ĐẠI HỌC THÁI NGUYÊN

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

MÔNG QUỐC TUẤN

NGHIÊN CỨU MÔ HÌNH NGƯỜI SỬ DỤNG MỞ TRONG CÁC HỆ THỐNG GỢI Ý THÔNG TIN THEO NHU CẦU

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

THÁI NGUYÊN - 2017

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

MÔNG QUỐC TUẤN

NGHIÊN CỨU MÔ HÌNH NGƯỜI SỬ DỤNG MỞ TRONG CÁC HỆ THỐNG GỢI Ý THÔNG TIN THEO NHU CẦU

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 01 01

LUẬN VĂN THẠC SĨ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: TS. NGUYỄN VIỆT ANH

THÁI NGUYÊN - 2017

LỜI CẢM ƠN

Luận văn này được hoàn thành tại Trường Đại học Công nghệ Thông tin và Truyền thông dưới sự hướng dẫn của TS. Nguyễn Việt Anh. Tác giả xin bày tỏ lòng biết ơn tới các thầy cô giáo thuộc Trường Đại học Công nghệ Thông tin và Truyền thông đã tạo điều kiện và giúp đỡ tác giả trong quá trình học tập và làm luận văn tại Trường, đặc biệt tác giả xin bày tỏ lòng biết ơn tới TS. Nguyễn Việt Anh đã tận tình hướng dẫn và cung cấp nhiều tài liệu cần thiết để tác giả có thể hoàn thành luận văn đúng thời hạn.

Xin chân thành cảm ơn anh chị em học viên cao học và bạn bè đồng nghiệp đã trao đổi, động viên và khích lệ tác giả trong quá trình học tập và làm luận văn tại Trường Đại học Công nghệ Thông tin và Truyền thông – Đại học Thái Nguyên.

Thái Nguyên, tháng 5 năm 2017

Học viên

Mông Quốc Tuấn

LỜI CAM ĐOAN

Tôi xin cam đoan luận văn này do chính tôi thực hiện, dưới sự hướng dẫn khoa học của TS. Nguyễn Việt Anh, các kết quả lý thuyết được trình bày trong luận văn là sự tổng hợp từ các kết quả đã được công bố và có trích dẫn đầy đủ, số liệu và kết quả của chương trình thực nghiệm trong luận văn này được tác giả thực hiện là hoàn toàn trung thực, nếu sai tôi hoàn toàn chịu trách nhiệm.

Thái Nguyên, tháng 5 năm 2017

Học viên

Mông Quốc Tuấn

MỤC LỤC

DANH MỤC HÌNH

DANH MỤC BẢNG BIỂU VÀ ĐỒ THỊ

PHẦN MỞ ĐẦU

Trong xã hội ngày nay, con người không những cần nắm bắt nhiều thông tin hơn, mà còn phải nhanh hơn. Internet là một trong những phương tiện quan trọng giúp con người có thể tiếp cận thông tin nhanh nhất. Một trong những tác dụng lớn của Internet trong thập kỷ vừa qua là Thương mại điện tử. Thương mại điện tử ra đời mở ra một kỉ nguyên mới trong thời kì thương mại trên Internet. Một trong những lợi thế lớn nhất của thương mại điện tử chính là khả năng cung cấp cho khách hàng mối liên hệ linh hoạt và mang tính cá nhân hóa.

Trên quan điểm của người sử dụng luôn có xu hướng muốn tìm được sản phẩm và dịch vụ thích hợp nhất đối với nhu cầu và sở thích của bản thân, nhưng mất càng ít thời gian tìm kiếm càng tốt, và với các thao tác càng đơn giản càng tốt. Trên quan điểm của những người thiết kế hệ thống và những nhà cung cấp dịch vụ, vấn đề đặt ra là làm sao xây dựng được các chiến lược kinh doanh và các giải pháp kỹ thuật tích hợp cho việc cung cấp các sản phẩm và dịch vụ đến cho các khách hàng tiềm năng. Các chiến lược kinh doanh tốt sẽ giúp mang lại hiệu quả đầu tư và tăng lợi nhuận. Hai mục tiêu này (của người sử dụng và của nhà cung cấp dịch vụ) có thể đạt được bằng cách cung cấp các hỗ trợ cho người sử dụng trong việc ra quyết định.

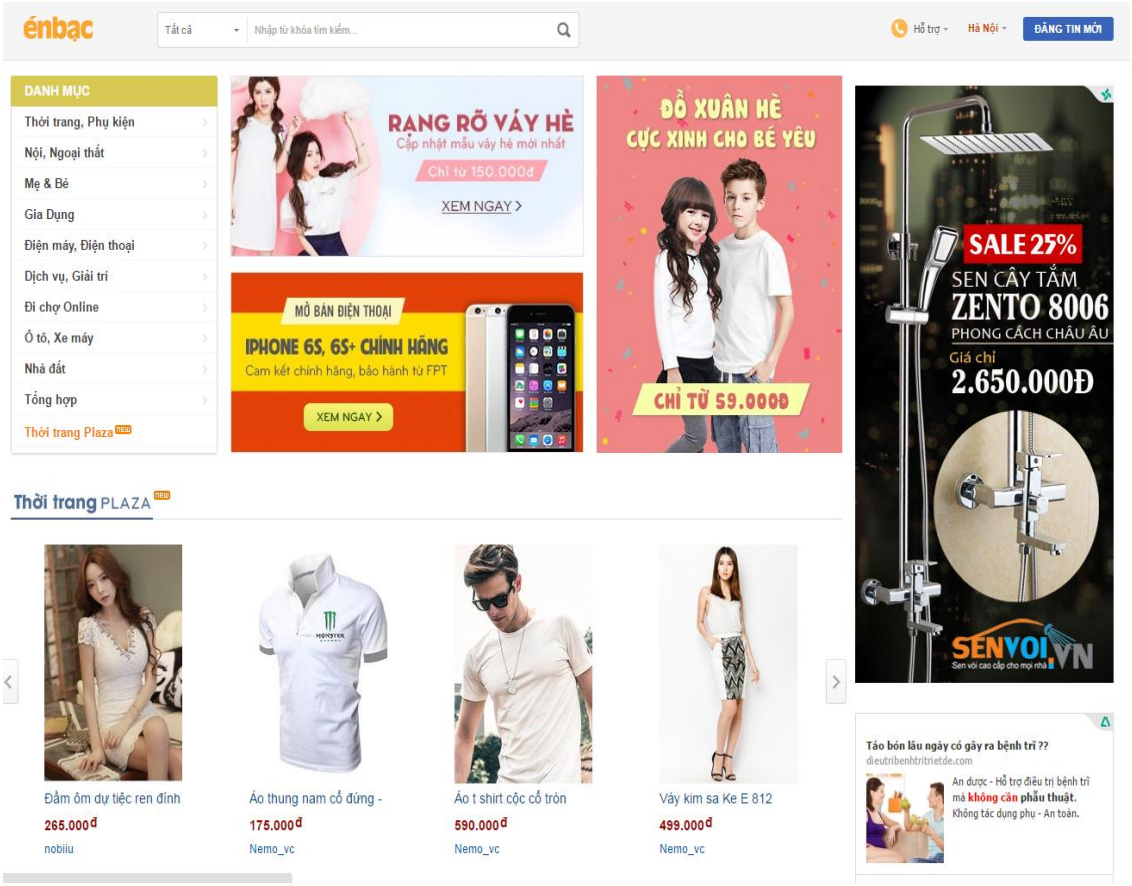
Tuy vậy, không phải hoàn toàn các website thương mại điện tử đều có thể đáp ứng được tất cả nhu cầu của người dùng và có thể giúp họ tìm kiếm được chính xác những sản phẩm mà họ cần mua. Lượng sản phẩm quá lớn, khiến người sử dụng không nhận được những thông tin cần thiết khi họ sử dụng công cụ tìm kiếm của sản phẩm. Phải duyệt qua tất cả các kết quả của quá trình tìm kiếm là công việc mệt mỏi đối với người dùng. Trong những năm gần đây, hệ thống gợi ý (recommender system) được biết đến như là một sự phát triển quan trọng trong việc giúp người dùng đối mặt

với sự bùng nổ thông tin. Hệ thống này được ứng dụng trong nhiều lĩnh vực như thương mại điện tử với Amazon, Netflix, Ebay trong lĩnh vực giải trí với MovieLens, Last.fm, Film-Conseil; trong lĩnh vực khác như tin tức trực tuyến Netnews,…

Hệ gợi ý (recommender systems) là một dạng của hệ hỗ trợ ra quyết định, cung cấp giải pháp mang tính cá nhân hóa mà không phải trải qua quá trình tìm kiếm phức tạp. Hệ gợi ý học từ khách hàng và gợi ý các sản phẩm tốt nhất trong số các sản phẩm phù hợp. Sự phát triển nhanh chóng của thương mại điện tử, sự bùng nổ thông tin khiến cho việc tìm kiếm sản phẩm thích hợp để mua của khách hàng khó khăn hơn.

Hiện nay, việc áp dụng hệ gợi ý vào các website thương mại điện tử là điều tất yếu nhằm tiết kiệm thời gian, công sức và chi phí cho khách hàng, giúp họ tìm ra sản phẩm ưng ý nhất để mua. Hệ gợi ý sử dụng các tri thức về sản phẩm, các tri thức của chuyên gia hay tri thức khai phá học được từ hành vi của người tiêu dùng để đưa ra các gợi ý về sản phẩm mà họ thích trong hàng ngàn hàng vạn sản phẩm có trong hệ thống. Các website thương mại điện tử, ví dụ như sách, phim, nhạc, báo,... sử dụng hệ thống gợi ý để cung cấp các thông tin giúp cho người sử dụng quyết định sẽ lựa chọn sản phẩm nào.

Các sản phẩm được gợi ý dựa trên số lượng sản phẩm đó đã được bán, dựa trên các thông tin người sử dụng, dựa trên sự phân tích hành vi mua hàng trước đó của người sử dụng để đưa ra các dự đoán về hành vi mua hàng trong tương lai của chính khách hàng đó. Các dạng gợi ý gồm: Gợi ý sản phẩm tới người tiêu dùng, các sản phẩm mang tính cá nhân hóa, tổng kết các ý kiến cộng đồng, và cung cấp các chia sẻ, các phê bình, đánh giá mang tính cộng đồng liên quan tới yêu cầu, mục đích của người sử dụng đó.



Hình 1.1: Ví dụ về giao diện một hệ thống thương mại điện tử

Chính vì những lý do trên, tôi nhận thấy sự cần thiết xây dựng một *Hệ thống gợi ý thương mại điện tử với người sử dụng mở*. Nội dung quá trình nghiên cứu nhằm hiện thực hoá Hệ thống này được trình bày trong phần tiếp theo.

Bố cục luận văn

Nội dung nghiên cứu của luận văn gồm:

Chương 1: Đưa ra các khái niệm chung một cách tổng quan về hệ thống gợi ý thông tin, tìm hiểu về thương mại điện tử và ứng dụng của thương mại điện tử

Chương 2: Thảo luận các nghiên cứu liên quan về học máy (machine learning), tìm hiểu về các loại hình học máy trong đó có 3 loại học máy quan trọng được áp dụng trong luận văn của tôi đó là học máy có giám

sát(Supervised Learning) và học máy không giám sát (UnSupervised Learning) và học máy bán giám sát (Semi UnSupervised Learning) từ đó áp dụng vào hệ thống thử nghiệm trong chương 3

Chương 3: Chương này sẽ trình bày các thử nghiệm nhằm đánh giá hiệu quả hoạt động tổng thể của hệ thống gợi ý trong thương mại điện tử, với thành phần then chốt là các tương tác của người dùng với hệ thống. Nội dung chương này bao gồm:

- Các yêu cầu với hệ thống cần xây dựng
- Mô tả hồ sơ đối tượng thông tin – trong trường hợp cụ thể của các sản phẩm trong hệ thống
- Trình bày về việc ứng dụng thuật toán K-Means trong phân cụm dữ liệu và ứng dụng thuật toán vào sản phẩm thực tế.
- Mô tả thiết kế hệ thống gợi ý thương mại điện tử, dựa trên nộidung, bao gồm hai phần chính: Lõi xử lý OFFSHORE MODE, tầng giao tiếp trung gian ONLINE MODE.
- Đưa ra được các độ đo dùng để đánh giá hệ thống

Phương pháp nghiên cứu

Để hoàn thành nội dung nghiên cứu đã đề ra, đầu tiên, tôi tiến hành tìm hiểu kiến thức cơ bản về các mô hình hệ thống gợi ý nói chung. Sau đó, dựa trên những đặc điểm riêng của thương mại điện tử và điều kiện thực tế mà chọn hướng tiếp cận phù hợp.

Khi đã xác định được hướng tiếp cận, tôi tiến hành nghiên cứu các thuật toán và xây dựng hệ thống. Song song với đó, các thói quen tìm kiếm sản phẩm của mọi người trên các website thương mại điện tử hiện nay cũng được điều tra, thông qua hình thức hỏi đáp trực tiếp. Các thông tin này sẽ giúp ích rất nhiều cho quá trình xây dựng hệ thống thử nghiệm, đặc biệt là giao diện và tương tác trên website.

Cuối cùng, một số thử nghiệm khác nhau sẽ được tiến hành, nhằm đánh giá khả năng của hệ thống, từ đó, đề xuất các hướng nghiên cứu tiếp theo trong tương lai.

Kết quả thu được

Sau quá trình nghiên cứu cơ bản, tôi đã quyết định xây dựng một hệ thống gợi ý (hay chọn lọc) thương mại điện tử, hoàn chỉnh, theo hướng tiếp cận dựa trên nội dung (content-based approach). Trong đó, thành phần quan trọng nhất là việc gán nhãn cho các sản phẩm của hệ thống để có thể phân cụm và gợi ý cho người dùng những sản phẩm gần với nhu cầu của họ nhất, có khả năng: Mô hình hoá thông tin dựa trên các thuật toán học máy đó là các thuật toán học có giám sát và học không giám sát, trong luận văn này tôi sử dụng thuật toán K-means là thuật toán học máy không giám sát.

Qua thử nghiệm, thuật toán K-means đã chứng tỏ rằng nó hoạt động hiệu quả hơn so với từng mô hình thông tin (ngắn hạn, dài hạn) độc lập, có khả năng nắm bắt nhanh sở thích của người dùng và theo dõi được những gì họ đã đọc tuy nhiên sự chính xác vẫn chưa cao, có đôi khi còn đưa ra gợi ý không chính xác do cách tính hàm khoảng cách giữa các đối tượng chưa hoàn toàn chính xác.

Do dữ liệu của hệ thống dạng này phụ thuộc nhiều vào phản hồi, tương tác của người dùng, nên một phương pháp xử lý dữ liệu đầu vào được đưa vào nhằm giải quyết vấn đề của bài toán đó là: Do là sản phẩm thương mại điện tử nên tất cả các thuộc tính của sản phẩm đều rất phức tạp trong vấn đề phân tích dữ liệu nếu không thể đưa vào cùng một hệ tọa độ để tính khoảng cách, do vậy trước khi dùng thuật toán K-means để phân cụm dữ liệu các dữ liệu đầu vào tôi sử dụng Mạng Neuron để huấn luyện các đối tượng và quy hoàn toàn các đối tượng có dữ liệu đầu vào không phải dạng số về cùng 1 dạng dữ liệu để tính khoảng cách.

CHƯƠNG 1:
KHÁI QUÁT CHUNG VỀ HỆ THỐNG GỢI Ý THÔNG TIN VÀ THƯƠNG MẠI ĐIỆN TỬ

Chương đầu tiên sẽ bắt đầu từ việc giải thích lý do, động lực thực hiện đề tài luận văn này. Nội dung của chương xoay quanh trình bày bối cảnh và sự cần thiết của một *Hệ thống gợi ý thông tin,* cách tiếp cận của hệ thống gợi ý thông tin trong thương mại điện tử, mô tả các nội dung nghiên cứu trong luận văn, cùng với sơ lược các kết quả đã đạt được.

1.1 Hệ thống gợi ý thông tin (Recommender Systems)

1.1.1 Khái niệm hệ thống gợi ý thông tin

Hệ thống gợi ý (Recommender Systems - RS) là một dạng của hệ thống lọc thông tin (information filtering), nó được sử dụng để dự đoán sở thích (preferences) hay xếp hạng (rating) mà người dùng có thể dành cho một mục thông tin (item) nào đó mà họ chưa xem xét tới trong quá khứ (item có thể là bài hát, bộ phim, đoạn video clip, sách, bài báo,..).

1.1.2 Một số ứng dụng của hệ thống gợi ý thông tin

Hiện nay với lượng dữ liệu quá lớn cho các hệ thống lớn, việc đưa ra được các gợi ý chính xác nhằm tiết kiệm thời gian cho người dùng là rất quan trọng và nó giúp hệ thống có thể hữu ích hơn rất nhiều so với nhứng hệ thống khác.

Ví dụ, trong hệ thống bán hàng trực tuyến (chẳng hạn như Amazon), nhằm tối ưu hóa khả năng mua sắm của khách hàng (user), người ta quan tâm đến việc những khách hàng nào đã ‘yêu thích’ những sản phẩm (item) nào bằng cách dựa vào dữ liệu quá khứ của họ (dữ liệu này có thể là xếp hạng mà người dùng đã bình chọn trên sản phẩm, thời gian duyệt (browse) trên sản phẩm, số lần click chuột trên sản phẩm,..) từ đó hệ thống sẽ dự

đoán được người dùng có thể thích sản phẩm nào và đưa ra những gợi ý phù hợp cho họ.

Hệ thống gợi ý không chỉ đơn thuần là một dạng Hệ thống thông tin mà nó còn là cả một lĩnh vực nghiên cứu hiện đang rất được các nhà khoa học quan tâm. Kể từ năm 2007 đến nay, hàng năm đều có hội thảo chuyên về hệ thống gợi ý của ACM (ACM RecSys) cũng như các tiểu bang dành riêng cho RS trong các hội nghị lớn khác như ACM KDD, ACM CIKM,..

1.2 Tổng quát chung về thương mại điện tử

1.2.1 Thương mại điện tử là gì ?

Cho đến hiện tại có nhiều định nghĩa khác nhau về thương mại điện tử(TMĐT). Các định nghĩa này xem xét theo các quan điểm, khía cạnh khác nhau. Theo quan điểm truyền thông, thương mại điện tử là khả năng phân phối sản phẩm, dịch vụ, thông tin hoặc thanh toán thông qua một mạng ví dụ Internet hay world wide web.

Theo [19], thương mại điện tử liên quan đến nhiều hình thức trao đổi thông tin giữa doanh nghiệp với nhau, giữa khách hàng với doanh nghiệp và giữa khách hàng với khách hàng.

Theo quan điểm quá trình kinh doanh: thương mại điện tử bao gồm các hoạt động được hỗ trợ trực tiếp bởi liên kết mạng.

Theo quan điểm môi trường kinh doanh: Thương mại điện tử là một môi trường cho phép có thể mua bán các sản phẩm, dịch vụ và thông tin trên Internet. Sản phẩm có thể hữu hình hay vô hình.

Theo quan điểm cấu trúc: Thương mại điện tử liên quan đến các phương tiện thông tin để truyền: văn bản, trang web, điện thoại Internet, video Internet.

Sau đây là một số định nghĩa khác về thương mại điện tử:

Thương mại điện tử là tất cả các hình thức giao dịch được thực hiện

thông qua mạng máy tính có liên quan đến chuyển quyền sở hữu về sản phẩm hay dịch vụ.

Theo diễn đàn đối thoại xuyên Đại tây dương, thương mại điện tử là các giao dịch thương mại về hàng hoá và dịch vụ được thực hiện thông qua các phương tiện điện tử.

Cục Thống kê Hoa kỳ định nghĩa thương mại điện tử là việc hoàn thành bất kỳ một giao dịch nào thông qua một mạng máy tính làm trung gian mà bao gồm việc chuyển giao quyền sở hữu hay quyền sử dụng hàng hoá và dịch vụ.

Theo nghĩa rộng có nhiều định nghĩa khác về thương mại điện tử như thương mại điện tử là toàn bộ chu trình và các hoạt động kinh doanh liên quan đến các tổ chức hay cá nhân hay thương mại điện tử là việc tiến hành hoạt động thương mại sử dụng các phương tiện điện tử và công nghệ xử lý thông tin số hoá.

UNCITAD định nghĩa về thương mại điện tử bao gồm việc sản xuất, phân phối, marketing, bán hay giao hàng hoá và dịch vụ bằng các phương tiện điện tử.

Bao gồm các giao dịch thương mại thông qua các mạng viễn thông Liên minh châu Âu định nghĩa thương mại điện tử và sử dụng các phương tiện điện tử. Nó bao gồm thương mại điện tử gián tiếp (trao đổi hàng hoá hữu hình) và thương mại điện tử trực tiếp (trao đổi hàng hoá vô hình).

Thương mại điện tử cũng được hiểu là hoạt động kinh doanh điện tử, bao gồm: mua bán điện tử hàng hoá, dịch vụ, giao hàng trực tiếp trên mạng với các nội dung số hoá được, chuyển tiền điện tử - EFT(electronic fund transfer), mua bán cổ phiếu điện tử - EST (electronic share trading), vận đơn điện tử - E B/L (electronic bill of lading) đấu giá thương mại - Commercial auction, hợp tác thiết kế và sản xuất, tìm kiếm các nguồn lực

trực tuyến, mua sắm trực tuyến - Online procurement, marketing trực tiếp, dịch vụ khách hàng sau khi bán...

UN đưa ra định nghĩa đầy đủ nhất để các nước có thể tham khảo làm chuẩn, tạo cơ sở xây dựng chiến lược phát triển thương mại điện tử phù hợp. Định nghĩa này phản ánh các bước thương mại điện tử , theo chiều ngang: "thương mại điện tử là việc thực hiện toàn bộ hoạt động kinh doanh bao gồm marketing, bán hàng, phân phối và thanh toán (MSDP) thông qua các phương tiện điện tử".

Định nghĩa của WTO Thương mại điện tử bao gồm việc sản xuất, quảng cáo, bán hàng và phân phối sản phẩm được mua bán và thanh toán trên mạng Internet, nhưng được giao nhận có thể hữu hình hoặc giao nhận qua Internet dưới dạng số hoá.

Định nghĩa của OECD (Tổ chức Hợp tác và Phát triển Kinh tế): Thương mại điện tử là việc làm kinh doanh thông qua mạng Internet, bán những hàng hoá và dịch vụ có thể được phân phối không thông qua mạng hoặc những hàng hoá có thể mã hoá bằng kỹ thuật số và được phân phối thông qua mạng hoặc không thông qua mạng.

Định nghĩa của AEC(Hiệp hội thương mại điện tử): Thương mại điện tử là làm kinh doanh có sử dụng các công cụ điện tử. Định nghĩa này rộng, coi hầu hết các hoạt động kinh doanh từ đơn giản như một cú điện thoại giao dịch đến những trao đổi thông tin EDI phức tạp đều là thương mại điện tử.

Trong Luật mẫu về thương mại điện tử, UNCITRAL (Ủy ban của LHQ về thương mại quốc tế) nêu định nghĩa để các nước tham khảo: Thương mại điện tử là việc trao đổi thông tin thương mại thông qua các phương tiện điện tử, không cần phải in ra giấy bất cứ công đoạn nào của toàn bộ quá trình giao dịch.

Kinh doanh điện tử (E-business): cũng có nhiều quan điểm khác nhau, về cơ bản kinh doanh điện tử được hiểu theo góc độ quản trị kinh doanh, đó là việc ứng dụng công nghệ thông tin và Internet vào các quy trình, hoạt động của doanh nghiệp.

Ngoài khái niệm E-commerce và E-business, đôi khi người ta còn sử dụng khái niệm M-commerce. M-commerce(mobile commerce) là kinh doanh sử dụng mạng điện thoại di động.

Ở đây "Thông tin" được hiểu là bất cứ thứ gì có thể truyền tải bằng kỹ thuật điện tử, bao gồm cả thư từ, các file văn bản, các cơ sở dữ liệu, các bản tính, các bản thiết kế, hình đồ hoạ, quảng cáo, hỏi hàng, đơn hàng, hoá đơn, bảng giá, hợp đồng, hình ảnh động, âm thanh...

"Thương mại" được hiểu theo nghĩa rộng bao quát mọi vấn đề nảy sinh từ mọi mối quan hệ mang tính thương mại, dù có hay không có hợp đồng. Các mối quan hệ mang tính thương mại bao gồm, nhưng không chỉ bao gồm, các giao dịch sau đây: bất cứ giao dịch nào về cung cấp hoặc trao đổi hàng hoá hoặc dịch vụ, đại diện hoặc đại lý thương mại, uỷ thác hoa hồng; cho thuê dài hạn, xây dựng các công trình, tư vấn, kỹ thuật công trình, đầu tư cấp vốn, ngân hàng, bảo hiểm, thoả thuận khai thác hoặc tô nhượng, liên doanh và các hình thức khác về hợp tác công nghiệp hoặc kinh doanh, chuyên chở hàng hoá hay hành khách bằng đường biển, đường không, đường sắt hoặc đường bộ.

Mạng trong thương mại điện tử được hiểu là bao gồm các máy tính, máy fax, điện thoại,TV… được kết nối với nhau để trao đổi thông tin dưới dạng điện tử.

1.2.2 Lợi ích của TMĐT

Lợi ích lớn nhất mà TMĐT đem lại chính là sự tiết kiệm chi phí và tạo thuận lợi cho các bên giao dịch. Giao dịch bằng phương tiện điện tử

nhanh hơn so với giao dịch truyền thống, ví dụ gửi fax hay thư điện tử thì nội dung thông tin đến tay người nhận nhanh hơn gửi thư. Các giao dịch qua Internet có chi phí rất rẻ, một doanh nghiệp có thể gửi thư tiếp thị, chào hàng đến hàng loạt khách hàng chỉ với chi phí giống như gửi cho một khách hàng. Với TMĐT, các bên có thể tiến hành giao dịch khi ở cách xa nhau, giữa thành phố với nông thôn, từ nước này sang nước kia, hay nói cách khác là không bị giới hạn bởi không gian địa lý. Điều này cho phép các doanh nghiệp tiết kiệm chi phí đi lại, thời gian gặp mặt trong khi mua bán. Với người tiêu dùng, họ có thể ngồi tại nhà để đặt hàng, mua sắm nhiều loại hàng hóa, dịch vụ thật nhanh chóng.

Những lợi ích nhưtrên chỉ có được với những doanh nghiệp thực sự nhận thức được giá trị của TMĐT. Vì vậy, TMĐT góp phần thúc đẩy sự cạnh tranh giữa các doanh nghiệp để thu được nhiều lợi ích nhất. Điều này đặc biệt quan trọng trong bối cảnh hội nhập kinh tế quốc tế, khi các doanh nghiệp trong nước phải cạnh tranh một cách bình đẳng với các doanh nghiệp nước ngoài.

1.2.3 Các loại hình ứng dụng TMĐT

TMĐT được phân chia thành một số loại như B2B, B2C, C2C dựa trên thành phần tham gia hoạt động thương mại. Có thể sử dụng hình sau để minh họa cách phân chia này.

Bảng 1.1: Các loại hình TMĐT

	Government	Business	Consumer
Government	G2G	G2B	G2C
Business	B2G	B2B	B2C
Consumer	C2G	C2B	C2C

Hình thức giao dịch thương mại điện tử doanh nghiệp với khách hàng (Business to Customer B2C) thành phần tham gia hoạt động thương mại gồm người bán là doanh nghiệp và người mua là người tiêu dùng. Sử dụng trình duyệt (web browser) để tìm kiếm sản phẩm trên Internet. Sử dụng giỏ hàng (shopping cart) để lưu trữ các sản phẩm khách hàng đặt mua. Thực hiện thanh toán bằng điện tử.

Hình thức giao dịch thương mại điện tử doanh nghiệp với doanh nghiệp (Business to Business - B2B): Thành phần tham gia hoạt động thương mại là các doanh nghiệp, tức người mua và người bán đều là doanh nghiệp. Sử dụng Internet để tạo mối quan hệ giữa nhà cung cấp và các cửa hàng thông qua các vấn đề về chất lượng, dịch vụ. Marketing giữa hai đối tượng này là marketing công nghiệp. Hình thức này phổ biến nhanh hơn B2C. Khách hàng là doanh nghiệp có đủ điều kiện tiếp cận và sử dụng Internet hay mạng máy tính. Thanh toán bằng điện tử.

Giao dịch giữa doanh nghiệp với cơ quan chính quyền (Business to Government - B2G) và giao dịch giữa doanh nghiệp với cơ quan chính quyền (B2G). Các giao dịch này gồm khai hải quan, nộp thuế, báo cáo tài chính và nhận các văn bản pháp qui... Giao dịch giữa các cá nhân với cơ quan chính quyền (Custmer to Government C2G). Các giao dịch này gồm xin giấy phép xây dựng, trước bạ nhà đất…

Hai loại giao dịch này thuộc về một hình thức được gọi là chính phủ điện tử. Chính phủ điện tử là cách thức qua đó các Chính phủ sử dụng các công nghệ mới trong hoạt động để làm cho người dân, Doanh nghiệp tiếp cận các thông tin và dịch vụ do Chính phủ cung cấp một cách thuận tiện hơn, để cải thiện chất lượng dịch vụ và mang lại các cơ hội tốt hơn cho người dân, Doanh nghiệp trong việc tham gia vào xây dựng các thể chế và tiến trình phát triển đất nước.

Mục đích của chính phủ điện tử là của dân, do dân và vì dân, có ảnh hưởng mang tính cách mạng đến sức mạnh và sự sống còn của các Chính phủ và nền dân chủ thực sự ở mỗi quốc gia. Việc phát triển chính phủ điện tử theo lộ trình được hoạch định sẽ mở ra khả năng phát huy sự đóng góp trí tuệ của tất cả người dân tham gia vào quá trình thúc đẩy sự phát triển đất nước. Chính phủ điện tử sẽ cải thiện chính phủ theo 4 cách thức quan trọng:

- Người dân có thể đóng góp ý kiến một cách dễ dàng hơn đối với Chính phủ.

- Người dân sẽ nhận được các dịch vụ tốt hơn từ các cơ quan tổ chức Chính phủ bất kỳ lúc nào, bất kỳ ở đâu (tại nhà, ở công sở, trạm điện thoại…) và vì bất kỳ lý do gì.

Đây là hình thức phát triển mới của mô hình Chính phủ một cửa: Chính phủ có nhiều cửa và khách hàng có thể thông qua một cửa bất kỳ để tiếp cận được các dịch vụ của chính phủ.

- Người dân sẽ nhận được nhiều dịch vụ thích hợp hơn từ các cơ quan Chính phủ, bởi các cơ quan này sẽ phối hợp một cách hiệu quả hơn với nhau.

- Người dân sẽ có được thông tin một cách tốt hơn vì họ có thể nhận được các thông tin cập nhật và toàn diện về các luật lệ, quy chế, chính sách và dịch vụ của chính phủ.

Các dịch vụ chính phủ trực tuyến:

- Trước đây các cơ quan chính phủ cung cấp dịch cho dân chúng tại trụ sở của mình, thì nay nhờ vào công nghệ thông tin và viễn thông, các trung tâm dịch vụ trực tuyến được thiết lập, hoặc là ngay trong trụ sở cơ quan chính phủ hoặc gần với dân.

- Qua các cổng thông tin cho công dân, người dân nhận được thông tin, có thể hỏi đáp pháp luật, được phục vụ giải quyết các việc trong cuộc

sống hàng ngày: Chuyển quyền sử dụng đất, cấp phép xây dựng, cấp đăng ký kinh doanh, chứng thực, và xác nhận chính sách xã hội…mà không phải đến trực tại trụ sở các cơ quan Chính phủ như trước đây.

Ngoài các hình thức kể trên, còn phải kể đến hình thức giao dịch giữa các cá nhân với nhau hay còn gọi là giao dịch Customer to Customer (C2C) hoặc Peer to Peer (P2P). Thành phần tham gia hoạt động thương mại là các cá nhân, tức người mua và người bán đều là cá nhân.

B2C là loại hình giao dịch giữa doanh nghiệp và người tiêu dùng qua các phương tiện điện tử.

Hình 1.2: Sơ đồ chu trình hệ thống TMĐT B2C

Doanh nghiệp sử dụng các phương tiện điện tử để bán hàng hóa, dịch vụ tới người tiêu dùng. Người tiêu dùng thông qua các phương tiện điện tử để lựa chọn, mặc cả, đặt hàng, thanh toán, nhận hàng. Giao dịch B2C tuy chiếm tỷ trọng ít (khoảng 10%) trong TMĐT nhưng có sự phạm vi ảnh

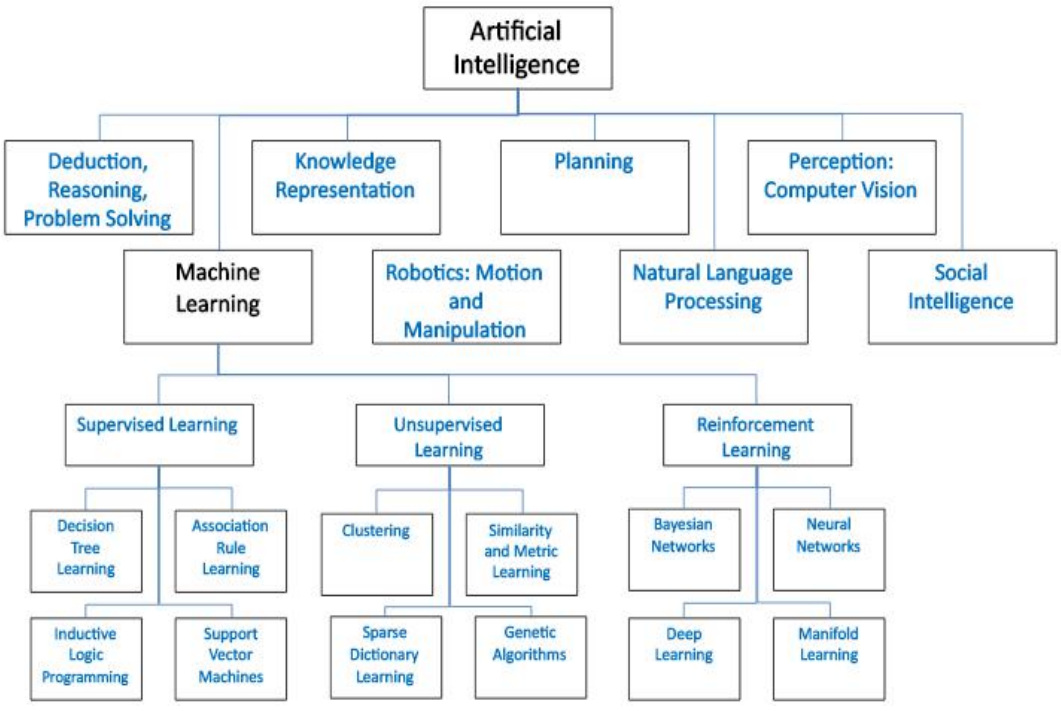
hưởng rộng. Để tham gia hình thức kinh doanh này, thông thường doanh nghiệp sẽ thiết lập website, hình thành cơ sở dữ liệu về hàng hoá, dịch vụ; tiến hành các quy trình tiếp thị, quảng cáo, phân phối trực tiếp tới người tiêu dùng. TMĐT B2C đem lại lợi ích cho cả doanh nghiệp lẫn người tiêu dùng. Doanh nghiệp tiết kiệm nhiều chi phí bán hàng do không cần phòng trưng bày hay thuê người giới thiệu bán hàng, chi phí quản lý cũng giảm hơn. Người tiêu dùng sẽ cảm thấy thuận tiện vì không phải tới tận cửa hàng, có khả năng lựa chọn và so sánh nhiều mặt hàng cùng một lúc…

CHƯƠNG 2:
HỌC MÁY VÀ CÁC PHƯƠNG PHÁP PHÂN CỤM DỮ LIỆU

2.1 Tổng quan về học máy(Machine learning)

2.1.1 Học máy là gì?

Machine learning gây nên cơn sốt công nghệ trên toàn thế giới trong vài năm nay. Trong giới học thuật, mỗi năm có hàng ngàn bài báo khoa học về đề tài này. Trong giới công nghiệp, từ các công ty lớn như Google, Facebook, Microsoft đến các công ty khởi nghiệp đều đầu tư vào machine learning. Hàng loạt các ứng dụng sử dụng machine learning ra đời trên mọi linh vực của cuộc sống, từ khoa học máy tính đến những ngành ít liên quan hơn như vật lý, hóa học, y học, chính trị. AlphaGo, cỗ máy AI với khả năng tính toán tối ưu hơn bất kì đại kì thủ nào trong một không gian có số lượng phần tử còn nhiều hơn số lượng hạt trong vũ trụ, là một trong rất nhiều ví dụ hùng hồn cho sự vượt trội của machine learning so với các phương pháp cổ điển.



Hình 2.1: Sơ đồ tổng quát về học máy

Vậy thực chất, machine learning là gì?

Để giới thiệu về machine learning, theo [1] tôi xin dựa vào mối quan hệ của nó với ba khái niệm sau:

Machine learning và trí tuệ nhân tạo (Artificial Intelligence hay AI)

Machine learning và Big Data

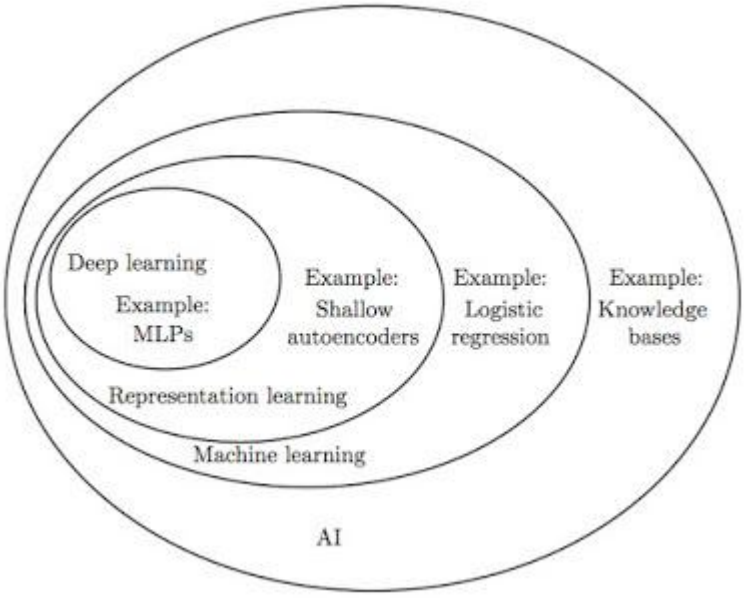
Machine learning và dự đoán tương lai

Trí tuệ nhân tạo, AI, một cụm từ vừa gần gũi vừa xa lạ đối với chúng ta. Gần gũi bởi vì thế giới đang phát sốt với những công nghệ được dán nhãn AI. Xa lạ bởi vì một AI thực thụ vẫn còn nằm ngoài tầm với của chúng ta. Nói đến AI, hẳn mỗi người sẽ liên tưởng đến một hình ảnh khác nhau. Các bạn có để ý rằng vài thập niên gần đây có một sự thay đổi về diện mạo của AI trong các bộ phim quốc tế. Trước đây, các nhà sản xuất phim thường xuyên đưa hình ảnh robot hoặc terminator vào phim, nhằm gieo vào đầu người xem suy nghĩ rằng trí tuệ nhân tạo như một phương thức nhân bản con người bằng máy móc. Tuy nhiên, trong những bộ phim gần đây nhất về đề tài này, ví dụ như Transcendence do Johny Depp vào vai chính, ta không thấy hình ảnh của một con robot nào cả. Thay vào đó là một bộ não điện toán khổng lồ chỉ huy hàng vạn con Nanobot, được gọi là Singularity. Tất nhiên cả hai hình ảnh đều là hư cấu và giả tưởng, nhưng sự thay đổi như vậy cũng một phần nào phản ánh sự thay đổi ý niệm của con người về AI. AI bây giờ được xem như vô hình vô dạng, hay nói cách khác có thể mang bất cứ hình dạng nào.

Trong giới hàn lâm, theo hiểu biết chung, AI là một ngành khoa học được sinh ra với mục đích làm cho máy tính có được trí thông minh. Mục tiêu này vẫn khá mơ hồ vì không phải ai cũng đồng ý với một định nghĩa thống nhất về trí thông minh. Thế nên các nhà khoa học phải định nghĩa một số mục tiêu cụ thể hơn, một trong số đó là việc làm cho máy tính lừa

được Turing Test. Turing Test được tạo ra bởi Alan Turing (1912-1954), người được xem là cha đẻ của ngành khoa học máy tính hiện đại, nhằm phân biệt xem người đối diện có phải là người hay không (xem phim The Imitation Game về nhân vật này, nhưng đừng tin hết những gì trong phim).

AI thể hiện một mục tiêu của con người, Machine learning là một phương tiện được kỳ vọng sẽ giúp con người đạt được mục tiêu đó và thực tế thì machine learning đã mang nhân loại đi rất xa trên quãng đường chinh phục AI nhưng vẫn còn một quãng đường xa hơn cần phải đi. Machine learning và AI có mối quan hệ chặt chẽ với nhau nhưng không hẳn là trùng khớp vì một bên là mục tiêu (AI), một bên là phương tiện (machine learning).



Hình 2.2: Sơ đồ các lớp trí tuệ nhân tạo

Chinh phục AI mặc dù vẫn là mục đích tối thượng của machine learning, nhưng hiện tại machine learning tập trung vào những mục tiêu ngắn hạn hơn như:

Làm cho máy tính có những khả năng nhận thức cơ bản của con người như nghe, nhìn, hiểu được ngôn ngữ, giải toán, lập trình, …

Hỗ trợ con người trong việc xử lý một khối lượng thông tin khổng lồ mà chúng ta phải đối mặt hàng ngày, hay còn gọi là Big Data.

Big Data thực chất không phải là một ngành khoa học chính thống. Đó là một cụm từ dân gian và được giới truyền thông tung hô để ám chỉ thời kì bùng nổ của dữ liệu hiện nay. Nó cũng không khác gì với những cụm từ như "cách mạng công nghiệp", "kỉ nguyên phần mềm". Big Data là một hệ quả tất yếu của việc mạng Internet ngày càng có nhiều kết nối. Với sự ra đời của các mạng xã hội nhưng Facebook, Instagram, Twitter, nhu cầu chia sẻ thông của con người tăng trưởng một cách chóng mặt. Youtube cũng có thể được xem là một mạng xã hội, nơi mọi người chia sẻ video và comment về nội dung của video. Để hiểu được quy mô của Big Data, hãy xem qua những con số sau đây:

Hơn 900 triệu người thật sự sử dụng Facebook mỗi ngày, 82.8% trong số đó ở ngoài Mỹ và Canada (theo http://newsroom.fb.com/company-info/)

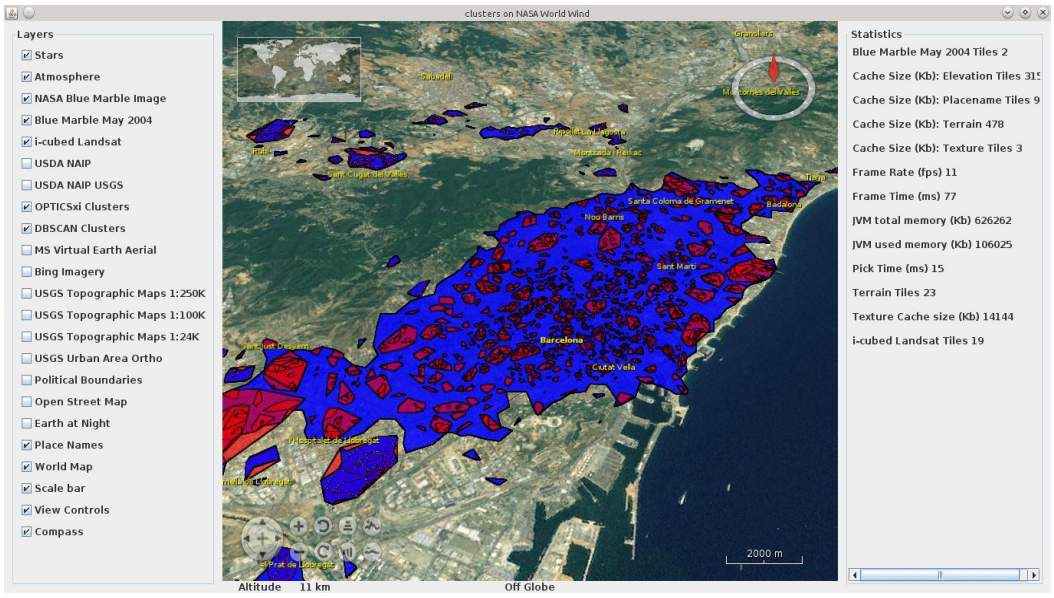
Nhu cầu chia sẻ tăng đi đôi với việc nhu cầu tìm kiếm thông tin cũng tăng. Google phải xử lý 100 tỉ lượt tìm kiếm mỗi tháng, tức là 3,3 tỉ lượt mỗi ngày(theo http://www.internetlivestats.com/google-search-statistics/).

Bùng nổ thông tin không phải là lý do duy nhất dẫn đến sự ra đời của cụm từ Big Data. Nên nhớ rằng Big Data xuất hiện mới từ vài năm gần đây nhưng khối lượng dữ liệu tích tụ kể từ khi mạng Internet xuất hiện vào cuối thế kỉ trước cũng không phải là nhỏ. Thế nhưng, lúc ấy con người ngồi quanh một đống dữ liệu và không biết làm gì với chúng ngoài lưu trữ và sao chép. Cho đến một ngày, các nhà khoa học nhận ra rằng trong đống dữ liệu ấy thực ra chứa một khối lượng tri thức khổng lồ. Những tri thức ấy có thể giúp cho ta hiểu thêm về con người và xã hội. Từ danh sách bộ phim yêu thích của một cá nhân chúng ta có thể rút ra được sở thích của người đó

và giới thiệu những bộ phim người ấy chưa từng xem, nhưng phù hợp với sở thích. Từ danh sách tìm kiếm của cộng đồng mạng chúng ta sẽ biết được vấn đề nóng hổi nhất đang được quan tâm và sẽ tập trung đăng tải nhiều tin tức hơn về vấn đề đó. Big Data chỉ thực sự bắt đầu từ khi chúng ta hiểu được gía trị của thông tin ẩn chứa trong dữ liệu, và có đủ tài nguyên cũng như công nghệ để có thể khai thác chúng trên quy mô khổng lồ. Và không có gì ngạc nhiên khi machine learning chính là thành phần mấu chốt của công nghệ đó. Ở đây ta có một quan hệ hỗ tương giữa machine Learning và Big Data: machine learning phát triển hơn nhờ sự gia tăng của khối lượng dữ liệu, ngược lại, thành công của Big Data phụ thuộc vào khả năng khai thác tri thức từ dữ liệu.

Ngược dòng lịch sử, machine Learning đã xuất hiện từ rất lâu trước khi mạng Internet ra đời. Một trong những thuật toán machine learning nổi tiếng đó chính là thuật toán phân cụm dữ liệu K-Means.

Trước khi tìm hiểu sâu hơn về thuật toán K-Means chúng ta hãy tìm hiểu thế nào là phân cụm dữ liệu và phân cụm có ý nghĩa như thế nào, theo [22] tôi sẽ đưa ra giả thiết về 1 bức tranh phân cụm tổng quát như sau:



Hình 2.3: Mô phỏng khái quát về phân cụm dữ liệu

Việc phân cụm dữ liệu thực sự quan trọng trong học máy, tôi xin đưa ra khái niệm về phân cụm dữ liệu:

Phân cụm dữ liệu là một trong những hành vi nguyên thủy nhất của con người nhằm nắm giữ lượng thông tin khổng lồ họ nhận được hằng ngày vì xử lý mọi thông tin như một thực thể đơn lẻ là không thể. Phân cụm là một kỹ thuật được sử dụng để kết hợp các đối tượng quan sát thành các cụm sao cho mỗi cụm có cùng một số đặc điểm tương đồng ở một số đặc điểm đang xét. Ngược lại các đối tượng trong các nhóm khác nhau thì độ tương đồng khác nhau (ít tương đồng hơn) ở một số đặc điểm đang xét.

Phân cụm là kỹ thuật rất quan trọng trong khai phá dữ liệu, nó thuộc lớp các phương pháp Unsupervised Learning trong Machine Learning. Có rất nhiều định nghĩa khác nhau về kỹ thuật này, nhưng về bản chất ta có thể hiểu phân cụm là các qui trình tìm cách nhóm các đối tượng đã cho vào các cụm (clusters), sao cho các đối tượng trong cùng 1 cụm tương tự (similar) nhau và các đối tượng khác cụm thì không tương tự (Dissimilar) nhau.

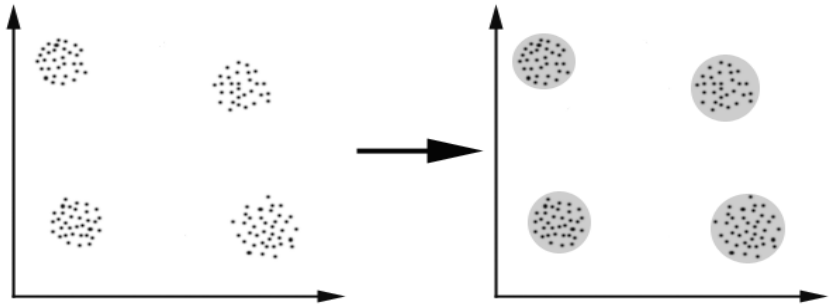
Khi bài toán khai phá dữ liệu đã được phân cụm thì tất cả các bước xử lý dữ liệu sau đó sẽ trở nên thuận lợi và ứng dụng được vào rất nhiều các linh vực khác nhau.

Mục đích của phân cụm là tìm ra bản chất bên trong các nhóm của dữ liệu. Các thuật toán phân cụm (Clustering Algorithms) đều sinh ra các cụm (clusters). Tuy nhiên, không có tiêu chí nào là được xem là tốt nhất để đánh hiệu của của phân tích phân cụm, điều này phụ thuộc vào mục đích của phân cụm như: data reduction, “natural clusters”, “useful” clusters, outlier detection.

Phân cụm dữ liệu nghĩa là quá trình phân chia 1 tập dữ liệu ban đầu thành các cụm dữ liệu thỏa mãn:

Các đối tượng trong 1 cụm “tương tự” nhau.

Các đối tượng khác cụm thì "không tương tự" nhau.



Hình 2.4: Mô phỏng dữ liệu sau khi đã được phân cụm

Nếu X : 1 tập các điểm dữ liệu

Ci : cụm thứ i

X = C1 …Ck…Cngoại lai

Ci $\cap$ Cj = $\varnothing$

Một số độ đo trong phân cụm:

Minkowski

$$\sum_{i=1}^{n} (|| x_i - y_i ||^p)^{\frac{1}{p}}$$

Euclidean – p = 2

Độ đo tương tự (gần nhau): cosin hai vectơ

$$\cos\mu = \frac{v.w}{||v|| . ||w||}$$

Mục đích của phân cụm là định được bản chất của việc nhóm các đối tượng trong 1 tập dữ liệu không có nhãn.

Một số phương pháp phân cụm điển hình:

Phân cụm phân hoạch

Phân cụm phân cấp

Phân cụm dựa trên mật độ

Phân cụm dựa trên lưới

Phân cụm dựa trên mô hình

Phân cụm có ràng buộc

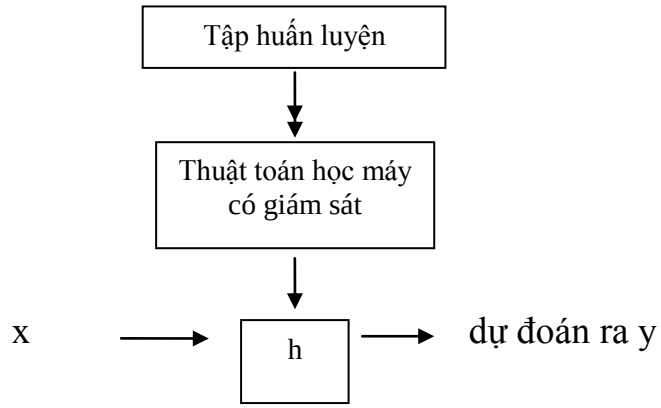
Phân cụm không dựa trên 1 tiêu chuẩn chung nào, mà dựa vào tiêu chí mà người dùng cung cấp trong từng trường hợp.

2.2 Các dạng học máy và các thuật toán liên quan

2.2.1 Các dạng học máy

2.2.2.1 Học máy có giám sát(Supervised Learning)

[21] Học có giám sát là một kỹ thuật của ngành học máy để xây dựng một hàm từ dữ liệu huấn luyện. Dữ liệu huấn luyện bao gồm các cặp đối tượng đầu vào (thường dạng vector) và đầu ra thực sự. Đầu ra của một hàm có thể là một giá trị liên tục (gọi là hồi quy), hay có thể là dự đoán một nhãn phân lớp cho một đối tượng đầu vào (gọi là phân lớp). Nhiệm vụ của chương trình học có giám sát là dự đoán giá trị của hàm cho một đối tượng bất kỳ là đầu vào hợp lệ, sau khi đã xem xét một số ví dụ huấn luyện (nghĩa là, các cặp đầu vào và đầu ra tương ứng). Mục đích chính của bài toán học có giám sát là để học một ánh xạ x tới y. Mô hình chung của học có giám sát được khái quát như hình sau:



Hình 2.5: Mô hình thuật toán học có giám sát

Học có giám sát có thể tạo ra 2 loại mô hình. Phổ biến nhất, học có

giám sát tạo ra một mô hình toàn cục (global model) để ánh xạ đối tượng đầu vào đến đầu ra mong muốn. Tuy nhiên, trong một số trường hợp, việc ánh xạ được thực hiện dưới dạng một tập các mô hình cục bộ (như trong phương pháp lập luận theo tình huống (case-based reasoning) hay giải thuật láng giềng gần nhất).

Để có thể giải quyết một bài toán nào đó của học có giám sát (ví dụ: học để nhận dạng chữ viết tay) người ta phải xem xét nhiều bước khác nhau:

Xác định loại của các ví dụ huấn luyện. Trước khi làm bất cứ điều gì, người kĩ sư nên quyết định loại dữ liệu nào sẽ được sử dụng làm ví dụ. Chẳng hạn, đó có thể là một ký tự viết tay đơn lẻ, toàn bộ một từ viết tay, hay toàn bộ một dòng chữ viết tay.

Thu thập tập huấn luyện. Tập huấn luyện cần đặc trưng cho thực tế sử dụng của hàm chức năng. Vì thế, một tập các đối tượng đầu vào được thu thập và đầu ra tương ứng được thu thập, hoặc từ các chuyên gia hoặc từ việc đo đạc tính toán.

Xác định việc biểu diễn các đặc trưng đầu vào cho hàm chức năng cần tìm. Sự chính xác của hàm chức năng phụ thuộc lớn vào cách các đối tượng đầu vào được biểu diễn. Thông thường, đối tượng đầu vào được chuyển đổi thành một vec-tơ đặc trưng, chứa một số các đặc trưng nhằm mô tả cho đối tượng đó. Số lượng các đặc trưng không nên quá lớn, do sự bùng nổ tổ hợp (curse of dimensionality), nhưng phải đủ lớn để dự đoán chính xác đầu ra.

Xác định cấu trúc của hàm chức năng cần tìm và giải thuật học tương ứng. Ví dụ, người kĩ sư có thể lựa chọn việc sử dụng mạng nơ-ron nhân tạo hay cây quyết định.

Hoàn thiện thiết kế. Người kĩ sư sẽ chạy giải thuật học từ tập huấn luyện thu thập được. Các tham số của giải thuật học có thể được điều chỉnh

bằng cách tối ưu hóa hiệu năng trên một tập con (gọi là tập kiểm chứng - validation set) của tập huấn luyện, hay thông qua kiểm chứng chéo (cross-validation). Sau khi học và điều chỉnh tham số, hiệu năng của giải thuật có thể được đo đạc trên một tập kiểm tra độc lập với tập huấn luyện.

2.2.2.2 Học máy không giám sát(Unsupervised Learning)

[21] Unsupervised Learning(UL)là một kĩ thuật của máy học nhằm tìm ra một mô hình hay cấu trúc bị ẩn bơi tập dữ liệu không được gán nhãn cho trước. UL khác với SL là không thể xác định trước output từ tập dữ liệu huấn luyện được. Tùy thuộc vào tập huấn luyện kết quả output sẽ khác nhau. Trái ngược với SL, tập dữ liệu huấn luyện của UL không do con người gán nhãn, máy tính sẽ phải tự học hoàn toàn. Có thể nói, học không giám sát thì giá trị đầu ra sẽ phụ thuộc vào thuật toán UL.

Ứng dụng phổ biến nhất của học không giám sát là gom cụm (cluster). Đương nhiên sẽ có nhiều ứng dụng khác, có cơ hội tôi sẽ đề cập thêm.

Ứng dụng này dễ nhận ra nhất là Google và Facbook. Google có thể gom nhóm các bài báo có nội dung gần nhau, hoặc Facebook có thể gợi ý kết bạn có nhiều bạn chung cho bạn.

Các bài báo có cùng nội dung sẽ được gom lại thành một nhóm (cluster) phân biệt với các nhóm khác. Dữ liệu huấn luyện là các bài báo từ quá khứ tới hiện tại và tăng dần theo thời gian. Dễ nhận ra rằng dữ liệu không thể gán nhãn bởi con người.

Khi một bài báo mới được cho vào input, nó sẽ tìm cụm (cluster) gần nhất với bài báo đó và gợi ý những bài liên quan.

2.2.2.3 Học máy bán giám sát

Trong khoa học máy tính, học nửa giám sát là một lớp của kỹ thuật học máy, sử dụng cả dữ liệu đã gán nhãn và chưa gán nhãn để huấn luyện -

điển hình là một lượng nhỏ dữ liệu có gán nhãn cùng với lượng lớn dữ liệu chưa gán nhãn. Học nửa giám sát đứng giữa học không giám sát (không có bất kì dữ liệu có nhãn nào) và có giám sát (toàn bộ dữ liệu đều được gán nhãn). Nhiều nhà nghiên cứu nhận thấy dữ liệu không gán nhãn, khi được sử dụng kết hợp với một chút dữ liệu có gán nhãn, có thể cải thiện đáng kể độ chính xác. Để gán nhãn dữ liệu cho một bài toán học máy thường đòi hỏi một chuyên viên có kĩ năng để phân loại bằng tay các ví dụ huấn luyện. Chi phí cho quy trình này khiến tập dữ liệu được gán nhãn hoàn toàn trở nên không khả thi, trong khi dữ liệu không gán nhãn thường tương đối rẻ tiền. Trong tình huống đó, học nửa giám sát có giá trị thực tiễn lớn lao.

Một ví dụ cho kỹ thuật học máy nửa giám sát là đồng huấn luyện (co-training), trong đó một hay nhiều bộ học được huấn luyện cùng một tập ví dụ nhưng mỗi bộ sử dụng một tập đặc trưng khác nhau, lý tưởng nhất là độc lập với nhau.

Một cách tiếp cận khác là mô hình hoá phân phối xác suất đồng thời của các đặc trưng và nhãn. Với dữ liệu chưa gán nhãn, có thể coi nhãn là"dữ liệu còn thiếu". Các kỹ thuật xử lý dữ liệu còn thiếu như là lấy mẫu Gibbs và tối ưu kỳ vọng có thể được sử dụng để ước lượng tham số.

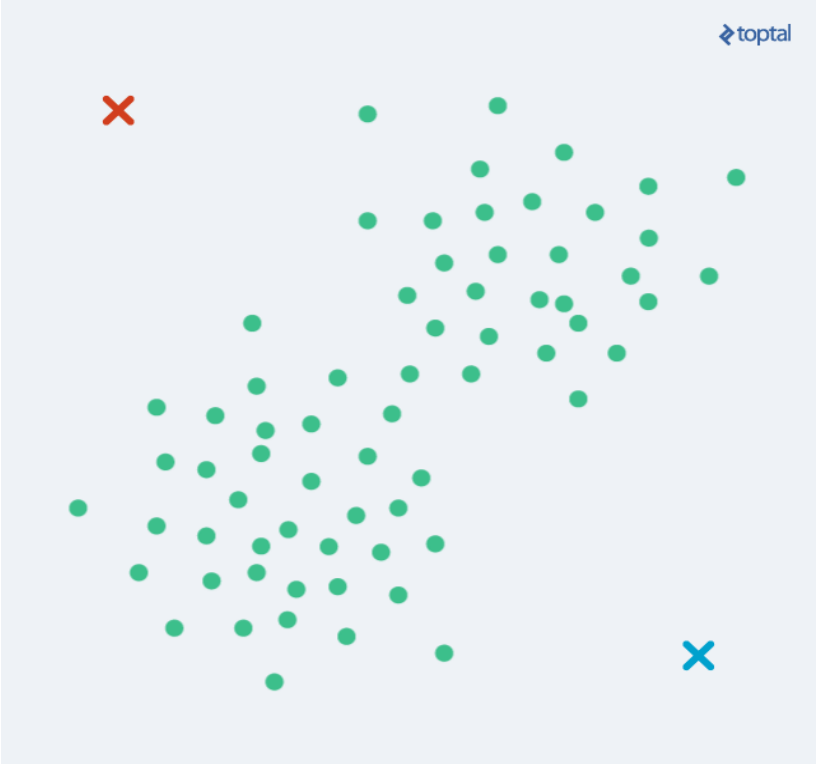
2.2.2 Thuật toán K-Means và ứng dụng

Có rất nhiều cách để phân cụm dữ liệu, chúng ta có thể dùng một số phương pháp như phân cụm phân cấp, phân cụm phân hoạch…..

Tuy nhiên trong nghiên cứu này tôi sử dụng học máy để giải quyết bài toàn phân cụm dữ liệu và một trong những thuật toán đó là thuật toán học máy không giám sát K-Means, chúng ta sẽ cùng tìm hiểu về thuật toán K-Means và ứng dụng của thuật toán.

Theo [20] Thuật toán phân cụm K-Means do MacQueen đề xuất trong lĩnh vực thống kê năm 1967. Thuật toán dựa trên độ đo khoảng cách của các

đối tượng dữ liệu trong cụm. Trong thực tế, nó đo khoảng cách tới các giá trị trung bình của các dữ liệu trong cụm. Nó được xem như là trung tâm của cụm. Như vậy, nó cần khởi tạo một tập trung tâm các trung tâm cụm ban đầu và thông qua đó nó lặp lại các bước gồm gán mỗi đối tượng tới cụm mà trung tâm gần nhất và tính toán lại trung tâm của mỗi cụm trên cơ sở gán mới cho các đối tượng. Quá trình lặp này dừng khi các trung tâm cụm hội tụ.



Hình 2.6: Mô phỏng tâm của các cụm được tính toán trong thuật toán K-Means

Để hiểu rõ ràng hơn về thuật toán K-Means tôi sẽ trình bày cụ thể về bài toán như sau:

Phát biểu bài toán:

Input

Tập các đối tượng X = {xi| i = 1, 2, …, N}, $x_i \in R^d$

Số cụm: K

Output

Các cụm Ci (i = 1 ÷ K) tách rời và hàm tiêu chuẩn E đạt giá trị tối thiểu.

Thuật toán hoạt động trên 1 tập vectơ d chiều, tập dữ liệu X gồm N phần tử:

$$X = \{xi \mid i = 1, 2, \ldots, N\}$$

K-Mean lặp lại nhiều lần quá trình:

Gán dữ liệu.

Cập nhật lại vị trí trọng tâm.

Quá trình lặp dừng lại khi trọng tâm hội tụ và mỗi đối tượng là 1 bộ phận của 1 cụm.

Hàm đo độ tương tự sử dụng khoảng cách Euclidean:

$$E = \sum_{i=1}^{N} \sum_{x_i \in C_j} (\| x_i - c_j \|^2)$$

trong đó cj là trọng tâm của cụm Cj

Hàm trên không âm, giảm khi có 1 sự thay đổi trong 1 trong 2 bước: gán dữ liệu và định lại vị trí tâm.

Các bước của thuật toán:

Bước 1: Khởi tạo

Chọn K trọng tâm {ci} (i = 1÷K).

Bước 2: Tính toán khoảng cách

$$S_i^{(t)} = x_j : \| x_j - c_i^{(t)} \| \leq \| x_j - c_{i^*}^{(t)} \| \text{ for all } = 1, \ldots, k\}$$

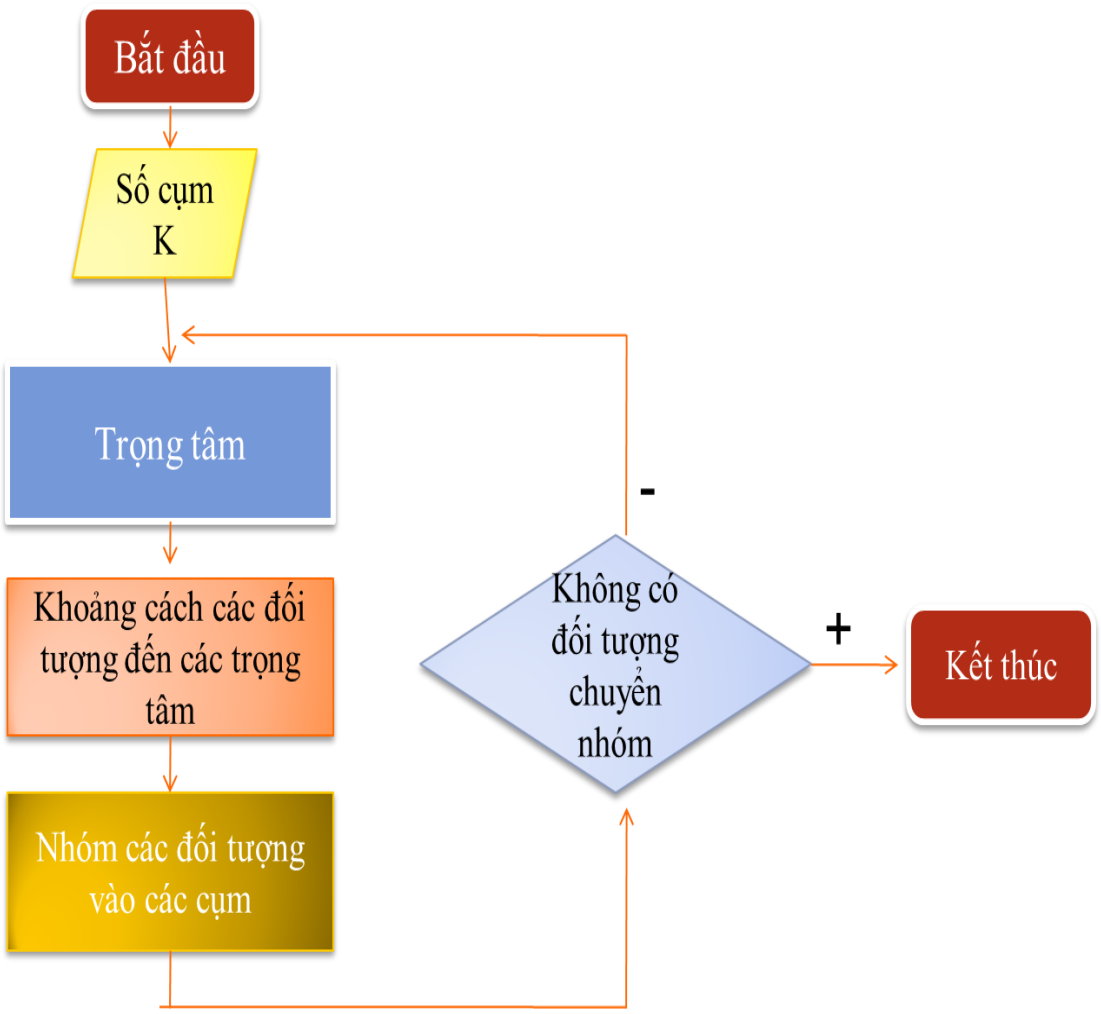
Bước 3: Cập nhật lại trọng tâm

$$c_i^{(t+1)} = \frac{1}{| S_i^{(t)} |} \sum_{x_j \in S_i^{(t)}} x_j$$

Bước 4: Điều kiện dừng

Lặp lại các bước 2 và 3 cho tới khi không có sự thay đổi trọng tâm của cụm.

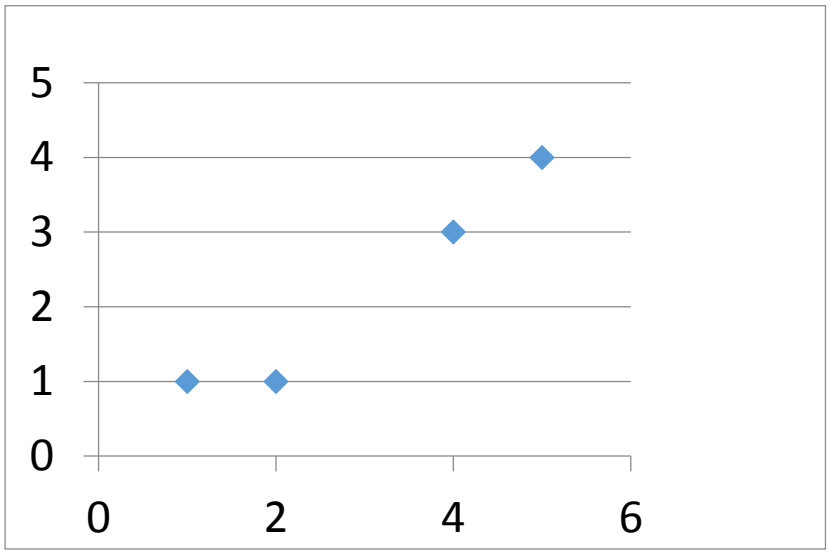
Thuật toán K-Means được mô tả như sau:



Hình 2.7: Mô tả thuật toán K-Means

Ví dụ minh họa về thuật toán K-Means:

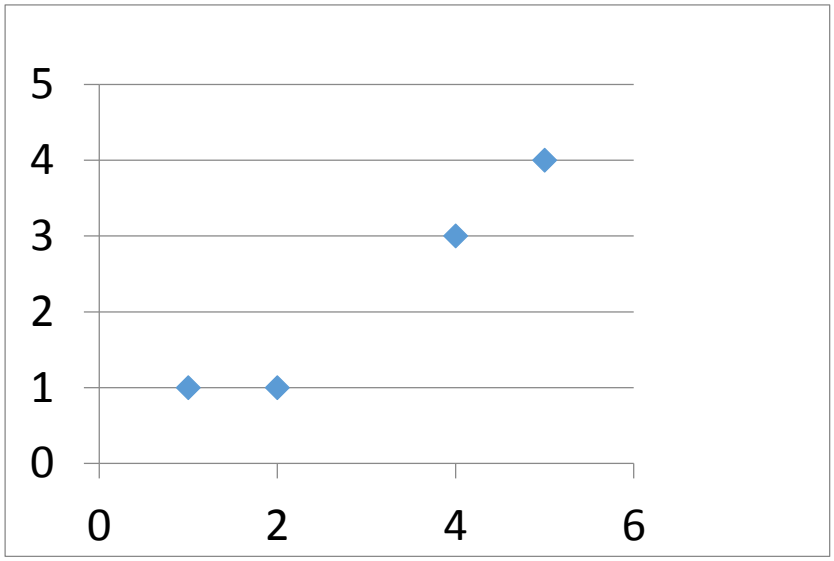
Đối tượng	Thuộc tính 1 (X)	Thuộc tính 2 (Y)
A	1	1
B	2	1
C	4	3
D	5	4



Bước 1: Khởi tạo

Chọn 2 trọng tâm ban đầu:

c1(1,1) ≡ A và c2(2,1) ≡ B, thuộc 2 cụm 1 và 2



Bước 2: Tính toán khoảng cách

$d(C, c1) = (4-1)^2 + (3-1)^2$

$= 13$

$d(C, c2) = (4-2)^2 + (3-1)^2$

$= 8$

$d(C, c1) > d(C, c2) \longrightarrow$ C thuộc cụm 2

$d(D, c1) = (5-1)^2 + (4-1)^2$

$= 25$

$d(D, c2) = (5-2)^2 + (4-1)^2$

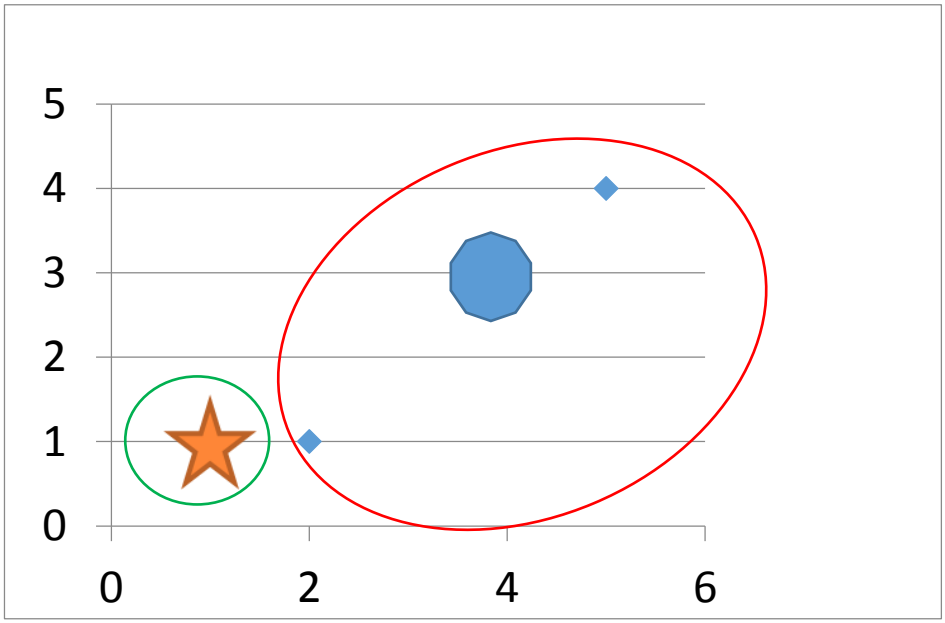
$= 18$

$d(D,c1) > d(D, c2) \longrightarrow$ D thuộc cụm 2

Bước 3: Cập nhật lại vị trí trọng tâm

Trọng tâm cụm 1 c1 ≡ A (1, 1)

Trọng tâm cụm 2 c2 $(x,y) = (\frac{2+4+5}{3}, \frac{1+3+4}{3})$



Bước 4-1: Lặp lại bước 2 – Tính toán khoảng cách

d(A, c1) = 0 < d(A, c2) = 9.89

A thuộc cụm 1

d(B, c1) = 1 < d(B, c2) = 5.56

B thuộc cụm 1

d(C, c1) = 13 > d(C, c2) = 0.22

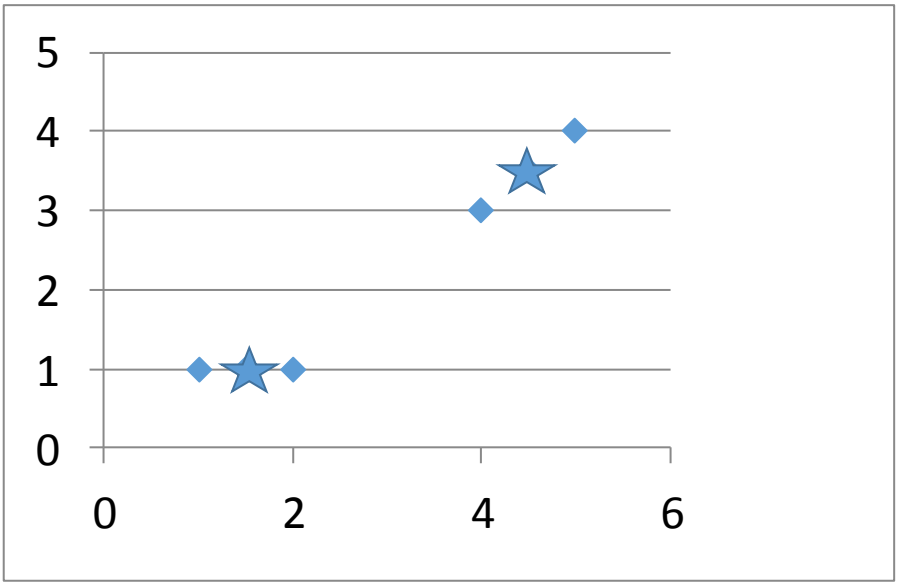
C thuộc cụm 2

d(D, c1) = 25 > d(D, c2) = 3.56

D thuộc cụm 2

Bước 4-2: Lặp lại bước 3-Cập nhật trọng tâm

c1 = (3/2, 1) và c2 = (9/2, 7/2)



Bước 4-3: Lặp lại bước 2

d(A, c1) = 0.25 < d(A, c2) = 18.5

A thuộc cụm 1

d(B, c1) = 0.25 < d(B, c2) = 12.5

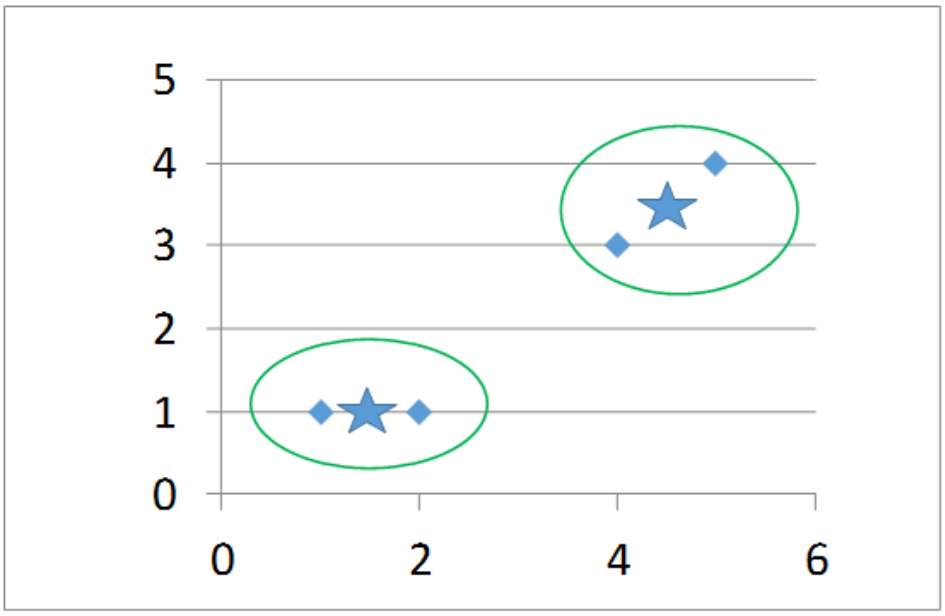
B thuộc cụm 1

d(C, c1) = 10.25 < d(C, c2) = 0.5

C thuộc cụm 2

d(D, c1) = 21.25 > d(D, c2) = 0.5

D thuộc cụm 2



Độ phức tạp: O($K.N.l$) với l: số lần lặp

Có khả năng mở rộng, có thể dễ dàng sửa đổi với những dữ liệu mới.

Bảo đảm hội tụ sau 1 số bước lặp hữu hạn.

Luôn có K cụm dữ liệu

Luôn có ít nhất 1 điểm dữ liệu trong 1 cụm dữ liệu.

Các cụm không phân cấp và không bị chồng chéo dữ liệu lên nhau.

Mọi thành viên của 1 cụm là gần với chính cụm đó hơn bất cứ 1 cụm nào khác.

CHƯƠNG 3
MÔ PHỎNG HỆ THỐNG GỢI Ý THÔNG TIN TRONG THƯƠNG MẠI ĐIỆN TỬ

3.1 Hướng tiếp cận và kiến trúc hệ thống

3.1.1 Hướng tiếp cận

Ngày nay, việc đưa Machine Learning vào trong vào trong các hệ thống thông tin thông minh ngày càng được phổ biến rộng rãi vì Machine Learning có những ưu điểm vượt bậc mà các giải pháp thông thường không giải quyết được.

Đối với những hệ thống thương mại điện tử bình thường, chúng ta chỉ có thể tìm kiếm những sản phẩm mà ta cần theo danh mục các sản phẩm đã có sẵn trong hệ thống điều đó đồng nghĩa với việc nếu hệ thống có ít sản phẩm thì người dùng có thể tìm kiếm ngay ra sản phẩm mình cần tìm nhưng với hệ thống có rất nhiều sản phẩm giống nhau và có những thuộc tính giống nhau tuy nhiên trong các thuộc tính đó lại có những sản phẩm phù hợp với người dùng hơn thì những hệ thống thương mại điện tử bình thường không thể đưa ra một các nhanh chóng cho người dùng hoặc người dùng phải mất rất nhiều thời gian để tìm kiếm được những gì mình cần.

Điển hình là bài toán gợi ý cho người dùng các đối tượng mà họ có thể quan tâm. Chẳng hạn ơ những ứng dụng thương mại điện tử thông thường, việc gợi ý này chỉ dựa trên một vài tiêu chí nhất định như nhóm đối tượng, mức giá, mầu sắc…

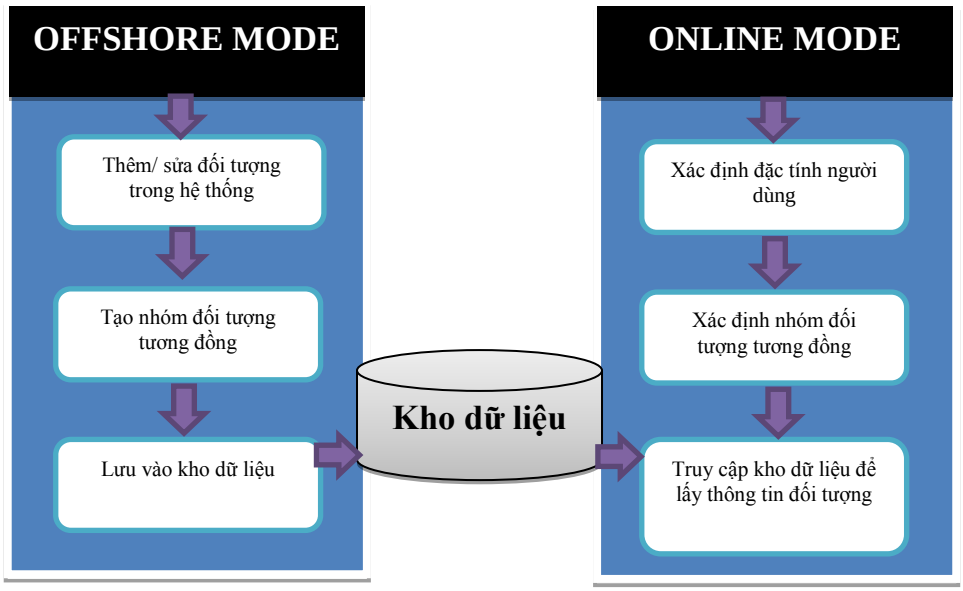
Hình 3.1: Gợi gợi ý đối tượng tương tự

Tuy nhiên, một đối tượng sẽ có rất nhiều thuộc tính khác nhau, chỉ dựa vào một vài thuộc tính sẽ không đánh giá hết được độ tương đồng của các đối tượng. Hơn nữa, khi lượng dữ liệu trở nên vô cùng lớn, việc truy vấn theo nhiều điều kiện phức tạp sẽ không đảm bảo được hiệu năng hệ thống. Theo khảo sát của Amazon, đa số người dùng trả lời rằng họ không muốn tiếp tục khám phá ứng dụng thương mại điện tử nếu thời gian phản hồi của ứng dụng quá chậm.

Để giải quyết được cấn vấn đề này, chúng ta phải xây dựng trước được mối liên hệ giữa các đối tượng tương tự nhau dựa trên các thuộc tính của chúng và thực hiện đánh chỉ mục trước khi thực hiện truy vấn.

3.1.2 Kiến trúc hệ thống

Trong chương này tôi sẽ trình bày một mô hình tổng thể cho việc lưu vết hành vi của người dùng với một hệ thống thông tin để đưa ra gợi ý đối tượng thông tin mà người sử dụng quan tâm, và một giải pháp sử dụng học máy nhằm đưa các đối tượng tương đồng vào vào các nhóm cụ thể dựa theo tập thuộc tính của các đối tượng.



Hình 3.2: Sơ đồ luồng hệ thống

Tại OffShore mode:

Đây là chế độ dành cho người quản trị. Mỗi khi thêm hoặc sửa dữ liệu về đặc tính của đối tượng, hệ thống sẽ thực hiện đưa đối tượng này vào đúng nhóm mà chúng thuộc về.

Sau khi đã thiết lập nhóm cho các đối tượng, toàn bộ dữ liệu này sẽ được đánh chỉ mục và lưu vào kho dữ liệu.

Tôi sử dụng học máy trong quá trình phân nhóm đối tượng. Có 3 phương pháp học máy là:

- *Học có giám sát*
- *Học không giám sát*
- *Học bán giám sát*

Nếu sử dụng phương pháp học có giám sát, yêu cầu phải có tập dữ liệu đã được gán nhãn là rất lớn thì mới cho hiệu quả cao. Vấn đề là thật khó để định nghĩa các nhãn lớp cụ thể cho từng đối tượng, hơn nữa, quá trình này tốn rất nhiều công sức thủ công.

Do không có nhiều dữ liệu đã được gán nhãn lớp và việc gán nhãn lớp trong trường hợp này là không có tiêu chí cụ thể, vì thế, tôi sự dụng phương pháp học không giám sát bằng thuật toán Kmean để thực hiện phân nhóm cho các đối tượng.

Trong quá trình vận hành hệ thống, tôi đã thêm một module lưu lại các đối tượng mà cùng một người dùng lựa chọn để xem thông tin hoặc like đối tượng. Sau đó, để tăng độ chính xác cho quá trình phân nhóm, tôi sử dụng học máy để huấn luyện ra hàm khoảng cách và dùng hàm khoảng cách này để phân nhóm cho toàn bộ các đối tượng trong ứng dụng.

Như vậy, trong luận văn này áp dụng học bán giám sát bằng cách kết hợp học không giám sát và huấn luyện hàm khoảng cách bằng học có giám sát dựa trên tập dữ liệu ít ỏi được phân nhóm có sẵn.

Tại Online mode:

Khi người dùng truy cập vào ứng dụng thương mại điện tử, hệ thống sẽ xem xét đó có phải là người dùng đã từng truy cập vào ứng dụng hay là một người dùng hoàn toàn mới.

- Nếu là người dùng mới và ứng dụng chưa có thông tin gì về người dùng này thì sẽ liệt kê ngẫu nhiên các đối tượng thuộc các nhóm khác nhau để gợi ý cho người dùng.
- Nếu trước đây người dùng đã từng truy cập vào ứng dụng thì sẽ tìm cách liệt kê các đối tượng dựa theo thông tin mà trước đó người dùng đã để lại.

Một người dùng khi truy cập ứng dụng sẽ có các hành động: xem, like hoặc đặt hàng đối tượng. Như vậy, tôi đưa ra cách thức lưu vết người dùng qua cách hành động của họ như sau:

Với mỗi phiên truy cập, hệ thống sẽ lưu lại IP của người dùng như là một định danh (ID) của người dùng. Khi người dùng thực hiện đăng kí tài khoản, IP đó sẽ chuyển thành ID thực sự của người dùng. Mỗi khi người dùng chọn xem đối tượng, hệ thống sẽ lưu lại thông tin vào bảng sau:

Bảng 3.1: Mô tả cấu trúc bảng lưu trữ hành vi người sử dụng

IP hoặc ID	Định danh người dùng
Mã đối tượng	Mã đối tượng trên ứng dụng
View	Lượt view, x1 (nhân 1) mỗi khi người dùng view đối tượng đó
Like	x3 (nhân 3) nếu người dùng thích

Trong đó, cặp **IP/ID** và **Mã đối tượng** là cặp khóa duy nhất trong bảng. Mỗi đối tượng tương ứng với **Mã đối tượng** đã được phân lớp từ trước. Từ vết khách hàng này, hệ thống sẽ tính ra lớp đối tượng mà người dùng thích, sau đó sẽ liệt kê các đối tượng trong lớp này làm mục tiêu gợi ý

cho khách hàng.

Công thức tính như sau:

$$Point\ of\ P_i = Hit_i + 3*Like_i)$$

$$P = Max(Point\ of\ P_i)$$

Giả sử có P đối tượng người dùng đã từng View hoặc Vote, ta lấy được đối tượng Pi có lượng View + Like lớn nhất. Lớp đối tượng người dùng thích chính là lớp của đối tượng thứ Pi.

Ví dụ:

Bảng 3.2: Ví dụ lưu trữ hành vi người sử dụng

IP/ ID	Mã đối tượng	View	Like
IP1	P1	1	1
IP1	P2	2	0
IP1	P3	3	0

Như bảng trên, ta tính $Point\ of\ P_i$ như sau:

$Point\ of\ P_1 = 1 + 3 * 1 = 4;$

$Point\ of\ P_2 = 2 + 3 * 0 = 2;$

$Point\ of\ P_3 = 3 + 3 * 0 = 3;$

Vậy, đối với người dùng có IP là *IP1* thì đối tượng *P1* của điểm cao nhất, do đó hệ thống sẽ liệt kê các đối tượng thuộc class của *P1*

3.2 Thiết kế và cài đặt chi tiết các thành phần hệ thống

3.2.1 Phân nhóm đối tượng bằng phương pháp học bán giám sát

Trong luận văn này tôi sử dụng thuật toán Kmean để phân cụm, trong đó, sử dụng phương pháp học máy để xây dựng hàm khoảng cách.

a. Lựa chọn số cụm

Làm cách nào mà ta chọn ra được số cụm (k) thích hợp?

Nếu lựa chọn nhiều cụm, dữ liệu sẽ được chia nhỏ ra, và giá trị *error*

(tổng khoảng cách) cũng sẽ nhỏ hơn. Vậy, có thể chọn $k = m$ (số đối tượng), như thế mỗi đối tượng sẽ trở thành tâm của chính nó và mỗi cụm sẽ chỉ có 1 điểm.

Rõ ràng rằng điều đó là không sai vì *error* sẽ bằng 0, nhưng ta sẽ không thể tìm được mô tả đơn giản cho dữ liệu, và mô hình thu được cũng gặp vấn đề *overfitting* và không thể phủ được những điểm mới thêm vào.

Để giải quyết vấn đề này là bổ sung thêm hàm phạt (*penalty*) cho số lượng cụm. Từ đó, mục tiêu lúc này không chỉ còn giảm thiểu error, mà phải cân bằng cả *error* + *penalty*. Giá trị *error* sẽ tiến dần tới 0 khi tăng số lượng cụm, nhưng đồng thời *penalty* cũng tăng theo. Quá trình tăng số lượng cụm sẽ dừng lại khi mà lượng *error* giảm đi thấp hơn so với giá trị *penalty*, và kết quả thu được là kết quả tối ưu.

b. Xây dựng tập dữ liệu mẫu

Quá trình xây dựng tập mẫu được thực hiện một cách thủ công hoàn toàn dựa trên bộ dữ liệu là các đối tượng ở trong kho dữ liệu. Trên kho dữ liệu đó ta chọn ra k xu hướng đối tượng, thiết lập n thuộc tính điển hình cho mỗi k xu hướng đó. Tập thuộc tính điển hình này sẽ là tâm của mỗi cụm.

Tiếp đó, lựa chọn các đối tượng tương đồng trong kho dữ liệu đối với bộ dữ liệu ở tâm của mỗi cụm. Việc lựa chọn này yêu cầu kinh nghiệm chuyên nghành về lĩnh vực thông tin mà hệ thống đang thao tác. Ví dụ đối với các đối tượng là đồng hồ đeo tay, việc lựa chọn các đối tượng tương đồng phải dựa trên kinh nghiệm của người làm trong lĩnh vực marketing và nghiên cứu thị trường.

Trong luận văn này, tôi sử dụng tập mẫu là 30 chiếc đồng hồ được lựa chọn từ kho dữ liệu gồm 300 chiếc đồng hồ khác nhau và phân vào 6 cụm được gán nhãn từ 1 đến 6.

Việc lựa chọn tập mẫu cũng cần được sự trợ giúp từ hành vi người

sử dụng. Ví dụ, nếu một người xem đối tượng x mà luôn có xu thế xem thêm đối tượng y thì nhiều khả năng x và y sẽ thuộc chung vào một cụm

c. Quy đổi miền không gian giá trị của các thuộc tính

Một đối tượng có rất nhiều thông tin thuộc tính. Các thuộc tính này có miền giá trị rất khác nhau. Ví dụ với một chiếc đồng hồ:

Đường kính mặt: 50mm

Độ sâu ngập nước: 2m

Dễ dàng thấy rằng mặc dù 2 thuộc tính có độ quan trọng gần tương đương, tuy nhiên miền giá trị của chúng lại rất khác nhau. Điều này sẽ dẫn đến quá trình thực hiện thuật toán Kmean sẽ không chính xác bởi không gian n thuộc tính đã bị méo.

Để tránh bị méo miền không gian, tôi biến đổi các giá trị thuộc tính về miền [0,1] tương ứng.

Đối với các thuộc tính mà giá trị là số thực:

Giả sử thuộc tính của đối tượng thứ i là X_i (X lớn, là giá trị trước khi biến đổi của thuộc tính x), ta sẽ tính max của thuộc tính x trên N đối tượng cần phân cụm.

$$max = Max(X_i) \mid i = 1..N$$

Giá trị sau khi biến đổi của đối tượng thứ i là (quy định *x* nhỏ là giá trị sau khi biến đổi của thuộc tính x):

$$xi = \frac{Xi}{max} \; với \; i = 1..N$$

Đối với giá trị là các chuỗi:

Việc quy đổi một chuỗi thành số thực mà vẫn đảm bảo không gian thuộc tính không bị méo là rất khó khan, vì vậy, tôi sử dụng phân bố đều rời rạc trong khoảng [0, 1] cho các giá trị chuỗi này.

Ví dụ toàn bộ các giá trị khác nhau của thuộc tính x trên tập đối tượng cần phân lớp có giá trị là: “A”, “B”, “C”. Tiến hành phẩn bổ trên

khoảng [0, 1] sẽ được các giá trị số thực tương ứng như sau:

"A" – 0

"B" – 0.5

"C" – 1

Lưu ý rằng việc này cũng sẽ đảm bảo các giá trị giống nhau sẽ có cùng một vị trí trên trục tọa độ biểu diễn chiều thuộc tính x.

d. Xây dựng hàm khoảng cách

Một trong các yếu tố quan trọng nhất trong phân cụm đó là hàm khoảng cách. Có rất nhiều hàm khoảng cách có thể được sử dụng trong việc phân cụm dữ liệu như bảng mô tả dưới đây:

Bảng 3.3: Các hàm khoảng cách

Distance	Description
ArgMax	ArgMax distance (L0) distance.
Bhattacharyya	Bhattacharyya distance.
BrayCurtis	Bray-Curtis distance.
Canberra	Canberra distance.
Chebyshev	Chebyshev distance.
Cosine	Cosine distance.
Dice	Dice dissimilarity.
Euclidean	Euclidean distance metric.
Hamming	Hamming distance.

Hellinger	Herlinger distance.
Jaccard	Jaccard (Index) distance.
Kulczynski	Kulczynski dissimilarity.
Levenshtein	Levenshtein distance.
Mahalanobis	Mahalanobis distance.
Manhattan	Manhattan (also known as Taxicab or L1) distance.
Matching	Matching dissimilarity.
Minkowski	The Minkowski distance is a metric in a normed vector space which can be considered as a generalization of both the Euclidean distance and the Manhattan distance.
Modular	Modular distance (shortest distance between two marks on a circle).
PearsonCorrelation	Pearson Correlation similarity.
RogersTanimoto	Rogers-Tanimoto dissimilarity.
RusselRao	Russel-Rao dissimilarity.
SokalMichener	Sokal-Michener dissimilarity.
SokalSneath	Sokal-Sneath dissimilarity.

SquareEuclidean	Square-Euclidean distance and similarity. Please note that this distance is not a metric as it doesn't obey the triangle inequality.
SquareMahalanobis	Squared Mahalanobis distance.
Yule	Yule dissimilarity.

Hiển nhiên, với những điểm nằm trong không gian, khoảng cách Euclid rõ ràng là hiệu quả nhất, nhưng trong trường hợp này phải cần thêm vài thủ thuật cho những loại dữ liệu đặc trưng khác nhau. Có rất nhiều hàm khoảng cách phù hợp, việc này yêu cầu khá nhiều kiến thức chuyên ngành liên quan tới dữ liệu đó.

Do đó, tôi nhờ tới sự trợ giúp của Học máy để huấn luyện ra hàm khoảng cách thích hợp nhất. Hơn nữa, tôi đã có một tập dữ liệu đã được gán nhãn từ trước.

Với mỗi đối tượng, tôi chọn ra n đặc trưng có ảnh hưởng lớn nhất đến tính chất của đối tượng. n đặc trưng này tạo ra một không gian *vector* thuộc tính n chiều.

3.2.2 Huấn luyện mạng nơ ron để xây dựng hàm khoảng cách

Xác định cấu trúc mạng:

- Mạng noron được xây dựng theo phương pháp học có giám sát.

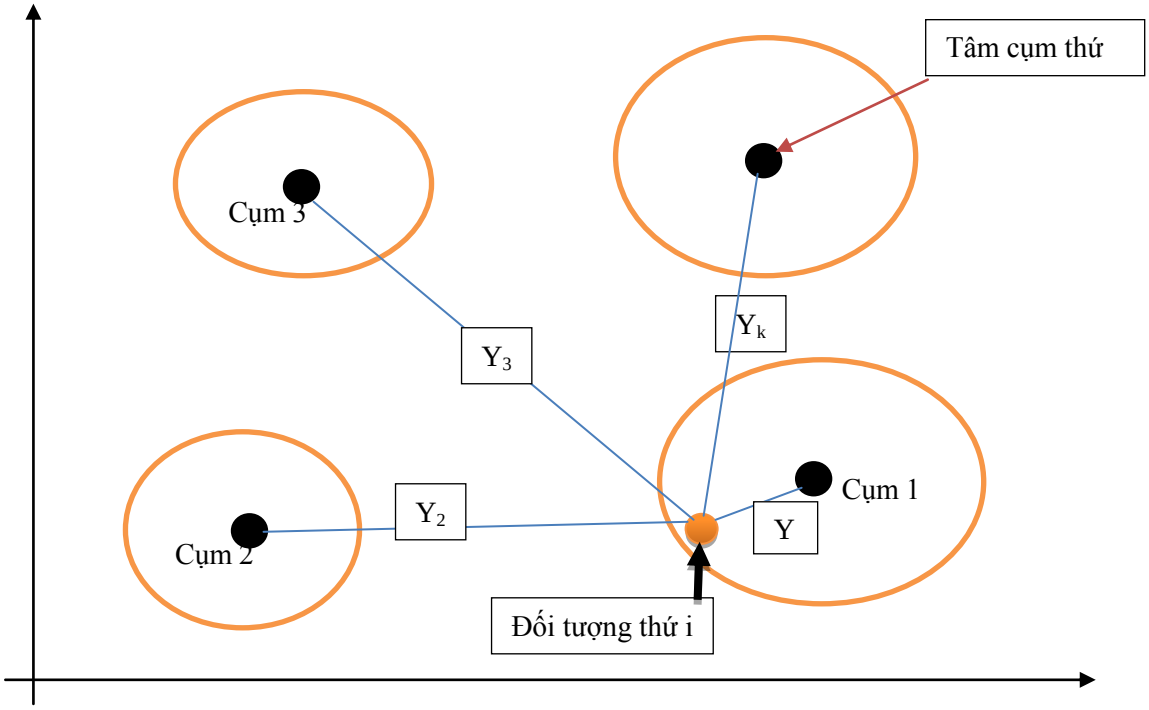
Bài toán lựa chọn mạng Feed-forward 3 lớp với cấu trúc như sau :

Số noron lớp đầu vào : n noron (tương ứng với số chiều của vector thuộc tính)

Số noron tầng ẩn: 500 noron. Con số 500 được turning dựa trên các lần thử sau mỗi quá trình học.

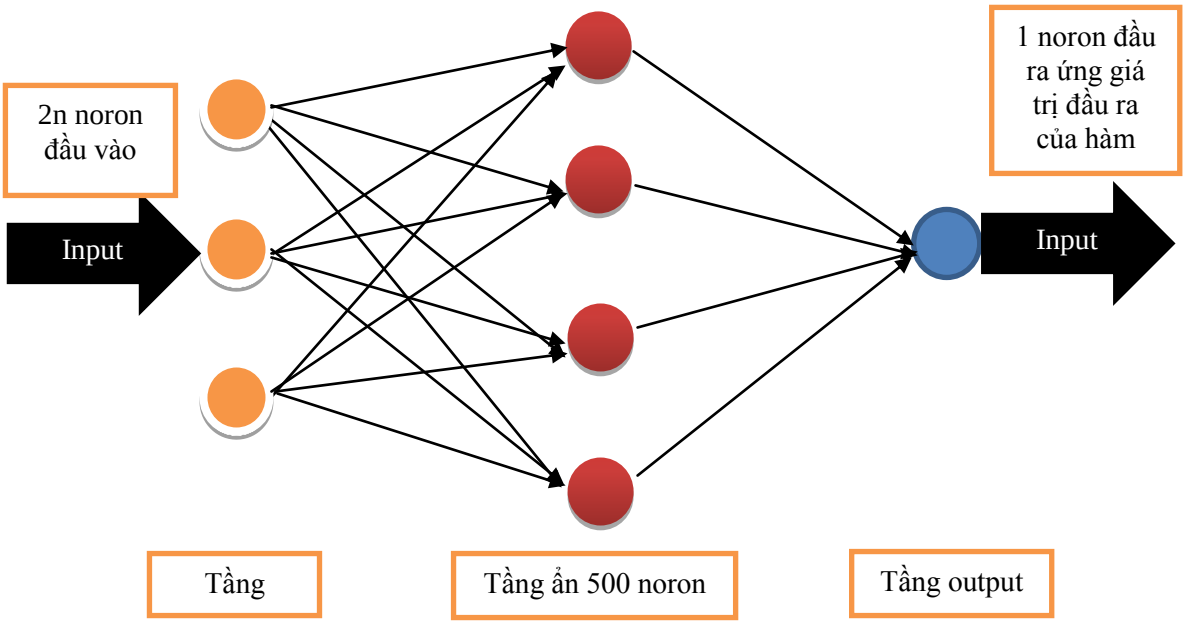
Số noron tầng đầu ra: 1 noron. Tương ứng với khoảng cách của mỗi đối tượng đến cụm thứ k trong bộ dữ liệu huấn luyện.

Trong hình dưới đây biểu diễn khoảng cách của một đối tượng thứ i tới tâm của k cụm đã được phân nhóm từ trước.



Hình 3.3: Mô hình khoảng cách đến tâm cụm của tập dữ liệu mẫu

Việc huấn luyện cho mạng học là một vòng lặp duyệt qua lần lượt các đối tượng giúp mạng noron nhớ các khoảng cách đến các tâm cụm.Với mỗi vòng lặp, một đối tượng sẽ được đưa vào giảng dạy cho mạng noron học.



Hình 3.4: Mô hình mạng nơ ron để huấn luyện hàm khoảng cách

Giá trị đầu ra của mỗi nơ ron được cho bởi công thức:

$$y_i = f(net_i - \theta_i) \text{ và } net_i = \sum_{j=1}^{n} w_{ij} x_j$$

trong đó: x_1, x_2, …x_m là các tín hiệu đầu vào tương ứng với giá trị của vector thuộc tính cho mỗi đối tượng, còn w_{i1}, w_{i2},…,w_{im} là các trọng số kết nối của nơron thứ i, net_i là hàm tổng, f là hàm truyền, θ_i là một ngưỡng, y_i là tín hiệu đầu ra của nơron.

Đầu ra

Thuật toán huấn luyện mạng:

Mạng feed- forward sử dụng giải thuật lan truyền ngược sai số Back Propagation.

Giải thuật gồm 2 bước :

Bước 1:

-Lan truyền xuôi đầu vào qua mạng.

Sử dụng các công thức :

Công thức chung tính đầu ra của một noron thứ i tại lớp thứ L:

$$y_{i=} \quad f\left(\sum_{j=0}^{n} w_{ij}x_j - b_i\right)$$

Với f là hàm chuyển sigmoid lưỡng cực được tính theo công thức :

$$f=\frac{2}{1-e^{-\alpha t}}-1$$

α: hệ số góc của hàm chuyển

t: biến net-input

b_i: hệ số ngưỡng hay độ lệch

Áp dụng đối với mô hình mạng của chương trình :

-Công thức cho đầu ra của một noron thứ i (1→500) tại lớp ẩn

$$a_{i=} f\left(\sum_{j=1}^{n} w_{ij}x_j - b_i\right)$$

Với w_{ij} : trọng số tại noron thứ i của lớp ẩn kết nối với đầu vào thứ j của lớp vào

x_j : giá trị đầu vào của noron thứ j tại lớp vào

b_i : giá trị ngưỡng hay độ lệch của noron thứ i của đầu vào

-Công thức cho đầu ra của một noron tại lớp output

$$y_{=} f\left(\sum_{j=1}^{n} w_j a_j - b\right)=f\left[\sum_{j=1}^{n} w_j\left(f\left(\sum_{j=1}^{n} w_{ij}x_j\right)\right)\right]$$

Bước 2 : Lan truyền ngược

Tính toán sai lệch giữa đầu ra thực và đầu ra mong muốn của noron tại đầu ra. Đây chính là sai số của mạng ứng với mẫu học (Xs, Ts):

$$e_{=} t- y$$

Nếu e > ε thì:

Thông tin sai số sẽ được lan truyền ngược qua mạng để điều chỉnh lại trọng số tại vòng lặp L.

- Công thức điều chỉnh trọng số với liên kết giữa noron thứ j trong lớp ẩn và noron thứ i trong lớp ra tại lần lặp l+1: (l+1<số lần dạy(epochs)

$$w_{ij}(l+1) = w_{ij}(l) + \eta.e(l).y_j.f'(y_i(l))$$

Với η: hệ số học

E(l): giá trị sai lệch của noron thứ i trong lớp ra , trong lần dạy (lặp) thứ L.

f': đạo hàm của hàm chuyển lưỡng cực ,công thức $f'=\frac{1-x^2}{2}$

η : hệ số học

y(l): giá trị đầu ra của noron thứ i trong lớp ra tại vòng lặp thứ l.

$y_j(L)$: giá trị đầu ra của noron thứ j trong lớp ẩn tại vòng lặp thứ L.

- Công thức điều chỉnh trọng số với liên kết giữa noron vào thứ j và noron ẩn thứ i , tại lần lặp thứ l+L (l+L <epochs)

$$w_{ij}(l+1) = w_{ij}(l) + \eta.x_j.f'(y_i).\sum_{k=1}^{m} w_{ki}(l+1).e(l).f'(y_k(l))$$

Với :

η: hệ số học

x_j : giá trị đầu ra của noron thứ j trong lớp vào.

y_i : giá trị đầu ra của noron thứ i trong lớp ẩn

$w_i(l+L)$: trọng số liên kết giữa noron trong lớp ra và noron thứ i trong lớp ẩn trong lần lặp thứ l + L.

y(L) : giá trị đầu ra của noron thứ k trong lớp ra.

Các tham số sử dụng trong chương trình

Tốc độ học η = 150.

Hệ số góc α Sigmoid = 0.014.

Giá trị ngưỡng hay độ lệch: 0.2

Số lần dạy epochs = 600.

Ngưỡng của lỗi = 0.0002.

e. Thực hiện phân cụm các đối tượng

Quá trình huấn luyện hàm khoảng cách sẽ thu được 2 ma trận trọng số:

+ Ma trận hai chiều W1, có số dòng là 2n, tương ứng với 2n giá trị đầu vào của hàm khoảng cách

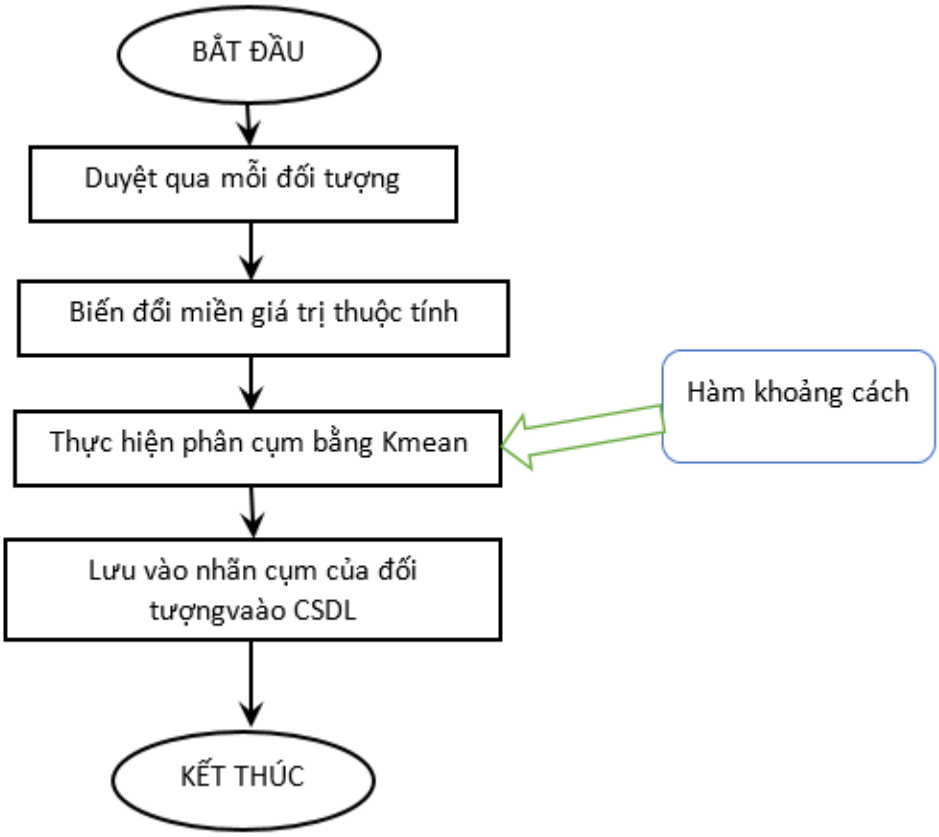
- n thuộc tính của một đối tượng
- n thuộc tính của tâm

$$n + n = 2n$$

+ Ma trận 1 chiều W2, có số dòng là H = 500, chính là số noron của tầng ẩn

Hàm khoảng cách chính là việc cho chạy các tín hiệu đầu vào qua mạng noron với 2 bộ trọng số này.

Các bước thực hiện phân cụm như sau:



Hình 3.5: Quá trình phân cụm các đối tượng

3.2.3 Đánh giá mức độ hiệu quả

Tôi tiến hành xây dựng một ứng dụng khảo sát người dùng nhằm đánh giá kết quả của mô hình như sau:

Ứng dụng bao gồm có 2 màn hình chính

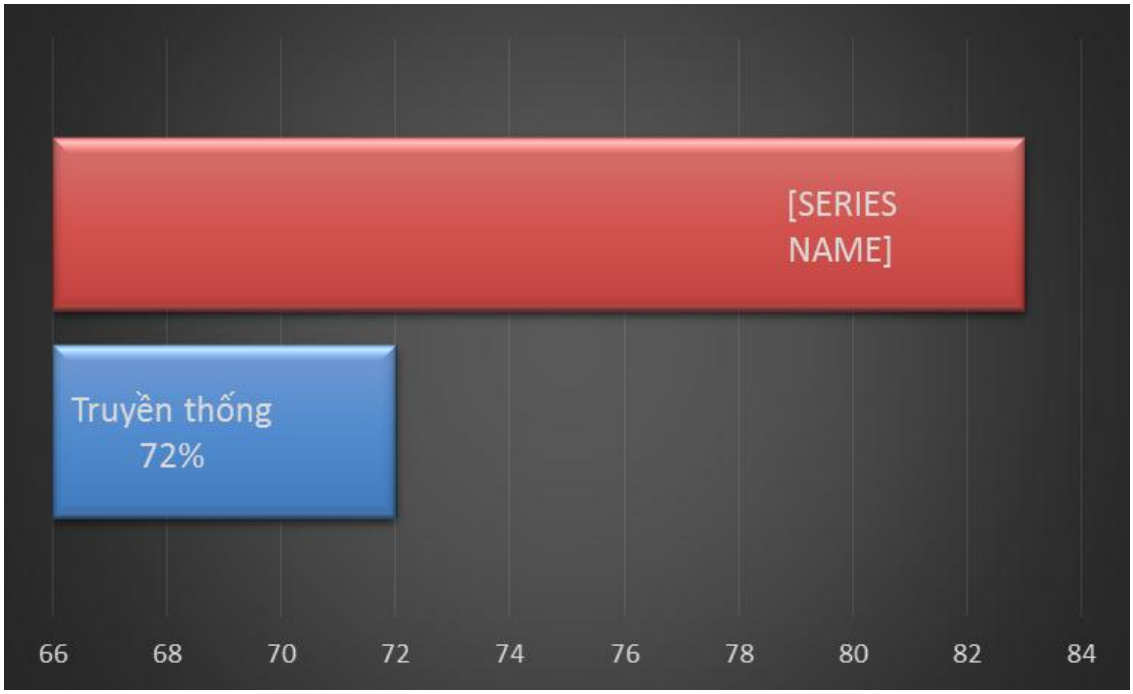
Màn hình đầu tiên sẽ lựa chọn ngẫu nhiên 5 sản phẩm thuộc mỗi cụm trong toàn bộ tập dữ liệu

Màn hình thứ 2: Sau người dùng chọn đối tượng ở màn hình đầu tiên, sẽ chuyển màn hình thứ 2 gồm:

- Thông tin đối tượng mà họ lựa chọn
- Danh sách các đối tượng trong cùng một cụm với đối tượng người dùng đã lựa chọn.

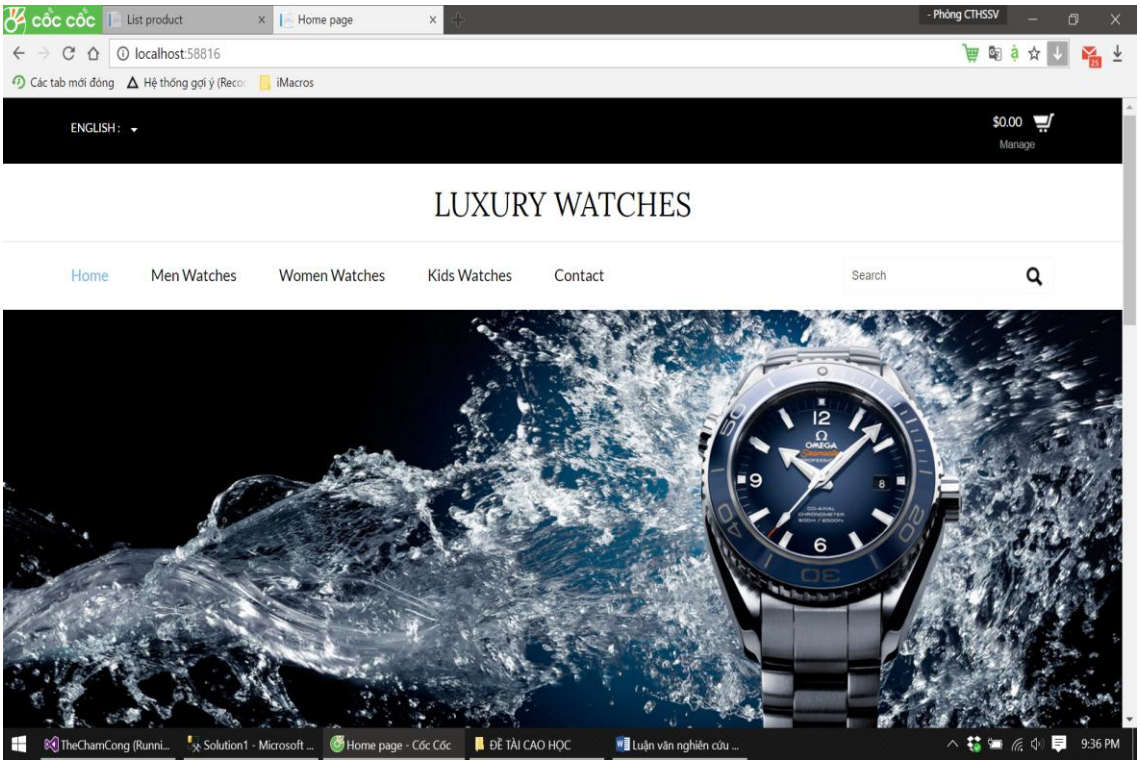
Kết quả cho thấy 83% người dùng sẽ click vào sản phẩm tương tự.

Cùng một phương pháp khảo sát này, nếu như tôi lựa chọn các sản phẩm tương tự bằng phương pháp truyền thống như truy vấn theo mầu sắc, chủng loại, giá cả.. thì kết quả thu được là 72%.

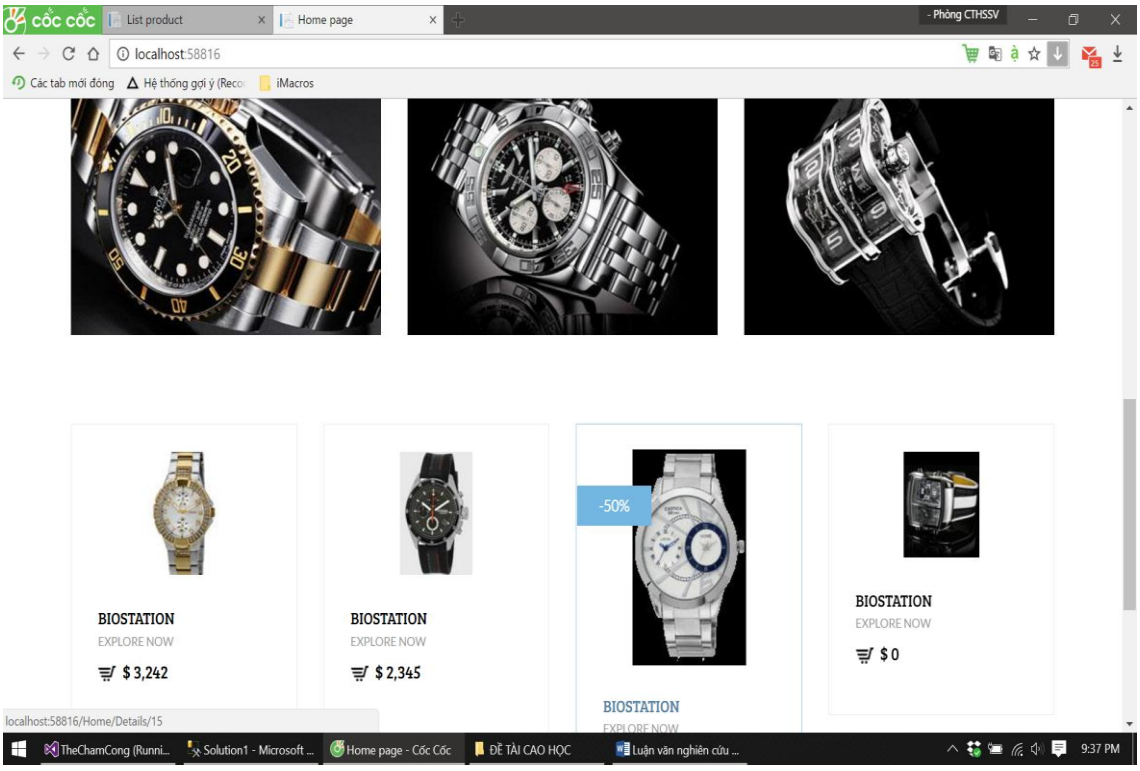


Hình 3.6: Đánh giá mức độ hiệu quả

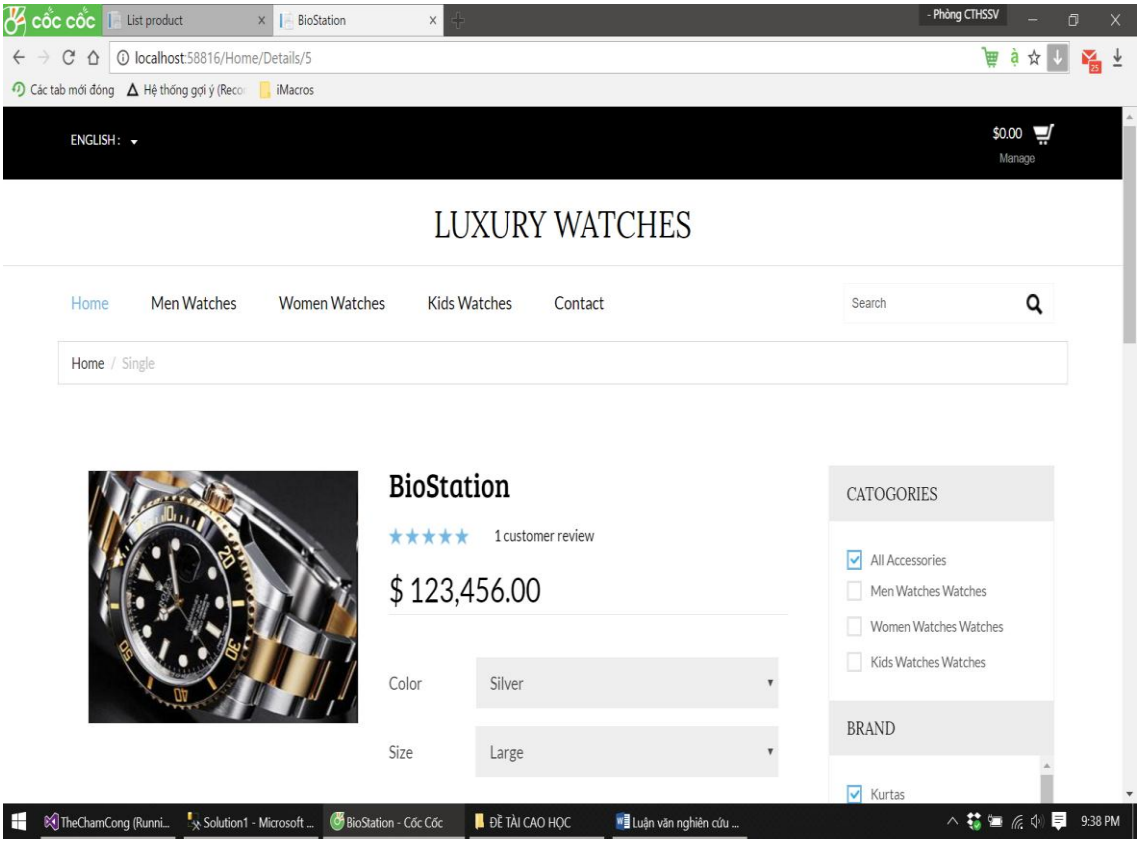
Một số hình ảnh demo hệ thống đã xây dựng:



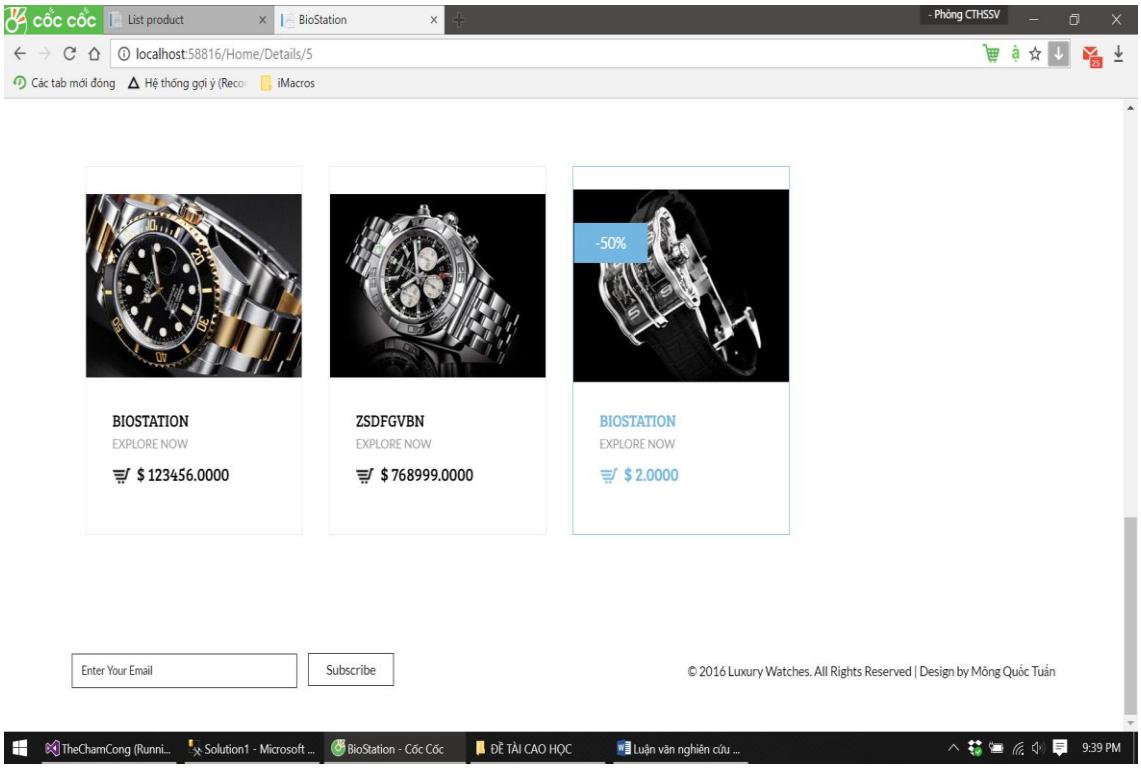
Hình 3.7: Giao diện tổng quan hệ thống khi truy cập



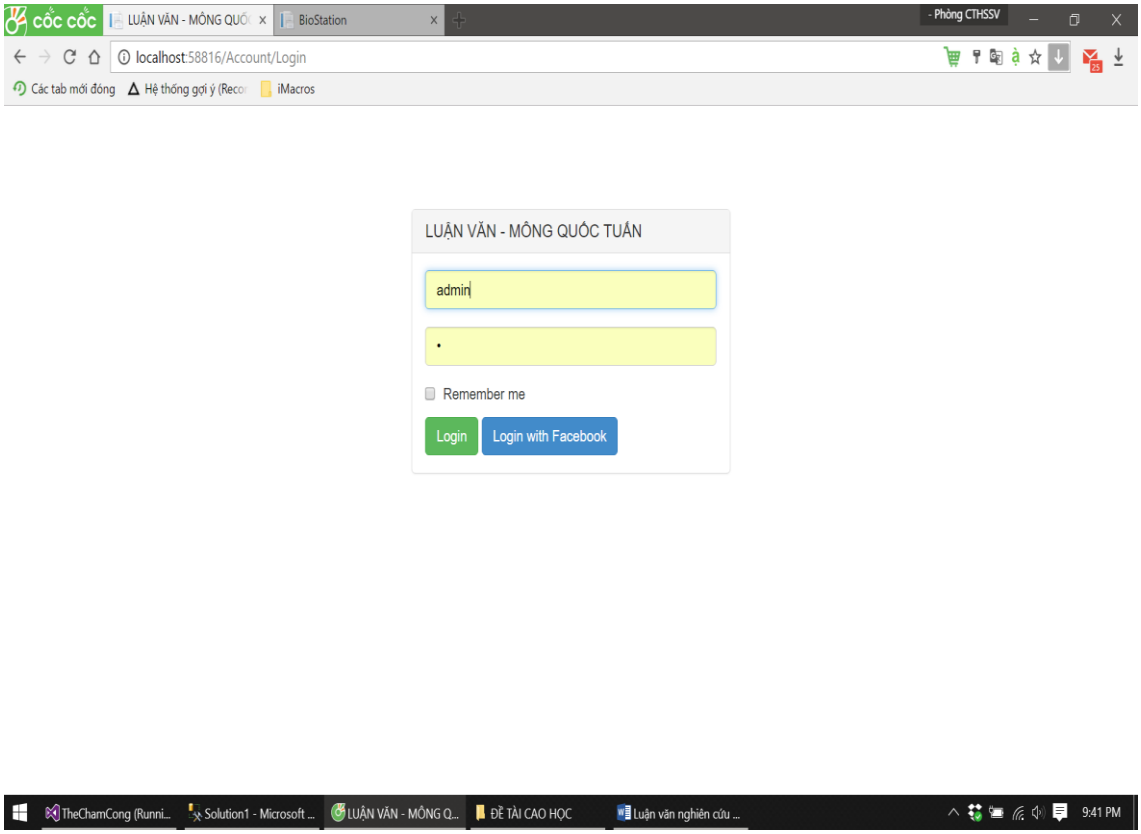
Hình 3.8: Giao diện tổng quan hệ thống khi ở trạng thái Online Mode



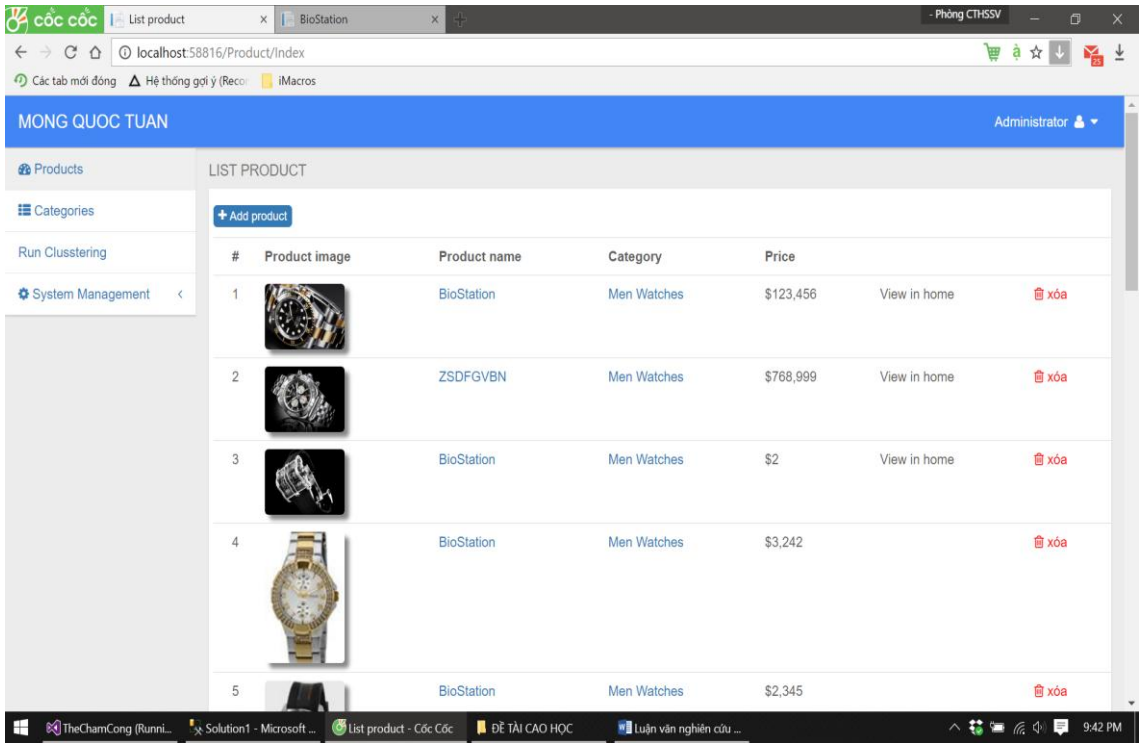
Hình 3.9: Giao diện chi tiết sản phẩm khi truy cập



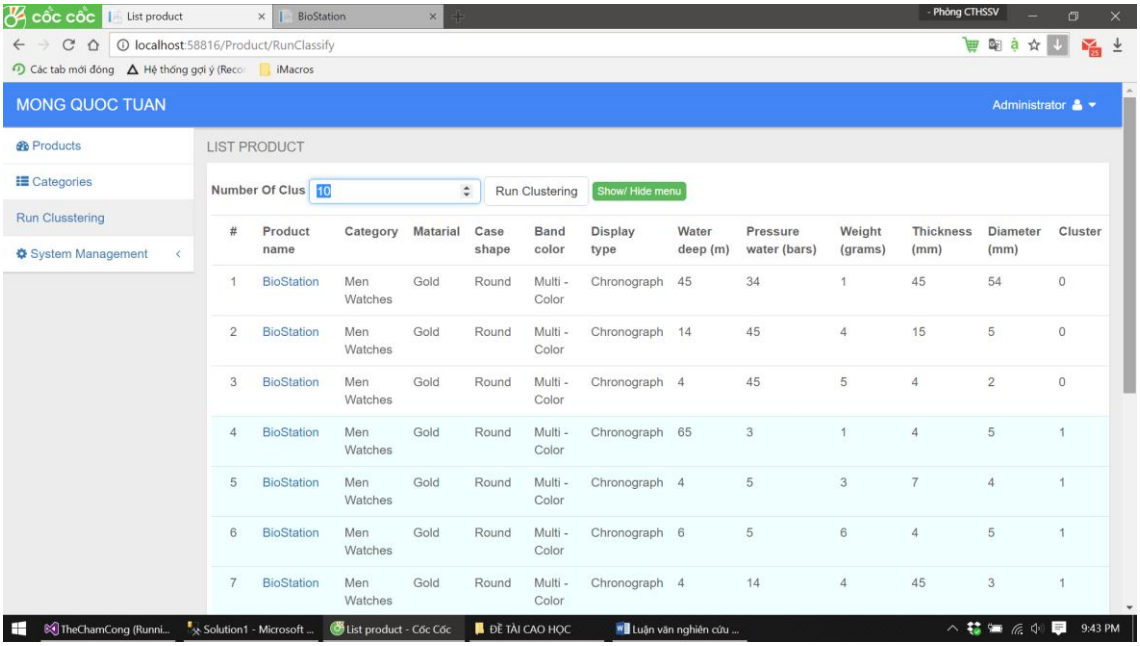
Hình 3.10: Những sản phẩm tương tự đã được gợi ý trong hệ thống



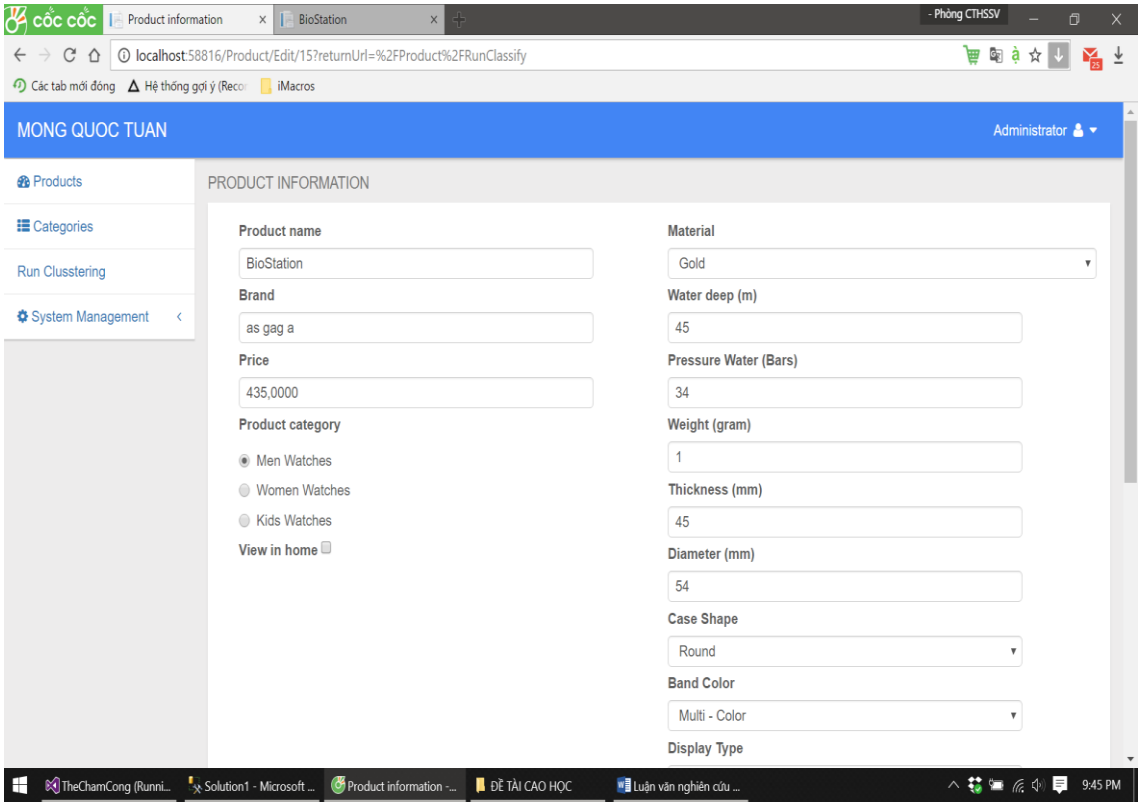
Hình 3.11: Đăng nhập vào Offshore mode trên hệ thống



Hình 3.12: Tổng quan hệ thống quản lý sản phẩm



Hình 3.13: Lựa chọn số cụm để phân cụm cho thuật toán K-Means



Hình 3.14: Chi tiết quản lý thông tin cho từng sản phẩm

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết quả đạt được

Trong luận văn này, tác giả tập trung giải quyết các nội dung sau:

- Nghiên cứu tổng quan về thương mại điện tử, lý thuyết học máy.
- Nghiên cứu sâu về thuật toán học không giám sát K-Means.
- Xây dựng ứng dụng thử nghiệm cho thuật toán K-Means kết hợp với một số kỹ thuật khác ứng dụng trong hệ thống thương mại điện tử.

2. Hướng phát triển

- Tiếp tục phát triển chương trình nhằm hoàn thiện các chức năng một cách đầy đủ nhất sao cho đây thực sự là một công cụ hữu ích hỗ trợ cho nhiều hệ thống hơn nữa và rút ngắn thời gian tìm kiếm hơn nữa.
- Tiếp tục nghiên cứu các phương pháp khác kết hợp cùng thuật toán K-Means để đưa ra được kết quả chính xác hơn nữa.

TÀI LIỆU THAM KHẢO

[1] M. Pazzani and D. Billsus, "Learning and Revising User Profiles/: The Identification of Interesting Web Sites," *Machine Learning 27*, pp. 313–331, 1997.

[2] W. Hill, L. Stead, M. Rosenstein, and G. Furnas, "Recommending and Evaluating choices in a Virtual Community of use," *Proceedings of CHI'95*.

[3] M. Balabanovic and Y. Shoham, "Fab: Content-based, Collaborative Recommendation," *Communications of the ACM*, vol. 40, no. 3, pp. 66–72, 1997.

[4] K. Lang, "NewsWeeder: Learning to filter news," *Proceedings of the Twelfth International Conference on Machine Learning*, pp. 331–339, 1995.

[5] J. S. Breese, D. Heckerman, and C. Kadie, "Empirical analysis of predictive algorithms for collaborative filtering," *Proceedings of 14th Conference on Uncertainty in Artificial Intelligence*, 1998.

[6] J. Delgado and N. Ishii, "Memory-Based Weighted-Majority Prediction for Recommender Systems," *ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.

[7] L. H. Ungar and D. P. Foster, "Clustering Methods for Collaborative Filtering," *Papers from 1998 Workshop. Technical Report WS-98-08. AAAI Press*, 1998.

[8] G. Shani, D. Heckerman, and R. I. Brafman, "An MDP-Based Recommender System," *Proceedings of 18th Conference on Uncertainty in Artificial Intelligence*, vol. 6, pp. 1265–1295, 2002.

[9] D. M. Pennock, S. Lawrence, and C. L. Giles, "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-

Based Approach," *IJCAI'99 Workshop: Machine Learning for Information Filtering*, 1999.

[10] K. Yu, A. Schwaighofer, V. Tresp, X. Xu, and H. Kriegel, "Probabilistic Memory- based Collaborative Filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 1, pp. 56–69, 2004.

[11] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining content-based and collaborative filters in an online newspaper," *ACM SIGIR'99 Workshop on Recommender Systems: Algorithms and Evaluation*, 1999.

[12] M. J. Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering," *Artificial Intelligence Review*, pp. 393–408, 1999.

[13] P. Melville, R. J. Mooney, and R. Nagarajan, "Content-Boosted Collaborative Filtering for Improved Recommendations," *Proceedings of the 18th National Conference on Artificial Intelligence, Edmonton, Canada*, 2002.

[14] I. Soboroff and C. Nicholas, "Combining content and collaboration in text filtering," *IJCAI'99 Workshop: Machine Learning for Information Filtering*, 1999.

[15] C. Basu, H. Hirsh, and W. Cohen, "Recommendation as classification: Using social and content-based information in recommendation," *Recommender Systems. Papers from 1998 Workshop. Technical Report WS-98-08. AAAI Press*, 1998.

[16] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and Metrics for Cold-Start Recommendations," *Proceedings of the 25th Annual International ACM SIGIR*

Conference, 2002.

[17] N. Belkin, "User Modeling in Information Retrieval," *Sixth International Conference on User Modeling*, 1997.

[18] N. Belkin, J. Kay, and C. Tasso, "Special Issue on User Modeing and Information Filtering," *User Modeling and User Adapted Interaction*, 1997.

[19] J.Schumpeter, *Theory of Economic Development.* Harvard University Press,1961.

[20] MacQueen, "J. Some methods for classification and analysis of multivariate observations,"*Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967.

[21] http://bis.net.vn

[22] https://techmaster.vn

[23] http://viet.jnlp.org