ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐINH ĐỨC LONG

KỸ THUẬT DATAMINING ĐỂ KHUYẾN NGHỊ KHÁCH HÀNG TRONG HỆ THỐNG BI (BUSINESS INTELLIGENCE)

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2015

ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

ĐINH ĐỨC LONG

KỸ THUẬT DATAMINING ĐỂ KHUYẾN NGHỊ KHÁCH HÀNG TRONG HỆ THỐNG BI (BUSINESS INTELLIGENCE)

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

HƯỚNG DẪN KHOA HỌC: TS HOÀNG ĐỖ THANH TÙNG

THÁI NGUYÊN - 2015

LỜI CAM ĐOAN

Luận văn là kết quả nghiên cứu và tổng hợp các kiến thức mà học viên đã thu thập được trong quá trình học tập tại trường Đại học Công nghệ Thông tin và Truyền thông - Đại học Thái Nguyên, dưới sự hướng dẫn, giúp đỡ của các thầy cô và bạn bè đồng nghiệp, đặc biệt là sự hướng dẫn, giúp đỡ của TS Hoàng Đỗ Thanh Tùng - Viện Công nghệ Thông tin, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Em xin cam đoan luận văn không phải là sản phẩm sao chép của bất kỳ tài liệu khoa học nào.

Thái Nguyên, ngày 30 tháng 6 năm 2015

Học viên

Đinh Đức Long

LỜI CẢM ƠN

Em xin gửi lời cảm ơn tới Trường Đại học Công Nghệ Thông Tin và Truyền thông - ĐHTN, Viện Công nghệ Thông tin - Viện Hàn lâm Khoa học và Công nghệ Việt Nam, nơi các Thầy cô đã tận tình truyền đạt các kiến thức quý báu cho em trong suốt quá trình học tập. Xin cảm ơn Ban chủ nhiệm khoa và các cán bộ khoa đã tạo điều kiện tốt nhất cho chúng em học tập và hoàn thành đề tài tốt nghiệp của mình.

Đặc biệt, em xin gửi lời cảm ơn sâu sắc nhất tới TS Hoàng Đỗ Thanh Tùng, người đã trực tiếp hướng dẫn, giúp đỡ để em hoàn thành luận văn của mình.

Mặc dù đã hết sức cố gắng hoàn thành luận văn với tất cả sự nỗ lực của bản thân, nhưng luận văn vẫn còn những thiếu sót. Kính mong nhận được những ý kiến đóng góp của quý thầy, cô và bạn bè đồng nghiệp.

Em xin chân thành cảm ơn!

Thái Nguyên, ngày 30 tháng 6 năm 2015

Học viên

Đinh Đức Long

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Viết đầy đủ	Ý nghĩa
BI	Business Intelligence	Hệ thống trí tuệ doanh nghiệp
DSS	Decision Support Systems	Hệ thống hỗ trợ ra quyết định
DM	Data Mining	Khai phá dữ liệu
IMDB	Internet Movies DataBase	Dữ liệu các bộ phim trên internet
KDD	Knowledge Discovery in Database	Khám phá tri thức trong cơ sở dữ liệu
OLAP	On – Line Analytical Processing	Phân tích dữ liệu trực tuyến đa chiều
RS	Recommender System	Hệ thống khuyến nghị

DANH MỤC CÁC HÌNH VẼ

MỞ ĐẦU

I. ĐẶT VẤN ĐỀ

Trong những năm gần đây, sự phát triển của thương mại điện tử (E-Commerce) đã đem lại nhiều lợi ích to lớn cho nền kinh tế toàn cầu. Thông qua thương mại điện tử, nhiều loại hình kinh doanh mới được hình thành, trong đó có mua bán hàng trên mạng. Với hình thức mới này, người tiêu dùng có thể tiếp cận với hàng hóa một cách dễ dàng và nhanh chóng hơn rất nhiều so với phương thức mua bán truyền thống trong môi trường cạnh tranh ngày càng tăng, các doanh nghiệp/tổ chức (DN/TC) đã nhận ra rằng để có thể thành công cũng như có được những kết quả tốt trong kinh doanh thì vấn đề nhận ra các xu hướng và cơ hội của thị trường là r ất quan trọng, từ đó đáp ứng nhanh cho các nhu cầu của khách hàng mới . Một cách ngắn gọn hơn mục tiêu chính của các doanh nghiệp là hướng tới mục tiêu của các khách hàng của mình.

Ngày nay việc lưu trữ , xử lý dữ liệu để tổng hợp thông tin và hỗ trợ ra quyết định đã trở nên phổ biến đối với nhiều doanh n ghiệp/tổ chức có nhiều giải pháp cho vấn đề này trong đó Business Intelligence (BI – giải pháp quản trị doanh nghiệp thông minh hay hệ thống trí tuệ doanh nghiệp) là một giải pháp tiêu biểu được nhiều DN /TC lựa chọn cho mục đ ích quản lý và điều hành các hoạt động của mình . Ở các nước phát triển , thuật ngữ Business Intelligence (BI) tạm dịch là giải pháp kinh doanh thông minh hay hệ thống trí tuệ doanh nghiệp không còn mới mẻ , tuy nhiên ở Việt N am chúng ta lĩnh vực này vẫn đang ở mức sơ khai. Vậy BI là gì ?

Business Intelligence (BI)

Có rất nhiều định nghĩa cũng như các quan điểm khác nhau về BI , mỗi định nghĩa đề cập đến một đặc trưng nổi bật của hệ thống BI nhưng chung qui lại tất cả đều đề cập đến khả năng hỗ trợ ra quyết định một cách hiệu quả hay BI còn được gọi là hệ thống hỗ trợ ra quyết định (Decision Support Systems

– DSS). Hoạt động dựa trên cở sở ứng dụng công nghệ thông tin, hệ thống BI là một tập hợp các quy trình và công nghệ mà các doanh nghiệp dùng để kiểm soát khối lượng dữ liệu khổng lồ, khai phá tri thức giúp cho các doanh nghiệp có thể đưa các quyết định hiệu quả hơn trong hoạt động kinh doanh của mình. Công nghệ BI (BI technology) cung cấp một cách nhìn toàn cảnh hoạt động của doanh nghiệp từ quá khứ, hiện tại và các dự đoán tương lai với mục đích là hỗ trợ ra quyết định. BI đã được sử dụng rộng rãi trên thế giới, đặc biệt là ở châu Âu từ nhiều năm nay. Ở Việt Nam hiện nay vẫn còn đang ở dạng sơ khai, mặc dù thị trường này cũng đã có sự góp mặt của nhiều hãng như Microsoft, Oracle, Cognos, Business Objects,.... Các tổ chức doanh nghiệp tại Việt Nam đang trong giai đoạn chuẩn hóa hệ thống thông tin của tổ chức, gồm có nhiều vấn đề dưới nhiều góc độ khác nhau trong hệ thống quản trị tổ chức. Mặc dù sự tăng trưởng, trưởng thành của một tổ chức hay còn gọi là tri thức của doanh nghiệp được tích lũy, thể hiện rõ ràng trên hệ thống dữ liệu hoạt động của doanh nghiệp trong quá khứ. Hệ thống trí tuệ doanh nghiệp là giải pháp toàn diện giúp tổ chức/doanh nghiệp chuẩn hóa hệ thống cơ sở dữ liệu quan hệ ở tầng ứng dụng trên nhiều nền tảng khác nhau, tích hợp dữ liệu vào DataWarehouse, phân tích và tích hợp tri thức nghiệp vụ để khai thác thông tin kinh doanh, thể hiện trên hệ thống báo cáo đa tương tác, nhằm giúp đội ngũ nhân viên kinh doanh, các cấp quản lý có thể ra quyết định và triển khai các giải pháp kinh doanh kịp thời trong môi trường kinh doanh đầy cạnh tranh ngày nay.

Hệ thống khuyến nghị

Hệ thống gợi ý có thể đưa ra những mục thông tin phù hợp cho người dùng bằng cách dựa vào dữ liệu về hành vi trong quá khứ của họ để dự đoán những mục thông tin mới trong tương lai mà người dùng có thể thích. Trong hệ thống gợi ý

Để khách hàng có thể đến và mua được một sản phẩm ưng ý thì một lời tư vấn, một sự trợ giúp là rất quan trọng. Trong phương thức bán hàng truyền thống những lời tư vấn như thế từ một người bán hàng sẽ tạo ra một lợi thế rất lớn cho cửa hàng. Do đó để phương thức bán hàng qua mạng thực sự phát triển thì bên cạnh các lợi thế vốn có của mình việc có thêm một "người trợ giúp" là hết sức cần thiết.

Một hệ thống gợi ý (Recommender System - RS) tốt có thể đóng vai trò như một người trung gian hỗ trợ khách hàng đưa ra các quyết định mua hàng đúng đắn. Bằng cách xác định mục đích và nhu cầu của khách hàng, hệ thống có thể đưa ra một tập hợp các gợi ý giúp cho người mua dễ dàng chọn lựa sản phẩm yêu thích hơn. Qua đó hiệu suất của việc mua bán hàng trực tuyến được tăng cao một cách đáng kể. Mặc dù vậy, việc xây dựng một hệ thống hoàn chỉnh để tư vấn cho người dùng vẫn còn chưa được quan tâm.

Data Mining (Khai phá dữ liệu).

Một ứng dụng công nghệ thông tin mô tả một quy trình tự động trích xuất các thông tin có giá trị ẩn chứa trong một khối lượng dữ liệu khổng lồ trong bằng cách dự đoán (Predictive Information).

Có nhiều cách định nghĩa cũng như quan điểm về khai phá dữ liệu (Data Mining) nhưng nhìn chung đó là một t huật ngữ rộng thường được sử dụng để mô tả một quá t rình sử dụng các công nghệ , các kỹ thuật khác nhau các ứng dụng phân tích thống kê, học máy để phân tích một khối lượng lớn dữ liệu một cách tự động để khám phá đ ược các thông tin có giá trị trong hàng loạt các thông tin và thực hiện bằng cách xây dựng các mô hình khai phá dữ liệu và sử dụng các mô hình này để dự đoán các dữ liệu mới. [8]

Trên cơ sở đó có thể nhận thấy được tầ m quan trọng của hệ thống trí tuệ doanh nghiệp (BI) cũng như vai trò của Data Mining trong việc phân tích xử lý dữ liệu . Đó cũng là lý do mà em chọn đề tài " *Kỹ Thuật datamining để*

khuyến nghị khách hàng trong hệ thống BI (business intelligence) ” với mục đích là tìm hiểu các k ỹ thuật, trên cơ sở ứng dụng công nghệ thông tin và lợi ích của việc kết hợp khai phá dữ liệu đ ể khuyến nghị khách hàng trong hệ thống BI.

II. ĐỐI TƯỢNG VÀ PHẠM VI NGHIÊN CỨU

Trong khuôn khổ của luận văn em sẽ thực hiện và giải quyết những vấn đề sau:

- Nghiên cứu hệ thống khuyến nghị khách hàng.
- Tìm hiểu một số thuật toán khai phá dữ liệu trong hệ thống khuyến nghị.
- Đánh giá thử nghiệm hệ thống khuyến nghị t

III. Ý NGHĨA KHOA HỌC VÀ THỰC TIỄN CỦA ĐỀ TÀI

- Tìm hiểu các phương pháp/kỹ thuật/thuật toán cho hệ thống khuyến nghị để định hướng nghiên cứu lâu dài và đưa vào thực tiễn.
- Phát triển hướng nghiên cứu đưa hệ thống BI khuyến nghị vào triển khai thực tiễn cho các công ty kinh doanh trực tuyến.

IV. PHƯƠNG PHÁP NGHIÊN CỨU

- Nghiên cứu các tài liệu liên quan đến các kỹ thuật khai phá dữ liệu trong hệ thống khuyến nghị.
- Phân tích và tổng hợp lý thuyết
- Phương pháp thực nghiệm qua chương trình thử nghiệm

CHƯƠNG 1: TÌM HIỂU HỆ THỐNG BUSINESS INTELLIGENCE

1.1 Business Intelligence (BI) là gì ?

Hệ thống trí tuệ doanh nghiệp (BI) hay còn được gọi là hệ thống hỗ trợ quyết định (Decision Support Systems – DSS). Có rất nhiều định nghĩa về hệ thống BI mỗi định nghĩa mô tả một đặc trưng nổi bật của hệ thống BI nhưng chung qui lại tất cả đều đề cập đến khả năng trợ giúp ra quyết định hiệu quả của BI.

Dưới đây là một số quan điểm về hệ thống BI:

Stackowiak (2007) định nghĩa hệ thống BI như một quá trình thu nhập dữ liệu với khối lượng lớn , phân tích các dữ liệu đó và thể hiện các kết quả bằng các báo cáo. Kết quả này có thể sử dụng để quản lý hoặc thực hiện quyết định một hành động nào đó khi có được các thông tin này . Cũng theo Cui và các cộng sự (2007) thì BI được xem như là một cách thức cải thiện hiệu quả kinh doanh bằng cách khi đã có được các thông tin quan trọng qua quá trình phân tích chẳng hạn như mặt hàng nào thường được bán chạy nhất hay khách hàng nào thường mua hàng với số lượng nhiều… .., từ thông tin đó doanh nghiệp hoặc tổ chức sẽ đưa ra một hành động tương ứng với thông tin có được. Như chăm sóc các khách hàng mua với số lượng nhiều , quan tâm đến các mặt hàng được bán nhiều… .., qua đó mang lại một giá trị gia tăng cho tổ chức của mình hay nói một cách khác BI như là phương pháp để cải thiện hiệu suất kinh doanh của các tổ chức/doanh nghiệp nói chung. [8]

Các khái niệm về BI cũng được đưa lên bởi Gartner Group từ những năm 1996, BI là ứng dụng của một tập hợp các phương pháp , các công nghệ như J2EE, DotNet, dịch vụ Web, XML, kho dữ liệu (Data warehouse), OLAP, khai thác dữ liệu , công nghệ biểu diễn vv…để nâng cao hiệu quả hoạt động của doanh nghiệp, hỗ trợ cho quá trình quản lý và quyết định để đạt được lợi thế cạnh tranh [8].

Gangadharan và Swamy (2004) xác định BI là kết quả của một quá trình phân tích chi tiết các dữ liệu kinh doanh . Họ đã mở rộng định nghĩa về BI như các công cụ quản lý có khả năng bao quát , hoạch định nguồn lực doanh nghiệp, hệ thống hỗ trợ quyết định và khai thác dữ liệu [8].

Berson cùng các cộng sự (2002) và Curt Hall (1999) định nghĩa BI bao gồm một số phần mềm để trích xuất chuyển đổi và nạp dữ liệu , kho dữ liệu, các cách thức truy vấn cơ sở dữ liệu và khả năng tạo báo cáo. Bên cạnh đó với kỹ thuật phân tích dữ liệu trực tuyến đa chiều OLAP (On – Line Analytical Processing), phân tích dữ liệu, khai thác dữ liệu và trực quan hóa dữ liệu.

Business Inteligence – BI (tạm dịch là giải pháp quản trị doanh nghiệp thông minh hay hệ thống trí tuệ doanh nghiệp) là một hệ thống báo cáo cho phép tổ chức/doanh nghiệp (TC/DN) khai thác dữ liệu từ nhiều nguồn khác nhau về khách hàng (KH), thị trường, nhà cung cấp, đối tác, nhân sự... và phân tích/sử dụng các dữ liệu đó thành các nguồn thông tin có ý nghĩa nhằm hỗ trợ việc ra quyết định. Thông thường cấu trúc một bộ giải pháp BI đầy đủ gồm một kho dữ liệu tổng hợp (datawarehouse) và các bộ báo cáo, bộ chỉ tiêu quản lý hiệu năng TC/DN (Key Perfomance Indicators – KPIs), các dự báo và phân tích giả lập (Balance Scorecards, Simulation and Forecasting...).

Business Intelligence đề cập đến các kỹ năng, qui trình, công nghệ, ứng dụng được sử dụng để hỗ trợ ra quyết định.

BI là các ứng dụng và công nghệ để chuyển dữ liệu doanh nghiệp thành hành động

BI là công nghệ mới giúp doanh nghiệp hiểu biết về quá khứ và dự đoán tương lai.

Tóm lại BI được xem như một giải pháp giúp cho tổ chức/doanh nghiệp (TC/DN) nắm bắt được thông tin , tri thức mà giúp cho TC /DN ra quyết định tốt hơn.

Vì vậy một hệ thống BI còn được gọi là hệ hỗ trợ quyết định (Decision Support System -DSS)

1.2 Vai trò của Data Mining trong hệ thống BI

Hệ thống trí tuệ doanh nghiệp (BI) như theo các định nghĩa đã nêu trên bao hàm một hệ thống đa dạng các ứng dụng phần mềm được sử dụng để phân tích dữ liệu của tổ chức/doanh nghiệp. BI được tạo nên từ các hoạt động có liên hệ chặt chẽ với nhau bao gồm :

Khai thác dữ liệu (Data Mining)

Xử lý phân tích trực tuyến (OLAP)

Truy vấn và báo cáo (Query and Report)

Mỗi doanh nghiệp /tổ chức dựa vào việc phân tích dữ liệu nhằm mục đích là gia tăng các hoạt động bán hàng cũng như khẳng định được vị trí của mình trong thị trường cạnh tranh . Kỹ thuật khai phá dữ liệu được sử dụng để phân tích lượng dữ liệu lớn bên cạnh đó khai phá dữ liệu đưa ra một số các kỹ thuật khác nhau đối với mục đích của hệ thống BI. Tại thời điểm hiện tại khai phá dữ liệu đã và đang được sử dụng nhiều hơn và được xem là một trong các giải pháp hàng đầu cho hệ thống BI.

Khai thác dữ liệu cung cấp một khuôn mẫu cho hệ thống BI trên cơ sở đó để phân tích và phát hiện ra các thông tin về các hoạt động dựa trên dữ liệu từ lịch sử hoạt động của doanh nghiệp trên mọi cấp độ . Kho dữ liệu (Data warehouse) và hệ thống BI cung cấp mộ t phương pháp cho người dùng để dự đoán các xu hướng trong tương lai từ việc phân tích dữ liệu từ quá khứ . Bản chất của khai phá dữ liệu mang nhiều tính năng chuyên biệt hơn nó đưa ra các nhìn nhận sâu sắc hơn về kho dữ liệ u, việc ứng dụng khai phá dữ liệu trong một doanh nghiệp sẽ giúp tìm ra được các xu hướng mới từ các dữ li ệu, thông tin trong quá khứ.[3]

1.2 1 Khai phá dữ liệu(Data Mining - DM)

Con người đã ghi lại các hiểu biết của mình từ lúc bắt đầu của cuộc sống. Đó là các hình vẽ trong các bức hang động từ cổ xưa để lại , nó ghi lại các hoạt động diễn ra thường ngày của con người như săn bắt , hái lượm sự sinh ra hoặc kết thúc một cuộc sống… ..vv. Ở bất cứ đâu con người luôn ghi nhận phản ánh lại thực tế cuộc sống được qui định bằng một số hình thức và các phương tiện khác nhau như các hình vẽ , các ngôn ngữ tượng hình … .vv. Họ có thể mô tả và dự đoán các yếu tố là m ảnh hưởng đến vụ thu hoạch cây ôliu ở vùng địa trung hải , ngày nay với các nhà khảo cổ học và nhân chủng học công bố các phát hiện và tìm kiếm của họ để từ đó có các suy đoán về quá khứ từ những vật chứng thu được.

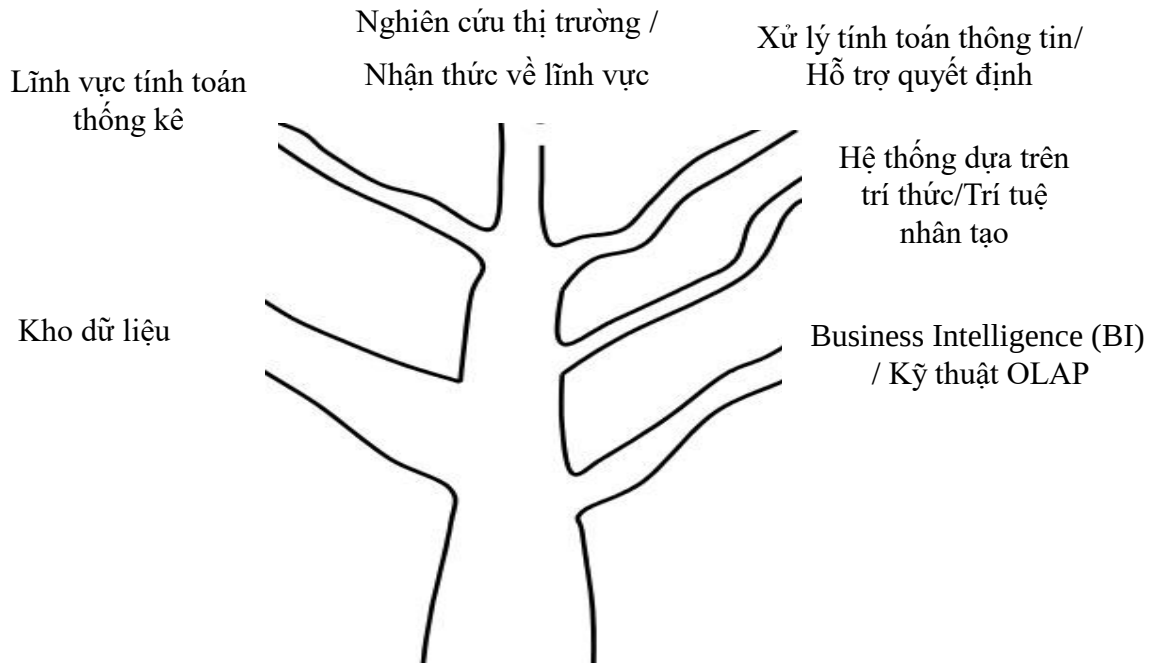
Đó là cách thu nhập thông tin từ xa xưa của con người. Vì vậy khai phá dữ liệu có nguồn gốc từ rất lâu đời với mong muốn tóm tắt lại các kinh nghiệm của cuộc sống , các hoạt động hàng ngày và thông qua một số hình thức như biểu tượng để mô tả chúng một cách tốt hơn.[3]

Data Mining được gọi là khai thác dữ liệu hay khám phá tri thức được xem như là một khái niệm mới lạ gần đây tuy nhiên nếu xét về bản chất thì khái niệm này cũng đã hình thành từ khi con người ghi nhận lại các hoạt động của mình từ khi nền văn minh bắt đầu hình thành.

Ngày nay khai thác dữ liệu là một thuật ngữ diễn tả việc máy tính thực hiện mô phỏng các hoạt động của con người theo hình thức vượt thời gian. Nó mô tả quá trình sử dụng các phương pháp để khám phá được ý nghĩa , các xu hướng, các mối quan hệ của dữ liệu trong một cơ sở dữ liệu dựa vào các dấu vết để lại một cách tự động . Việc sử dụng khai thác dữ liệu để đạt được mục đích là có được một cái nhìn sâu sắc hơn từ đó đưa ra một lựa chọn tốt hơn với từng hoàn cảnh cụ thể để cải thiện hình thức kinh doanh . Nhưng làm thế nào để thể hiện các th ông tin mà công việc khai thác dữ liệu thu được . Nó

được thể hiện thông qua các mô hình khai phá dữ liệu . Bằng cách xây dựng các mô hình khai phá dữ liệu có thể được dùng để đưa ra các dự đoán mô phỏng các sự kiện trong thực tế với phạm vi rất rộng đây chính là điểm mạnh của khai phá dữ liệu hay khám phá tri thức.[3]

Nguồn gốc của khai thác dữ liệu được thể hiện ở hình 1.1

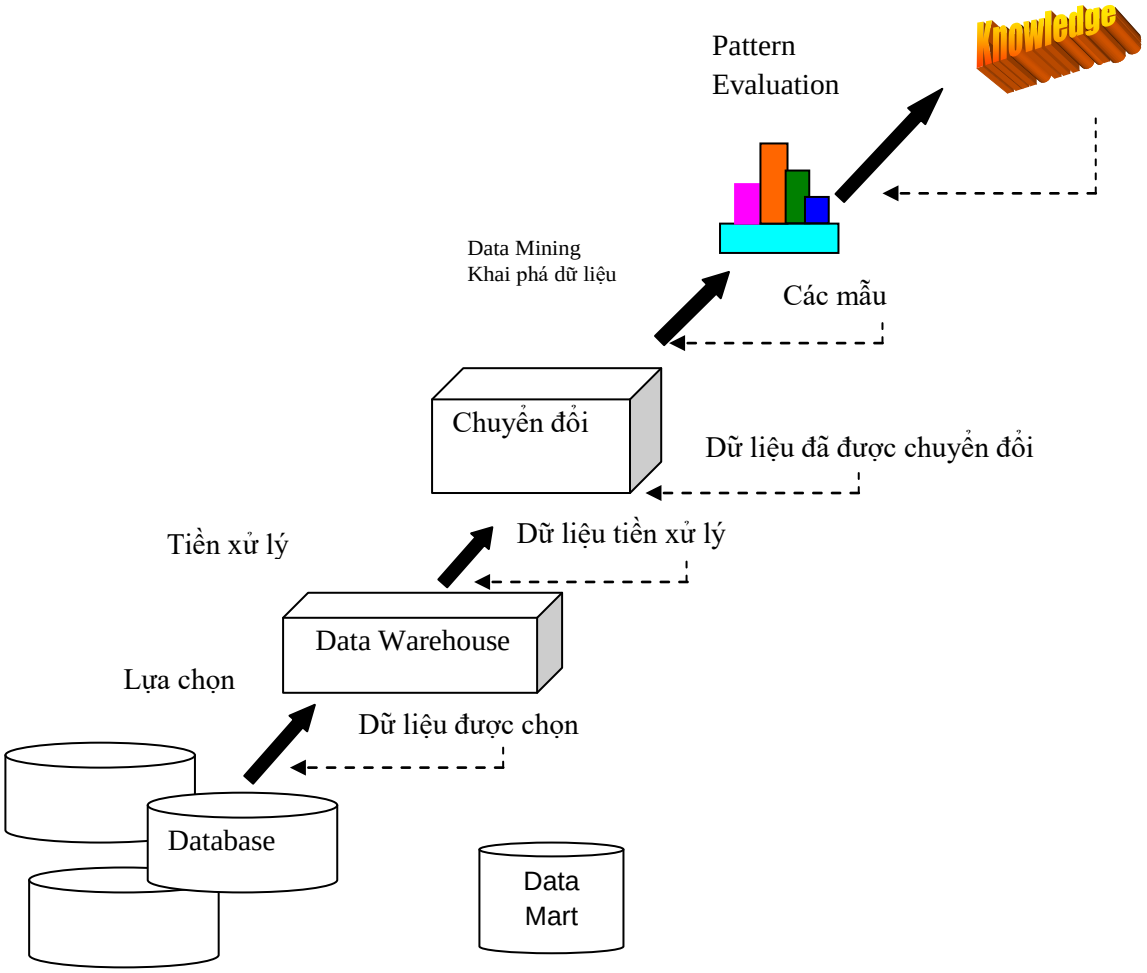


Hình 1.1. Nguồn gốc của khai thác dữ liệu

1.2.2 Khám phá tri thức trong CSDL (Knowledge Discovery in Database - KDD)

Việc phân tích dữ liệu để tìm ra được những thông tin tiềm ẩn có giá trị mà trước đó chưa được phát hiện hoặc bị che lấp , bên cạnh đó là các xu hướng phát triển cũng như yếu tố tác động lên chúng . Công việc này gọi là khám phá tri thức trong cơ sở dữ liệu (KDD) và kỹ thuật cho phép lấy được các tri thức chính là kỹ thuật khai phá dữ liệu (DM). Dữ liệu thường được cho bởi các giá trị mô tả các sự kiện , hiện tượng cụ thể. Còn tri thức (knowledge) khó có thể đưa ra định nghĩa chính xác và phân biệt với dữ liệu nhưng trong những ngữ cảnh nhất định thì có thể và rất cần thiết. Tuy nhiên chúng ta có thể

coi tri thức như là các thông tin được tích hợp bao gồm các sự kiện và các mối quan hệ giữa chúng. Các mối quan hệ này có thể nhận biết , phát hiện hay học được. Nói một cách khác tri thức có thể coi như là dữ liệu có độ trừu tượng và tổ chức cao ví dụ như các luật kết hợp mô tả các thuộc tính của dữ liệu , các mẫu thường xuyên xảy ra , hoặc các nhóm có chung thuộc tính trong CSDL….Các bước của qui trình khám phá tri thức được thể hiện trong hình1.2



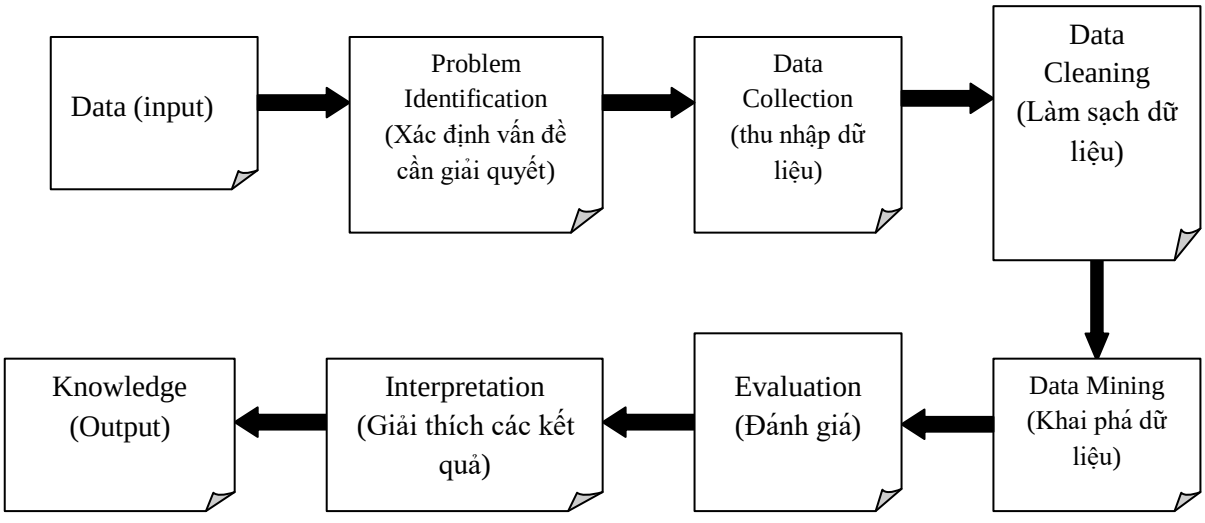
Hình 1.2. Các bước trong qui trình khám phá tri thức trong CSDL

Việc trích rút tri thức từ một khối lượng dữ liệu lớn được xem như một quá trình tương tác lặp đi lặp lại và không phải một hệ thống phân tích tự động. Quá trình này ám chỉ cách hiểu tổng thể về lĩnh vực ứng dụng bao gồm:

- Lựa chọn các dữ liệu cần thiết từ dữ liệu ban đầu : lựa chọn dữ liệu theo một số tiêu chí nhất định phục vụ cho mục đích yêu cầu đặt ra.

- Tích hợp dữ liệu vào kho dữ liệu.
- Tiền xử lý dữ liệu: xử lý các dữ liệu không đầy đủ, không mang tính nhất quán.
- Biến đổi dữ liệu: đưa dữ liệu về dạng thuận lợi nhất phục vụ cho các kỹ thuật khai phá dữ liệu ở bước sau.
- Khai phá dữ liệu: Đây là bước quan trọng áp dụng các kỹ thuật khai phá phần lớn là các kỹ thuật học máy (machine learning) để trích chọn được các mẫu (Pattern) thông tin, các mối liên hệ đặc biệt trong dữ liệu.
- Đánh giá các mẫu /mô hình: Dùng các kỹ thuật hiển thị để trình bày các mẫu hoặc mô hình, các mối liên hệ theo một dạng gần gũi với người sử dụng như đồ thị biểu đồ, bảng biểu, luật kết hợp dạng đơn giản… .đồng thời đánh giá những tri thức thu được theo những tiêu chí nhất định.
- Biểu diễn, sử dụng các tri thức thu được.

Bước quan trọng nhất trong quá trình khám phá tri thức trong CSDL là khai thác dữ liệu được mô tả như hình 1.3 [8]



Hình 1.3. Các bước của quá trình khai phá dữ liệu

Tóm lại ta có thể định nghĩa hai khái niệm DM và KDD như sau :

Khám phá tri thức trong CSDL là một quá trình của việc xác định các giá trị, các điều mới lạ , các thông tin tiềm ẩn kết quả cuối cùng của các mẫu/các mô hình trong dữ liệu. Khai phá dữ liệu là một bước trong quá trình khám phá tri thứ c bao gồm các thuật toán khai phá đặc biệt nằm trong giới hạn khả năng của máy tính để tìm ra các mẫu các mô hình trong dữ liệu.

1.2.3 Vai trò của DM trong hệ thống BI

Có rất nhiều lý do để giải thích cho sự cần thiết của v iệc khám phá tri thức và khai phá dữ liệu và lợi ích của chúng trong hệ thống BI và điển hình như sau :

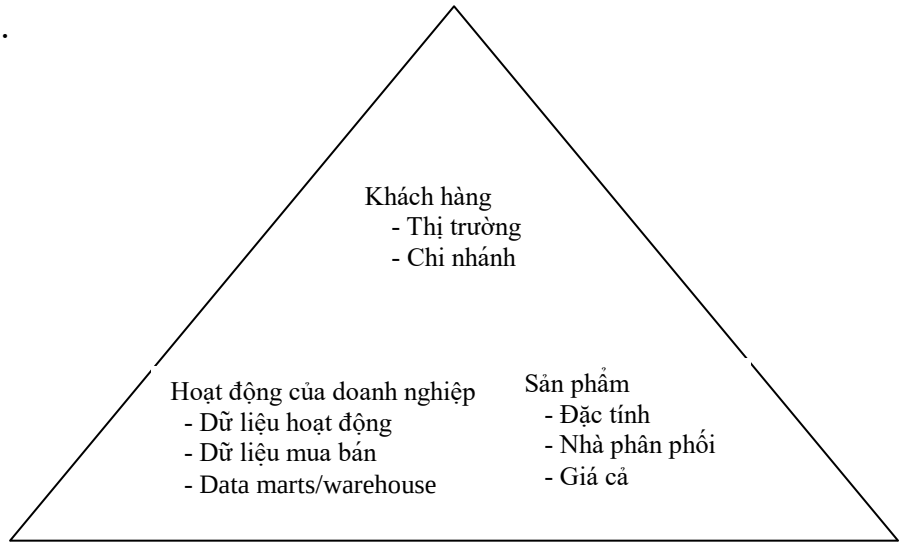
- Các dữ liệu trong hoạt động kinh doanh được lưu trữ rất nhiều vậy các doanh nghiệp phải làm gì với các dữ liệu này.
- Trong hoạt động kinh doanh cần thu nhập các thông tin về thị trường các công ty khác, các khách hàng…trong sản xuất cần thu nhập các số liệu về thời điểm hiệu quả phục vụ cho mục đích cải tiến các quy trình giải quyết c ác sự cố.
- Chỉ có một phần nhỏ của dữ liệu được đưa ra phân tích.
- Với sự gia tăng của dữ liệu cản trở các phương pháp phân tích truyền thống, cũng TC/DN không thể nhìn nhận một cách trọn vẹn các dữ liệu đã lưu.trữ
- Các người dùng đầu cuối không phải là những người am hiểu về các lĩnh vực chuyên môn , họ chỉ cần biết tri thức chứa trong CSDL mà họ đang lưu trữ.
- Cùng với việc lớn lên của CSLD , khả năng đưa ra quyết định và hỗ trợ phân tích thì rất khó thực hiện với các truy vấn CSDL truyền thống.

Khai thác dữ liệu cung cấp ba lợi thế lớn cho các doanh nghiệp :

- Khai thác dữ liệu cung cấp các thông tin về quy trình kinh doanh , các thông tin về khách hàng và hành vi của thị trường.

▪ Tận dụng dữ liệu có sẵn trong quá trình thu nhập dữ liệu từ hoạt động của doanh nghiệp, các data Mart, data warehouse.

▪ Nó cung cấp một mẫu về hành vi được phản ánh trong dữ liệu từ đó tích lũy thêm c ác kinh nghiệm tri thức và khả năng dự đoán các xu hướng tương lai.



Hình 1.4. Vai trò của khai thác dữ liệu và khám phá tri thức trong 3 lĩnh vực chính của một doanh nghiệp.

Bằng cách cung cấp th êm các thông tin về thị trường nó thúc đẩy gia tăng khả năng cạnh tranh của các doanh nghiệp.

Kể từ khi có khai thác dữ liệu có thể khai thác được các thông tin trong dữ liệu tổng hợp của doanh nghiệp cũng như phản ánh lại bất kỳ một thông tin nào thuộc một trong ba lĩnh vực chính này nó có thể cung cấp lợi thế kinh doanh trong các lĩnh vực kể trên với phạm vi rất rộng của dữ liệu và có liên quan tới các lĩnh vực bao gồm tiếp thị, bán hàng, kỹ thuật, công nghệ, các yếu tố về tài chính và con người…..

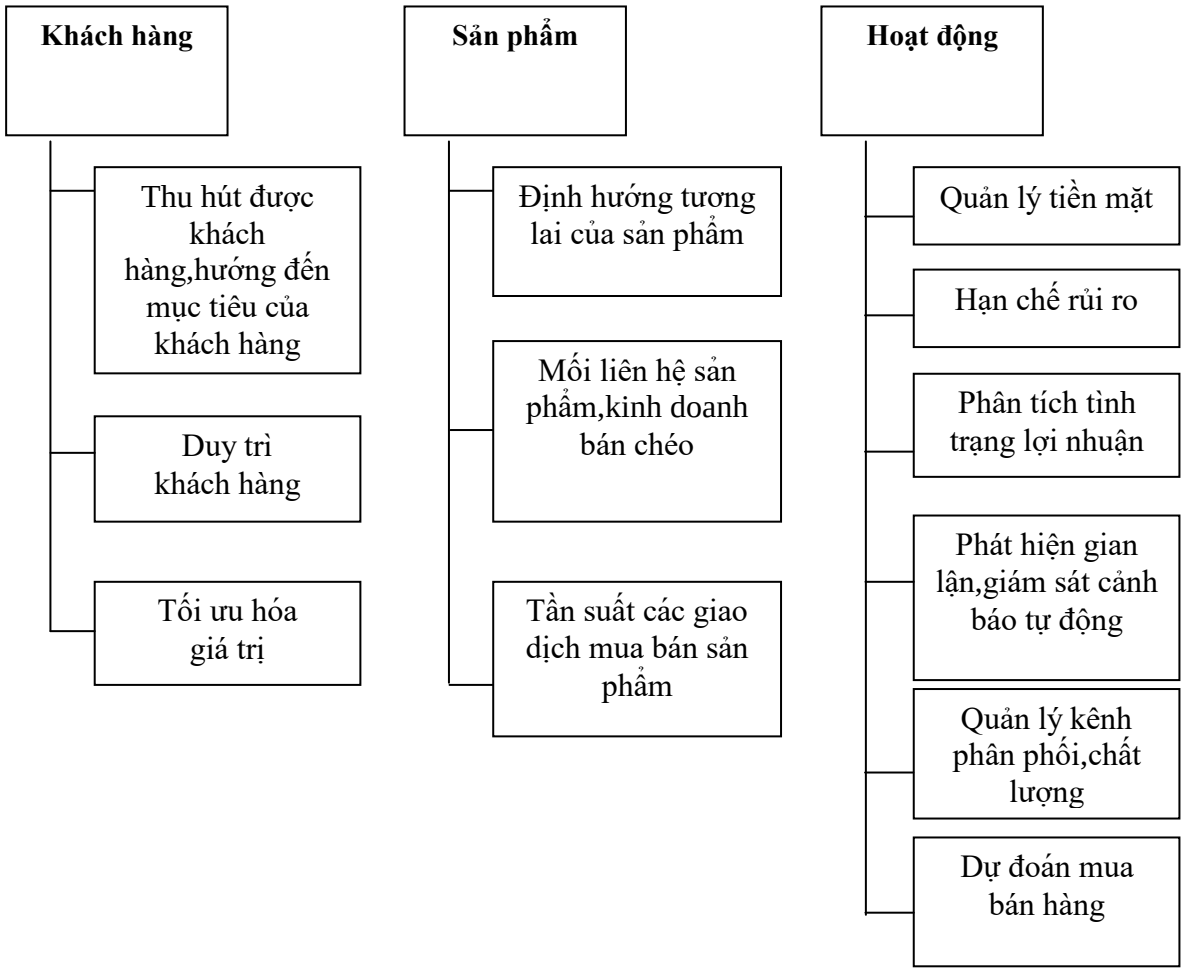
Nhiều nhóm khai thác dữ liệu thường xuyên được triển khai việc khai thác dữ liệu trong ba lĩnh vực như trong các ví dụ sau [3]:

▪ *Xác định khách hàng t iềm năng (Customer Excellence):* Các nhà Băng ở Mỹ sử dụng khai thác dữ liệu để xác định được các khách hàng tiềm

năng bằng cách khai thác các đặc điểm cụ thể của từng khách hàng mục đích tìm thu được lợi nhuận cao nhất cũng như việc đưa ra được các triển vọng mới cho các hợp đồng cho vay một cách hợp lý . Thông qua các hồ sơ của khách hàng họ có thể xác định về các khoản vay với các nhu cầu khác nhau của khách hàng như mua nhà , mua xe …, từ đó có thể tìm được khách hàng tiềm năng cho khoản vay đó cũng như việc xác định các khác hàng đã vay rồi.

- *Hoạt động một cách hiệu quả (Operation Excellence):* Doanh nghiệp American Express sử dụng một kho dữ liệu để lưu trữ toàn bộ hoạt động của các doanh nghiệp trên toàn thế giới với mục đích thương lượng để giảm chi phí từ các nhà cung cấp sản phẩm để loại bỏ những chi phí cao và ngược lại để xác định và thúc đẩy các hoạt động mang lại lợi nhuận cao cho doanh nghiệp.

- *Cung cấp các dịch vụ hàng đầu :* Một trong những nhà cung cấp nhiều dịch vụ trong lĩnh vực viễn thông như hãng Bell ở Canada sử dụng hệ thống BI như một công cụ quản lý các quan hệ khách hàng để đảm bảo rằng cung cấp cho các khách hàng đúng sản phẩm mà họ cần tại đúng thời điểm . Khai phá dữ liệu với mục đích có thể khám phá được các mẫu thông tin hỗ trợ cho việc phát triển và tiếp thị sản phẩm cho khách hàng. Khai phá dữ liệu và khám phá tri thức trong CSDL cung cấp các giải pháp phân tích dữ liệu có được các tri thức và thông tin cho nhiều lĩnh vực khác nhau như khoa học, địa lý, ngân hàng….., đặc biệt đối với doanh nghiệp với các mục tiêu chính được thể hiện trong hình 1.5



Hình 1.5. Vai trò của DM và KDD và các lĩnh vực quan tâm của DN

Tóm lại các ứng dụng tốt nhất của khai phá dữ liệu trong lĩnh vực kinh doanh có thể được đưa ra như sau :

- Có được nhiều khách hàng và biết được mục đích của khách hàng
- Dự đoán xác xuất và giảm bớt rủi ro
- Phân tích hoạt động và tối ưu hóa các hoạt động
- Tiếp thị và các quan hệ
- Phát hiện gian lận và quản lý các chiến lược
- Quản lý hàng tồn kho, các kênh phân phối
- Nghiên cứu thị trường
- Phát triển các sản phẩm, kỹ thuật và kiểm soát chất lượng sản phẩm
- Quản lý bán hàng

1.3 Hệ thống khuyến nghị khách hàng

Trong phần này sẽ tìm hiểu về mô hình hệ thống khuyến nghị dựa trên ma trận khả dụng. Giải thích các ưu điểm của người bán hàng trên mạng so với người bán hàng thông thường, (các cửa hàng truyền thống: siêu thị, đại lý….). Cuối cùng khảo sát ngắn gọn các kiểu ứng dụng mà các hệ thống khuyến nghị hỗ trợ hiệu quả. [1]

1.3.1 Ma trận khả dụng

Trong ứng dụng hệ thống khuyến nghị có 2 lớp thực thể, thông thường là **người dùng** và **mặt hàng**. Người dùng thường có những ưu tiên cho các mặt hàng nhất định và những ưu tiên này phải lấy được ra từ dữ liệu. Bản thân dữ liệu được thể hiện dưới dạng ma trận khả dụng, theo từng cặp người dùng và mặt hàng, giá trị ma trận thể hiện mức độ ưu tiên người dùng đối với một mặt hàng cụ thể. Các giá trị được lấy từ một tập có thứ tự, ví dụ tập các số tự nhiên từ 1-5 thể hiện số ngôi sao mà người dùng đã đánh giá cho sản phẩm đó trên website, điện thoại di động. Ma trận này được giả thiết là thưa, có nghĩa là phần lớn các phần tử là chưa biết. Một đánh giá chưa biết ngụ ý là thông tin về độ ưu tiên của người dùng về mặt hàng đó là chưa rõ ràng.

Ví dụ: Hình 1.6 chỉ ra 1 ví dụ về ma trận khả dụng, đại diện cho đánh giá của người dùng về các bộ phim theo thang 1 – 5, với cấp độ 5 là cao nhất. Phần tử trống là tình huống người dùng chưa đánh giá cho bộ phim đó. Tên bộ phim là HP1, HP2, and HP3 cho bộ **Harry Potter** I, II, và III, TW cho bộ phim **Twilight**, và SW1, SW2, và SW3 cho các tập phim **Star Wars** 1, 2, and 3. Những người dùng được đại diện bằng các chữ cái từ A đến D

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Hình 1.6. Một ma trận khả dụng đại diện cho việc đánh giá các bộ phim theo thang từ 1-5

Chú ý rằng hầu hết các cặp người dùng – bộ phim có giá trị trống, có nghĩa là người dùng không đánh giá bộ phim đó. Thực tế, ma trận thậm chí còn thưa hơn bởi vì những người dùng bình thường chỉ đánh giá phần nhỏ các bộ phim đưa ra.

Mục tiêu của hệ thống khuyến nghị là để dự đoán các giá trị trống trong ma trận khả dụng. Ví dụ: người dùng A có thích SW2 không?. Hệ thống khuyến nghị có thể được thiết kế để đánh giá các thuộc tính của các bộ phim, như là nhà sản xuất, đạo diễn, các ngôi sao, hoặc thậm chí là sự giống nhau về tên của chúng. Nhờ đó, có thể thấy sự tương đồng giữa SW1 và SW2, và kết luận là do A đã không thích SW1 thì khả năng A cũng không thích SW2. Tương tự vậy với nhiều dữ liệu hơn sẽ thấy là những người dùng mà đánh giá cả SW1 và SW2 thì đều có xu hướng đánh giá chúng tương tự nhau. Do đó có thể kết luận rằng A sẽ đánh giá thấp SW2, tương tự như A đánh giá về SW1.

Nhiều ứng dụng có mục tiêu khác nhau, do đó không cần phải dự đoán mọi phần tử trống trong ma trận khả dụng. Thay vì đó chỉ cần tìm ra những phần tử trên một hàng mà có khả năng được đánh giá cao. Trong hầu hết các ứng dụng, hệ thống khuyến nghị không cho người dùng đánh giá tất cả các mặt hàng, mà gợi ý 1 vài mặt hàng mà người dùng đánh giá cao. Thậm chí không cần tìm ra tất cả các mặt hàng với đánh giá được hy vọng là cao nhất, mà chỉ cần tìm 1 tập hợp con của những mặt hàng có đánh giá cao nhất.

Các hệ thống phân phối có giới hạn không gian về kho, và chỉ có thể cho khách hàng xem một phân số nhỏ trong số tất cả các lựa chọn đang có. Mặt khác, các cửa hàng trên mạng có thể cung cấp cho khách hàng tất cả mọi thứ mà cửa hàng có. Do đó, một cửa hàng sách thực thế có thể có vài ngàn quyển sách trên giá, nhưng Amazon có hàng triệu quyển sách. Một tờ báo thực tế có thể in vài tá bài báo mỗi ngày trong khi các dịch vụ tin tức trên mạng cho ra hàng ngàn bài báo mỗi ngày.

Hệ thống khuyến nghị trong thế giới thực tế là khá đơn giản. Đầu tiên, không thể lắp đặt cửa hàng cho từng khách hàng. Do đó, việc lựa chọn nên đưa ra mặt hàng nào sẽ bị chi phối bởi con số có tính chất tổng hợp. Ví dụ, 1 cửa hàng sách sẽ chỉ trưng bày những quyển sách phổ biến nhất, và 1 tờ báo sẽ chỉ in những bài báo mà họ tin rằng hầu hết mọi người sẽ thích. Trong trường hợp đầu tiên, con số bán hàng chi phối sự lựa chọn, trong trường hợp thứ 2, đánh giá của tổng biên tập sẽ chi phối.[2]

1.3.2 Các ứng dụng của hệ thống khuyến nghị

Phần này sẽ đưa ra một số ứng dụng quan trọng của hệ thống khuyến nghị.

1. Ứng dụng của hệ thống khuyến nghị sản phẩm: Có lẽ ứng dụng này được dùng nhiều nhất trong các hệ thống bán lẻ. Amazon hoặc những người bán lẻ qua mạng đưa cho mỗi người dùng một vài gợi ý về sản phẩm mà họ có thể thích mua. Những gợi ý này không phải ngẫu nhiên, mà dựa trên các quyết định mua hàng của các khách hàng khác hoặc dựa vào các kỹ thuật khác mà luận văn này sẽ viết ở chương sau.

2. Các ứng dụng giới thiệu phim: Netflix gợi ý cho khách hàng các bộ phim mà họ có thể thích. Những gợi ý này dựa trên các đánh giá mà người dùng cung cấp, rất giống các đánh giá được gợi ý trong ví dụ ma trận khả dụng ở hình 1.6. Dự đoán đánh giá chính xác là rất quan trọng, do đó Netflix đưa ra 1 giải thưởng 1 triệu đô la cho thuật toán đầu tiên nào mà có thể đánh bại hệ thống gợi ý của Netflix khoảng 10%. Sau 3 năm của cuộc thi giải thưởng được trao cho đội nghiên cứu có tên là "Bellkor's Pragmatic Chaos," sau khi cuộc thi xuất hiện trên 3 năm.[4]

3. Ứng dụng bài báo tin tức: Các dịch vụ tin tức đã nỗ lực để nhận dạng các bài báo mà độc giả ưa thích, dựa trên các bài báo mà họ đã đọc trước đây. Sự giống nhau có thể dựa trên sự giống nhau về các từ quan trọng trong các tài liệu, hoặc dựa trên các bài báo mà những người có cùng thị hiếu đọc. Các nguyên tắc

tương tự áp dụng cho việc giới thiệu các blog từ hàng triệu các blog sẵn có, các videos trên YouTube, hoặc các trang khác mà nội dung được cung cấp đều đặn.

1.3.3 Xây dựng ma trận khả dụng

Nếu không có ma trận khả dụng thì hệ thống gần như không thể gợi ý các mặt hàng. Tuy nhiên, để lấy được dữ liệu để xây dựng ma trận khả dụng thường rất khó. Nhìn chung có 2 cách tiếp cận để khám phá giá trị mà người dùng đánh giá các mặt hàng.

1. Hỏi người dùng để đánh giá các mặt hàng. Nhìn chung đánh giá về phim được thực hiện theo cách này. Các trang mạng cung cấp nội dung, như 1 vài trang tin tức hoặc YouTube cũng yêu cầu người dùng đánh giá các mặt hàng. Hướng tiếp cận này hạn chế về hiệu quả vì nhìn chung người dùng không hài lòng khi đưa ra phản ứng của mình và thông tin từ những người như thế này có thể sai lệch so với thông tin từ những người sẵn lòng đưa ra đánh giá.

2. Tham khảo từ hành vi của người dùng. Nu 1 người dùng mua 1 sản phẩm tại Amazon, xem 1 bộ phim trên YouTube, hoặc đọc 1 bài báo tin tức, thì có thể nói là người dùng "thích" sản phẩm đó. Lưu ý rằng loại hệ thống đánh giá này thực sự chỉ có 1 giá trị: 1 có nghĩa là người dùng thích mặt hàng. Thông thường, một ma trận khả dụng với thông tin dữ liệu là 0, tức là người dùng không mua cũng không xem mặt hàng. Tuy nhiên, trong trường hợp này 0 không phải là đánh giá thấp hơn 1, mà là không đánh giá. Khái quát hơn, có thể thấy sự ưa thích của khách hàng từ hành vi chứ không phải việc mua hàng. Ví dụ, nếu 1 khách hàng Amazon xem thông tin về 1 mặt hàng, có thể hiểu là họ thích mặt hàng, thậm chí cả khi họ không mua nó.[5]

1.4. Kết luận chương 1

Chương 1 đã trình bày những nghiên cứu về hệ thống BI và vai trò của Data Mining trong hệ thống BI. Từ đó triển khai hệ thống khuyến nghị khách hàng và các ứng dụng quan trọng của nó.

CHƯƠNG 2. KHAI PHÁ DỮ LIỆU TRONG HỆ THỐNG BI

2.1 Giới thiệu một số kỹ thuật khai phá dữ liệu dùng trong BI

Các kỹ thuật khai phá dữ liệu thường được chia làm hai nhóm chính đó là kỹ thuật khai phá dữ liệu mô tả và kỹ thuật khai phá dữ liệu dự đoán.

Kỹ thuật khai phá dữ liệu mô tả có nhiệm vụ mô tả về các tính chất các đặc trưng chung trong dữ liệu hiện có bao gồm phân cụm (clustering), tóm tắt (summarization), trực quan hóa (visualization), phân tích phát hiện độ lệch (Evolution and deviation analysis), phát hiện luật kết hợp (association rules),…

Kỹ thuật khai phá dữ liệu dự đoán có nhiệm vụ đưa ra các dự đoán vào các suy diễn trên dữ liệu hiện thời . Các kỹ thuật này gồm có : phân lớp (classification), hồi quy (regression)….

Khai phá dữ liệu có thể được dùng để giải quyết nhiều bài toán với những mục đích và nhiệm vụ khác nhau. Dựa trên bản chất của bài toán có thể chia thành những nhóm bài toán sau:

2.1.1 Phân cụm

Phân cụm (Clustering) là việc nhóm các đối tượng dữ liệu thành các lớp đối tượng có sự tương tự nhau dựa trên thuộc tính của chúng . Mỗi lớp đối tượng được gọi là một cụm (Cluster). Một cụm bao gồm các đối tượng mà giữa bản thân chúng có sự ràng buộc và khác biệt so với các lớp đối tượng khác. Phân cụm còn được gọi là học không giám sát (unsupervised learning). Trong phương pháp này ta không thể biết kết quả của các cụm thu được sẽ thế nào khi bắt đầu quá trình, các cụm có thể tách rời nhau hoặc gối lên nhau hay là một mục dữ liệu có thể vừa thuộc cụm này vừa thuộc cụm kia , vì vậy cần phải có một chuyên gia về lĩnh vực này để đánh giá các cụm thu được.

Phân cụm thường được áp dụng nhiều trong các ứng dụng phân loại thị trường, phân loại khách hàng, nhận dạng mẫu, phân loại trang web,…

2.1.2 Luật kết hợp

Phát hiện luật kết hợp (Association Rules) là một trong các nội dung cơ bản và phổ biến trong khai phá dữ liệu . Phương pháp này nhằm phát hiện ra các luật kết hợp giữa các thành phần dữ liệu trong CSDL. Mẫu đầu ra của khai phá dữ liệu là tập luật kết hợp tìm được.

Ví dụ: Phân tích một CSD L bán hàng và kết quả là những hành khách mua mặt hàng A có xu hướng mua mặt hàng B trong cùng một lần mua được miêu tả trong luật kết hợp sau :

" mua A → mua B" [Độ hỗ trợ 4%, độ tin cậy 70%]

Độ hỗ trợ và độ tin cậy là hai độ đo dùng để đo lường tính hữu dụng của luật trong ví dụ được giải thích như sau :

Độ hỗ trợ 4%: 4% của tất cả các tác vụ đã phân tích chỉ ra rằng A và B đã được mua cùng nhau

Độ tin cậy 70% : 70% các khách hàng mua A thì cũng mua B.

Qua ví dụ trên có thể thấy xét trên quan điểm kinh doanh bán hàng ta có được thông tin mang tính hữu ích mà luật kết hợp mang lại để từ đó đưa ra được quyết định tương ứng với thông tin nhận được.

Ở mức đơn giản nhất BI được xem là các yêu cầu đặt ra của nhà quản lý đối với mỗi hệ thống phần mềm quản lý . Một giải pháp hỗ trợ quyết định hiệu quả cho các doanh nghi ệp ở các mức độ khác nhau đặc biệt là trong các doanh nghiệp có hoạt động bán hàng . Trong hoạt động kinh doanh bán hàng , các nhà quản lý với các thông tin có được mang tính chất thống kê như " 70% khách hàng là khách lẻ khi mua TV thì thường mua loại TV 21 inches"…. những thông tin này rất hữu dụng trong việc đưa ra quyết định , định hướng kinh doanh. Vì vậy việc tìm ra được các luật như vậy bằng kỹ thuật khai phá dữ liệu cụ thể là luật kết hợp là rất quan trọng đối với hệ thống BI . Đó chính là lợi ích của việc áp dụng luật kết hợp trong hệ thống BI với vai trò phân tích

dữ liệu và hỗ trợ quyết định . Với mục đích chính là các tri thức thu được sẽ được sử dụng trong dự báo thông tin trợ giúp trong hoạt động kinh doanh.

2.1.3 Lý thuyết luật kết hợp

Cho trước một tập các giao tác, trong đó mỗi giao tác là một tập các mục, tìm sự tương quan giữa các mục như là một luật và kết quả của giải thuật là tập luật kết hợp tìm được. Luật kết hợp thường có dạng $X \rightarrow Y$.

Trong đó: X là tiền đề, Y là hệ quả (X, Y là hai tập của mục). Ý nghĩa trực quan của luật là các giao tác của cơ sở dữ liệu mà trong đó nội dung X có khuynh hướng đến nội dung Y.

Có hai thông số quan trọng của luật kết hợp là độ hỗ trợ (support) và độ tin cậy (confidence). Độ hỗ trợ và độ tin cậy là hai độ đo của sự đáng quan tâm của luật. Chúng tương ứng phản ánh sự hữu ích và sự chắc chắn của luật đã khám phá. Khai phá các luật kết hợp từ cơ sở dữ liệu là việc tìm các luật có độ hỗ trợ và độ tin cậy lớn hơn ngưỡng mà người dùng xác định trước.

Cho cơ sở dữ liệu gồm các giao dịch T là tập các giao dịch t1, t2, …, tn.

T = {t1, t2, …, tn}. T gọi là cơ sở dữ liệu giao dịch (Transaction Database)

Mỗi giao dịch ti bao gồm tập các đối tượng I (gọi là itemset). I = {i1, i2, …, im}. Một itemset gồm k items gọi là k-itemset

Mục đích của luật kết hợp là tìm ra sự kết hợp (association) hay tương quan giữa các items. Những luật kết hợp này có dạng X =>Ycó thể hiểu rằng những người mua các mặt hàng trong tập X cũng thường mua các mặt hàng trong tập Y. (X và Y gọi là itemset).

Ví dụ, nếu X = {A, B} và Y = {C, D} và ta có luật kết hợp X =>Y có thể nói rằng những người mua A và B thì cũng thường mua C và D.

Độ hỗ trợ (Support) của luật kết hợp X =>Y là tần suất của giao dịch chứa tất cả các items trong cả hai tập X và Y. Ví dụ, *support của luật* X =>Y là 5% có nghĩa là *5% các giao dịch X và Y được mua cùng nhau.*

Công thức để tính độ hỗ trợ (support) của luật X =>Y như sau :

Support(X =>Y) = P(X U Y) = $\frac{n(X \cup Y)}{N}$

N Trong đó N là tổng số giao dịch

Độ tin cậy (Confidence) của luật kết hợp X =>Y là xác suất xảy ra Y khi đã biết X. Ví dụ độ tin cậy của luật kết hợp {A} =>{B} là 80% có nghĩa là 80% khách hàng mua A cũng mua B.

Công thức để tính độ tin cậy của luật kết hợp X => Y là xác suất có điều kiện Y khi đã biết X như sau :

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

Trong đó n(X) là số giao dịch chứa X

Để thu được các luật kết hợp, thường áp dụng 2 tiêu chí: minimum support (min_sup) và minimum confidence (min_conf)

Các luật thỏa mãn có support và confidence thỏa mãn (lớn hơn hoặc bằng) cả Minimum support và Minimum confidence gọi là các luật

Minimum support và Minimum confidence gọi là các giá trị ngưỡng (threshold) và phải xác định trước khi sinh các luật kết hợp.

Các luật thỏa mãn có support và confidence thỏa mãn (lớn hơn hoặc bằng) cả Minimum support và Minimum confidence gọi là các luật kết hợp tốt

Minimum support và Minimum confidence gọi là các giá trị ngưỡng (threshold) và phải xác định trước khi sinh các luật kết hợp.

Một itemsets mà tần suất xuất hiện của nó >= min_sup gọi là frequent itemsets

2.1.4 Thuật toán Apriori sinh luật kết hợp

Tư tưởng chính của thuật toán Apriori là:

- Tìm tất cả frequent itemsets:

k-itemset (itemsets gồm k items) được dùng để tìm (k+1)- itemset.

Đầu tiên tìm 1-itemset (ký hiệu L1). L1 được dùng để tìm L2 (2-itemsets). L2 được dùng để tìm L3 (3-itemset) và tiếp tục cho đến khi không có k-itemset được tìm thấy.

- Từ frequent itemsets sinh ra các luật kết hợp mạnh (các luật kết hợp thỏa mãn 2 tham số min_sup và min_conf)

1. Duyệt (Scan) toàn bộ transaction database để có được support S của 1-itemset, so sánh S với min_sup, để có được 1-itemset (L1)

2. Sử dụng Lk-1 nối (join) Lk-1 để sinh ra candidate k-itemset. Loại bỏ các itemsets không phải là frequent itemsets thu được k-itemset

3. Scan transaction database để có được support của mỗi candidate k-itemset, so sánh S với min_sup để thu được frequent k –itemset (Lk)

4. Lặp lại từ bước 2 cho đến khi Candidate set (C) trống (không tìm thấy frequent itemsets)

5. Với mỗi frequent itemset I, sinh tất cả các tập con s không rỗng của I

6. Với mỗi tập con s không rỗng của I, sinh ra các luật **s => (I-s)** nếu độ tin cậy (Confidence) của nó > =min_conf

Đầu vào: CSDL các items.

Đầu ra: các luật kết hợp và độ tin cậy của mỗi luật.

Ví dụ minh họa mô tả các bước thuật toán Apriori

Giả sử có CSDL giao dịch sau

Giao dịch	Items
1	A,C,D
2	B,C,E
3	A,B,C,E
4	B,E

Xác định min-sup = 2/4

Bước 1: scan toàn bộ các item so sánh với min-sup.

Sau bước này loại D có giá trị nhỏ hơn min-sup

Giao dịch	Items
{A}	2/4
{B}	¾
{C}	¾
{D}	¼
{E}	¾

Thực hiện nối

Giao dịch	Items
{A}	2/4 (số lần xuất hiện của A)
{B}	3/4 (số lần xuất hiện của B)
{C}	3/4 (số lần xuất hiện của C)
{E}	3/4 (số lần xuất hiện của E)

Itemset
{A,B}
{A,C}
{A,E}
{B,C}
{B,E}

Scan lần 2

Itemset	Sup
{A,B}	¼
{A,C}	2/4
{A,E}	¼
{B,C}	2/4
{B,E}	¾
{C,E}	2/4

Sau bước này loại {A,B} và{A,E}

Itemset
{A,B,C}
{A,B,E}
{A,C,E}
{B,C,E}

thực hiện nối

itemset	Sup
{A,C}	2/4
{B,C}	2/4
{B,E}	¾
{C,E}	2/4

Itemset	sup
{A,B,C}	1/4
{A,B,E}	1/4
{A,C,E}	1/4
{B,C,E}	2/4

itemset	sup
{B,C,E}	2/4

Bước 2: Lặp lại các bước nối và scan có được kết quả itemset {B,C,E} thỏa min-sup=2

Bước 3: Từ đó ta có các luật sau :

Áp dụng công thức tính độ tin cậy cho các luật kết hợp

$$\{B\}=>\{C,E\}= \frac{\text{số lần xuất hiện của (B,C,E)}}{\text{số lần xuất hiện của (B)}}$$

Association	Độ tin cậy
{B}=>{C,E}	2/3= 67%
{C}=>{B,E}	2/3= 67%
{E}=>{B,C}	2/3= 67%
{B,C}=>{E}	2/2 = 100%
{B,E}=>{C}	2/3 = 67%
{C,E}=>{B}	2/2 = 100 %

So sánh độ tin cậy thu với độ tin cậy xác đ ịnh trước sẽ đo lường được các luật kết hợp thu được

2.2 Hệ thống khuyến nghị dựa trên nội dung

Như đã đề cập ở chương I, có 2 kiến trúc cơ bản cho 1 hệ thống khuyến nghị :

1. Các hệ thống khuyến nghị dựa trên nội dung tập trung vào đặc tính của mặt hàng. Sự giống nhau của các mặt hàng được xác định bằng cách ước lượng sự tương đồng về các thuộc tính của chúng.

2. Các hệ thống lọc cộng tác tập trung vào mối quan hệ giữa người dùng và các mặt hàng. Sự giống nhau của các mặt hàng được quyết định bằng sự giống nhau về cách mà những người dùng đánh giá về 2 mặt hàng đó

2.2.1 Hồ sơ hàng hóa

Trong 1 hệ thống khuyến nghị dựa trên nội dung, cần phải xây dựng hồ sơ cho mỗi mặt hàng, hồ sơ thể hiện đặc tính của các mặt hàng đó. Trong các

trường hợp đơn giản, hồ sơ bao gồm một vài đặc điểm dễ phát hiện của mặt hàng đó. Ví dụ, cân nhắc đặc điểm của 1 bộ phim mà có thể liên quan đến hệ thống khuyến nghị

1. Dàn diễn viên của bộ phim. Một vài khán giả thích các bộ phim có các diễn viên mà họ yêu thích

2. Đạo diễn. 1 vài khán giả thích tác phẩm của các đạo diễn nhất định

3. Năm bộ phim được sản xuất. Một vài khán giả thích các bộ phim cũ, những người khác thích các bộ phim mới nhất.

4. Thể loại phim. Một vài khán giả chỉ thích hài kịch, những người khác thích phim truyền hình hoặc các tác phẩm lãng mạn

Có rất nhiều các đặc điểm của bộ phim cũng được sử dụng ngoại trừ thể loại phim tùy thông tin của nó đã có sẵn trong phần miêu tả của các bộ phim. Thể loại là một khái niệm mơ hồ. Tuy nhiên, nhìn chung nhiều khán giả gắn tên thể loại phim theo các thuật ngữ hay dùng nhất. Ví dụ **Internet Movie Database** (IMDB) gắn với một thể loại hoặc các thể loại cho tất cả các bộ phim.

Nhiều loại mặt hàng khác cũng cho phép ta có được các đặc điểm từ các dữ liệu sẵn có, mặc dù dữ liệu đó, tại 1 thời điểm nào đó, phải được nhập bằng tay. Ví dụ, các sản phẩm thường có các miêu tả do người sản xuất viết ra, đưa ra các đặc điểm tương ứng với loại sản phẩm đó (ví dụ, kích thước màn hình và màu sắc vỏ TV). Các quyển sách có các miêu tả tương tự như việc miêu tả của các bộ phim, do vậy có thể có các đặc điểm như tác giả, năm xuất bản, và thể loại. Các sản phẩm âm nhạc như đĩa CD và MP3 có các đặc điểm như nghệ sĩ, nhà soạn nhạc và thể loại.

2.2.2 Khám phá đặc điểm của các dữ liệu

Có những lớp mặt hàng mà không dễ gì xác định được các giá trị đặc điểm của chúng. Xét 2 trong số chúng là: Các tập tài liệu và hình ảnh.

Có rất nhiều loại tài liệu mà hệ thống khuyến nghị có thể sử dụng. Ví dụ, có nhiều bài báo tin tức được xuất bản mỗi ngày, mà người dùng không thể đọc tất cả chúng. Một hệ thống khuyến nghị có thể gợi ý các bài báo về các chủ đề mà người dùng ưa thích, nhưng làm thế nào để có thể phân loại các chủ đề? Các trang Web cũng là 1 bộ sưu tập các tài liệu. Có thể gợi ý các trang mà người sử dụng muốn xem không? Giống như vậy, blog cũng có thể được giới thiệu cho những người dùng ưa thích, nếu các blog được phân loại theo chủ đề.

Thật không may, những lớp tài liệu này không có xu hướng có các thông tin sẵn có để đưa ra được các đặc điểm. Một cách khác có ích trong thực tế là nhận dạng các từ mà thể hiện đặc tính chủ đề của tài liệu. Cách thức nhận dạng, đầu tiên loại bỏ các từ thừa – vài trăm từ thông thường nhất, các từ này có xu hướng nói rất ít về chủ đề của tài liệu. Đối với các từ còn lại, tính toán điểm TF.IDF cho mỗi từ trong tài liệu. Những từ có điểm cao nhất là những từ mang đặc điểm của tài liệu.

Sau đó có thể lấy các đặc điểm của một dữ liệu n từ với các điểm TF.IDF cao nhất. Có thể nhặt n là giống nhau cho tất cả các tài liệu, hoặc để n là 1 tỷ lệ phần trăm cố định cho tất cả các từ trong tài liệu. Cũng có thể chọn tất cả các từ mà các điểm TF.IDF ở trên ngưỡng cố định.

Bây giờ, các tài liệu được đại diện bởi bộ các từ. Bằng trực giác, có thể mong đợi các từ này diễn đạt các chủ đề hoặc các ý tưởng chính của tài liệu. Ví dụ, trong 1 bài báo tin tức, có thể mong đợi các từ có điểm TF.IDF cao nhất là những từ chỉ tên người được nói tới trong bài báo, các đặc điểm bất thường của sự kiện được miêu tả, và địa điểm của sự kiện. Để tính toán sự giống nhau của 2 tài liệu, có thể sử dụng 1 vài cách tính toán khoảng cách tự nhiên:

1. Sử dụng khoảng cách Jaccard giữa bộ các từ
2. Sử dụng khoảng cách cosin giữa các bộ được xem như các vector

Để tính toán khoảng cách cosin trong lựa chọn (2), hãy coi các bộ từ - TF.IDF cao như 1 vector với 1 phần tử cho mỗi từ có thể. Vector là 1 nếu từ ở trong bộ và là 0 nếu từ không ở trong bộ. Bởi vì giữa 2 tài liệu, chỉ có số giới hạn các từ nhất định giữa 2 bộ, chiều không giới hạn của vector thì không quan trọng. Phần lớn các phần tử là 0. Trong cả 2 và phần tử 0 không ảnh hưởng đến giá trị tích vô hướng. Để chính xác, tích vô hướng là kích thước giao của hai tập từ và chiều dài của vector là căn bậc hai của số từ trong mỗi bộ. Cách tính đó tính toán cosin của góc giữa các vector khi tích vô hướng được chia bởi phép nhân các độ dài vector.

2.2.3 Lấy đặc điểm của mặt hàng từ thẻ (Tag)

Giả thiết có cơ sở dữ liệu các hình ảnh và làm sao lấy được các đặc điểm của các mặt hàng từ đó. Vấn đề với hình ảnh, dữ liệu của chúng là một mảng các điểm không mang lại thông tin hữu ích về đặc điểm của mặt hàng. Có thể tính toán các đặc tính đơn giản của pixel, lượng trung bình của màu đỏ trung bình trong hình ảnh, nhưng rất ít người dùng tìm các hình ảnh màu đỏ hoặc đặc biệt là thích các hình ảnh màu đỏ.

Có nhiều nỗ lực để thu được thông tin về các đặc điểm của các mặt hàng bằng cách mời gọi những người dùng gắn kết-tag các hàng hóa với các từ hoặc cụm từ miêu tả chúng. Do đó, một hình ảnh với nhiều màu đỏ có thể được gắn thẻ-tag là "quảng trường Ba Đình," hoặc "hoàng hôn ở biển." Sự phân biệt không phải là 1 thứ gì đó mà có thể được khám phá bởi các chương trình phân tích hình ảnh hiện có.

Gần như bất kỳ dữ liệu nào có thể có các đặc điểm được mô tả bởi các thẻ - tag. Một trong những nỗ lực đầu tiên là gắn thẻ khối lượng dữ liệu khổng lồ là trang del.icio.us, sau đó trang này được Yahoo! Mua lại, Yahoo mời những người dùng gắn thẻ vào các trang Web. Mục tiêu của việc gắn thẻ này là để tìm ra 1 phương pháp tìm kiếm mới để người dùng nhập 1 bộ thẻ khi họ

yêu cầu tìm kiếm, và hệ thống sẽ truy hồi các trang Web mà được gắn thẻ theo cách đó. Tuy nhiên cũng có thể sử dụng thẻ như 1 hệ thống khuyến nghị. Quan sát thấy rằng người dùng truy hồi hoặc đánh dấu nhiều trang với một bộ thẻ nhất định, thì có thể giới thiệu các trang khác với cùng thẻ tương tự. Vấn đề của việc gắn thẻ để tiếp cận việc khám phá đặc điểm là quá trình chỉ hoạt động nếu người dùng sẵn lòng mất công để tạo ra thẻ, và có vừa đủ các thẻ gây sai số sẽ không làm ảnh hưởng tới hệ thống.[7]

2.2.4 Trình bày hồ sơ hàng hóa

Mục tiêu cuối cùng cho hệ thống khuyến nghị dựa trên nội dung là để tạo ra cả bộ hồ sơ mặt hàng bao gồm các cặp đặc điểm – giá trị và bộ hồ sơ người dùng mà tổng hợp sự ưa thích của người dùng dựa trên hàng ma trận khả dụng. Trong mục 2.2.2 đã gợi ý làm thế nào xây dựng hồ sơ mặt hàng. Tưởng tượng một vecto 0 và 1, trong đó 1 đại diện cho sự xuất hiện cao của từ TF.IDF trong tài liệu. Vì đặc điểm của các tài liệu là tất cả đều bằng từ nên rất dễ trình bày hồ sơ theo cách này.

Khái quát hóa hướng tiếp cận vecto đối với tất cả các loại đặc điểm. Rất dễ làm như vậy đối với các đặc điểm mà là tập hợp các giá trị rời rạc. Ví dụ, nếu 1 đặc điểm của bộ phim là dàn diễn viên thì tưởng tượng rằng có 1 thành phần cho mỗi diễn viên, với 1 nếu diễn viên tham gia trong phim, và 0 nếu diễn viên không tham gia trong phim. Tương tự như vậy, có thể có thành phần cho từng đạo diễn và từng thể loại. Tất cả các đặc điểm này chỉ sử dụng 0 hoặc 1.

Có 1 bộ các đặc điểm khác mà không được các vecto logic biểu diễn: các đặc điểm đó thuộc về số. Ví dụ, có thể lấy đánh giá trung bình cho các bộ phim là một đặc điểm, giá trị trung bình là số thực. Không có nghĩa khi có 1 thành phần cho mỗi đánh giá trung bình, và làm như vậy sẽ khiến chúng mất cấu trúc ẩn về số. Đó là, 2 đánh giá mà gần nhau nhưng không giống nhau

nên được cân nhắc giống nhau hơn so với các đánh giá khác. Giống như vậy, các đặc điểm trị số của mặt hàng, chẳng hạn như kích cỡ màn hình, dung lượng ổ đĩa PC nên được xem là giống nhau nếu các giá trị của chúng không khác nhau lắm.

Các đặc điểm trị số nên được biểu diễn bởi các thành phần đơn vector đại diện cho các mặt hàng. Các thành phần này có giá trị chính xác của đặc điểm đó.

Không có hại gì nếu 1 số thành phần vector logic và các thành phần khác có giá trị thực hoặc nguyên. Ta vẫn có thể tính toán khoảng cách cosin giữa các vecto, mặc dù nếu làm vậy, ta nên suy nghĩ 1 chút về tỉ lệ phù hợp của các thành phần phi logic để chúng không chi phối việc tính toán cũng như chúng không liên quan.

Ví dụ phần 2.2.1: Giả sử các đặc điểm duy nhất của các bộ phim là dàn diễn viên và đánh giá trung bình. Cân nhắc 2 bộ phim, mỗi bộ phim 5 diễn viên. 2 diễn viên xuất hiện trong cả 2 bộ phim. Cũng như vậy, 1 bộ phim có đánh giá trung bình là 3 và bộ phim còn lại có đánh giá trung bình là 4. Vector sẽ có dạng giống như thế này:

$$\begin{array}{cccccccccc} 0 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 3\alpha \\ 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 & 4\alpha \end{array}$$

Tuy nhiên, về nguyên tắc có một số lượng không giới hạn các thành phần bổ sung, mỗi thành phần có 0 cho cả 2 vector, thể hiện cho tất cả các diễn viên mà có thể không xuất hiện trong bộ phim nào cả. Do khoảng cách cosin của các vecto không bị ảnh hưởng bởi các thành phần trong đó cả 2 vector có 0 nên không cần lo lắng về tác động của các diễn viên mà không xuất hiện trong bộ phim nào. Thành phần cuối cùng biểu diễn đánh giá trung bình. Coi nó có yếu tố tỷ lệ không được biết tới là α.Về α có thể tính toán

cosin của các góc giữa các vector. Tích vô hướng là 2 + 12α2 và Chiều dài của các vector là$\sqrt{5+9\alpha^2}$ và $\sqrt{5+16\alpha^2}$. do đó góc cosin giữa các vector là:

$$\frac{2+12\alpha^2}{\sqrt{25+125\alpha^2+144\alpha^4}}$$

Nếu chọn α = 1, có nghĩa là lấy đánh giá trung bình đúng như vậy, vậy giá trị của phương trình trên là 0.816. Nếu sử dụng α = 2, có nghĩa gấp đôi mức đánh giá, thì cosin là 0.940. Có nghĩa vector xuất hiện gần hướng hơn nếu sử dụng α = 1. Tương tự như vậy, nếu sử dụng α = 1/2, thì cosin là 0.619, khiến vector trông khá khác. Không thể nói giá trị α là "đúng", nhưng thấy rằng việc lựa chọn yếu tố tỷ lệ cho các đặc điểm trị số ảnh hưởng đến quyết định của ta về các mặt hàng giống nhau như thế nào.

2.2.5 Hồ sơ người dùng

Hồ sơ người dùng không chỉ cần tạo ra các vector miêu tả các mặt hàng, cần tạo ra các vector với các thành phần giống nhau miêu tả sự ưa thích của người dùng. Ma trận khả dụng thể hiện sự kết nối giữa người dùng và mặt hàng. Khôi phục các phần tử ma trận trống có thể là thể hiện người dùng mua hàng hoặc có sự liên quan hoặc là một con số tùy biến nào đó thể hiện sự đánh giá ảnh hưởng của người dùng đối với sản phẩm. Với thông tin này, dự đoán tốt nhất có thể đưa ra khi xem xét người dùng thích mặt hàng nào là sự tổng hợp hồ sơ của những mặt hàng đó. Nếu ma trận khả dụng chỉ có giá trị 1 thì sự tổng hợp tự nhiên là trung bình của các thành phần vector đại diện cho hồ sơ của các mặt hàng trong đó ma trận khả dụng là 1 cho người dùng đó.

Giả sử các mặt hàng là các bộ phim, được các hồ sơ logic biểu diễn với các thành phần tương ứng với các diễn viên. Cũng như vậy, ma trận khả dụng là 1 nếu người dùng đã xem bộ phim và trống nếu ngược lại. Nếu 20% bộ phim mà người

dùng U thích có Julia Roberts làm 1 trong những diễn viên của bộ phim, thì hồ sơ của người dùng U sẽ có 0.2 trong thành phần cho Julia Roberts.

Nếu ma trận khả dụng không logic, ví dụ các đánh giá từ 1–5, thì có thể tính các vector đại diện cho hồ sơ các mặt hàng theo giá trị khả dụng. Rất có ý nghĩa khi chuẩn hóa độ khả dụng bằng cách trừ giá trị trung bình cho người dùng. Bằng cách đó có các giá trị âm với những đánh giá dưới mức trung bình và giá trị dương cho những đánh giá trên mức trung bình.

Kết quả này sẽ hữu ích khi ta đề cập đến cách để tìm các mặt hàng mà người dùng thích ở phần sau.

Cân nhắc thông tin bộ phim như trong ví dụ trên, nhưng giả thiết ma trận khả dụng có các phần tử không trống được đánh giá từ 1-5. Giả sử người dùng U đưa ra đánh giá trung bình là 3. Có 3 bộ phim có Julia Robert làm diễn viên, và các bộ phim đó có các đánh giá là 3,4 và 5. Vậy thì trong hồ sơ người dùng U thành phần cho Julia Roberts sẽ có các giá trị trung bình của 3 − 3, 4 − 3, và 5 – 3, là 1.

Mặt khác người dùng V đưa ra mức đánh giá trung bình là 4, và cũng đánh giá 3 bộ phim của Julia Roberts (không vấn đề gì nếu chúng là cùng 3 bộ phim mà U đánh giá hay không). Người dùngV đánh giá 3 bộ phim này theo mức 2, 3, và 5. Về thành Julia Roberts, hồ sơ người sử dụng V có giá trung bình 2 − 4, 3 − 4, và 5 − 4, là −2/3.

2.2.6 Khuyến nghị sản phẩm cho người dùng dựa trên nội dung

Với các vector hồ sơ cho cả người dùng và các mặt hàng, có thể dự đoán cấp độ ưa thích 1 mặt hàng của người dùng bằng cách tính toán khoảng cách cosin giữa vector người dùng và vector mặt hàng như trong ví dụ 2.2. Cũng có thể tính tỷ lệ các thành phần khác nhau mà giá trị không phải là giá trị logic. Mặt phẳng ngẫu nhiên và kỹ thuật theo địa phương (LSH) có thể được sử dụng để đưa các hồ sơ hàng hóa vào các khu vực. Theo cách đó, với

1 người dùng mà muốn giới thiệu 1 vài mặt hàng, có thể áp dụng 2 kỹ thuật đó để xác định trong khu vực nào có thể tìm các mặt hàng có khoảng cách cosin nhỏ với người dùng.

Đầu tiên sử dụng dữ liệu của ví dụ 2.3 Hồ sơ của người dùng sẽ có các thành phần cho các diễn viên theo tỷ lệ về khả năng diễn viên sẽ xuất hiện trong 1 bộ phim mà người dùng thích. Do đó, sự khuyến nghị cao nhất (khoảng cách cosin thấp nhất) thuộc về các bộ phim với nhiều diễn viên xuất hiện trong nhiều bộ phim mà người dùng thích. Miễn là các diễn viên là các thông tin duy nhất mà ta có về đặc điểm của bộ phim thì đó có thể là điều tốt nhất mà ta có thể làm.

Vector cho người dùng sẽ có các con số khả quan cho các diễn viên có xu hướng xuất hiện trong các bộ phim người dùng thích và có các con số không khả quan cho các diễn viên xuất hiện trong các bộ phim mà người dùng không thích.

Xem xét ví dụ 2.4 thấy rằng vector cho người dùng có các con số dương đối với các diễn viên mà có xu hướng trong các bộ phim mà người dùng thích và các con số âm đối với các diễn viên mà có xu hướng trong các bộ phim mà người dùng không thích. Xem xét một bộ phim với nhiều diễn viên mà người dùng ưa thích, và chỉ 1 vài hoặc không có bộ phim nào mà người dùng không thích. Cosin của góc giữa vector của người dùng và vector của bộ phim sẽ là 1 phân số dương lớn. Điều đó ngụ ý 1 góc gần với 0, và do đó khoảng cách cosin giữa các vector là nhỏ.

Với bộ phim mà số lượng diễn viên được nhiều người dùng thích nhiều như số diễn viên người dùng không. Trong tình huống này, cosin của góc giữa người dùng và bộ phim là gần 0, và do đó góc giữa 2 vector là khoảng gần 90 độ. Cuối cùng, xem xét 1 bộ phim với hầu hết các diễn viên mà người dùng

không thích. Trong trường hợp đó cosin sẽ là 1 phân số âm lớn, và góc giữa 2 vector sẽ gần với 180 độ - khoảng cách cosin có thể là lớn nhất.

2.2.7 Các thuật toán phân lớp

Một hướng tiếp cận hoàn toàn khác đối với 1 hệ thống khuyến nghị sử dụng các hồ sơ mặt hàng và các ma trận khả dụng để giải quyết vấn đề như hỏng máy. Xem xét các dữ liệu đã cho như 1 bộ tập luyện, và đối với từng người dùng, xây dựng 1 bộ phân loại để dự đoán đánh giá của tất cả các mặt hàng. Có 1 số lượng lớn các bộ phân loại khác nhau và mục tiêu không phải là để luyện các đối tượng này. Tuy nhiên, nhận thức được sự lựa chọn để phát triển 1 bộ phân loại cho khuyến nghị, vì vậy ta sẽ đề cập đến một bộ phân loại thông thường- cây quyết định.

Một cây quyết định là 1 bộ sưu tập các nốt mạng, được sắp xếp như 1 cây nhị phân. Các lá diễn tả các quyết định, trong trường hợp này, quyết định sẽ là "thích" hoặc "không thích". Mỗi nốt mạng bên trong là 1 điều kiện về các đối tượng được phân loại, trong trường hợp này, điều kiện là tính chất liên quan đến 1 hoặc nhiều đặc điểm của 1 mặt hàng.

Để phân loại 1 mặt hàng, bắt đầu từ gốc, và tại gốc áp dụng thuộc tính vào mặt hàng. Nếu thuộc tính là đúng, đến nhánh con bên trái và nếu thuộc tính sai, đến nhánh con bên phải. Sau đó lặp lại cùng 1 quá trình tại nốt mạng đi qua cho đến khi đến được lá. Lá đó phân biệt mặt hàng là thích hay không thích. Việc xây dựng 1 cây quyết định đòi hỏi lựa chọn thuộc tính cho mỗi nốt mạng bên trong. Có nhiều cách để chọn thuộc tính tốt nhất.

Nhưng chúng đều cố gắng sắp xếp để một trong những nhánh để có tất cả hoặc phần lớn các mẫu dương (những mặt hàng mà người dùng thích) và nhánh khác có tất cả hoặc phần lớn các mẫu âm (những mặt hàng mà người dùng không thích)

Khi đã chọn thuộc tính cho nốt mạng N thì ta phân chia các mặt hàng thành 2 nhóm: nhóm thỏa mãn các thuộc tính và nhóm không. Đối với mỗi nhóm, lại tìm ra thuộc tính mà phân chia tốt nhất các mẫu dương và mẫu âm trong nhóm đó. Các thuộc tính này được gắn cho các nhánh con của N. Quá trình phân chia các mẫu và xây dựng các nhánh con có thể tiến tới thực hiện ở bất kỳ cấp độ nào. Như vậy có thể dừng và tạo ra 1 lá, nếu nhóm của mặt hàng đó cho 1 nốt mạng là đồng nhất, có nghĩa là chúng đều là các mẫu dương hoặc mẫu âm.

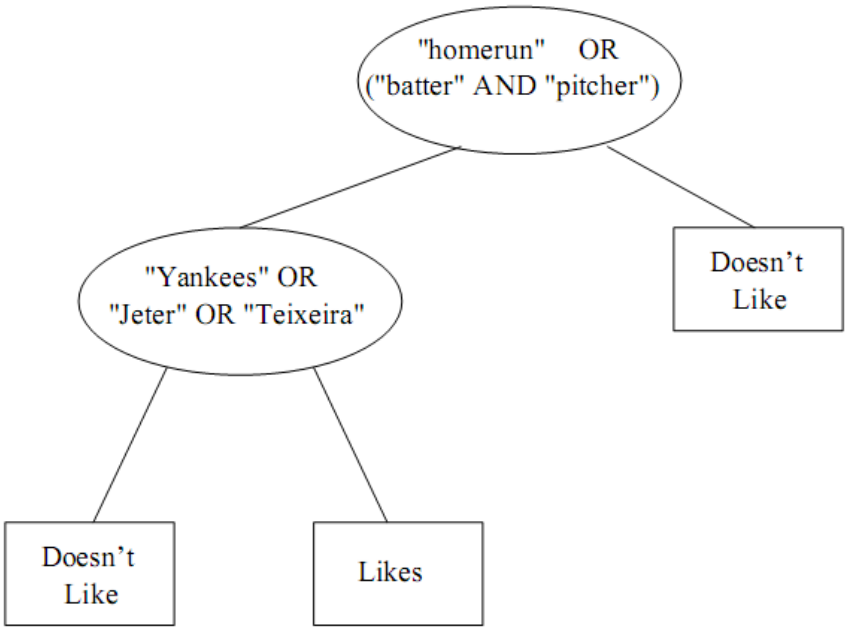
Tuy nhiên, có thể cần phải dừng để tạo ra một nốt lá với quyết định chính cho 1 nhóm mặc dù nhóm đó chứa cả mẫu âm và dương. Lý do là tính thống kê của một nhóm nhỏ có thể không đủ độ tin cậy. Vì lý do đó, có 1 chiến lược khác để tạo ra 1 quần thể các cây quyết định, mỗi cây sử dụng các thuộc tính khác nhau, nhưng cho phép cây sâu hơn những gì dữ liệu sẵn có chứng minh được. Những cây như vậy được gọi là quá hợp (**overfitted)**. Để phân loại 1 mặt hàng, hãy áp dụng tất cả các cây trong quần thể để chúng đưa ra kết quả. Ở đây sự lựa chọn này sẽ không được xem xét, nhưng đưa ra 1 ví dụ giả định đơn giản về 1 cây quyết định.

Giả sử mặt hàng là các bài báo tin tức, và các đặc điểm là trong các tài liệu đó các từ TF.IDF cao (**từ khóa**). Giả sử thêm nữa là có 1 người dùng U thích các bài báo về bóng chày, ngoại trừ các bài báo về New York Yankees. Hàng ma trận khả dụng cho U là 1 nếu U đã đọc bài báo và là trống nếu U không đọc bài báo. Giả sử lấy 1 là “thích” và trống là “không thích.” Các thuộc tính sẽ là phương trình logic của các từ khóa. Do nhìn chung U thích bóng chày nên có thể thấy rằng thuộc tính tốt nhất cho gốc là “homerun” hoặc (“batter” và “pitcher”). Các mặt hàng mà thỏa mãn được các thuộc tính sẽ có xu hướng là các mẫu dương (các bài báo với 1 trong hàng cho U trong ma trận khả dụng), và các mặt hàng mà không thỏa mãn được các thuộc tính sẽ có

xu hướng là những ví dụ mẫu âm (trống trong hàng ma trận khả dụng cho U). Hình 2.1 cho thấy gốc cũng như phần còn lại của cây quyết định.

Giả sử nhóm các bài báo không thỏa mãn được các thuộc tính bao gồm rất ít các mẫu dương nên kết luận rằng tất cả các mặt hàng này là ở trong lớp "không thích". Sau đó có thể đặt 1 lá với quyết định "không thích" là nhánh con bên phải của gốc. Tuy nhiên, các bài báo mà thỏa mãn thuộc tính bao gồm 1 số bài báo mà người dùng U không thích; đây là những bài báo đề cập đến Yankees. Do đó tại nhánh con bên trái của gốc, sẽ xây dựng 1 thuộc tính khác. Có thể nhận ra rằng thuộc tính "Yankees" hoặc "Jeter" hoặc "Teixeira" là chỉ số tốt nhất có thể của 1 bài báo về bóng chày và về Yankees. Do vậy trong hình 2.1 nhánh con bên trái của gốc áp dụng thuộc tính này. Cả 2 nhánh con của nốt mạng này là các lá, do vậy có thể giả sử rằng các mặt hàng thỏa mãn thuộc tính này chiếm ưu thế hơn hẳn về mặt không khả quan và những mặt hàng không thỏa mãn thuộc tính này âm là chủ yếu và thỏa mãn thuộc tính này dương là chủ yếu.

Thật không may, các bộ phân loại của tất cả các loại có xu hướng mất thời gian dài để xây dựng. Ví dụ, nếu muốn sử dụng cây quyết định, cần 1 cây cho 1 người dùng. Xây dựng 1 cây không chỉ yêu cầu xem tất cả các hồ sơ của mặt hàng, mà còn phải xem xét nhiều thuộc tính khác nhau, mà có thể liên quan đến tổ hợp các đặc điểm. Do đó hướng tiếp cận này có xu hướng chỉ được dùng cho các kích cỡ vấn đề tương đối nhỏ.



Hình 2.1. Một cây quyết định

2.3. Lọc cộng tác (collaborative filtering).

Bây giờ bắt đầu với một cách tiếp cận khác trong hệ thống khuyến nghị khác. Thay vì sử dụng đặc điểm của các mặt hàng để xác định sự tương đồng của chúng, ta tập trung vào sự tương đồng trong đánh giá của người dùng cho 2 mặt hàng. Đó là, trong không gian vector hồ sơ-hàng hóa sử dụng mỗi mặt hàng. Hơn nữa, thay vì nghĩ ra 1 vector hồ sơ cho người dùng, thì ta biểu diễn chúng theo hàng trong ma trận khả dụng. Người dùng giống nhau nếu vector của họ gần nhau theo độ đo khoảng cách như Jaccard hoặc khoảng cách cosin. Việc khuyến nghị cho người dùng U được thực hiện bằng cách tìm người dùng giống với U nhất, và khuyến nghị các mặt hàng mà những người dùng này thích. Quá trình nhận dạng những người dùng giống nhau và khuyến nghị những gì mà những người dùng giống nhau thích, được gọi là lọc cộng tác (**collaborative filtering).**[5]

2.3.1 Đo độ tương đồng

Câu hỏi đầu tiên phải giải quyết là làm thế nào để đo độ tương đồng của những người dùng hoặc các mặt hàng từ hàng hoặc cột trong ma trận khả

dụng. Biểu diễn lại hình 1.6 như hình 2.2. Dữ liệu này quá nhỏ nên không thể đưa ra bất kỳ kết luận đáng tin cậy nào nhưng kích cỡ nhỏ sẽ làm rõ một vài khó khăn trong việc chọn ra một độ đo khoảng cách. Quan sát cụ thể người sử dụng A và C. Họ đánh giá 2 bộ phim giống nhau, nhưng dường như họ có quan điểm gần như trái ngược nhau về các bộ phim này. Một độ đo khoảng cách tốt sẽ làm chúng trông tách rời nhau. Sau đây là 1 vài độ đo thay thế để xem xét:

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	4			5	1		
B	5	5	4				
C				2	4	5	
D		3					3

Hình 2.2. Ma trận khả dụng được gợi ý trong hình 1.6

Khoảng cách Pearson:

Khoảng cách Pearson đo mức độ của một quan hệ tuyến tính tồn tại giữa hai đối tượng. Xuất phát từ mô hình hồi quy tuyến tính, khoảng cách Pearson dựa trên một tập các giả định về dữ liệu: thứ nhất mối quan hệ phải là tuyến tính, thứ hai là các lỗi phải độc lập, phân bố xác suất với kì vọng là 0 và độ biến đổi không đổi trên tất cả các biến độc lập. Khi các giả định không còn đúng nữa thì khoảng cách Pearson không còn phản ánh chính xác độ tương tự. Rất may mắn dữ liệu cho các thuật toán lọc cộng tác thường không vi phạm các giả định trên [6]

Khoảng cách Pearson giữa hai người dùng u_i *và* u_k được tính như sau:

$$sim_{ik} = corr_{ik} = \frac{\sum_{j=1}^{l}(r_{i,j} - \overline{r_i})(r_{k,j} - \overline{r_k})}{\sqrt{\sum_{j=1}^{l}(r_{i,j} - \overline{r_i})^2 \sum_{j=1}^{l}(r_{k,j} - \overline{r_k})^2}}$$

Trong đó l là lực lượng của tập các sản phẩm mà cả hai người dùng u_i và u_k đều đã có đánh giá. Các tổng đều chạy trên tập sản phẩm này.

- ✓ $\bar{r_i}$, $\bar{r_k}$ là đánh giá trung bình của người dùng u_i và u_k.
- ✓ $\bar{r_i} = \frac{1}{|I_{u_i}|} \sum_{j \in I_{u_i}} r_{i,j}$, với I_{u_i} là tập các sản phẩm mà người dùng u_i đã đánh giá.
- ✓ Giá trị của khoảng cách Pearson nằm trong đoạn [-1,1].

Khoảng cách Jaccard

Có thể lờ đi các giá trị trong ma trận và chỉ tập trung vào tập các giá trị đã được đánh giá. Nếu ma trận khả dụng chỉ phản ánh việc mua hàng thì độ đo này sẽ là 1 sự lựa chọn tốt. Tuy nhiên, khi độ khả dụng là các đánh giá chi tiết hơn thì khoảng cách Jaccard mất đi thông tin quan trọng.

VD: A và B có giao điểm cỡ 1 và hợp của cỡ 5. Do vậy, sự giống nhau Jaccard là 1/5, và khoảng cách Jaccard là 4/5; tức là, chúng rất cách xa nhau. So sánh, A và C có sự giống nhau Jaccard 2/4, vì vậy khoảng cách Jaccard là giống nhau 1/2. Vì thế, A gần C hơn là gần B. Kết luận đó dường như sai về mặt trực giác. A và C không thích 2 bộ phim mà họ xem, trong khi A và B dường như đều thích 1 bộ phim giống nhau mà họ đã xem.

Khoảng cách Cosine

Có thể coi khoảng trống có giá trị 0. Sự lựa chọn này là đáng ngờ bởi vì nó coi việc thiếu đánh giá giống với việc không thích bộ phim hơn là việc thích bộ phim.

VD: Cosine của góc giữa A và B là

$$\frac{4 \times 5}{\sqrt{4^2 + 5^2 + 1^2}\sqrt{5^2 + 5^2 + 4^2}} = 0.380$$

Cosine của góc giữa A và C là:

$$\frac{5 \times 2 + 1 \times 4}{\sqrt{4^2 + 5^2 + 1^2}\sqrt{2^2 + 4^2 + 5^2}} = 0.322$$

Do cosine lớn hơn (dương) ngụ ý góc nhỏ hơn và do đó khoảng cách nhỏ hơn nên độ đo này cho ta biết rằng A gần với B hơn là với C.

Làm tròn số liệu

Cố gắng loại bỏ sự giống nhau giữa các bộ phim mà người dùng đánh giá cao và những bộ phim có các đánh giá thấp bằng cách làm tròn các đánh giá. VD: ta có thể xem xét các đánh giá 3, 4 và 5 là "1" và xem xét các đánh giá 1 và 2 là không đánh giá. Ma trận khả dụng sẽ trông giống như trong hình 2.3. Bây giờ, khoảng cách Jaccard giữa A và B là 3/4, trong khi giữa A và C là 1; tức là, C có vẻ xa A hơn so với B, điều này đúng về mặt trực giác. Áp dụng khoảng cách cosine vào hình 2.3 cho phép ta đưa ra kết luận tương tự

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	1			1			
B	1	1	1				
C					1	1	
D		1					1

Hình 2.3. Độ khả dụng 3, 4 và 5 được thay thế bằng 1, trong khi các đánh giá 1 và 2 bị loại bỏ

Chuẩn hóa đánh giá

Nếu chuẩn hóa đánh giá bằng cách trừ mỗi đánh giá cho đánh giá trung bình của người dùng đó thì ta biến các đánh giá thấp thành các số âm (-) và các đánh giá cao thành các số dương (+). Nếu dùng khoảng cách cosine, thì thấy rằng người dùng với các quan điểm đối lập về bộ phim giống nhau mà họ đã xem sẽ có các vector theo hướng gần như ngược lại, và có thể được coi là càng cách xa càng tốt. Tuy nhiên, những người dùng với quan điểm tương tự về bộ phim họ đánh giá chung sẽ có 1 góc tương đối nhỏ giữa chúng.

VD: Hình 2.4 chỉ ra ma trận của hình 2.2 với tất cả các đánh giá được chuẩn hóa. Tác động thú vị là đánh giá của D biến mất 1 cách hiệu quả bởi vì

0 giống như trống khi khoảng cách cosine được tính toán. Lưu ý rằng D chỉ đưa ra đánh giá là 3 và không phân biệt giữa các bộ phim, do vậy có thể quan điểm của D không đáng xem xét.

Tính toán cosine của góc giữa A và B:

$$\frac{(2/3)\times(1/3)}{\sqrt{(2/3)^2+(5/3)^2+(-7/3)^2}\sqrt{(1/3)^2+(1/3)^2+(-2/3)^2}} = 0.092$$

	HP1	HP2	HP3	TW	SW1	SW2	SW3
A	$2/3$			$5/3$	$-7/3$		
B	$1/3$	$1/3$	$-2/3$				
C				$-5/3$	$1/3$	$4/3$	
D		0					0

Hình 2.4. Ma trận khả dụng được gợi ý trong hình 2.2

Cosine của góc giữa A và C là:

$$\frac{(5/3)\times(-5/3)+(-7/3)\times(1/3)}{\sqrt{(2/3)^2+(5/3)^2+(-7/3)^2}\sqrt{(-5/3)^2+(1/3)^2+(4/3)^2}} = -0.559$$

Lưu ý rằng theo độ đo này, A và C tách xa hơn là A và B, và không cặp nào là gần nhau. Cả 2 quan sát này có nghĩa về mặt trực giác, A và C không đồng ý về 2 bộ phim họ đánh giá, trong khi A và B đưa ra đánh giá tương tự cho 1 bộ phim họ đánh giá chung.

2.3.2 Tính đối ngẫu của sự tương đồng

Ma trận khả dụng cho biết về người dùng, hoặc các mặt hàng hoặc cả hai. Quan trọng để nhận ra rằng bất kỳ phương pháp nào được gợi ý trong phần 2.3.1 để tìm người dùng giống nhau có thể được dùng trên cột của ma trận khả dụng để tìm các mặt hàng giống nhau. Có hai cách tính đối xứng bị phá vỡ trong thực tế

1. Có thể sử dụng thông tin về người dùng để gợi ý các mặt hàng. Có nghĩa, đưa ra một người dùng, có thể tìm 1 vài người dùng giống nhau nhất, có thể dựa vào khuyến nghị trên các quyết định tạo ra bởi những người dùng

giống nhau, ví dụ, gợi ý các mặt hàng mà có số lượng người lớn nhất mua hoặc đánh giá cao. Tuy nhiên, không có tính đối xứng. Thậm chí nếu tìm thấy các cặp mặt hàng giống nhau, thì vẫn cần thêm 1 bước nữa để khuyến nghị các mặt hàng cho người dùng.

2. Có sự khác nhau về hành vi điển hình của người dùng và các mặt hàng, do nó gắn liền với sự giống nhau. Về mặt trực giác, các mặt hàng có xu hướng có thể phân loại theo các thuật ngữ đơn giản. Ví dụ, âm nhạc có xu hướng thuộc về 1 thể loại nhất định. Tức là 1 đoạn nhạc không thể vừa là nhạc rock của năm 60 vừa là phong cách baroque của năm 1700. Mặt khác, có những cá nhân vừa thích nhạc rock năm 60 vừa thích nhạc baroque năm 1700, những người này mua các tác phẩm điển hình của 2 thể loại nhạc này. Kết quả sẽ dễ dàng hơn khi khám phá các mặt hàng giống nhau bởi chúng thuộc về thể loại giống nhau hơn là khám phá ra 2 người dùng giống nhau bởi họ thích chung 1 thể loại nhạc trong khi mỗi người cũng thích 1 vài thể loại mà người kia không thích.

Như đã gợi ý ở trên (1), có 1 cách để dự đoán giá trị của đầu vào ma trận khả dụng cho người dùng U và mặt hàng I là tìm ra n người dùng (với n được xác định trước) giống U nhất và trung bình đánh giá của họ cho mặt hàng I, chỉ tính toán trong n người dùng đánh giá về I. Nhìn chung thì chuẩn hóa ma trận đầu tiên sẽ tốt hơn. Có nghĩa, với mỗi người trong n người dùng, đánh giá cho I trừ đi đánh giá trung bình cho các mặt hàng. Trung bình sự khác nhau giữa những người dùng có đánh giá về I, sau đó cộng trung bình này với trung bình đánh giá mà U đánh giá tất cả các mặt hàng. Sự điều chỉnh các dự đoán trong trường hợp U có xu hướng đưa ra các đánh giá rất cao hoặc rất thấp, hoặc phần lớn những người dùng giống nhau đánh giá I (trong số đó có thể chỉ có vài người) là những người có xu hướng đánh giá rất cao hoặc rất thấp.

Có thể sử dụng sự giống nhau của mặt hàng để dự đoán đầu vào cho người dùng U và mặt hàng I. Tìm ra m mặt hàng giống I nhất, với 1 vài m, và tính đánh giá trung bình giữa m mặt hàng trong số các đánh giá mà U đưa ra. Về sự giống nhau giữa người dùng – người dùng, ta chỉ xét các mặt hàng trong số m mặt hàng mà U đã đánh giá và sẽ là khôn ngoan để chuẩn hóa các đánh giá mặt hàng trước.

Lưu ý là bất kể sử dụng phương pháp nào trong việc dự đoán phần tử trong ma trận khả dụng thì việc chỉ tìm 1 phần tử là không đủ. Để khuyến nghị các mặt hàng cho người dùng U, ta cần dự đoán mọi phần tử trong hàng của ma trận khả dụng với U, hoặc ít nhất tìm ra tất cả hoặc hầu hết các phần tử trong dãy trống nhưng có giá trị được dự đoán cao. Có 1 sự cân nhắc liệu có nên bắt đầu từ người dùng giống nhau hay các mặt hàng giống nhau.

Nếu tìm được những người dùng giống nhau, thì chỉ phải làm quá trình này 1 lần cho người dùng U. Từ tập những người dùng tương tự có thể dự đoán tất cả các khoảng trống trong ma trận khả dụng cho U. Nếu bắt đầu từ các mặt hàng giống nhau, thì phải tính toán các mặt hàng giống nhau cho gần như tất cả các mặt hàng, trước khi dự đoán hàng cho U.

Mặt khác, sự giống nhau giữa mặt hàng – mặt hàng thường cung cấp nhiều thông tin đáng tin cậy hơn, do hiện tượng được quan sát ở trên, nên sẽ dễ hơn khi tìm các mặt hàng cùng thể loại hơn là tìm người dùng chỉ thích các mặt hàng của 1 thể loại nhất định. [5]

Bất kể chọn phương pháp nào, nên tính toán trước các mặt hàng mà mỗi người dùng ưa thích, hơn là đợi cho đến khi ta cần đưa ra quyết định. Do ma trận khả dụng tiến triển rất chậm, nên nhìn chung chỉ cần tính toán không thường xuyên và giả định rằng nó vẫn cố định giữa các lần tính toán lại.

2.3.3 Phân cụm những người dùng và các mặt hàng

Rất khó để có thể phát hiện ra sự giống nhau giữa mặt hàng hoặc là người dùng, vì có rất ít thông tin về cặp người dùng – mặt hàng trong ma trận khả dụng thưa. Trong phần 2.3.2 nếu 2 mặt hàng cùng thể loại có khả năng rất ít người dùng mua hoặc đánh giá cả hai.

Mặc dù 2 sản phẩm cùng thể loại, nhưng có khả năng rất ít người dùng mua hoặc đánh giá cả hai mặt hàng. Giống như vậy, mặc dù cả 2 người dùng thích 1 hay nhiều thể loại, nhưng họ có thể không cùng mua chung 1 mặt hàng nào cả.

Một cách để giải quyết khó khăn này là phân cụm các mặt hàng và/ hoặc người dùng. Chọn bất kỳ độ đo khoảng cách nào khác, sử dụng nó để thực hiện việc phân cụm các mặt hàng. Tuy nhiên, có ít lý do để cố gắng phân cụm thành số lượng nhỏ các cụm. Hơn nữa, 1 hướng tiếp cận phân cấp, bước đầu tiên ta để nhiều cụm chưa kết hợp. Ví dụ, có thể để lại 1 nửa các cụm so với các mặt hàng.

	HP	TW	SW
A	4	5	1
B	4.67		
C		2	4.5
D	3		3

Hình 2.5. Ma trận khả dụng cho người dùng và cụm các mặt hàng

VD: Hình 2.5 cho thấy điều gì xảy ra đối với ma trận khả dụng của hình 2.2 nếu cụm 3 bộ phim Harry-Potter vào 1 cụm, ký hiệu là HP, và cũng cụm 3 bộ phim Star-Wars vào 1 cụm SW.

Có các mặt hàng được phân cụm các hàng hóa theo mức độ, có thể xem lại ma trận khả dụng sao cho các cột đại diện cho các cụm của các mặt hàng, phần tử của người dùng U và cụm C là đánh giá trung bình mà U cho các

thành viên của cụm C mà U đã đánh giá. Lưu ý U có thể không đánh giá thành viên của cụm các thành viên, trong trường hợp đầu vào cho C và U vẫn trống.

Có thể sử dụng ma trận khả dụng để phân cụm những người dùng sử dụng lại độ đo khoảng cách cho là phù hợp nhất. Sử dụng thuật toán phân cụm mà loại bỏ nhiều cụm, ví dụ một nửa cụm so với số người dùng. Xem lại ma trận khả dụng, các hàng tương ứng với cụm người dùng, cột tương đương với cụm các mặt hàng. Về cụm các mặt hàng, tính toán đầu vào cho cụm người dùng bằng cách tính trung bình đánh giá của những người dùng trong cụm đó.

Bây giờ, nếu muốn ta có thể lặp lại quá trình này vài lần. Có nghĩa, là có thể phân cụm các cụm mặt hàng và kết hợp các cột của ma trận khả dụng mà thuộc về 1 cụm 1 lần nữa. Sau đó chuyển sang người dùng lần nữa và phân cụm các cụm người dùng. Quá trình có thể lặp lại cho đến khi có số lượng cụm hợp lý cho mỗi loại về mặt trực giác.

Một khi phân cụm những người dùng và/hoặc các mặt hàng tới một cấp độ mong muốn và một ma trận khả dụng cụm – cụm đã được tính toán thì có thể dự đoán các phần tử đầu vào trong ma trận khả dụng ban đầu như sau. Giả sử muốn dự đoán đầu vào cho người dùng U và mặt hàng I:

(a) Tìm các cụm mà U và I thuộc về cụm đó, chẳng hạn lần lượt là cụm C và D.

(b) Nếu đầu vào trong ma trận khả dụng cụm – cụm cho C và D là 1 thứ gì đó chứ không phải trống, thì sử dụng giá trị này như là giá trị được dự đoán cho đầu vào U – I trong ma trận khả dụng ban đầu.

(c) Nếu đầu vào cho C–D là trống, thì sử dụng phương pháp được chỉ ra trong mục 2.3.2 để dự đoán cụm đó bằng cách cân nhắc các cụm tương tự C hoặc D. Sử dụng dự đoán kết quả như dự đoán cho đầu vào U – I.

2.4 Kết luận chương 2

Chương 2 tập trung nghiên cứ các kỹ thuật khai phá dữ liệu . Các kỹ thuật khai phá dữ liệu thường được chia làm hai nhóm chính đó là k ỹ thuật khai phá dữ liệu mô tả và kỹ thuật khai phá dữ liệu dự đoán. Từ đó xây dựng hệ thống khuyến nghị dựa trên nội dung và phương pháp lọc cộng tác để phân cụm người dùng và các mặt hàng trong hệ thống.

CHƯƠNG 3: ỨNG DỤNG TRIỂN KHAI THỬ NGHIỆM HỆ THỐNG TƯ VẤN CHỌN PHIM

3.1 Bài toán

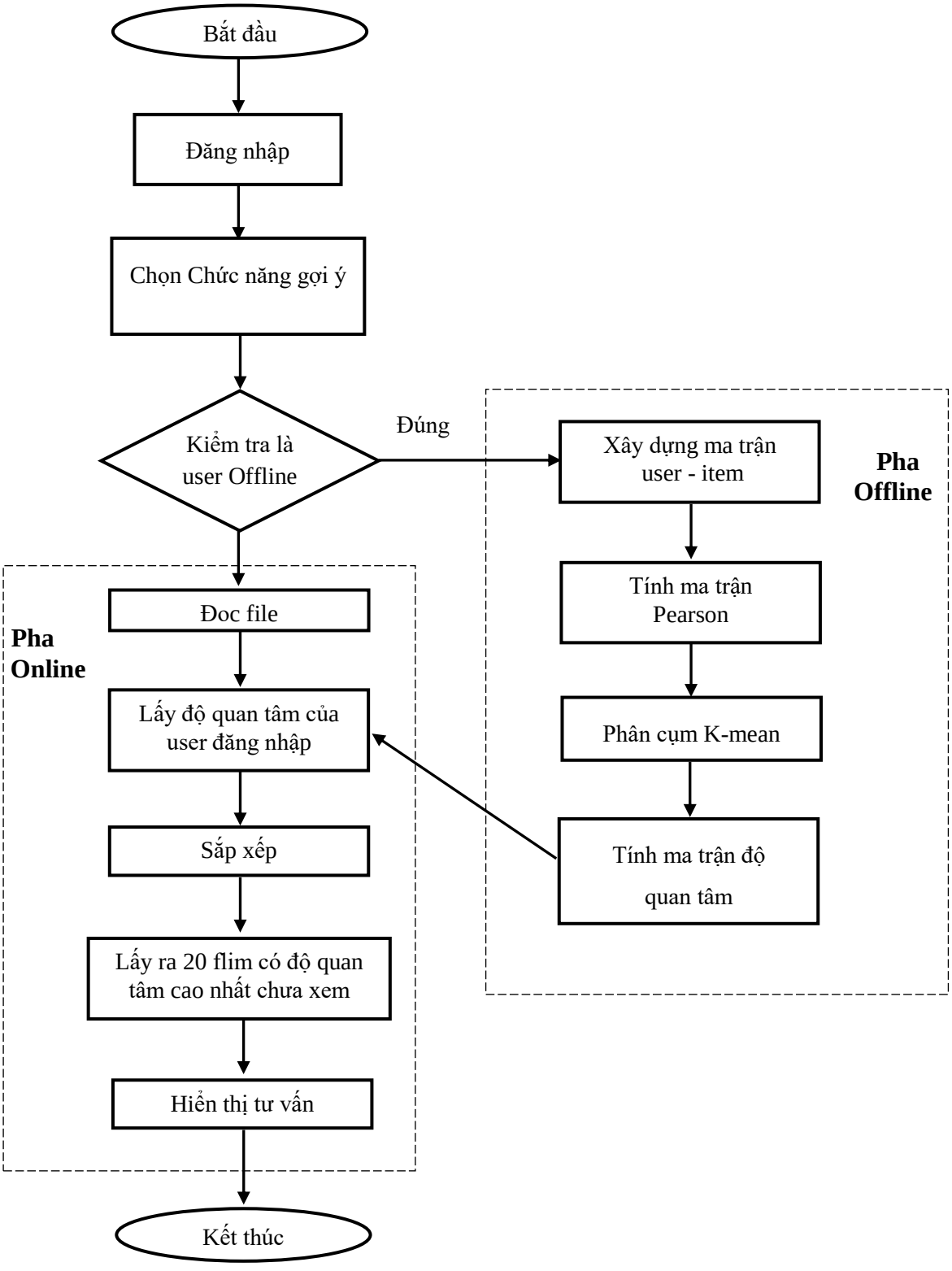
Với lượng thông tin về phim khổng lồ như hiện nay, việc người dùng phải tự mình thử, tìm kiếm các phim trên mạng để biết thông tin về những phim mình nên xem quả là một vấn đề khó và không thực sự hiệu quả, tốn kém thời gian, công sức. Trước khi chọn bộ phim nào đó để xem, họ sẽ cố tìm hiểu về bộ phim đó bằng cách hỏi những người khác hoặc đọc các bài giới thiệu trước khi chọn bộ phim đó để xem, để giảm sự quá tải thông tin đó, hệ tư vấn là một hướng tiếp cận đang ngày càng được ứng dụng rộng rãi không chỉ cho phim mà còn giải quyết được nhiều vấn đề của "quá tải thông tin" nói chung

Đặc biệt là ở Việt Nam, các website gần như chỉ cung cấp nội dung phim chứ không có chức năng tư vấn ví dụ phim3s.net, bomtan.org, …

Từ nhu cầu đó, chương 3 của luận văn sẽ tập trung xây dựng chức năng tư vấn phim cho người dùng. Mục tiêu là xây dựng một module tư vấn trên website. Có khả năng dựa vào các thông tin đánh giá của người dùng trên các phim đã xem, từ đó đưa ra gợi ý các phim có khả năng sẽ phù hợp sở thích.

Dự kiến module này sẽ sử dụng thuật toán phân cụm K-mean và phương pháp lọc cộng tác giải quyết bài toán trên.

Biểu đồ hoạt động:



Hình 3.1. Biểu đồ hoạt động hệ thống tư vấn phim

3.2 Xây dựng hệ tư vấn phim

3.2.1 Chuẩn bị dữ liệu

Danh sách phim, người dùng và đánh giá

Hệ thống tư vấn phim là một Website thông tin về phim, để chức năng hoạt động tư vấn có thể thực hiện được hệ thống phải có dữ liệu về một số lượng phim và người dùng ban đầu lớn, giải pháp cho vấn đề này là sử dụng nguồn dữ liệu sẵn có:

Danh sách phim, người dùng và đánh giá được lấy từ tập dữ liệu của MovieLens, các đánh giá từ 1 đến 5. Đối với phim, các tập dữ liệu này chỉ cung cấp thông tin về tên, năm sản xuất, thể loại. Còn đối với người dùng chỉ có thông tin về ID, giới tính, độ tuổi và mã vùng (Zipcode)

✓Tập dữ liệu MovieLens có 6040 người dùng và 3883 phim, số đánh giá đã có là 1.000.209 trong đó mỗi người dùng đánh giá ít nhất 20 phim.

✓Tại đây, MovieLens còn có một tập dữ liệu thu gọn của tập dữ liệu trên (706 người dùng, 3882 phim, 71398 đánh giá) có thể được sử dụng thử nghiệm các thuật toán trước khi thực hiện trên các tập dữ liệu lớn.

Tập dữ liệu trên của MovieLens có thể tải tại địa chỉ. [9]

http://www.cs.umn.edu/research/grouplens/data.

Cấu trúc dữ liệu trong tập dữ liệu của MovieLens như sau:

✓*Dữ liệu đánh giá*: được lưu trong file văn bản "ratings.dat"

File này gồm nhiều dòng, mỗi dòng ứng với một đánh giá và có định dạng:

UserID::MovieID::Rating::Timestamp

Ví dụ: **1::3408::4::978300275**

- **UserID**: Nằm trong khoảng [1,6040]
- **MovieID**: Nằm trong khoảng [0,3952]
- **Rating**: Nằm trong khoảng [1,5]

- **Timestamp**: Nhãn thời gian tại thời điểm đánh giá. Trường này không được sử dụng trong tư vấn phim.
- Mỗi người dùng có ít nhất 20 đánh giá

✓*Dữ liệu người dùng*: được lưu trong file văn bản "users.dat"

File này gồm nhiều dòng, mỗi dòng ứng với một người dùng và có định dạng:

UserID::Gender::Age::Occupation::Zip-code

Ví dụ: **1::F::1::10::48067**

- **UserID**: Mã người dùng.
- **Gender**: Giới tính người dùng, thể hiện bằng chữ cái. "M" là nam, "F" là nữ.
- **Age**: Độ tuổi người dùng, nhận giá trị trong tập {1,18,24,35,45,50,56}. Ý nghĩa được tham chiếu trong bảng dưới

Bảng tham chiếu tuổi trong tập dữ liệu MovieLens

Giá trị trường	Ý nghĩa
1	Dưới 18 tuổi
18	18-24 tuổi
25	25-34 tuổi
35	35-44 tuổi
45	45-49 tuổi
50	50-55 tuổi
56	Từ 56 tuổi trở lên

- **Occupation**: Nghề nghiệp người dùng, nhận giá trị trong khoảng [0,20], ý nghĩa được tham chiếu trong bảng dưới.

Bảng tham chiếu nghề nghiệp trong tập dữ liệu MovieLens

Giá trị trường	Ý nghĩa
0	Không xác định tuổi
1	Nhà nghiên cứu/Giáo viên
2	Nghệ sĩ/hoạ sĩ
3	Nhân viên văn phòng
4	Sinh viên đã tốt nghiệp
5	Dịch vụ khách hàng
6	Bác sĩ/Chăm sóc sức khoẻ
7	Quản lí/Giám đốc
8	Nông dân
9	Nội trợ
10	Sinh viên
11	Luật sư
12	Lập trình viên
13	Nghỉ hưu
14	Nhân viên bán hàng/tiếp thị
15	Nhà khoa học
16	Kinh doanh tư nhân
17	Kĩ thuật viên / Kĩ sư
18	Thợ thủ công
19	Thất nghiệp
20	Nhà văn

Zip-code: Mã khu vực người dùng sống. Trường này không được sử dụng trong tư vấn phim.

✓Dữ liệu về phim: được lưu trong file "movies.dat".

File này gồm nhiều dòng, mỗi phim được lưu trong một dòng theo định dạng sau:

MovieID::Title::Genres

Ví dụ: **12::Dracula: Dead and Loving It (1995)::Comedy|Horror**

- **MovieID**: Mã của phim.
- **Title**: Tên của phim (bao gồm cả năm sản xuất).
- **Genres**: Thể loại phim. Các dữ liệu trên sẽ được lưu vào cơ sở dữ liệu của tư vấn phim.

Thông tin chi tiết về phim

Thông tin chi tiết về phim được tách từ các trang Web của Website IMDB: us.imdb.com. IMDB cung cấp đường dẫn để có thể truy xuất trang thông tin về một phim với tên và năm sản xuất đã xác định như sau:

http://us.imdb.com/M/title-exact?**Moviename**+(**Year**)

- Moviename là tên của phim đã được mã hoá theo chuẩn URL
- Year là năm sản xuất của phim. [10]

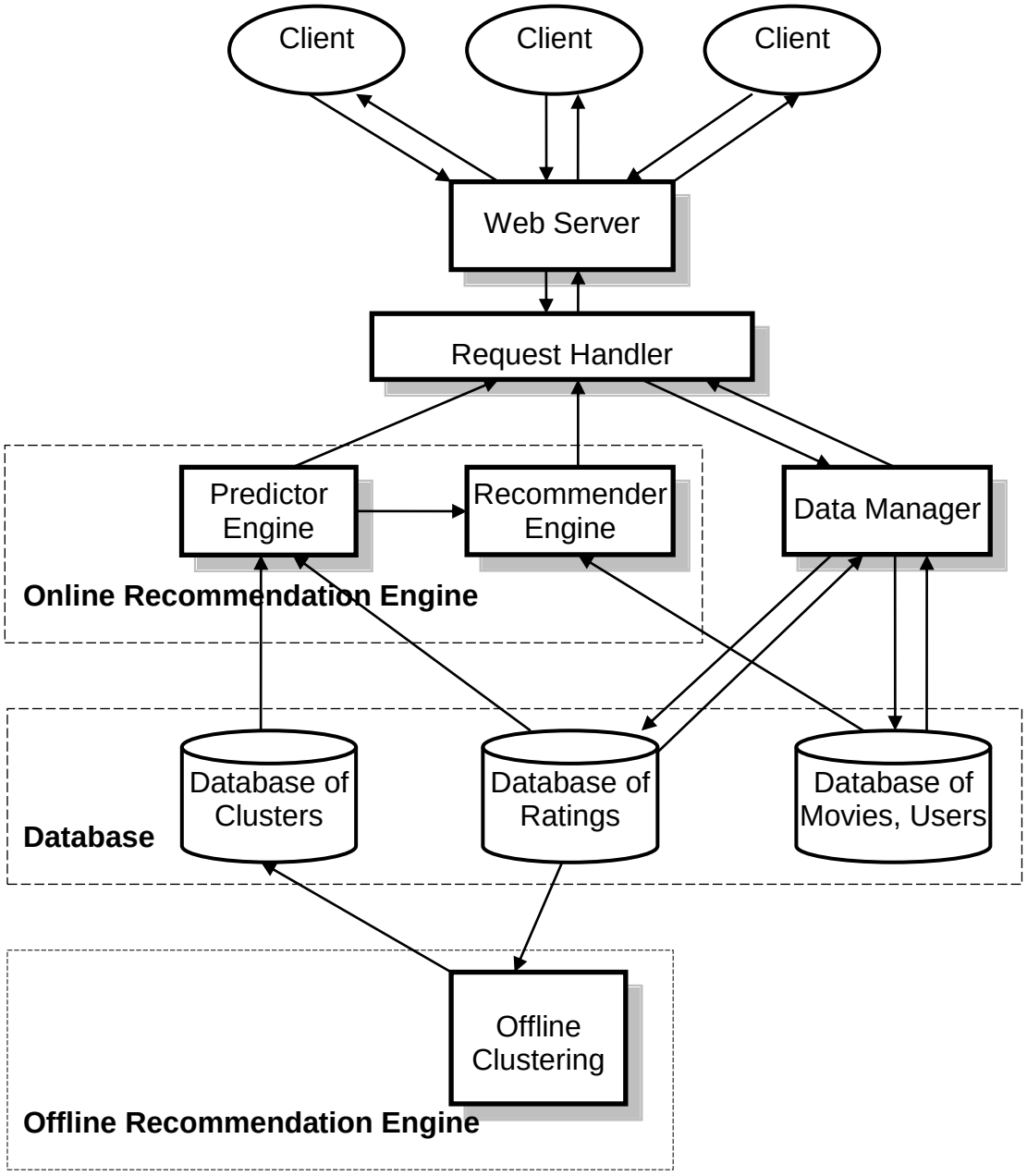
Chẳng hạn ta cần tìm thông tin chi tiết về phim "Toy Story" được sản xuất vào năm 1995 bằng cách vào trang: [11]

http://us.imdb.com/M/title-exact?Toy+Story+(1995)

3.2.3 Thiết kế hệ thống

3.2.3.1 Kiến trúc tổng quan của hệ thống

Hệ thống tư vấn phim có kiến trúc như sau:



Hình 3.2. Kiến trúc hệ tư vấn phim

Offline Clustering

✓ Đây là khối chức năng thực hiện phân cụm tập người dùng dựa trên dữ liệu đánh giá của người dùng trong hệ thống tại thời điểm tiến hành phân cụm.

✓ Kết quả của thực hiện phân cụm là dữ liệu về các cụm người dùng lưu trong cơ sở dữ liệu hoạt động của hệ thống Website.

✓ Chức năng phân cụm chỉ cần thực hiện một lần. Khi dữ liệu thay đổi nhiều, quản trị viên có thể thực hiện lại chức năng này.

Predictor Engine

✓ Đây là khối chức năng thực hiện đánh giá dự đoán của một người dùng đối với một phim. Đầu vào là cặp (người dùng, phim), đầu ra là dự đoán đánh giá của hệ thống

✓ Khối này được sử dụng trong hai trường hợp

- Người dùng xem thông tin về phim, hệ thống cung cấp thông tin về dự đoán đánh giá đối với những phim đang xem
- Khối chức năng "Recommender Engine" sinh ra tập gợi ý cho người dùng dựa vào dự đoán đánh giá của người dùng đối với những phim chưa xem.

Recommender Engine

✓ Đây là khối chức năng thực hiện sinh ra tập gợi ý cho người dùng hiện thời. Đầu vào là người dùng, đầu ra là tập phim gợi ý tương ứng với người dùng đó.

✓ Khối chức năng thực hiện dựa trên việc dự đoán đánh giá của người dùng hiện thời đối với từng phim chưa xem thông qua việc triệu gọi chức năng "Predictor Engine". Những phim được dự đoán cao nhất sẽ được gợi ý cho người dùng.

Data Manager

Khối chức năng Data Manager thực hiện việc vào ra dữ liệu trong quá trình giao tiếp với người dùng bao gồm truy xuất thông tin phim, thông tin

người dùng, cập nhật đánh giá, đăng kí người dùng mới, quản lý người dùng, quản lý phim của hệ thống.

Request Handler

Khối chức năng Request Handler thực hiện vai trò trung gian giữa người dùng với hệ thống.

✓ Nhận các yêu cầu của người dùng và thực hiện chức năng tương ứng

✓ Sinh ra các trang Web động hiển thị kết quả trả về cho người dùng

Database

Cơ sở dữ liệu để hệ thống hoạt động bao gồm cơ sở dữ liệu về người dùng, về phim, cơ sở dữ liệu đánh giá của người dùng đối với phim và cơ sở dữ liệu về các phân cụm phục vụ cho chức năng tư vấn.

3.2.2 Lựa chọn giải pháp

Thuật toán

Hệ thống tư vấn phim sử dụng phương pháp phân cụm trên tập người dùng nhằm nhóm những người dùng có sở thích giống nhau vào cùng một nhóm. Hoạt động của hệ thống gồm hai pha:

✓ Pha offline: Tiến hành phân cụm người dùng

✓ Pha online: Đưa ra tư vấn và dự đoán đánh giá cho người dùng dựa vào việc phân cụm đã được thực hiện ở pha offline

Độ tương đồng

Độ tương đồng được lựa chọn sử dụng trong hệ thống là độ đo khoảng cách Pearson:

$$sim_{ik} = corr_{ik} = \frac{\sum_{j=1}^{l}(r_{ij} - \overline{r_i})(r_{kj} - \overline{r_k})}{\sqrt{\sum_{j=1}^{l}(r_{ij} - \overline{r_i})^2 \sum_{j=1}^{l}(r_{kj} - \overline{r_k})^2}}$$

Trong đó

✓ sim_{ik} là độ tương đồng giữa hai người dùng u_i và u_k

- ✓ l là số phim mà cả u_i và u_k đều đã có đánh giá.
- ✓ $\overline{r_i}$, $\overline{r_k}$ là đánh giá trung bình của người dùng u_i và u_k.
- ✓ $\overline{r_i} = \frac{1}{|I_{u_i}|} \sum_{j \in I_{u_i}} r_{i,j}$, với I_{u_i} là tập các phim mà người dùng u_i đã đánh giá.

Thuật toán phân cụm

Thuật toán phân cụm được lựa chọn ở đây dựa trên ý tưởng của thuật toán *K-means* có thay đổi.

Đầu vào: tập N người dùng

Đầu ra: k cụm ổn định khác nhau

Thuật toán *K-means* được mô tả như sau:

✓Giả sử cần phân tập người dùng vào k cụm khác nhau.

✓Chọn ra k người dùng bất kì coi như các tâm cụm.

✓Thực hiện lặp

- Gán người dùng vào cụm mà độ tương đồng của người dùng với tâm cụm là lớn nhất.
- Tính toán lại tâm của các cụm. Tâm cụm được xác định là phần tử của cụm có trung bình toàn phương của các độ tương đồng với các phần tử còn lại trong cụm là lớn nhất.
- Quá trình lặp dừng khi các cụm ổn định (không có người dùng nào chuyển từ cụm này sang cụm khác) hoặc số lần lặp vượt quá giới hạn tối đa.

Trong thuật toán *K-means* chuẩn ta cần gán phần tử vào cụm nào mà phần tử có khoảng cách tới tâm cụm nhỏ nhất. Ở đây, hai người dùng càng giống nhau khi độ tương đồng càng lớn nên người dùng sẽ được gán vào cụm nào mà tâm cụm có độ tương đồng với người dùng lớn nhất. Một điểm khác nữa là với *K-means*, tâm cụm được xác định bằng trọng tâm của các phần tử trong cụm, tức là tâm cụm là điểm có trung bình toàn phương của các khoảng cách tới tất cả các phần tử của cụm nhỏ nhất. Còn với khoảng cách Pearson, do hai người dùng càng giống nhau nếu độ

tương đồng càng lớn nên phải có cách khác để xác định tâm. Ở đây, tôi chọn phương pháp lấy tâm là phần tử trong nhóm có trung bình toàn phương độ tương đồng đến các phần tử khác lớn nhất.

Dự đoán đánh giá của người dùng

Sau khi phân cụm thì công việc dự đoán đánh giá của người dùng khá nhẹ nhàng.

Sử dụng cụm như tập láng giềng

Theo cách này có thể coi tất cả các người dùng trong cụm là một tập láng giềng của từng người dùng trong cụm đó. Khi tính toán dự đoán cho một người dùng ta cần xác định cụm tương tự nhất với anh ta, cụ thể là: người dùng sẽ thuộc về cụm nào mà tâm cụm có độ tương tự với người dùng đó là lớn nhất. Sau đó sử dụng thuật toán lọc cộng tác cơ bản với tập người dùng đầu vào ban đầu là cả cụm đã chọn. Nghĩa là tập láng giềng được chọn ra để tính toán dự đoán cho người dùng sẽ là tập con của cụm này.

Như vậy, phương pháp trên nhằm mục đích xác định tập láng giềng của người dùng hiện thời như trong thuật toán dựa trên bộ nhớ truyền thống. Sau khi xác định được tập láng giềng này, dự đoán đánh giá của người dùng hiện thời u_a đối với phim i_j được xác định như sau:

$$pr_{aj} = \overline{r_a} + \frac{\sum_{i=1}^{l} (r_{ij} - \overline{r_i}) * sim_{ai}}{\sum_{i=1}^{l} |sim_{ai}|}$$

Trong đó:

- ✓ pr_{aj} là dự đoán cho đánh giá của người dùng u_a đối với phim i_j
- ✓ sim_{ai} là độ tương đồng giữa người dùng u_a và u_i
- ✓ l là số người dùng trong tập láng giềng của người dùng u_a và đã đánh giá i_j
- ✓ $\overline{r_a}$ là đánh giá trung bình hiện thời của người dùng u_a

Đưa ra gợi ý cho người dùng

Khi người dùng yêu cầu danh sách phim gợi ý, hệ thống sẽ thực hiện các bước:

- ✓ Xây dựng tập phim là hợp của các phim mà các láng giềng của người dùng hiện thời đã đánh giá nhưng người dùng hiện thời lại chưa đánh giá.
- ✓ Tính dự đoán cho tập phim đã chọn ra. Lấy ra những phim có độ dự đoán đánh giá cao nhất (top-N) và giới thiệu cho người dùng.

3.2.3.2 Công cụ và môi trường phát triển

Ngôn ngữ lập trình

Sau khi đã xác định được kiến trúc của hệ thống, ta cần cài đặt hệ thống bằng những ngôn ngữ lập trình cụ thể. Ngôn ngữ lập trình được lựa chọn để xây dựng hệ thống là ASP.NET

ASP.NET được lựa chọn trước tiên là do những ưu điểm sau:

- ✓ ASP.NET cho phép ta lựa chọn một trong các ngôn ngữ lập trình mà ta quen thuộc: Visual Basic.Net, J#, C#,… Biên dịch những trang web động thành những tập tin DLL mà Server có thể thi hành nhanh chóng và hiệu quả.
- ✓ ASP.NET hỗ trợ mạnh mẽ bộ thư viện phong phú và đa dạng của .Net Framework, làm việc với XML, Web Service, truy cập cơ sở dữ liệu qua ADO.Net,…
- ✓ ASP.NET sử dụng phong cách lập trình mới: Code behide. Tách code riêng, giao diện riêng điều này giúp ta dễ quản lý và bảo trì chương trình.
- ✓ Kiến trúc lập trình giống ứng dụng trên Windows.
- ✓ Tự động phát sinh mã HTML cho các Server control tương ứng với từng loại Browser.

ASP.NET giúp chúng ta phát triển và triển khai các ứng dụng về mạng trong một thời gian kỷ lục vì nó cung cấp cho ta một kiểu mẫu lập trình dễ dàng và gọn gàng nhất. Ngoài ra, các trang ASP.NET còn làm việc với mọi browsers hiện nay như Internet Explorer (IE), FireFox, Chrome, Netscape, Opera, AOL,… mà không cần phải thay đổi lại các nguồn mã.

Với tất cả những ưu điểm trên, ngôn ngữ ASP.NET được lựa chọn để cài đặt tư vấn phim. Phiên bản ASP.NET hệ thống hỗ trợ là phiên bản Visual Studio 2010

Hệ quản trị cơ sở dữ liệu

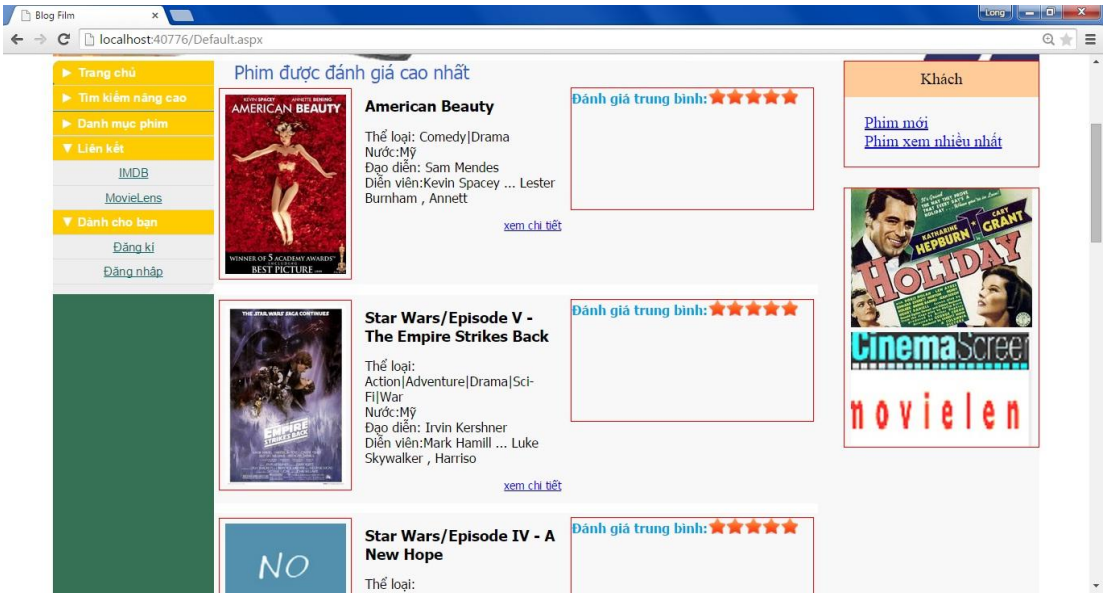
Hệ quản trị cơ sở dữ liệu được lựa chọn là Microsoft SQL Server 2008. Ưu điểm của hệ quản trị cơ sở dữ liệu này là có thể nhập/xuất (import/export) rất dễ dàng giữa dữ liệu của hệ thống với các file văn bản. Tính năng này rất cần thiết cho tư vấn phim khi trao đổi dữ liệu với các thành phần bên ngoài lưu trữ dữ liệu ở dạng file văn bản.

SQL Server 2008 cùng với .NET Framework đã giảm được sự phức tạp trong việc phát triển các ứng dụng mới. Các mở rộng của ngôn ngữ truy vấn tích hợp (LINQ) mới trong .NET Framework đã cách mạng hóa cách các chuyên gia phát triển truy vấn dữ liệu bằng việc mở rộng Visual C#.NET và Visual Basic.NET để hỗ trợ cú pháp truy vấn giống SQL vốn đã có.

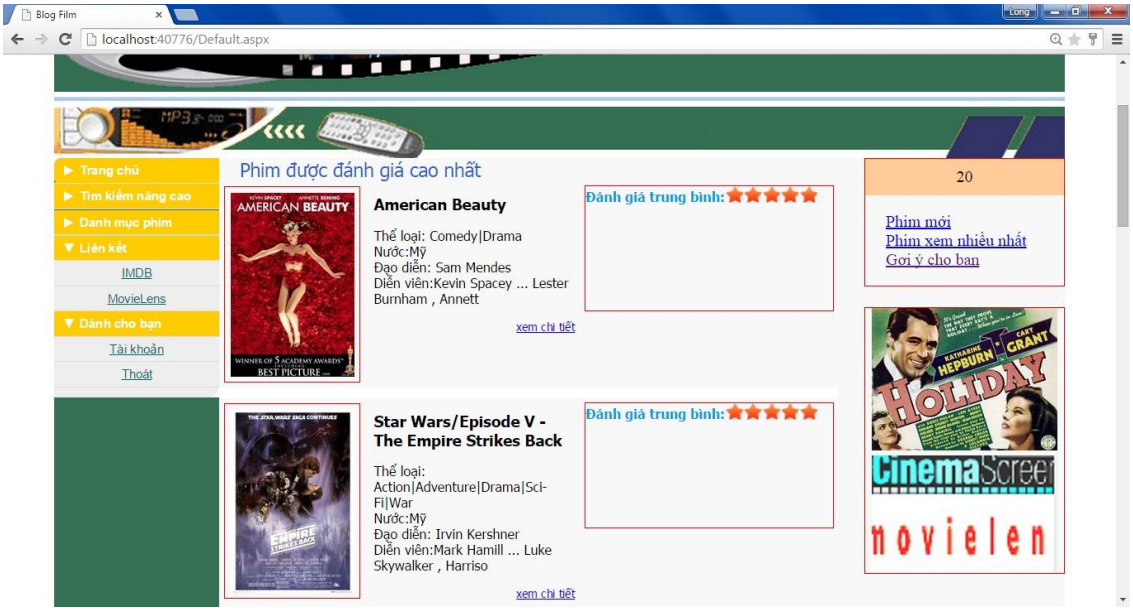
Kết quả

Giao diện chương trình:

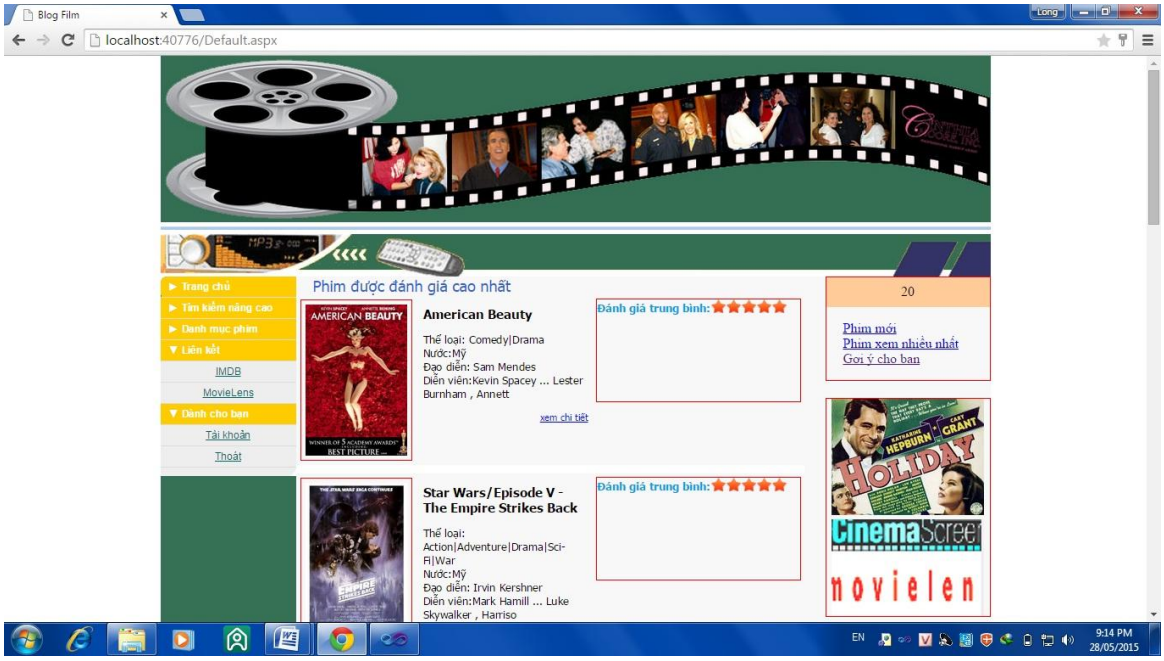
- Giao diện khi chưa đăng nhập: website mặc định hiển thị danh sách các phim được đánh giá cao nhất.



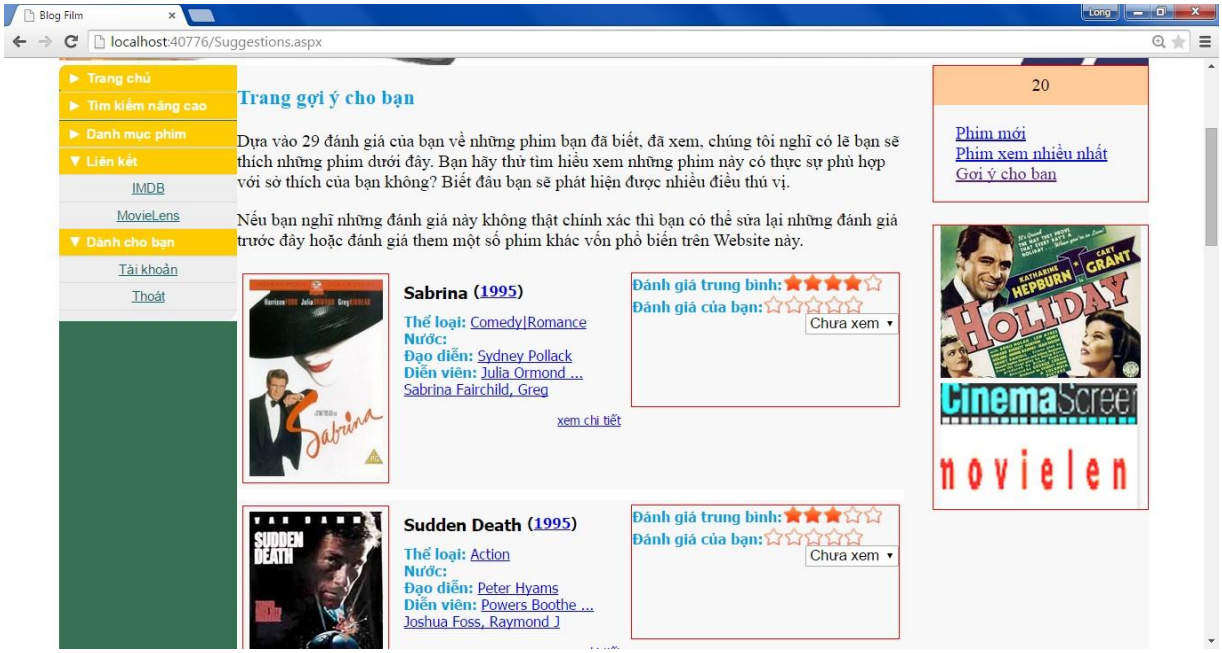
- Đăng nhập: ví dụ, Tên đăng nhập: 20; Mật khẩu: 1



- Sau khi đăng nhập vào tài khoản 20, website vẫn giữ giao diện mặc định và hiển thị thêm chức năng "Gợi ý cho bạn" ở menu bên phải:



- Chọn chức năng "Gợi ý cho bạn", module tư vấn phim sẽ tính toán và đưa ra gợi ý cho người dùng danh sách 20 phim chưa xem và có khả năng sẽ thích dựa trên độ tương đồng với người dùng khác.



3.3 Kết luận chương 3

Chương 3 xây dựng ứng dụng tư vấn chọn phim sử dụng hệ thống khuyến nghị khách hàng BI.

- Vai trò hệ thống BI là mang lại cho người dùng sự tiện lợi khi lựa chọn phim, bằng các gợi ý phù hợp, nhanh chóng, và giúp họ dễ dàng đưa ra quyết định xem phim khác phù hợp với họ.

- Vai trò Datamining là công cụ lõi, bằng các thuật toán khai thác từ các dữ liệu thu thập được từ người dùng để tìm ra những gợi ý phù hợp.

Như vậy Datamining là công cụ hỗ trợ BI thực hiện nhiệm vụ trợ giúp khách hàng ra quyết định.

Giải quyết được vấn đề

- Cung cấp cho người dùng tính năng có thể thể hiện quan điểm, đánh giá của mình về các bộ phim trong hệ thống.

- Áp dụng các thuật toán phân cụm, lọc công tác để trợ giúp tìm ra được những phim người dùng có thể thích.

Chưa giải quyết được

- Hệ thống hiện chưa giải quyết được vấn đề khi thêm phim mới và người dùng mới. Cách giải quyết hiện thời là giới thiệu phim mới trên trang chủ của người dùng không triệt để.

KẾT LUẬN VÀ KIẾN NGHỊ

Những vấn đề giải quyết được:

Tìm hiểu tổng quan về khái niệm, vai trò của Datamining trong hệ thống BI nói chung và hệ thống khuyến nghị nói riêng.

Đi sâu tìm hiểu phương pháp lọc cộng tác và một số kỹ thuật khai phá dữ liệu như K-Means, luật kết hợp, thuật toán Apriori.

Xây dựng website tư vấn phim dựa vào phương pháp lọc cộng tác kết hợp với phân cụm dữ liệu.

Những hạn chế của luận văn:

Do hạn chế về mặt thời gian nghiên cứu cũng như trình độ học thuật, nên một số vẫn đề được phân tích nghiên cứu trong luận văn chắc chắn vẫn còn những thiếu sót.

Luận văn mới chỉ đưa ra được tư vấn dựa trên những dữ liệu sẵn có, pha offline tính toán còn chậm. Dẫn đến chưa giải quyết được vấn đề khi thêm phim mới và người dùng mới. Cách giải quyết hiện thời là giới thiệu phim mới trên trang chủ của người dùng không triệt để.

Học viên hy vọng những hạn chế thiếu sót trong luận văn này sẽ được khắc phục trong các nghiên cứu sâu hơn. Rất mong nhận được ý kiến đóng góp từ quý thầy cô và đồng nghiệp.

Hướng phát triển

Triển khai và hoạt động thực sự trên Internet, việc cập nhập bổ sung thông tin mới cho hệ thống một cách tự động từ các trang chứa dữ liệu.

Tìm hiểu thêm các phương pháp /kỹ thuật để giải quyết được vấn đề khi thêm phim mới và người dùng mới.

TÀI LIỆU THAM KHẢO

Tiếng Anh:

[1] **Adomavicius and A. Tuzhilin,** *Towards the next generation of recommender systems: a survey of the state-of-the-art and possible exten- sions*, IEEE Trans. on Data and Knowledge Engineering 17:6, 734–749, 2005.

[2] **Anderson**, *The Long Tail:* Why the Future of Business is Selling Less of More, Hyperion Books, New York, 2006.

[3] **Barry de Ville**, *Microsoft ® Data Mining Integrated Business Intelligence for e-Commerce and Knowledge Management*,Digital Press,USA, 2001.

[4] **Koren**, *The BellKor solution to the Netflix grand prize*, 2009.

[5] **Linden, B. Smith, and J. York**, *Amazon.com recommendations: item-to-item collaborative filtering*," Internet Computing 7:1, pp. 76–80, 2003.

[6] **Li, Q. & Kim, B.M.** "*An approach for combining content-based and collaborative filters*", Korea Research Foundation Grant, KRF-2002-041-D00459, 2002.

[7] **M. Piotte and M. Chabbert**, *The Pragmatic Theory solution to the Net- flix grand prize*, 2009.

[8] **Ruchira Bhargava,Yogesh Kumar Jakhar,** *Knowledge Base Data Mining for Business Intelligence*", National Monthly Refereed Journal of Reasearch in Science & Technology,1(11), 1-5, 2003.

Website:

[9] http://www.cs.umn.edu/research/grouplens/data.

[10] http://us.imdb.com/M/title-exact?**Movienam**e+(**Year**)

[11] http://us.imdb.com/M/title-exact?Toy+Story+(1995)