

ỨNG DỤNG KỸ THUẬT PHÂN LỚP VÀ PHÂN CỤM TRONG KHAI PHÁ DỮ LIỆU PHÂN TÍCH HÀNH VI SỬ DỤNG ĐIỆN THOẠI DI ĐỘNG CỦA SINH VIÊN TRƯỜNG ĐẠI HỌC KINH TẾ - ĐẠI HỌC ĐÀ NẴNG

APPLYING CLASSIFICATION AND CLUSTERING TECHNIQUES IN DATA MINING TO ANALYZE MOBILE PHONE USAGE BEHAVIOR OF STUDENTS AT DANANG UNIVERSITY OF ECONOMICS – THE UNIVERSITY OF DANANG

Nguyễn Văn Chức, Nguyễn Thị Thùy Dương, Nguyễn Bá Thế, Phan Thị Bích Vân

Trường Đại học Kinh tế, Đại học Đà Nẵng

chuc.nv@due.edu.vn, thuyduongthoquang@gmail.com, thenb@due.edu.vn, van.pth@due.edu.vn

Tóm tắt - Hiện nay, trên thị trường điện thoại Việt Nam có rất nhiều dòng điện thoại di động từ nhiều thương hiệu với đủ mẫu mã, chức năng, giá cả... khiến cho người mua gặp nhiều khó khăn khi lựa chọn. Nghiên cứu hành vi sử dụng điện thoại của sinh viên nhằm hiểu được xu hướng và các yếu tố ảnh hưởng đến quyết định mua điện thoại của sinh viên, giúp tư vấn cho sinh viên có được quyết định đúng đắn khi lựa chọn điện thoại. Các tri thức này cũng rất có ích đối với các đại lý bán điện thoại, giúp họ tiếp cận tốt hơn với khách hàng của mình. Bài báo nghiên cứu kỹ thuật phân lớp dựa vào cây quyết định và phân cụm dữ liệu để phân tích hành vi sử dụng điện thoại của sinh viên trên cơ sở thu thập dữ liệu của sinh viên Trường Đại học Kinh tế - ĐHQĐ. Từ 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 hành vi sử dụng điện thoại của sinh viên.

Từ khóa - khai phá dữ liệu; phân tích hành vi; phân lớp; cây quyết định; phân cụm; điện thoại di động.

Abstract - Nowadays, there are so many kinds of mobile phone from various brands with many models, functions, prices... that buyers may encounter many difficulties when they want to buy a mobile phone. Studying behavior of students in using mobile phone to understand the trends and factors that affect students' decision to buy a mobile phone will help students make the right decisions to choose one. This knowledge is also useful with phone dealers to help them reach their customers more effectively. This paper focuses on studying classification based on decision tree and clustering technique to analyze mobile phone usage behavior of students with data collected from students at Danang University of Economics - the University of Danang. Based on knowledge discovered, an interface is built on a web platform to help users use that knowledge for analysing data of students using the mobile phone.

Key words - data mining; behavior analysis; classification; decision tree; clustering; mobile phone.

1. Đặt vấn đề

Hiện nay, cùng với mức sống của người dân ngày càng được cải thiện, việc sở hữu một chiếc điện thoại di động (ĐTDD) đã không còn là một giấc mơ xa vời như cách đây hơn chục năm. Báo cáo về thị trường ĐTDD châu Á/Thái Bình Dương của Công ty Nghiên cứu thị trường IDC công bố khảo sát quý IV/2014 cho thấy, thị trường Việt Nam năm 2014 đã tiêu thụ 28,7 triệu chiếc ĐTDD, tăng 13% so với năm trước. Trong đó, ĐTDD thông minh (smartphone) có mức tăng trưởng cao nhất với 11,6 triệu chiếc, tăng 57% so với năm 2013. Thêm vào đó, việc giới trẻ sử dụng đồng thời hai chiếc điện thoại (hoặc nhiều hơn) hoặc thay đổi điện thoại mới từ 6 tháng đến 1 năm đã tăng lên rõ rệt [5]. Điều này cho thấy nhu cầu rất lớn về ĐTDD của người dân Việt Nam hiện nay. Nhận thấy tiềm năng lớn từ thị trường Việt Nam, ngày càng nhiều hãng điện thoại nước ngoài “đổ bộ” vào Việt Nam cùng với sự xuất hiện của những điện thoại thương hiệu Việt như FPT, Viettel hay Bphone. Giữa một “rừng” ĐTDD hẳn rất nhiều người có nhu cầu mua điện thoại sẽ gặp khó khăn, phân vân chưa biết nên lựa chọn những dòng điện thoại của hãng sản xuất nào, chức năng và giá cả như thế nào là hợp lý, phù hợp nhất với yêu cầu sử dụng của bản thân. Đó cũng là câu hỏi mà nhiều bạn sinh viên băn khoăn. Đứng trước thực tế đó, cần có một hệ thống tư vấn, giúp sinh viên có dự định mua ĐTDD dễ dàng hơn trong việc lựa chọn. Ngoài ra, các hãng ĐTDD rất cần hiểu được xu hướng, đặc điểm và các yếu tố ảnh hưởng đến quyết định mua điện thoại của sinh viên để đề ra kế hoạch kinh doanh hiệu quả và xây dựng chiến lược quảng cáo, tiếp thị nhằm mở rộng thị phần, thu hút khách hàng, đặc biệt là sinh

viên, hay rộng hơn là giới trẻ, những người đóng vai trò rất lớn cho sự thành bại của một thương hiệu điện thoại. TIEP

2. Sơ lược về kỹ thuật phân lớp dữ liệu bằng cây quyết định và phân cụm dữ liệu

2.1. Giới thiệu kỹ thuật phân lớp dữ liệu dựa vào cây quyết định

Trong lĩnh vực khai phá dữ liệu, cây quyết định (Decision Tree – DT) là một mô hình dự đoán thuộc lớp các bài toán phân lớp dùng để xác định lớp của các đối tượng cần dự đoán [3]. Bản chất cây quyết định dựa vào dãy các luật IF ... THEN để dự đoán lớp của đối tượng. Mỗi nút trong (internal node) của DT tương ứng với một biến, đường nối giữa một nút trong với nút con của nó thể hiện một giá trị cụ thể của biến đó. Mỗi nút lá (leaf) đại diện cho giá trị dự đoán. Cây quyết định học để dự đoán giá trị của các biến phân loại bằng cách dựa vào tập dữ liệu huấn luyện (training data) để chọn ra nút gốc (root node) để phân tách cây bằng cách tính độ lợi thông tin (Information Gain - IG); quá trình phân tách cây được thực hiện một cách đệ quy cho đến khi không thể tiếp tục thực hiện việc phân tách cây được nữa [1], [3].

Cây quyết định được chia thành hai loại:

Cây hồi quy dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu định lượng như dự đoán doanh thu, lợi nhuận, giá thành sản phẩm... Thuật toán phổ biến dùng để xây dựng cây hồi quy là CART [3].

Cây phân lớp dùng để dự đoán giá trị của biến phân loại có kiểu dữ liệu định danh như dự đoán khả năng mua hàng (mua, không mua), khả năng bị bệnh (bị bệnh, không

bị bệnh), kết quả học tập của sinh viên (xuất sắc, giỏi, khá, trung bình, yếu)... Thuật toán phổ biến dùng để xây dựng cây phân lớp là ID3, J48, C4.5, C5.0 [3].

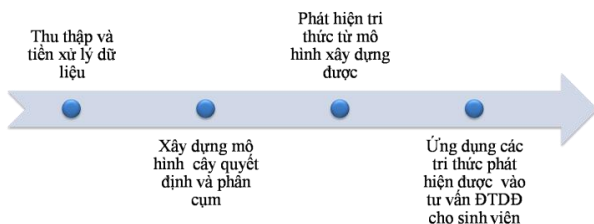
2.2. Giới thiệu phân cụm dữ liệu

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 [2].

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 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 [2].

3. Kịch bản triển khai hệ thống

Hệ thống phân tích hành vi sử dụng điện thoại của sinh viên được tiến hành theo các bước như Hình 1.



Hình 1. Kịch bản triển khai hệ thống phân tích hành vi sử dụng điện thoại di động

Bước 1. Thu thập và tiền xử lý dữ liệu

Dữ liệu thu thập được hơn 1300 mẫu qua khảo sát các sinh viên trường Đại học Kinh tế Đà Nẵng thông qua bảng hỏi online từ tháng 2 - 4/2016.

Sau khi thu thập dữ liệu, tiến hành tiền xử lý dữ liệu:

loại bỏ các dữ liệu bị thiếu, không phù hợp, chuyển đổi dữ liệu. Dữ liệu sau tiền xử lý có cấu trúc được xác định như Bảng 1.

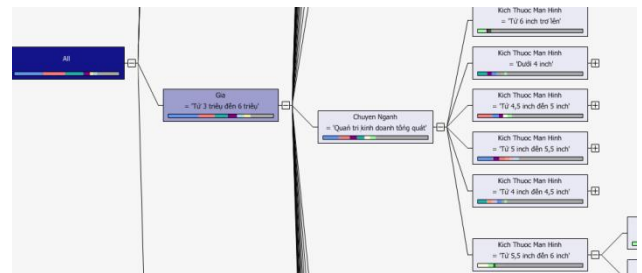
Bước 2. Xây dựng mô hình cây quyết định và phân cụm

Mô hình cây quyết định và phân cụm dữ liệu phân tích dữ liệu sinh viên trường Đại học Kinh tế Đà Nẵng sử dụng điện thoại di động được xây dựng trên công cụ SQL Server Data Tools (SSDT) của Microsoft. SSDT là công cụ rất mạnh cho phép triển khai hầu hết các kỹ thuật khai phá dữ liệu như cây quyết định, phân cụm dữ liệu, hồi qui, phát hiện luật kết hợp, phân tích dãy số thời gian... với giao diện dễ sử dụng, khả năng tích hợp dữ liệu mạnh mẽ và nhất là khả năng biểu diễn tri thức phát hiện được rất trực quan, dễ hiểu, dễ sử dụng [4], [6].

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 Decision Tree và Microsoft Clustering với thuật toán K-mean trong SSDT để xây dựng mô hình phân tích dữ liệu sinh viên sử dụng điện thoại di động.

Bước 3. Phát hiện tri thức từ mô hình cây quyết định và phân cụm

a. Mô hình cây quyết định



Hình 2. Kết quả một nhánh của cây quyết định

Bảng 1. Dữ liệu để xây dựng mô hình

STT	Tên thuộc tính	Kiểu dữ liệu	Miền giá trị	Giải thích
1	ID	Nominal	M0001-M1302	Mã sinh viên, khóa
2	GioiTính	Nominal	Nam, Nữ	Giới tính của sinh viên
3	ChuyenNganh	Nominal	Hành chính công, Kế toán, Kiểm toán, Kinh tế chính trị, Kinh tế đầu tư, Kinh tế lao động, Kinh tế phát triển, Kinh tế và quản lý công, Luật học, Luật kinh doanh, Marketing, Ngân hàng, Ngoại thương, Quản trị chuỗi cung ứng, Quản trị hệ thống thông tin, Quản trị khách sạn, Quản trị kinh doanh du lịch, Quản trị kinh doanh thương mại, Quản trị kinh doanh tổng quát, Quản trị nguồn nhân lực, Quản trị sự kiện và lễ hội, Quản trị tài chính, Tài chính công, Tài chính doanh nghiệp, Thống kê kinh tế xã hội, Thương mại điện tử, Tin học quản lý.	Chuyên ngành của trường ĐHKT-ĐHĐN
4	NgheNhac	Nominal	Hoàn toàn không thích, Không thích lắm, Bình thường, Thích, Rất thích	Mức độ yêu thích nghe nhạc của sinh viên
5	LuotWeb	Nominal	Hoàn toàn không thích, Không thích lắm, Bình thường, Thích, Rất thích	Mức độ yêu thích lướt web của sinh viên
6	ChupAnh-QuayPhim	Nominal	Hoàn toàn không thích, Không thích lắm, Bình thường, Thích, Rất thích	Mức độ yêu thích chụp ảnh/quay phim của sinh viên
7	ChoiGame	Nominal	Hoàn toàn không thích, Không thích lắm, Bình thường, Thích, Rất thích	Mức độ yêu thích chơi game của sinh viên
8	ThichSuMoiLaThuThach	Nominal	Hoàn toàn không thích, Không thích lắm, Bình thường, Thích, Rất thích	Mức độ ham muốn trải nghiệm sự mới lạ và thử thách của sinh viên
9	HangDienThoai	Nominal	Apple, Samsung, ASUS, BlackBerry...	Hãng ĐTDĐ sinh viên sử dụng

10	AnhHuong	Nominal	Cấu hình, Giá cả, Mẫu mã, Người thân hoặc bạn bè gợi ý, Thời gian sử dụng của pin hoặc độ bền của máy, Thương hiệu, Tính năng, Ảnh hưởng khác	Yếu tố ảnh hưởng lớn nhất trong việc lựa chọn ĐTDĐ của sinh viên
11	Gia	Nominal	Dưới 1 triệu, Từ 1 triệu đến 3 triệu, Từ 3 triệu đến 6 triệu, Từ 6 triệu đến 10 triệu, Từ 10 triệu trở lên	Giá chiếc ĐTDĐ sinh viên đang sử dụng
12	KichThuocManHinh	Nominal	Dưới 4 inch, Từ 4 inch đến 4,5 inch, Từ 4,5 inch đến 5 inch, Từ 5 inch đến 5,5 inch, Từ 5,5 inch đến 6 inch, Từ 6 inch trở lên	Kích thước màn hình chiếc ĐTDĐ sinh viên đang sử dụng
13	MucChiTieu	Nominal	Dưới 1,5 triệu, Từ 1,5 triệu đến 2 triệu, Từ 2 triệu đến 2,5 triệu, Từ 2,5 triệu đến 3 triệu, Từ 3 triệu trở lên	Mức chi tiêu trung bình hàng tháng của sinh viên

Từ mô hình cây quyết định đã xây dựng, có thể rút ra một số luật như sau:

IF (Gia = 'Dưới 1 triệu' and Chuyen Nganh = 'Quản trị kinh doanh tổng quát' and Luot Web not = 'Rất thích') THEN Hang Dien Thoai = 'Microsoft/Nokia' (xác suất 22.73%).

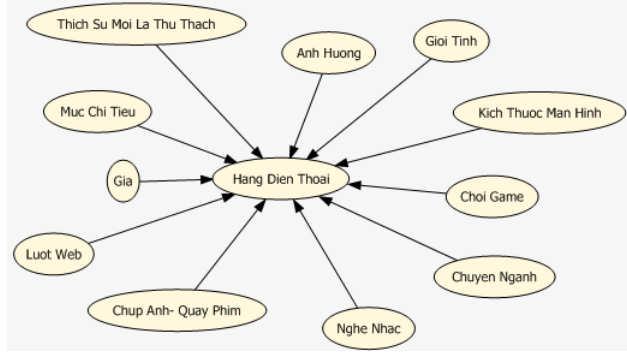
IF (Gia = 'Từ 10 triệu trở lên' and Chuyen Nganh = 'Marketing') THEN Hang Dien Thoai = 'Apple' (xác suất 19.05%).

IF (Gia = 'Từ 1 triệu đến 3 triệu' and Chuyen Nganh = 'Kinh tế phát triển' and Kich Thuoc Man Hinh = 'Từ 4 inch đến 4.5 inch' and Anh Huong not = 'Mẫu mã') THEN Hang Dien Thoai = 'Microsoft/Nokia' (xác suất 19.05%).

IF (Gia = 'Dưới 1 triệu' and Chuyen Nganh = 'Kế toán' and Choi Game not = 'Thích') THEN Hang Dien Thoai = 'Microsoft/Nokia' (xác suất 17.07%).

IF (Gia = 'Từ 10 triệu trở lên' and Chuyen Nganh = 'Kinh tế đầu tư' and Anh Huong not = 'Mẫu mã') THEN Hang Dien Thoai = 'Apple' (xác suất 17.07%).

Ngoài ra, mô hình cây quyết định cũng cho biết mức độ ảnh hưởng của từng tiêu chí đối với quyết định lựa chọn hãng ĐTDĐ như Hình 3 và Bảng 2.



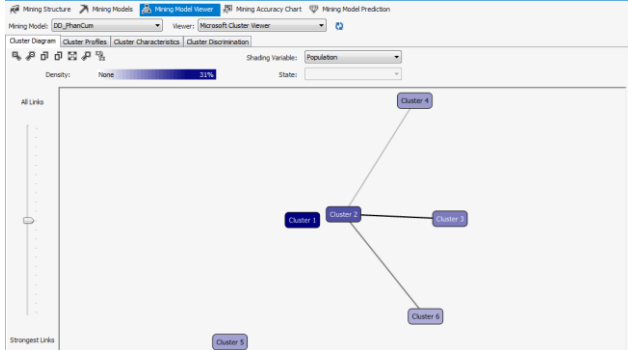
Hình 3. Các yếu tố ảnh hưởng đến quyết định lựa chọn hãng ĐTDĐ

Bảng 2. Mức độ ảnh hưởng của các tiêu chí đến quyết định lựa chọn hãng ĐTDĐ

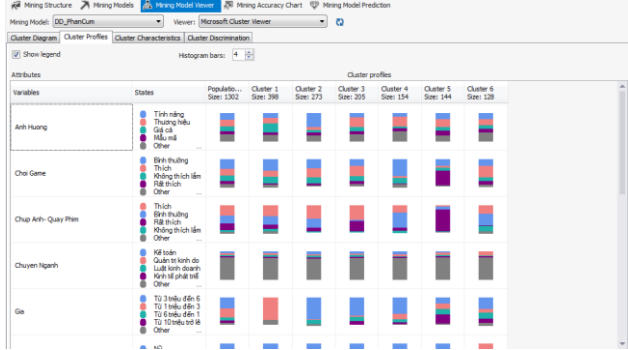
Tiêu chí	Mức độ phụ thuộc (giảm dần)
Giá	1
Chuyên ngành	2
Choi game	3
Ảnh hưởng lựa chọn	4
Chụp ảnh – Quay phim	5
Kích thước màn hình	6
Giới tính	7
Thích sự mới lạ và thử thách	8

Nghe nhạc	9
Mức chi tiêu trung bình hàng tháng	10
Lướt web	11

b. Mô hình phân cụm



Hình 4. Kết quả mô hình phân cụm



Hình 5. Đặc trưng các cụm dữ liệu

Từ mô hình phân cụm đã xây dựng, ta có 6 nhóm sinh viên với các đặc trưng như sau:

Cụm 1: Chủ yếu là nữ, giá mua điện thoại là từ 1 triệu đến 3 triệu, đều thích lướt web và nghe nhạc, thích sự mới lạ – thử thách, mức chi tiêu trung bình hàng tháng là dưới 1,5 triệu, kích thước màn hình sử dụng là dưới 4 inch. Hãng điện thoại chủ yếu được lựa chọn trong cụm này là Microsoft/Nokia và SamSung.

Cụm 2: Chủ yếu là nữ, giá điện thoại từ 3 triệu đến 6 triệu, thích lướt web, nghe nhạc và chụp ảnh – quay phim, kích thước màn hình từ 5 inch đến 5.5 inch, quan tâm đến tính năng nhất khi mua điện thoại. Hãng điện thoại chủ yếu được lựa chọn trong cụm này là SamSung.

Cụm 3: Chủ yếu là nữ, giá điện thoại từ 3 triệu đến 6 triệu, rất thích nghe nhạc, thích quay phim – chụp ảnh và lướt web, thích sự mới lạ thử thách. Hãng điện thoại chủ yếu được lựa chọn trong cụm này là Apple và SamSung.

Cụm 4: Chủ yếu là nữ, mức độ yêu thích đối với lướt web, nghe nhạc, chơi game và chụp ảnh – quay phim là bình thường, giá điện thoại từ 3 triệu đến 6 triệu, mức chi tiêu từ 1,5 triệu đến 2 triệu/tháng. Hãng điện thoại chủ yếu được sử dụng là Microsoft/Nokia và Samsung.

Cụm 5: Sinh viên trong cụm này chủ yếu rất thích chụp ảnh – quay phim, lướt web, nghe nhạc, chơi game, rất thích sự mới lạ và thử thách. Hãng điện thoại chủ yếu là Apple.

Cụm 6: Chủ yếu là nam, thích lướt web, nghe nhạc, chơi game, mức độ yêu thích chụp ảnh – quay phim là bình thường, thích sự mới lạ và thử thách, giá điện thoại từ 3 triệu đến 6 triệu. Hãng điện thoại chủ yếu là Apple.

Hình 6. Giao tiếp người dùng với hệ thống phân tích dữ liệu

Bước 4. Ứng dụng tri thức phát hiện được vào phân tích dữ liệu sinh viên sử dụng điện thoại di động

Dựa vào các tri thức phát hiện được từ mô hình cây quyết định và phân cụm dữ liệu, một hệ thống giao tiếp được xây dựng trên nền web cho phép người dùng có thể sử dụng các tri thức này vào việc tư vấn lựa chọn ĐTDĐ phù hợp cũng như tìm hiểu các đặc trưng của các nhóm người dùng ĐTDĐ (sinh viên).

4. Kết luận và hướng phát triển

Bài báo đã tập trung tìm hiểu kỹ thuật phân lớp dữ liệu dựa vào cây quyết định và phân cụm dữ liệu, từ đó nghiên cứu ứng dụng các kỹ thuật này vào xây dựng mô hình khai phá dữ liệu sinh viên sử dụng ĐTDĐ.

Dựa vào mô hình khai phá dữ liệu với hai kỹ thuật phân lớp, dựa vào kỹ thuật cây quyết định và phân cụm dữ liệu 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 sinh viên, giúp lựa chọn được chiếc ĐTDĐ phù hợp nhất với bản thân. Thêm vào đó, kết quả nghiên cứu cũng giúp cho các nhà phân phối ĐTDĐ nắm được xu hướng, đặc trưng của sinh viên trong việc sử dụng ĐTDĐ, từ đó đề ra chiến lược kinh doanh hiệu quả. Trong thời gian tới sẽ tiến hành thu thập thêm dữ liệu từ nhiều nguồn khác nhau, không chỉ từ sinh viên trường Đại học Kinh tế – Đại học Đà Nẵng, mà sẽ nghiên cứu mở rộng đối với các sinh viên trên toàn địa bàn thành phố Đà Nẵng, làm cho dữ liệu đầy đủ và đa dạng hơn, nhằm hoàn thiện thêm mô hình và nâng cao hiệu quả phân tích và dự đoán.

TÀI LIỆU THAM KHẢO

- [1] Nguyễn Đức Thuần, *Nhập môn khai phá dữ liệu và quản trị tri thức*, NXB Thông tin và Truyền thông, 2013.
- [2] Jiawei Han and Micheline Kamber, *Data Mining: Concepts and Techniques*, Simon Fraser University, 2011.
- [3] Rokach Lior; Maimon O., *Data mining with decision trees: theory and applications*, World Scientific Pub Co Inc. ISBN 978-9812771711, 2008.
- [4] Nguyễn Văn Chức, “Ứng dụng kỹ thuật cây quyết định trong khai phá dữ liệu xây dựng hệ thống tư vấn chọn ngành tuyển sinh đại học”, *Tạp chí KH&CN DHDN*, số 1(74).2014, Quyển 2.
- [5] Mai Phương, “Năm 2014, Việt Nam tiêu thụ hơn 28 triệu chiếc điện thoại di động”, *Báo Thanh Niên Online*, 2015.
- [6] JamieMacLennan, Z.T., Bogdan Crivat, *Data Mining with Microsoft SQL Server 2008*, Indianapolis, Indiana: Wiley Publishing, Inc, 2008.
- [7] <http://bis.net.vn/forums/36.aspx>

(BBT nhận bài: 20/05/2016, phản biện xong: 30/05/2016)