

ỨNG DỤNG KỸ THUẬT PHÂN CỤM VÀ LUẬT KẾT HỢP KHAI PHÁ DỮ LIỆU KHÁCH HÀNG SỬ DỤNG DỊCH VỤ KHÁCH SẠN

APPLYING CLUSTERING TECHNIQUE AND ASSOCIATION RULE TO MINE DATA OF CUSTOMERS USING HOTEL SERVICES

Nguyễn Văn Chức, Đào Thị Giang

Trường Đại học Kinh tế, Đại học Đà Nẵng; chuc.nv@due.edu.vn, giang.dt37K14@gmail.com

Tóm tắt - Ngày nay, ngành công nghiệp du lịch, đặc biệt là kinh doanh khách sạn đang phát triển mạnh mẽ. Dữ liệu khách hàng lưu trú tại khách sạn được lưu trữ chứa rất nhiều tri thức giá trị. Tuy nhiên, dữ liệu này lại không được khai thác triệt để. Trong khi đó, các nhà quản trị khách sạn đang rất cần được hỗ trợ về vấn đề chăm sóc khách hàng, đưa ra chính sách tối ưu trong quản lý từ những tri thức có được nếu dữ liệu của khách hàng được khai thác và sử dụng hiệu quả. Bài báo tập trung nghiên cứu kỹ thuật phân cụm và luật kết hợp trong khai phá dữ liệu để phân tích dữ liệu khách hàng lưu trú tại khách sạn, giúp các nhà quản trị có thể nắm được đặc điểm khách hàng và đưa ra được các quyết định kinh doanh hợp lý. Trên cơ sở các tri thức phát hiện được, một giao tiếp trên nền web được xây dựng để người dùng sử dụng các tri thức này vào phân tích dữ liệu khách hàng sử dụng dịch vụ khi lưu trú tại khách sạn.

Từ khóa - khai phá dữ liệu, phân tích hành vi, phân cụm, luật kết hợp, dịch vụ khách sạn.

Abstract - Nowadays, tourism industry, especially hospitality business has been developing strongly. Customer information saved in hotels contains much valuable knowledge. However, this data has not been exploited thoroughly. Meanwhile, the hospitality business managers are yearning for being supported in customer service to make optimal policies on hotel management from the knowledge if the data has been exploited and used effectively. This paper focuses on studying clustering technique and association rule in data mining to analyse customer data in hotels to help the managers understand customer characteristics and make better decisions in business. Based on the knowledge discovered, the research group also builds an interface on a web platform to help users use that knowledge for analysing data of customers using hotels services.

Key words - data mining, behavior analysis, clustering, association rule, hotel services.

1. Đặt vấn đề

Cùng với sự phát triển của ngành du lịch, các khách sạn được liên tục xây dựng với nhiều quy mô khác nhau đã tạo nên sự cạnh tranh không ngừng. Để đáp ứng nhu cầu ngày càng cao của khách du lịch nước ngoài và nội địa, ngoài việc đáp ứng nhu cầu lưu trú của khách hàng thì vấn đề nâng cao chất lượng các dịch vụ khác tại khách sạn cũng rất quan trọng. Du khách sẵn sàng trả thêm các khoản chi phí hợp lý cho các dịch vụ cộng thêm của khách sạn khi nó thực sự cần thiết hoặc khiến họ hài lòng. Để tồn tại và phát triển trong thị trường đầy cạnh tranh này, các doanh nghiệp kinh doanh khách sạn cần hiểu được đối tượng du khách thường xuyên lưu trú tại khách sạn họ là ai và họ mong muốn những gì. Mang lại đúng giá trị mà du khách cần là phương pháp đơn giản nhất để phục vụ du khách ít tốn kém chi phí nhất và làm du khách hài lòng nhất. Để làm được điều này, các nhà quản lý cần có một chiến lược hợp lý dài hạn. Trên thực tế, các thông tin về khách hàng lưu trú tại khách sạn được lưu trữ với rất nhiều tri thức có giá trị, song các dữ liệu này lại không được khai thác hiệu quả. Do đó rất cần có một hệ thống giúp người dùng phân tích các thông tin lưu trú của khách hàng tại khách sạn. Từ đó, các nhà quản trị có thể đưa ra được một số giải pháp hỗ trợ chăm sóc khách hàng, giúp khách sạn chủ động hơn với khách hàng trong việc cung cấp dịch vụ, nhờ đó giúp tăng chất lượng dịch vụ, nâng cao mức độ hài lòng của khách hàng, tăng cơ hội kinh doanh cho khách sạn. Bài báo tập trung nghiên cứu về kỹ thuật phân cụm và luật kết hợp để tìm ra tìm ra được các đặc điểm của từng nhóm khách hàng, biết được mối quan hệ giữa các thuộc tính liên quan đến hành vi lưu trú của khách hàng và dự báo được khả năng sử dụng dịch vụ cũng như khả năng đặt tour du lịch của khách hàng tại khách sạn.

2. Sơ lược về kỹ thuật phân cụm và luật kết hợp

2.1. Phân cụm dữ liệu

Phân cụm dữ liệu là qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm càng giống nhau (similar) càng tốt và các đối tượng khác cụm thì càng khác nhau nhau (Dissimilar) càng tốt [3].

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Có rất nhiều kỹ thuật phân cụm như phân cụm phân hoạch, phân cụm phân cấp, phân cụm dựa trên mật độ... 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 bài toán phân cụm [3].

2.2. Luật kết hợp

Trong lĩnh vực Data Mining, mục đích của luật kết hợp (Association Rule - AR) là tìm ra các mối quan hệ giữa các đối tượng trong khối lượng lớn dữ liệu. Nội dung cơ bản của luật kết hợp được tóm tắt như dưới đây [3].

Cho cơ sở dữ liệu giao dịch T gồm tập các giao dịch $t_1, t_2, \dots, t_n$.

$T = \{t_1, t_2, \dots, t_n\}$. Mỗi giao dịch t_i bao gồm tập các đối tượng I (gọi là itemset).

$I = \{i_1, i_2, \dots, i_m\}$. Một itemset gồm k items gọi là k-itemset.

Mục đích của luật kết hợp là tìm ra sự kết hợp (tương quan) giữa các items.

Những luật kết hợp này có dạng $X \rightarrow Y$

Hai tiêu chí rất quan trọng trong việc đánh giá luật kết hợp đó là độ hỗ trợ (support) và độ tin cậy (confidence).

Công thức tính độ hỗ trợ và độ tin cậy của luật kết hợp $X \rightarrow Y$ [1]:

$$\text{Support}(X \rightarrow Y) = P(X \cup Y) = \frac{n(X \cup Y)}{N}$$

$$\text{Confidence}(X \rightarrow Y) = P(Y|X) = \frac{n(X \cup Y)}{n(X)}$$

Trong đó:

$n(X)$: Số giao dịch chứa X

N: Tổng số giao dịch

Các luật kết hợp có độ hỗ trợ và độ tin cậy lớn hơn hoặc bằng độ hỗ trợ tối thiểu (min_sup) và độ tin cậy tối thiểu (min_conf) gọi là các luật mạnh. min_sup và min_conf gọi là các giá trị ngưỡng (threshold) được xác định trước khi sinh các luật kết hợp. [1]

3. Xây dựng hệ thống khai phá dữ liệu khách hàng sử dụng dịch vụ khách sạn

3.1. Mô tả hệ thống

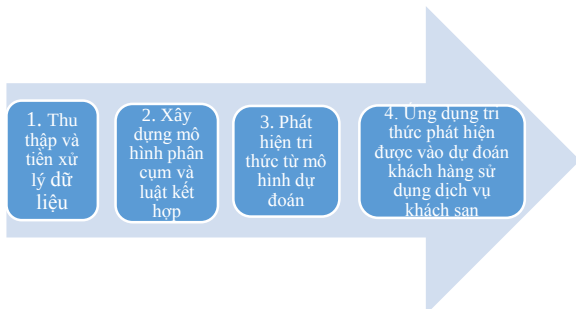
Mục đích: Ứng dụng kỹ thuật phân cụm và luật kết hợp phân tích dữ liệu khách hàng sử dụng dịch vụ khách sạn.

Đầu vào: Gồm các thông tin khách hàng sử dụng dịch vụ khách sạn: giới tính, quốc tịch, tuổi, nguồn khách,...; thông tin về các dịch vụ khách sạn: giặt là, thuê xe, đặt tour....

Đầu ra: Đưa ra đặc trưng của từng nhóm khách hàng, dự báo khả năng sử dụng dịch vụ và mối liên hệ giữa các thuộc tính liên quan đến hành vi lưu trú tại khách sạn của khách hàng

3.2. Kịch bản triển khai hệ thống phân tích dữ liệu khách hàng:

Hệ thống phân tích dữ liệu khách hàng được tiến hành theo các bước chính như Hình 1.



Hình 1. Kịch bản triển khai hệ thống phân tích dữ liệu khách hàng sử dụng dịch vụ khách sạn

Bước 1. Thu thập và tiền xử lý dữ liệu

Dữ liệu thu thập được gần 2000 mẫu, từ hệ thống quản lý khách sạn Headway của khách sạn Thời Đại ở Đà Nẵng từ đầu năm 2014 đến 3/2015.

Dữ liệu ban đầu gồm rất nhiều thuộc tính, sau quá trình tiền xử lý dữ liệu (sử dụng phương pháp trích chọn thuộc tính) để đánh giá mức độ ảnh hưởng của các thuộc tính đến

hành vi lưu trú tại khách sạn của khách hàng, mô hình xác định được các thuộc tính như Bảng 1.

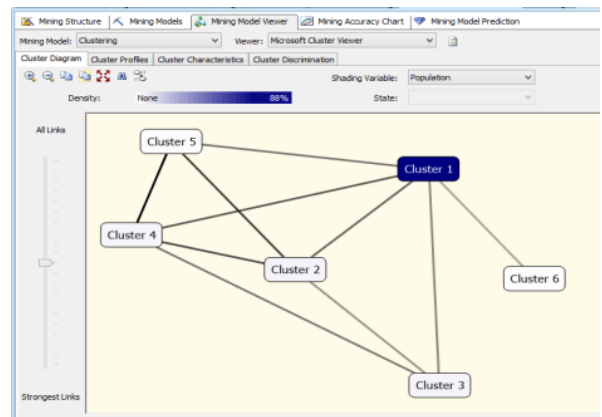
Bước 2. Xây dựng mô hình phân cụm và luật kết hợp

Mô hình phân cụm và luật kết hợp phân tích dữ liệu khách hàng sử dụng dịch vụ khách sạn được xây dựng trên công cụ khai phá dữ liệu Business Intelligence Development Studio (BIDS) của Microsoft. BIDS là công cụ rất mạnh cho phép triển khai các mô hình khai phá dữ liệu, được sử dụng rộng rãi hiện nay bởi khả năng kết nối dễ dàng với nhiều nguồn dữ liệu, giao diện dễ sử dụng và nhất là khả năng biểu diễn tri thức phát hiện được rất trực quan, dễ hiểu, dễ sử dụng. BIDS được tích hợp vào SQL SERVER 2005 trở về sau trong các phiên bản Enterprise hoặc Development [4].

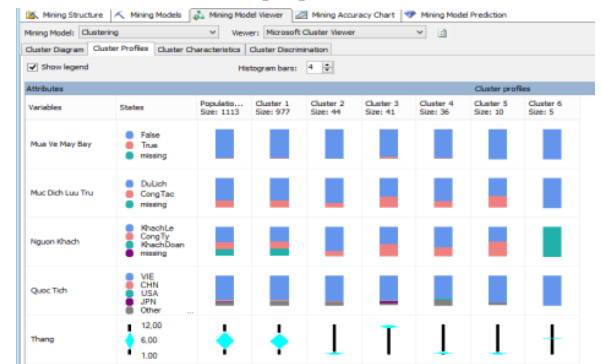
Sau khi thực hiện các thao tác tiền xử lý dữ liệu để phù hợp với mô hình khai phá dữ liệu, sử dụng Microsoft Clustering với thuật toán K-means và Microsoft Association Rule với thuật toán Apriori trong BIDS để xây dựng mô hình phân tích dữ liệu khách hàng sử dụng dịch vụ khách sạn.

Bước 3. Phát hiện tri thức từ mô hình phân cụm và luật kết hợp

a. Mô hình phân cụm



Hình 2. Kết quả phân cụm dữ liệu



Hình 3. Đặc trưng từng cụm dữ liệu

Bảng 1. Mô tả dữ liệu

STT	Tên thuộc tính	Kiểu dữ liệu	Miền giá trị	Giải thích
1	ID	Nominal	000001-002000	Mã khách hàng, khóa
2	GioiTinh	Nominal	Nu, Nam	Giới tính của khách hàng
3	Tuoi	Interval	>=18	Độ tuổi của khách hàng
4	QuocTich	Nominal	VIE, USA, CHN, JPN, AUS,	Quốc tịch của khách hàng
5	NguonKhach	Nominal	KhachLe, KhachDoan, CongTy	Nguồn khách hàng

6	MucDichLuuTru	Nominal	DuLich, CongTac	Mục đích lưu trú của khách hàng
7	ThoiGianLuuTru	Nominal	>=1, <=12	Thời gian lưu trú của khách hàng (tháng lưu trú)
8	SoNgayLuuTru	Interval	>=1	Số ngày lưu trú tại khách sạn
9	LoaiPhong	Nominal	StandardSingle, StandardDouble, StandardTwin, DeluxeSingle, DeluxeDouble, DeluxeTwin, VipSingle, VipDouble	Loại phòng khách sạn
10	HinhThucThanhToan	Nominal	TienMat, The, ChuyenKhoan	Hình thức thanh toán
11	GiatLa	Nominal	Co, Khong	Dịch vụ giặt là
12	ThueXe	Nominal	Co, Khong	Dịch vụ thuê xe
13	MuaVeMayBay	Nominal	Co, Không	Dịch vụ mua vé máy bay cho khách
14	DVMiniBar	Nominal	Co, Khong	Dịch vụ mini bar
15	MuaDoGiupKhach	Nominal	Co, Khong	Dịch vụ mua đồ giúp khách
16	ChenhLechHoaDon	Nominal	Co, Khong	Dịch vụ thanh toán hóa đơn cho khách
17	DoiTien	Nominal	Co, Khong	Dịch vụ đổi tiền cho khách
18	DatTourHue	Nominal	Co, Khong	Đặt tour Huế
19	DatTourBaNa	Nominal	Co, Khong	Đặt tour Bà Nà
20	DatTourNHS	Nominal	Co, Khong	Đặt tour Ngũ Hành Sơn
21	CityTour	Nominal	Co, Khong	Đặt City tour
22	DatTourSonTra	Nominal	Co, Khong	Đặt tour Sơn Trà
23	DatTourHoiAn	Nominal	Co, Khong	Đặt tour Hội An
24	DatTourCLC	Nominal	Co, Khong	Đặt tour Cù Lao Chàm

Từ mô hình phân cụm, cho ra được đặc trưng 6 nhóm khách hàng như sau:

Cụm 1: Cụm này đa số là khách đoàn người nước ngoài, có thời gian lưu trú nhiều ngày, độ tuổi từ 20 đến 40, sử dụng nhiều dịch vụ như giặt là, thuê xe, đặt tour, đổi tiền...

Cụm 2: Cụm này đa số là khách lẻ đi du lịch, lưu trú nhiều ngày, vào tháng 1 đến tháng 5, độ tuổi trên 30, thường ở loại phòng Deluxe, thanh toán bằng tiền mặt, sử dụng nhiều dịch vụ như giặt là, minibar, mua vé..., thường đặt Citytour, tour Bà Nà, Hội An.

Cụm 3: Cụm này đa số là khách lẻ, đi công tác, tập trung vào tháng 6 đến tháng 12, độ tuổi từ 18 đến 35, thường ở loại phòng Deluxe, thanh toán bằng thẻ hoặc chuyển khoản, sử dụng nhiều dịch vụ như giặt là, thuê xe, đặt tour, đổi tiền...

Cụm 4: Cụm này đa số là khách công ty, chỉ lưu trú 1 ngày tại khách sạn, vào tháng 1 đến tháng 5, thường thuê xe, mua vé máy bay..., thanh toán chủ yếu bằng tiền mặt hoặc chuyển khoản.

Cụm 5: Cụm này đa số là khách công ty, trong khoảng thời gian từ tháng 1 đến tháng 9, độ tuổi từ 26 đến 38, thường ở các phòng Single, thời gian lưu trú nhiều ngày, nên ngoài mục đích công tác, khách hàng thuộc nhóm này còn kết hợp đi du lịch, khách sạn thường thanh toán chênh lệch hóa đơn cho nhóm khách này.

Cụm 6: Cụm này chủ yếu là khách đoàn Việt Nam, có thời gian lưu trú nhiều ngày, thường vào tháng 6 đến tháng 10, tập trung nhiều vào tháng 7, sử dụng nhiều dịch vụ như giặt là, thuê xe, họ thường đặt tour Bà Nà, Cù Lao Chàm, Hội An, hình thức thanh toán bằng tiền mặt.

b. Mô hình luật kết hợp

Từ mô hình luật kết hợp đã xây dựng, phát hiện được

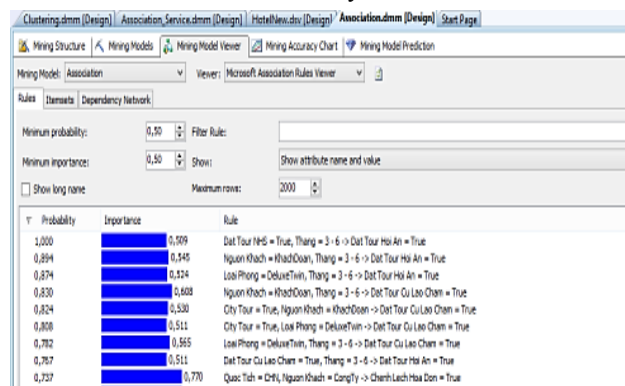
tri thức về mối quan hệ giữa các thuộc tính liên quan đến hành vi lưu trú của khách hàng tại khách sạn. Sau đây là một số luật được trích ra từ mô hình mô hình:

L1: Khách hàng đặt tour Ngũ Hành Sơn, vào tháng 3 đến tháng 6 thì đặt tour Hội An, với độ tin cậy 100%.

L2: Khách hàng là khách đoàn, lưu trú tại khách sạn vào tháng 3 đến tháng 6 thì đặt tour Hội An, với độ tin cậy 89,4%.

L3: Khách ở loại phòng Deluxe Twin, vào tháng 3 đến tháng 6 thì đặt tour Hội An với độ tin cậy 87,4%.

L4: Khách hàng là khách đoàn, đặt City Tour thì sẽ đặt tour Cù Lao Chàm với độ tin cậy 82,4%.



Hình 4. Kết quả mô hình luật kết hợp

Bước 4. Ứng dụng tri thức phát hiện được vào phân tích dữ liệu khách hàng lưu trú tại khách sạn

Dựa vào các tri thức phát hiện được từ mô hình phân cụm và luật kết hợp, một hệ thống giao tiếp được xây dựng trên nền web cho phép người dùng sử dụng để tìm ra đặc điểm từng nhóm khách hàng, dự báo khả năng sử dụng dịch vụ của khách hàng.

Trang chủ | Nhập thông tin | Dự đoán khách hàng | Phân nhóm khách hàng | Xem luật kết hợp



HỆ THỐNG PHÂN TÍCH DỮ LIỆU KHÁCH HÀNG

- Nhập thông tin
- Xem thông tin
- Dự đoán
- Phân nhóm khách hàng
- Xem luật kết hợp

Trang chủ | Nhập thông tin | Dự đoán khách hàng | Phân nhóm khách hàng | Xem luật kết hợp

THÔNG TIN KHÁCH HÀNG

ID	Quốc tịch	Tuổi	Giới tính	Ngành khách	Thương hiệu	Loại phòng	Phân nhóm	Dự đoán	Dự đoán độ tin cậy	Chức vụ
1	VIE	23	Nữ	Khách lẻ	3	StandardSingle	Khách lẻ	✓	✓	✓
2	VIE	45	Nam	Khách lẻ	3	StandardSingle	Khách lẻ	✓	✓	✓
4	VIE	26	Nam	Khách lẻ	1	StandardSingle	Khách lẻ	✓	✓	✓
1	VIE	30	Nữ	Khách lẻ	9	StandardSingle	Khách lẻ	✓	✓	✓
6	CHN	35	Nữ	Khách đoàn	10	StandardSingle	Khách đoàn	✓	✓	✓
7	JPN	28	Nam	Khách lẻ	1	VipSingle	Khách lẻ	✓	✓	✓
8	KOR	23	Nữ	Công ty	6	VipSingle	Khách lẻ	✓	✓	✓
9	KOR	26	Nam	Công ty	8	VipSingle	Khách lẻ	✓	✓	✓
10	VIE	30	Nam	Khách lẻ	10	StandardSingle	Khách lẻ	✓	✓	✓
11	VIE	30	Nam	Khách lẻ	1	StandardSingle	Khách lẻ	✓	✓	✓

1234

Trang chủ | Nhập thông tin | Dự đoán | Phân nhóm khách hàng | Xem luật kết hợp

PHÂN CỤM KHÁCH HÀNG

THÔNG TIN KHÁCH HÀNG

Họ tên	Hồ Minh Đức	Quốc tịch	Việt Nam
Tuổi	30	Giới tính	Nam
Thương hiệu	10	Ngành khách	Công ty
Số ngày lưu trú	4	Loại phòng	DeluxeSingle

KẾT LUẬN

Nhóm khách Thuộc cụm 3

Đặc trưng Cụm này đa số là khách công ty, trong khoảng thời gian từ tháng 1 đến tháng 9, độ tuổi từ 26 đến 38, thường ở các phòng Single, thời gian lưu trú nhiều ngày, ăn ngoài mục đích công tác, khách hàng thuộc nhóm này còn kết hợp đi du lịch, khách sạn thường thanh toán chênh lệch hóa đơn cho nhóm khách này

Đề xuất Khách hàng lưu trú tại sạn chủ yếu là khách đi công tác dài ngày. Vì khách chủ yếu là liên việc nên ban quản lý cần thường xuyên kiểm tra chất lượng dịch vụ.

Trang chủ | Nhập thông tin | Dự đoán | Phân nhóm khách hàng | Xem luật kết hợp

DỰ ĐOÁN KHẢ NĂNG ĐẶT TOUR

Tour	Xác suất
City tour	0,3919157496307278
Bà Nà	0,610062693081761
Hoi An	0,691626844364241
Cù Lao Chàm	0,783935109073046
Huế	0,979335130278527
Ngũ Hành Sơn	0,966736113926123
Sơn Trà	0,922142744950584

Trang chủ | Nhập thông tin | Dự đoán | Phân nhóm khách hàng | Xem luật kết hợp

LUẬT KẾT HỢP

Số luật: 0

Xác suất tối thiểu: 0.5

Lọc Luật:

OK

STT	LUẬT	XÁC SUẤT	ĐỘ HỖ TRỢ
1	Dat Tour Hoi An = Existing, Thương = 3 - 6 -> Dat Tour Hoi An = Existing	1,000	19
2	Ngành Khách = Khách Đoàn, Thương = 3 - 6 -> Dat Tour Hoi An = Existing	0,894	84
3	Loại Phòng = Deluxe Twin, Thương = 3 - 6 -> Dat Tour Hoi An = Existing	0,874	76

Hình 5. Giao tiếp người dùng với hệ thống phân tích dữ liệu khách hàng

4. Kết luận và hướng phát triển

Bài báo đã tìm hiểu về lý thuyết kỹ thuật phân cụm và luật kết hợp, từ đó nghiên cứu ứng dụng các kỹ thuật này vào xây dựng mô hình khai phá dữ liệu khách hàng lưu trú tại khách sạn

Dựa vào mô hình khai phá dữ liệu với hai kỹ thuật phân cụm dữ liệu và phát hiện luật kết hợp đã xây dựng, nhóm tác giả đã xây dựng thành công một giao tiếp trên nền web để hỗ trợ việc ra quyết định của các nhà quản lý khách sạn, giúp đưa ra được các chính sách riêng cho từng nhóm khách hàng, dự báo hành vi sử dụng dịch vụ khách sạn cũng như đặt tour du lịch của khách hàng. Trong thời gian tới sẽ nghiên cứu mở rộng sang lĩnh vực du lịch nhằm nâng cao hiệu quả trong việc phân tích hành vi du lịch của khách du lịch tại các công ty du lịch lữ hành.

TÀI LIỆU THAM KHẢO

- [1] Nguyễn Đức Thuận, *Nhập môn khai phá dữ liệu và quản trị tri thức*, NXB Thông tin và truyền thông, 2013.
- [2] Trần Thị Kim Hằng, *Luận văn Ứng dụng khai phá dữ liệu dự đoán bệnh đái tháo đường*, 2014
- [3] Jiawei Han and Micheline Kamber, *Datamining: Concepts and Techniques*, Simon Fraser University, 2011.
- [4] Jamie MacLennan, Z.T., Bogdan Crivat, *Data Mining with Microsoft SQL Server 2008*, Indianapolis, Indiana: Wiley Publishing, Inc, 2008.
- [5] <http://bis.net.vn/data+mining>
- [6] <http://www.sqlserverdatamining.com>

(BBT nhận bài: 24/07/2015, phản biện xong: 23/09/2015)