MỤC LỤC

DANH SÁCH CÁC HÌNH

DANH SÁCH BẢNG BIỂU

GIỚI THIỆU

Web là kho tài nguyên dữ liệu khổng lồ, không ngừng tăng trưởng với tốc độ cao. Ngày càng nhiều thông tin trong cuộc sống được đưa lên Internet. Trong đó, Internet chứa nhiều thông tin có giá trị liên quan đến cộng đồng nói chung, và các hoạt động sản xuất kinh doanh nói riêng.Xuất phát từ thực tế đó, vậy có phương pháp nào? Chúng ta có thể khai thác thông tin từ Internet để phục vụ cuộc sống. Hiện nay có nhiều công trình nghiên cứu các phương pháp khai thác thông tin từ Internet.

Xuất phát từ bài toán thực tế trong hoạt động kinh doanh thương mại, liệu có phương pháp nào đánh giá thông tin về sản phẩm thông qua các nhận xét của người dùng trên Internet? Đây là một bài toán khó cần kết hợp nhiều kiến thức để giải quyết bài toán này. Do đó em chọn đề tài: ***"Bài toán khai thác thông tin về sản phẩm từ Web".***Khóa luận tập trung tìm hiểu các lý thuyết liên quan nhằm phần nào giải quyết được vấn đề đặt ra.

Một hệ thống tổng hợp thông tin từ Internet cho phép người dùng đưa vào các thông tin cần đánh giá về sản phẩm được quan tâm. Sau đó, hệ thống đưa được ra các thông tin liên quan đến sản phẩm để có thể hỗ trợ các doanh nghiệp có thêm một kênh thông tin về các sản phẩm trên thị trường. Hệ thống được mô tả như sau:

1. Nhập thông tin sản phẩm: Người dùng nhập các thuật ngữ về *thông tin sản phẩm* vào ô thông tin sản phẩm cần đánh giá. Hệ thống trả về các thông tin sản phẩm mà hệ thống khai thác, phân loại, thống kê được thông qua máy tìm kiếm

2. Tìm kiếm thông tin:Hệ thống dựa vào các *thông tin sản phẩm* được nhập vào và gửi vào máy tìm kiếm để tìm các *Ý kiến người dùng sản phẩm*hoặc *Xu hướng.*

3. Hỗ trợ đánh giá:Kết quả trả về từ máy tìm kiếm được đem phân loại, thống kê các thông tin cần thiết về sản phẩm nhằm đánh giá cảm nhận của *người tiêu dùng* đối với sản phẩm được đưa vào đánh giá.

4. Báo cáo:Hệ thống đưa ra các bản báo cáo về ý kiến của người sử dụng *sản phẩm* bằng các số liệu theo chuyên môn.

5. Giao diện hệ thống:Hệ thống có giao diện thân thiện, thuận lợi cho người dùng và người quản lý.

Qua cách đặt vấn đề trên, khóa luận được trình bày như sau:

Giới thiệu: *Giới thiệu chung về bài toán và phạm vi của khóa luận.*

Chương 1:*Trình bày cơ sở lý thuyết để thục hiện khóa luận.*

Chương 2:*Trình bày các kiến thức liên quan đến bài toán tìm kiếm thông tin trên Internet dùng để trợ giúp các hoạt động trong kinh doanh.*

Chương 3:*Trình bày phần phân tích thiết kế một ứng dụng mang tính chất thử nghiệm.*

Kết luận

Tài liệu tham khảo

CHƯƠNG 1: CƠ SỞ LÝ THUYẾT

1.1CÁC KHÁI NIỆM CƠ BẢN

Khai phá dữ liệu là một hướng nghiên ra đời hơn một thập niên trở lại đây.Các kỹ thuật chính được áp dụng trong lĩnh vực này phần lớn được thừa kế từ lĩnh vực Cơ sở dữ liệu,học máy, trí tuệ nhân tạo, lý thuyết thông tin, xác suất thống kê, và tính toán hiệu năng cao. Do sự phát triển nhanh của *Khai phá dữ liệu* về phạm vi áp dụng và các phương pháp tìm kiếm tri thức, nên đã có nhiều quan điểm khác nhau về *Khai phá dữ liệu*. Tuy nhiên, ở một mức trừu tượng nhất định, theo [1] khái niệm *Khai phá dữ liệu* như sau:

"Khai phá dữ liệu là một quá trình tìm kiếm, phân tích, phát hiện các tri thức mới, tiềm ẩn, hữu dụng trong các cơ sở dữ liệu lớn."

Trong lĩnh vực *khai phá dữ liệu*, có nhiều hướng nghiên cứu được đưa ra trong đó có một số hướng chính được nhiều nhà nghiên cứu quan tâm.

Khai phá dữ liệu văn bản,Web, Trích chọn thông tin, Phân tích mạng xã hội, khai phá quan điểm, Phân tích dữ liệu kinh tế - tài chính, Khai phá dữ liệu sinh học, y tế,…

Khóa luận này hướng đến việc tìm hiểu và xây dựng hệ thống thống nhằm khai thác thông tin các sản phẩm mà người tiêu dùng đã đánh giá trên Internet, cụ thể là các trang web. Người tiêu dùng có mua, sử dụng các sản phẩm và họ đã có cảm nhận và suy nghĩ về sản phẩm đó. Đôi lúc họ đưa những cảm nhận về sản phẩm nào đó lên các trang web cá nhân, các diễn đàn, … trên Internet. Bài toán mà khóa luận này tìm cách giải quyết là tìm cách khai thác các thông tin về một sản phẩm cụ thể nào đó trên Internet, thống kê để phục vụ cho công tác khảo sát và đánh giá sản phẩm trên thị trường. Sau đây là một số khái niệm được sử dụng.

Thông tin sản phẩm: Là thông tin mô tả về nguồn gốc, tính năng kỹ thuật, tính chất lý hóa tính, công dụng chính, giá thành, màu sắc, hình dáng, kích thước, … của sản phẩm.

Ý kiến người dùng sản phẩm: Là các thông tin người dùng phản ánh về sản phẩm được thể hiện qua các từ đánh giá về ưu điểm như: Tốt, thuận tiện, tiết kiệm, bền, rẻ, đẹp, phong phú, đa dạng, mượt mà, mịn, … hoặc được đánh giá nhược điểm: Xấu, kém, mau hỏng, hàng giả, hàng nhái, … hoặc được thể hiện mong muốn về sản phẩm qua các từ như: giá như, giá mà, cần, phải, để tốt hơn, …

Xu hướng: Là các từ liên quan đến các mong muốn của người dùng về sản phẩm. Được chi thành xu hướng tốt hoặc xấu hoặc không thiện cảm.

+ Xu hướng tốt: Xu hướng đánh giá thông tin sản phẩm tốt

+ Xu hướng xấu: Xu hướng đánh giá thông tin sản phẩm xấu

+ Xu hướng không thiện cảm: Xu hướng không khen, không chê sản phẩm.

Người tiêu dùng: Là người mua hoặc người sử dụng sản phẩm hoặc người có ý định mua hay sử dụng sản phẩm có gửi thông tin lên Internet.

Người dùng: Người có tương tác với hệ thống.

Người quản trị: Người có nhiệm vụ quản lý hệ thống.

Máy tìm kiếm:Các cỗ máy tìm kiếm thông tin trên Internet: Google, Yahoo, Bing, …

Sản phẩm: Là tất cả các mặt hàng đang được tiêu thụ trên thị trường bao gồm thị trường trong nước và nước ngoài.

1.2 KHÁM PHÁ TRI THỨC TRONG CƠ SỞ DỮ LIỆU

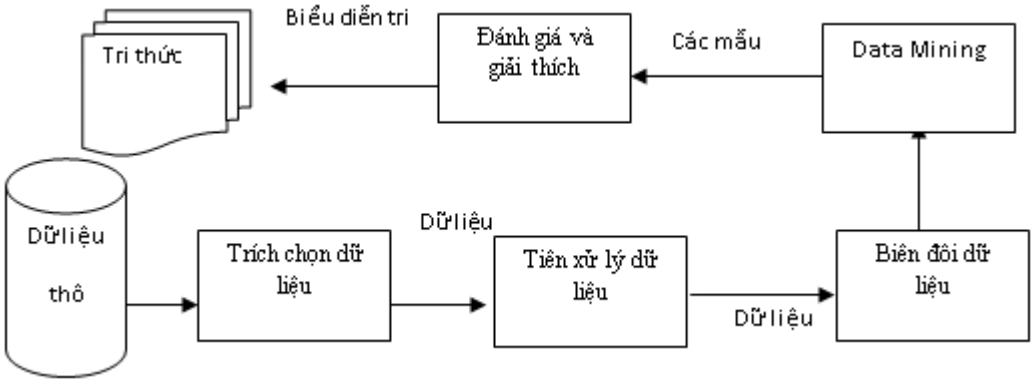
Khai phá dữ liệu là lĩnh vực có liên quan đến nhiều ngành khác nhau như: thống kê, học máy, tính toán phân tán,cơ sở dữ liệu, thuật toán, mô hình hóa dữ liệu,…

Mục tiêu của khai phá dữ liệu là khám phá tri thức từ đó dùng hỗ trợ ra quyết định, trong lĩnh vực hẹp này có thể được chia thành một số giai đoạn [3][4]:

- *Trích chọn dữ liệu:* bước này trích những bộ dữ liệu cần được khám phá từ các hệ thống dữ liệu (databases, data warehouses, data repositories) ban đầu theo một số tiêu chí nhất định.

- *Tiền xử lý dữ liệu:*Bước này làm sạch dữ liệu (xử lý những dữ liệu dư thừa, nhiễu, .v.v.), rút gọn dữ liệu (áp dụng các thuật toán lấy mẫu, .v.v.), rời rạc hóa dữ liệu. Kết quả là dữ liệu sẽ nhất quán, đầy đủ, được rút gọn, và được rời rạc hóa.
- *Biến đổi dữ liệu:*Đây là bước chuẩn hóa dữ liệu, tinh chỉnh dữ liệu để đưa dữ liệu về dạng chuẩn để giúp kỹ thuật khai phá dữ liệu ở bước sau.
- *Khai phá dữ liệu*: Áp dụng những kỹ thuật phân tích nhằm để trích chọn thông tin, những mối liên hệ đặc biệt của dữ liệu. Bước này rất quan trọng và cần nhiều tài nguyên nhất của toàn bộ quá trình khai phá trin thức.
- *Đánh giá và biểu diễn tri thức*: Các mẫu tin và quan hệ giữa chúng đã được rút trích ở bước trên được mã hóa và biểu diễn theo dạng dễ quan sát như đồ thị, cây, bảng biểu, luật, .v.v. Bước này cung cấp thông tin cho các nhà quản trị ra quyết định.

Các giai đoạn trong KDD được thể hiện trực quan như hình 1 dưới đây:

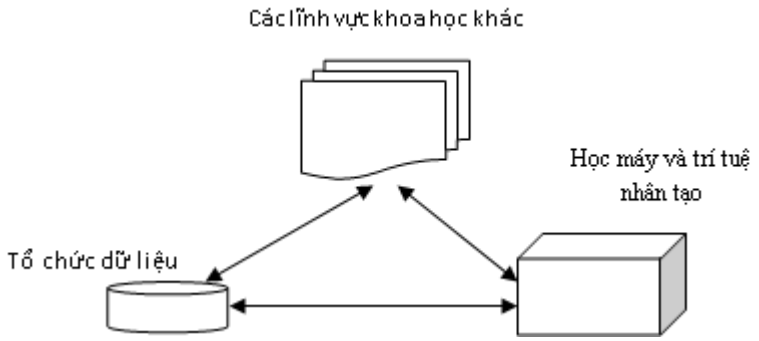


Hình 1.1:Các bước thực hiện trong quá trình khám phá tri thức

1.3 CÁC KỸ THUẬT ÁP DỤNG TRONG KHAI PHÁ DỮ LIỆU

1.3.1 Các kỹ thuật tiếp cận trong Khai phá dữ liệu

Khai phá tri thức là một lĩnh vực liên ngành, bao gồm: Tổ chức dữ liệu, học máy, trí tuệ nhân tạo và các khoa học khác, sự kết hợp này có thể được diễn tả như trong hình 1.2 dưới đây:



Hình 1.2 Các lĩnh vực liên quan đến Khám phá tri thức trong CSDL

Dựa trên quan điểm của học máy thì các kỹ thuật trong *Khai phá dữ liệu*, bao gồm:

- *Học có giám sát*: Là quá trình gán nhãn lớp cho các phần tử trong CSDL dựa trên một tập các ví dụ huấn luyện và các thông tin về nhãn lớp đã biết.
- *Học không có giám sát:* Là quá trình phân chia một tập dữ liệu thành các lớp hay là cụm (clustering) dữ liệu tương tự nhau mà chưa biết trước các thông tin về lớp hay tập các ví dụ huấn luyện.
- *Học nửa giám sát:* Là quá trình phân chia một tập dữ liệu thành các lớp dựa trên một tập nhỏ các ví dụ huấn luyện và một số các thông tin về một số nhãn lớp đã biết trước.

Theo lớp các bài toán cần giải quyết, thì Khai phá dữ liệu bao gồm các kỹ thuật áp dụng sau:

- *Phân lớp và dự đoán(classification and prediction)*: xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ: phân lớp các bệnh nhân dữ liệu trong hồ sơ bệnh án. Hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định (decision tree), mạng nơ ron nhân

tạo (neural network), .v.v. Phân lớp và dự đoán còn được gọi là học có giám sát.

- *Luật kết hợp (association rules)*: là dạng luật biểu diễn tri thứ ở dạng khá đơn giản. Ví dụ: "60 % nữ giới vào siêu thị nếu phấn thì có tới 80% trong số họ sẽ mua thêm son". Luật kết hợp được ứng dụng nhiều trong lĩnh vực kinh doanh, y học, tin-sinh, tài chính và thị trường chứng khoán, .v.v.
- *Phân tích chuỗi theo thời gian (sequential/ temporal patterns)*: tương tự như khai phá luật kết hợp nhưng có thêm tính thứ tự và tính thời gian. Hướng tiếp cận này được ứng dụng nhiều trong lĩnh vực tài chính và thị trường chứng khoán vì nó có tính dự báo cao.
- *Phân cụm (clustering/ segmentation)*: xếp các đối tượng theo từng cụm dữ liệu tự nhiên. *Phân cụm* còn được gọi là học không có giám sát (unsupervised learning).
- *Mô tả khái niệm (concept description and summarization)*: thiên về mô tả, tổng hợp và tóm tắt khái niệm. Ví dụ: tóm tắt văn bản.

1.3.2 Các dạng dữ liệu có thể khai phá

Do *Khai phá dữ liệu* được ứng dụng rộng rãi trên nhiều lĩnh vực có thể làm việc với rất nhiều kiểu dữ liệu khác nhau. Ví dụ: *Cơ sở dữ liệuquan hệ, dữ liệu đa chiều, Cơ sở dữ liệu quan hệ - hướng đối tượng, Cơ sở dữ liệu có thuộc tính không gian và thời gian, Cơ sở dữ liệu chuỗi thời gian, Cơ sở dữ liệu đa phương tiện, ...*

1.4TÌM KIẾM THÔNG TIN TRÊN INTERNET

Theo [thụy1] máy tìm kiếm là một hệ thống phần mềm được xây dựng nhằm tiếp nhận yêu cầu tìm kiếm của người dùng, sau đó phân tích yêu cầu này và tìm kiếm thông tin trong cơ sở dữ liệu được tải xuống từ Internet và đưa ra kết quả là danh sách các trang Web lên quan với yêu cầu người dùng.

Về cơ bản, mọi kỹ thuật tập trung vào máy tìm kiếm (*Searche engine*). Hiện nay trên thế giới có rất nhiều máy tìm kiếm, điển hình là *Google*, *Bing*, *Yahoo*,…,và một số máy tìm kiếm có cách thực hiện rất đặc biệt không chỉ đưa ra kết quả tìm kiếm là các địa chỉ chứa thông tin mà tìm kiếm và tổng hợp tri thức như

Wolframanpha, máytìm kiếm này còn biết cách trả lời các vấn đề mang tính chất đặc thù chuyên ngành như toán học, lý, hóa, lịch sử, địa lý,…

Kiến trúc cơ bản của máy tìm kiếm gồm các khối như truy vấn dữ liệu, đánh chỉ mục, phân loại dữ liệu….Nói chung, máy tìm kiếm thực hiện một số thao tác cơ bản sau:

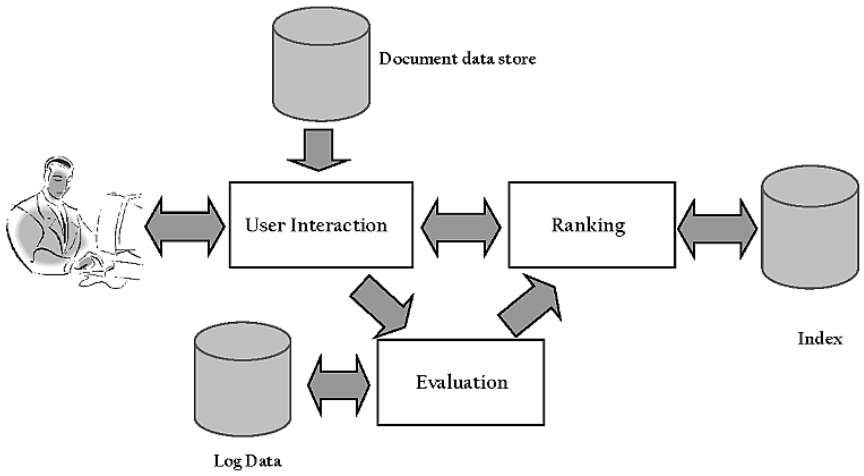
Bước 1: *Phân tích các yêu cầu của người dùng, phân loại và đánh chỉ mục các yêu cầu này, lưu vào hệ thống.*

Bước 2:*Các kết quả tìm kiếm cũng được phân tích, đánh chỉ mục và lưu vào hệ thống.*

Bước 3: *Khi có yêu cầu tìm kiếm thông tin, máy tìm kiếm so khớp yêu cầu với các yêu cầu đã có sẵn nếu phù hợp sẽ đưa kết quả ra luôn, nếu yêu cầu này chưa có thì sẽ tìm thông tin rồi thao tác lại bước 1. Đối với kết quả tìm kiếm mới sẽ bổ sung như bước 2.*

Sau đây là sơ đồ kiến trúc chung của một số khối trong máy tìm kiếm [2].

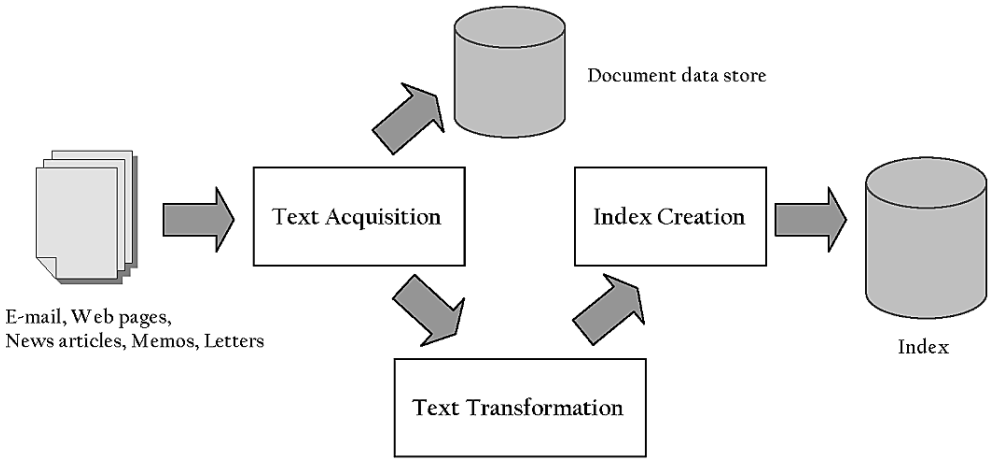
Khối Truy vấn



Hình 1.3: Sơ đồ khối Khối truy vấn

Khối truy vấn nhận thông tin từ người dùng theo dạng văn bản. Từ đó phân loại, xác định yêu cầu của câu truy vấn xem thuộc nhóm nào? Đánh giá và phân tích câu. Tìm kiếm trong cơ sở dữ liệu chỉ mục rồi trả lại kết quả tìm kiếm cho người dùng.

Khối **Đánh chỉ mục**



Hình1.4: Sơ đồ khối Đánh chỉ mục

Trong khóa luận này máy tìm kiếm Google được sử dụng làm công cụ để lấy thông tin. Do đó, phần sau sẽ khảo sát kỹ hơn về máy tìm kiếm của *Google*.

Tên gọi của máy tìm kiếm Google có nguồn gốc từ chữ "Googol". Sau một thời gian không lâu máy tìm kiếm này trở nên nổi tiếng vì đáp ứng tốt yêu cầu người dùng. Google đã áp dụng những kỹ thuật tiên tiến để nâng cao khả năng sản phẩm của họ như:

- *Công nghệ crawling có tốc độ cao khi thu thập tài liệu và cập nhật chúng.*
- *Hệ thống lưu trữ không những lưu trữ chỉ số mà lưu trữ toàn bộ nội dung tài liệu.*
- *Hệ thống đánh chỉ số hiệu quả khi làm việc trên hàng trăm tetrabyte dữ liệu.*
- *Câu hỏi cần được tiếp nhận và đáp ứng nhanh theo cỡ hàng trăm nghìn câu hỏi trong một giây.*

Máy tìm kiếm này có một số đặc trưng

- ***Boolean**: việc cho phép ngầm định các phép toán logic(and, or, not, (),+,-) trong câu hỏi tìm kiếm và thực hiện.*

- ***Default**:Phép toán logic được thi hành ngầm định.*

- ***Proxymity**:Thực hiện tìm theo cụm từ.*

- ***Truncation**: Tiến hành tìm kiếm theo từ gốc,cho phép có kí hiệu đại diện trong câu hỏi.*

- ***Fields****: Cho phép đặt tham số tìm kiếm theo một số trường theo tiêu đề, địa chỉ URL, liên kết, miền/site, kiểu file, ...*
- ***Limits****: Cho phép đưa ra một số hạn chế về thời gian, lĩnh vực , nội dung, đa phương tiện ...*
- ***Stop****(stop word): Cho phép loại bỏ từ dừng, một số trường hợp không tiến hành tìm kiếm từ quá thông dụng.*
- ***Sorting****: Sắp xếp kết quả tìm kiếm theo độ liên quan, phân cụm theo địa chỉ web, sắp theo thứ tự thời gian, kích thước.*

1.5 PHÂN LOẠI THÔNG TIN TÌM KIẾM

Bài toán phâncụm thông tinlà một trong những bài toán quan trọng nhất trong lĩnh vực khai phá dữ liệu. Phân cụm dữ liệu là dựa vào các mục tiêu tức là cụ thể các tiêu chí phân cụm để tự động sinh ra các lớp(cụm) thông tin.

Khi áp dụng các thuật toán phân cụm dữ liệu nhằm mục đích quan trọng là khai pháccấu trúc của mẫu dữ liệu để từ đó tạo ra các cụm dữ liệu từ kho dữ liệu gốc, theo đó, cho phép phân tích, nghiên cứu cho từng cụm dữ liệu để khám phá và trích xuất các thông tin tiềm ẩn, có ích hỗ trợ ra quyết định.

Ví dụ: *Sau khi tìm kiếm các văn bản trên Internet về các thông tin sản phẩm, hệ thống được xây dựng phải khám phá ra các thông tin về sự đánh giá của người tiêu dùng về sản phẩm đó là "tốt" hay "xấu" hoặc xu hướng mong muốn về sản phẩm cảu người tiêu dùng.*

Như vậy, phân cụm dữ liệu là phương thức xử lý thông tin nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm.Hiện nay, các kỹ thuật phân cụm đã được ứng dụng rộng rãi trong các ứng dụng như: nhận dạng mẫu, xử lý ảnh, nghiên cứu thị trường, trực quan hoá, …Trong nội dung tiếp theo, khóa luận sẽ đề cập đến các hướng phân cụm dữ iệu, đây là phần quan trọng trong lĩnh vực khai phá dữ liệu.

Các hướng giải quyết phân cụm:

Theo [thụy1], có một số cách phân cụm như sau:

- ***Phương pháp phân cụm theo mô hìnhvà phân vùng (partitioning):*** Phương pháp thứ nhất tạo ra các mô hình biểu diễn các cụm; phương pháp thứ hai chỉ đơn giản là tập hợp các phần tử dữ liệu vào các cụm.

- Phân cụm đơn định và phân cụm xác suất: Trong phân cụm đơn định, mỗi một phần tử dữ liệu (thông tin trên trang Web) chỉ phụ thuộc vào một cụm. Có thể xem xét việc gán thông tin *d* thuộc cụm *i* như là việc đặt một giá trị trong mảng hai chiều *Z* Boolean $Z_{d,i}$là l. Trong phân cụm xác suất. mỗi phần tử dữ liệu sẽ có xác suất nào đó đối với mỗi cụm. Trong ngữ cảnh này, $Z_{d,i}$ có giá trị là một số thực trongkhoảng[0,1]. Tức là, giá trị trong bảng là một ánh xạ *z: S ×S→ [0, 1]* và các vector c_i, làm cực tiểu hóa $\sum_i \sum_{d \in Si} \delta(d, c_i)$hoặc cực đại hóa $\sum_i \sum_{d \in S_i} sim(d, c_i)$.

- ***Phân cụm phẳng và phân cụm phân cấp:*** Phân cụm phẳng chỉ đơn giản là chia tập dữ liệu thành một số tập con. Còn phân cụm phân cấp tạo ra một cây phân cấp của các cụm. Việc phân hoạch có thể thực hiện theo hai cách,a) cách thứ nhất bắt đầu bằng việc cho mỗi mẫu tin vào một cụm của nó và tiến hành kết hợp các cụm lại với nhau cho đến khi số các cụm là phù hợp, cách này được gọi là phân cụm từ dưới lên (bottom - up). b) Cách thứ hai bắt đầu bằng việc khai báo các cụm nguyên thủy và sau đó gán các mẫu tin vào các cụm, cách này dược gọi là phân cụm từ trên xuống (top - down). Như vậy, có thể xem xét kỹ thuật phân cụm bottom - up dựa vào quá trình lặp lại việc trộn các cụm tương tự nhau cho đến khi đạt được sổ cụm mongmuốn; kỹ thuật phân cụm top - down làm mịn dần bằng cách gán các mẫu tin vào các cụm được thiết đặt trước. Kỹ thuật bottom - up thường chậm hơn, nhưng có thể được dùng trộn một tập nhỏ các mẫu có trước để khởi tạo các cụm nguyên thủy trước khi tiến hành kỹ thuật từ trên xuống.

- Phân cụm theo lô và phân cụm gia tăng: Trong phân theo lô, toàn bộ tập dữ liệu được sử dụng để tạo ra các cụm. Trong phân cụm gia tăng. giải thuật phân cụm lấy từng phần tử dữ liệu và cập nhật các cụm để phân vào cụm thích hợp.

Trong khóa luận này, các mẫu tin được phân cụm theo các tiêu chí đem vào tìm kiếm. Nghĩa là, các tiêu chí tìm kiếm bao gồm tên sản phẩm, các thuộc tính của sản phẩm. Các sản phẩm được phân loại theo loại sản phẩm. Các loại sản phẩm

thuộc một nhóm sản phẩm nào đó. Các tiêu chí này được gán một mã xác định(mã tìm kiếm) nhằm phân biệt các tiêu chí khác nhau, dễ dàng cho việc phân cụm.

1.6 TỔ CHỨC LƯU TRỮ THÔNG TIN TÌM KIẾM

Khi có kết quả tìm kiếm các hệ thống cần lưu trữ theo một định dạng nào đó để phục vụ các nghiệp vụ tiếp theo. Hiện nay người ta thường dùng hệ quản trị cơ sở dữ liệu lớn để lưu trữ như: SQL server, MySQL, Postgre, Oracle,… Đặc biệt hiện nay định dạng XML là một trong những chuẩn dữ liệu được dùng phổ biến. Khóa luận này sử dụng hệ quản trị cơ sở dữ liệu SQL server để lưu trữ.

Dữ liệu khai thác về được phân loại theo các tiêu chí tìm kiếm, các thông tin từ các trang web khi lấy về được đánh mã để phân biệt cho mỗi lần lấy kết quả. Các thông tin này được gắn với *mã tìm kiếm*. Các *url* chính xác của từng bản tin cũng được lưu trữ để thuận tiện cho việc lấy lại nội dung sau này.

Ví dụ: Lưu trữ thông tin sau khi tìm kiếm:

WebsiteID	SearchID	Url	Content
97	26	Vanphongphamt2.com	

WebsiteID là mã của trang Web chứa bản tin thỏa mãn tiêu chí tìm kiếm có mã *SearchID* là 26 (chứa các từ khóa về sản phẩm các loại bút bi). Thuộc tính *Url* chứa địa chỉ của Website có chứa thông tin về *bút bi*, Thuộc tính *Content*chứa các văn bản về thông tin các loại bút bị có trong Website *Vanphongphamt2.com,* đôi khi còn có lẫn các thẻ định dạng HTML của trang Web đó. Dữ liệu này mới chỉ là dữ liệu thô.

Các bản tin được nhóm theo mục tiêu tìm kiếm (phụ thuộc vào nội dung của khóa tìm kiếm) do vậy các bản tin thường chứa các thông tin về một loại sản phẩm cụ thể.

1.7XỬ LÝ THÔNG TIN

Các bản tin nhận được từ các máy tìm kiếm được lưu trữ trong hệ quản trị cơ sở dữ liệu SQL Server. Các dữ liệu này được gọi là dữ liệu thô. Về mặt hình thức văn bản này được coi là văn bản phi cấu trúc, trong đó các đối tượng được diễn tả

bằng các danh từ và các thuộc tính của đối tượng được mô tả bằng các tính từ, trạng từ,…

Khi xử lý thông tin được máy tìm kiếm trả về, dựa vào bộ từ khóa tìm kiếm *SearchKeys*trong bảng *SearchTable* theo hình sau:

SearchID	SearchKeys	ProductID	SearchEngineID
26	Bút + bi + ngoại + Giá + tiền + Bền + Rẻ	10	www.google.com

Dữ liệu được phân cụm theo mã sản phẩm *ProductID = 10*và các thuộc tính của sản phẩm này. Hệ thống phân tích các thông tin rồi phân cụm chúng theo các tiêu chí được lưu trong *SearchKeys* đối với sản phẩm có mã *ProductID = 10.*

CHƯƠNG 2: KHAI PHÁ VÀ TỔNG HỢP DỮ LIỆU

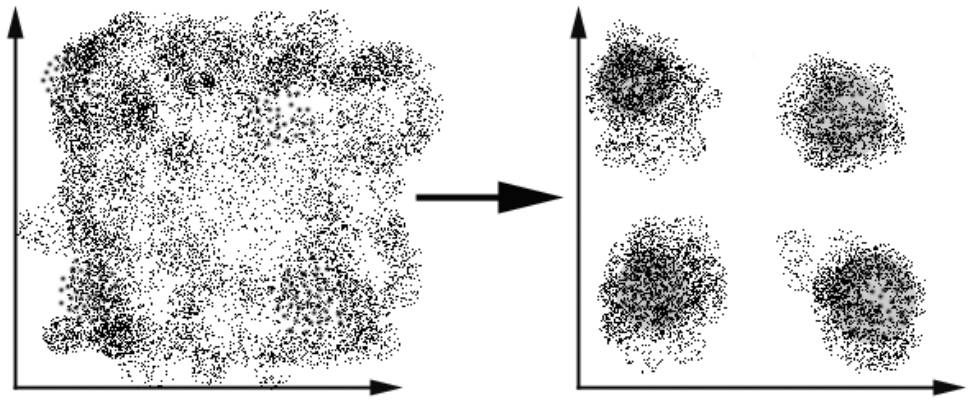
Chương này khóa luận trình bày một số kiến thức cơ bản liên quan đến thống kê và khai phá dữ liệu, theo đó làm sáng tỏ cách thức tổng hợp thông tin từ các mẫu tin khai thác được

2.1 PHÂN CỤM DỮ LIỆU

Phân cụm dữ liệu áp dụng nhiều kiến thức trong các ngành học máy, thống kê, nhận dạng, … Có rất nhiều khái niệm khác nhau về phân cụm, tuy nhiên có khái niệm chung nhất về phân cụm [2].

"Phân cụm dữ liệu *là một phương pháp trong khai phá dữ liệu, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên tiềm ẩn, hấp dẫn trong tập dữ liệu lớn, từ đó cung cấp thông tin, tri thức hữu ích cho người sử dụng.*"

Thật vậy, phân cụm dữ liệu là quá trình phân chiatập dữ liệu thành các khần khác nhau dựa trên một tập các tiêu chí cho trước. Phương pháp phân cụm có thể được xác định trước theo kinh nghiệm hoặc có thể được tự động xác định bằng phương pháp phân cụm.



Hình2.1:Hình minh họa phân cụm dữ liệu

Ở hình trên, khi áp dụng phương pháp phân cụm dù thủ công hay tự động, sẽ thu được các cụm trong đó các phần tử "*gần nhau*" hay là "*tương tự*" thì chúng thuộc về các cụm khác nhau.

Phân cụm dữ liệu phải giải quyết đó là hầu hết các dữ liệu chứa dữ liệu "nhiễu" (noise) do các bước lấy mẫu chưa đầy đủ hoặc thiếu chính xác, do đó cần phải lập kế hoạch chiến lược ngay tại bước tiền xử lý dữ liệu để loại bỏ "nhiễu"

trước khi đưa vào giai đoạn tiếp theo. Khái niệm "nhiễu" được hiểu là thông tin về các đối tượng chưa chính xác, hoặc là khuyết thiếu thông tin về một số thuộc tính. Một trong các kỹ thuật xử lý nhiễu phổ biến là việc thay thế giá trị của các thuộc tính của đối tượng "nhiễu" bằng giá trị thuộc tính tương ứng của đối tượng dữ liệu gần nhất.

Do vậy, phân cụm dữ liệu cần giải quyết một số vấn đề sau:

- *Xây dựng hàm tính độ đo tương tự*
- *Xây dựng tập các tiêu chí phân cụm*
- *Thiết lập các cấu trúc dữ liệu cho cụm dữ liệu*
- *Xây dựng thuật toán phân cụm dữ liệu*
- *Xây dựng hệ thống phân tích và đánh giá kết quả*

Ngày nay, chưa có một phương pháp phân cụm nào có thể giải quyết trọn vẹn cho tất cả các dạng cấu trúc cụm dữ liệu.

2.2 CÁC ỨNG DỤNG CỦA PHÂN CỤM DỮ LIỆU

Phân cụm dữ liệu được ứng dụng trong nhiều lĩnh vực kinh tế, y học, thương mại, khoa học,... Các phương pháp phân cụm được áp dụng cho một số ứng dụng điển hình trong các lĩnh vực sau**:**

- *Thương mại*: Trong thương mại, các hệ thống thông tin áp dụng phương pháp phân cụm dữ liệu có thể giúp các doanh nhân có đủ thông tin về nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và từ đó ra quyết định chính xác hơn.
- *Khoa học tự nhiên*: Các lĩnh vực như sinh học, môi trường, địa lý, toán học,… các phương pháp phân cụm giúp cho các nhà nghiên cứu cô lập được các thông tin đặc thù của từng đối tượng để phục vụ cho nghiên cứu.
- *Nghiên cứu trái đất*: Phân cụm để theo dõi các hoạt động của các vùng trên trái đất nhằm cung cấp thông tin cho nhận dạng các vùng nguy hiểm.
- *Khai phá dữ liệu Web*: Phân cụm dữ liệu có thể khai phá các nhóm dữ liệu có nhiều ý nghĩa trong môi trường Web, như khai thác quan điểm người dùng, xu hướng tiếp cận và giải quyết vấn đề.

2.3 CÁC KIỂU DỮ LIỆU VÀ ĐỘ ĐO TƯƠNG TỰ

Khi phân cụm dữ liệu cần có một "thước đo" nào đó để đo các sự vật. Như vậy với các đối tượng khác nhau thì cần "thước đo" cũng khác nhau. Sau đây là cách phân lớp dựa trên hai đặc trưng là: kích thước miền và hệ đo.

Cho một *Cơ sở dữ liệuD* chứa *n*phần tử trong không gian *k* chiều, trong đó x,y,z là các phần tử thuộc *D*: $x=(x_1,x_2,..,x_k)$;$y=(y_1,y_2,..,y_k)$;$z=(z_1,z_2,..,z_k)$, trong đó x_i, y_i, z_i với $i=\overline{1,k}$ là các thuộc tính tương ứng của các đối tượng x,y,z. Vì vậy, hai khái niệm "các kiểu dữ liệu" và "các kiểu thuộc tính dữ liệu" được xem là tương đương với nhau, như vậy, chúng ta sẽ có các kiểu dữ liệu sau [2].

2.3.1 Phân loại các kiểu dữ liệu dựa trên kích thước miền

- *Thuộc tính liên tục (Continuous Attribute)*: Thuộc tính này có miền giá trị là vô hạn không đếm được, nghĩa là giữa hai giá trị tồn tại vô số giá trị khác. Thí dụ như trường số thực.
- *Thuộc tính rời rạc (DiscretteAttribute)*: Miền giá trị của thuộc tính này là đếm được. Thí dụ như số nguyên.

Lớp các thuộc tính nhị phân là trường hợp đặc biệt của thuộc tính rời rạc mà miền giá trị của nó chỉ có 2 phần tử được diễn tả như:*Yes/No* hoặc *Nam/Nữ*, *False/true*,…

2.3.2 Phân loại các kiểu dữ liệu dựa trên hệ đo

Giả sử rằng chúng ta có hai đối tượng x, y và các thuộc tính x_i, y_i tương ứng với thuộc tính thứ i của chúng. Chúng ta có các lớp kiểu dữ liệu như sau:

- *Thuộc tính định danh(nominal Scale)*: đây là dạng thuộc tính khái quát hoá của thuộc tính nhị phân, trong đó miền giá trị là rời rạc không phân biệt thứ tự và có nhiều hơn hai phần tử - nghĩa là nếu x và y là hai đối tượng thuộc tính thì chỉ có thể xác định là $x \neq y$ hoặc x=y. Thí dụ như thuộc tính về *nơi sinh* hoặc thuộc tính *các đội bóng chơi cho giải vô địch quốc gia Việt Nam*.
- *Thuộc tính có thứ tự (Ordinal Scale)*: là thuộc tính định danh có thêm tính *thứ tự,* nhưng chúng không được định lượng. Nếu x và y là hai thuộc tính thứ tự thì ta có thể xác định là $x \neq y$ hoặc x=y hoặc x>y hoặc x<y. Thí dụ như thuộc tính *Huy chương* của vận động viên thể thao.

- *Thuộc tính khoảng (Interval Scale)*: Nhằm để đo các giá trị theo xấp xỉ tuyến tính. Với thuộc tính khoảng, chúng ta có thể xác định một thuộc tính là đứng trước hoặc đứng sau thuộc tính khác với một khoảng là bao nhiêu. Nếu $x_i > y_i$ thì ta nói x cách y một khoảng $x_i - y_i$ tương ứng với thuộc tính thứ i. Một thí dụ về thuộc tính khoảng như thuộc tính *số Serial* của một đầu sách trong thư viện hoặc thuộc tính *số kênh* trên truyền hình.
- *Thuộc tính tỉ lệ (Ratio Scale):* là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc đầy ý nghĩa, *thí dụ như thuộc tính chiều cao hoặc cân nặng lấy điểm 0 làm mốc.*

Trong các thuộc tính dữ liệu trình bày ở trên, thuộc tính định danh và thuộc tính có thứ tự gọi chung là thuộc tính hạng mục (Categorical), trong khi đó thì thuộc tính khoảng và thuộc tính tỉ lệ được gọi là thuộc tính số (Numeric).

Người ta còn đặc biệt quan tâm đến *dữ liệu không gian (Spatial Data)*. Đây là loại dữ liệu có các thuộc tính số khái quát trong không gian nhiều chiều, dữ liệu không gian mô tả các thông tin liên quan đến không gian chứa đựng các đối tượng, thí dụ như thông tin về hình học,… Dữ liệu không gian có thể là dữ liệu liên tục hoặc rời rạc:

- *Dữ liệu không gian rời rạc*: có thể là một điểm trong không gian nhiều chiều và cho phép ta xác định được khoảng cách giữa các đối tượng dữ liệu trong không gian.
- *Dữ liệu không gian liên tục*: bao chứa một vùng trong không gian.

Thông thường, các thuộc tính số được đo bằng các đơn vị xác định như là *kilogams* hay là *centimeter*. Tuy nhiên, các đơn vị đo có ảnh hưởng đến các kết quả phân cụm. Thí dụ như thay đổi độ đo cho thuộc tính cân nặng từ kilogams sang *Pound* có thể mang lại các kết quả khác nhau trong phân cụm. Để khắc phục điều này người ta phải *chuẩn hoá dữ liệu*, tức là sử dụng các thuộc tính dữ liệu không phụ thuộc vào đơn vị đo. Thực hiện chuẩn hoá phụ thuộc vào ứng dụng và người dùng, thông thường chuẩn hoá dữ liệu được thực hiện bằng cách thay thế mỗi một thuộc tính bằng thuộc tính số hoặc thêm các trọng số cho các thuộc tính.

2.4 CÁC YÊU CẦU CẦN THIẾT CHO TẠO DỤNG KỸ THUẬT PCDL

Dựa vào mục đích của ứng dụng thực tế hoặc yêu cầu về chất lượng số liệu mà các phương pháp phân cụm có thể khác nhau. Đây là bước quan trọng cho việc

giải quyết vấn đề phân cụm. Các phương pháp đều thỏa mãn tiêu chuẩn chung như sau:

- *Có khả năng mở rộng* (Scalability): Một số thuật toán có thể ứng dụng tốt cho tập dữ liệu nhỏ (khoảng 200 bản ghi dữ liệu) nhưng không hiệu quả khi áp dụng cho tập dữ liệu lớn (Khoảng 1 triệu bản ghi).
- *Thích nghi với các kiểu dữ liệu khác nhau:* Thuật toán có thể áp dụng hiệu quả cho việc phân cụm các tập dữ liệu với nhiều kiểu dữ liệu khác nhau như dữ liệu kiểu số, kiểu nhị phân, dữ liệu kiểu hạng mục, .. và thích nghi với kiểu dữ liệu hỗn hợp giữa các dữ liệu đơn trên.
- *Khám phá ra các cụm với hình thù bất kỳ*: do hầu hết các CSDL có chứa nhiều cụm dữ liệu với các hình thù khác nhau như: hình lõm, hình cầu, hình que, …Vì vậy, để khám phá được các cụm có tính tự nhiên thì các thuật toán phân cụm cần phải có khả năng khám phá ra các cụm có hình thù bất kỳ.
- *Tối thiểu lượng tri thức cần cho xác định các tham số vào*: do các giá trị đầu vào thường rất ảnh hưởng đến thuật toán phân cụm và rất phức tạp để xác định các giá trị vào thích hợp đối với các CSDL lớn.
- *Ít nhạy cảm với thứ tự của dữ liệu vào:* Cùng một tập dữ liệu, khi đưa vào xử lý cho thuật toán PCDL với các thứ tự vào của các đối tượng dữ liệu ở các lần thực hiện khác nhau thì không ảnh hưởng lớn đến kết quả phân cụm.
- *Khả năng thích nghi với dữ liệu nhiễu cao:* Hầu hết các dữ liệu phân cụm trong Data Mining đều chứa đựng các dữ liệu lỗi, dữ liệu không đầy đủ, dữ liệu rác. Thuật toán phân cụm không những hiệu quả đối với các dữ liệu nhiễu mà còn tránh dẫn đến chất lượng phân cụm thấp do nhạy cảm với nhiễu.
- *Ít nhạy cảm với các tham số đầu vào:* Nghĩa là giá trị của các tham số đầu vào khác nhau ít gây ra các thay đổi lớn đối với kết quả phân cụm.
- *Thích nghi với dữ liệu đa chiều*: Thuật toán có khả năng áp dụng hiệu quả cho dữ liệu có số chiều khác nhau.
- *Dễ hiểu, cài đặt và khả dụng.*

Các yêu cầu này đồng thời là các tiêu chí để đánh giá hiệu quả của các phương pháp phân cụm dữ liệu, đây là các thách thức cho các nhà nghiên cứu trong lĩnh vực PCDL. Các yêu cầu này sẽ được đề cập đến cụ thể hơn khi đi vào khảo cứu chi tiết một số thuật toán PCDL được trình bày ở các chương sau.

2.5 MỘT SỐ THUẬT TOÁN PHÂN CỤM DỮ LIỆU ĐIỂN HÌNH

Có rất nhiều thuật toán được áp dụng trong phân cụm dữ liệu. Do đo trong phần này khóa luận trình bày một số thuật toán cơ bản, rất kinh điển trong phân cụm dữ liệu. Các thuật toán này được chia thành các họ thuật toán: Họ các thuật toán phân cụm phân hoạch (Patitional), họ các thuật toán phân cụm phân cấp (Hierachical), họ các thuật toán phân cụm dựa trên lưới và các thuật toán PCDL đặc thù khác như: các thuật toán phân cụm dựa trên mật độ, các thuật toán phân cụm dựa trên mô hình,…

2.5.1 Họ các thuật toán phân hoạch

Họ các thuật toán phân cụm phân hoạch bao gồm các thuật toán đề xuất đầu tiên trong lĩnh vực Data Mining cũng là các thuật toán được áp dụng nhiều trong thực tế như k-means, PAM (Partioning Around Medoids), CLARA (Clustering LARge Applications), CLARANS (Clustering LARge ApplicatioNS). Trước hết chúng ta đi khảo cứu thuật toán k-means, đây là một thuật toán kinh điển được kế thừa sử dụng rộng rãi.

2.5.1.1 Thuật toán k-means

Thuật toán phân hoạch K-means do MacQeen đề xuất trong lĩnh vực thống kê năm 1967, mục đích của thuật toán k-means là sinh ra k cụm dữ liệu $\{C_1, C_2, \ldots, C_k\}$ từ một tập dữ liệu chứa n đối tượng trong không gian d chiều $X_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ ($i = \overline{1,n}$), sao cho hàm tiêu chuẩn: $E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2(x - m_i)$ đạt giá trị tối thiểu. Trong đó: m_i là trọng tâm của cụm C_i, D là khoảng cách giữa hai đối tượng.

Trọng tâm của một cụm là một véc tơ, trong đó giá trị của mỗi phần tử của nó là trung bình cộng của các thành phần tương ứng của các đối tượng vectơ dữ liệu trong cụm đang xét. Tham số đầu vào của thuật toán là số cụm k, và tham số đầu ra của thuật toán là các trọng tâm của các cụm dữ liệu. Độ đo khoảng cách D giữa các đối tượng dữ liệu thường được sử dụng dụng là khoảng cách Euclide, bởi vì đây là mô hình khoảng cách dễ để lấy đạo hàm và xác định các cực trị tối thiểu. Hàm tiêu chuẩn và độ đo khoảng cách có thể được xác định cụ thể hơn tuỳ vào ứng dụng

hoặc các quan điểm của người dùng. Thuật toán k-means bao gồm các bước cơ bản như sau:

InPut: Số cụm k và các trọng tâm cụm $\{m_j\}_{j=1}^{k}$;
OutPut: Các cụm C_i ($i = \overline{1,k}$) và hàm tiêu chuẩn E đạt giá trị tối thiểu;

Begin

Bước 1: Khởi tạo:

Chọn k trọng tâm $\{m_j\}_{j=1}^{k}$ ban đầu trong không gian R^d (d là số chiều của dữ liệu) . Việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm.

Bước 2: Tính toán khoảng cách:

Đối với mỗi điểm Xi (1<=i<=n), tính toán khoảng cách của nó tới mỗi trọng tâm mj j=1,k. Và sau đó tìm trọng tâm gần nhất đối với mỗi điểm.

Bước 3: Cập nhật lại trọng tâm:

Đối với mỗi j=1,k, cập nhật trọng tâm cụm mj bằng các xác định trung bình cộng của các vectơ đối tượng dữ liệu.

Bước 4: Điều kiện dừng

Lặp các bước 2 và 3 cho đến khi các trọng tâm của cụm không thay đổi.

End.

Hình sau minh họa về một số hình dạng cụm dữ liệu khám phá được bởi k-means:

Hình2.2: Hình dạng cụm dữ liệu khám phá được bởi k-means

2.5.1.2 Thuật toán CLARA

CLARA (Clustering LARge Application) được Kaufman đề xuất năm 1990, thuật toán này nhằm khắc phục nhược điểm của thuật toán PAM trong trường hợp giá trị của k và n là lớn. CLARA tiến hành trích mẫu cho tập dữ liệu có n phần tử, nó áp dụng thuật toán PAM cho mẫu này và tìm ra các các đối tượng tâm *medoid* cho mẫu được trích từ dữ liệu này. Người ta thấy rằng, nếu mẫu dữ liệu được trích theo cách ngẫu nhiên, thì các medoid của nó xấp xỉ với các medoid của toàn bộ tập dữ liệu ban đầu. Để tiến tới một xấp xỉ tốt hơn, CLARA đưa ra nhiều cách lấy mẫu và thực hiện phân cụm cho mỗi trường hợp và tiến hành chọn kết quả phân cụm tốt

nhất khi thực hiện phân cụm trên các mẫu này. Để cho chính xác, chất lượng của các cụm được đánh giá thông độ phi tương tự trung bình của toàn bộ các đối tượng dữ liệu trong tập đối tượng ban đầu. Kết quả thực nghiệm chỉ ra rằng, 5 mẫu dữ liệu có kích thước 40+2k cho các kết quả tốt. Các bước thực hiện của thuật toán CLARA như hình sau:

Bước 1. Duyệt dãy i = 1 đến i= 5

Bước 2. Lấy một mẫu có 40 + 2k đối tượng dữ liệu ngẫu nhiên từ tập dữ liệu và áp dụng thuật toán PAM cho mẫu dữ liệu này nhằm để tìm các đối tượng medoid đại diện cho các cụm.

Bước 3. Đối với mỗi đối tượng O_j trong tập dữ liệu ban đầu, xác định đối tượng medoid tương tự nhất trong số k đối tượng medoid.

Bước 4. Tính độ phi tương tự trung bình cho phân hoạch các đối tượng dành ở bước trước, nếu giá trị này bé hơn giá trị tối thiểu hiện thời thì sử dụng giá trị này thay cho giá trị tối thiểu ở trạng thái trước, như vậy, tập k đối tượng medoid xác định ở bước này là tốt nhất cho đến thời điểm này.

Bước 5. Quay về bước 1.

Độ phức tạp tính toán của nó là $O(k(40+k)^2 + k(n-k))$, và CLARA có thể thực hiện đối với tập dữ liệu lớn. Chú ý đối với kỹ thuật tạo mẫu trong PCDL: kết quả phân cụm có thể không phụ thuộc vào tập dữ liệu khởi tạo nhưng nó chỉ đạt tối ưu cục bộ. Thí dụ như: Nếu các đối tượng *medoid* của dữ liệu khởi tạo không nằm trong mẫu, khi đó kết quả thu được không đảm bảo là tốt nhất được.

2.5.1.4 Thuật toán CLARANS

Thuật toán CLARANS nhằm để cải tiến cho chất lượng cũng như mở rộng áp dụng cho tập dữ liệu lớn. CLARANS cũng sử dụng các đối tượng trung tâm medoids làm đại diện cho các cụm dữ liệu. Ý tưởng cơ bản của CLARANS là không xem xét tất cả các khả năng có thể thay thể các đối tượng tâm medoids bởi một đối tượng khác, nó ngay lập tức thay thế các đối tượng tâm này nếu việc thay thế này có tác động tốt đến chất lượng phân cụm chứ không cần xác định cách thay thể tối ưu nhất. Một phân hoạch cụm phát hiện được sau khi thay thế đối tượng trung tâm được gọi là một láng giềng (Neighbor) của phân hoạch cụm trước đó. Số các láng giềng được hạn chế bởi tham số do người dùng đưa vào là **Maxneighbor,** quá trình lựa chọn các láng giềng này là hoàn toàn ngẫu nhiên. Tham số **Numlocal** cho phép người dùng xác định số vòng lặp tối ưu cục bộ được tìm kiếm. Không phải

tất các các láng giềng được duyệt mà chỉ có **Maxneighbor** số láng giềng được duyệt.

Thuật toán chi tiết CLARANS như biểu diễn như sau:

Input: *O, k, dist, numlocal,* and *maxneighbor;*
Output: k cụm dữ liệu;

```
CLARANS (int k, function dist, int numlocal, int maxneighbor)
BEGIN
    for (i = 1; i <= numlocal; i++) {
        current.create_randomly(k);
        j = 1;
        while (j < maxneighbor) {
            current.select_randomly(old, new);
            diff = current.calculate_distance_difference(old, new);
            if (diff < 0) {
                current.exchange(old, new);
                j = 1;
            }
            else j++; // end if
        } // end while
        dist = current.calculate_total_distance();
        if (dist < smallest_dist) {
          best = current;
           smallest_dist = dist;
        } // end if
    } // end for
END;
```

Trong đó:

Create_Randomly(k): tạo ngẫu nhiên k cụm dữ liệu, nghĩa là thuật toán lựa chọn ngẫu nhiên k đối tượng medoid từ n đối tượng dữ liệu.

Select_randomly(old, new): Thay thế một đối tượng tâm cụm medoid **old** bởi đối tượng khác **new.**

Calculate_distance_difference(old, new): Tính toán sự khác nhau về tổng khoảng cách giữa phân hoạch hiện thời và láng giềng của nó.

Exchange(old, new): Hoán đổi giữa đối tượng tâm cụm medoid old với đối tượng không phải là medoid new, sau khi hoán đổi vai trò của chúng cũng được hoán đổi.

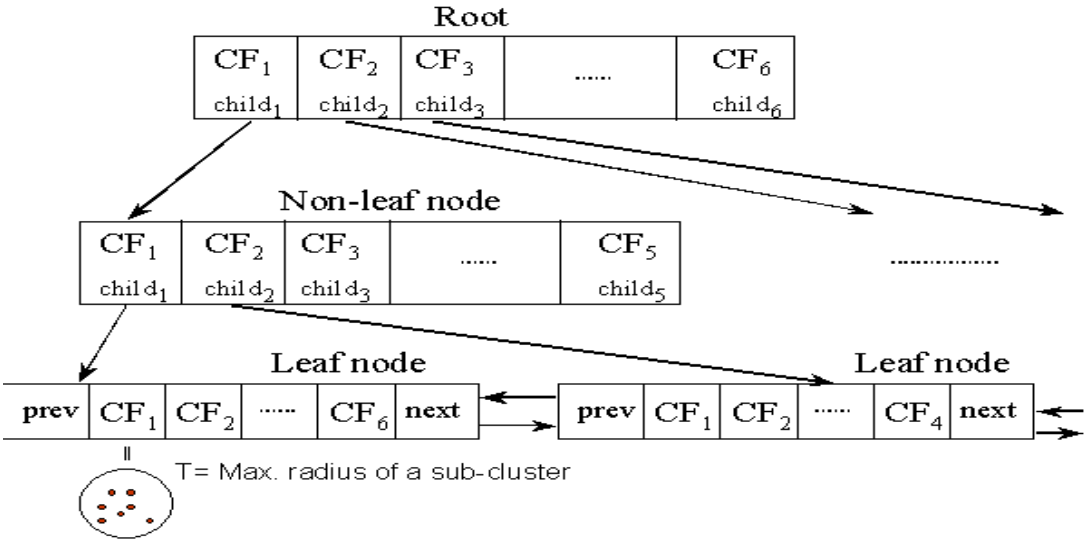
Calculate_total_distance(): Tính tổng khoảng cách cho mỗi phân hoạch.

Như vậy, quá trình hoạt động của CLARANS tương tự với quá trình hoạt động của thuật toán CLARA. Tuy nhiên, ở giai đoạn lựa chọn các trung tâm **medoid** của cụm dữ liệu, CLARANS lựa chọn một giải pháp tốt hơn bằng cách lấy ngẫu nhiên một đối tượng của k đối tượng trung tâm **medoid** của cụm và cố gắng thay thế nó với một đối tượng được chọn ngẫu nhiên trong (n-k) đối tượng còn lại, nếu không có giải pháp nào tốt hơn sau một số cố gắng lựa chọn ngẫu nhiên xác định, thuật toán dùng và cho kết quả phân cụm tối ưu cục bộ.

2.5.2 Các thuật toán phân cụm phân cấp

2.5.2.1 Thuật toán BIRCH

BIRCH *(Balanced Iterative Reducing and Clustering Using Hierarchies)* là thuật toán phân cụm phân cấp sử dụng chiến lược phân cụm trên xuống (top down). Ý tưởng của thuật toán là không cần lưu toàn bộ các đối tượng dữ liệu của các cụm trong bộ nhớ mà chỉ lưu các đại lượng thống kê. Đối với mỗi dữ liệu, BRICH chỉ lưu một bộ ba (n, LS, SS), trong đó n là số đối tượng trong cụm, LS là tổng các giá trị thuộc tính của các đối tượng trong cụm và SS là tổng bình phương của các giá trị thuộc tính của các đối tượng trong cụm. Các bộ ba này được gọi là các đặc trưng của cụm (*Cluster Features - CF*) và được lưu giữ trong một cây được gọi là cây CF (CF-tree). Hình dưới đây biểu thị một ví dụ về cây CF.



Hình 2.3:Cây CF được sử dụng bởi thuật toán BIRCH

Cây CF là cây cân bằng, nhằm để lưu trữ các đặc trưng của cụm (CF). Cây CF chứa các nút trong và nút lá, nút trong là nút chứa các nút con và nút lá thì không có

con. Nút trong lưu giữ tổng các đặc trưng cụm (CF) của các nút con của nó. Một cây CF được đặc trưng bởi hai tham số:

- *Yếu tố nhánh (Branching Factor -B)*: Nhằm xác định số tối đa các nút con của mỗi nút trong của cây, và
- *Ngưỡng (Threshold - T)*: Khoảng cách tối đa giữa bất kỳ một cặp đối tượng trong nút lá của cây, khoảng cách này còn gọi là đường kính của các cụm con được lưu tại các nút lá.

Hai tham số này có ảnh hưởng đến kích thước của cây CF. Thuật toán BIRCH thực hiện qua giai đoạn sau:

Giai đoạn 1: BIRCH duyệt tất cả các đối tượng trong CSDL và xây dựng một cây CF khởi tạo. Trong giai đoạn này, các đối tượng lần lượt được chèn vào nút lá gần nhất của cây CF (nút lá của cây đóng vai trò là cụm con), sau khi chèn xong thì tất cả các nút trong cây CF được cập nhật thông tin. Nếu đường kích của cụm con sau khi chèn là lớn hơn ngưỡng T, thì nút là được tách. Quá trình này lặp cho đến khi tất cả các đối tượng đều được chèn vào trong cây. Ở đây ta thấy rằng, mỗi đối tượng trong cây chỉ được đọc một lần, để lưu toàn bộ cây CF trong bộ nhớ thì cần phải điều chỉnh kích thước của cây CF thông qua điều chỉnh ngưỡng T.

Giai đoạn hai: BIRCH lựa chọn một thuật toán PCDL (như thuật toán phân cụm phân hoạch chẳng hạn) để thực hiện PCDL cho các nút lá của cây.

Tư tưởng thuật toán BIRCH được minh họa như sau:

1. Các đối tượng dữ liệu lần lượt được chèn vào cây CF, sau khi chèn hết các đối tượng ta thu được cây CF khởi tạo. Một đối tượng được chèn vào nút lá gần nhất tạo thành cụm con. Nếu đường kính của cụm con này lớn hơn T thì nút lá được tách. Khi một đối tượng thích hợp được chèn vào nút lá, tất cả các nút trỏ tới gốc của cây được cập nhật với các thông tin cần thiết.

2.Nếu cây CF hiện thời không có đủ bộ nhớ trong thì tiến hành cây dựng một cây CF nhỏ hơn: Kích thước của cây CF được điều khiển bởi tham số T và vì vậy việc chọn một giá trị lớn hơn cho nó sẽ hoà nhập một số các cụm con thành một cụm, điều này làm cho cây CF nhỏ hơn. Bước này không cần yêu cầu bắt đầu đọc dữ liệu lại từ đầu nhưng vẫn đảm bảo hiệu chỉnh cây dữ liệu nhỏ hơn.

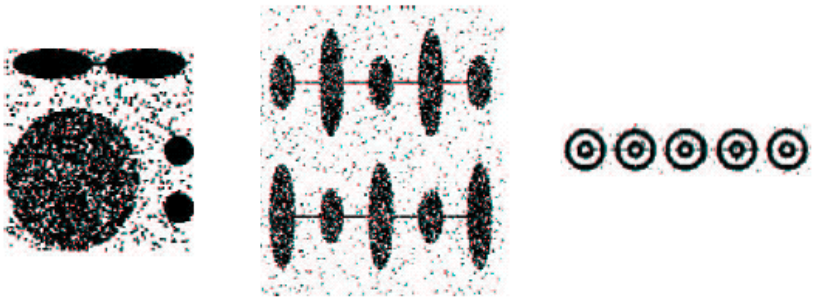
3. Thực hiện phân cụm: Các nút lá của cây CF lưu giữ các đại lượng thống kê của các cụm con. Trong bước này, BIRCH sử dụng các đại lượng thống kê này để áp dụng một số kỹ thuật phân cụm thí dụ như k-means và tạo ra một khởi tạo cho phân cụm.

4. Phân phối lại các đối tượng dữ liệu bằng cách dùng các đối tượng trọng tâm cho các cụm đã được khám phá từ bước 3: Đây là một bước tuỳ chọn để duyệt lại tập dữ liệu và gán nhãn lại cho các đối tượng dữ liệu tới các trọng tâm gần nhất. Bước này nhằm để gán nhãn cho các dữ liệu khởi tạo và loại bỏ các đối tượng ngoại lai.

2.5.2.2 Thuật toán CURE

Việc chọn một cách biểu diễn cho các cụm có thể nâng cao chất lượng phân cụm. Thuật toán CURE *(Clustering Using REpresentatives)* là thuật toán sử dụng chiến lược dưới lên (Bottom up) của kỹ thuật phân cụm phân cấp.CURE sử dụng nhiều đối tượng để diễn tả cho mỗi cụm dữ liệu. Các đối tượng đại diện cho cụm này ban đầu được lựa chọn rải rác đều ở các vị trí khác nhau, sau đó chúng được di chuyển bằng cách co lại theo một tỉ lệ nhất định. Tại mỗi bước của thuật toán, hai cụm có cặp đối tượng đại diện gần nhất (đối tượng thuộc về mỗi cụm) sẽ được trộn lại thành một cụm.

Với cách thức sử dụng nhiều hơn một điểm đại diện cho các cụm, CURE có thể khám phá được các cụm có các dạng hình thù và kích thước khác nhau trong CSDL lớn. Việc co các đối tượng đại diện lại có tác dụng làm giảm tác động của các phần tử ngoại lai, vì vậy, CURE có khả năng xử lý đối với các phần tử ngoại lai. Hình 17 dưới đây là thí dụ về các dạng và kích thước cụm dữ liệu được khám phá bởi CURE:



Hình 2.4:Các cụm dữ liệu được khám phá bởi CURE

Để áp dụng với CSDL lớn, CURE sử dụng lấy mẫu ngẫu nhiên và phân hoạch. Mẫu dữ liệu được xác định ngẫu nhiên là phân hoạch đầu tiên, CURE tiến hành phân cụm trên mỗi phân hoạch. Quá trình này lặp lại cho đến khi ta thu được phân hoạch đủ tốt. Các cụm thu được sau đó lại được phân cụm nhằm để thu được các cụm con cần quan tâm. Thuật toán CURE được thực hiện qua các bước cơ bản như sau:

1. Chọn một mẫu ngẫu nhiên từ tập dữ liệu ban đầu;

2. Phân hoạch mẫu này thành nhiều nhóm dữ liệu có kích thước bằng nhau: Ý tưởng chính ở đây là phân hoạch mẫu thành p nhóm dữ liệu bằng nhau, kích thước của mỗi phân hoạch là n'/p (n' là kích thước của mẫu) ;

3. Phân cụm các điểm của mỗi nhóm: chúng ta thực hiện PCDL cho các nhóm cho đến khi mỗi nhóm được phân thành n'/pq cụm (với q>1);

4. Loại bỏ các phần tử ngoại lai: Trước hết, khi các cụm được hình thành cho đến khi số các cụm giảm xuống một phần so với số các cụm ban đầu. Sau đó, trong trường hợp các phần tử ngoại lai được lấy mẫu cùng với quá trình pha khởi tạo mẫu dữ liệu, thuật toán sẽ tự động loại bỏ các nhóm nhỏ.

5. Phân cụm các cụm không gian: Các đối tượng đại diện cho các cụm di chuyển về hướng trung tâm cụm, nghĩa là chúng được thay thế bởi các đối tượng gần trung tâm hơn.

6. Đánh dấu dữ liệu với các nhãn tương ứng

Độ phức tạp tính toán của thuật toán CURE là $O(n^2\log(n))$. CURE là thuật toán tin cậy trong việc khám phá các cụm với hình thù bất kỳ và có thể áp dụng tốt đối trên các tập dữ liệu hai chiều. Tuy nhiên, nó lại rất nhạy cảm với các tham số như là tham số các đối tượng đại diện, tham số co của các phần tử đại diện. Nhìn chung thì BIRCH tốt hơn so với CURE về độ phức tạp, nhung kém về chất lượng phân cụm. Các hai thuật toán này có thể xử lý các phần tử ngoại lai tốt.

2.5.3 Các thuật toán phân cụm dựa trên mật độ

Các cụm có thể được xem như các vùng có mật độ cao, được tách ra bởi các vùng không có hoặc ít mật độ. Khái niệm mật độ ở đây được xem như là các số các đối tượng láng giềng.

2.5.3.1 Thuật toán DBSCAN

Thuật toán phân cụm dựa trên mật độ thông dụng nhất là thuật toán DBSCAN (*Density - Based Spatial Clustering of Applications with noise*). Thuật toán đi tìm các đối tượng mà có số đối tượng láng giềng lớn hơn một ngưỡng tối thiểu. Tìm tất cả các đối tượng mà các láng giềng của nó thuộc về lớp các đối tượng đã xác định ở trên, một cụm được xác định bằng một tập tất cả các đối tượng liên thông mật độ với các láng giềng của nó. DBSCAN có thể tìm ra các các cụm với hình thù bất kỳ, trong khi đó tại cùng một thời điểm ít bị ảnh hưởng bởi thứ tự của các đối dữ liệu nhập vào. Khi có một đối tượng được chèn vào chỉ tác động đến một láng giềng xác định. Mặt khác, DBSCAN yêu cầu người dùng xác định bán kính Eps của các láng giềng và số các láng giềng tối thiểu Minpts, các tham số này khó mà xác định được tối ưu, thông thường nó được xác định bằng phép chọn ngẫu nhiên hoặc theo kinh nghiệm. Người ta áp dụng chỉ số không gian để giúp xác định các láng giềng của một đối tượng dữ liệu do vậy độ phức tạp của DBSCAN đã được cải tiến là O(nlogn) so với dộ phức tạp của DBSCAN là $O(n^2)$ trong trường hợp nếu không áp dụng cấu trúc chỉ số. Khoảng cách Euclide được sử dụng để đo sự tương tự giữa các đối tượng nhưng không hiệu quả đối với dữ liệu đa chiều.

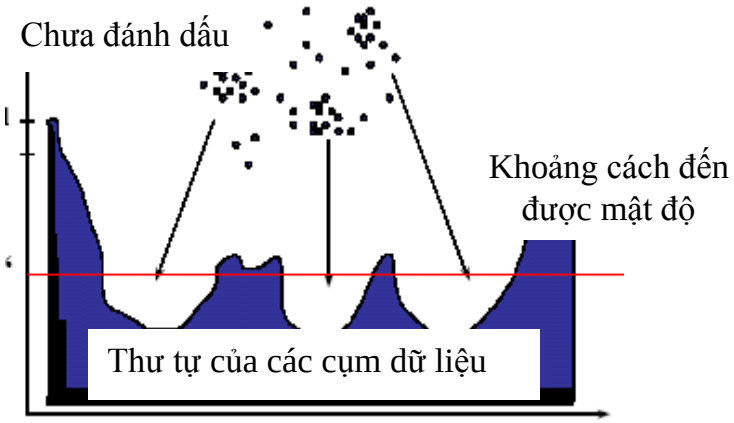
```
{----------Mô đun chương trình chính-------}
DBSCAN (SetOfPoints, Eps, MinPts)
// SetOfPoints is UNCLASSIFIED
ClusterId:= nextId(NOISE);
FOR i FROM 1 TO SetOfPoints.size DO
Point:= SetOfPoints.get(i);
IF Point.ClId = UNCLASSIFIED THEN
IF ExpandCluster(SetOfPoints, Point,
ClusterId, Eps, MinPts) THEN
ClusterId:= nextId(ClusterId)
END IF
END IF
```

```
SetOfPoints.changeClIds(seeds,ClId);
seeds.delete(Point);
WHILE seeds <> Empty DO
currentP:= seeds.first();
result:= SetOfPoints.regionQuery(currentP,
Eps);
IF result.size >= MinPts THEN
FOR i FROM 1 TO result.size DO
resultP:= result.get(i);
IF resultP.ClId
IN {UNCLASSIFIED, NOISE} THEN
IF resultP.ClId = UNCLASSIFIED THEN
```

END FOR
END; // DBSCAN
{----------Thủ tục Expand ----------}
ExpandCluster(SetOfPoints, Point, ClId, Eps, MinPts): Boolean;
seeds:=SetOfPoints.regionQuery(Point,Eps);
IF seeds.size<MinPts THEN // no core point
SetOfPoint.changeClId(Point,NOISE);
RETURN False;
ELSE // all points in seeds are density-
// reachable from Point
	seeds.append(resultP);
	END IF;
	SetOfPoints.changeClId(resultP,ClId);
	END IF; // UNCLASSIFIED or NOISE
	END FOR;
	END IF; // result.size >= MinPts
	seeds.delete(currentP);
	END WHILE; // seeds <> Empty
	RETURN True;
	END IF
	END; // ExpandCluster
	{----------End----------}

2.5.3.2 Thuật toán OPTICS

Đây là thuật toán mở rộng cho thuật toán DBSCAN, bằng cách giảm bớt các tham số đầu vào. OPTICS *(Ordering Points To Identify the Clustering Structure)* sắp xếp các cụm theo thứ tự tăng dần nhằm tự động phân cụm dữ liệu. Thứ tự này diễn tả cấu trúc dữ liệu phân cụm dựa trên mật độ chứa thông tin tương đương với phân cụm dựa trên mật độ với một dãy các tham số đầu vào. OPTICS xem xét bán kính tối thiểu nhằm xác định các láng giềng phù hợp với thuật toán. DBSCAN và OPTICS tương tự với nhau về cấu trúc và có cùng độ phức tạp: O(nLogn) (N là kích thước của tập dữ liệu).



Hình 2.5: Hình minh họa phân cụm OPTICS

2.5.3.3. Thuật toán DENCLUE

DENCLUE *(DENsity - Based CLUstEring)* là thuật toán PCDL dựa trên một tập các hàm phân phối mật độ. Ý tưởng chính của thuật toán này như sau:

- Sự tác động của một đối tượng tới láng giềng của nó được xác định bởi hàm ảnh hưởng (*Influence Function*).
- Mật độ toàn cục của không gian các đối tượng được mô hình như là tổng tất cả các hàm ảnh hưởng của các đối tượng.
- Các cụm được xác định bởi các đối tượng mật độ cao (*density attactors),* các đối tượng nàylà các điểm cực đại của hàm mật độ toàn cục.

Hình 2.6:Hình minh họa DENCLUE với hàm phân phối Gaussian

Chúng ta thấy rằng, DENCLUE phụ thuộc nhiều vào ngưỡng nhiễu ξ (*Noise Threshold*) và tham số mật độ δ, nhưng DENCLUE có các ưu điểm sau:

- Có cơ sở toán học vững chắc
- Có khả năng xử lý các phần tử ngoại lai.
- Cho phép khám phá ra các cụm với hình thù bất kỳ ngay cả đối với các dữ liệu đa chiều

Độ phức tạp tính toán của DENCLUE là O(nlogn). Các thuật toán dựa trên mật độ không thực hiện kỹ thuật phân mẫu trên tập dữ liệu như trong các thuật toán phân cụm phân hoạch, vì điều này có thể làm tăng thêm độ phức tạp do có sự khác nhau giữa mật độ của các đối tượng trong mẫu với mật độ của toàn bộ dữ liệu.

CHƯƠNG 3: HỆ THỐNG ĐÁNH GIÁ THÔNG TIN SẢN PHẨM

Trong chương này, khóa luận trình bày mô tả về bài toán, bản phân tích thiết kế bài toán. Sau đó tiến hành cài đặt chương trình sử dụng công cụ visual studio 2010 bằng ngôn ngữ lập trình C#. Sau khi cài đặt chương trình ta sẽ đánh giá thử nghiệm hệ thống có phù hợp với yêu cầu hay không.

3.1 Phát biểu bài toán

Một phần mềm cho phép người dùng đăng nhập vào để tìm kiếm thông tin về một sản phẩm được mô tả cụ thể như sau:

Hệ thống cho phép người dùng nhập thông tin về sản phẩm cần tìm.

+ Thông tin về sản phẩm:

Sản phẩm được phân chia làm nhiều nhóm sản phẩm,mỗi nhóm sản phẩm có nhiều loại sản phẩm.

Thông tin nhóm sản phẩm miêu tả như sau: Mã nhóm sản phẩm, Tên nhóm sản phẩm.

Thông tin về loại sản phẩm: Mã loại sản phẩm, Tên loại sản phẩm.

Trong loại sản phẩmthì có nhiều sản phẩm. Thông tin về sản phẩmgồm có: Mã sản phẩm, Tên sản phẩm.

Khi người dùng đăng nhập vào hệ thông thì kê khai sản phẩm mà mình cần tìm kiếm. Thông tin sản phẩm gồm có:

Mã sản phẩm

Tên sản phẩm

...

Mô tả hoạt động của hệ thống:

Người dùng đăng nhập vào hệ thống để nhập thông tin sản phẩm cần tìm kiếm.

Hệ thống cung cấp chức năng tụ động tìm kiếm sản phẩm trên mạng Internet thông qua các máy tìm kiếm.

Hệ thống có giao diện hợp lý, thuận lợi cho người dùng.

Có báo cáo kết quả tìm kiếm.

3.2 Xác định mô hình nghiệp vụ

3.2.1 Các chức năng nghiệp vụ

Ta có thể xác định các chức năng nghiệp vụ hệ thống như sau:

Tham chiếu	**Chức năng**
R_1	**Cập nhật các danh mục**
R_{11}	Cập nhật thông tin nhóm sản phẩm
R_{12}	Cập nhật thông tin loại sản phẩm
R_{13}	Cập nhật thông tin sản phẩm
R_{14}	Cập nhật thông tin Search Engine
R_{15}	Cập nhật thông số tìm kiếm
R_2	**Tìm kiếm**
R_3	**Báo cáo**

Bảng 3.1: Bảng xác định các chức năng nghiệp vụ của hệ thống

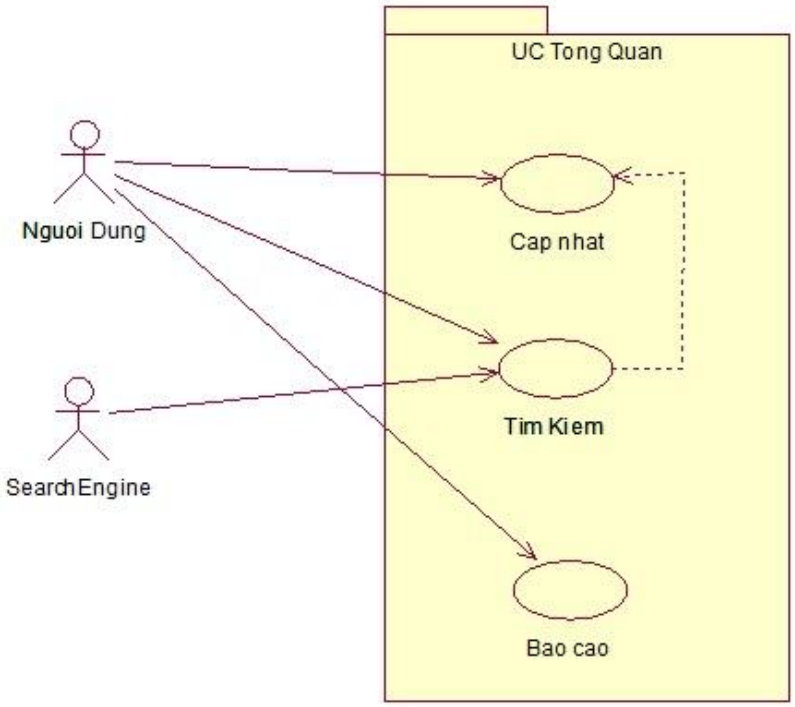
Các tác nhân hệ thống

Tác nhân là một bộ phận bên ngoài hệ thống nhưng có tương tác với hệ thống. Nó chính là đối tượng mà hệ thống phục vụ hoặc cần cung cấp dữ liệu. Hệ thống gồm các tác nhân sau:

Tác nhân	Vai trò
Người dùng	Là một tổ chức hoặc cá nhân đăng nhập vào hệ thống
Máy tìm kiếm	Là máy tìm kiếm trên mạng. Cụ thể là Google.

Bảng 3.2: Bảng xác định tác nhân của hệ thống.

3.2.2 Biểu đồ Use Case tổng quan



Hình 3.1: Biểu đồ Use Case tổng quan

Chương trình thử nghiệm tìm kiếm xác định được các ca sử dụng và tác nhân như sau:

Gói ca sử dụng	Các ca sử dụng chi tiết	Tác nhân
1.Cập nhật	UC1. Cập nhật thông tin nhóm sản phẩm UC2. Cập nhật thông tin loại sản phẩm UC3. Cập nhật thông tin sản phẩm UC4.Cập nhật thông tin Search Engine UC5. Cập nhật thông số tìm kiếm	Người dùng
2.Tìm kiếm	UC6. Tìm kiếm	Người dùng Máy tìm kiếm
3.Báo cáo	UC7. Lập báo cáo	Người dùng

Bảng 3.3: Bảng mô tả các ca sử dụng và tác nhân

3.2.3 Mô tả khái quát các hệ con

Hệ thống gồm ba hệ con:

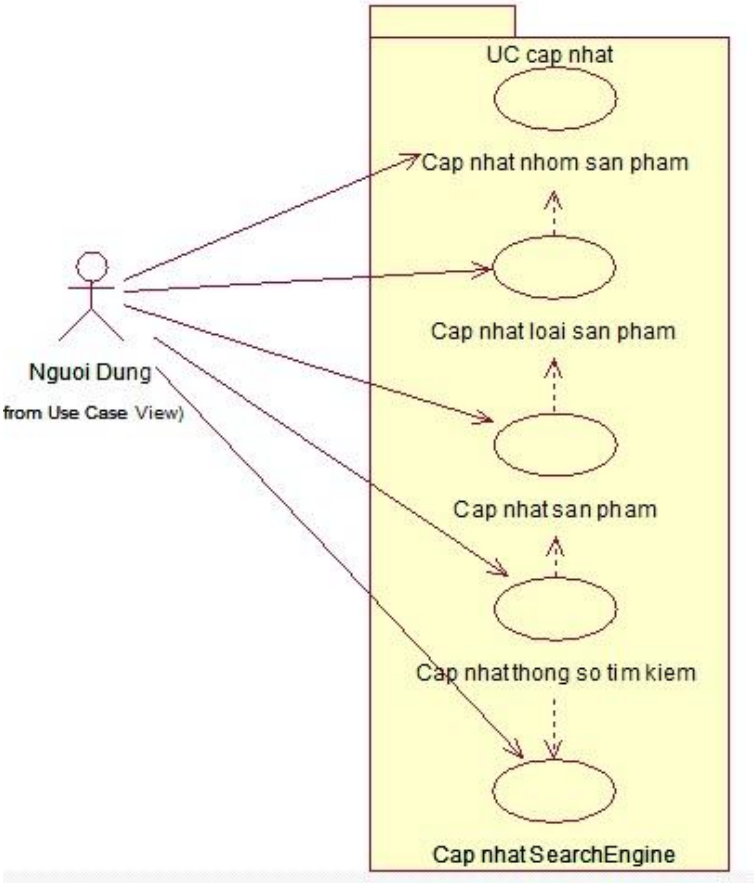
Cập nhật: Có tác nhân là người dùng. Có tác dụng cập nhật các thông tin đầu vào như: Nhóm sản phẩm, loại sản phẩm, sản phẩm, thông tin tìm kiếm.

Tìm kiếm: có tác nhân là máy tìm kiếm. Máy tìm kiếm nhận lệnh từ hệ thống và trả lại các kết quả tìm kiếm là các đường link

Báo cáo: Có các tác nhân là người dùng. Có tác dụng lập báo cáo và in báo cáo về các kết quả tìm kiếm, đánh giá.

3.2.4 Các mô hình ca sử dụng chi tiết

3.2.4.1 Gói ca sử dụng "Cập nhật"



Hình 3.2: Biểu đồ ca sử dụng gói "Cập nhật"

+ Mô tả chi tiết ca sử dụng

Ca sử dụng "Cập nhật nhóm sản phẩm"

Tên ca sử dụng	Cập nhật nhóm sản phẩm
Tác nhân	Người dùng
Mục đích	Cập nhật nhóm sản phẩm
Mô tả khái quát	Người dùng cập nhật thông tin các nhóm sản phẩm
Các tham chiếu	R11

Bảng 3.4: Bảng mô tả ca sử dụng cập nhật nhóm sản phẩm

Ca sử dụng "Cập nhật loại sản phẩm"

Tên ca sử dụng	Cập nhật loại sản phẩm
Tác nhân	Người dùng
Mục đích	Cập nhật loại sản phẩm
Mô tả khái quát	Người dùng cập nhật thông tin loại sản phẩm
Các tham chiếu	R_{12}

Bảng 3.5: Bảng mô tả ca sử dụng cập nhật loại sản phẩm

Ca sử dụng "Cập nhật sản phẩm"

Tên ca sử dụng	Cập nhật sản phẩm
Tác nhân	Người dùng
Mục đích	Cập nhật sản phẩm
Mô tả khái quát	Người dùng cập nhật thông tin các sản phẩm
Các tham chiếu	R_{13}

Bảng 3.6: Bảng mô tả ca sử dụng cập nhật sản phẩm

Ca sử dụng "Cập nhật search engine"

Tên ca sử dụng	Cập nhật search engine
Tác nhân	Người dùng
Mục đích	Cập nhật search engine
Mô tả khái quát	Người dùng cập nhật thông tin các search engine
Các tham chiếu	R_{14}

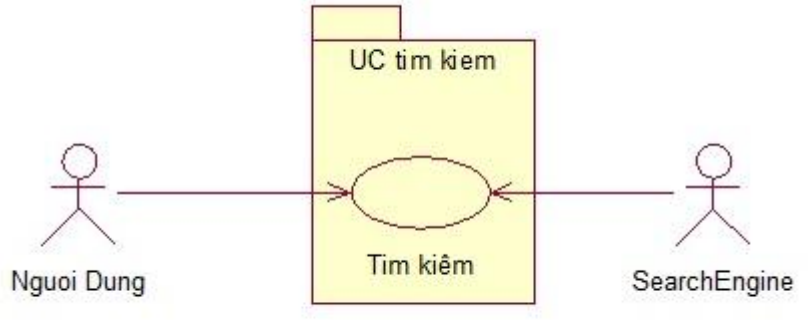
Bảng 3.7: Bảng mô tả ca sử dụng cập nhật Search Engine

Ca sử dụng "Cập nhật thông số tìm kiếm"

Tên ca sử dụng	Cập nhật thông số tìm kiếm
Tác nhân	Người dùng
Mục đích	Cập nhật các thông số cho việc tìm kiếm
Mô tả khái quát	Người dùng cập nhạt các thông số giúp cho việc tìm kiếm đạt kết quả mong muốn
Các tham chiếu	R_{15}

Bảng 3.8: Bảng mô tả ca sử dụng cập nhật thông số tìm kiếm

3.2.4.2 Gói ca sử dụng "Tìm kiếm"



Hình 3.3: Biểu đồ ca sử dụng gói "Tìm kiếm"

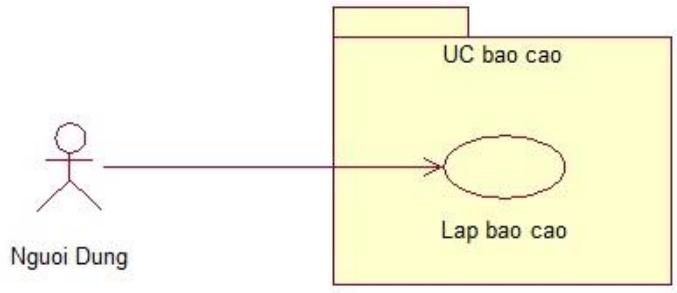
+ Mô tả chi tiết ca sử dụng

Ca sử dụng "Tìm kiếm"

Tên ca sử dụng	Nhận lệnh tìm kiếm
Tác nhân	Máy tìm kiếm, người dùng
Mục đích	Tìm kiếm thông tin
Mô tả khái quát	Máy tìm kiếm nhận các lệnh tìm kiếm và tiến hành tìm kiếm trên Internet
Các tham chiếu	R_2

Bảng 3.9: Bảng mô tả ca sử dụng tìm kiếm

3.2.4.3 Gói ca sử dụng "Báo cáo"



Hình 3.4: Biểu đồ ca sử dụng gói "Báo cáo"

+ Mô tả chi tiết ca sử dụng

Ca sử dụng "Báo cáo"

Tên ca sử dụng	Lập báo cáo
Tác nhân	Người dùng
Mục đích	Tổng hợp các kết quả tìm kiếm, tạo báo cáo
Mô tả khái quát	Hệ thống tổng hợp các kết quả tìm kiếm, tạo báo cáo cho người dùng
Các tham chiếu	R_3

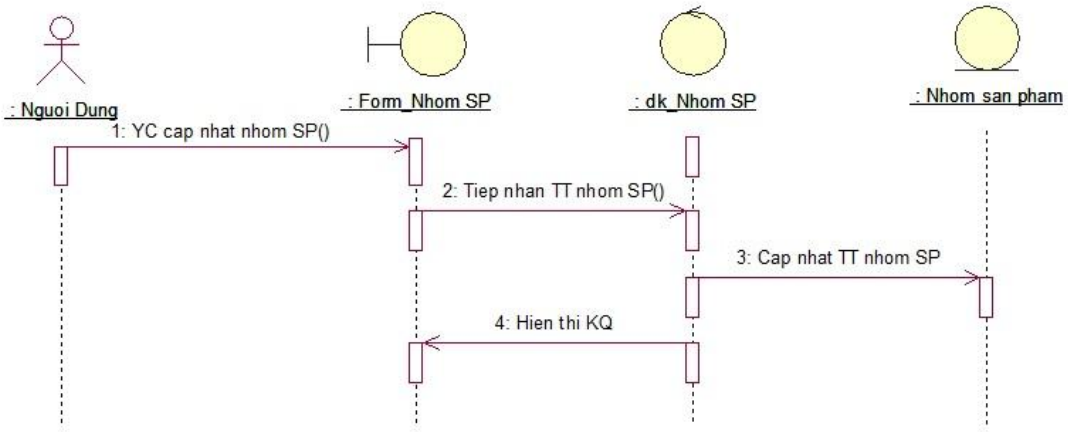
Bảng 3.10: Bảng mô tả ca sử dụng báo cáo

3.3 Phân tích hệ thống

3.3.2 Phân tích gói ca sử dụng "Cập nhật các danh mục"

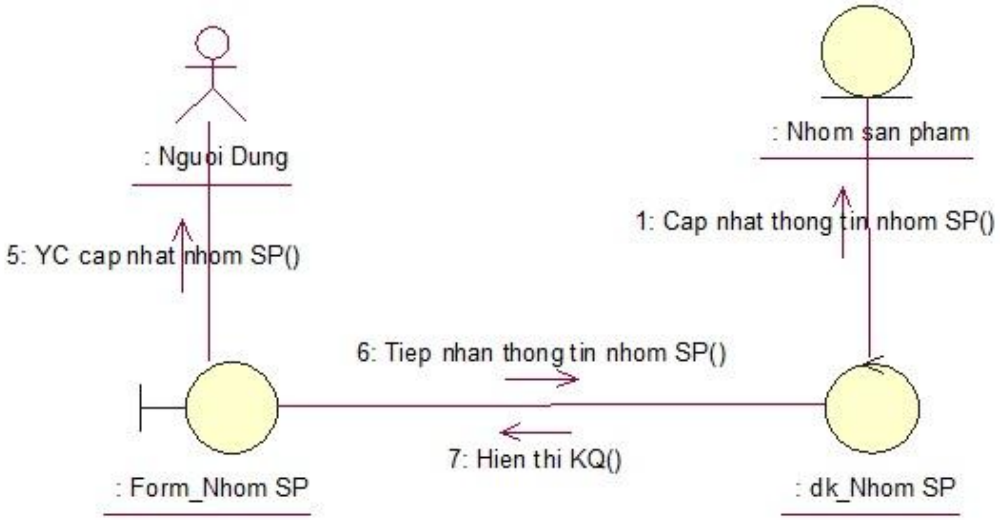
3.3.2.1 Ca sử dụng "Cập nhật nhóm sản phẩm"

Biểu đồ tuần tự thực thi ca sử dụng



Hình 3.5: Biểu đồ tuần tự thực thi ca sử dụng "Cập nhật nhóm sản phẩm"

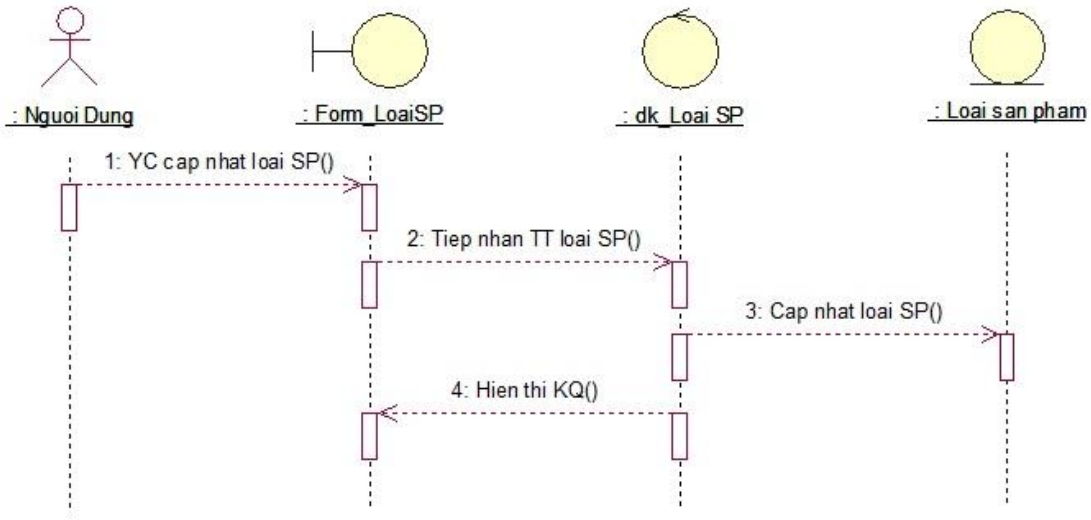
Biểu đồ cộng tác thực thi ca sử dụng



Hình 3.6: Biểu đồ cộng tác thực thi ca sử dụng "Cập nhật nhóm sản phẩm"

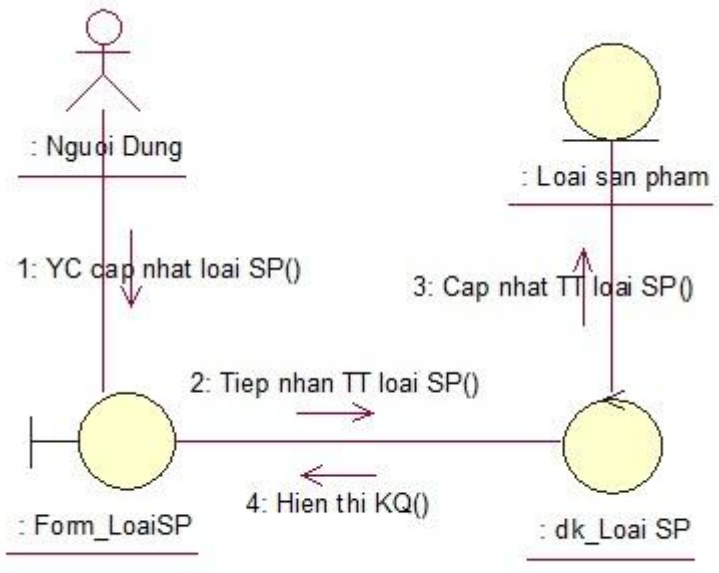
3.3.2.2 Ca sử dụng "Cập nhật loại sản phẩm"

Biểu đồ tuần tự thực thi ca sử dụng



Hình 3.7: Biểu đồ tuần tự thực thi ca sử dụng "Cập nhật loạisản phẩm"

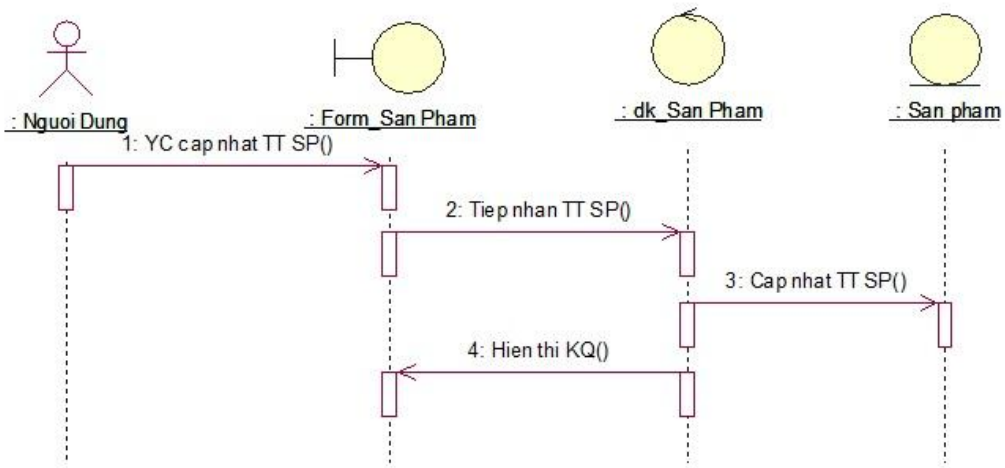
Biểu đồ cộng tác thực thi ca sử dụng



Hình 3.8: Biểu đồ cộng tác thực thi ca sử dụng "Cập nhật loạisản phẩm"

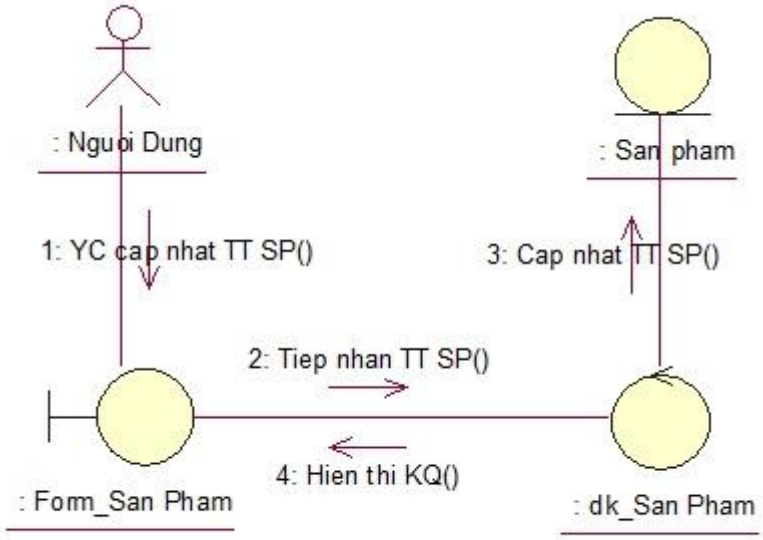
3.3.2.3 Ca sử dụng "Cập nhật sản phẩm"

Biểu đồ tuần tự thực thi ca sử dụng



Hình 3.9: Biểu đồ tuần tự thực thi ca sử dụng "Cập nhật sản phẩm"

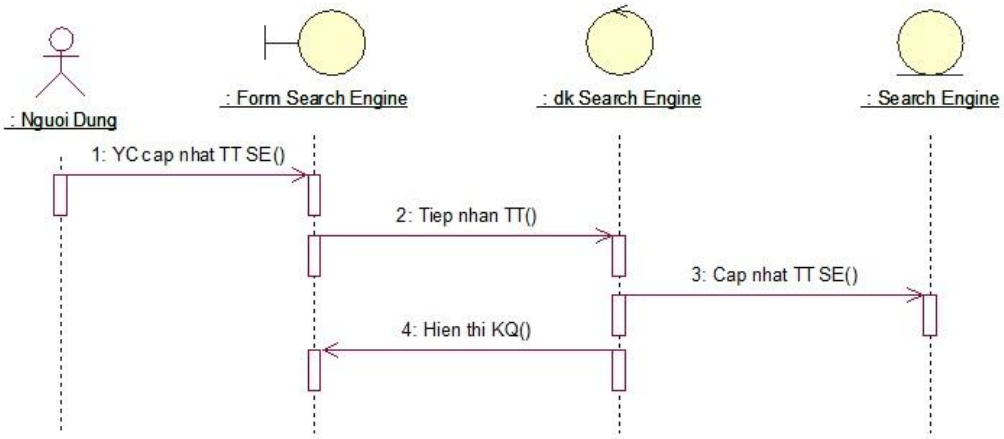
Biểu đồ cộng tác thực thi ca sử dụng



Hình 3.10: Biểu đồ cộng tác thực thi ca sử dụng "Cập nhật sản phẩm"

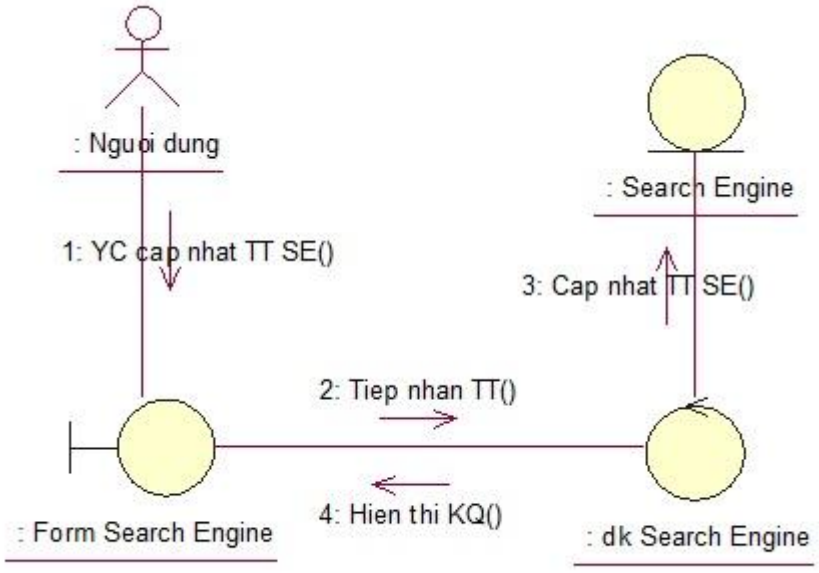
3.3.2.4 Ca sử dụng "Cập nhật Search engine"

Biểu đồ tuần tự thực thi ca sử dụng



Hình 3.11: Biểu đồ tuần tự thực thi ca sử dụng "Cập nhật Search Engine"

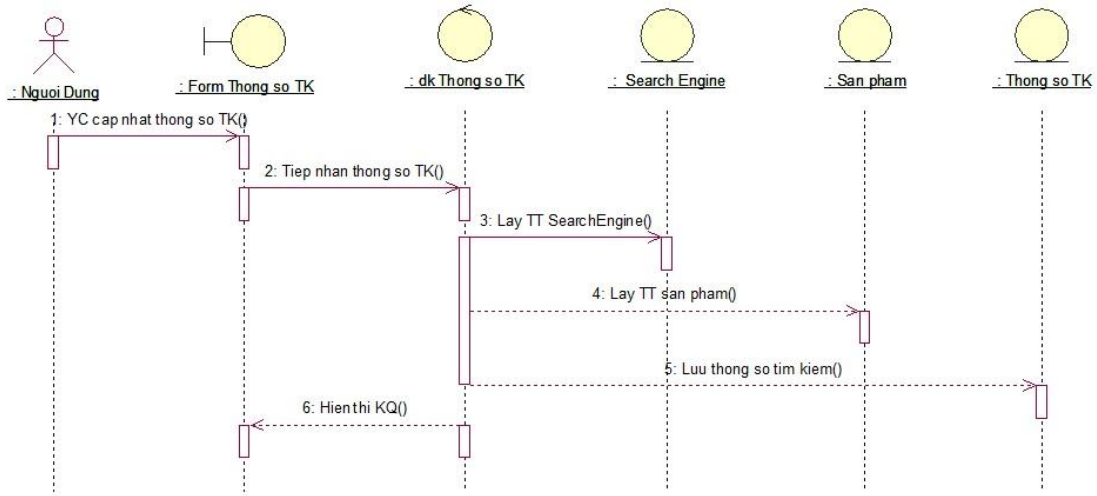
Biểu đồ cộng tác thực thi ca sử dụng



Hình 3.12: Biểu đồ cộng tác thực thi ca sử dụng "Cập nhật Search Engine"

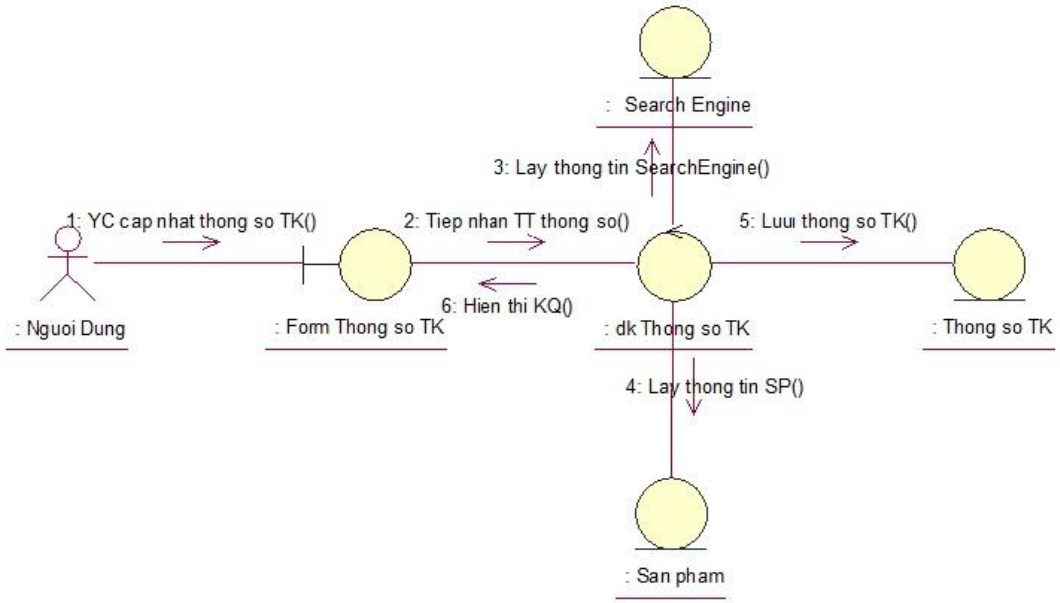
3.3.2.5 Ca sử dụng "Cập nhật thông số tìm kiếm"

Biểu đồ tuần tự thực thi ca sử dụng



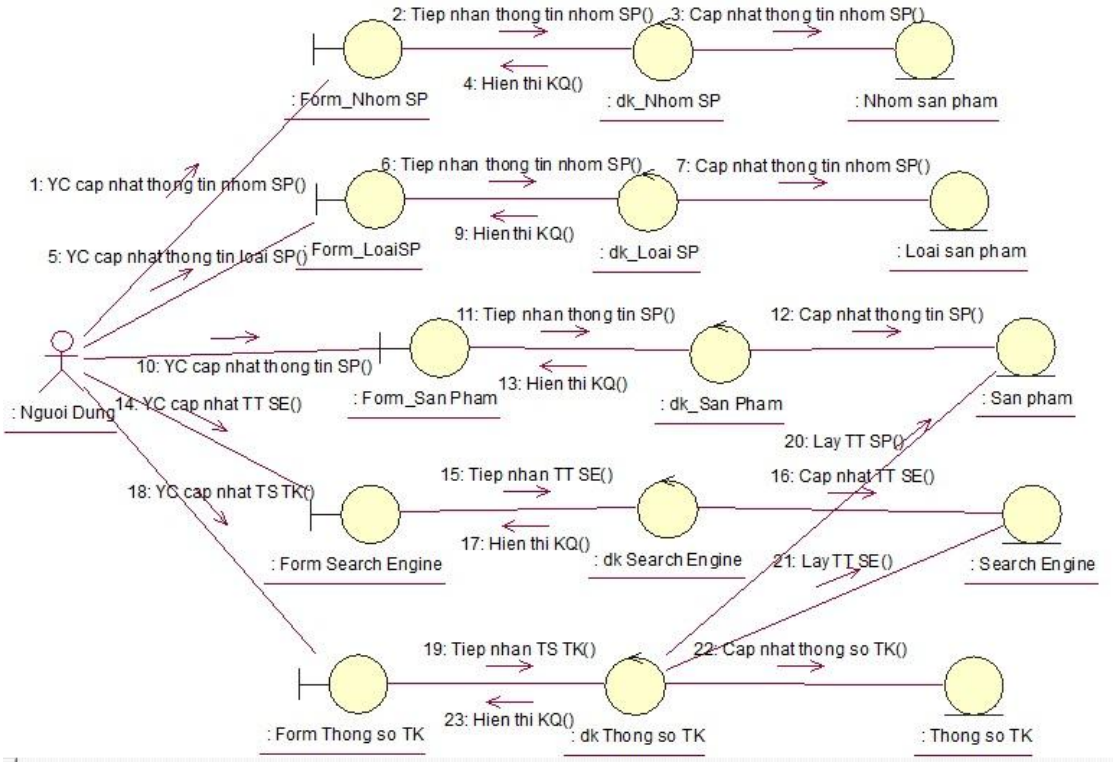
Hình 3.13: Biểu đồ tuần tự thực thi ca sử dụng "Cập nhật thông số tìm kiếm"

Biểu đồ cộng tác thực thi ca sử dụng



Hình 3.14: Biểu đồ cộng tác thực thi ca sử dụng "Cập nhật thông số tìm kiếm"

3.3.2.6 Mô hình phân tích gói ca sử dụng "Cập nhật"

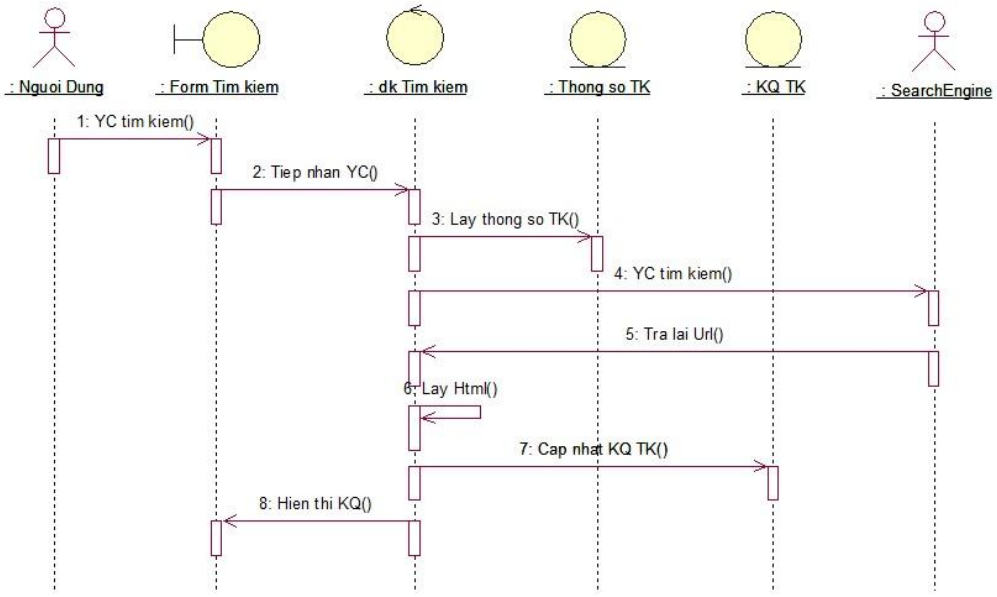


Hình 3.15: Mô hình phân tích gói ca "Cập nhật"

3.3.3 Phân tích gói ca sử dụng "Tìm kiếm"

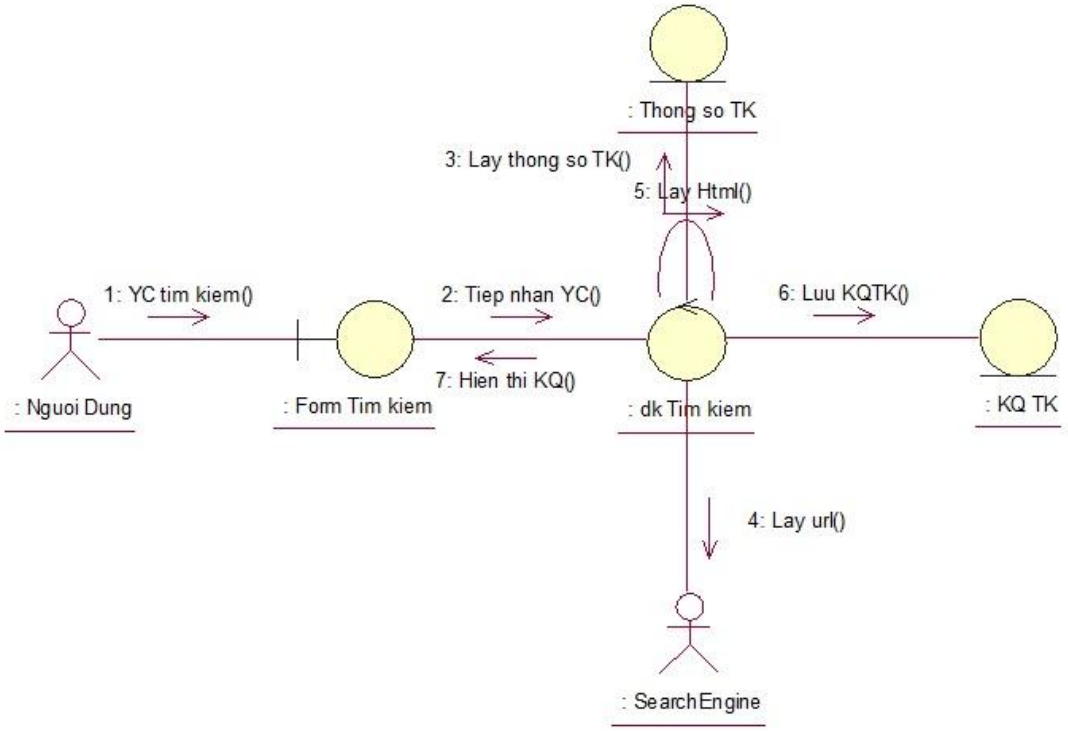
3.3.3.1 Ca sử dụng "Nhận lệnh tìm kiếm"

Biểu đồ tuần tự thực thi ca sử dụng



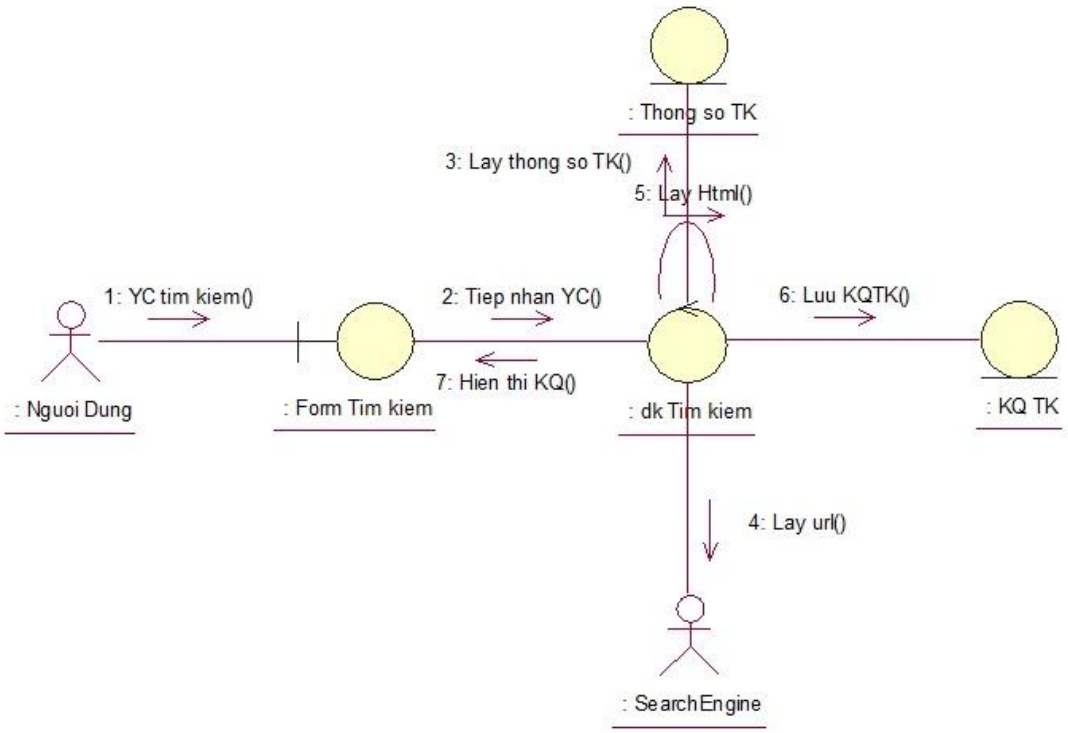
Hình 3.16: Biểu đồ tuần tự thực thi ca sử dụng "Tìm kiếm"

Biểu đồ cộng tác thực thi ca sử dụng



Hình 3.17: Biểu đồ cộng tác thực thi ca sử dụng "Tìm kiếm"

3.3.3.2 Mô hình phân tích gói ca sử dụng "Tìm kiếm"

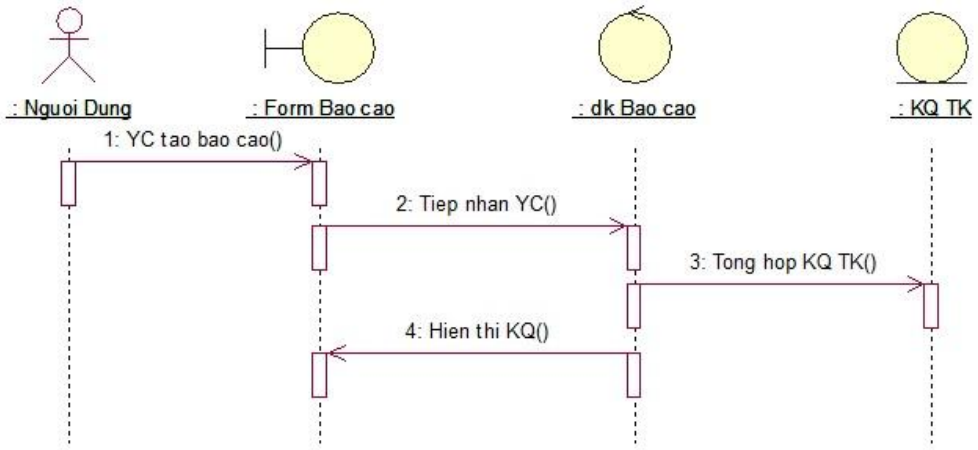


Hình 3.18: Mô hình phân tích gói ca"Tìm kiếm"

3.3.4 Phân tích gói ca sử dụng "Báo cáo"

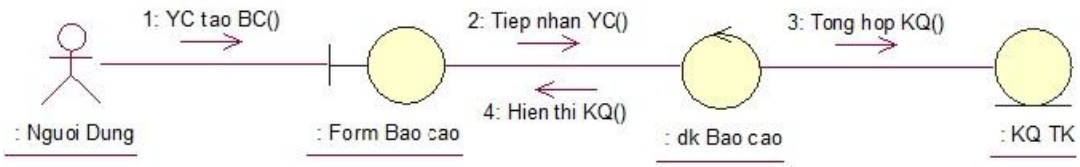
3.3.4.1 Ca sử dụng "Lập báo cáo"

Biểu đồ tuần tự thực thi ca sử dụng



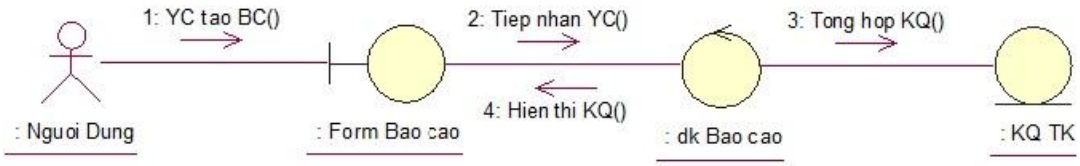
Hình 3.19: Biểu đồ tuần tự thực thi ca sử dụng "Lập báo cáo"

Biểu đồ cộng tác thực thi ca sử dụng



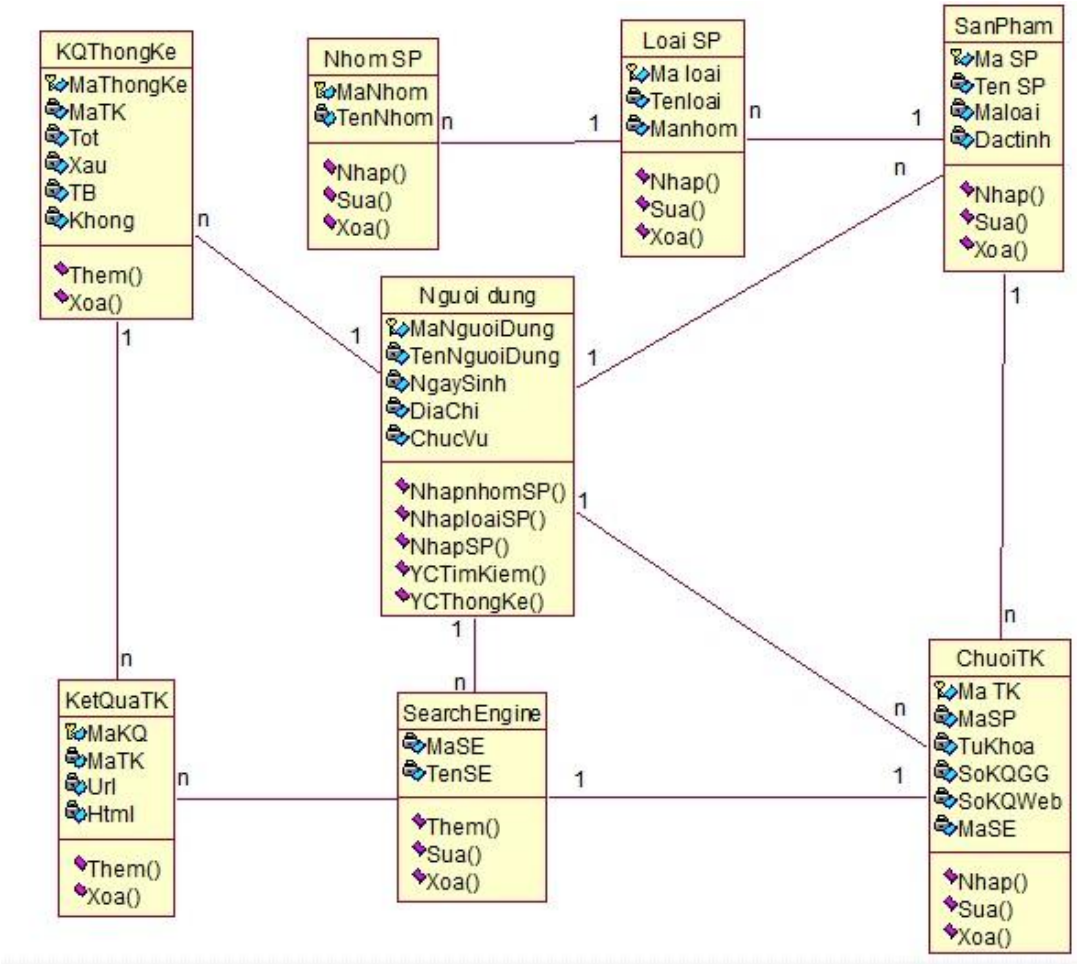
Hình 3.20: Biểu đồ cộng tác thực thi ca sử dụng "Báo cáo"

3.3.4.2 Mô hình phân tích gói ca sử dụng "Báo cáo"



Hình 3.21: Mô hình phân tích gói ca"Báo cáo"

3.4 Thiết kế hệ thống



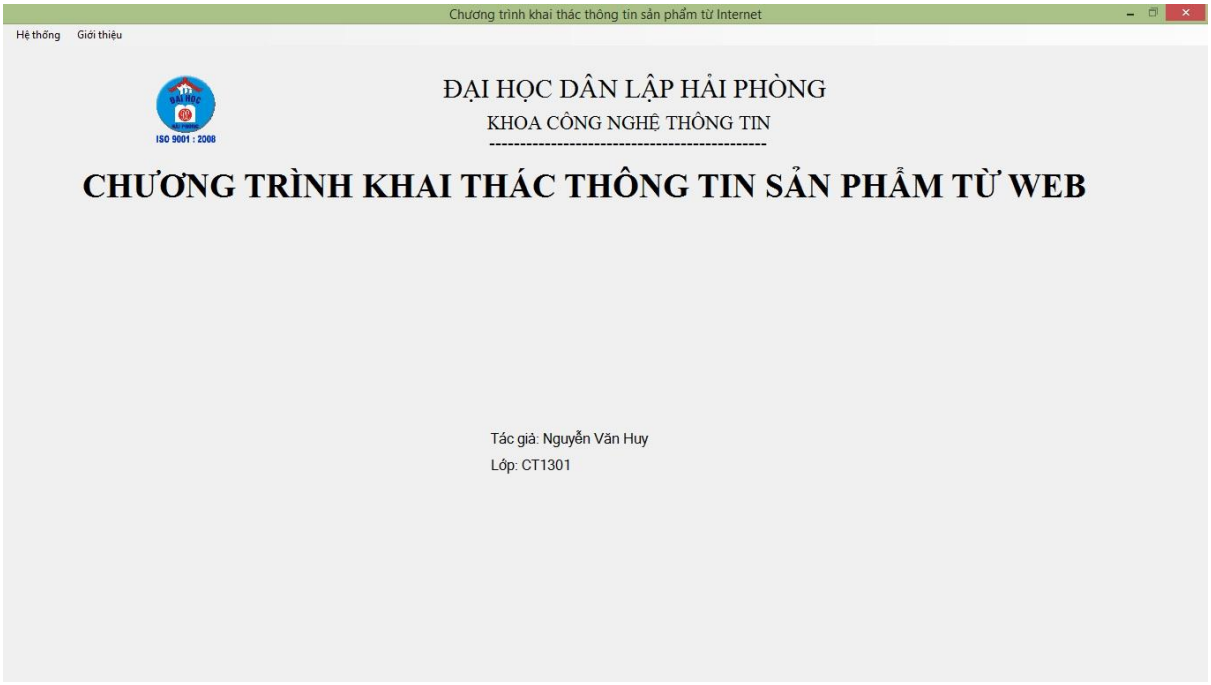
Hình 3.22: Mô hình lớp thiết kế hệ thống

3.5 Thiết kế chương trình

Chương trình viết bằng ngôn ngữ lập trình C#

Chương trình sử dụng hệ quản trị cơ sở dữ liệu SQL 2008.

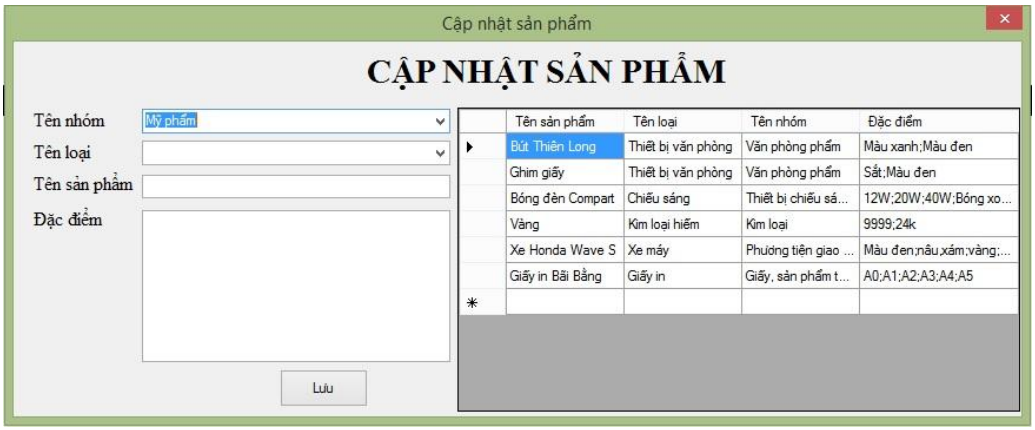
3.5.1 Giao diện chính của chương trình



Hình 3.23 Giao diện chính của chương trình

Giao diện cung cấp thông tin của chương trình và menu thực hiện công việc.

3.5.2 Giao diện cập nhật sản phẩm



Hình 3.24: Giao diện cập nhật sản phẩm

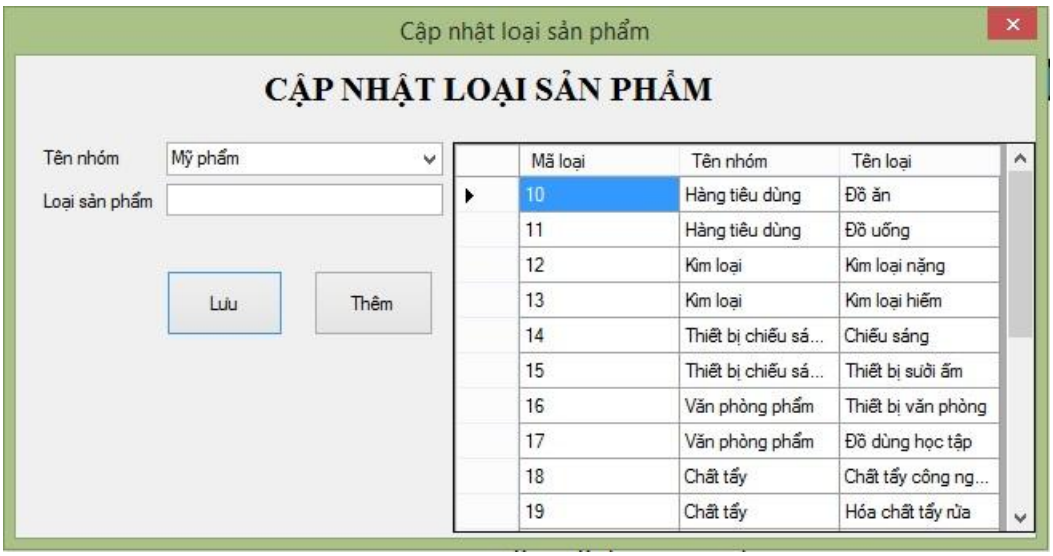
Giao diện này được mở từ menu "Sản phẩm" trên giao diện chính.

Khi cập nhật sản phẩm cần chọn tên nhóm từ ô "tên nhóm", tên loại sản phẩm từ ô "tên loại", nhập tên sản phẩm vào ô "tên sản phẩm", nhập đặc điểm sản phẩm vào ô "đặc điểm".

Nút "lưu" để lưu thông tin. Nhấn nút "thêm" để nhập sản phẩm mới.

Nhấn chuột phải vào loại sản phẩm , chọn "Xóa" nếu muốn xóa sản phẩm, chọn "Sửa" nếu muốn sửa sản phẩm

3.5.3 Giao diện cập nhật loại sản phẩm



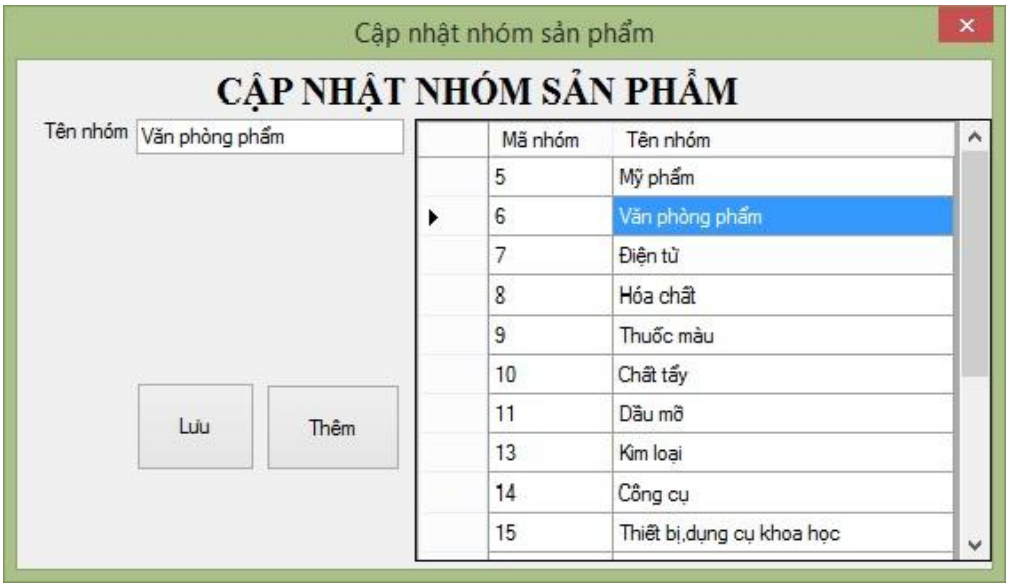
Hình 3.25: Giao diện cập nhật loại sản phẩm

Giao diện này được gọi từ menu "Loại sản phẩm " trên giao diện chính.

Khi cập nhật loại sản phẩm cần chọn tên nhóm sản phẩm từ ô "tên nhóm", nhập tên loại sản phẩm vào ô "loại sản phẩm". Nhấn nút "Lưu" để lưu thông tin. Nút thêm để thêm một loại mới.

Nhấn chuột phải vào loại sản phẩm , chọn "Xóa" nếu muốn xóa nhóm sản phẩm, chọn "Sửa" nếu muốn sửa loại sản phẩm

3.5.4 Giao diện cập nhật nhóm sản phẩm



Hình 3.26: Giao diện cập nhật nhóm sản phẩm

Giao diện này được gọi từ menu "Nhóm SP" từ giao diện chính.

Khi cập nhật nhóm sản phẩm cần nhập tên nhóm vào ô "tên nhóm". Nút lưu để lưu thông tin. Nhấn nút "Thêm" để thêm nhóm mới.

Nhấn chuột phải vào nhóm sản phẩm , chọn "Xóa" nếu muốn xóa nhóm sản phẩm, chọn "Sửa" nếu muốn sửa nhóm sản phẩm

3.5.5 Giao diện tìm kiếm thông tin sản phẩm

Hình 3.27: Giao diện tìm kiếm thông tin sản phẩm

Ô "Số KQ Google" dung để nhập số kết quả tìm kiếm mong muốn. Chọn nhóm sản phẩm từ ô "Tên nhóm", chọn tên loại sản phẩm từ ô "Tên loại", chọn sản phẩm từ ô "Tên sản phẩm", tích chọn đặc điểm trong ô đặc điểm. Nút "..." để sửa sản phẩm đã chọn. Nhấn "Tim kiếm" để bắt đâu tìm kiếm.

Chọn chuỗi tìm kiếm trong ô "Chuỗi tìm kiếm", nhấn nút "Thống kê" để xem kết quả thống kê cho chuỗi tìm kiếm.

Nút "Thoát" để thoát khỏi giao diện này trở về giao diện chính

3.5.6 Kết quả của chương trình minh họa

- Chương trình đã giải quyết được vấn đề cơ bản mà bài toán đưa ra là tìm kiếm được các đánh giá vế sản phẩm trên web.
- Chương trình đã xây dựng được các chức năng cơ bản mà bài toán đưa ra
  - Cập nhật, thêm mới, xóa lưu các thông tin về sản phẩm, chuỗi tìm kiếm, kết quả tìm kiếm...
  - Cho phép tìm kiếm các đánh giá tự động

Có chức năng hiển thị số lượng các đánh giá.

KẾT LUẬN

Đề tài tìm kiếm thông tin sản phẩm trên Web là một công việc rất thú vị. Nó thể hiện rất rõ việc áp dụng công nghệ thông tin vào các lĩnh vực khác sẽ đem lại kết quả cao. Khóa luận đã đạt một số kết quả như sau:

1. Trình bày các khái niệm cơ bản trong lĩnh vực khai phá dữ liệu từ đó giúp cho việc hiểu rõ bài toán cần được giải quyết như thế nào.

2. Khóa luận tổng hợp một số thuật toán liên quan đến lĩnh vực khai phá dữ liệu. và phân tích,thiết kế, xây dựng chương trình minh họa ứng dụng tìm kiếm thông tin sản phẩm trên Web, nhằm giúp các nhà quản lý ra quyết định tốt hơn về sản phẩm hoặc nhóm sản phẩm họ định triển khai trên thị trường.

Với cách trình bày từ khái niệm cơ bản đến cách thức xây dựng hệ thống tìm kiếm thông tin sản phẩm trên Web đã giúp cho em bổ sung nhiều kiến thức liên quan đến thực tế nghề nghiệp.

TÀI LIỆU THAM KHẢO

Tiếng Việt

[1]. Hà Quang Thụy, Phan Xuân Hiếu - Đoàn Sơn - Nguyên Trí Thành Nguyễn Thu Trang - Nguyễn Cầm Tú, *Giáo trình khai phá dữ liệu web*, NXBGDVN, 2009, 313 tr.

[2] Lê Văn Phùng, Quách Xuân Trưởng, *Khai phá dữ liệu*, NXB TT-TT, 2012, 320 tr.

Tiếng Anh

[3] W. Bruce Croft,Donald Metzler,Trevor Strohman, *Search Engines: Information Retrieval in Practice,*Addison-Wesley, 2008,524p.

[4] Alan Rea, *Data Mining – An Introduction. The Parallel Computer Centre*, Nor of The Queen's University of Belfast, 1995.

[5] Jiawei Han and Micheline Kamber, Data Mining:Concepts andTechniques, Hacours Science and Technology Company, USA, 2005.

[6] Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI Press- The MIT Press, 1996.

[7] Jiawei Han and Micheline Kamber (2001), Data Mining: Concepts and Techniques, Hacours Science and Technology Company, USA.

[8] J.Han, M.Kamber and A.K.H. Tung (2002), Spatial Clustering Methods in Data mining: A Survey, Simon Fraster University, Canada.