BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC HOA SEN

KHOA KHOA HỌC VÀ CÔNG NGHỆ

KHÓA LUẬN TỐT NGHIỆP

Tên đề tài:

XÂY DỰNG HỆ THỐNG GIỚI THIỆU BÁN HÀNG TRÊN MẠNG XÃ HỘI

Giảng viên hướng dẫn : Tiến sĩ. Trần Vũ Bình

Nhóm sinh viên thực hiện : Nguyễn Hoàng Phúc (09258L)

Vũ Thị Quỳnh (09259L)

Huỳnh Thị Minh Tâm (09262L)

Lớp : QL092L

Tháng 12 /năm 2010

ĐẠI HỌC
HOA SEN
KHOA KH&CN

PHIẾU GIAO ĐỀ TÀI KHÓA LUẬN TỐT NGHIỆP

1. **Mỗi sinh viên phải viết riêng một báo cáo**
2. **Phiếu này phải dán ở trang đầu tiên của báo cáo**

1. Họ và tên sinh viên/ nhòm sinh viên được giao đề tài (sĩ số trong nhóm: 3)

(1) Huỳnh Thị Minh Tâm MSSV: 09262L khóa: QL092L

(2) Vũ Thị Quỳnh MSSV: 09259L khóa: QL092L

(3) Nguyễn Hoàng Phúc MSSV: 09258L khóa: QL092L

Chuyên ngành : CNTT .....Khoa : KH&CN

2. Tên đề tài : Xây Dựng Hệ Thống Giới Thiệu Bán Hàng Trên Mạng Xã Hội.

3. Các dữ liệu ban đầu:

Tài liệu về các phương pháp khai thác dữ liệu.

Thông tin kỹ thuật liên quan đến các mạng xã hội.

4. Các yêu cầu đặc biệt:

Kết hợp các phương pháp khai thác dữ liệu để đáp ứng nhu cầu dữ liệu đa dạng biến đổi liên tục trên mạng xã hội.

So sánh việc sử dụng các phương pháp khai thác dữ liệu đơn thuần và sự kết hợp để đánh giá hiệu quả của việc kết hợp.

5. Kết quả tối thiểu phải có:

1.Phương pháp khai thác dữ liệu kết hợp ít nhất 2 phương pháp khai thác dữ liệu truyền thống

2.Phương pháp xử lý dữ liệu online, offline để tăng hiệu suất hệ thống

3.Xây dựng website thử nghiệm ý tưởng phát triển

Ngày giao đề tài:……../……./…… Ngày nộp báo cáo: ……/……/……..

Họ tên GV hướng dẫn 1: …..……Chữ ký: ……

Họ tên GV hướng dẫn 2: ……..…….. Chữ ký: …..……..

Ngày .... tháng ... năm

1. TRÍCH YẾU

Trong những năm gần đây, bán hàng trên mạng xã hội là xu hướng phổ biến của nhiều người vì nơi đó là một cộng đồng lớn, nơi kết nối những người có chung một sở thích, bán hàng ở đây sẽ gây sự chú ý hơn. Hiện tại vẫn chưa có phương pháp, công cụ cụ thể giúp người dùng mạng xã hội mua hàng trực truyến một cách hiệu quả. Bên cạnh đó, mạng xã hội là mô hình tương đối đa dạng và phức tạp vì có nhiều người kết nối với nhau, tồn tại nhiều mối quan hệ, nhiều sở thích luôn luôn thay đổi theo thời gian. Vậy làm thế nào để có thể nắm bắt được những xu hướng mua hàng hiện tại của những người dùng này?

Với các doanh nghiệp lớn, họ luôn làm công việc thu thập thông tin khách hàng để phân tích, tìm ra xu hướng của khách hàng. Công việc này hay còn gọi là "khai thác dữ liệu" đã được xem là một môn khoa học để ứng dụng có nhiều ngành nghề khác nhau. Khai thác dữ liệu không những là một hướng nghiên cứu lớn của ngành khoa học máy tính mà còn là một giải pháp hữu ích cho nhiều doanh nghiệp. Ứng dụng của khai thác dữ liệu rất đa dạng từ tiếp thị, phân tích hành vi sử dụng của con người thông qua việc mua hàng để đưa ra những quyết định kinh doanh phù hợp…Thêm vào đó, phương pháp này có đặc điểm phù hợp với lượng dữ liệu tĩnh, cố định giải quyết vấn đề trong một lĩnh vực mà thông tin trên mạng xã hội thì luôn luôn biến đổi. Với sự thay đổi về thông tin như thế, hệ thống sẽ thực hiện khai thác dữ liệu như thế nào để giới thiệu bán hàng một cách hiệu quả hơn?

Sau khi nghiên cứu khai thác dữ liệu dựa trên lý thuyết khoa học, chúng tôi chọn giải pháp là kết hợp phương pháp gom cụm và phương pháp cây quyết định để giải quyết bài toán đã đặt ra. Với phương pháp gom cụm, kết quả trả về là thông tin các cụm khách hàng với các loại sản phẩm tương ứng. Phương pháp cây quyết định lấy kết quả này để phân tích và xác định xu hướng cụ thể tương ứng với từng cụm này, từ đó đưa ra gợi ý thông tin mua hàng phù hợp.

Chúng tôi đã thực hiện giải pháp này trên hệ thống giới thiệu bán hàng trực tuyến face4shop.com với thông tin người dùng từ mạng xã hội facebook. Sau khi thử nghiệm độ chính xác của phương pháp kết hợp trên bằng cách sử dụng dữ liệu ngẫu nhiên có giả định một xu hướng mua hàng cụ thể, kết quả thu được là 81% dự đoán chính xác nhu cầu của khách hàng. Với kết quả nghiên cứu này, giúp chúng tôi có nền tảng nghiên cứu khai thác dữ liệu, có cơ hội phát triển sâu ứng dụng về sau nhằm giúp đỡ việc mua hàng trên mạng xã hội.

2. MỤC LỤC

3. LỜI CẢM ƠN

Trong mười bốn tuần thực hiện đồ án này tại Trường Đại học Hoa Sen từ ngày 13/09/2010 đến ngày 18/12/2010, chúng tôi đã được Thầy Trần Vũ Bình – Trưởng khoa Khoa Học và Công nghệ trực tiếp hướng dẫn chúng tôi trong thời gian thực hiện đề tài, đưa ra những yêu cầu cụ thể, định hướng rõ ràng và tạo điều kiện tốt cho chúng tôi thực hiện đề tài này.

Bên cạnh đó, chúng tôi gửi lời cảm ơn Thầy Nguyễn Kim Long người đã nhiệt tình truyền đạt kiến thức về kỹ thuật. Thời gian vừa qua đã giúp nhóm chúng tôi ứng dụng những kiến thức đã học, đồng thời có cơ hội tìm hiểu những nét mới về lĩnh vực khai phá dữ liệu và đạt được những kết quả nhất định.

Mong rằng các kiến thức nhận được từ trong quá trình làm khóa luận sẽ giúp ích cho chúng tôi trong công việc sau này.

Trân trọng.

Nhóm sinh viên thực hiện đề tài
Nguyễn Hoàng Phúc
Vũ Thị Quỳnh
Huỳnh Thị Minh Tâm

4. NHẬN XÉT CỦA GIẢNG VIÊN HƯỚNG DẪN

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

.....

Tp. Hồ Chí Minh, ngày tháng năm 2010

Giảng viên hướng dẫn

Tiến sĩ. Trần Vũ Bình

5. NHẬP ĐỀ

Ngày nay với sự phát triển của Internet, bán hàng trực tuyến qua mạng xã hội ngày càng có nhiều người tham gia. Câu hỏi đặt ra là ai cũng có thể bán được, nhưng làm thể nào để bán được nhiều hơn, hiệu quả hơn. Nhiều doanh nghiệp chọn cách đặt banner quảng cáo trên các trang web khác để mọi người chú ý, mục đích là để nhiều người chú ý đến thương hiệu, sản phẩm của mình. Nhiều người chú ý đến hệ thống bán hàng trực tuyến thì khả năng mua hàng sẽ cao hơn. Tuy nhiên, hệ thống có thật sự giữ chân được khách hàng hay không hay khách hàng chỉ mua một lần duy nhất.

Đặt vấn đề

- ***Với phương pháp kinh doanh truyền thống***: các doanh nghiệp lớn thường phân tích dữ liệu bán hàng hoặc thông qua một công ty chuyên làm nghiên cứu thị trường để biết được nhu cầu của khách hàng về sản phẩm của doanh nghiệp. Những thông tin đã thu thập sẽ được phân tích, thống kê, đánh giá để doanh nghiệp có một cái nhìn tổng quan về mối liên hệ giữa doanh nghiệp, khách hàng và những nhu cầu sản phẩm thực sự của khách hàng để đưa ra những chiến lược kinh doanh phù hợp. Từ đó nhận thấy rằng việc khai thác dữ liệu là một công việc quan trọng trong chiến lược phát triển kinh doanh của từng doanh nghiệp.

- ***Với xu hướng bán hàng trực tuyến***: thông tin khách hàng cũng là nguồn dữ liệu quan trọng cho việc phân tích, thống kê vá đánh giá, nhưng làm thế nào để thu thập được thông tin cần thiết của khách hàng giúp ích cho việc bán hàng? Câu trả lời chính là ***mạng xã hội. Mạng xã hội là nơi kết nối của nhiều người khác nhau cùng chung các sở thích dựa trên mối quan hệ.*** Mạng xã hội tồn tại vô số thông tin, mỗi thông tin có những ý nghĩa khác nhau, thậm chí có nhiều cách kết hợp các thông tin lại với nhau và tạo ra nhiều ý nghĩa khác nhau. Điều này cho thấy ***mạng xã hội là nền tảng dữ liệu tốt để thực hiện khai thác dữ liệu.***

  Ví dụ: mối quan hệ giữa những người trong một cộng đồng mạng xã hội, mối quan hệ này ban đầu có thể chỉ là hai người xa lạ, nhưng khi hai người có chung một sở thích về thể thao thì hai người có thể là bạn của nhau, hoặc có thể thông qua một người bạn

trung gian khác, mối quan hệ gia đình… Từ những sở thích đó và họ có mua hàng trực tuyến, chúng ta sẽ khai thác thông tin gì đối với những người thích thể thao? Họ có xu hướng mua loại sản phẩm gì? Và sản phẩm có thuộc tính như thế nào?

Với các vấn đề được đặt ra, chúng tôi chọn đề tài "Xây dựng hệ thống giới thiệu bán hàng trên mạng xã hội" với mong muốn giải quyết được bài toán giới thiệu bán hàng trên cộng đồng mạng xã hội. Bài toán cụ thể được đặt ra là làm thế nào gợi ý tặng quà sinh nhật cho những người có mối quan hệ với khách hàng trên mạng xã hội.

Phương pháp thực hiện

Các phương pháp phân tích dữ liệu kinh doanh hiện tại (thống kê, khảo sát..) không thể giải quyết bài toán có độ phức tạp thông tin mua bán hàng qua mạng xã hội đang gặp phải. Vì vậy, việc ứng dụng khai thác dữ liệu vào bài toán tiếp thị sản phẩm là giải pháp được chúng tôi quan tâm.

Có nhiều phương pháp thực hiện khai thác dữ liệu để thực hiện *nhưng mỗi phương pháp có ưu khuyết điểm khác nhau khi giải quyết một vấn đề cụ thể.* Sau quá trình tìm hiểu cách phương pháp nghiên cứu lý thuyết khoa học, chúng tôi chọn ra hai phương pháp tương đối hiệu quả, phù hợp với các công việc mà bài toán chúng tôi cần giải quyết, đó là sự kết hợp giữa phương pháp gom cụm và cây quyết định.

Lý do lựa chọn phương pháp gom cụm là phương pháp thực hiện đầu tiên

- Về đặc điểm, phương pháp gom cụm là tìm những điểm giống nhau của các tập thuộc tính đầu vào và gom lại thành nhiều cụm có dữ liệu đồng nhất. Đồng thời làm giảm độ phức tạp của dữ liệu khi thực hiện tiếp thuật toán khác. Còn cây quyết định giúp phân loại và đưa ra dự đoán về khả năng mua hay không mua sản phẩm của khách hàng.

- Về cách thực hiện, thực hiện gom cụm với số thuộc tính ít hơn, không đi sâu vào chi tiết từng thuộc tính sản phẩm để gom cụm khách hàng mà chỉ dùng phương pháp gom cụm để tạo ra các cụm khách hàng và những loại sản phẩm tương ứng mà khách hàng đã mua. Từ đó, tạo dữ liệu huấn luyện của từng cụm và đi sâu vào thông tin chi tiết của các mặt hàng trong cụm để thực hiện khai thác dữ liệu bằng phương pháp cây

quyết định, dữ liệu lúc này đã được chia nhỏ ra nên sự chênh lệch giữa khả năng mua và không mua thấp hơn so với việc chỉ thực hiện cây quyết định.

Lý do kết hợp hai phương pháp khi thực hiện

- Với phương pháp gom cụm, mục đích của thuật toán này là tìm ra các cụm dữ liệu có đặc điểm chung của từng cụm và có thể dự đoán khách hàng mua cái gì trong từng cụm riêng biệt. Dữ liệu đầu vào bao gồm thông tin khách hàng và thông tin sản phẩm khách hàng đã mua. Vấn đề của phương pháp này là các thuộc tính đầu vào phải có số lượng giá trị/trạng thái của từng thuộc tính nhất định nhưng theo thời gian thì số lượng giá trị/trạng thái thuộc tính sẽ tăng dần dẫn đến kết quả dự đoán các cụm không cao.

- Với phương pháp cây quyết định, nhiệm vụ của thuật toán này là dự đoán. Dự đoán khả năng mua/không mua của một cụm khách hàng. Nếu số thuộc tính tham gia dữ liệu đầu vào càng lớn (không thực hiện gom cụm khách hàng trước), giá trị bên trong từng thuộc tính rời rạc, không có sự lặp lại, tỷ lệ giữa khả năng mua và không mua chênh lệch quá lớn sẽ dẫn đến kết quả bị chia nhỏ ra và việc dự đoán chính xác sẽ thấp.

Như vậy, chúng tôi dựa vào nhiệm vụ công việc mà từng thuật toán đó có thể giải quyết được, so sánh với các thuật toán khác để chọn ra phương pháp kết hợp hai thuật toán lại với nhau để làm giải pháp thực hiện. Hy vọng với giải pháp đưa ra sẽ giúp sự tương tác giữa doanh nghiệp bán hàng trên cộng đồng mạng xã hội và nhu cầu cụ thể của khách hàng ngày càng gần nhau hơn dựa trên kết quả và đánh giá của giải pháp mà chúng tôi đã thực hiện.

6. TỔNG QUAN VỀ MẠNG XÃ HỘI VÀ KHAI PHÁ DỮ LIỆU

6.1 Mạng xã hội

6.1.1 Nguồn gốc mạng xã hội ra đời

Mở đầu cho trào lưu cộng đồng mạng ảo phải kể đến forum, rồi tới blog, đó chính là những tiền đề cho mạng xã hội, có thể tóm tắt sơ lược như sau:

- Forum là một diễn đàn trực tuyến nơi mà mọi người tự do thảo luận về một chủ đề nào đó trong khuôn khổ cho phép của diễn đàn. Điển hình của Forum là phân chia được nội dung. Nếu phân chia các loại nội dung từ, người dùng sẽ dễ dàng nắm bắt được các bài viết…Một điểm thuận lợi là người dùng cá nhân cũng có thể tạo cho mình một forum và thu hút những người khác quan tâm. Nhưng có một vấn đề là khi có một bài viết mới nhưng lại không nằm trong bất kỳ phân loại nào hết thì sẽ gây bối rối cho người quản trị lẫn người viết.
- Blog là nhật ký cá nhân dùng để đăng tải nội dung ngắn của một cá nhân riêng biệt. Điển hình blog đơn giản chỉ là một website nơi cá nhân chia sẽ ý kiến, đánh giá, kể chuyện sự kiện của riêng họ. Ngày nay Blog còn sử dụng trong nhiều lĩnh vực khác nhau, các tổ chức, cá thể, một tờ báo có thể dùng nó như một công cụ để thăm dò ý kiến, tạo ra kênh giao tiếp với thành phần bên ngoài. Tag là một khái niệm quen thuộc của người viết blog, một bài viết có thể được gắn nhiều tag theo nhiều chủ để khác nhau mà người viết muốn đưa vào. Ví dụ một bài viết liên quan đến âm nhạc có thể được gắn tag là Âm nhạc, và có thể tag là tên của một ca sĩ nào đó có trong bài viết. Với ưu điểm đó blog trở nên thịnh hành hơn forum, nhưng khi blog phát triển thì vấn đề quản lý và phân chia nội dung trở nên khó khăn.

Bảng 1- Một vài đặc điểm giúp phân biệt giữa forum và blog

	Forum	**Blog**
Nguồn gốc	Nhu cầu trao đổi thông tin về một lĩnh vực, vấn đề quan tâm.	Nhu cầu chia sẻ thông tin mang tính chất cá nhân.
Quản lý/điều hành	Quản trị.	Cá nhân.

Nội dung	Phân chia nội dung theo các tiêu chí đã được quy định trước.	Người dùng tự quy định loại nội dung của bài viết và có thể gắn nhiều phân loại cho cùng một bài viết.
Tổ chức / Sắp xếp	Theo trình tự được thiếp lập một cách cụ thể.	Theo trình tự thời gian.
Cập nhật	Câu trả lời/đối thoại chưa chắc được cập nhật thường xuyên.	Cập nhật thường xuyên do tác giả.

Forum và blog đều giúp mọi người chia sẻ thông tin cho nhau mỗi cái đều có ưu khuyết điểm khác nhau. Forum có thể phân loại thông tin rõ ràng nhưng lại không đáp ứng được nhu cầu của từng cá nhân cụ thể khi viết bài. Blog đáp ứng nhu cầu cá nhân của mỗi người nhưng lại không có tính riêng tư. Mạng xã hội ra đời đã giải quyết phần nào những vướng mắc mà các forum và blog gặp phải.

6.1.2 Giới thiệu mạng xã hội (1)

Mạng xã hội là dịch vụ nối kết các thành viên cùng sở thích trên Internet lại với nhau. Các thành viên sử dụng các tính năng như chat, e-mail, phim ảnh, voice chat, chia sẻ file, blog và xã luận. Họ liên kết với nhau và tạo nên xã hội ảo với hàng trăm triệu thành viên khắp thế giới.

Mạng xã hội cung cấp nhiều cách thức để các thành viên tìm kiếm bạn bè, đối tác: dựa theo group (ví dụ như tên trường hoặc tên thành phố), dựa trên thông tin cá nhân (như địa chỉ e-mail hoặc screen name), hoặc dựa trên sở thích cá nhân (như thể thao, phim ảnh, sách báo, hoặc ca nhạc), lĩnh vực quan tâm: kinh doanh, mua bán.

Mạng xã hội được xây dựng theo các mục tiêu sau:

- Tạo ra một hệ thống ảo trên nền Internet cho phép người dùng giao lưu và chia sẻ thông tin một cách có hiệu quả, vượt ra ngoài những giới hạn về địa lý và thời gian.

- Xây dựng lên một mẫu định danh trực tuyến nhằm phục vụ những yêu cầu công cộng chung và những giá trị của cộng đồng.

- Nâng cao vai trò của mỗi công dân trong việc tạo lập quan hệ và tự tổ chức xoay quanh những mối quan tâm chung trong những cộng đồng thúc đẩy sự liên kết các tổ chức xã hội.

Mạng xã hội tiêu biểu "Facebook": Điểm tiêu biểu của mạng xã hội Facebook là xây dựng trên nền tảng mối quan hệ cá nhân, tổ chức và gia đình và thiết lập tính riêng tư về nội dung của cá nhân theo những mối quan hệ này. Ví dụ như tôi thích sản phẩm của trang web này, nhưng tôi chỉ muốn bạn bè của tôi biết điều này, Facebook làm được điều đó. Đó là điểm mà blog không giải quyết được. Mạng xã hội Facebook hiện có hơn 500 triệu thành viên tích cực trên khắp thế giới (2). Với con số ấy, Facebook là mạng xã hội phổ biến nhất hiện nay, tiếp theo sau là MySpace và Twitter (3).

Tiếp thị qua mạng xã hội:

- Với số lượng thành viên lớn, việc kinh doanh qua mạng xã hội đang được nhiều doanh nghiệp nhắm đến trong việc quảng bá thương hiệu và tiếp thị sản phẩm. Theo thống kê, với cách quảng cáo truyền thống để đạt 500 triệu người sử dụng, Radio mất 38 năm, Ti vi mất 13 năm, Internet mất 4 năm, Ipod mất 3 năm, còn Facebook - mạng xã hội được nhiều người dùng nhất hiện nay đã đạt 100 triệu người dùng chỉ sau chưa đầy 9 tháng.

- Thực tế theo kết quả khảo sát của Cty Nielsen (Mỹ) cho thấy, chỉ có 14% số người tin vào quảng cáo trên các phương tiện truyền thông và gần 80% người xem truyền hình chuyển kênh khi tới phần quảng cáo. Tại Việt Nam, doanh nghiệp muốn có 20-30 giây quảng cáo trên truyền hình vào giờ vàng (19h40 – 20h10) thì phải bỏ ra từ 30-55 triệu đồng.

Như vậy thực hiện chiến dịch marketing trên mạng xã hội, doanh nghiệp sẽ có thể làm nhiều hơn chứ không chỉ là một tiếp thị sản phẩm trong thời gian ngắn ngủi.

6.1.3 Cấu trúc mạng xã hội (1)

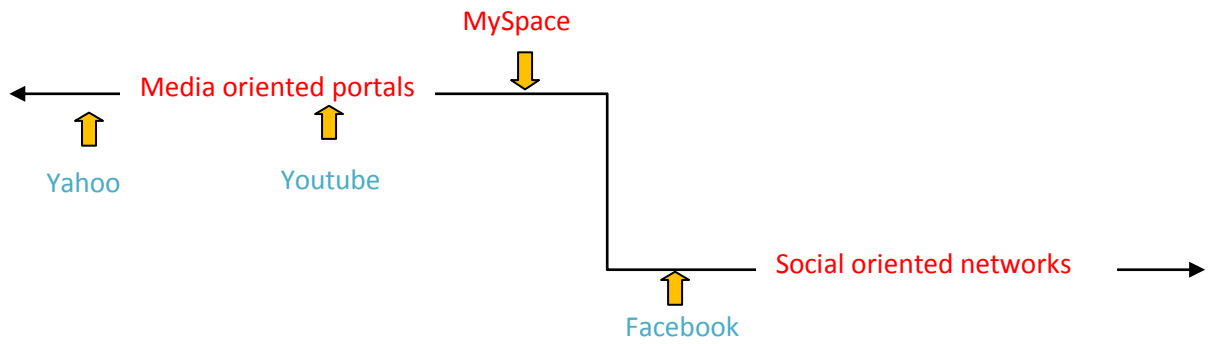
Một số thuật ngữ liên quan:

- Nút (node): Là một thực thể trong mạng. Thực thể này có thể là một cá nhân, một doanh nghiệp hoặc một tổ chức bất kỳ.
- Liên kết (tie): là mối quan hệ giữa các thực thể. Trong mạng có thể có nhiều kiểu liên kết. Ở dạng đơn giản nhất, mạng xã hội là một đơn đồ thị vô hướng các mối liên kết phù hợp giữa các nút. Ta có thể biểu diễn mạng liên kết này bằng một biểu đồ mà các nút được biểu diễn bởi các điểm còn các liên kết được biểu diễn bởi các đoạn thẳng.

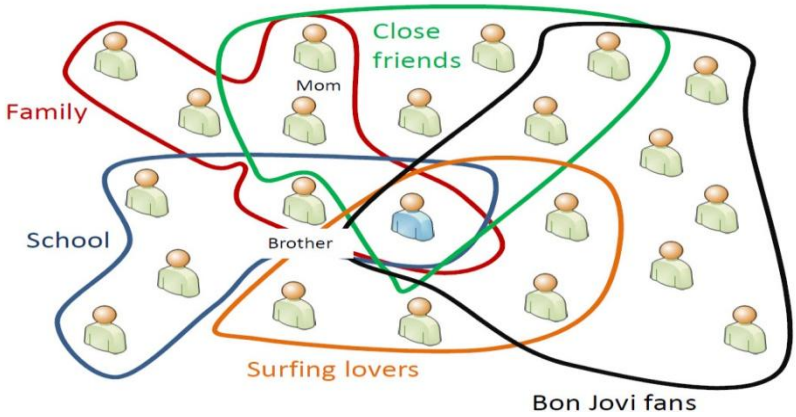
So sánh mô hình các mạng xã hội

Facebook	**MySpace**
Mô hình quan hệ xã hội: các thành viên nối kết với nhau dựa trên mối quan hệ bạn bè, gia đình, đồng nghiệp,..	Mô hình gom nhóm: các thành viên nối kết với nhau dựa trên sở thích.

Bảng 2- Mô hình mạng xã hội facebook và myspace



Hình 1 – Khuynh hướng mạng xã hội



Hình 2 – Cấu trúc mạng xã hội – liên kết các thành phần quan hệ trong Facebook

6.1.4 Phân loại quan hệ của mạng xã hội

Phân loại quan hệ của mạng xã hội tiêu biểu facebook và MySpace

Facebook	**MySpace**
Liên kết quan hệ dựa vào thuộc tính, khoảng cách địa lý, nơi ở, trường học, đồng nghiệp… Liên kết mối quan hệ chặt chẽ với nhau: một mối quan hệ mới và duy trì với quan hệ cũ Liên kết dàn trải rộng khắp: xu hướng, hoạt động, sự kiện, thể thao, ca sĩ, phim yêu thích… Duy trì tất cả mối quan hệ: bạn bè, gia đình… Facebook tự tìm ra các mối liên hệ trung gian để tăng sự liên kết giữa các thành viên (ví dụ như hai người không có mối quan hệ bạn bè nhưng lại có chung một người bạn vẫn có thể trở thành bạn của nhau)	Liên kết quan hệ bạn bè thông thường Liên kết trong lĩnh vực giải trí: nhạc, phim… Duy trì mối quan hệ: bạn bè, nghệ sĩ với các fan

Bảng 3- So sánh phân loại quan hệ xã hội facebook và myspace

Hình 3 – Cấu trúc mạng xã hội – thành phần với một vài giá trị thuộc tính: sở thích, thói quen…

6.1.5 Khả năng tiếp cận hệ thống mạng xã hội từ bên ngoài

Dựa vào Open Platform: nền tảng mở giúp nhà phát triển phát triển ứng dụng.

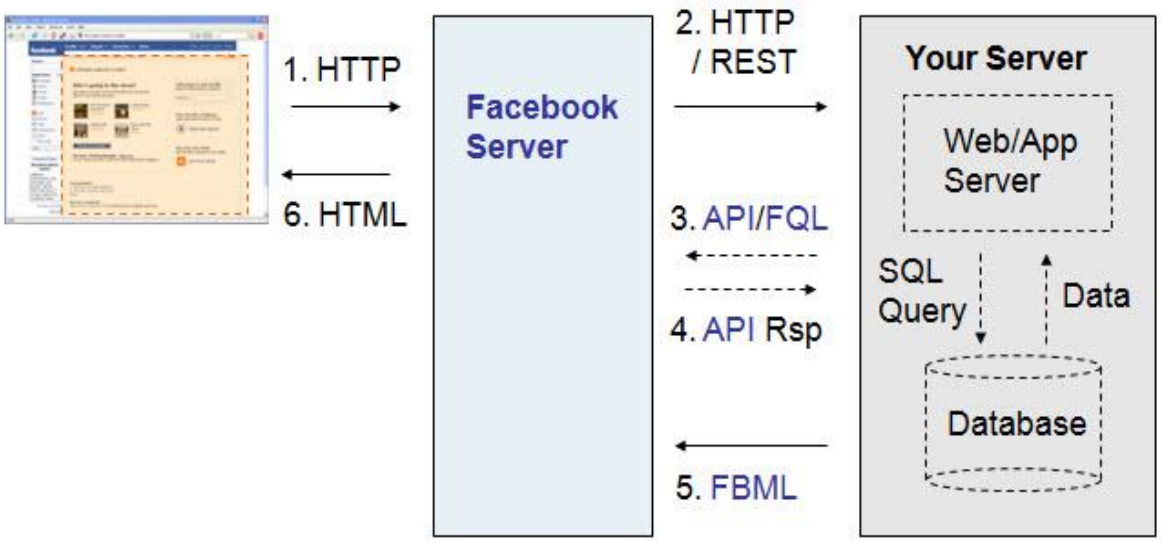
Ví dụ: ZYNGA - http://www.zynga.com/ hay JIBJAB - http://sendables.jibjab.com/

Facebook	**MySpace**	**Yahoo**
Open flatform. Ứng dụng ngoài được phát triển mạnh mẽ tiếp cận thông qua API. Cung cấp đầy đủ hàm, thư viện lấy thông tin của người dùng.	Open flatform. Chỉ có ứng dụng bên trong được xây dựng.	Close flatform.

Bảng 4- So sánh khả nănng tiếp cận hệ thống mạng từ bên ngoài

Facebook	MySpace
API giúp ứng dụng lấy được thông tin người dùng phục vụ cho đánh giá/kiểm tra cho mục đích cụ thể của nhà phát triển. Ví dụ: marketing, data mining.	Chỉ tiếp cận với người dùng trong hệ thống, tương tác giữa các thành viên – nghệ sĩ, nhà tổ chức. Ví dụ: Giúp các nghệ sĩ, nhà tổ chức định hình xu hướng, phong cách thưởng thức âm nhạc.

Bảng 5- So sánh Hệ thống Facebook và MySpace tiếp cận với hệ thống bên ngoài qua API



Hình 4 – Quy trình giao tiếp giữa facebook API với ứng dụng bên ngoài

6.1.6 Kết luận

Forum và Blog hoàn toàn có thể đáp ứng nhu cầu mua hàng, xem thông tin sản phẩm nhưng điểm bất lợi là từ đây ta không thể tìm ra được xu hướng, thói quen mua hàng của người dùng vì hai loại hình này chỉ dừng ở mức cung cấp thông tin.

Với mạng xã hội, đặc điểm của cộng đồng này là sự hình thành dựa trên tiêu chí là cùng chung một sở thích nào đó và ta có thể lấy được thông tin này cũng như những đối tượng liên quan, đây là điều quan trọng trong việc phân tích xu hướng của người dùng. Vì thế chúng tôi thấy mạng xã hội là có đầy đủ điều kiện để phục vụ cho việc khai thác dữ liệu cho bài toán bán hàng trực tuyến.

MySpace và Facebook là hai trong số những mạng xã hội chúng tôi chọn để thực hiện khai thác dữ liệu. Tuy nhiên, MySpace phân nhóm khách hàng dựa trên sở thích cá nhân như nhóm thành viên yêu thích âm nhạc, nghệ thuật nói chung, đối tượng trẻ là chủ yếu và không bao quát hết các thành phần ở nhiều độ tuổi khác nhau cũng như thuộc tính của từng đối tượng khách hàng. Ngược lại, Facebook lại có đầy đủ những thành phần cũng như thuộc tính của khách hàng, điểm đặc biệt là mối quan hệ trong Facebook: gia đình, bạn bè, đồng nghiệp, … mà MySpace còn thiếu nên mạng xã hội Facebook thích với yêu cầu đề tài của chúng tôi.

Bên cạnh đó, điểm mạnh của Facebook cũng là khó khăn của những người muốn kinh doanh trên mạng xã hội vì không xác định được xu hướng của nhiều nhóm đối tượng cụ thể và sở thích tương ứng. Với đề tài "Xây dựng hệ thống giới thiệu bán hàng trên mạng xã hội", chúng tôi chọn mạng xã hội Facebook với mong muốn khai thác nguồn dữ liệu phong phú này, từ đó tìm ra, gom nhóm lại xu hướng mua hàng của nhiều nhóm đối tượng liên quan đến một hệ thống bán hàng trực tuyến.

6.2 Khai phá dữ liệu

6.2.1 Nguồn gốc hình thành (4)

- 1960s: Hê thống xử lý tập tin đơn giản. Hệ thống cơ sở dữ liệu.
- 1970s: Cơ sở dữ liệu quan hệ, mô hình hóa, câu truy vấn, ...
- 1980s: Lý thuyết mô hình hướng đối tượng, CSDL phân tán, ...

Các số liệu cho thấy trong khoảng hai thập niên, công nghệ phát triển, dữ liệu cũng được quản lý, phân tích dưới nhiều hình thức, công nghệ khác nhau. Điều này dẫn tới sự bùng nổ kho dữ liệu khổng lồ với dạng "giàu dữ liệu, nghèo thông tin". Mục đích của khai thác dữ liệu là nhằm phát hiện ra các thông tin có giá trị tiềm ẩn trong các tập dữ liệu lớn.

Như vậy, đối với việc kinh doanh sản phẩm. Việc khai phá dữ liệu trên lượng đơn hàng là rất quan trọng. Từ đó, doanh nghiệp có thể khám phá ra các đối tượng khách hàng của mình họ cần gì? Tại sao mặc hàng này lại bán chạy, mặc hàng kia lại không bán được? Khi một sản phẩm mới được tung ra, họ sẽ bắt đầu tiếp thị loại đối tượng khách hàng như thế nào? Hàng loạt các câu hỏi được đặt ra đối với chiến lược kinh doanh của các doanh nghiệp. Chính vì thế khai phá dữ liệu đang trở thành một hướng nghiên cứu mới trong lĩnh vực khoa học máy tính và công nghệ tri thức.

6.2.2 Chức năng của khai phá dữ liệu

Khai phá dữ liệu có nhiệm vụ mô tả hay dự đoán tùy thuộc vào quá trình khai phá.(5)

- ***Mô tả (Descriptive):*** mô tả đặc trưng các thuộc tính chung của dữ liệu được khai phá.
  Ví dụ: Người đang sử dụng thẻ ID = 1234 thật sự là chủ nhân của thẻ hay là một tên trộm?
  Loại chức năng sử dụng: Phân nhóm (Clustering), Kết hợp (Association), Phân tích trình tự (Sequence Analysis).
- ***Dự đoán (predictive):*** có khả năng suy luận từ dữ liệu hiện có để dự đoán
  Ví dụ: Ông A (Tid = 100) có khả năng trốn thuế??? Ngày mai cổ phiếu HSBC sẽ tăng??? Làm sao xác định được khả năng tốt nghiệp của một sinh viên hiện tại?
  Loại chức năng sử dụng: Phân loại (Classification), Hồi quy (Regression), Phân tích độ lệch (Deviation Analysis).

6.2.3 Mô tả chi tiết chức năng

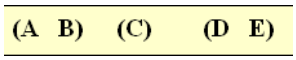
- **Clustering (D):** cho 1 tập các điểm dữ liệu (data points) với các thuộc tính và 1 đơn vị tương đương (similarity measure), tìm các nhóm sao cho:
  - Dữ liệu trong cùng 1 nhóm (cluster) thì giống nhau hơn nhóm khác.
  - Dữ liệu trong các nhóm khác nhau thì ít giống nhau hơn.

  Ứng dụng:
  - Phân loại tài liệu text (phân loại theo giải trí, học tập, …)
  - Thăm dò kết quả mua bán hằng ngày, nhận thấy người mua bia thường mua khoai tây, vì thế họ sẽ đặt 2 sản phẩm cạnh nhau để tăng doanh thu.

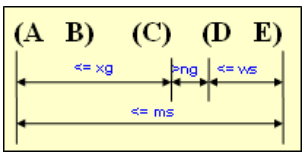
- **Association (D):** Cho trước 1 nhóm records có chứa một số các items từ 1 bộ nhất định. Thiết lập các luật phụ thuộc để mô tả sự xuất hiện của 1 số item dựa trên những item khác.

  Ứng dụng: Trong tiếp thị/quảng cáo, khi khách hàng mua máy tính, gợi ý khách hàng mua thêm sản phẩm khác (chuột, bàn phím, máy in) dựa vào thông tin sản phẩm được mua ở những lần trước đó.

- **Sequence Analysis (D):** Cho trước 1 tập các đối tượng, trong đó mỗi đối tượng có riêng 1 chuỗi các sự kiện. Cần tìm các quy luật dự đoán sự phụ thuộc tuần tự giữa các sự kiện.

  Các quy luật được lập nên bằng cách đầu tiên tìm ra các kiểu mẫu. Các sự kiện xảy ra trong các mẫu này bị giới hạn về mặt thời gian.

- **Classification (P):** xác định mô hình/mẫu nhằm mô tả các lớp quan trọng hay dự đoán khuynh hướng dữ liệu trong tương lai.

  Cách thực hiện: sử dụng 1 tập các records có sẵn, mỗi record có chứa nhiều thuộc tính, trong đó có 1 thuộc tính là class. Thực hiện 2 bước:

  - Tập huấn luyện (training set): tập dữ liệu có sẵn đã được phân loại bằng cách dùng cây quyết định và máy học ➔ đưa ra mô hình cụ thể
  - Tập kiểm tra (test set): tập dữ liệu đưa vào dựa vào mô hình cụ thể có sẵn để xác định xem tập dữ liệu thuộc loại nào.

    Ví dụ: khi đưa vào thông tin mua hàng mới bằng credit card ta có thể xác định credit card này có bị gian lận hay không.

  Ứng dụng :

  - Có một lượng khách hàng. Dựa vào thông tin đưa ra phân loại khách hàng tiềm năng, khách hàng không tiềm năng để giảm thiểu chi phí liên lạc (gửi email, tin nhắn, gọi điện,...)

- Phát hiện gian lận trong credit card dựa vào thông tin đã được phân loại (mua gì, ở đâu, ai mua,…), giả sử mua ở nơi xa chỗ thường hay mua hàng, từ đó xác định khả năng mất thẻ của khách hàng.

- **Regression (P):** Dự đoán giá trị của 1 thông số được cho liên tục, dựa trên giá trị của những thông số khác. (Giả sử cho trước 1 mô hình phụ thuộc tuyến tính hay phi tuyến tính).

  Ứng dụng: Nhiều trong lĩnh vực thống kê, mạng lưới thần kinh.
- **Deviation Analysis (P):** Nhận ra những thay đổi khác biệt so với hành vi bình thường.

  Ứng dụng: Phát hiện xâm nhập network…

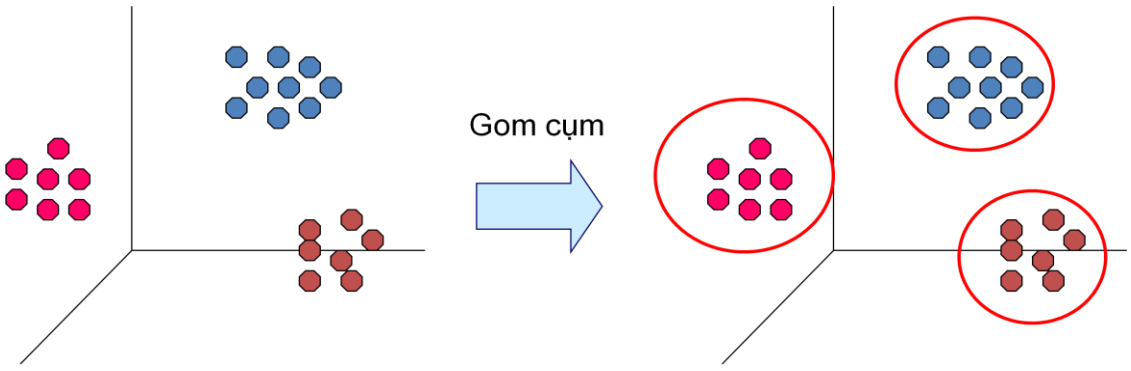
6.3 Các phương pháp trong khai phá dữ liệu

6.3.1 Phương pháp phân cụm (6)

6.3.1.1 Mô tả khái niệm

- Là quá trình gom nhóm/cụm dữ liệu/đối tượng vào các lớp/cụm.
- Các đối tượng trong cùng một cụm tương tự với nhau hơn so với đối tượng ở các cụm khác.

Ví dụ: Obj1, Obj2 ở cụm C1; Obj3 ở cụm C2 → Obj1 tương tự Obj2 hơn so với tương tự Obj3.



Hình 5 – Ví dụ về gom cụm

6.3.1.2 Mục đích của gom cụm

- Khảo sát sự phân bố của tập dữ liệu mẫu.
- Làm bước tiền xử lý cho các thuật toán khác.

6.3.1.3 Ứng dụng sử dụng

- **Tiếp thị:** khám phá các nhóm khách hàng phân biệt trong cơ sở dữ liệu mua hàng.
- **Sử dụng đất:** nhận dạng các vùng đất sử dụng giống nhau khi khảo sát CSDL.
- **Bảo hiểm:** nhận dạng các nhóm công ty có chính sách bảo hiểm mô tô với chi phí đền bù trung bình cao
- **Hoạch định thành phố:** nhận dạng các nhóm nhà cửa theo loại nhà, giá trị và vị trí địa lý.

6.3.1.4 Phương pháp phân cụm

Phương pháp phân hoạch

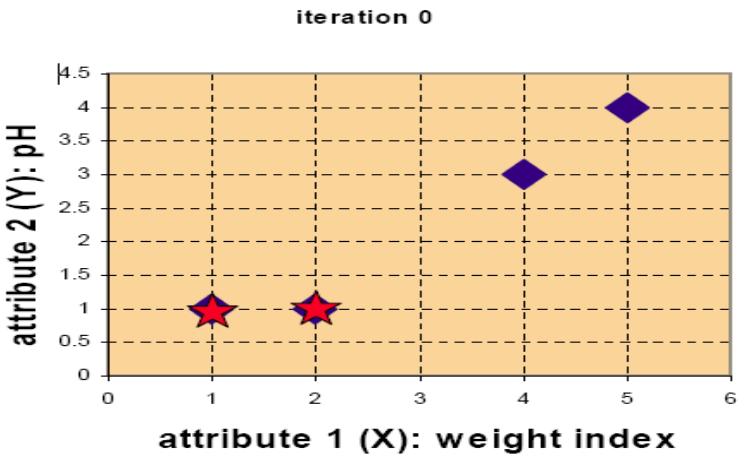
- Đầu vào:
  - Tập dữ liệu mẫu X
  - Số cụm k
  - Hàm độ đo
- Kết quả: Phân hoạch của X gồm k lớp tương đương (k cụm)
- Các bước thực hiện:
  - Bước 1: Chọn (bất kỳ) k phần tử của X làm các tâm ban đầu.
  - Bước 2: Với mỗi phần tử còn lại của X,
    - Gán phần tử này vào cụm có tâm gần nhất.
    - Tính toán lại tâm của các cụm.
  - Bước 3: Nếu không có sự thay đổi nào của các tâm → Quay lại Bước 2.
- Ví dụ minh họa: Khảo sát bảng dữ liệu về nồng độ pH và trọng lượng của một số loại thuốc. Ta có tập huấn luyện như sau:

Object	Feature 1 (X): weight index	Feature 2 (Y): pH
Medicine A	1	1
Medicine B	2	1
Medicine C	4	3
Medicine D	5	4

Bảng 6- Mô tả tập huấn luyện phân cụm loại thuốc

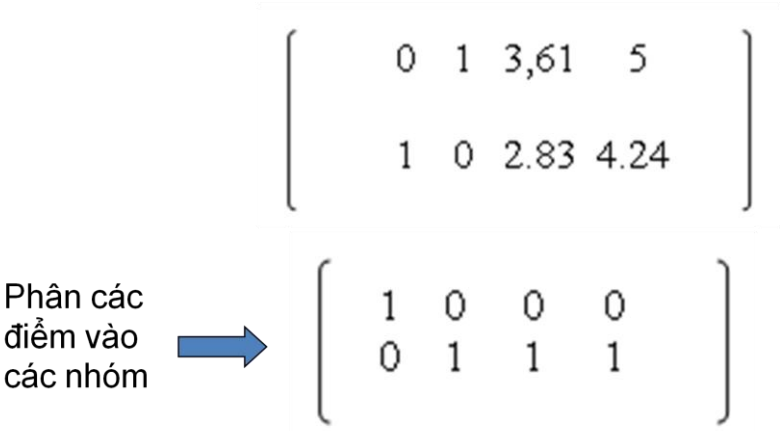
Các bước thực hiện:

Bước 1: Phân hoạch tập mẫu thành hai nhóm. Chọn ngẫu nhiên hai điểm làm tâm cho mỗi nhóm: (1,1) và (2,1).



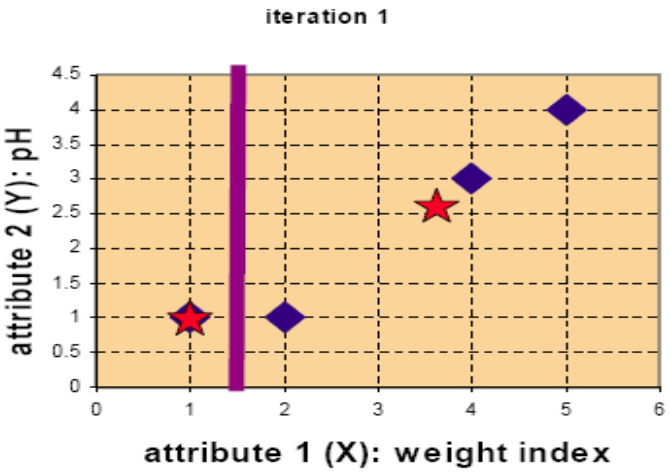
Hình 6 – Phân nhóm ngẫu nhiên

Bước 2: Sử dụng độ đo Euclide. Tính khoảng cách từ các điểm đến các tâm.



Hình 7 – Sử dụng độ đo Euclide

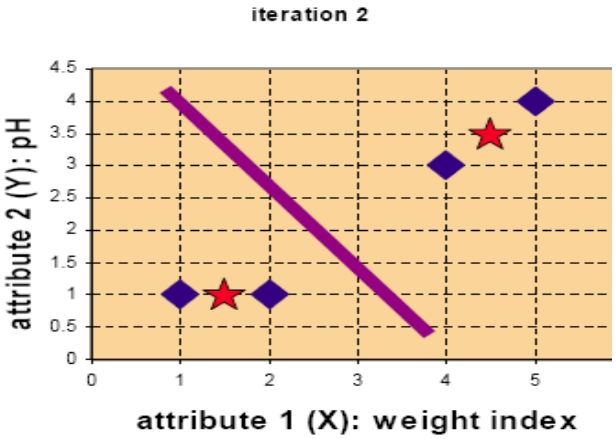
Bước 3: Tính toán lại tâm của các cụm, theo trung bình tọa độ.



Hình 8 – Tính lại tâm của các cụm

Bước 4: Có sự thay đổi các điểm giữa các nhóm, nên lặp lại bước 2.

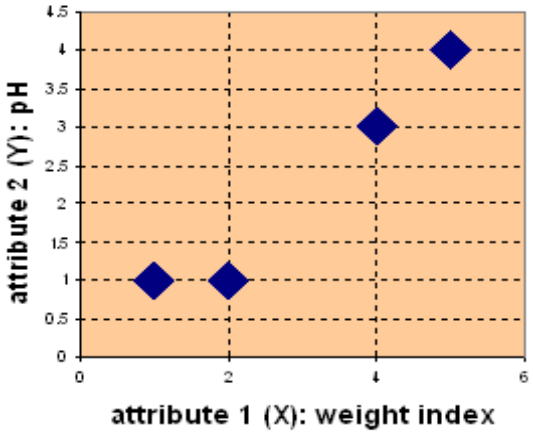
0	1	3,61	5
3.14	2.36	0.47	1.89
1	1	0	0
0	0	1	1



Hình 9 – Thay đổi các điểm giữa các nhóm lặp lần 1

Có sự thay đổi các điểm giữa các nhóm, nên lặp lại bước 2.

0.5	0.5	3,20	4.61
4.30	3.54	0.71	0.71
1	1	0	0
0	0	1	1



Hình 10 – Thay đổi các điểm giữa các nhóm lần 2

Bước 5: Không có sự thay đổi của các nhóm → Kết thúc thuật toán.

Object	Feature 1 (X): weight index	Feature 2 (Y): pH	Group (result)
Medicine A	1	1	1
Medicine B	2	1	1
Medicine C	4	3	2
Medicine D	5	4	2

Bảng 7- Kết quả phân cụm

Phương pháp phân cấp

- Đầu vào:
  - Tập dữ liệu mẫu X
  - Số cụm k
  - Hàm độ đo
- Kết quả: Phân hoạch của X gồm k lớp tương đương (k cụm)
- Các bước thực hiện
  - Bước 1: Gán N phần tử vào N nhóm.
  - Bước 2: Xây dựng ma trận khoảng cách M giữa các nhóm.
  - Bước 3: Tìm cặp nhóm gần nhau nhất, gom hai nhóm này lại thành một.
  - Bước 4: Nếu số nhóm đạt đến k → Kết thúc thuật toán, ngược lại, → Quay về B2.
- Ví dụ minh họa: Khảo sát khoảng cách giữa sáu thành phố, thực hiện phân nhóm với k =2

Hình 11 – Bản đồ khoảng cách giữa các thành phố

Các bước thực hiện:

Bước 1: Xây dựng bảng khoảng cách giữa các cụm

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Hình 12 – Hình khoảng cách giữa các cụm lần 1

Bước 2: Cặp nhóm ngắn nhất là d (1, 6) = 138, gom nhóm 1 và 6 thành một nhóm. Số nhóm hiện tại là 5.

Hình 13 – Gom cụm theo khoảng cách, số nhóm 5

Bước 3: Xây dựng lại bảng khoảng cách giữa các nhóm.

	BA	FI	MI/TO	NA	RM
BA	0	662	877	255	412
FI	662	0	295	468	268
MI/TO	877	295	0	754	564
NA	255	468	754	0	219
RM	412	268	564	219	0

Hình 14 – Bảng khoảng cách giữa các nhóm sau khi xây dựng lại lần 2

Cặp nhóm ngắn nhất là d (1, 4) = 255, gom nhóm 1 và 4 thành một nhóm. Số nhóm hiện tại là 3

Hình 15 – Gom cụm theo khoảng cách, số nhóm 3

Xây dựng lại bảng khoảng cách giữa các nhóm.

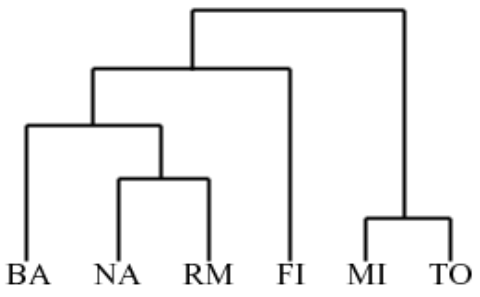
	BA/NA/RM	FI	MI/TO
BA/NA/RM	0	268	564
FI	268	0	295
MI/TO	564	295	0

Hình 16 – Bảng khoảng cách giữa các nhóm sau khi xây dựng lại lần 3

Cặp nhóm ngắn nhất là d (1, 2) = 268, gom nhóm 1 và 2 thành một nhóm. Số nhóm hiện tại là 2

Hình 17 – Gom cụm theo khoảng cách, số nhóm 2

Số nhóm đã bằng 2, nên kết thúc thuật toán.



Hình 18 – Gom cụm kết thúc

6.3.2 Phương pháp phân lớp dựa vào cây quyết định

6.3.2.1 Mô tả khái niệm

Là một kiểu mô hình dự báo (predictive model): từ các quan sát về một sự vật/hiện tượng. Từ đó, có thể kết luận về giá trị mục tiêu của sự vật/hiện tượng đó.

Cây quyết định là cấu trúc cây sao cho:

- Mỗi nút trong ứng với một phép kiểm tra trên một thuộc tính.
- Mỗi nhánh biểu diễn kết quả phép kiểm tra.
- Các nút lá biểu diễn các lớp hay các phân bố lớp
- Nút cao nhất trong cây là nút gốc.

6.3.2.2 Mục đích

Để xây dựng kế hoạch nhằm đạt được mục tiêu mong muốn, hỗ trợ quá trình ra quyết định. Phương tiện mô tả việc tính toán các xác suất có điều kiện (xác suất của A, biết B).

6.3.2.3 Thuật toán trên cây quyết định

Thuật toán căn bản

- Xây dựng một cây đệ quy phân chia và xác định các đặc tính từ trên xuống.
- Các đặc tính được xem là rõ rang, rời rạc
- Tham lam (có thể có tình trạng cực đại cục bộ)

Nhiều dạng khác nhau: ID3, C4.5, CART, CHAID. Điểm khác biệt phụ thuộc vào tiêu chuẩn/ thuộc tính phân chia, độ đo để chọn lựa

Các độ đo để lựa chọn thuộc tính, dựa vào:

- Độ lợi thông tin (Information Gain): chọn thuộc tính có chỉ số có độ lợi thông tin lớn nhất, dựa vào khối lượng thông tin cần thiết

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Với p là phần tử lớp P và n là phần tử lớp N cùng thuộc một tập dữ liệu S. Entropy (thông tin mong muốn cần thiết để phân lớp các đối tượng trong tất cả các cây con)

$$E(A) = \sum_{i=1}^{v} \frac{pi + ni}{p + n} I(pi, ni)$$

Thông tin có được bởi việc phân nhánh trên thuộc tính A

Gain (A) = I (p,n) – E (A)

- λ2 – số thống kê bảng ngẫu nhiên
- G – thống kê

Ví dụ minh họa: Giải quyết bài toán chơi tennis ID3, ta có 1 tập huấn luyện như sau:

Day	Outlook	Temp	Humidity	Wind	Play Tennis ?
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Bảng 8- Mô tả tập huấn luyện bài toán "Play tennis"

- Bước 1: Xác nhận đâu là Lớp P, N

  Lớp P: play tennis = "yes"

  Lớp N: play tennis = "no"

  → Thông tin cần thiết để phân lớp cho một mẫu: I (p,n) = I (9,5) = 0.940

- Bước 2: Tính Entropy cho thuộc tính "Outlook". Ta có:

Outlook	**pi**	**ni**	**I (pi, ni)**
Sunny	2	3	0.971
Overcast	4	0	0
Rain	3	2	0.971

Bảng 9- Mô tả Entropy cho thuộc tính "Outlook"

Từ đó:

$$E(\text{Outlook}) = \frac{5}{14}.I\,(2,3) + \frac{4}{14}.I\,(4,0) + \frac{5}{14}.I\,(3,2) = 0.694$$

- Bước 3: Tính Gain

Gain (Outlook) = I (9,5) – E (Outlook) = 0.246

- Bước 4: Tương tự bước 2, bước 3 cho các thuộc tính còn lại:

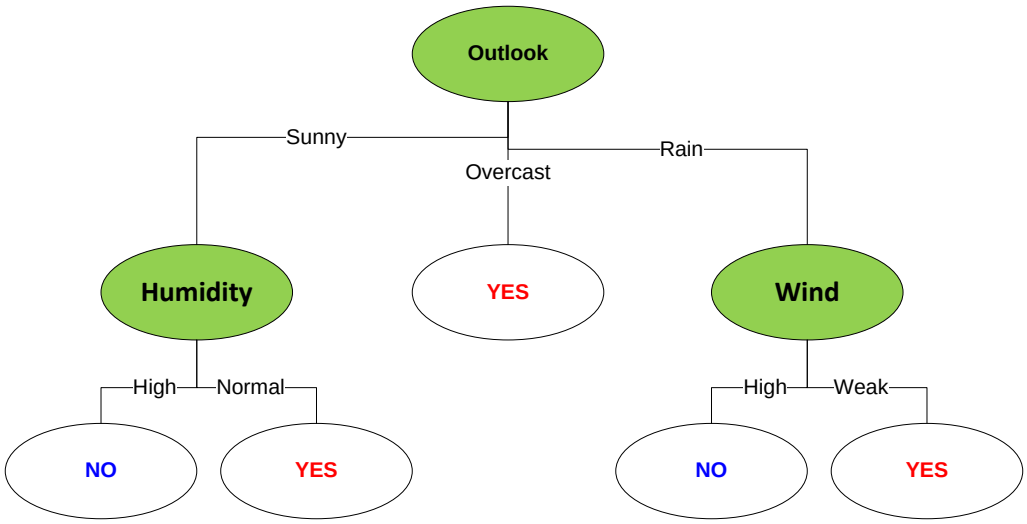
Gain (Temp) = I (9,5) – E (Temp) = 0.029

Gain (Humidity) = I (9,5) – E (Humidity) = 0.151

Gain (Wind) = I (9,5) – E (Wind) = 0.048

Ta xác nhận root đầu tiên của cây sẽ là thuộc tính Outlook vì có độ lợi thông tin lớn nhất

- Bước 5: Tiếp tục các nhánh còn lại lặp lại từ bước 1 đến bước 4 để xác định các nút lá. Từ đó ta xây dựng được một cây quyết định như sau:



Hình 19 – Cây quyết định cho bài toán Play tennis

Từ cây quyết định, ta hình thành các luật như sau:

- Quy luật 1: If Outlook = 'Over cast' then PlayTennis = 'Yes'
- Quy luật 2: If Outlook = 'Sunny' and Humidity = 'Normal' then PlayTennis = 'Yes'
- Quy luật 3: If Outlook = 'Rain' and Wind = 'Weak' then PlayTennis = 'Yes'
- Quy luật 4: If Outlook = 'Sunny' and Humidity = 'High' then PlayTennis = 'No'
- Quy luật 5: If Outlook = 'Rain' and Wind = 'High' then PlayTennis = 'Yes'

6.3.2.4 Tránh việc quá khít (Overfitting) trong cây quyết định

- Cây tạo được có thể overfit dữ liệu huấn luyện: quá nhiều nhánh, độ chính xác kém cho những mẫu chưa biết
- Lý do overfit: dữ liệu nhiễu và tách rời nhóm, dữ liệu huấn luyện quá ít, các giá trị cục bộ trong tìm kiếm tham lam (greedly search).

Cách giải quyết để tránh overfitting:

- Rút gọn trước: ngừng sớm
- Rút gọn sau: loại bỏ bớt các nhánh sau khi xây xong toàn bộ cây.

6.3.2.5 Ưu điểm

- Tốc độ nhanh hơn các phương pháp khác.

- Có thể chuyển đổi CÂY thành LUẬT phân lớp đơn giản và dễ hiểu.

Ví dụ: If Outlook = 'Over cast' then PlayTennis = 'Yes'

- Có thể dùng các truy vấn SQL phục vụ truy cập CSDL.
- Độ chính xác (accuracy) trong phân lớp có thể so sánh (phần trăm của các mẫu trong tập kiểm tra được bộ phân lớp xếp đúng lớp)

$$\text{Độ } chính\ xác = \frac{phân\ lớp\ kiểm\ tra\ đúng}{tổng\ số\ tập\ kiểm\ tra}$$

6.3.2.6 Khuyết điểm

Giới hạn là miêu tả cây và luật chỉ có thể biểu diễn với một số dạng chức năng. Từ đó, giới hạn độ chính xác của mô hình.

6.4 So sánh phương pháp khai thác dữ liệu trên MS SQL datamining server

Các mô hình khai thác có thể dự đoán các giá trị, đưa ra bảng tóm tắt dữ liệu, và tìm ra sự tương quan. Để giúp cho việc lựa chọn thuật toán cho giải pháp khai thác dữ liệu. (5)

Công việc	**Thuật toán áp dụng**
Dự đoán một thuộc tính rời rạc Ví dụ: Dự đoán người nhận thư của cuộc vận động sẽ mua sản phẩm hay không?	Thuật toán Decision Trees Thuật toán Naïve Bayes Thuật toán Clustering Thuật toán Neural Network (SSAS)
Dự đoán một thuộc tính liên tục Ví dụ: Dự đoán doanh thu năm tiếp theo.	Thuật toán Decision Trees Thuật toán Time Series
Dự đoán một trình tự. Ví dụ: Thực hiện phân tích một clickstream cho một web site của công ty.	Thuật toán Sequence Clustering
Tìm nhóm của những mục chọn (item) trong các	Thuật toán Association

các giao tác (transaction). Ví dụ: Sử dụng phân tích thị trường để đưa thêm các sản phẩm cho khách hàng	Thuật toán Decision Trees
Tìm những mục (item) giống nhau. Ví dụ: Phân chia các dữ liệu vào các nhóm để hiểu dễ hơn các mối quan hệ giữa các thuộc tính	Thuật toán Clustering Thuật toán Sequence Clustering

Bảng 10- Bảng lựa chọn thuật toán khai thác dữ liệu

6.4.1 Dự đoán một thuộc tính rời rạc

Thuộc tính rời rạc là thuộc tính:

- Là tập hợp các số hữu hạn
- Thường đại diện bởi các số nguyên.
  Ví dụ: Giới tính = {Nam, Nữ}; Tháng = {1,2,3….,11,12}

6.4.1.1 Thuật toán Decision tree

- Phục vụ việc phân loại và dự đoán.
- Dự đoán thuộc tính rời rạc của thuật toán Decision Trees giúp có cái nhìn tổng quan về những khả năng có thể xảy ra của luật để đưa ra những quyết định phù hợp.

Ví dụ: dự đoán khách hàng nào có khả năng mua xe đạp, thuật toán Decision Trees dựa trên mối quan hệ của các cột dữ liệu đầu vào của tập dữ liệu. Giá trị của các cột dữ liệu có thể hiểu là các trạng thái tương ứng với từng cột dữ liệu được sử dụng trong thuật toán để dự đoán giá trị/trạng thái của cột được chọn làm cột dự đoán của thuật toán. Thuật toán sẽ xem xét tần số xuất hiện của các trạng thái của các cột dữ liệu đầu vào có tần số cao và ảnh hưởng đến trạng thái của cột dự đoán thì thuật toán sẽ chọn cột dữ liệu làm nhân tố chính cho việc dự đoán kết quả.

Ví dụ: Trạng thái tuổi: “Trẻ”, sản phẩm mua: xe đạp có tần số xuất hiện là 9/10 người.

Trạng thái tuổi: “Già”, sản phẩm: truyện tranh có tần số xuất hiện là 2/10. Như vậy, thuật toán đưa ra kết luận trạng thái tuổi là thuộc tính dự đoán tốt về khả năng mua xe đạp.

6.4.1.2 Thuật toán Naïve Bayes

- Thuật toán này xây dựng mô hình khai thác nhanh hơn các thuật toán khác.
- Phục vụ việc phân loại và dự đoán.

Thuật toán này cho ta 1 mô hình khai thác đơn giản (có thể được coi là điểm xuất phát của DataMining), bởi vì hầu như tất cả các tính toán sử dụng trong khi thiết lập mô hình, được sinh ra trong xử lí của cube (mô hình kích thước hợp nhất), kết quả được trả về nhanh chóng. Điều này tạo cho mô hình 1 lựa chọn tốt để khai phá dữ liệu khám phá các thuộc tính input được phân bố trong các trường khác nhau của thuộc tính dự đoán như thế nào?

Nó tính toán khả năng có thể xảy ra trong mỗi trường hợp lệ của thuộc tính đầu vào, gán cho mỗi trường một thuộc tính có thể dự đoán. Mỗi trường này có thể sau đó được sử dụng để dự đoán kết quả của thuộc tính dự đoán dựa vào những thuộc tính input đã biết. Các khả năng sử dụng để sinh ra các mô hình được tính toán và lưu trữ trong suốt quá trình xử lý của khối lập phương (cube: các mô hình được dựng lên từ các khối lập phương).

6.4.1.3 Thuật toán Clustering

- Phuc vụ việc phân nhóm, cụm.
- Có thể dự đoán dự đoán từ các liên cung được tạo ra từ thuật toán.

Đối với các thuộc tính rời rạc. Thuật toán này sử dụng kỹ thuật lặp để nhóm các bản ghi từ 1 tập hợp dữ liệu vào một liên cung có giá trị rời rạc cùng đặc điểm giống nhau. Sử dụng liên cung này có thể khám phá dữ liệu, tìm hiểu về các quan hệ đã tồn tại, mà các quan hệ này không dễ dàng tìm được một cách hợp lý thông qua quan sát ngẫu nhiên.

Ví dụ: Xem xét một nhóm người sống trong cùng một vùng, sử dụng một loại xe, ăn một loại thức ăn và mua cùng một sản phẩm. Đây là một liên cung của dữ liệu, một liên cung khác có thể bao gồm những người cùng đến một nhà hàng, cùng mức lương, và được đi nghỉ ở nước ngoài 2 lần trong năm. Hãy quan sát những liên cung này được phân phối. Từ đó, có thể biết rõ hơn sự ảnh hưởng của các bản ghi trong một tập hợp dữ liệu. Cũng như sự ảnh hưởng này có ảnh hưởng đến kết quả của thuộc tính dự đoán.

6.4.1.4 Thuật toán Neural Network (SSAS)

- Phuc vụ việc phân loại và dự đoán.

Thuật toán này tạo các mô hình khai thác hồi quy và phân loại bằng cách xây dựng đa lớp perceptom của các neuron.

- Giống như thuật toán cây quyết định, đưa ra mỗi tình trạng của thuộc tính có giá trị rời rạc có thể dự đoán. Thuật toán này tính toán khả năng có thể của mỗi trạng thái có thể của thuộc tính đầu vào. Thuật toán sẽ xử lý toàn thể các trường hợp. Sự lặp đi lặp lai so sánh các dự đoán phân loại của các trường với sự phân loại của các trường đã biết. Sai số từ sự phân loại ban đầu (của phép lặp ban đầu) của toàn bộ các trường hợp được trả về network và được sử dụng để thay đổi sự thực thi của network cho các phép lặp kế theo,v.v.. Có thể sau đó sử dụng những khả năng này để dự đoán kết quả cảu các thuộc tính dự đoán, dựa trên thuộc tính đầu vào.
- Một sự khác biệt chính giữa thuật toán này và thuật toán Cây quyết định là các kiến thức xử lí là những tham số network tối ưu nhằm làm nhỏ nhất các lỗi có thể trong khi cây quyết định tách các luật, mục đích để cực đại hoá thông tin có lợi.

6.4.2 Dự đoán một thuộc tính liên tục

Thuộc tính liên tục là thuộc tính:

- Có số thực như là giá trị của thuộc tính
  Ví dụ: Nhiệt độ, chiều cao hay cân nặng.
- Gía trị được thực hiện bằng việc đo hay biểu diễn bằng cách một số hữu hạn của một dãy số.
  Ví dụ: Nhiệt độ = 15, 16, 17, 18…40, chiều cao = 1,6; 1,7; 18.5, cân nặng= 35.5kg, 40.5kg

6.4.2.1 Thuật toán Decision Trees

Khi xây dựng một cây sẽ dựa trên thuộc tính dự đoán có giá trị liên tục, mỗi một nhánh sẽ là một công thức hồi quy. Việc chia cây sẽ thực hiện trên một điểm của công thức không hồi quy.

Ví dụ: Trạng thái tuổi: 15 +, trạng thái sản phẩm: xe đạp có tần số xuất hiện là 9/10 người.

Trạng thái tuổi: 40 +, sản phẩm: truyện tranh có tần số xuất hiện là 2/10. Như vậy, thuật toán đưa ra kết luận phân chia độ tuổi là thuộc tính dự đoán tốt về khả năng mua xe đạp để thực hiện tiếp công việc phân chia tiếp theo.

6.4.2.2 Thuật toán Time Series

Sử dụng thuật toán này có thể chọn 1 hoặc nhiều biến để dự đoán (nhưng các biến là phải liên tục). Có thể có nhiều trường hợp cho mỗi mô hình. Tập các trường hợp xác định vị trí của 1 nhóm, như là ngày tháng khi xem việc bán hàng thông qua vài tháng hoặc vài năm trước.

Một trường hợp có thể bao gồm 1 tập các biến (ví dụ như bán hàng tại các cửa hàng khác nhau). Thuật toán này có thể sử dụng sự tương quan của thay đổi biến số (cross-variable) trong dự đoán.

Ví dụ: Bán hàng trước kia tại 1 cửa hàng có thể rất hữu ích trong việc dự báo bán hàng hiện tại tại những cửa hàng.

Ví dụ: Việc giám sát hoạt động kinh doanh trên một sản phẩm (Iphone) có ảnh hưởng đến việc dự đoán đến hoạt động kinh doanh của các sản phẩm khác (Ipad)

6.4.3 Dự đoán một trình tự.

Thuật toán Sequence Clustering sử dụng kết hợp kỹ thuật gom cụm và kỹ thuật phân tích chuỗi để xác định các cụm và trình tự của các cụm. Điểm nổi bật của thuật toán này là sử dụng dữ liệu tuần tự. Dữ liệu này đại diện cho một chuỗi các sự kiện hoặc sự chuyển đổi giữa các trạng thái trong tập dữ liệu ví dụ như một loạt các sản phẩm được mua hoặc các thao tác nhấp chuột của một người dùng cụ thể. Thuật toán kiểm tra tất cả những khả năng chuyển đổi có thể xảy ra giữa tất cả các trình tự bên trong tập dữ liệu để quyết định trình tự nào là tốt nhất cho các cụm.

Ví dụ: Một website thu thập thông tin về những trang mà người dùng đã xem và thứ tự các trang được click chọn để xem. Khách hàng phải đăng nhập trước khi xem các thông tin trên website nên website sẽ ghi nhận được thông tin về thứ tự các trang đã được click xem của từng khách hàng cụ thể. Bằng cách sử dụng thuật toán Sequence Clustering, ta có thể tìm ra các cụm khách hàng có cùng đặc điểm chung và cùng trình tự click chọn các trang đã xem.

Từ đây ta có thể phân tích cách khách hàng di chuyển tuần tự qua các trang như thế nào vì có thể liên quan đến việc bán một sản phẩm cụ thể trên mạng và cũng có thể dự đoán được những trang có nhiều khả năng được truy cập.

6.4.4 Tìm nhóm của những mục chọn (item) trong các các giao tác (transaction).

6.4.4.1 Thuật toán Association

Dữ liệu đầu vào của thuật toán là các giao tác (transactions) và một nhóm các mục chọn (itemset) tương ứng với từng giao tác, mỗi mục chọn gọi là một item. Thuật toán tìm ra sự kết hợp giữa các mục chọn trong từng giao tác và số lần xuất hiện của sự kết hợp này trong tập dữ liệu đưa vào, dựa vào độ hỗ trợ và xác suất để tìm ra quy luật kết hợp giữa các mục chọn.

Ví dụ: Trong bài toán phân tích giỏ hàng, các giao tác chính là các đơn hàng đã phát sinh trong quá khứ, itemset là danh sách các mặt hàng trong từng đơn hàng hoặc thông tin khách hàng mua hàng. Thuật toán Association dựa vào các thông tin trên và tìm ra các luật kết hợp như:

- Khách hàng có thuộc tính gì sẽ mua mặt hàng gì.
- Khách hàng mua mặt hàng A nào đó sẽ mua mặt hàng nào khác nữa.
- Mặt hàng có thuộc tính sản phẩm gì có thể kết hợp với thuộc tính sản phẩm gì

Với những quy luật mà thuật toán Association, ta có thể đưa ra nhiều chiến lược kinh doanh phù hợp như:

- Tăng số lượng một số mặt hàng dành cho một số đối tượng khách hàng.
- Sắp xếp lại vị trí các mặt hàng để khách hàng có thể lựa chọn thoải mái.
- Tăng/thêm mới số lượng mặt hàng có những giá trị thuộc tính cụ thể mà nhiều người đang quan tâm.

6.4.4.2 Thuật toán Decision Trees

Dữ liệu đầu vào của thuật toán là các thuộc tính có ảnh hưởng đến thuộc tính dự đoán, từ đó tìm ra luật là khả năng có thể xảy ra với một số mối liên hệ giữa các thuộc tính.

Ví dụ: Cũng với bài toán phân tích giỏ hàng, luật của thuật toán Decision Trees sẽ là dự đoán khả năng xảy ra dựa vào thông tin khách hàng và mặt hàng khách mua.

Với thuật toán Decision Trees, luật đưa ra sẽ giúp các doanh nghiệp tăng số lượng mặt hàng có một số thông tin mà nhiều người quan tâm hoặc đưa ra mặt hàng mới dựa vào những tiêu chí về đặc

6.4.5 Tìm những mục (item) giống nhau

6.4.5.1 Thuật toán Clustering

Thuật toán Clustering sử dụng kỹ thuật lặp nhiều lần để tìm ra sự giống nhau của các thuộc tính dữ liệu và gom dữ liệu lại thành nhiều cụm dữ liệu có các đặc điểm tương tự nhau.

Ví dụ: Với một tập dữ liệu bán hàng, thuật toán Clustering thực hiện khai thác dữ liệu và tìm ra được nhiều cụm dữ liệu khác nhau, trong đó có cụm khách hàng tuổi teen thường quan tâm đến các loại hình giải trí còn cụm khách hàng lớn tuổi thường quan tâm đến sản phẩm liên quan đến sức khoẻ,…

6.4.5.2 Thuật toán Sequence Clustering

Thuật toán này phân tích các đối tượng dữ liệu có trình tự, các dữ liệu này bao gồm một chuỗi các giá trị rời rạc. Cách phân tích sự chuyển tiếp giữa các tình trạng của một chuỗi, thuật toán có thể dự đoán tương lai trong các chuỗi có quan hệ với nhau. Thuật toán nhóm tất cả các sự kiện phức tạp với các thuộc tính trình tự vào một phân đoạn dựa vào sự giống nhau của những chuỗi này.

Ví dụ: Phân tích khách hàng web của một cổng thông tin (portal site). Một cổng thông tin là một tập các tên miền liên kết như: tin tức, thời tiết, giá tiền, mail, và thể thao.. . Mỗi khách hàng được liên kết với một chuỗi các click web trên các tên miền này. Thuật toán này có thể nhóm các khách hàng web về một hoặc nhiều nhóm dựa trên kiểu hành động của họ. Những nhóm này có thể được trực quan hoá, cung cấp một bản chi tiết để biết được mục đích sử dụng trang web này của khách hàng.

6.4.6 Kết luận

Từ các thuật toán trên, phân loại các thuật toán khai thác dữ liệu như sau (5):

- Thuật toán phân loại (Classification): Dự đoán 1 hoặc nhiều biến rời rạc (không liên tục), dựa trên các thuộc tính trong tập hợp dữ liệu (Decision Trees Algorithm).

- Thuật toán hồi quy (Regression): Dự đoán 1 hoặc nhiều biến liên tục, kiểu như những lợi nhuận và những tổn thất, dựa trên các thuộc tính khác nhau của tập hợp dữ liệu (Time Series Algorithm).
- Thuật toán phân đoạn (Segmentation): Chia dữ liệu thành 2 nhóm, hoặc các liên cung, hoặc các danh mục có thuộc tính giống nhau (Clustering Algorithm).
- Thuật toán kết hợp (Association): Tìm những sự tương quan giữa các thuộc tính khác nhau trong 1 tập hợp dữ liệu. Ứng dụng phổ biến nhất của loại thuật toán này là tạo ra các luật kết hợp, có thể được dùng trong market basket (Association Algorithm).
- Thuật toán phân tích tiến trình (Sequence analysis): Tổng kết những tiến trình thường xảy ra hoặc ít xảy ra trong dữ liệu. Thuật toán phổ biến (Sequence Clustering Algorithm).

Kết luận: chọn một thuật toán đúng để sử dụng cho các nghiệp vụ riêng biệt là một nhiệm vụ rất khó khăn. Các thuật toán khác nhau để thực thi cùng một nghiệp vụ, mỗi thuật toán tạo ra một kết quả khác nhau và một vài thuật toán có thể tạo ra nhiều hơn một kết quả.

Ví dụ 1: Có thể sử dụng thuật toán Clustering, nhận ra các mẫu, đưa dữ liệu vào nhóm đồng nhất, và sau đó sử dụng các kết quả để tạo ra mô hình cây quyết định tốt hơn.

Ví dụ 2: Có thể sử dụng thuật toán Decision Trees không chỉ để dự đoán mà còn là một cách để giảm số lượng cột trong dataset, vì cây quyết định có thể xác định các cột mà không ảnh hưởng đến mô hình khai thác cuối cùng.

Như vậy, tùy vào một bài toán cụ thể, việc phân tích xác định vấn đề rất quan trọng. Cần phải phân tách công việc thực hiện, xem xét tình huống, ưu và khuyết điểm của từng thuật toán để việc chọn lựa thuật toán phù hợp với yêu cầu đề ra.

7. PHÂN TÍCH VẤN ĐỀ & GIẢI PHÁP

7.1 Xác định vấn đề

7.1.1 Xác định các định nghĩa liên quan

Các khái niệm được sử dụng trong mua hàng trên mạng xã hội:

- *Chỉ mình tôi (khách hàng):* người trực tiếp mua hàng cho chính họ trên hệ thống.

Ví dụ: Khách hàng là Nguyễn Văn A.

- *Bạn bè*: bạn bè của khách hàng. Dữ liệu này lấy từ mã tài khoản (uid) của người dùng trên mạng xã hội facebook.
  Ví dụ: Bạn của Nguyễn Văn A là Nguyễn Văn B, Nguyễn Thị C.
- *Mối quan hệ giữa các đối tượng*: gồm: cha/mẹ, con trai/con gái, anh/em trai, chị/em gái, người yêu, vợ/chồng.
  Ví dụ: User Trần Văn Y có mối quan hệ cha con với User Nguyễn Văn A.
- *Mọi người*: Tất cả mọi người có tài khoản tại hệ thống website bán hàng trực tuyến.
- *Đối tượng tặng và đối tượng được tặng quà sinh nhật*
  Ví dụ: Bạn của Nguyễn Văn A là Nguyễn Văn B, Nguyễn Thị C.
  Ngày 10/12/2010 sắp tới, là ngày sinh nhật của Nguyễn Văn B. Ta gọi Nguyễn Văn A là đối tượng tặng và Nguyễn Văn B là đối tượng được tặng.

7.1.2 Xác định bài toán

Dựa vào hành vi thao tác trên hệ thống và mua hàng của khách hàng trong lịch sử mua hàng. Hệ thống cần xác định:

Trường hợp *đối tượng được tặng tồn tại* trên hệ thống và *họ có mua hàng*:

- Họ mua trong khoảng thời gian nào?
- Đối tượng được tặng họ hay mua loại sản phẩm gì?
- Loại sản phẩm đó có những thuộc tính ra sao?

Trường hợp *đối tượng được tặng* tồn tại hay không tồn tại trên hệ thống và họ *chưa mua hàng*:

- *Đối tượng giống đối tượng được tặng* họ mua cái gì? Loại sản phẩm đó có những thuộc tính gì? Họ mua trong khoảng thời gian nào?
- *Hệ thống muốn dự đoán gì* mà đối tượng được tặng có xu hướng mua?
- *Tập khách hàng như thế nào* gọi là giống với đối tượng được tặng?
- *Tập thuộc tính sản phẩm* gồm những thuộc tính như thế nào?
- *Mối quan hệ giữa đối tượng tặng và đối tượng được tặng* gồm những mối quan hệ như thế nào?

7.1.3 Xác định công nghệ sử dụng

- Microsoft SQL server 2008 R2.
- Data mining của MS SQL server phiên bản 2008.
- Ngôn ngữ lập trình PHP version 5.2.13
- Frame work CodeIgniter version 1.7.3
- Mô hình cấu trúc MVC.
- Kỹ thuật AJAX.

7.2 Khó khăn khi thực hiện

Khó khăn đầu tiên bao gồm:

- Cần dữ liệu thông tin lớn của khách hàng và đơn hàng. Thông tin phải xác thực và có ý nghĩa. Dữ liệu này có đặc thù riêng, thông tin khách hàng không chỉ thông tin của khách hàng đó mà phải đi kèm với danh sách bạn bè của khách hàng từ mạng xã hội facebook.

  Ví dụ: Đối với khách hàng Nguyễn Văn A, khách hàng A này có 50 bạn bè. Thông tin mỗi bạn bè của khách hàng A gồm: mã tài khoản facebook, tên đầy đủ, ngày sinh, giới tính và hình ảnh.

  Những thông tin bạn bè này rất có ích vì khả năng tiếp cận cao, giới thiệu sản phẩm mà khách hàng A đã mua cho những bạn bè của mình thông qua mạng xã hội.
- Cần xây dựng một hệ thống chạy trực tuyến: đối tượng mua hàng có tồn tại nhiều mối quan hệ với các đối tượng khác (danh sách bạn bè của họ). Mà API của mạng xã hội facebook chỉ cung cấp hàm dữ liệu trả về thông tin của khách hàng thông qua Web Services.
- Không có nhiều thời gian để tạo danh sách sản phẩm phong phú và đa dạng, làm thế nào để kích thích người dùng trên mạng xã hội tham gia mua hàng tại hệ thống trong thời gian ngắn (4 tháng).
- Không lấy đủ thông tin của người dùng do chế độ bảo mật từ mạng xã hội facebook. Trong mạng xã hội facebook có chế độ bảo mật có các ứng dụng truy cập để lấy thông tin ở dạng công khai (public) và riêng tư (private).

Ví dụ: Không lấy dữ liệu trả về tình trạng quan hệ của người dùng Vũ Q (kết hôn, đang có tình trạng quan hệ với ai...) do người dùng Vũ Q xét chế độ bảo mật thông tin quan hệ cá nhân ở chế độ "private".

Khó khăn thứ hai là xử lý giá trị bị nhiễu (missing value). Cho tình huống như sau:

- *Tình huống 1*: Không tồn tại dữ liệu của đối tượng mua hàng làm dữ liệu kiểm tra (test set).
  Ví dụ: Hệ thống chỉ tồn tại các đối tượng mua hàng qua mạng xã hội từ 17 đến 50 tuổi. Người dùng A muốn tặng quà sinh nhật cho cha (độ tuổi: 60 tuổi) của mình. Như vậy, không tồn tại lứa tuổi 60 trong hệ thống hiện tại. Gợi ý thông tin gì cho người tặng?
- *Tình huống 2*: Tồn tại dữ liệu của đối tượng mua hàng nhưng giá trị đó bị thiếu.
  Ví dụ: Hệ thống có khách hàng A, B, C đã mua hàng nhưng thông tin của khách hàng đó lại không có giá trị về tuổi (3/12/<không có gí trị năm sinh>), hoặc không có ngày tháng năm sinh? Gợi ý thông tin gì cho người tặng?

Khó khăn thứ ba là xây dựng hệ thống gợi ý tặng quà cho các đối tượng trong hệ thống và các đối tượng đó có các mối quan hệ với đối tượng đi mua hàng. Những gợi ý bao gồm thói quen mua hàng của cá nhân người được tặng và thói quen mua hàng của những người giống người được tặng. Ta phải thực hiện khai thác dữ liệu dựa trên thông tin đơn hàng đã tồn tại trên hệ thống và việc gợi ý phải thực hiện trên hệ thống truy cập trực tiếp (Online Mode). Điều này đòi hỏi phải có một server riêng mới đảm bảo yêu cầu gợi ý ngay tại thời điểm mua hàng. Bên cạnh đó, không tồn tại host thuê có hỗ trợ Analysis Services của MS SQL Server và chức năng truy cập từ xa (remote) trong thời điểm thực hiện dự án (tháng 09 năm 2010). Nếu muốn sử dụng Analysis Services của MS SQL Server thì phải mua server riêng phục vụ đề tài thì chi phí mua Server và đặt Server ở Data Center dự tính ở mức thấp nhất là 52 triệu/ 4 tháng.

7.3 Giải pháp bài toán

Với những khó khăn nêu trên và yêu cầu của bài toán, giải pháp phát triển của chúng tôi đối với khó khăn đầu tiên:

- Xây dựng hệ thống chạy trực tuyến với tên miền là www.face4shop.com, sử dụng API để lấy thông tin khách hàng thực từ mạng xã hội facebook bằng việc đăng nhập hệ thống bằng tài khoản của facebook.
- Hệ thống cho phép người dùng facebook có thể đăng bán sản phẩm của mình tạo thành một Shop với tên Shop là tên của người dùng. Bên cạnh đó, người dùng có thể mua hàng ở các Shop khác, hoạt động của người dùng từ mạng xã hội facebook hoàn toàn miễn phí. Việc này tạo sự kích thích trao đổi mua bán, có cùng sở thích mua bán hàng trực tuyến trên mạng xã hội.
- Đối với chế độ bảo mật của mạng xã hội facebook. Đây là bài toán tâm lý về sự tin cậy cho người dùng facebook.

  Giải pháp vấn đề này: Cho người dùng *nhập thêm thông tin* mà hệ thống cần khai thác (địa chỉ, tình trạng quan hệ…) tại trang www.face4shop.com sau khi người dùng này sử dụng hệ thống với tham số trên 3 lần (đăng nhập) trong một khoảng thời gian (tham số: 2 ngày). Việc này thể hiện, khách hàng đã tin tưởng hệ thống và họ sử dụng thường xuyên (3 lần/2 ngày).

Giải pháp cho khó khăn thứ hai: dữ liệu bị nhiễu (missing value)

- *Tình huống 1*: Dữ liệu ở huấn luyện lần đầu không tồn tại (đối tượng đó chưa mua hàng ở hệ thống face4shop.com). Như vậy, sẽ không gợi ý mua quà sinh nhật tại thời điểm này.
- *Tình huống 2*: Dữ liệu ở huấn luyện lần hai (đối tượng đã mua hàng nhưng không có tuổi hoặc không có ngày tháng năm sinh), xem xét:
  - *Đối tượng nhiễu ít: bỏ qua trường hợp này, không thực hiện gợi ý.*
  - *Đối tượng nhiễu nhiều:*

  Cách 1: Xem xét họ thiếu thông tin gì (ví dụ tuổi)? Từ đó, tập huấn luyện dữ liệu lần sau lấy dữ liệu thiếu.

  Ví dụ: Đối tượng như khách hàng A, B, C.. → tập huấn luyện dữ liệu phân cụm khách hàng lần sau, bỏ qua đặc trưng "Tuổi".

  Cách 2: Quy những đối tượng bị nhiễu về 1 giá trị mặc định. Đặt tên giá trị "cluster Missing".

Ví dụ: Khách hàng đã mua hàng nhưng thông tin về khách hàng bị nhiễu năm sinh. Thống kê cho thấy họ hay mua loại sản phẩm là Sách. Cho họ vào 1 cụm khách hàng mặc định với loại hàng mua mặc định (Sách). Đối với dữ liệu kiểm tra đưa vào, nếu đối tượng được tặng bị giá trị nhiễu về tuổi thì cho họ vào cụm khách hàng mặc định này.

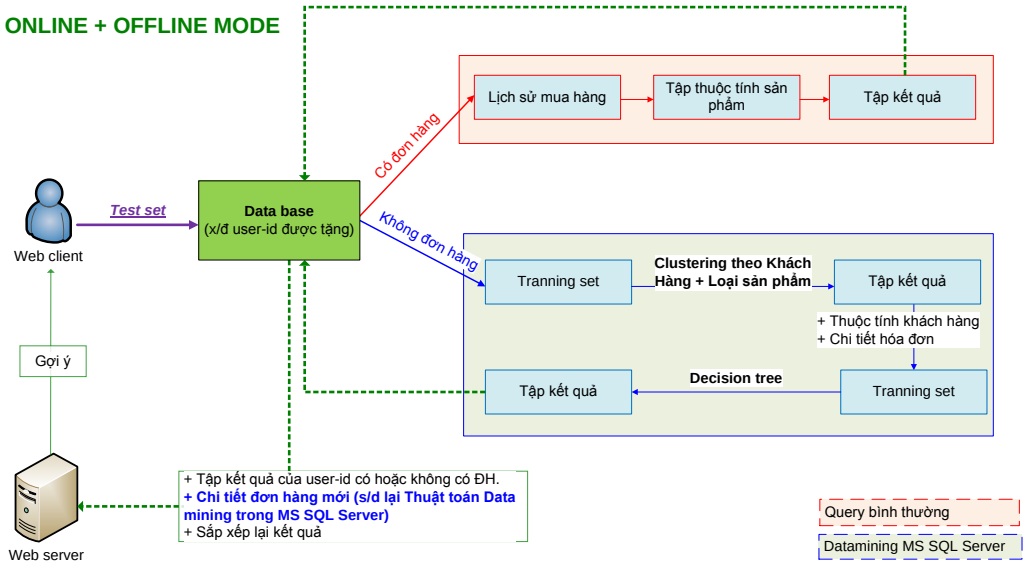
Và khó khăn thứ ba:

- Vấn đề thực hiện khai thác dữ liệu sẽ được thực hiện ở chế độ Offline định kỳ (7 ngày/ huấn luyện dữ liệu một lần). Bên cạnh đó, chế độ Offline định kỳ này thích hợp với bài toán có *dữ liệu tĩnh* (chỉ khai thác dữ liệu lịch sử mua hàng của người dùng trước đó).

  Quá trình khai thác dữ liệu sẽ được thực hiện tại cơ sở dữ liệu cục bộ (database local). Sau đó, đẩy kết quả từ cục bộ lên cơ sở dữ liệu máy chủ (Database server) để thực hiện gợi ý.

  Sau một định kỳ (7 ngày), sẽ lấy dữ liệu của lần huấn luyện dữ liệu định kỳ trước đó kết hợp với dữ liệu của lần huấn luyện định kỳ lần này. Từ đó, tạo ra một tập huấn luyện mới và sử dụng khai thác dữ liệu đối với tập huấn luyện. Việc này sẽ lặp lại đối với lần huấn luyện ở định kỳ tiếp theo.
- Khi thực hiện khai thác dữ liệu ở chế độ Offline sẽ giải quyết được vấn đề giảm được chi phí dự án (không phải mua server riêng) nhưng không giải quyết được lượng dữ liệu phát sinh sau thời điểm thực hiện khai thác dữ liệu. Nếu chọn cách kết hợp tập kết quả khai thác dữ liệu và lượng dữ liệu sau đó rồi tiến hành khai thác lại ngay trên hệ thống face4shop.com bằng thuật toán tự viết thì sẽ không đồng bộ được với thuật toán thực hiện trong Data mining Microsoft SQL Server 2008.

  Giải pháp vấn đề này là thực hiện gợi ý thói quen mua hàng của cá nhân người được tặng (Online tab) và xu hướng mua hàng của những người giống với người được tặng (Offline tab). Sau đây là mô hình phác thảo giải pháp đối với những khó khăn khi thực hiện:



Hình 20 – Hình mô tả giải pháp Online và Offline Mode

Gợi ý theo "Thói quen mua hàng của người được tặng" – Online tab

- *Điều kiện*: đối tượng được tặng đã mua hàng trong hệ thống.
- *Mô tả*: khi người mua hàng chọn mua hàng tặng cho một đối tượng xác định, tìm thông tin đơn hàng mà đối tượng này đã từng mua trong hệ thống. Từ đó tìm thông tin tập thuộc tính sản phẩm đối tượng đã từng mua. Thực hiện sắp xếp tập thuộc tính sản phẩm đã mua theo thứ tự mua nhiều nhất giảm dần.
- *Kết quả*: kết quả trả về là tập thuộc tính sản phẩm gợi ý cho người tặng.
- *Ưu điểm*: xử lý được lượng dữ liệu tức thời ngay tại thời điểm người tặng mua hàng. Thực hiện thống kê tập thuộc tính sản phẩm mà đối tượng được tặng từng mua.

Gợi ý theo "Xu hướng mua hàng của những người giống với người được tặng" – Offline tab

- *Điều kiện*: đối tượng được tặng quà chưa mua hàng hoặc có mua hàng.
- *Mô tả*: thực hiện khai thác dữ liệu
    - *Bước 1: Thực hiện gom cụm khách hàng*

      Tạo dữ liệu huấn luyện gồm thông tin khách hàng và loại sản phẩm khách hàng mua. Thực hiện khai thác dữ liệu huấn luyện theo phương phép gom cụm với mong muốn sẽ phân cụm khách hàng cùng mua một số loại sản phẩm.

Ví dụ: Cụm khách hàng A thể hiện rằng vào tháng 10, khách hàng có độ tuổi 20, giới tính Nam hay mua sản phẩm quần jean.

- *Bước 2: Thực hiện phân lớp thuộc tính sản phẩm cho từng cụm khách hàng.*

Tạo dữ liệu huấn luyện gồm thông tin loại sản phẩm và tập thuộc tính sản phẩm tương ứng theo từng cụm khách hàng. Thực hiện khai thác dữ liệu huấn luyện theo phương pháp phân lớp.

- *Kết quả*:
  - Kết quả bước 1: Tập các cụm khách hàng và loại sản phẩm tương ứng.
  - Kết quả bước 2: Tập thuộc tính sản phẩm của từng cụm khách hàng.
  - Kết quả bước 2 sẽ được lưu vào cơ sở dữ liệu hỗ trợ cho việc gợi ý.
- *Thực hiện gợi ý*: khi người tặng click chọn mua hàng tặng cho một đối tượng xác định, thông tin đối tượng này được tìm trong tập kết quả xem có trùng khớp hay không và kết quả trả về là tập thuộc tính sản phẩm gợi ý cho người mua hàng.
- *Ưu điểm*: không tốn chi phí dùng server riêng.
- *Khuyết điểm*: không thực hiện khai thác dữ liệu trên dữ liệu mới nhất vì thực hiện chế độ offline tại một thời điểm trước thời điểm người mua hàng thực hiện mua hàng. Khuyết điểm này đã được thay thế bằng lựa chọn gợi ý theo "Thói quen mua hàng của người được tặng" – Online tab.

7.4 Phương pháp thực hiện khai thác dữ liệu

7.4.1 Lý do chọn phương pháp khai thác dữ liệu

Dựa vào tổng quan lý thuyết về các phương pháp khai thác dữ liệu và bảng so sánh các phương pháp. Chúng tôi chọn kết hợp hai phương pháp khai thác dữ liệu sau:

Thuật toán gom cụm (Clustering) đáp ứng các công việc:

- Tìm những khách hàng có thuộc tính giống nhau thành những cụm khách hàng riêng biệt.

  Ví dụ: Cụm khách hàng A: có độ tuổi = {20, 21, 22}, giới tính là Nam, hay mua sản phẩm Quần jean.
- Dự đoán thuộc tính cụm khách hàng (cụm A, cụm B, cụm C và các cụm này đều là thuộc tính có giá trị rời rạc)

Ví dụ: Khách hàng X (tuổi: 20, giới tính Nam) sẽ được dự đoán nằm trong cụm khách hàng A.

Thuật toán cây quyết định (Decision Trees) đáp ứng các công việc:

- Dự đoán thuộc tính mua hàng của sản phẩm (giá trị = {Mua, Không mua}). Đây là thuộc tính rời rạc. Thuộc tính này là thuộc tính dùng để phân lớp sản phẩm.
- Bên cạnh đó thuật toán cây quyết định còn hỗ trợ dự đoán thuộc tính liên tục, điều này có nghĩa sau này dự đoán thêm một giá trị liên tục sẽ không phải thay đổi bởi một thuật toán khác (tính dễ mở rộng).
- Từ mô hình cây quyết định, có thể chuyển đổi CÂY thành LUẬT phân lớp đơn giản và dễ hiểu.

  Ví dụ: Cụm khách hàng A có luật

  If Loại = 'Quần jean' and màu = "Xanh" then Mua = 'Mua'.
- Có thể dùng các truy vấn SQL phục vụ truy cập CSDL.

Kết hợp thuật toán gom cụm và cây quyết định vì:

- Thuật toán gom cụm nhận ra các mẫu, đưa dữ liệu vào nhóm đồng nhất. Từ đó, làm giảm độ phức tạp thuộc tính. Dẫn đến cụm khách hàng tạo ra sẽ trả về tập kết quả giảm bớt tập thuộc tính cho thuật toán khai thác dữ liệu tiếp theo (cây quyết định).

  Ví dụ: Thay vì sử dụng cây quyết định để phân lớp khách hàng và thuộc tính sản phẩm mua hàng. Nếu dùng cách này, sẽ có quá nhiều thuộc tính thực hiện phân lớp cây, lúc đó cây sẽ rơi vào tình trạng overfitting. Điều này dẫn đến cây phân nhánh quá rộng dẫn đến dự đoán sẽ không bảo đảm độ chính xác khi thực hiện phân lớp mua hàng.

7.4.2 Chi tiết thực hiện phương pháp

7.4.2.1 Phương pháp gom cụm

Tạo dữ liệu huấn luyện (trainning set): Trước khi tạo một tập huấn luyện để thực hiện khai thác dữ liệu, giai đoạn quan trọng nhất là cần phải thực hiện chính là làm sạch dữ liệu. Làm sạch dữ liệu được thực hiện đối với các trường hợp sau:

Dữ liệu bị thiếu:

- Giới tính, ngày tháng năm sinh: theo thống kê hiện tại số khách hàng sử dụng hệ thống có đầy đủ thông tin giới tính và ngày tháng năm sinh nên chúng tôi không xét những đối tượng bị thiếu thông tin này.
- Thời gian mua hàng: hệ thống face4shop.com ghi nhận thời điểm phát sinh đơn hàng của khách hàng.
- Hàng hoá: hàng hoá được phân loại trước khi thể hiện trên hệ thống nên việc chọn một mặt hàng bất kỳ là có thể xác định được tên loại sản phẩm, giới tính của loại sản phẩm.

  Như vậy, bộ thuộc tính khách hàng không tồn tại trường hợp dữ liệu bị thiếu.

Dữ liệu bị nhiễu: do bộ thuộc tính khách hàng sẽ được thực hiện khai thác dữ liệu bằng phương pháp gom cụm cũng là một cách làm mịn dữ liệu bị nhiễu nên ta bỏ qua bước này.

Làm sạch dữ liệu: về vấn đề dữ liệu không nhất quán, hiện tại hệ thống chạy trên một hệ quản trị cơ sở dữ liệu là SQL Server 2008, thông tin lấy từ nguồn khác là facebook đã được kiểm tra trước khi tích hợp vào cơ sở dữ liệu của hệ thống nên không xảy ra trường hợp không đồng bộ dữ liệu.

Tích hợp dữ liệu: dữ liệu được chọn để tạo dữ liệu huấn luyện thuộc một nguồn duy nhất là hệ quản trị cơ sở dữ liệu SQL Server 2008 nên chúng tôi không cần phải thực hiện tích hợp dữ liệu.

Chuyển hoá dữ liệu: điều kiện trước khi tạo dữ liệu huấn luyện là xác định bộ thuộc tính chuẩn. Các thuộc tính được chọn trong bộ thuộc tính chuẩn đã được kiểm tra và chọn lọc trước nên bước này sẽ bỏ qua.

Lựa chọn các tập thuộc tính trước khi thực hiện kết hợp

- Thuộc tính khách hàng gồm: giới tính, tuổi, tháng sinh nhật.
- Thuộc tính đơn hàng của khách hàng: năm mua hàng, tháng/quý mua hàng, loại sản phẩm, giới tính sản phẩm.
- Thực hiện kết hợp hai tập thuộc tính lại thành một tập thuộc tính duy nhất tạo thành bảng dữ liệu huấn luyện trước khi thực hiện khai thác dữ liệu.

- Tạo bảng ghi nhận các thông tin phục vụ cho việc thống kê sau này gồm:
    - Thời điểm thực hiện khai thác dữ liệu.
    - Số đơn hàng.
    - Số khách hàng.
    - Bảng dữ liệu huấn luyện.
    - Bảng danh sách gom cụm.
    - Bảng thông tin chi tiết gom cụm.

Thuật toán thực hiện thu gọn dữ liệu:

- Khai báo tham số thời gian thực hiện khai thác dữ liệu.
- Tính toán các số liệu: số đơn hàng, số khách hàng trong đơn hàng đến thời điểm thực hiện khai thác dữ liệu.
- Kiểm tra tồn tại và tạo mới (nếu không tồn tại) bảng ghi nhận thông tin các thời điểm thực hiện khai thác dữ liệu.
- Gán tên các thông tin vào các tham số đã khai báo trước có kết hợp với thời điểm thực hiện khai thác dữ liệu: bảng dữ liệu huấn luyện, bảng danh sách gom cụm, bảng thông tin chi tiết gom cụm.
- Cập nhật các thông tin liên quan đến thời điểm thực hiện khai thác dữ liệu vào bảng quy định.
- Tạo cấu trúc bảng dữ liệu huấn luyện.
- Cập nhật thông tin vào bảng dữ liệu huấn luyện.
    - Thông tin cập nhật: tháng sinh nhật, giới tính, tuổi, tháng/quý mua hàng, năm mua hàng, loại sản phẩm, giới tính loại sản phẩm.
    - Thông tin bảng dữ liệu kết hợp:

      Chi tiết đơn hàng: kết hợp với bảng sản phẩm để lấy thông tin sản phẩm.

      Đơn hàng: thông tin là mã khách hàng, tháng/quy mua hàng (chuyển dữ liệu tháng của ngày thực hiện đơn hàng theo quý trong năm), năm mua hàng (lấy năm của ngày thực hiện đơn hàng), điều kiện là đến thời điểm thực hiện khai thác dữ liệu và kết hợp với chi tiết đơn hàng.

      Khách hàng: thông tin là mã khách hàng, tháng sinh nhật (lấy tháng của ngày sinh), giới tính (chuyển đổi kiểu dữ liệu bit thành dữ liệu chuỗi phân biệt giới tính khách hàng), tuổi (tính tuổi của khách hàng đến thời điểm khai

thác dữ liệu), điều kiện là kết hợp với bảng đơn hàng.

Sản phẩm: thông tin là loại sản phẩm, giới tính sản phẩm, điều kiện là kết hợp với chi tiết đơn hàng.

- Kết thúc giải thuật, kết quả trả về:
  - Nội dung bảng quy định tại thời điểm thực hiện khai thác dữ liệu bao gồm cả tên bảng dữ liệu huấn luyện.
  - Nội dung bảng dữ liệu huấn luyện.

Thực hiện gom cụm: Tạo dự án (project) thực hiện khai thác dữ liệu (nếu chưa tồn tại)/ mở project thực hiện khai thác dữ liệu để thực hiện gom cụm.

Cập nhật các thông tin:

- Thông tin dữ liệu: nguồn dữ liệu (datasource), bảng dữ liệu huấn luyện (datasource view).
- Thông tin dữ liệu huấn luyện:
  - Khoá: mã tự phát sinh trong bảng dữ liệu huấn luyện.
  - Dữ liệu đầu vào: giới tính, tuổi, tháng sinh nhật, tháng/quý mua hàng, năm mua hàng, loại sản phẩm, giới tính sản phẩm.
  - Dữ liệu dự đoán: phân cụm.
  - Tham số thực hiện gom cụm: số cụm (CLUSTER_COUNT), số mẫu (CLUSTER_SEED), phương thức gom cụm (CLUSTER_METHOD), số thuộc tính đầu vào tối đa (MAXIMUM_INPUT_ATTRIBUTE), số trạng thái tối đa của thuộc tính (MAXIMUM_STATE), độ hỗ trợ tối thiểu (MINIMUM_SUPPORT),số mô hình mẫu (MODELLING_CARDINALITY), kích thước mẫu (SAMPLE_SIZE), STOPPING_TOLERANCE.

Thực hiện khai thác dữ liệu trên bảng dữ liệu huấn luyện.

Lấy kết quả trả về: Lấy kết quả sau khi thực hiện khai thác dữ liệu đổ về cơ sở dữ liệu local bao gồm:

- Danh sách cụm
- Thông tin chi tiết các cụm

Thuật toán thực hiện đổ kết quả về cơ sở dữ liệu local.

- Khai báo tham số: thời gian thực hiện khai thác dữ liệu, nguồn dữ liệu trong project, tên mô hình khai thác.
- Lấy nội dung khai thác dữ liệu trong bảng quy định khai thác dữ liệu gồm: bảng danh sách gom cụm, bảng thông tin chi tiết gom cụm.
- Tạo cấu trúc các bảng: bảng danh sách gom cụm, bảng thông tin chi tiết gom cụm.
- Cập nhật danh sách cụm vào bảng danh sách gom cụm và tên các bảng dữ liệu huấn luyện, bảng kết quả cho phương pháp phân lớp từng cụm tương ứng.
- Cập nhật thông tin chi tiết gom cụm vào bảng thông tin chi tiết gom cụm.
- Kết thúc giải thuật, kết quả trả về là:
    - Nội dung bảng danh sách cụm bao gồm tên bảng dữ liệu huấn luyện, bảng kết quả cho phương pháp phân lớp từng cụm tương ứng.
    - Nội dung bảng thông tin chi tiết gom cụm.

7.4.2.2 Phương pháp phân lớp

Tạo dữ liệu huấn luyện (tranning set): Giai đoạn làm sạch dữ liệu để tạo dữ liệu huấn luyện ở phương pháp gom cụm qua các trường hợp sau:

Dữ liệu bị thiếu

- Mã cụm: lấy từ kết quả gom cụm.
- Loại sản phẩm: lấy từ kết quả gom cụm.
- Giới tính loại sản phẩm: lấy từ kết quả gom cụm.
- Mã sản phẩm: lấy từ thông tin đơn hàng theo điều kiện của cụm (mã khách hàng, năm mua hàng, tháng mua hàng, loại sản phẩm, giới tính loại sản phẩm).
- Thương hiệu: lấy từ thông tin sản phẩm theo mã sản phẩm.
- Đơn vị tính: lấy từ thông tin sản phẩm theo mã sản phẩm.
- Màu sắc: lấy từ thông tin sản phẩm theo mã sản phẩm.
- Kích thước: lấy từ thông tin sản phẩm theo mã sản phẩm.

Tóm lại, bộ thuộc tính sản phẩm không tồn tại dữ liệu bị thiếu.

Dữ liệu bị nhiễu: để giảm nhiễu thì phải làm mịn dữ liệu trước khi làm sạch, việc thực hiện gom cụm dữ liệu khách hàng có xác định loại sản phẩm và giới tính loại sản phẩm là một cách để giảm số chiều dữ liệu hiệu quả.

Làm sạch dữ liệu: về vấn đề dữ liệu không nhất quán, hiện tại hệ thống chạy trên một hệ quản trị cơ sở dữ liệu là SQL Server 2008, mọi thông tin sản phẩm đều được quy định kiểu dữ liệu thống nhất nên thông tin sản phẩm luôn được kiểm tra hợp lệ trước khi thực hiện các giao dịch có liên quan.

Tích hợp dữ liệu: dữ liệu được chọn để tạo dữ liệu huấn luyện thuộc một nguồn duy nhất là hệ quản trị cơ sở dữ liệu SQL Server 2008 nên không cần phải thực hiện tích hợp dữ liệu.

Chuyển hoá dữ liệu: điều kiện trước khi tạo dữ liệu huấn luyện là xác định bộ thuộc tính chuẩn. Các thuộc tính được chọn trong bộ thuộc tính chuẩn đã được kiểm tra và chọn lọc trước nên ta bỏ qua bước này.

Thu gọn dữ liệu: Lựa chọn các tập thuộc tính trước khi thực hiện kết hợp

- Thuộc tính gom cụm: mã cụm, loại sản phẩm, giới tính, loại sản phẩm.
- Thuộc tính sản phẩm: mã sản phẩm, đơn vị tính, thương hiệu, màu sắc, kích thước.

Thực hiện kết hợp hai tập thuộc tính lại thành một tập thuộc tính duy nhất tạo thành bảng dữ liệu huấn luyện cho từng cụm trước khi thực hiện khai thác dữ liệu.

Thuật toán thu gọn dữ liệu:

- Khai báo tham số: mã thực hiện khai thác dữ liệu.
- Duyệt bảng danh sách cụm. Tại từng cụm, tạo cấu trúc bảng dữ liệu huấn luyện tương ứng. Cập nhật thông tin cụm và thông tin sản phẩm vào bảng dữ liệu huấn luyện với cờ phân biệt là TRUE vì dữ liệu phát sinh từ đơn hàng. Cập nhật thông tin sản phẩm còn lại cũng thuộc các loại sản phẩm trong từng cụm với với cờ phân biệt là FALSE vì dữ liệu sản phẩm không phát sinh từ đơn hàng.
- Kết thúc giải thuật, kết quả trả về là:
    - Nội dung bảng danh sách các cụm.
    - Nội dung các bảng dữ liệu huấn luyện tương ứng với từng cụm.

Thực hiện phân lớp: Cập nhật các thông tin:

- Thông tin dữ liệu: nguồn dữ liệu (datasource), bảng dữ liệu huấn luyện (datasource view). Mỗi một cụm khách hàng sẽ có kết quả phân lớp thuộc tính tương ứng cho cụm khách hàng đó.
- Thông tin dữ liệu huấn luyện:
    - Khoá: mã tự phát sinh trong bảng dữ liệu huấn luyện.
    - Dữ liệu đầu vào: loại sản phẩm, thương hiệu kích thước, màu sắc, giới tính sản phẩm, phân lớp xác định (chỉ tồn tại giá trị 2 lớp là mua và không mua).
    - Dữ liệu dự đoán: phân lớp thuộc tính sản phẩm ở hai lớp Mua (TRUE) và Không mua (FALSE).
    - Tham số thực hiện cây quyết định:

      Kiểm soát sự tăng trưởng của cây quyết định (COMPLEXITY_PENALTY).

      Xác định số lượng đầu vào tối đa các thuộc tính mà các thuật toán có thể xử lý (MAXIMUM_INPUT_ATTRIBUTES).

      Xác định số lượng đầu ra tối đa của các thuộc tính trên cây quyết định (MAXIMUM_OUTPUT_ATTRIBUTES).

      Xác định số lượng tối thiểu các trường hợp nút lá để tạo ra một sự rẽ nhánh trong cây quyết định (MINIMUM_SUPPORT). Giá trị tham số này lớn thì cây sẽ tránh bị huấn luyện quá nhiều (over tranining)

      Xác định phương pháp thực hiện phân chia cây (SCORE_METHOD). Các phương pháp như Entropy, Bayesian with K2 Prior, Bayesian Dirichlet Equivalent (BDE) Prior.

Thực hiện khai thác phân lớp thuộc tính sản phẩm trên bảng dữ liệu huấn luyện.

Lấy kết quả trả về: Lấy kết quả sau khi thực hiện khai thác dữ liệu đổ về cơ sở dữ liệu local bao gồm: Thông tin chi tiết thuộc tính sản phẩm của từng cụm.

Thuật toán thực hiện trả kết quả về cơ sở dữ liệu local.

- Khai báo tham số: tên model thực hiện phân lớp.
- Viết câu truy vấn có liên kết đến dự án (project) thực hiện phân lớp để lấy kết quả trả về

- Thông tin từ bảng dữ liệu huấn luyện phân lớp của từng cụm
- Điều kiện là chỉ lấy kết quả dự đoán có mua hàng.
- Sắp xếp danh sách thuộc tính sản phẩm theo thứ tự xác suất giảm dần.
- Lưu thông tin vừa truy vấn vào bảng kết quả phân lớp tương ứng với từng cụm (thông tin tên bảng kết quả phân lớp lấy từ danh sách gom cụm).

- Kết thúc giải thuật, kết quả trả về là: Thông tin chi tiết thuộc tính sản phẩm của từng cụm.

8. ỨNG DỤNG MINH HỌA

8.1 Mô tả ứng dụng

Ứng dụng này cho phép người dùng mạng xã hội Facebook, có tài khoản hệ thống face4shop bằng cách đăng nhập chung từ tài khoản Facebook mà không cần tạo thêm một tài khoản nữa.

Người dùng có thể mua hàng, bán hàng trên hệ thống website face4shop.com. Mọi thông tin của người dùng thao tác trên sản phẩm như chọn thích (Like), nhận xét sản phẩm đều được thông báo trên tường (Wall) của mạng xã hội Facebook. Việc này nhằm mục đích, thông báo với bạn bè của người đó về sản phẩm trong hệ thống face4shop, bên cạnh đó còn thể hiện việc kích thích mua hàng của những khách hàng tiềm năng.

Ứng dụng còn thực hiện gợi ý tiếp thị sản phẩm được thể hiện ở chức năng cơ bản như hiển thị sản phẩm có chung thuộc tính với sản phẩm mà người dùng đang xem theo tiêu chí giống nhau về loại sản phẩm, cùng thuộc một thương hiệu hoặc cùng bán ở một shop. Bên cạnh đó, hệ thống còn gợi ý những sản phẩm đi kèm với sản phẩm đi xem tạo nên cái nhìn tổng thể cho người mua hàng khi họ không biết thông tin sản phẩm họ mua sẽ đi chung với sản phẩm nào. Chức năng đánh giá, nhận xét mang lại cho người dùng thông tin sản phẩm của khách hàng đã qua sử dụng.

Với các gợi ý các tiêu chí chỉ quan tâm đến tất cả mọi đối tượng khách hàng. Vậy còn cụ thể một khách hàng mua hàng thì như thế nào. Chức năng nâng cao sẽ hỗ trợ phần này. Chức năng tạo cho người dùng thông tin tặng quà sinh nhật cho bạn bè, cho người thân theo mối quan hệ. Chức năng được thực hiện bằng khai thác dữ liệu, từ những thông tin khách hàng,

đơn hàng tồn tại trên hệ thống. Từ đó, khai phá dữ liệu mua hàng của họ, tạo ra những quy luật nhằm đưa ra sự hỗ trợ quyết định giúp cho khách hàng có sự lựa chọn tốt hơn.

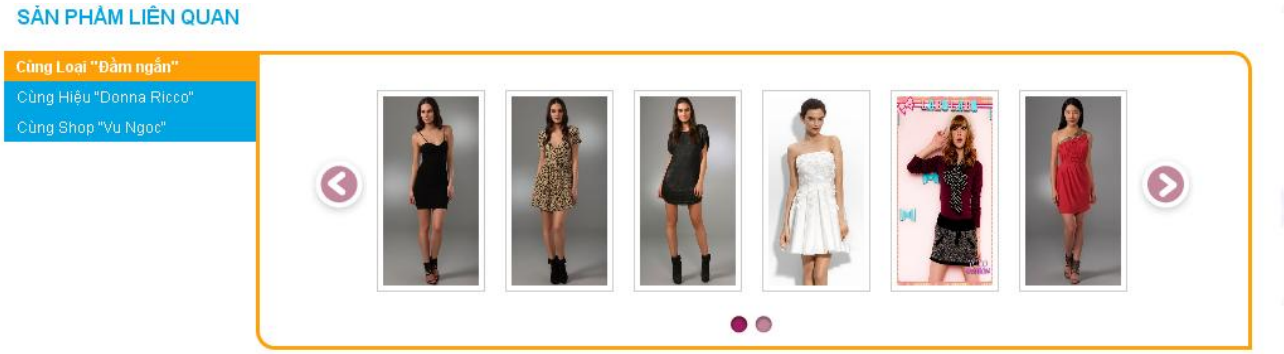
8.2 Chức năng cơ bản

8.2.1 Hiển thị sản phẩm chung một thuộc tính (Related products)

Giao diện: thể hiện bên dưới màn hình tại thông tin chi tiết của một sản phẩm bất kỳ.

Ví dụ: sản phẩm đầm ngắn hiệu Donna

http://www.face4shop.com/product/view/499#relatedProductCategory



Hình 21 – Giao diện chức năng related products

Các tiêu chí được chọn cho related product: loại sản phẩm, thương hiệu, cửa hàng cá nhân. Ví dụ: sản phẩm đang xem là một đôi giày cao gót hiệu Bonicii của cửa hàng VuNgocShop, sản phẩm liên quan sẽ thể hiện là những sản phẩm cùng loại sản phẩm giày cao gót, những sản phẩm cùng thương hiệu Bonicii, những sản phẩm thuộc cửa hàng VuNgocShop.

- Loại sản phẩm: các sản phẩm trong danh sách related product phải cùng loại sản phẩm và giới tính sản phẩm với sản phẩm đang xem, những thông tin khác như thương hiệu, màu sắc, kích thước... của từng sản phẩm nếu giống với thông tin sản phẩm đang xem sẽ được sắp xếp theo thứ tự giống nhau nhiều nhất giảm dần.
- Thương hiệu: các sản phẩm trong danh sách related product phải cùng thương hiệu và giới tính sản phẩm với sản phẩm đang xem, những thông tin khác như loại sản phẩm, màu sắc, kích thước,.. của từng sản phẩm nếu giống với thông tin sản phẩm đang xem sẽ được sắp xếp theo thứ tự giống nhau nhiều nhất giảm dần.

- Cửa hàng cá nhân: các sản phẩm chung một cửa hàng cá nhân với sản phẩm đang xem, những thông tin khác như loại sản phẩm, thương hiệu, màu sắc, kích thước... của từng sản phẩm nếu giống với thông tin sản phẩm đang xem sẽ được sắp xếp theo thứ tự giống nhau nhiều nhất giảm dần.
- Về giới tính sản phẩm là một trường hợp đặc biệt: nếu sản phẩm chỉ dành cho nam hoặc nữa, nếu tồn tại sản phẩm dành cho nam và nữa và thoả điều kiện của ba tiêu chí trên thì cũng được thể hiện.

Giải pháp thực hiện

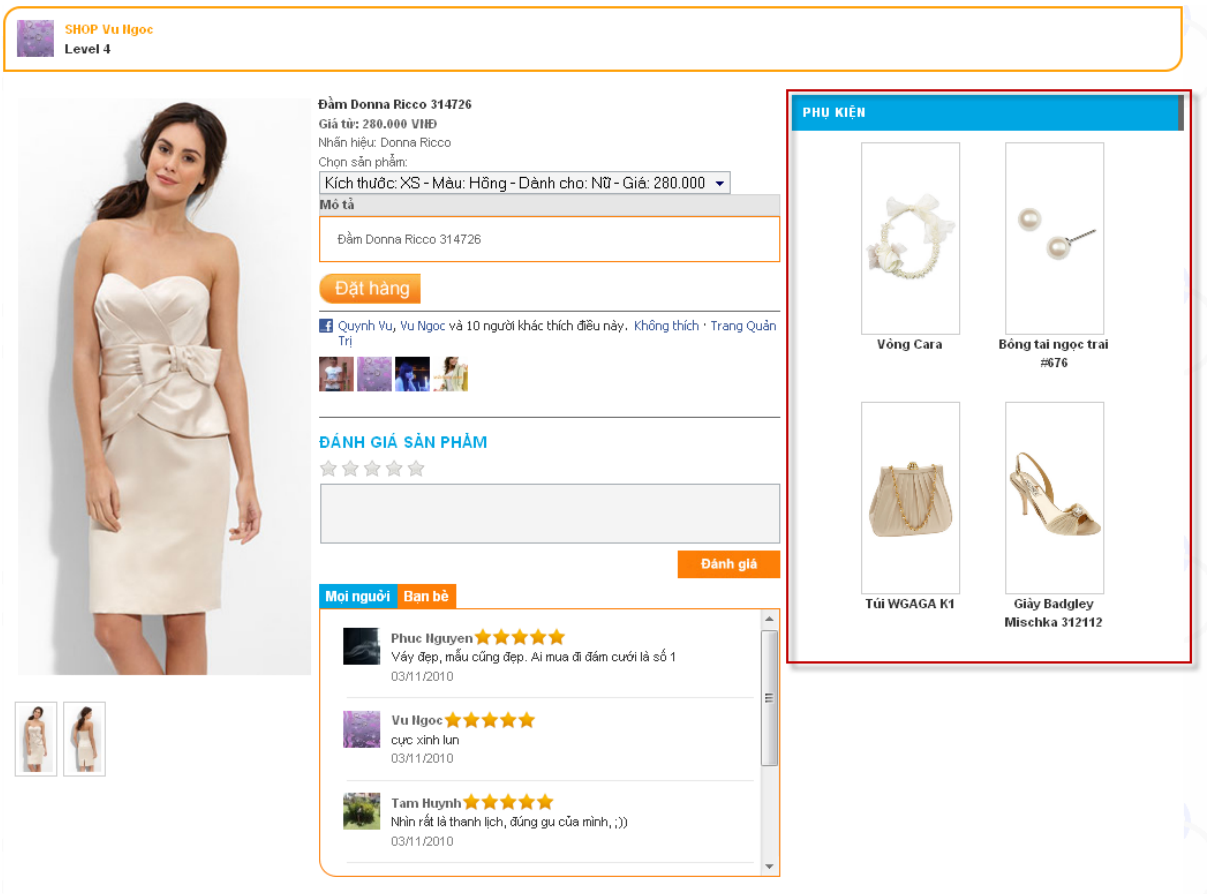
- Khai báo tham số: mã giao dịch với khách hàng (phân biệt hàng hoá giữa các session khác nhau), mã sản phẩm (mã sản phẩm đang được xem), giới tính sản phẩm (giới tính sản phẩm đang được xem).
- Kiểm tra tồn tại bảng related product cho một session, thực hiện tạo cấu trúc bảng gồm mã sản phẩm, cờ phân biệt ba tiêu chí thể hiện, ba thuộc tính loại, thương hiệu, cửa hàng tương ứng với cờ, biến đếm.
- Lần lượt cập nhật thông tin sản phẩm theo ba tiêu chí cùng loại sản phẩm, cùng thương hiệu, cùng cửa hàng, biến đếm lúc này tăng lên 1 (vì giống với sản phẩm đang xem 1 thuộc tính)
- Tạo biến con trỏ duyệt danh sách bảng thuộc tính sản phẩm. Lần lượt kiểm tra mã thuộc tính sản phẩm có tồn tại như một trường dữ liệu trong bảng related product hay không, nếu không thì thực hiện tạo mới một trường dữ liệu có tên là mã thuộc tính với kiểu dữ liệu là kiểu logic. Đồng thời thực hiện cập nhật cho trường này giá trị là TRUE/FALSE nếu có giá trị thuộc tính giống/không giống với giá trị thuộc tính đang xem, đồng thời biến đếm tăng lên 1 nếu giá trị giống nhau. Việc cập nhật này được thực hiện cho cả ba tiêu chí. Kết thúc duyệt con trỏ là kết thúc việc tạo thêm trường dữ liệu mới và cập nhật thông tin giá trị thuộc tính giống nhau.
- Kết quả trả về sau khi thực hiện là nội dung bảng dữ liệu related product theo ba tiêu chí và sắp xếp theo thứ tự giống nhau nhiều nhất giảm dần.
- Ưu điểm của thuật toán này là số thuộc tính sản phẩm nếu có tăng lên cũng không làm thay đổi nội dung thuật toán, không phải customize lại cho người dung.

8.2.2 Hiển thị danh sách sản phẩm liên quan (Related Accessories)

Hiển thị danh sách sản phẩm phụ kiện đi kèm với sản phẩm đang xem. Ví dụ sản phẩm đang xem là một đôi giày cao gót, sản phẩm phụ kiện đi kèm có thể là váy, áo, túi xách,…

Giao diện: thể hiện bên phải màn hình tại thông tin chi tiết của một sản phẩm bất kỳ.

Ví dụ: sản phẩm đầm ngắn Donna
http://www.face4shop.com/product/view/499#relatedProductCategory

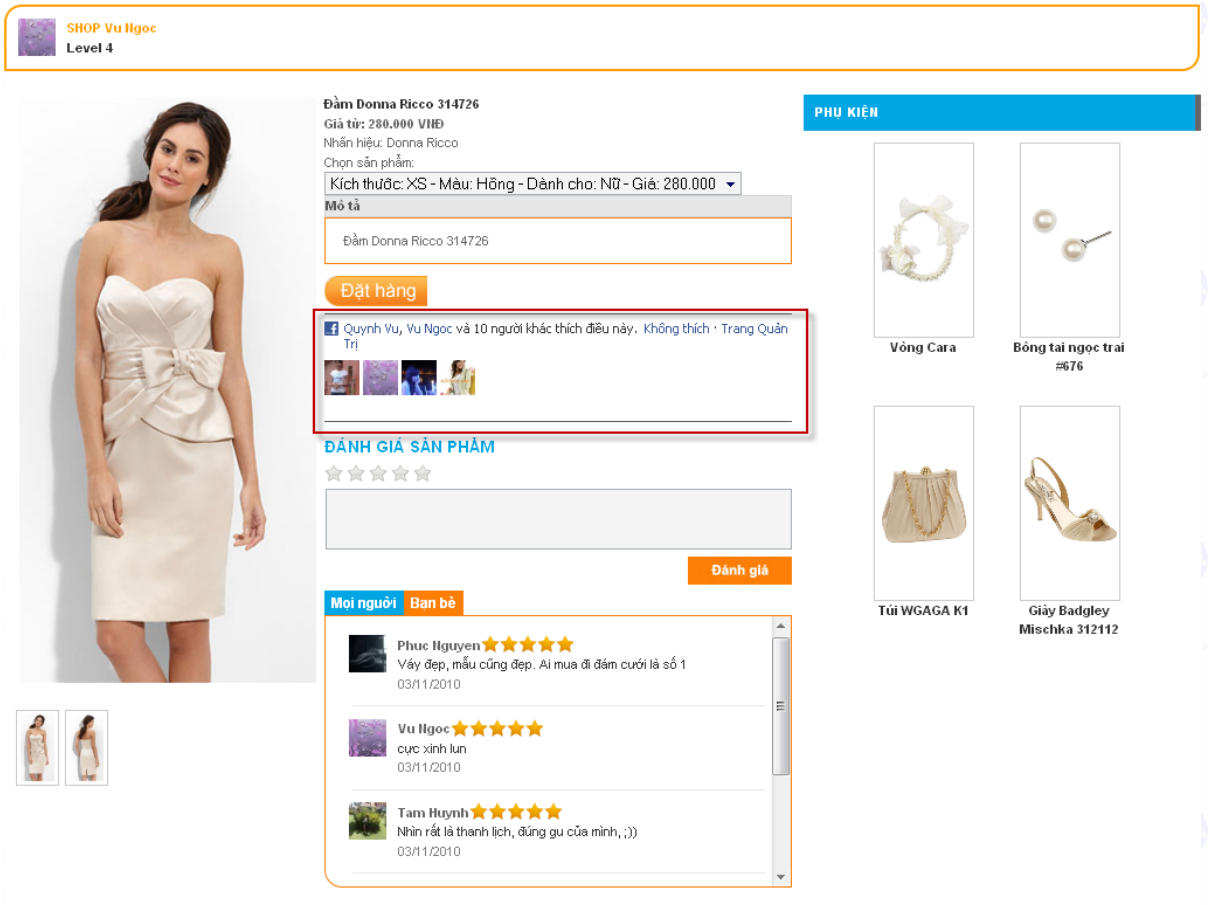


Hình 22 – Giao diện chức năng related accessories

8.2.3 Chức năng Like của Facebook

Nhằm kích thích người dùng Facebook tham gia vào hệ thống Face4shop.com, màn hình chi tiết sản phẩm thể hiện chức năng Like của Facebook để người dùng có thể click vào nếu thích. Thông tin này sẽ được Facebook ghi nhận và hiển thị trên trang Facebook cá nhân của

người dùng và bạn bè của họ, từ đó nhiều người sẽ biết đến các sản phẩm từ hệ thống face4shop.com.



Hình 23 – Giao diện chức năng Like sản phẩm

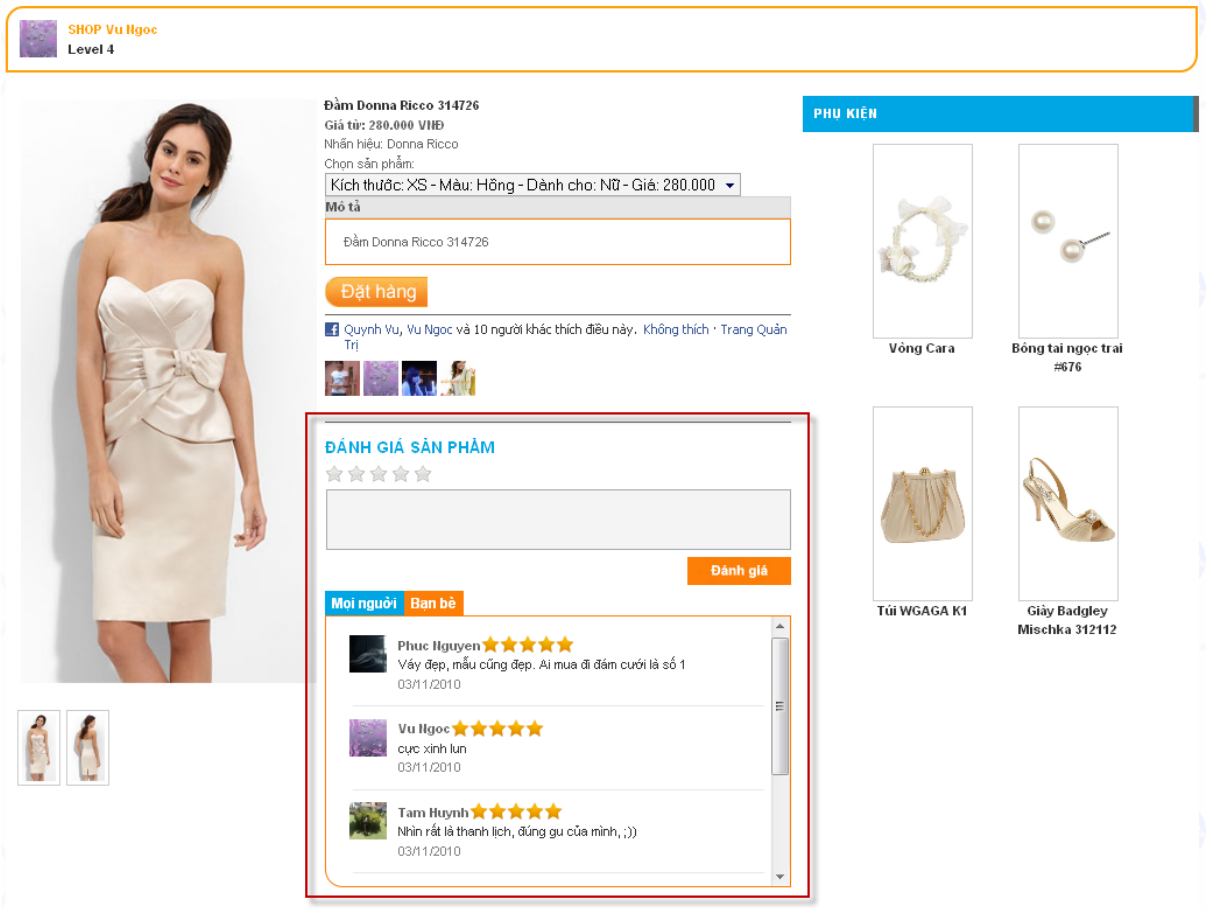
8.2.4 Đánh giá sản phẩm

Thể hiện thông tin đánh giá sản phẩm của mọi người và sắp xếp theo hai tiêu chí là Mọi người và Bạn bè.

- Mọi người: nếu người dùng chưa đăng nhập vào hệ thống và chọn xem một sản phẩm bết kỳ thì thông tin đánh giá sẽ thể hiện đánh giá của tất cả mọi người về sản phẩm đang xem và người dùng không thấy được đánh giá của bạn bè mình.
- Bạn bè: nếu người dùng đã đăng nhập vào hệ thống và chọn xem một sản phẩm bất kỳ thì sẽ thấy được thông tin đánh giá sản phẩm của Mọi người (trình bày ở trên), ngoài

ra còn thấy được thông tin đánh giá sản phẩm của bạn bè mình về sản phẩm đang xem.

Chức năng này kích thích người dùng quan tâm đến sản phẩm của face4shop.com đồng thời cũng có thể tạo được niềm tin của người mua hàng vì thông qua nội dung đánh giá sản phẩm của mọi người và bạn bè của mình, người dùng có thể biết được chất lượng của sản phẩm trước khi chọn.



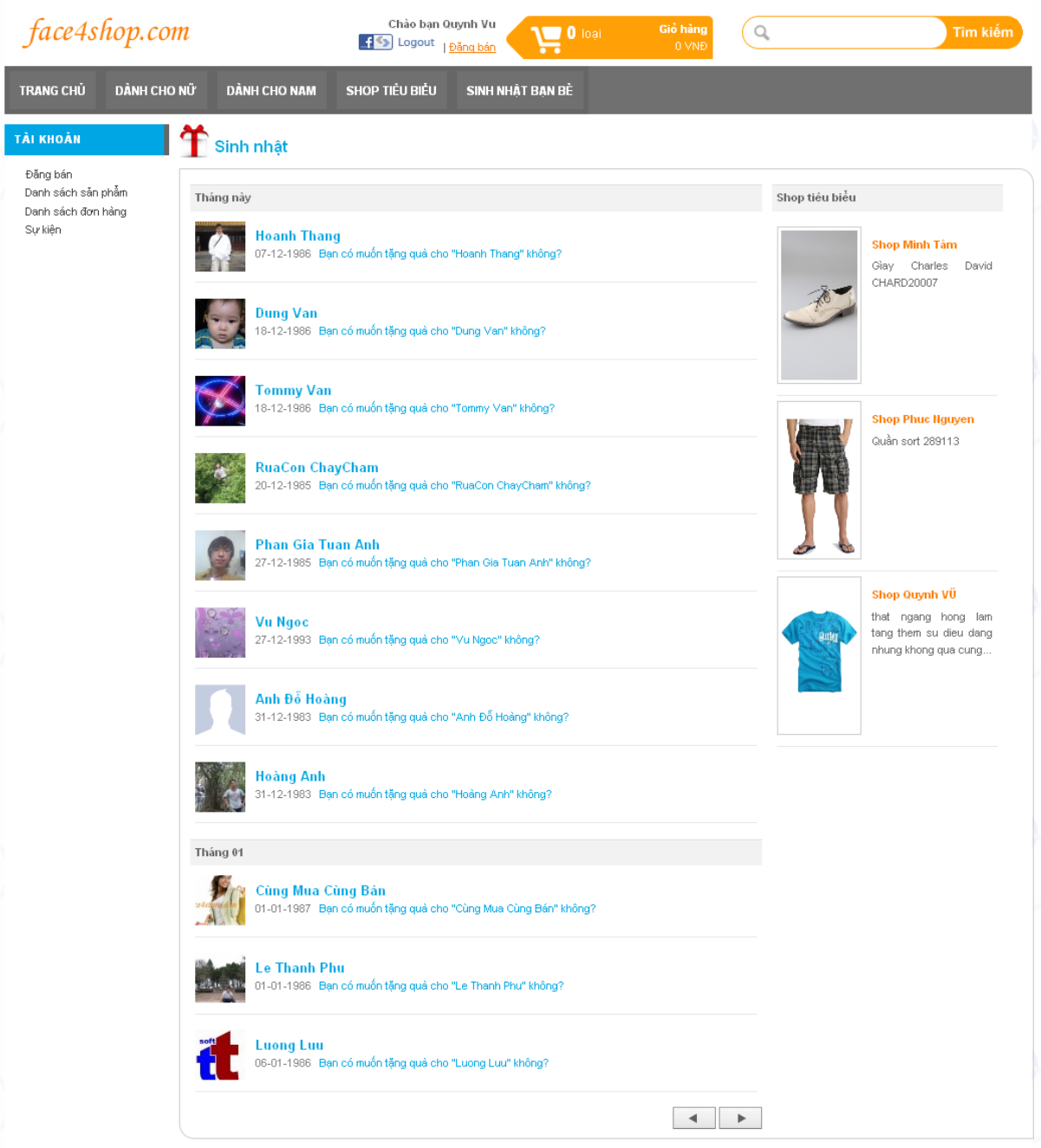
Hình 24 – Giao diện chức năng đánh giá và nhận xét sản phẩm

8.3 Chức năng nâng cao

Tên chức năng: "Gợi ý mua hàng cho bạn bè của khách hàng vào dịp sinh nhật".

Ví dụ: Quỳnh là người đi mua hàng, Phúc là bạn của Quỳnh và sắp đến sinh nhật của Phúc, hệ thống sẽ thông báo cho Quỳnh và gợi ý cho Quỳnh những món quà có thể mua cho Phúc

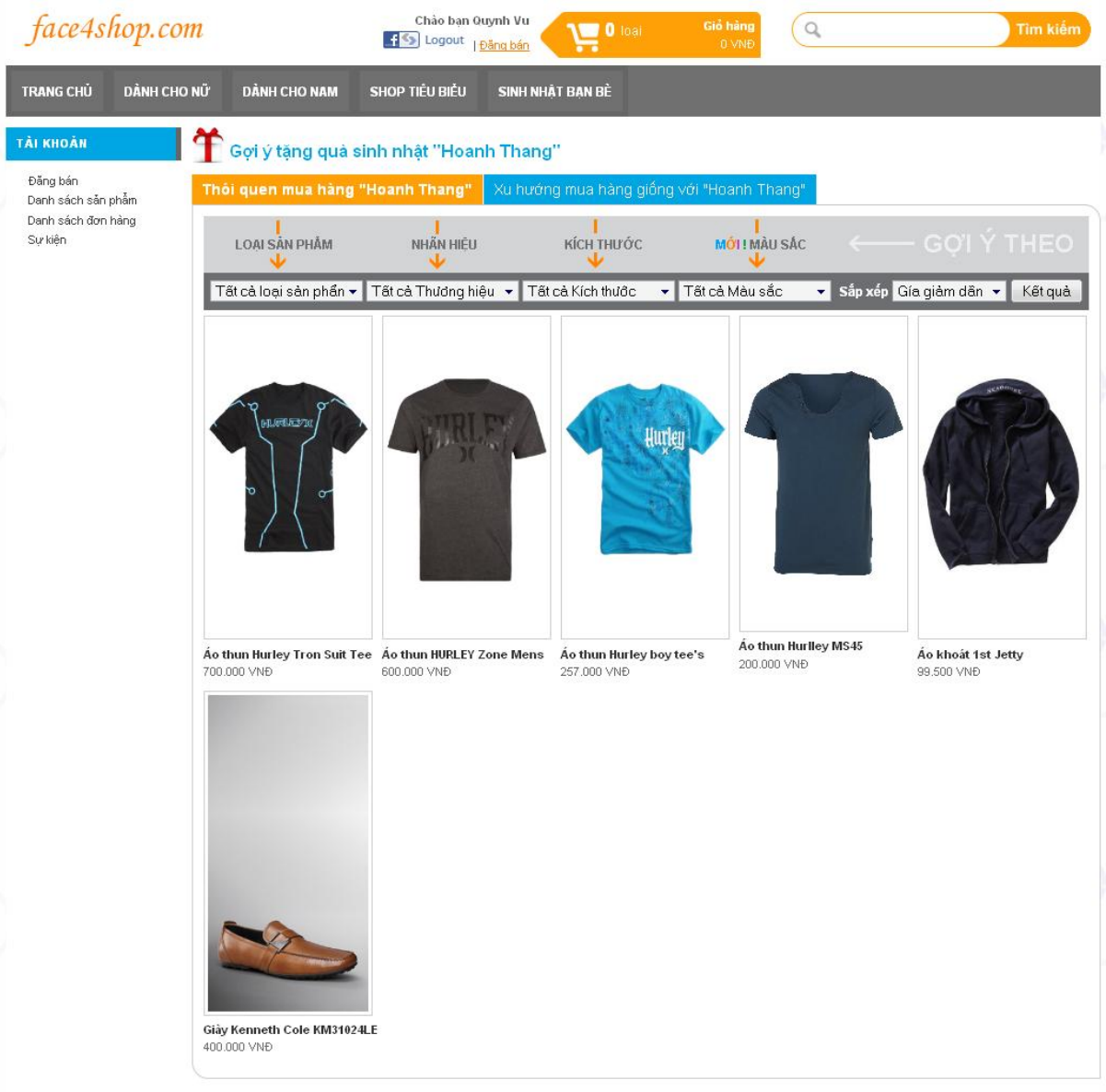
(dựa vào thông tin của Phúc như giới tính độ tuổi…, hệ thống sẽ gợi ý những sản phẩm mà Phúc thích hoặc những người giống Phúc thích)



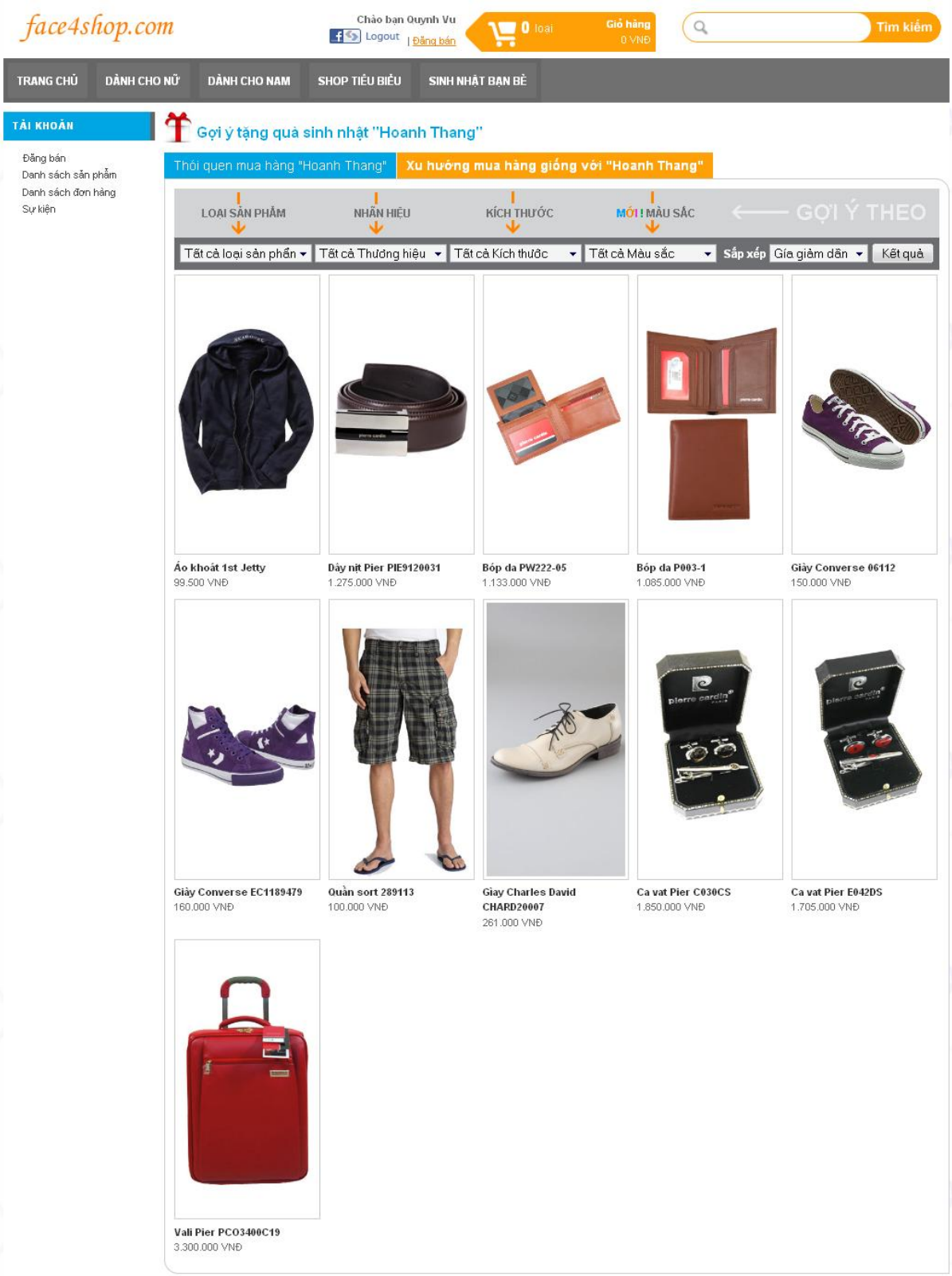
Hình 25 – Giao diện gợi ý danh sách bạn bè theo tháng sinh nhật

Trường hợp nhiều người mua hàng cùng có một người bạn sắp đến sinh nhật, hệ thống gợi ý như nhau sẽ dẫn đến người bạn đó nhận cùng một món quà từ nhiều người khác nhau. Cách

giải quyết là hệ thống sẽ kiểm tra các mối quan hệ có liên quan đến người mua và người nhân quà và từ đó đưa ra cảnh báo nếu người nhận qua đã được nhận từ một người khác và cũng là bạn của người nhận.



Hình 26 – Giao diện thuộc tính sản phẩm theo Online tab



Hình 27 – Giao diện thuộc tính sản phẩm theo Offline tab

9. KẾT QUẢ & ĐÁNH GIÁ

9.1 Thống kê dữ liệu thực từ face4shop.com

Theo thống kê đơn hàng phát sinh từ ngày 27/10/2010 đến ngày 5/12/2010 được lấy từ hệ thống bán hàng trực tuyến face4shop.com

- Số đơn hàng: 270 đơn hàng.
- Số khách hàng tham gia mua hàng: 61 khách.
- Độ tuổi khách hàng tham gia mua hàng: chủ yếu từ 17 đến 30 tuổi.
- Số mặt hàng phát sinh trong đơn hàng: 157 sản phẩm.

Từ thống kê này cho thấy:

- Dữ liệu đơn hàng và số khách hàng mua không không nhiều trong thời gian ngắn, độ tuổi khách mua hàng chủ yếu từ 17 đến 30 tuổi.
- Khi gợi ý cho khách hàng sẽ gặp khó khăn nếu như khách hàng không nằm trong độ tuổi đã được khai thác.
- Độ chính xác của thuật toán sẽ giảm.

Như vậy, với dữ liệu quá ít, sẽ làm ảnh hưởng đến việc thực hiện khai phá dữ liệu tìm gợi ý mua hàng. Giải pháp vấn đề này, cụ thể là giải pháp đánh giá độ chính xác của thuật toán:

- Dữ liệu: xây dựng dữ liệu random theo kịch bản bao gồm
    - Dữ liệu khách hàng: Xây dựng dữ liệu random khách hàng ở nhiều độ tuổi, dữ liệu random được thực hiện dựa theo cơ cấu độ tuổi của Bách khoa toàn thư mở (Kết quả sơ bộ điều tra dân số năm 2009 _ Mục 4 _ Tỷ số giới tính) (7).
    - Dữ liệu đơn hàng: bao gồm hai nhóm đơn hàng chính là đơn hàng không định hướng sản phẩm và đơn hàng có định hướng sản phẩm (*sản phẩm có định hướng được chọn từ thống kê sản phẩm được mua nhiều nhất* trên đơn hàng thực của face4shop.com).
- *Dữ liệu huấn luyện*: thực hiện random hai phần ba đơn hàng của hai loại đơn hàng theo kịch bản để tạo dữ liệu huấn luyện.
- *Đánh giá các phương pháp thực hiện khai thác dữ liệu*: tiến hành thực hiện trên ba phương pháp gồm:

- Phương pháp gom cụm: thực hiện khai thác dữ liệu huấn luyện chỉ bằng phương pháp gom cụm.
- Phương pháp cây quyết định: thực hiện khai thác dữ liệu huấn luyện chỉ bằng phương pháp cây quyết định.
- Kết hợp hai phương pháp gồm gom cụm và cây quyết định: thực hiện khai thác dữ liệu huấn luyện bằng cách kết hợp cả hai phương pháp gom cụm và cây quyết định.

- *Dữ liệu kiểm tra*: dữ liệu kiểm tra được tạo từ một phần ba dữ liệu đơn hàng random còn lại để kiểm tra độ chính xác của thuật toán.

9.2 Kịch bản xây dựng dữ liệu ngẫu nhiên

9.2.1 Dữ liệu ngẫu nhiên khách hàng

Thông tin dữ liệu ngẫu nhiên, giả định:	**Tỉ lệ %**
Giới tính	Nam (49%), Nữ (51%)
Nhóm tuổi	Giả định chia làm 4 nhóm tuổi gồm nhóm 17-32 (45%), nhóm 33-50 (30%), nhóm 51-60 (20%), và nhóm 61-70 (5%).
Mối quan hệ giữa các khách hàng: giả định các quan hệ sau	
Quan hệ bạn bè (khách hàng có mối quan hệ bạn bè)	Nhóm 17-32 (65%), nhóm 33-50 (20%), nhóm 51-60 (10%), và nhóm 61-70 (5%)
Quan hệ người yêu (khách hàng có mối quan hệ là người yêu của nhau)	Nhóm 17-32 (65%), nhóm 33-50 (20%), nhóm 51-60 (10%), và nhóm 61-70 (5%).
Quan hệ vợ chồng (khách hàng có mối quan hệ vợ chồng với nhau)	Nhóm 17-32 (65%), nhóm 33-50 (20%), nhóm 51-60 (10%), và nhóm 61-70 (5%).
Quan hệ anh/em trai (khách hàng có mối quan hệ anh/em)	Nhóm 17-32 (50%), nhóm 33-50 (35%), nhóm 51-60 (10%), và nhóm 61-70 (5%).

Quan hệ chị/em gái (khách hàng có mối quan hệ chị/em)	Nhóm 17-32 (50%), nhóm 33-50 (35%), nhóm 51-60 (10%), và nhóm 61-70 (5%).
Quan hệ cha/mẹ - con trai/con gái (khách hàng có mối quan hệ cha/mẹ - con trai/con gái)	Nhóm 17-32 (80%), nhóm 33-50 (20%)

9.2.2 Dữ liệu ngẫu nhiên đơn hàng

9.2.2.1 Đơn hàng không định hướng sản phẩm

Thuộc tính giả định	**Tỉ lệ %**
Đơn hàng: giả định theo hai mối quan hệ	
Bạn bè mua cho nhau	Giả định 70% trên tổng số đơn hàng là bạn bè mua cho nhau gồm nhóm 17-32 (65%), nhóm 33-50 (20%), nhóm 51-60 (10%), và nhóm 61-70 (5%).
Cha mẹ mua cho con	Giả định 30% trên tổng số đơn hàng là cha mẹ mua cho con ở nhóm tuổi 17-32
Sản phẩm: không định hướng sản phẩm cho đơn hàng	
Sản phẩm không định hướng	Random bất kỳ sản phẩm có trong hệ thống của face4shop.com

9.2.2.2 Đơn hàng có định hướng sản phẩm

Thuộc tính giả định	Tỉ lệ %
Đơn hàng: giả định theo hai mối quan hệ	
Bạn bè mua cho nhau	Giả định 70% trên tổng số đơn hàng là bạn bè mua cho nhau gồm nhóm 17-32 (65%), nhóm 33-50 (20%), nhóm 51-60 (10%), và nhóm 61-70 (5%).
Cha mẹ mua cho con	Giả định 30% trên tổng số đơn hàng là cha mẹ mua cho con ở nhóm tuổi 17-32
Sản phẩm: định hướng sản phẩm cho đơn hàng	
Sản phẩm có định hướng	Dựa vào thống kê loại sản phẩm phát sinh trong dữ liệu đơn hàng thực để thực hiện random sản phẩm theo thứ tự từ cao xuống thấp tương ứng với % đơn hàng quy định trên phần đơn hàng.

9.3 Thông tin dữ liệu huấn luyện

Thực hiện ngẫu nhiên hai phần ba đơn hàng theo kịch bản bao gồm đơn hàng không định hướng sản phẩm và đơn hàng có định hướng sản phẩm để tạo dữ liệu huấn luyện. Thực hiện khai thác dữ liệu theo ba phương pháp gồm:

9.3.1 Phương pháp gom cụm – Clustering

Bảng thống kê thông tin dữ liệu huấn luyện để thực hiện khai thác dữ liệu theo phương pháp gom cụm.

Tiêu đề	Tên thuộc tính	Giá trị thuộc tính
Số thuộc tính	11	
Danh sách thuộc tính	Tuổi	17, 21 – 30, 110
	Tháng sinh nhật	1-12
	Giới tính	Nam – nữ
	Tháng/quý mua hàng	4
	Năm mua hàng	2010
	Loại sản phẩm	Áo, quần, váy, đầm, giày, túi xách, thắt lưng, phụ kiện
	Thương hiệu	3.1 Phillip Lim, 7 For All Mankind, Adidas, Alexander Wang, An Phước, …
	Màu sắc	Bạc, nâu, cam, đen, đỏ, hồng, tím, trắng, vàng, xám, xanh.
	Kích thước	5-9, 24-30, 37-39, XS, S, M, L, Một size.
	Giới tính sản phẩm	Nam, nữ, nam và nữ
	Đơn vị tính sản phẩm	Cái, đôi
Dự đoán	Cụm khách hàng	

9.3.2 Phương pháp cây quyết định– Decision Tree

Bảng thống kê thông tin dữ liệu huấn luyện để thực hiện khai thác dữ liệu theo phương pháp cây quyết định.

Tiêu đề	Tên thuộc tính	Giá trị thuộc tính
Số thuộc tính	12	
Danh sách thuộc tính	Tuổi	17, 21 – 30, 110
	Tháng sinh nhật	1-12
	Giới tính	Nam – nữ
	Tháng/quý mua hàng	4
	Năm mua hàng	2010
	Loại sản phẩm	Áo, quần, váy, đầm, giày, túi xách, thắt lưng, phụ kiện
	Thương hiệu	3.1 Phillip Lim, 7 For All Mankind, Adidas, Alexander Wang, An Phước, …
	Màu sắc	Bạc, nâu, cam, đen, đỏ, hồng, tím, trắng, vàng, xám, xanh.
	Kích thước	5-9, 24-30, 37-39, XS, S, M, L, Một size.
	Giới tính sản phẩm	Nam, nữ, nam và nữ
	Đơn vị tính sản phẩm	Cái, đôi

	Tuỳ chọn mua	1-Có mua, 0-Không mua
Dự đoán	Mua	

9.3.3 Kết hợp hai phương pháp gom cụm và cây quyết định

Bảng thống kê thông tin dữ liệu huấn luyện để thực hiện khai thác dữ liệu theo phương pháp gom cụm.

Tiêu đề	**Tên thuộc tính**	**Giá trị thuộc tính**
Số thuộc tính	7	
Danh sách thuộc tính	Tuổi	17, 21 – 30, 110
	Tháng sinh nhật	1-12
	Giới tính	Nam – nữ
	Tháng/quý mua hàng	4
	Năm mua hàng	2010
	Loại sản phẩm	Áo, quần, váy, đầm, giày, túi xách, thắt lưng, phụ kiện
	Giới tính sản phẩm	Nam, nữ, nam và nữ
Dự đoán	Cụm	

Từ kết quả khai thác dữ liệu theo phương pháp gom cụm, tạo bảng thông tin dữ liệu huấn luyện tương ứng với từng cụm để thực hiện khai thác dữ liệu theo phương pháp cây quyết định.

Tiêu đề	Tên thuộc tính	Giá trị thuộc tính
Số thuộc tính	7	
Danh sách thuộc tính tương ứng với từng cụm	Loại sản phẩm	Áo, quần, váy, đầm, giày, túi xách, thắt lưng, phụ kiện
	Thương hiệu	3.1 Phillip Lim, 7 For All Mankind, Adidas, Alexander Wang, An Phước, …
	Màu sắc	Bạc, nâu, cam, đen, đỏ, hồng, tím, trắng, vàng, xám, xanh.
	Kích thước	5-9, 24-30, 37-39, XS, S, M, L, Một size.
	Giới tính sản phẩm	Nam, nữ, nam và nữ
	Đơn vị tính sản phẩm	Cái, đôi
	Tuỳ chọn mua	1-Có mua, 0-Không mua
Dự đoán	Mua	

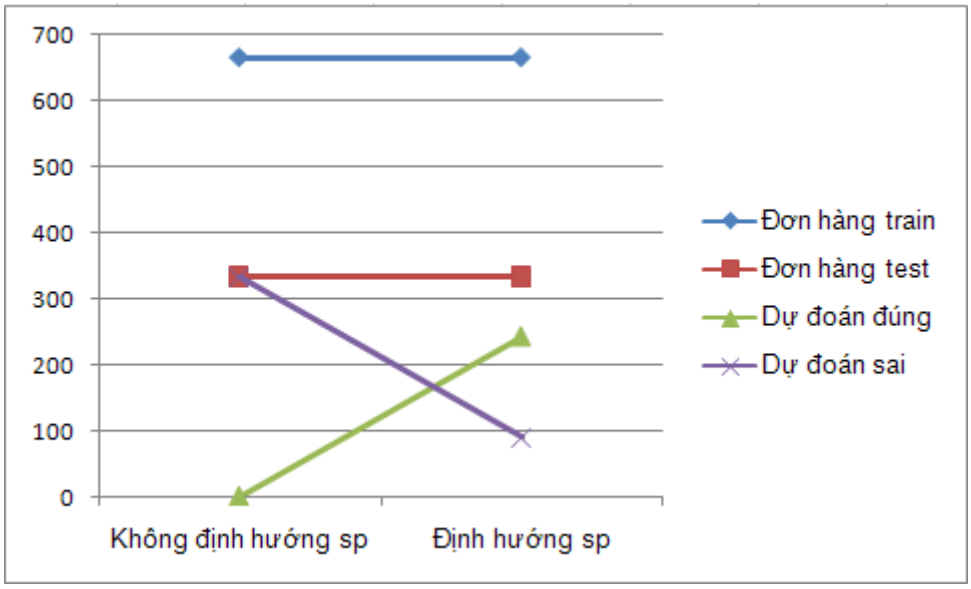
9.4 Kết quả và đánh giá

9.4.1 Đánh giá dựa trên một phương pháp khai thác dữ liệu

9.4.1.1 Phương pháp gom cụm

Bảng so sánh kết quả đánh giá tỷ lệ dự đoán đúng/sai trên hai nhóm dữ liệu đơn hàng không định hướng sản phẩm và dữ liệu đơn hàng có định hướng sản phẩm theo phương pháp gom cụm.

Loại đơn hàng	Đơn hàng train	Đơn hàng test	Dự đoán đúng	Dự đoán sai
Không định hướng	666	334	1 (0,3%)	333 (99,7%)
Định hướng	666	334	244 (73,05%)	90 (26,95%)



Hình 28 –Biểu đồ tỷ lệ dự đoán đúng/sai của hai nhóm dữ liệu theo phương pháp gom cụm

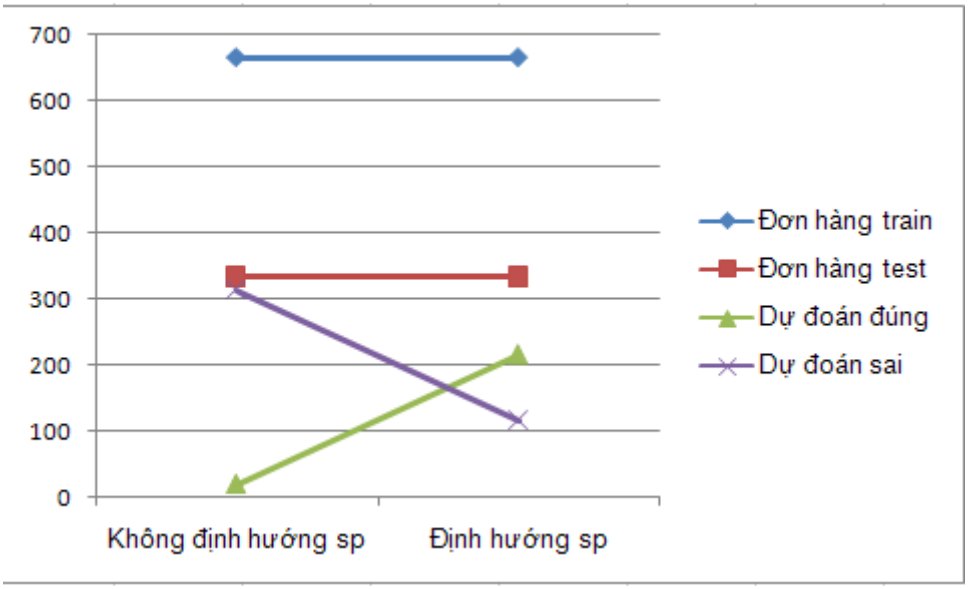
Nhận xét: Đặc điểm của phương pháp gom cụm là tạo ra các cụm có đặc điểm giống nhau, nếu dữ liệu huấn luyện đã tồn tại những mối liên hệ bên trong giữa các thuộc tính đầu vào thì kết quả gom cụm sẽ chính xác hơn. Với dữ liệu đơn hàng không định hướng sản phẩm, thông tin sẽ bị dàn trải dẫn đến kết quả là thông tin các cụm không chính xác, kết quả cho thấy là việc dự đoán chỉ có 1% tính trên 334 đơn hàng kiểm tra. Nhưng với đơn hàng đã được định hướng sản phẩm, giả định này gần với thực tế là xu hướng mua hàng của một số nhóm người ở một số độ tuổi xác định, việc lấy một phần ba đơn hàng còn lại để kiểm tra đã cho kết quả khả quan hơn là 73,05% dự đoán đúng tính trên 334 đơn hàng kiểm tra. Tỷ lệ giữa dự đoán sai và đúng có chiều hướng từ sai nhiều chuyển thành đúng nhiều, theo số liệu là tỷ lệ sai

nhiều gấp 333 lần so với tỷ lệ đúng ở dữ liệu đơn hàng không định hướng sản phẩm trở thành tỷ lệ đúng nhiều gấp 2,7 lần so với tỷ lệ sai ở đơn hàng định hướng sản phẩm.

9.4.1.2 Phương pháp cây quyết định

Bảng so sánh kết quả đánh giá tỷ lệ dự đoán đúng/sai trên hai nhóm dữ liệu đơn hàng không định hướng sản phẩm và dữ liệu đơn hàng có định hướng sản phẩm theo phương pháp cây quyết định.

Loại đơn hàng	Đơn hàng train	Đơn hàng test	Dự đoán đúng	Dự đoán sai
Không định hướng	666	334	20 (6%)	314 (94%)
Định hướng	666	334	217 (65%)	117 (35%)



Hình 29 –Biểu đồ tỷ lệ dự đoán đúng/sai của hai nhóm dữ liệu theo phương pháp cây quyết định

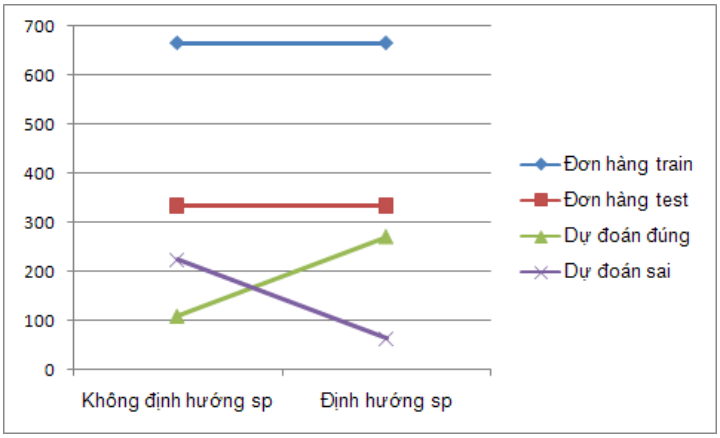
Nhận xét: Đặc điểm của phương pháp cây quyết định là sự chênh lệch tỷ lệ dự đoán của nhóm dữ liệu có khả năng và không có khả năng nằm bên trong tập dữ liệu dẫn đến kết quả dự đoán có chính xác hay không. Với dữ liệu đơn hàng không định hướng sản phẩm, thông

tin sẽ bị dàn trải dẫn đến kết quả là khả năng dự đoán không chính xác, chỉ có 6% dự đoán đúng tính trên 334 đơn hàng kiểm tra. Nhưng với đơn hàng đã được định hướng sản phẩm, giả định này gần với thực tế là xu hướng mua hàng của một số nhóm người ở một số độ tuổi xác định, việc lấy một phần ba đơn hàng còn lại để kiểm tra đã cho kết quả khả quan hơn là 65% dự đoán đúng tính trên 334 đơn hàng kiểm tra. Tỷ lệ giữa dự đoán sai và đúng có chiều hướng từ sai nhiều chuyển thành đúng nhiều, theo số liệu là tỷ lệ sai nhiều gấp 15,7 lần so với tỷ lệ đúng ở dữ liệu đơn hàng không định hướng sản phẩm trở thành tỷ lệ đúng nhiều gấp khoảng 2 lần so với tỷ lệ sai ở đơn hàng định hướng sản phẩm.

Nếu so sánh giữa hai phương pháp gom cụm và cây quyết địnhthì khả năng dự đoán của phương pháp gom cụm cao hơn (73,05%) so với phương pháp cây quyết định (65%).

9.4.1.3 Phương pháp gom cụm và cây quyết định

Loại đơn hàng	Đơn hàng train	Đơn hàng test	Dự đoán đúng	Dự đoán sai
Không định hướng	666	334	110 (33%)	224 (67%)
Định hướng	666	334	271 (81%)	63 (19%)



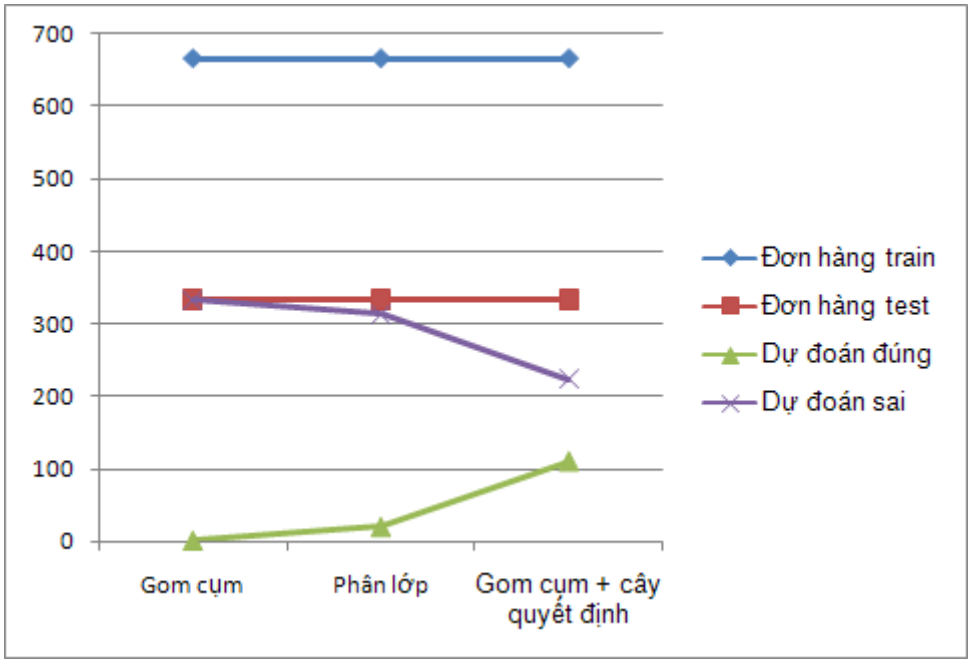
Hình 30 –Biểu đồ tỷ lệ dự đoán đúng/sai của hai nhóm dữ liệu theo phương pháp kết hợp giữa gom cụm và cây quyết định

Nhận xét: kết hợp hai phương pháp để thực hiện khai thác dữ liệu cho thấy với dữ liệu đơn hàng không định hướng sản phẩm, chỉ có 33% dự đoán đúng tính trên 334 đơn hàng kiểm tra. Nhưng với đơn hàng đã được định hướng sản phẩm, kết quả khả quan hơn là 81% dự đoán đúng tính trên 334 đơn hàng kiểm tra. Tỷ lệ giữa dự đoán sai và đúng có chiều hướng từ sai nhiều chuyển thành đúng nhiều, theo số liệu là tỷ lệ sai nhiều gấp 2,03 lần so với tỷ lệ đúng ở dữ liệu đơn hàng không định hướng sản phẩm trở thành tỷ lệ đúng nhiều gấp khoảng 4,3 lần so với tỷ lệ sai ở đơn hàng định hướng sản phẩm.

Nếu so sánh với hai phương pháp gom cụm và cây quyết địnhthì phương pháp kết hợp cả hai là tối ưu nhất vì tỷ lệ dự đoán đúng cao nhất (81%) so với gom cụm (73,05%) và cây quyết định (65%).

9.4.2 Đánh giá dựa trên thông tin đơn hàng không định hướng sản phẩm.

Phương pháp	Đơn hàng train	Đơn hàng test	Dự đoán đúng	Dự đoán sai
Gom cụm	666	334	1 (0,3%)	333 (99,7%)
Cây quyết định	666	334	20 (6%)	314 (94%)
Gom cụm + cây quyết định	666	334	110 (33%)	224 (67%)

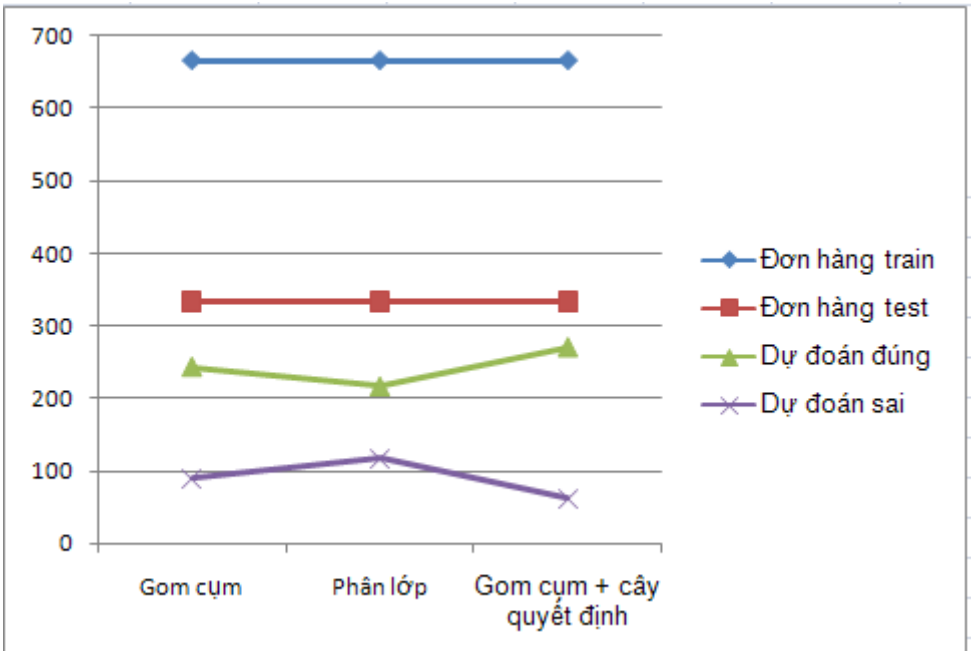


Hình 31 –Biểu đồ đánh giá dựa trên thông tin đơn hàng không định hướng sản phẩm ở các phương pháp

Nhận xét: với dữ liệu đơn hàng không định hướng sản phẩm thì phương pháp cây quyết định (6% đúng) trội hơn so với phương pháp gom cụm (0,3%) nhưng phương pháp kết hợp vẫn chiếm ưu thế (33% đúng) với tỷ lệ dự đoán đúng tăng dần (từ 0,3% lên 33%) và tỷ lệ dự đoán sai giảm dần (từ 99,7% còn 67%).

9.4.3 Đánh giá dựa trên thông tin đơn hàng có định hướng sản phẩm

Phương pháp	Đơn hàng train	Đơn hàng test	Dự đoán đúng	Dự đoán sai
Gom cụm	666	334	244 (73,05%)	90 (26,95%)
Cây quyết định	666	334	217 (65%)	117 (35%)
Gom cụm và cây quyết định	666	334	271 (81%)	63 (19%)



Hình 32 –Biểu đồ đánh giá dựa trên thông tin đơn hàng định hướng sản phẩm ở các phương pháp

Nhận xét: với dữ liệu đơn hàng không định hướng sản phẩm thì phương pháp gom cụm (73,05% đúng) trội hơn so với phương pháp cây quyết định (65%), điều này có thể giải thích là dữ liệu có định hướng thì việc tìm điểm giống nhau sẽ dễ dàng hơn nên phương pháp gom cụm sẽ chính xác hơn, còn với phương pháp cây quyết địnhthì tỷ lệ dự đoán đúng/sai có chênh lệch khá lớn vì chỉ có một lượng nhỏ mặt hàng thực sự được mua, cây cây quyết địnhsẽ bị tẻ nhánh dẫn đến kết quả dự đoán không chính xác. Nhưng phương pháp kết hợp vẫn chiếm ưu thế (81% đúng) do đã làm giảm số thuộc tính trong từng phương pháp và kết hợp lại nên tỷ lệ dự đoán đúng cao nhất (81%) và tỷ lệ dự đoán sai nhỏ nhất (19%).

10. KẾT LUẬN & ĐỀ NGHỊ

10.1 Kết luận

Trong khoảng thời gian nghiên cứu đề tài, chúng tôi nhận ra rằng:

- Phương pháp khai thác dữ liệu không chỉ sử dụng một phương pháp để có thể giải quyết bài toán mà chúng tôi đã đặt ra, mỗi phương pháp chỉ giải quyết một vấn đề cụ thể, có ưu điểm và khuyết điểm riêng ở từng phương pháp.
- Mỗi phương pháp cho ra kết quả là một quy luật được khai thác từ tập huấn luyện truyền vào.

  Ví dụ: Đối với phương pháp gom cụm xác định đối tượng gom cụm gồm những thuộc tính khách hàng như thế nào? Nếu đối tượng kiểm tra thỏa điều kiện được đặt ra, xem xét kết quả gợi ý mua loại hàng đó có đúng hay không?
- Việc kiểm tra quy luật sẽ được thống kê xem xét có bao nhiêu trường hợp đúng và trường hợp sai ở bộ kiểm tra. Từ đó, phân tích cải tiến từ mỗi lần huấn luyện dữ liệu.

  Ví dụ: Nếu điều kiện thỏa mà quyết định mua hàng không đúng với gợi ý mà hệ thống giới thiệu thì luật tạo ra đã sai. Vậy việc tiếp theo giải pháp sẽ như thế nào?
- Đối với việc gợi ý sai, việc phân tích xác định các yếu tố có thể thay đổi quy luật của hệ thống như:
    - Thay đổi mô hình khai khác dữ liệu.
    - Thay đổi thuật toán sử dụng.
    - Thay đổi tham số khai thác cấu trúc dữ liệu.
    - Thay đổi thuộc tính dữ liệu truyền vào (ví dụ: dữ liệu bị nhiễu giá trị tuổi)

Như vậy, việc khai thác dữ liệu không thể gợi ý chính xác 100% hoàn toàn ở lần huấn luyện đầu tiên. Do đó, chúng tôi liên tục đặt ra các tình huống giả định trên dữ liệu ngẫu nhiên (do vấn đề thiếu dữ liệu) có ý nghĩa, nhằm huấn luyện tập dữ liệu với các tham số rõ ràng. Từ đó, đưa ra tỉ lệ % số gợi ý chính xác ngày càng cao hơn, nhằm mục đích phục vụ cho chiến lược tiếp thị của doanh nghiệp với khách hàng.

10.2 Đề nghị

10.2.1 Đề nghị hướng nghiên cứu

Nghiên cứu, phân tích các khía cạnh thông tin có thể thực hiện khai thác dữ liệu ở mức độ chi tiết hơn. Cụ thể là khai thác dữ liệu có quan tâm đến mối quan hệ trong dữ liệu đơn hàng.

Phân tích, đánh giá tốc độ thực hiện khai thác dữ liệu huấn luyện của phương pháp gom cụm và cây quyết định. So sánh thời gian thực hiện khai thác dữ liệu của các trường hợp gồm:

- Thực hiện khai thác dữ liệu chưa được làm sạch
- Thực hiện khai thác dữ liệu chưa được làm sạch
- Tăng/giảm số thuộc tính thực hiện khai thác dữ liệu
- Tăng/giảm giá trị/trạng thái thuộc tính khi thực hiện khai thác dữ liệu.

10.2.2 Đề nghị hướng ứng dụng

Vấn đề thực hiện việc khai phá dữ liệu quan trọng nhất đó chính là nguồn dữ liệu. Trong khoảng thời gian ngắn thực hiện đề tài, chúng tôi gặp khó khăn lớn nhất đó là vấn đề thiếu nguồn dữ liệu. Mặc dù chúng tôi đã có giải pháp:

- Xây dựng hệ thống giới thiệu bán hàng trực tuyến trên mạng xã hội face4shop.com để lấy dữ liệu từ thực tế thông tin người dùng nhưng do thời gian thực hiện ngắn nên lượng dữ liệu thu thập không nhiều.
- Để giải quyết vấn đề dữ liệu thiếu. Chúng tôi xây dựng dữ liệu ngẫu nhiên và đưa ra các luật tốt hơn cho dự án.

Nhưng bên cạnh đó, chúng tôi vẫn mong muốn thực thao tác trực tiếp với nguồn dữ liệu thực lớn có ý nghĩa. Vì như vậy, chúng tôi mới phân tích và khám phá hết các tình huống có thể xảy ra trong thực tế.

Việc sử dụng mạng xã hội facebook mục đích ngoài lấy thông tin cá nhân của khách hàng mà quan trọng hơn đối với thông tin bạn bè của khách hàng đó. Như vậy việc xây dựng xác định độ tin cậy giữa các thành viên trong mạng xã hội là khá là quan trọng. Nếu độ tin cậy thông tin khách hàng đưa vào mạng xã hội càng đúng thì khả năng khai phá dữ liệu càng cao.

Do không có chi phí mua server để phục vụ cho đề tài, chúng tôi vẫn chưa sử dụng hết công cụ Immediate mode mà MS SQL server datamining hỗ trợ, vì nó sẽ tạo kết nối trực tiếp và liên tục đến Analysis Services server. Đặc biệt là việc sử dụng ngôn ngữ truy vấn DMX. Mọi thao tác từ việc đưa dữ liệu kiểm tra, dự đoán… sẽ được các hàm trong DMX cung cấp thực hiện một cách tự động và chính xác.

Xây dựng hệ thống giới thiệu bán hàng trên mạng xã hội facebook sẽ hấp dẫn hơn, đối với bài toán của chúng tôi, điển hình là tặng quà sinh nhật là sử dụng API của facebook thông báo cho người được tặng có người tặng quà sinh nhật bằng cách sử dụng tin nhắn riêng tư (facebook private message). Như vậy, sẽ làm tăng sự tương tác, khả năng đối tượng được tặng tham gia hệ thống giới thiệu bán hàng face4shop nhiều hơn.

Xây dựng công cụ thực hiện khai thác dữ liệu tự động: ứng dụng nhằm cải thiện thời gian thực hiện khai thác dữ liệu bằng tay. Cơ sở dữ liệu gồm dữ liệu của hệ thống và dữ liệu để khai thác dữ liệu. Ứng dụng đưa ra các lựa chọn phương pháp khai thác dữ liệu từ SQL Server 2008, các tham số đi kèm trước khi thực hiện khai thác dữ liệu, tự động trả về kết quả.

11. KINH NGHIỆM THU ĐƯỢC

Qua mười bốn tuần thực hiện, chúng tôi đã hoàn thành các yêu cầu đã đề ra, để làm được điều đó không chỉ phụ thuộc vào năng lực chuyên môn của mỗi thành viên mà còn phụ thuộc vào sự đoàn kết và quyết tâm làm việc nghiêm túc của nhóm. Sau đây là những điều chúng tôi làm được trong quá trình thực hiện đề án:

Về lý thuyết công nghệ:

- Hiểu được tầm quan trọng của khai thác dữ liệu trong ứng dụng thực tế. Nhất là đối với lĩnh vực e-marketing trực tuyến.
- Hiểu được các thuật toán áp khai thác dữ liệu.
- Nắm cách sử dụng MS Business Intelligence Development Studio, lựa chọn các thuật toán (cụ thể là phương pháp gom cụm và cây quyết định), cách truyền tham số, thay đổi mô hình và lấy kết quả trả về cho MS SQL server.
- Hiểu được kiến trúc làm việc của framework CI, mô hình kiến trúc MVC và ý nghĩa của nó trong việc xây dựng ứng dụng cụ thể.
- Nắm vững ngôn ngữ mã nguồn mở PHP và kỹ thuật AJAX.
- Nắm vững ngôn ngữ truy vấn dữ liệu SQL.
- Củng cố kiến thức về thiết kế web: HTML, CSS và Javascript.

Về ứng dụng giới thiệu bán hàng qua mạng xã hội:

Hệ thống website mua bán hàng có giao diện đơn giản, dễ sử dụng, có đầy đủ chức năng đáp ứng quy trình hoạt động cơ bản của một trang kinh doanh qua mạng phổ biến. Ngoài ra, hệ thống giúp khách hàng không tốn nhiều thời gian để tạo tài khoản, khách hàng có thể sử dụng chung với tài khoản mạng xã hội facebook do hệ thống sử dụng giao diện lập trình ứng dụng (API) mà mạng xã hội này cung cấp. Bên cạnh đó, hệ thống sử dụng khai thác dữ liệu là chức năng chính của ứng dụng, chức năng này cung cấp cho khách hàng mua hàng tặng quà cho bạn bè của mình dựa vào thông tin của khách hàng mua hàng trước đó. Quá trình xây dựng hệ thống bao gồm các bước sau:

- Tham khảo các site bán hàng trên mạng: shopbop, amazon.. → thiết kế các bước chức năng cơ bản → Xây dựng framework đáp ứng với yêu cầu đề tài đề ra.

- Phân tích bài toán và chọn công nghệ thực hiện trong khoảng thời gian thực hiện đề tài.
- Phân tích và thống kê kết quả gợi ý. Từ đó, tạo ra gợi ý giới thiệu sản phẩm càng đúng về sau.
- Phân tích thiết kế yêu cầu theo từng module (Khách hàng, Đơn hàng, Sản phẩm..) dễ quản lý khi tích hợp.

Về kỹ năng làm việc:

- Rèn luyện kỹ năng tự nghiên cứu, tìm hiểu công nghệ mới.
- Biết cách lập kế hoạch phân chia thời gian làm việc hợp lý.
- Kỹ năng đặt tình huống giả định và giải quyết tình huốn như khi gặp khó khăn về vấn đề thiếu dữ liệu.
- Rèn luyện kỹ năng truyền thông khi làm việc nhóm, nó giúp ích trong cách diễn đạt vấn đề trong quá trình thực hiện đề tài.
- Nâng cao kinh nghiệm làm việc nhóm cũng như làm quen được với áp lực về thời gian thực hiện.

12. PHỤ LỤC

12.1 Tài liệu tham khảo

1. **Wikipedia Community.** Social network. *Wikipedia.* [Online] Wikimedia Foundation, Inc, 12 16, 2010. http://en.wikipedia.org/wiki/Social_network.

2. Facebook. *Statistics Facebook.* [Online] http://www.facebook.com/press/info.php?statistics.

3. Bách khoa toàn thư mở Wikipedia. *Mạng xã hội - Bách khoa toàn thư mở Wikipedia.* [Online] Wikimedia Foundation, Inc. [Cited: Tháng 11 18, 2010.] http://en.wikipedia.org/wiki/Social_network.

4. Bách khoa toàn thư mở. *Khai phá dữ liệu.* [Online] Wikimedia Foundation, Inc, 10 1, 2010. http://vi.wikipedia.org/wiki/Khai_ph%C3%A1_d%E1%BB%AF_li%E1%BB%87u.

5. **MSDN community.** Data Mining Algorithms (Analysis Services - Data Mining). *Microsoft MSDN.* [Online] Microsoft Corporation. http://msdn.microsoft.com/en-us/library/ms175595.aspx.

6. **Nguyễn Kim Long.** Phâm cụm. *Elearning Hoa Sen - Site bài giảng.* [Online] Hoa Sen University, 11 1, 2010. http://www.elearning.hoasen.edu.vn/course/view.php?id=103.

7. **Wikipedia Community.** Thông tin nhân khẩu học Việt Nam. *Bách khoa toàn thư mở Wikipedia.* [Online] Wikimedia Foundation, Inc, 11 24, 2010. http://vi.wikipedia.org/wiki/Th%C3%B4ng_tin_nh%C3%A2n_kh%E1%BA%A9u_h%E1%BB%8Dc_Vi%E1%BB%87t_Nam.

8. **Jamie MacLennan, ZhaoHui Tang, Bogdan Crivat.** *Data Mining with Microsoft SQL Server 2008.* Canada : Wiley Publishing, Inc, 2009.

9. Quảng cáo trên mạng xã hội. *CleverAds.Vn.* [Online] CleverAds Viet Nam. http://cleverads.vn/home/quang-cao-mang-xa-hoi.html.

10. **Nguyễn Kim Long.** Tiền xử lý dữ liệu. *Elearning Hoa Sen - Site bài giảng* . [Online] Hoa Sen University, 10 9, 2010. [Cited: 10 9, 2010.] http://www.elearning.hoasen.edu.vn/course/view.php?id=103.

12.2 Từ điển thuật ngữ

Data mining: Khai phá dữ liệu là quá trình khám phá các tri thức mới và các tri thức có ích ở dạng tiềm năng trong nguồn dữ liệu đã có. Khai phá dữ liệu là một bước của quá trình khai phá tri thức (Knowledge Discovery Process).(5)

API: Application Programming Interface, hệ giao tiếp lập trình ứng dụng là một giao diện mà một hệ thống máy tính hay ứng dụng cung cấp để cho phép các yêu cầu dịch vụ có thể được tạo ra từ các chương trình máy tính khác, và/hoặc cho phép dữ liệu có thể được trao đổi qua lại giữa chúng.

SQL: Structure Query Language, ngôn ngữ truy cấn có cấu trúc.

PHP : Hypertext Preprocessor, ngôn ngữ lập trình kịch bản hay một loại mã lệnh chủ yếu được dùng để phát triển các ứng dụng viết cho máy chủ, mã nguồn mở, dùng cho mục đích tổng quát, thích hợp với web và có thể dễ dàng nhúng vào trang HTML.

HTML: Hypertext Markup Language, ngôn ngữ mã hóa định dạng, liên kết và những đặc tính khác trên web. Sử dụng những tag chuẩn như <h1>, <body>…

CSS: Cascading Style Sheets, được dùng để miêu tả cách trình bày các tài liệu viết bằng ngôn ngữ HTML và XHTML.

AJAX: Asynchronous JavaScript and XML, sự kết hợp của các công nghệ: CSS, DOM, XMLHttpRequest, JavaScript.

MVC: Model-View-Controller, là một mẫu kiến trúc phần mềm trong kỹ thuật kỹ sư phần mềm. Mẫu MVC giúp cô lập các nguyên tắc nghiệp vụ và giao diện người dùng một cách rõ ràng và hệ thống.

CI: CodeIgniter, là một framework PHP được phát triển bởi EllisLab . CodeIgniter cho phép xây dựng các ứng dụng web, có nhiều libraries và helpers hữu ích làm giảm bớt dòng code .

12.3 Danh mục bảng và hình ảnh

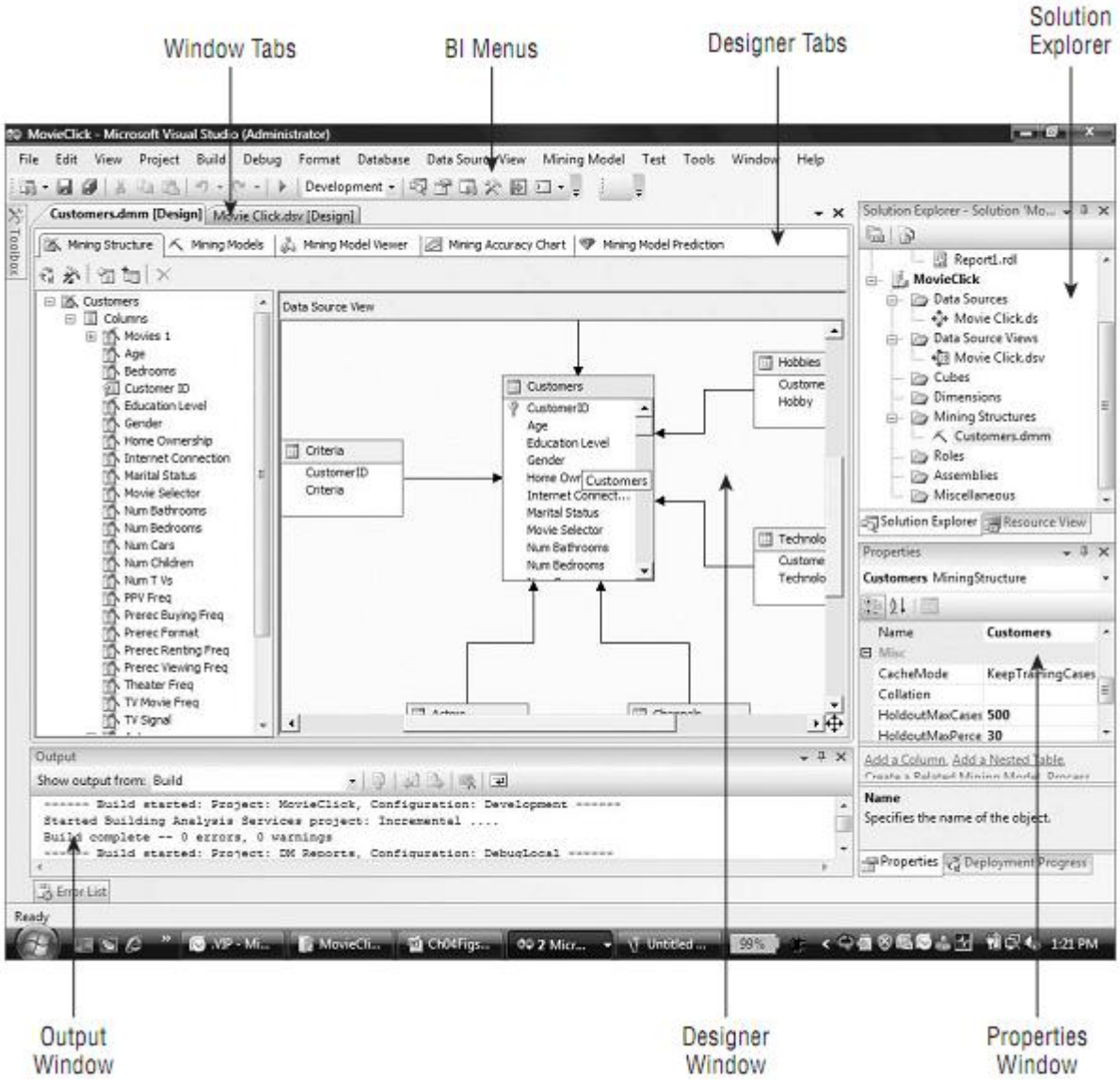
12.3.1.1 Thể loại bảng

12.3.1.2 Thể loại hình ảnh

12.4 Hướng dẫn sử dụng Datamining SQL Server version 2008

12.4.1.1 Business Intelligence Development Studio (BI Dev Studio)

Hầu hết thời gian sử dụng SQL Server data mining sẽ được thực hiện trong BI Dev Studio. Môi trường BI Dev Studio được tích hợp vào trong Microsoft Visual Studio.



Hình 33 – Môi trường BI Dev Studio

BI Dev Studio làm việc trong hai chế độ (các chế độ sử dụng dựa trên sở thích cá nhân và do sự cần thiết). Mỗi chế độ có lợi thế và nhược điểm và điều quan trọng là hiểu được sự khác biệt giữa chúng khi làm việc với BI Dev Studio.

- Immediate mode:
    - Kết nối trực tiếp và liên tục đến Analysis Services server.
    - Khi mở một đối tượng, đối tượng được mở từ máy chủ.
    - Khi thay đổi đối tượng và lưu nó, các đối tượng lập tức thay đổi trên máy chủ.
    - Tuy nhiên người khác có thể sửa đổi các đối tượng trong khi đang mở nó.
- Offline mode:
    - Các file được lưu trữ trên máy Client.
    - Khi thay đổi cho các đối tượng trong môi trường này, các thay đổi được lưu trữ ở định dạng XML trên ổ cứng.
    - Các mô hình và các đối tượng không được tạo ra trên máy chủ cho đến khi người dùng quyết định triển khai chúng đến các máy chủ.
    - Người dùng phải là một quản trị viên máy chủ để triển khai một dự án từ chế độ Offline cho máy chủ.

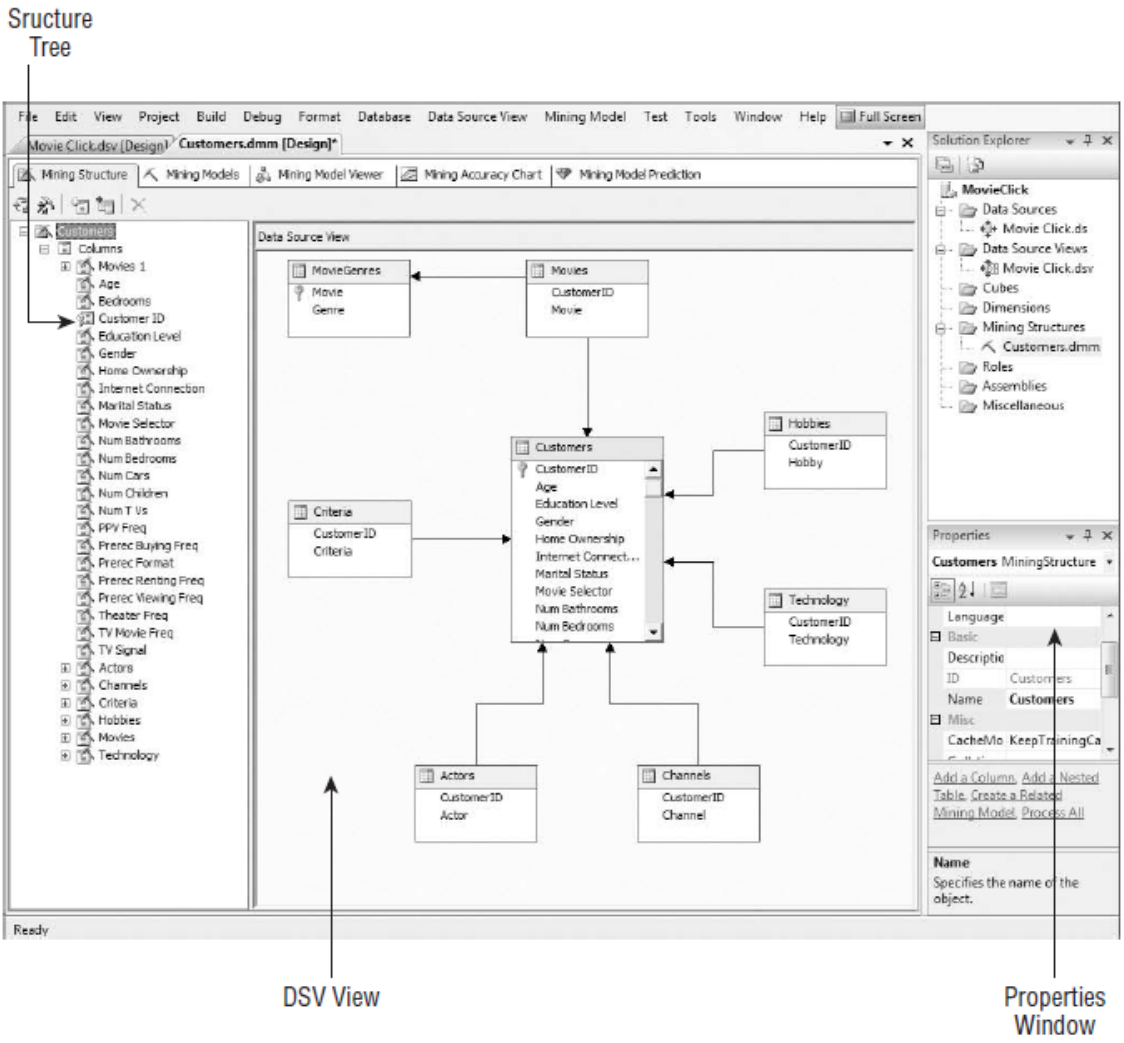
12.4.1.2 Data Sources

- Để thực hiện khai thác dữ liệu cần phải chỉ rõ, mô tả các nguồn dữ liệu. Sau đó tạo ra các cấu trúc khai thác và các mô hình.
- Data sources là một đối tượng khá đơn giản. Nó là một chuỗi kết nối, cộng với một số thông tin bổ sung cho thấy làm thế nào để kết nối.
- Một đối tượng data sources có thể được tạo ra với một trong bốn lựa chọn sau:
    - Impersonate Current User: Phương pháp an toàn nhất cho các nguồn dữ liệu, truy cập thông qua báo cáo truy vấn. Sử dụng thông tin người dùng hiện tại để truy cập dữ liệu từ xa.
    - Impersonate Account: Lựa chọn tốt thứ hai, cho phép chỉ định các thông tin tài khoản đó sẽ được sử dụng để truy cập vào data sources.
    - Impersonate Service Account: Tùy chọn này gây ra tất cả truy cập dữ liệu đến tài khoản Analysis Services đang chạy. Phương pháp chủ yếu cho mục đích thử nghiệm, và không được khuyến khích sử dụng sản xuất.
    - Default: Tùy chọn này gây ra các thông tin khác nhau được sử dụng, tùy thuộc vào data source được truy cập.

12.4.1.3 Mining Structure

Data Mining Designer gồm có 5 panes cho editing, browsing, querying, and comparing models:

- The Mining Structure pane
    - The Mining Models pane
    - The Mining Model Viewer pane
    - The Mining Accuracy pane
    - The Mining Model Prediction pane
- Mining Structure Editor cho phép thêm các cột và loại bỏ các cột từ mining structure, thiết lập các thuộc tính của từng cột.
- Ba thành phần của mining structure editor là structure tree, the DSV View Datasource View), và Properties window.



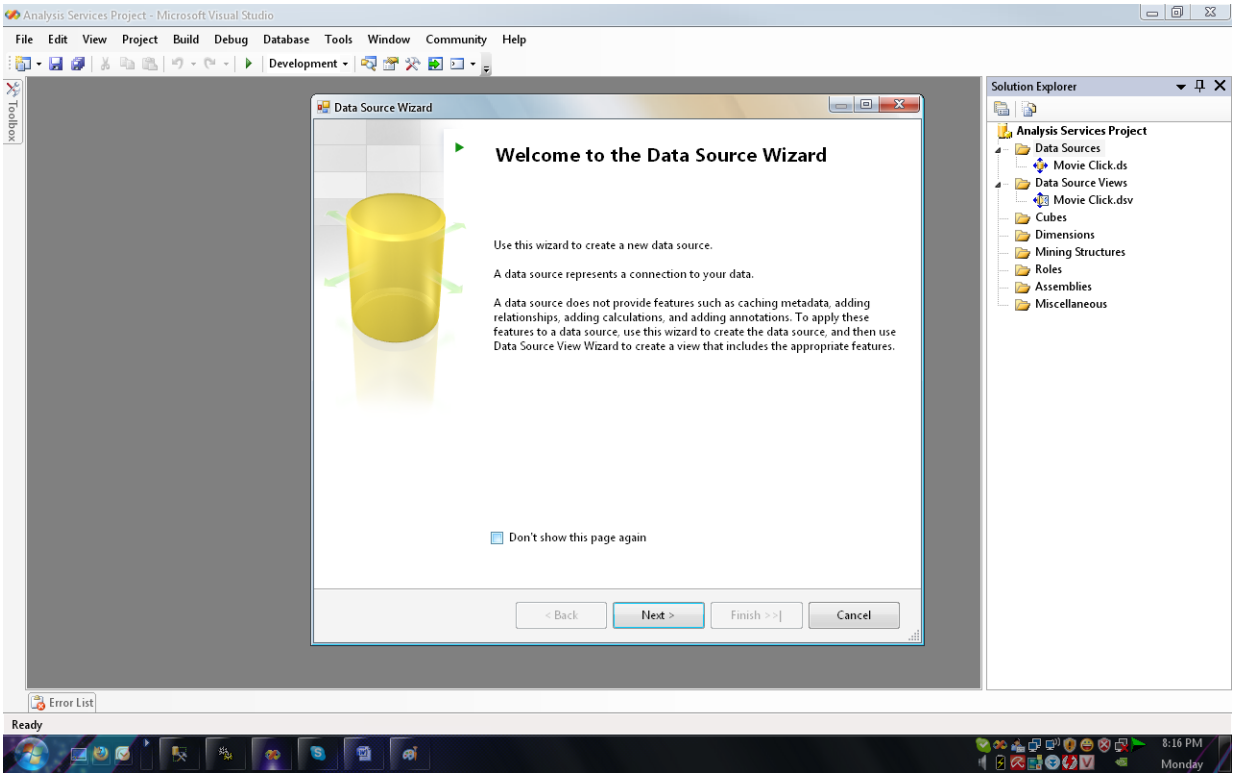
Hình 34 – Data Mining Designer

Ví dụ: các bước sử dụng datamining trong MS SQL Server 2008

- Bước 1: Import database vào MS SQL Server 2008 (sử dụng database MovieClick làm database mẫu)
- Bước 2: Chọn Start → Chọn mục MS SQL Server 2008.
- Bước 3: Menu bar chọn File → New project → Chọn SQL Server Business Intellligence Development Studio → Chọn Analysis Services Project → Chọn nơi lưu trữ file Solution
- Bước 4: Vào Solution Explorer (hoặc chọn menu View → Solution Explorer)

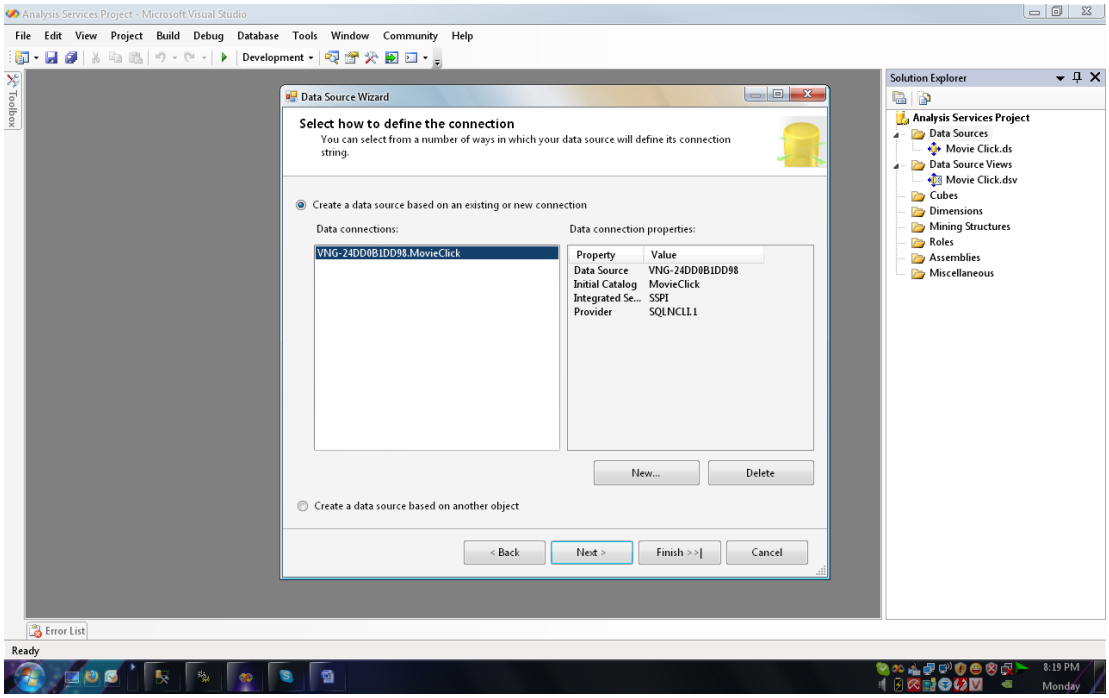
Tạo data source

- Bước 5: Chuột phải Data Source → Màn hình Data Source Wizard → Chọn Next



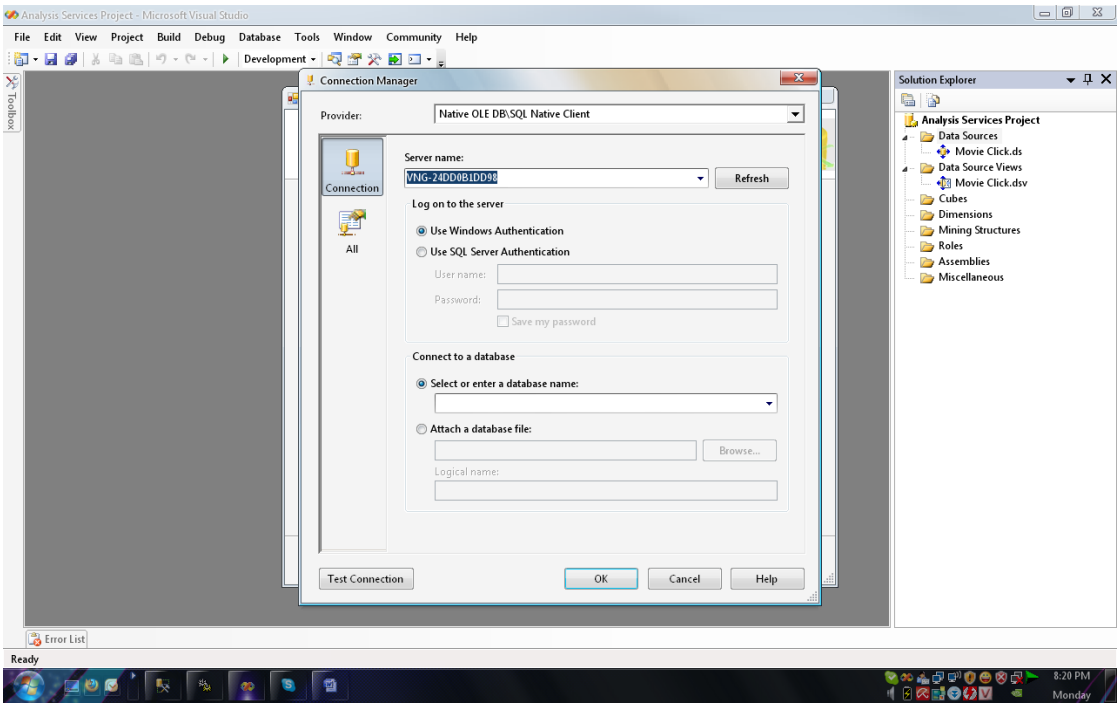
Hình 35 – Tạo data source

- Bước 6: Chọn New để có dữ liệu gốc data source.



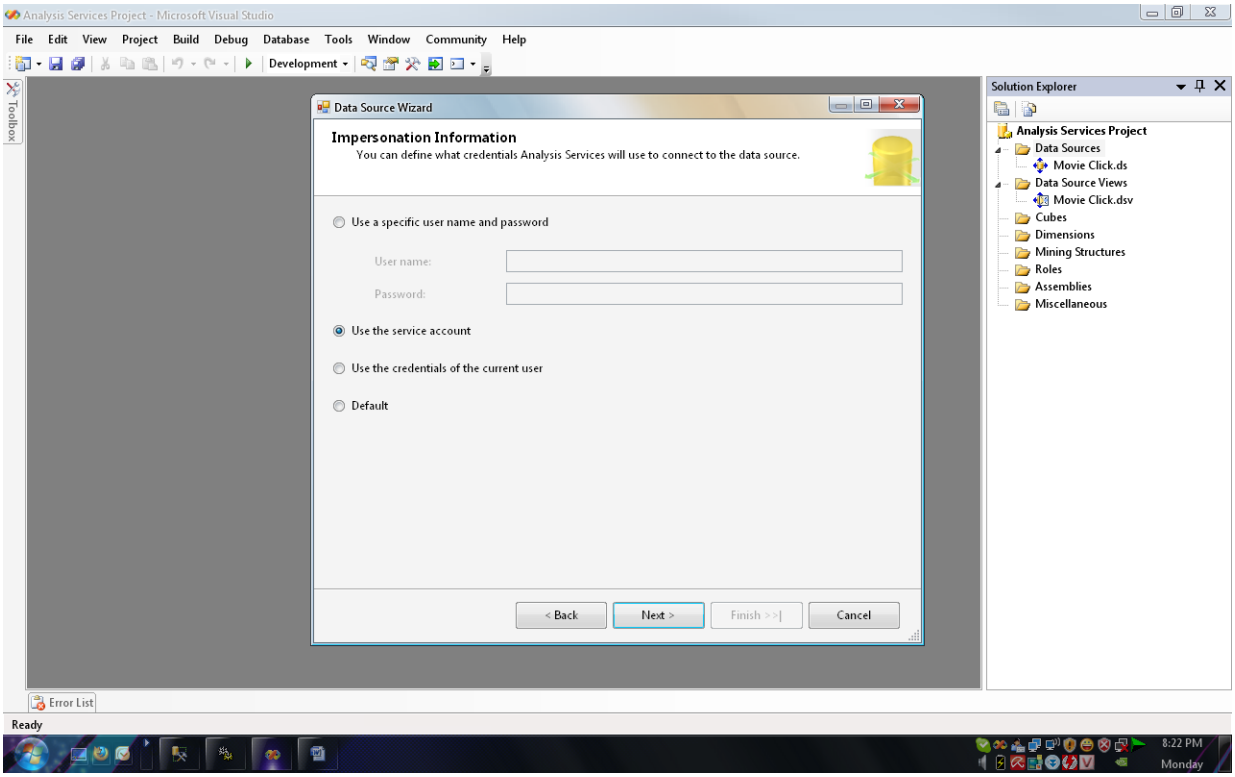
Hình 36 – Tạo mới data source

- Bước 7: Chọn server name



Hình 37 – Chọn server name

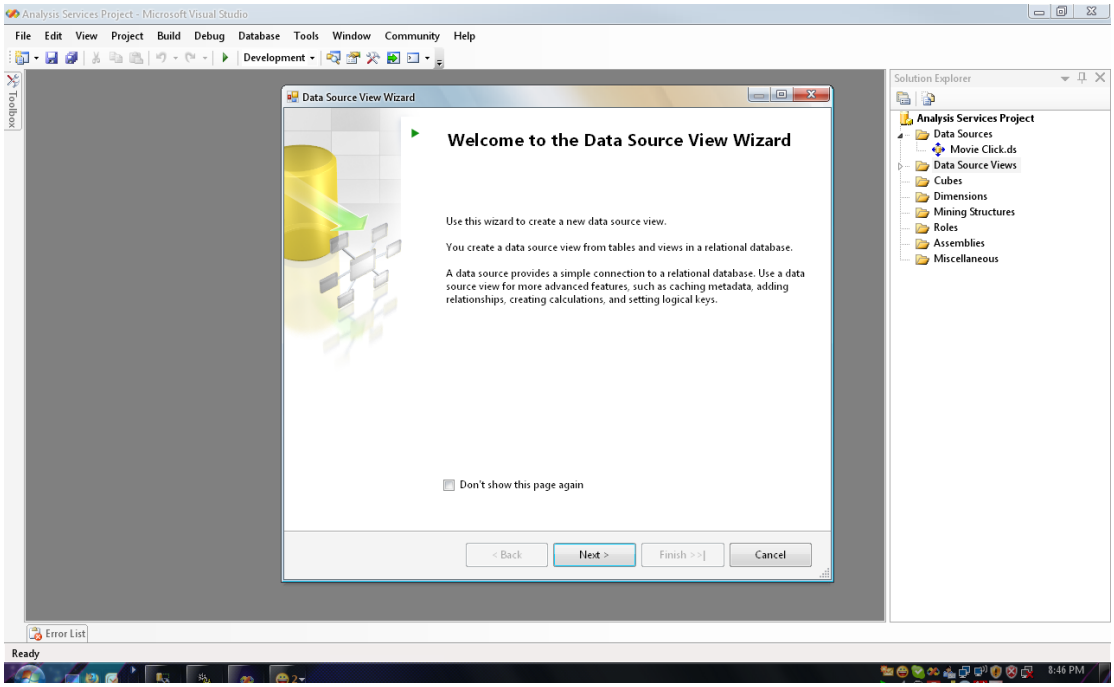
- Bước 8: Chọn data name → Chọn OK → Chọn Next.
- Bước 9: Chọn Use the Service Account hoặc Use a specific username and password (thông tin tài khoản SQLServer) → Chọn Next



Hình 38 – Chọn thông tin tài khoản

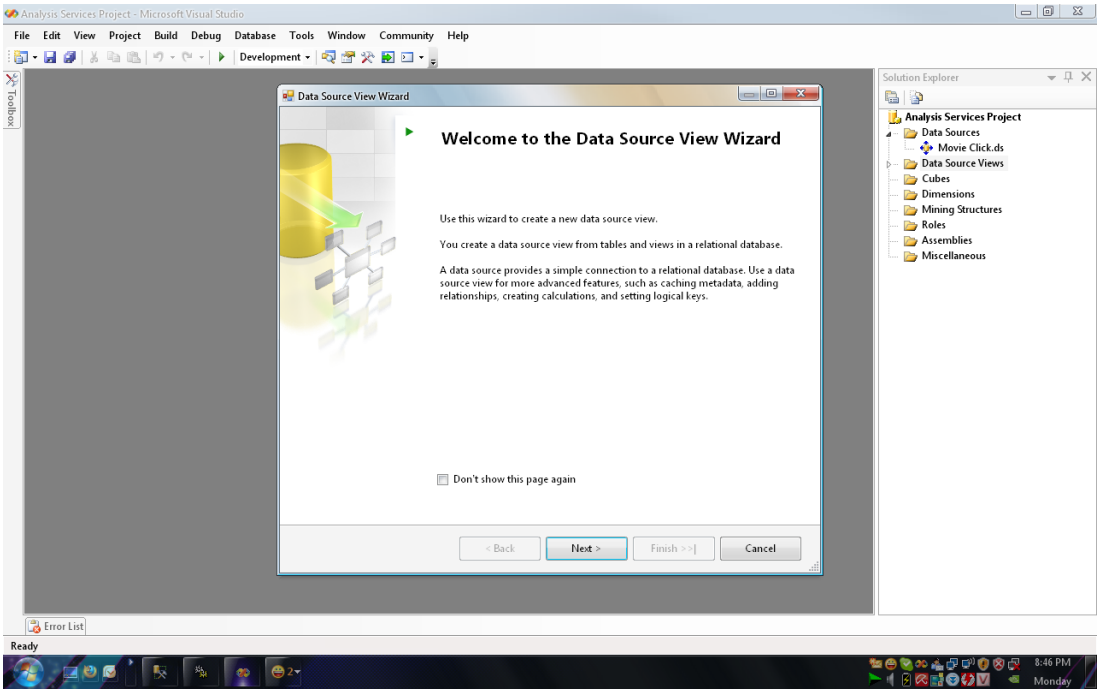
Tạo datasource View

- Bước 11: Chọn data Source Views → New datasource view → Màn hình Data Source View Winzard → Chọn Next



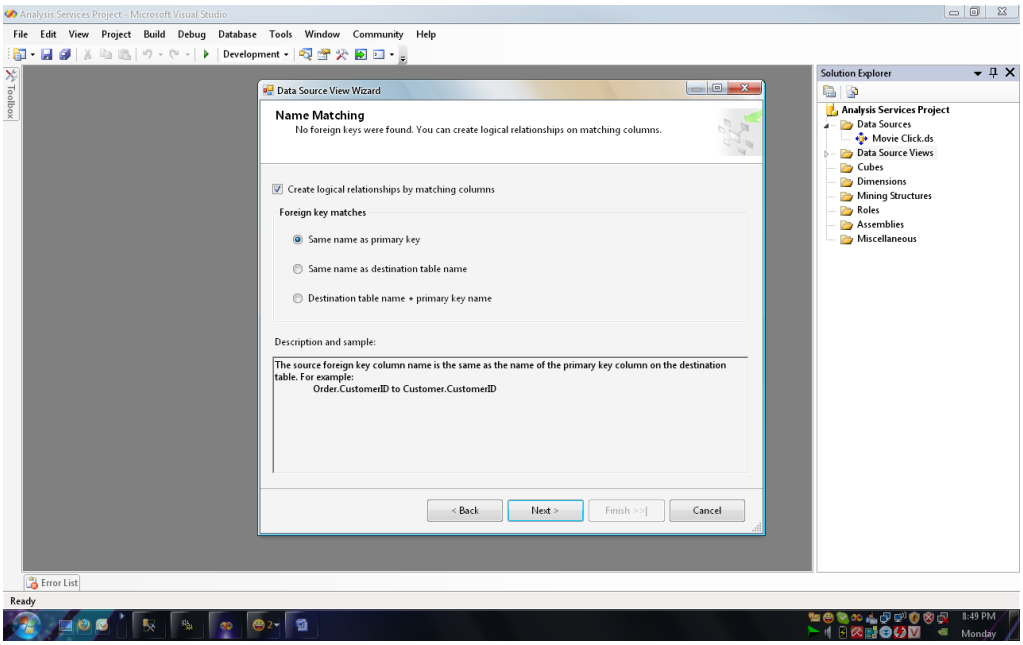
Hình 39 – Tạo data source view

- Bước 12: Click Next (Hoặc tạo mới một datasource khác)



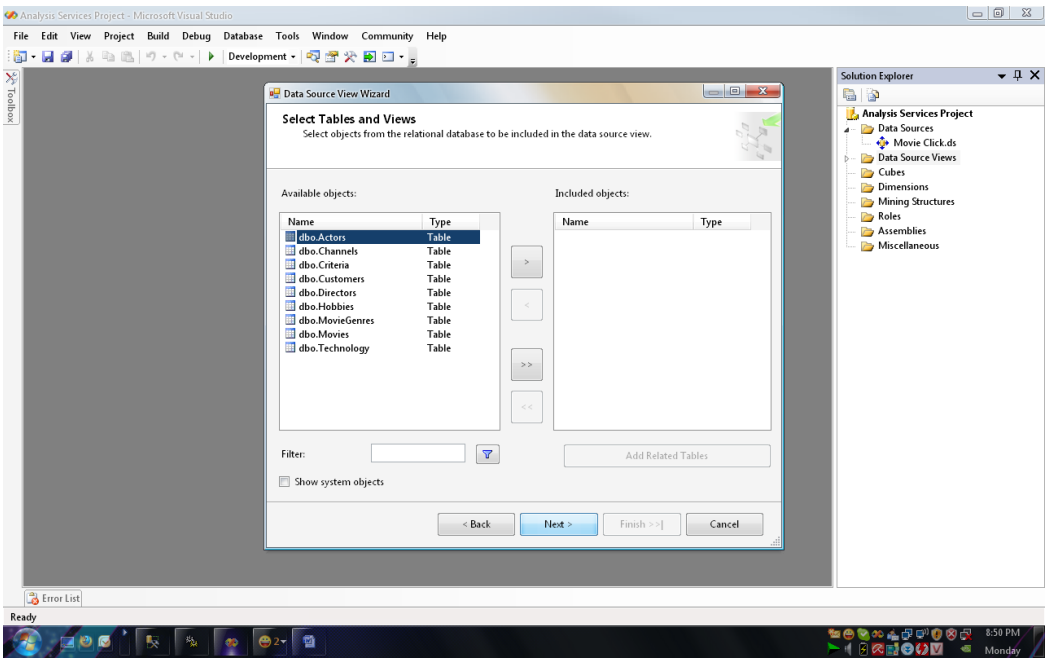
Hình 40 – Tạo mới data source view

- Bước 13: Màn hình hiện thị câu hỏi đã tìm thấy những mối quan hệ trong các bảng, bạn có muốn thay đổi gì không? → Chọn Next



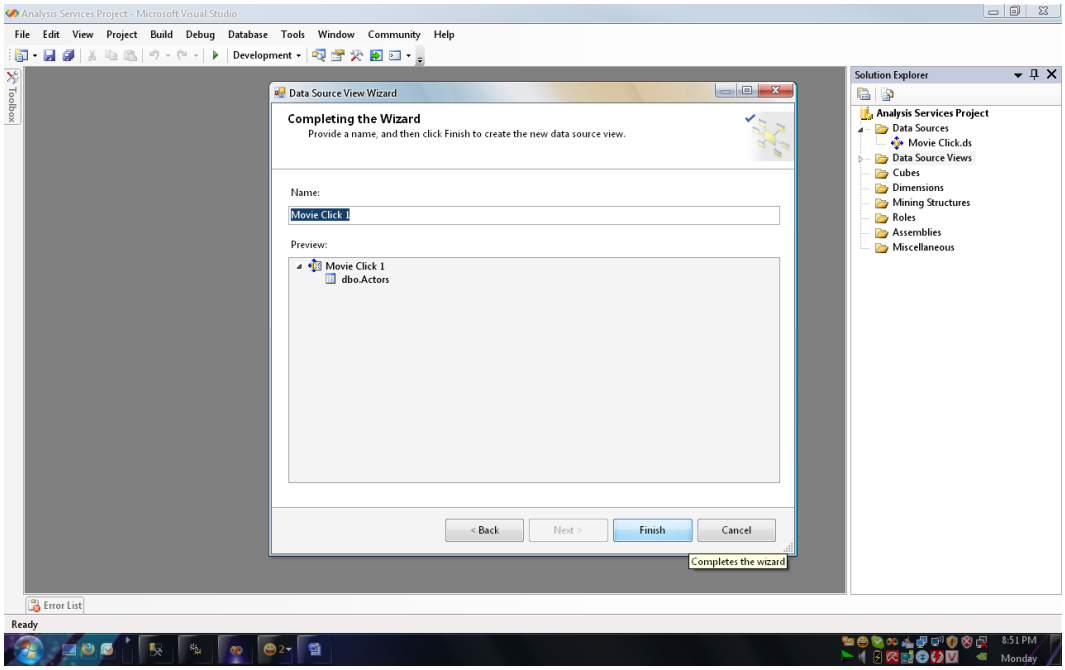
Hình 41 – Xác nhận mối quan hệ giữa các bảng

- Bước 14: Chọn các table muốn sử dụng → Chọn Next



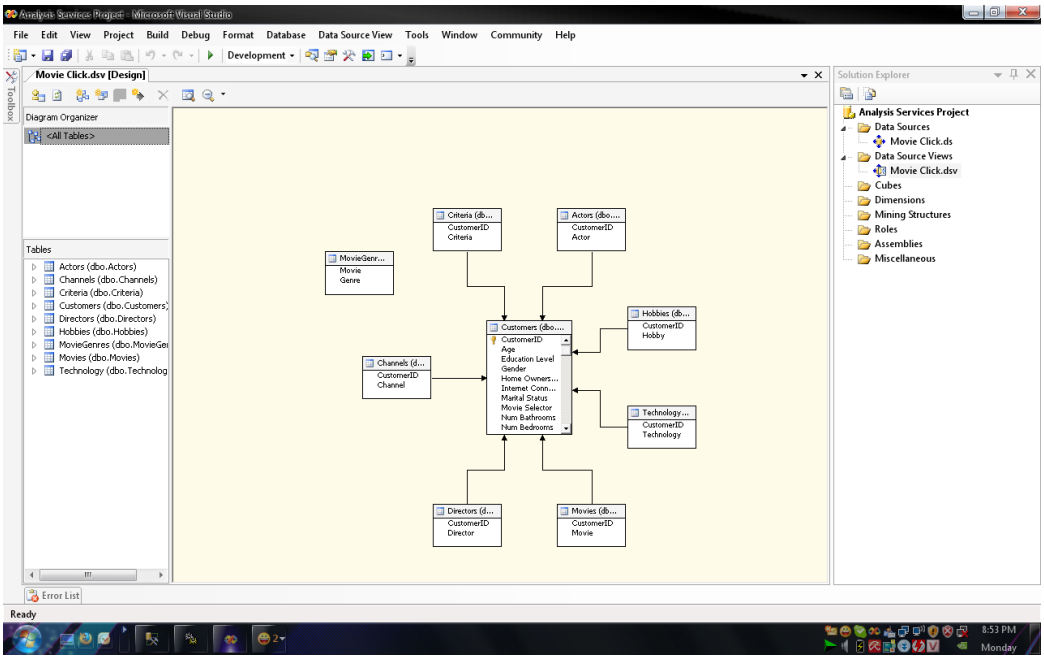
Hình 42 – Chọn bảng cần khai thác

- Bước 15: Đặt tên → Chọn Finish



Hình 43 – Đặt tên data source view

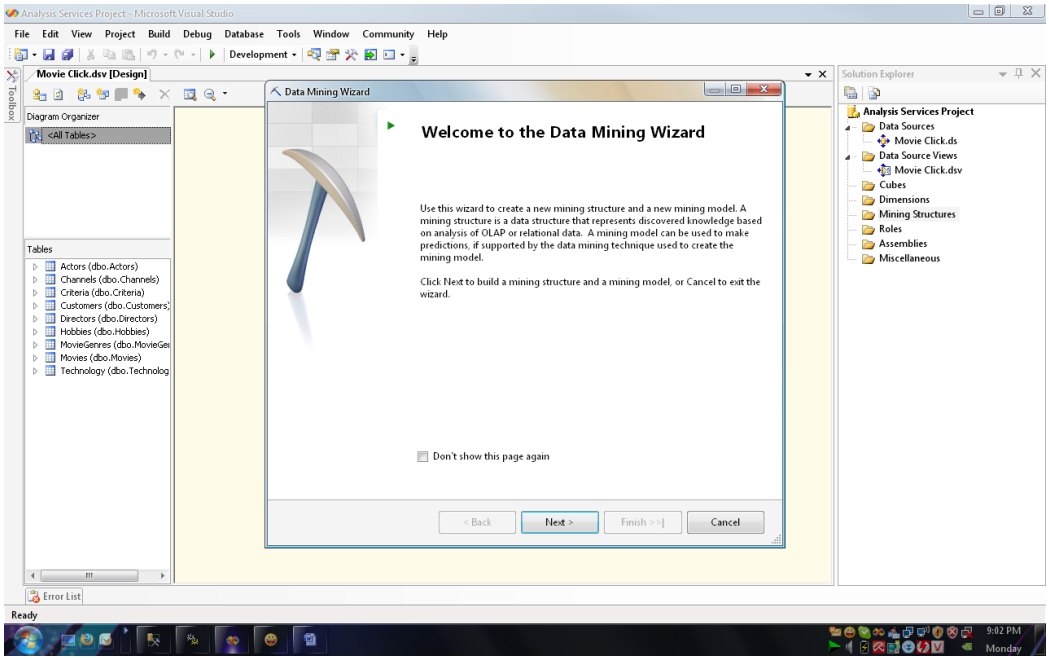
- Bước 16: Màn hình tự động hiển thị các bảng và các mối quan hệ



Hình 44 – Màn hình hiển thị các bảng và mối quan hệ

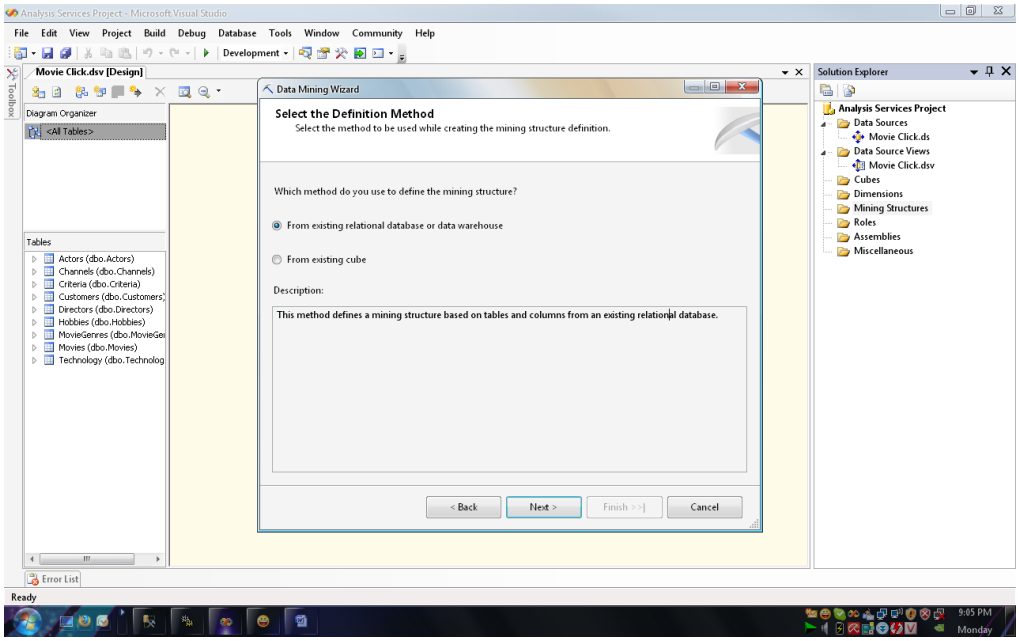
Tạo Mining Structure – Sử dụng thuật toán Decision Tree

- Bước 17: Khung Solition Explorer → Click phải Mining Structure → Màn hình Data Mining Wizard



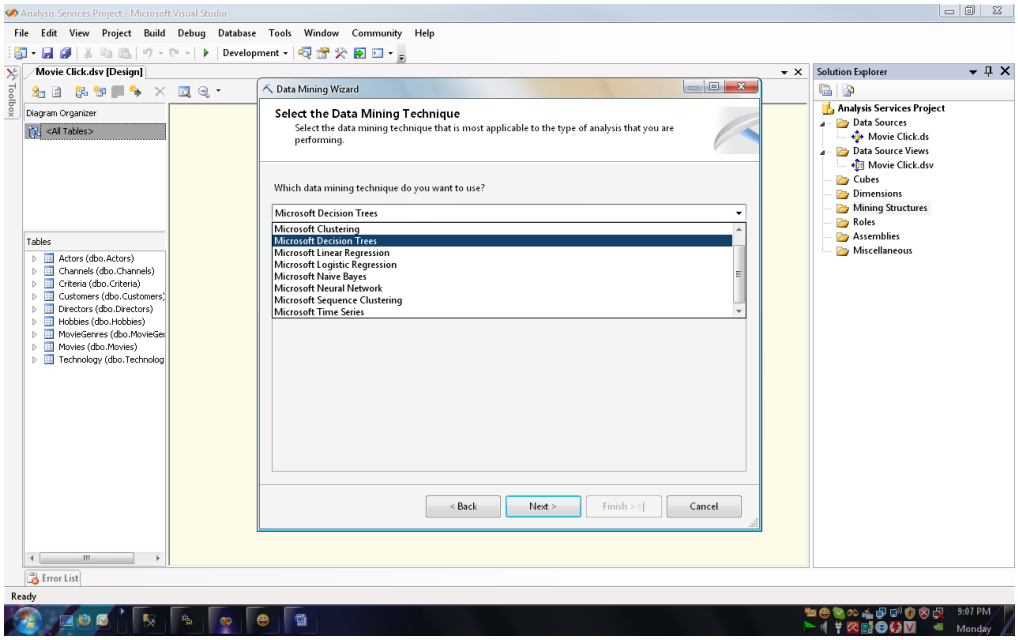
Hình 45 – Màn hình hiển thị tạo khai khác dữ liệu

- Bước 18: Màn hình hiển thị phương thức khai báo khai thác trên các tables đã có sẵn trong database hay là những phương thức khai thác trong cấu trúc đã có tồn tại. Trong trường hợp này, chọn "From exsting relation Database or data warehouse"



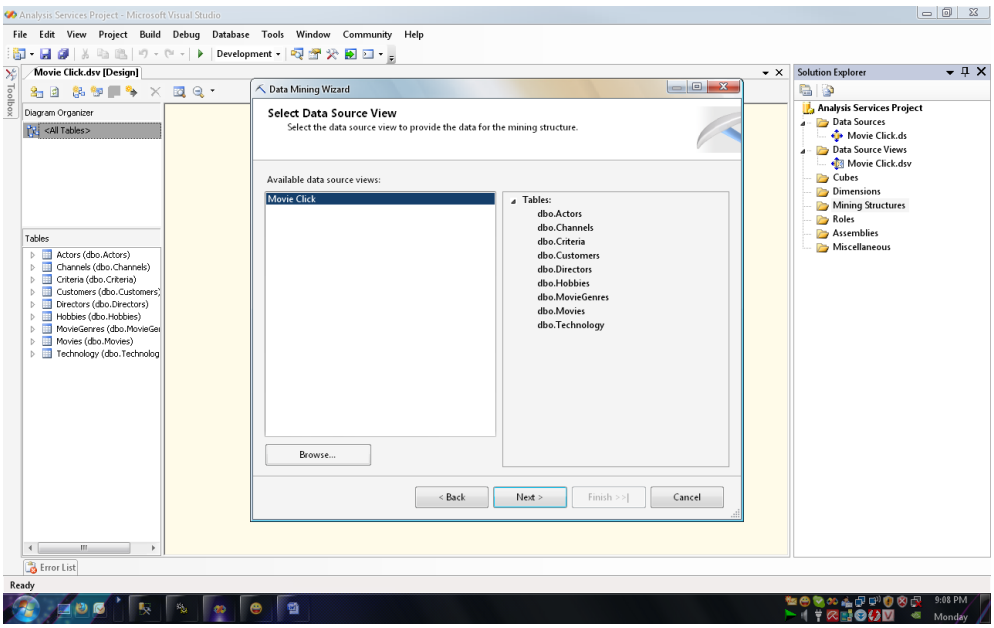
Hình 46 – Màn hình hiển thị các bảng và mối quan hệ

- Bước 19: Màn hình xác định mô hình mà mình muốn dùng, MS Visual Studio cung cấp nhiều mô hình data mining, trong trường hợp này ta chọn Microsoft Decision Tree.



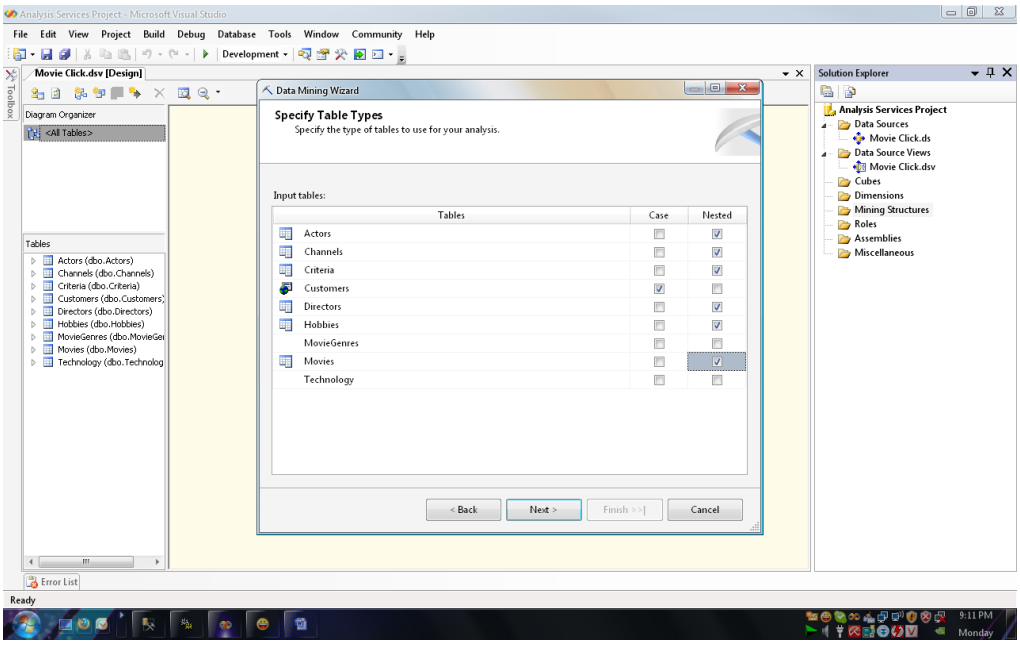
Hình 47 – Màn hình chọn phương pháp khai thác

- Bước 20: Màn hình xác định database muốn khai thác dữ liệu → Chọn Next



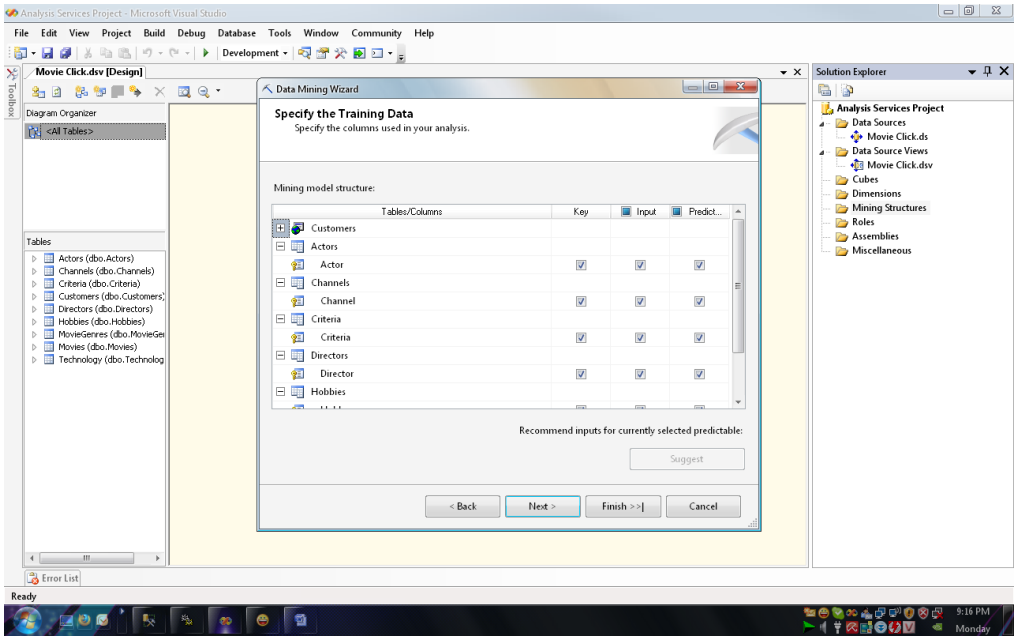
Hình 48 – Màn hình chọn data source view

- Bước 21: Màn hình chỉ định các cột được sử dụng trong cấu trúc, mô hình và xác định khóa



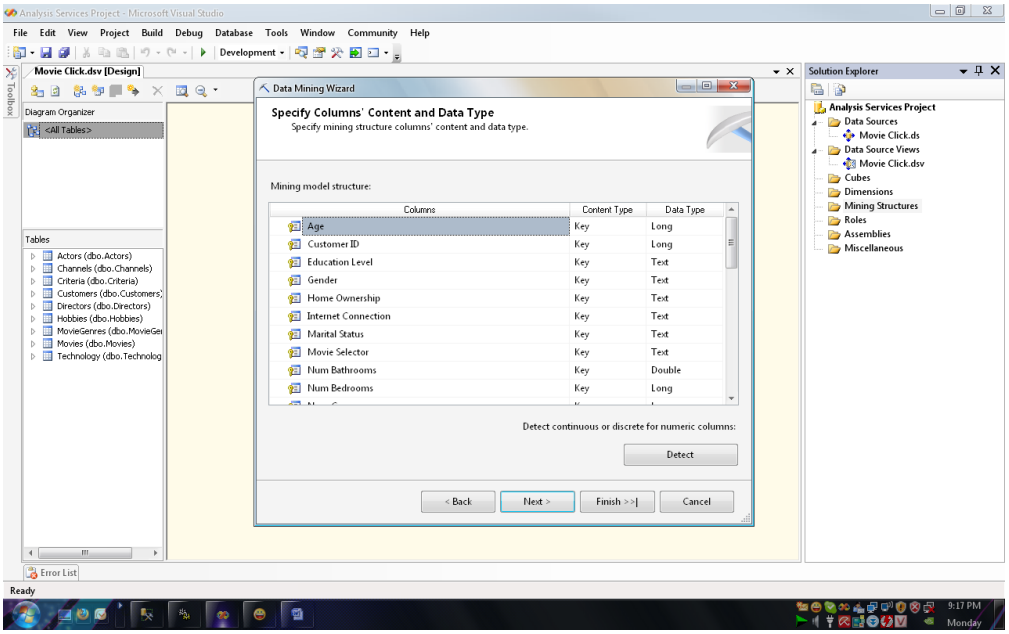
Hình 49 – Màn hình xác định cấu trúc các thuộc tính

- Bước 22: Màn hình xác định đầu vào và dự báo



Hình 50 – Màn hình xác định đầu vào và dự báo

- Bước 23: Màn hình trình bày một danh sách các cột đã lựa chọn và kiểu dữ liệu tương ứng



Hình 51 – Màn hình xác định kiểu dữ liệu

Chú ý: Các cột được xác định là input hay dự báo phụ thuộc vào giả thuyết và thuật toán đã chọn.

- Bước 24: (dành cho SQL Server 2008): Màn hình "Create Testing Set" xác định phần trăm dữ liệu up to maximum và các dự liệu đó được random dành cho việc sử dụng bởi các phương thức kiểm chứng
- Bước 25: Đặt tên cấu trúc datamining và tên mô hình → Chọn Finish.

12.4.1.4 Named Calculation - Named Query

Named Calculation

- Named Calculation là cột ảo được thêm vào một bảng bất kỳ trong Datasource View. Điều này cho phép ta khai thác thông tin từ nguồn dữ liệu ban đầu mà không làm thay đổi dữ liệu nguồn. Một Named Calculation bao gồm tên, biểu thức SQL có tính tóan và mô tả(nếu có).
- Biểu thức tính tóan có thể là bất kỳ biểu thức hợp lệ nào của SQL. Ngòai ra, có một số loại biểu thức được hỗ trợ trong việc khai thác dữ liệu gồm:
    - Biểu thức số học: +, -, *, /, %
    - Biểu thức tóan học: ABS, LOG, SIGN, SQRT,..
    - Biểu thức kết hợp: sự kết hợp giữa thông tin tình trạng hôn nhân và có con sẽ có giá trị hơn về mặt thông tin
    - Biểu thức CASE: cho phép gán giá trị kết quả dựa vào nhiều điều kiện khác nhau.

      Thay đổi giá trị

```
CASE [Category]
WHEN 1    THEN 'Food'
WHEN 2    THEN 'Beverage'
WHEN 3    THEN 'Goods'
END CASE
```

      Chia khỏang giá trị

```
CASE
```

```
WHEN [Age] < 20 THEN ‘Under 20’
WHEN [Age] <= 30 THEN ‘Between 20 and 30’
WHEN [Age] <= 40 THEN ‘Between 30 and 40’
ELSE ‘Over 40’
END
```

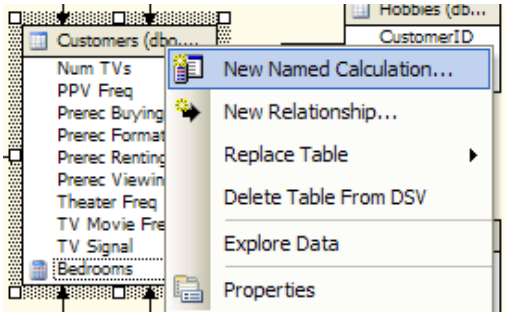
Giảm số trạng thái hợp lệ

```
CASE [Marital Status]
WHEN ‘Married’       THEN [Marital Status]
WHEN ‘Never Married’ THEN [Marital Status]
ELSE ‘Other’
END
```

Chuyển đổi một thuộc tính từ một bảng lồng nhau thành trường hợp chung.

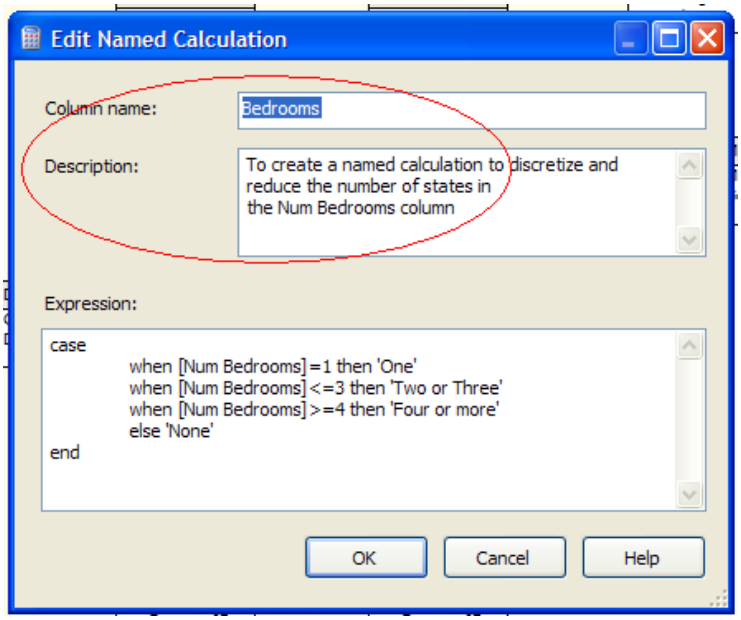
```
CASE
WHEN EXISTS
(SELECT [Movie] FROM [Movies]
WHERE [Movie]=’Star Wars’ AND
[Movies].[CustomerID]=[Customers].[CustomerID])
THEN ‘True’
ELSE ‘False’
END
```

- Tạo một Named Calculation tại một bảng bất kỳ
    - Click phải tại một bảng bất kỳ chọn New Named Calculation



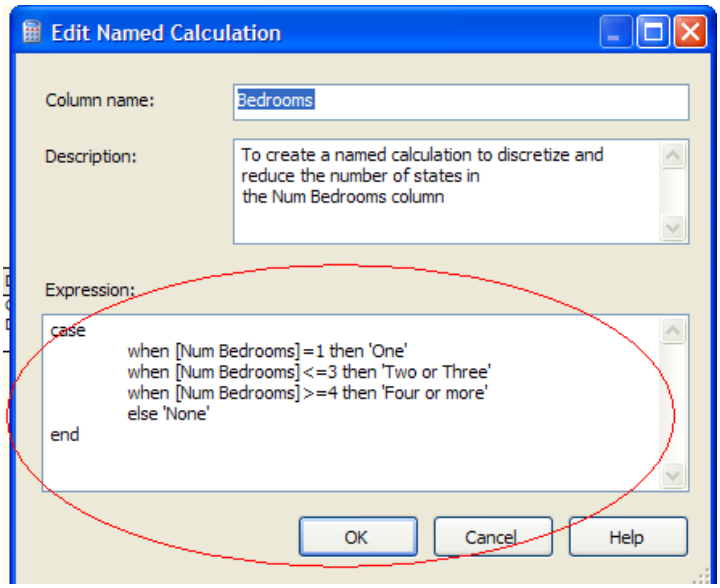
Hình 52 – Màn hình tạo mới Named Calculation

- Nhập tên, nhập mô tả



Hình 53 – Màn hình nhập tên và nhập mô tả

- Nhập biểu thức



Hình 54 – Màn hình nhập biểu thức

- Kết quả xem tại Explore

Prerec Viewin	CustomerID	Theater Freq	TV Movie Fre	TV Signal	Bedrooms
Monthly	877687	Monthly	Monthly	Cable	Two or Three
Weekly	877723	Rarely	Weekly	Cable	Two or Three
Weekly	877757	Rarely	Weekly	Cable	Four or more
Monthly	877792	Rarely	Daily	Cable	Four or more
Monthly	877840	Monthly	Weekly	Cable	Four or more
Weekly	877988	Weekly	Weekly	Digital Satellit	Four or more

Hình 55 – Màn hình xem kết quả khám phá

Named Query

- Giống Named Calculation, Named Query là một view ảo và view này cũng không làm thay đổi dữ liệu nguồn. Việc tạo Named Query trực tiếp trên Datasource View thì nhanh, dễ dàng và cho phép bạn duy trì view cùng với mô hình được sử dụng thay vì làm hỏng dữ liệu của bạn. Các query tiêu biểu cho việc khai thác dữ liệu như lọc, kết hợp, và tạo mẫu
  - Lọc dòng dựa trên điều kiện cột

```
SELECT * FROM [Movies] WHERE [Movie] != 'Star Wars'
```

- Lọc các dữ liệu không phổ biến từ một bảng lồng nhau

```
SELECT [CustomerID], [Movie] FROM [Movies]
WHERE [Movie] IN
(SELECT DISTINCT
[Movie]
FROM [Movies] GROUP BY [Movie]
HAVING COUNT([Movie]) > 20)
```

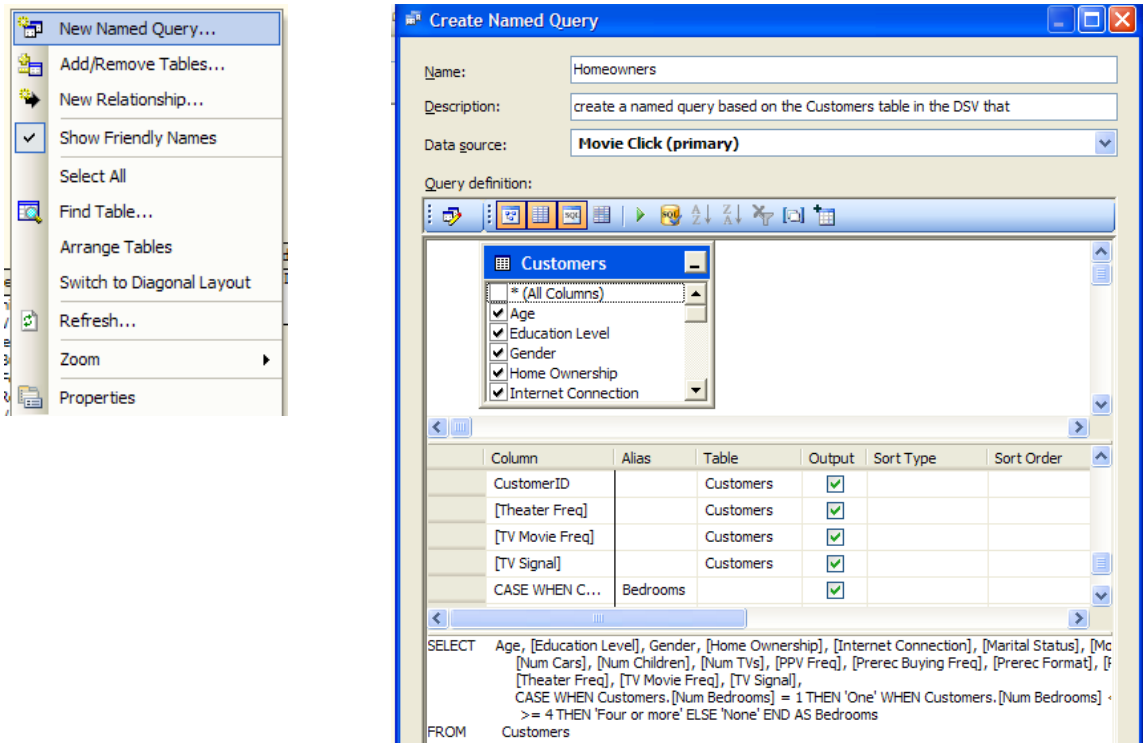
- Kết hợp thông tin từ một bảng khác

```
SELECT
Customers.*, Education.[Education Level]
FROM Customers JOIN Education
ON Customers.[Education Id] = Education.[Education Id]
```

- Tạo các dòng dữ liệu mẫu từ dữ liệu nguồn

```
SELECT * FROM CUSTOMERS
TABLESAMPLE (30 PERCENT)
REPEATABLE (1)
```

- Tạo một Named Query tại một bảng bất kỳ
  - Click phải trên Datasource View Designer, chọn New Named Query

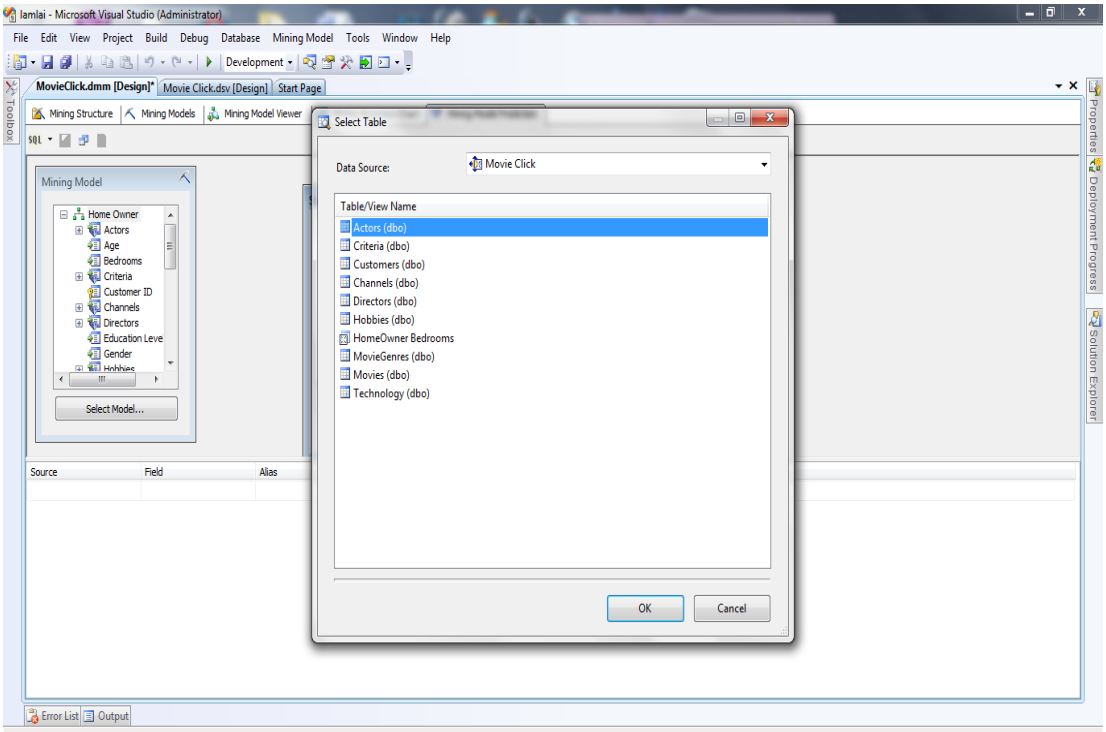


Hình 56 – Màn hình tạo mới Named query

- Nhập tên view, nhập mô tả
- Click Add table để chọn bảng dữ liệu
- Check chọn các cột thông tin cần thiết
- Thêm cột tính tóan bằng cách nhập biểu thức tại màn hình query
- Kết quả trả về là một view, xem kết quả cũng tương tự như một bảng dữ liệu nhưng có thêm cột tính tóan.
- Khi tạo thêm một Named Query, thì trên màn hình quan hệ không thể hiện mối quan hệ với bảng mới vừa tạo, nếu muốn tạo mối quan hệ với các bảng dữ liệu nguồn phải dùng chuột kéo thả khóa ngọai vào các bảng dữ liệu nguồn cần quan hệ hoặc là tạo mới một sơ đồ bao gồm quan hệ với bảng mới vừa tạo.

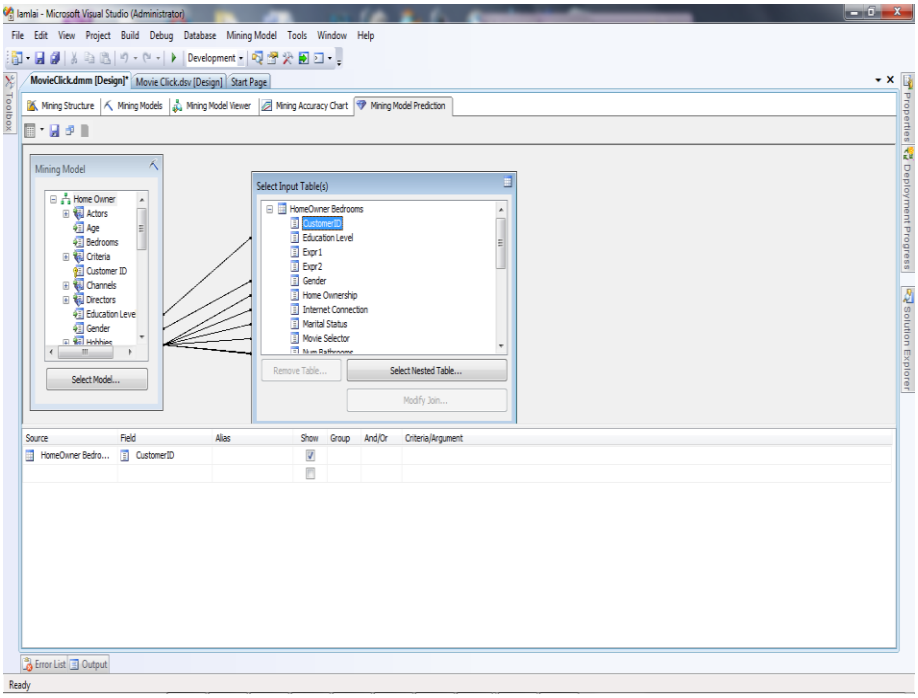
12.4.1.5 Mining prediction

- Chuyển sang tab mining model prediction
- Click select case table on select input table window
- Chọn homeOwners table in dialog box, click ok



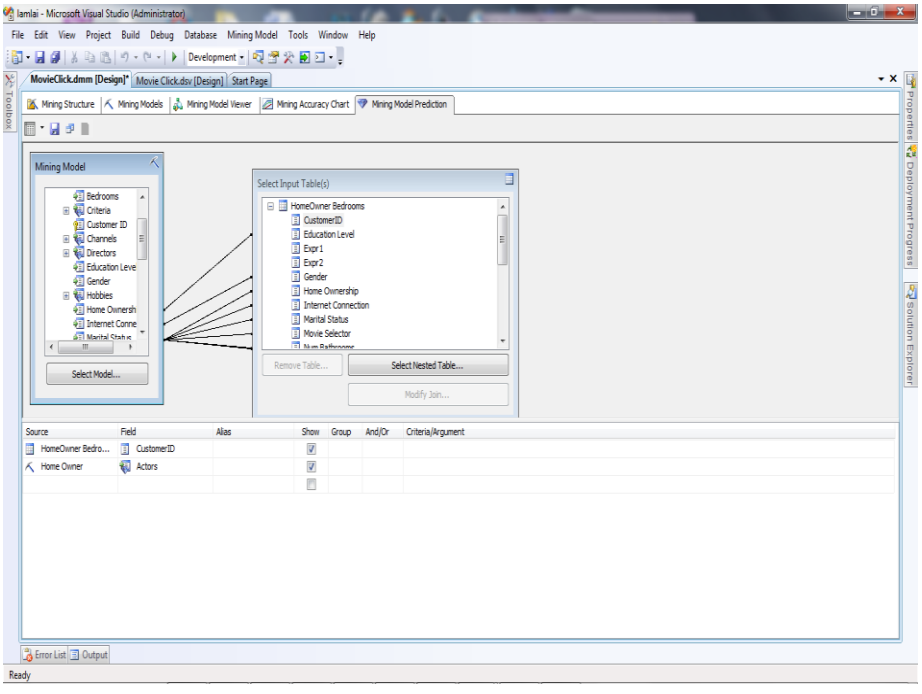
Hình 57 – Màn hình chọn bảng để dự đoán

- Kéo customerID column từ HomeOwners table, thả nó xuống grid bên dưới



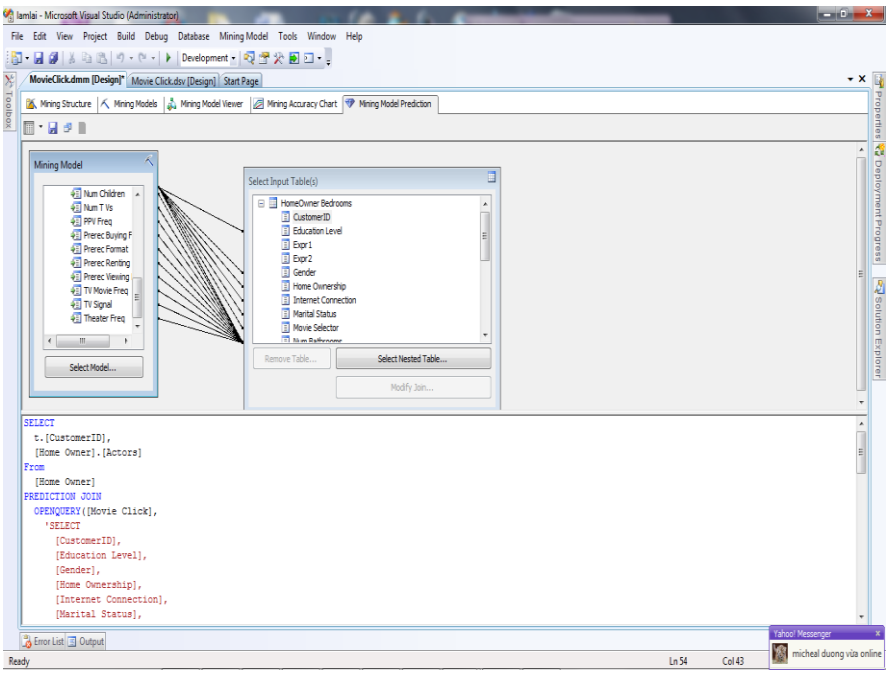
Hình 58 – Màn hình quan hệ giữa kết quả huấn luyện với giá trị dự đoán

- Kéo Actors column từ mining model và thả nó xuống grid bên dưới



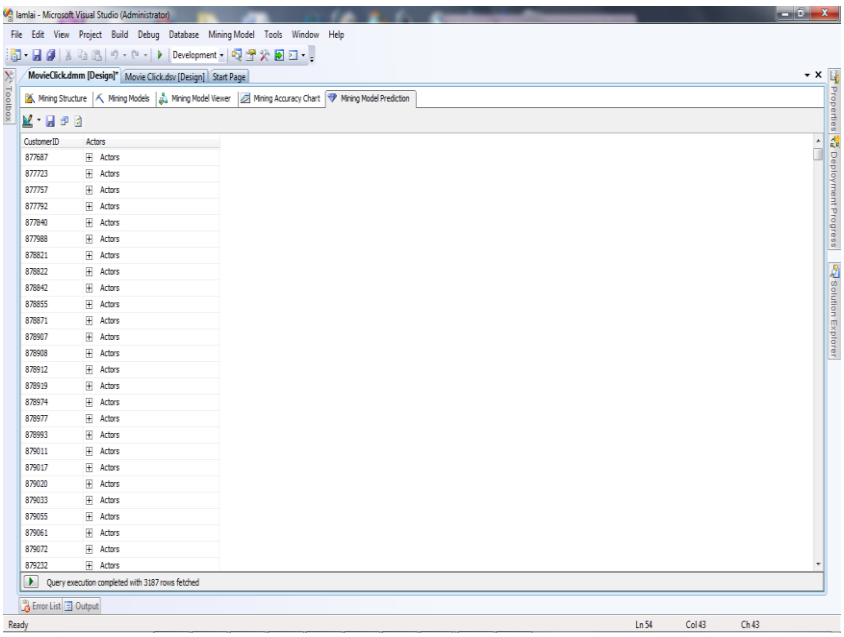
Hình 59 – Màn hình tạo truy vấn

- Chuyển sang chế độ query chúng ta có thể coi được câu query



Hình 60 – Màn hình xem truy vấn DMX

- Chuyển sang chế độ design để xem dữ liệu hiện ra



Hình 61 – Màn hình xem kết quả ở chế độ design