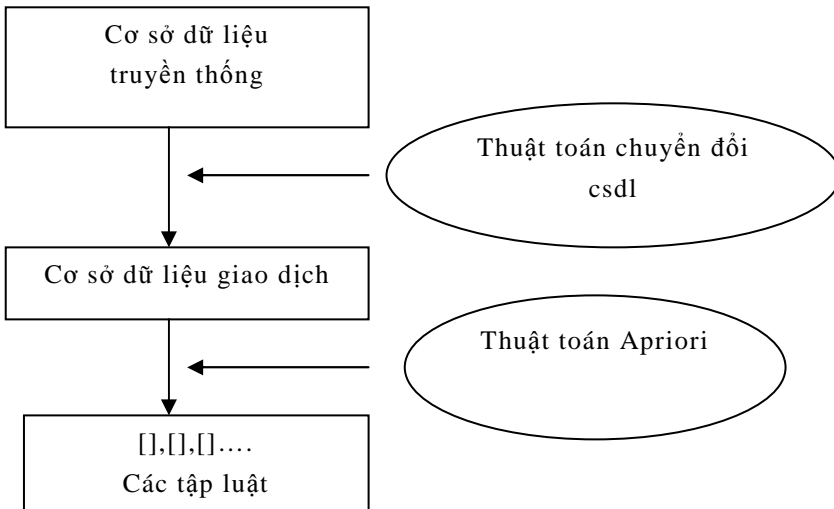
Giai đoạn tiền xử lý: Giai đoạn này nhằm thiết lập các đối tượng dữ liệu từ dữ liệu trong CSDL Khách. Dữ liệu được tiền xử lý đưa về dạng text, các thuộc tính (chính là các item) được ánh xạ bởi các số tự nhiên (tức là đánh số thứ tự các thuộc tính từ 1 đến hết). Mỗi dòng (bản ghi) được mô tả thành một dòng text như sau: liệt kê số thứ tự của các thuộc

tính (item) cách nhau một dấu cách (không liệt kê các thuộc tính mà liệt kê số thứ tự của nó). Tìm tập mục phổ biến và luật kết hợp dựa trên các số thứ tự này, kết quả được ánh xạ ngược trở lại tên các mục.

Giai đoạn khai phá: Đây là quá trình thực hiện các thuật toán (Apriori, Fp-growth) áp dụng đối với dữ liệu cung cấp sau giai đoạn tiền xử lý

3.3 Tóm lược cài đặt

3.3.1 Xây dựng chương trình



3.3.2 Giới thiệu chương trình

3.3.3 Kết quả chương trình

KẾT LUẬN

1. Kết quả đạt được

Luận văn đã trình bày tổng quan và các nét đặc trưng nhất trong lĩnh vực khai phá dữ liệu bao gồm các vấn đề cần khám phá tri thức, các hướng tiếp cận và nghiên cứu tiêu biểu, trong đó phát hiện luật kết hợp ứng dụng

trong việc xây dựng hệ thống trợ giúp triển khai chiến lược phát triển kinh doanh các loại hình dịch vụ du lịch tại Công Ty Cổ Phần Du Lịch Quảng Ngãi là một phương pháp khám phá tri thức quan trọng trong khai phá dữ liệu có nhiều ý nghĩa trong khoa học cũng như trong thực tiễn.

Về mặt lý thuyết, khai phá tri thức bao gồm các bước: Hình thành, xác định và định nghĩa bài toán; thu thập và tiền xử lý dữ liệu; khai phá dữ liệu, rút ra các tri thức; sử dụng các tri thức phát hiện được nhằm trợ giúp cho việc ra quyết định xây dựng chiến lược trong kinh doanh phát triển các loại hình dịch vụ du lịch.

Về thuật toán khai phá tri thức, luận văn trình bày một số thuật toán và minh họa một số thuật toán kinh điển về phát hiện tập chỉ báo phổ biến và khai phá luật kết hợp, như: Apriori, FP-Growth

Về mặt cài đặt thử nghiệm, luận văn giới thiệu kỹ thuật khai phá dữ liệu theo thuật toán Apriori áp dụng vào bài toán dự báo kinh doanh các loại hình dịch vụ du lịch tại Công Ty Cổ Phần Du Lịch Quảng Ngãi.

Trong quá trình thực hiện luận văn, tôi đã cố gắng tập trung tìm hiểu và tham khảo các tài liệu liên quan. Tuy nhiên, với thời gian và trình độ có hạn nên không tránh khỏi những hạn chế và thiếu sót. Tôi rất mong được sự nhận xét và góp ý của các thầy cô giáo và bạn bè, đồng nghiệp và những người cùng quan tâm để hoàn thiện hơn các kết quả nghiên cứu của mình.

2. Hướng nghiên cứu tiếp theo

Nghiên cứu sâu các thuật toán khai phá dữ liệu, tiếp tục hoàn thiện và mở rộng chương trình trong luận văn này để có thể áp dụng vào thực tế một cách triệt để.