BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

-------o0o-------

ĐỒ ÁN TỐT NGHIỆP

NGÀNH CÔNG NGHỆ THÔNG TIN

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

−−−−−−−o0o−−−−−−−

KẾT HỢP CÁC PHƯƠNG PHÁP PHÂN CỤM TRONG KHAI PHÁ DỮ LIỆU WEB

ĐỒ ÁN TỐT NGHIỆP HỆ ĐẠI HỌC CHÍNH QUY

Ngành: Công nghệ Thông tin

HẢI PHÒNG 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

−−−−−−−○0○−−−−−−−

KẾT HỢP CÁC PHƯƠNG PHÁP PHÂN CỤM TRONG KHAI PHÁ DỮ LIỆU WEB

ĐỒ ÁN TỐT NGHIỆP HỆ ĐẠI HỌC CHÍNH QUY

Ngành: Công nghệ Thông tin

Sinh viên thực hiện: Cao Hữu Hải

Giáo viên hướng dẫn: Nguyễn Trịnh Đông

Mã sinh viên: 1212101007

HẢI PHÒNG 2016

BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC DÂN LẬP HẢI PHÒNG

CỘNG HÒA XÃ HỘI CHỦ NGHĨA VIỆT NAM
Độc lập – Tự do – Hạnh phúc
−−−−−−−o0o−−−−−−−

NHIỆM VỤ THIẾT KẾ TỐT NGHIỆP

Sinh viên: Cao Hữu Hải Mã số: 1212101007

Lớp:CT1601 Ngành: Công nghệ Thông tin

Tên đề tài: Kết hợp các phương pháp phân cụm trong khai phá dữ liệu Web

NHIỆM VỤ ĐỀ TÀI

1. Nội dung và các yêu cầu cần giải quyết trong nhiệm vụ đề tài tốt nghiệp.

a. Nội dung:

- Tìm hiểu về khai phá dữ liệu, khai phá dữ liệu Web.
- Tìm hiểu các thuật toán phân cụm phổ biến.
- Áp dụng các thuật toán phân cụm trong tìm kiếm và phân cụm tài liệu Web.
- Đề ra phương pháp xây dựng hệ thống.
- Thử nghiệm với các công cụ để giải quyết bài toán.

b. Các yêu cầu cần giải quyết.

- Nắm được lý thuyết về khai phá dữ liệu Web.
- Nắm được các thuật toán phân cụm dữ liệu.
- Nắm được quá trình phân cụm dữ liệu Web.
- Xây đựng được mô hình phân cụm dữ liệu với phần mền *Orange*.

2. Các số liệu cần thiết để thiết kế, tính toán

3. Địa điểm thực tập

CÁN BỘ HƯỚNG DẪN ĐỀ TÀI TỐT NGHIỆP

Người hướng dẫn thứ nhất:

Họ và tên: Nguyễn Trịnh Đông

Học hàm, học vị: Thạc sĩ

Cơ quan công tác: Đại học Dân lập Hải Phòng

Nội dung hướng dẫn: Tìm hiểu các phương pháp phân cụm. Tìm hiểu một số phương pháp tạo các luật cơ bản và các giải thuật liên quan. Đề ra phương pháp xây dựng hệ thống. Thử nghiệm với các công cụ để giải quyết bài toán.

Đề tài tốt nghiệp được giao ngày 03 tháng 10 năm 2016

Yêu cầu phải hoàn thành trước ngày 24 tháng 12 năm 2016

Đã nhận nhiệm vụ: Đ.T.T.N	Đã nhận nhiệm vụ: Đ.T.T.N
Sinh viên	Cán bộ hướng dẫn Đ.T.T.N

Hải Phòng, ngày .....tháng.....năm 2016

HIỆU TRƯỞNG

GS.TS.NGƯT Trần Hữu Nghị

PHẦN NHẬN XÉT TÓM TẮT CỦA CÁN BỘ HƯỚNG DẪN

1. Tinh thần thái độ của sinh viên trong quá trình làm đề tài tốt nghiệp:

……
……
……
……
……

2. Đánh giá chất lượng của đề tài tốt nghiệp (so với nội dung yêu cầu đã đề ra trong nhiệm vụ đề tài tốt nghiệp):

……
……
……
……
……
……

1. Cho điểm của cán bộ hướng dẫn:

(Điểm ghi bằng số và chữ)

……
……

Ngày.....tháng.....năm 2016
Cán bộ hướng dẫn chính
(Ký, ghi rõ họ tên)

PHẦN NHẬN XÉT ĐÁNH GIÁ CỦA CÁN BỘ CHẤM PHẢN BIỆN ĐỀ TÀI TỐT NGHIỆP

1. Đánh giá chất lượng đề tài tốt nghiệp (về các mặt như cơ sở lý luận, thuyết minh chương trình, giá trị thực tế,…):

……
……
……
……
……
……
……
……
……
……
……
……
……

2. Cho điểm của cán bộ phản biện

(Điểm ghi bằng số và chữ)

……
……

Ngày.....tháng.....năm 2016
Cán bộ chấm phản biện
(Ký, ghi rõ họ tên)

LỜI CẢM ƠN

Trong lời đầu tiên của báo cáo đồ án tốt nghiệp "Áp dụng các phương pháp phân cụm trong khai phá dữ liệu Web", em muốn gửi những lời cám ơn và biết ơn chân thành nhất của mình tới tất cả những người đã hỗ trợ, giúp đỡ em về kiến thức và tinh thần trong quá trình thực hiện đồ án.

Trước hết, em xin chân thành cám ơn thầy giáo Ths. Nguyễn Trịnh Đông, giảng viên khoa Công nghệ Thông tin, Trường Đại học Dân lập Hải Phòng, người đã trực tiếp hướng dẫn, nhận xét, giúp đỡ em trong suốt quá trình thực hiện đồ án.

Xin chân thành cảm ơn GS.TS.NGƯT Trần Hữu Nghị Hiệu trưởng trường Đại học Dân lập Hải Phòng, ban giám hiệu nhà trường, các thầy cô trong khoa Công nghệ Thông tin và các phòng ban nhà trường đã tạo điều kiện tốt nhất cho em cũng như các bạn khác trong suốt thời gian học tập và làm tốt nghiệp.

Cuối cùng em xin gửi lời cảm ơn đến gia đình, bạn bè, người thân đã giúp đỡ động viên em rất nhiều trong quá trình học tập và làm đồ án tốt nghiệp.

Mặc dù em đã hết sức cố gắng để hoàn thiện báo cáo tốt nghiệp song khả năng còn hạn chế nên bài báo cáo vẫn còn thiếu nhiều sai sót. Vì vậy em rất mong được sự đóng góp của các thầy cô và bạn bè.

Em xin chân thành cảm ơn!

Hải Phòng,ngày 24 tháng 12 năm 2016

Sinh viên

Cao Hữu Hải

MỤC LỤC

DANH SÁCH HÌNH

DANH SÁCH BẢNG

DANH MỤC TỪ VIẾT TẮT

Stt	Từ viết tắt	Từ tiếng anh	Nghĩa tiếng việt
1	KPDL		Khai phá dữ liệu
2	PCDL		Phân cụm dữ liệu
3	CSDL		Cơ sở dữ liệu
4	KDD	Knowledge Discovery in Database	Khám phá tri thức trong cơ sở dữ liệu
5	KPVB		Khai phá văn bản
6	IF	Term Frequency	Tần số xuất hiện của từ trong 1 văn bản
7	IDF	Inverse Document Frequency	Tần số nghịch của 1 từ trong tập văn bản
8	PAM	Partitioning Around Medoids	Thuật toán phân cụm dựa trên ý tưởng k-medoid
9	BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies	Thuật toán phân cụm dựa trên ý tưởng cây phân cấp
10	DBSCAN	Density Based Spatial Clustering of Applications with Noise	Thuật toán phân cụm dựa trên mật độ
11	HTML	Hypertext Markup Language	Ngôn ngữ đánh dấu siêu văn bản
12	URL	Uniform Resource Locator	Định vị tài nguyên thống nhất
13	CF	Cluster Features	Đặc điểm cụm

LỜI MỞ ĐẦU

Trong những năm ngành công nghệ thông tin đã có những bước phát triển chóng mặt. Do việc ứng dụng công nghệ thông tin vào hầu hết các lĩnh vực trong đời sống như: giáo dục, văn hóa, kinh tế, giải trí,… và sự tăng nhanh về số lượng người dùng Intenet trên toàn cầu. Đẫn đến việc bùng nổ, sự cập nhật nhanh chóng, liên tục của kh8 dữ liệu số đã đặt ra thách thức về việc khai thác,sử lý thông tin từ kho dữ liệu khổng lồ thành các tri thức có ích một cách nhanh chóng để phục vụ cho việc quản lý, hoạt động kinh doanh,… Để đáp ứng yêu cầu này người ta đã xây dựng các công cụ tìm kiếm và xử lý thông tin để giúp người dùng tìm kiếm được các thông tin cần thiết, nhưng so với sự rộng lớn về nguồn tài nguyên Web thì dẫn đến sự khó khăn với những kết quả tìm được.

Với các phương pháp khai thác cơ sở dữ liệu truyền thống chưa đáp ứng được đầu đủ các yêu cầu từ người dùng. Vì vậy một hướng đi mới đó là nghiên cứu và áp dụng kỹ thuật khai phá dữ liệu và khám phá tri thức trong môi trường Web. Do đó, việc nghiên cứu các mô hình dữ pháp khai liệu mới và áp dụng các phương phá dữ liệu trong khai phá tài nguyên Web là một xu thế tất yếu vừa có ý nghĩa khoa học vừa mang ý nghĩa thực tiễn cao.

Vì vậy, em chọn đề tài đồ án tốt nghiệp “Kết hợp các phương pháp phân cụm trong khai phá dữ liệu Web”.

Bố cục đồ án gồm 3 chương:

Chương 1: Trình bày các kiến thức cơ bản về khám phá tri thức, khai phá dữ liệu, một số vấn đề về biểu diễn và xử lý dữ liệu văn bản áp dụng trong khai phá dữ liệu.

Chương 2 : Giới thiệu một số thuật toán phân cụm dữ liệu phổ biến và thường được sử dụng trong lĩnh vực khai phá dữ liệu Web.

Chương 3: Trình bày khai phá nội dung Web và tiếp cận theo hướng sử dụng các kỹ thuật phân cụm dữ liệu để giải quyết bài toán khai phá dữ liệu Web. Trong phần này cũng trình bày một mô hình áp dụng kỹ thuật phân cụm dữ liệu trong tìm kiếm và phân cụm tài liệu Web.

CHƯƠNG 1: GIỚI THIỆU VỀ KHAI PHÁ DỮ LIỆU WEB

1.1 Khai phá dữ liệu và khai phá tri thức

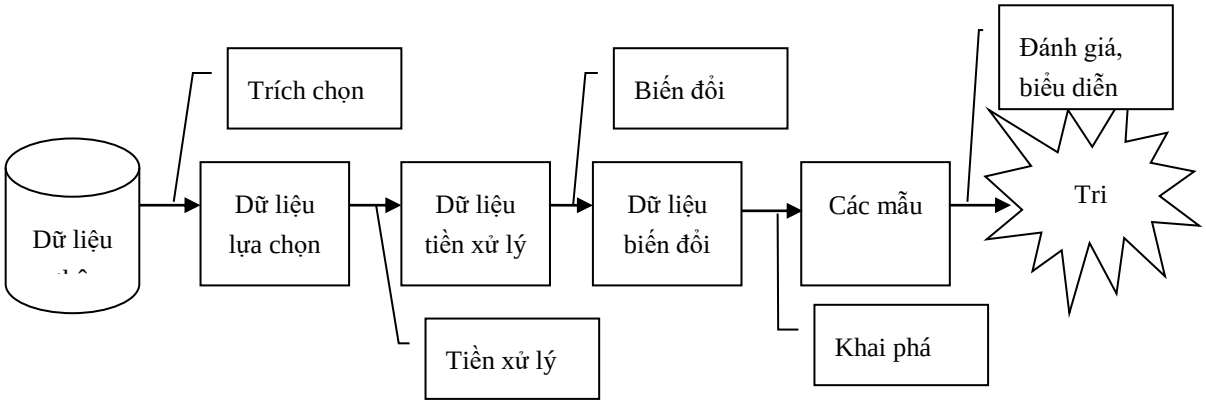
1.1.1 Khai phá dữ liệu

Khai phá dữ liệu là một lĩnh vực mới được nghiên cứu, nhằm tự động khai thác thông tin, tri thức mới hữu ích, tiềm ẩn từ những CSDL lớn cho các đơn vị, tổ chức, doanh nghiệp,…. từ đó thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh cho các đơn vị, tổ chức này. Các kết quả nghiên cứu khoa học cùng những ứng dụng thành công trong KDD cho thấy KPDL là một lĩnh vực phát triển bền vững, mang lại nhiều lợi ích và có nhiều triển vọng, đồng thời có ưu thế hơn hẳn so với các công cụ tìm kiếm phân tích dữ liệu truyền thống. Hiện nay, KPDL đã ứng dụng ngày càng rộng rãi trong các lĩnh vực như thương mại, tài chính, y học, viễn thông, tin – sinh…

Như vậy, Khai phá dữ liệu là quá trình khai phá, trích xuất, khai thác và sử dụng những dữ liệu có giá trị tiềm ẩn từ bên trong lượng lớn dữ liệu được lưu trữ trong các cơ sở dữ liệu (CSDL), kho dữ liệu, trung tâm dữ liệu…

1.1.2 Quá trình khám phá tri thức

Quá trình khá phá tri thức có thể chia thành 5 bước như sau [1]:



Hình 1-1: Quy trình khai phá tri thức

Quá trình KPDL có thể phân thành các giai đoạn sau:

Trích chọn dữ liệu: Đây là bước trích chọn những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn ban đầu theo một số tiêu chí nhất định.

Tiền xử lý dữ liệu: Đây là bước làm sạch dữ liệu (loại bỏ dữ liệu không đúng,xử lý dữ liệu thiếu sót,...), rút gọn dữ liệu (sử dụng hàm nhóm và tính tổng, các phương pháp nén dữ liệu, sử dụng histograms, lấy mẫu,...), rời rạc hóa dữ liệu (rời rạc hóa dựa

vào histograms, entropy,...). Sau bước này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và được rời rạc hóa.

Biến đổi dữ liệu: Đây là bước chuẩn hóa và làm mịn dữ liệu để đưa dữ liệu về cùng một kiểu, dạng thuận lợi nhất nhằm phục vụ quá trình xử lý ở bước sau.

Khai phá dữ liệu: Đây là bước áp dụng những kỹ thuật phân tích (như các kỹ thuật của học máy) nhằm để khai thác dữ liệu, trích chọn được những mẫu dữ liệu, những mối liên hệ đặc biệt trong dữ liệu. Đây được xem là bước quan trọng và tốn nhiều thời gian nhất của toàn quá trình KDD.

Đánh giá và biểu diễn tri thức: Những mẫu thông tin và mối liên hệ trong dữ liệu đã được khám phá ở bước trên được biến đổi và biểu diễn ở một dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật,... Đồng thời bước này cũng đánh giá những tri thức khám phá được theo những tiêu chí nhất định.

1.1.3 Khai phá dữ liệu và các lĩnh vực liên quan

KPDL là một lĩnh vực liên quan tới thống kê, học máy, CSDL, thuật toán, tính toán song song, thu nhận tri thức từ hệ chuyên gia và dữ liệu trừu tượng. Đặc trưng của hệ thống khám phá tri thức là nhờ vào các phương pháp, thuật toán và kỹ thuật từ những lĩnh vực khác nhau để KPDL. Với lĩnh vực học máy và nhận dạng mẫu thì KDD nghiên cứu các lý thuyết và thuật toán của hệ thống để trích ra các mẫu và mô hình từ dữ liệu lớn. KDD tập trung vào việc mở rộng các lý thuyết và thuật toán cho các vấn đề tìm ra các mẫu đặc biệt (hữu ích hoặc có thể rút ra tri thức quan trọng) trong CSDL lớn. Với lĩnh vực thống kê, hệ thống KDD thường gắn những thủ tục thống kê cho mô hình dữ liệu, đặc biệt là trong lĩnh vực thăm dò (Exploratory Data Analysis - EDA).

1.1.4 Các kỹ thuật áp dụng trong khai phá dữ liệu

Căn cứ vào các bài toán cần giải quyết thì KPDL gồm các kỹ thuật sau [5]:

Phân lớp và dự báo: Xếp một đối tượng vào một trong những lớp đã biết trước. Ví dụ như phân lớp các dữ liệu bệnh nhân trong hồ sơ bệnh án. Hướng tiếp cận này thường sử dụng một số kỹ thuật của học máy như cây quyết định, mạng nơron nhân tạo,... Phân lớp và dự báo còn được gọi là học có giám sát.

Luật kết hợp: Là dạng luật biểu diễn tri thức ở dạng khá đơn giản. Ví dụ: "60 % nữ giới vào siêu thị nếu mua 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,...

Phân tích chuỗi theo thời gian: 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: Xếp các đối tượng theo từng cụm dữ liệu tự nhiên. Phân cụm còn được gọi là học không có giám sát.

Mô tả và tóm tắt khái niệm: Thiên về mô tả, tổng hợp và tóm tắt khái niệm, ví dụ như tóm tắt văn bản.

1.1.5 Những chức năng chính của khai phá dữ liệu

KPDL có hai mục tiêu chính là: mô tả và dự báo . Dự báo là dùng một số biến hoặc trường trong CSDL để dự đoán ra các giá trị chưa biết hoặc sẽ có của các biến quan trọng khác. Việc mô tả tập trung vào tìm kiếm các mẫu mà con người có thể hiểu được để mô tả dữ liệu. Trong lĩnh vực KDD, mô tả được quan tâm nhiều hơn dự báo, nó ngược với các ứng dụng học máy và nhận dạng mẫu mà trong đó việc dự báo thường là mục tiêu chính. Trên cơ sở mục tiêu chính của KPDL, các chức năng chính của KDD gồm [1]:

Mô tả lớp và khái niệm: Dữ liệu có thể được kết hợp trong lớp và khái niệm. Ví dụ: trong kho dữ liệu bán hàng thiết bị tin học, các lớp mặt hàng bao gồm máy tính, máy in,…và khái niệm khách hàng bao gồm khách hàng mua sỉ và khách mua lẻ. Việc mô tả lớp và khái niệm là rất hữu ích cho giai đoạn tổng hợp, tóm lược và chính xác hoá. Mô tả lớp và khái niệm được bắt nguồn từ đặc trưng hoá dữ liệu và phân biệt dữ liệu. Đặc trưng hoá dữ liệu là quá trình tổng hợp những đặc tính hoặc các thành phần chung của một lớp dữ liệu mục tiêu. Phân biệt dữ liệu là so sánh lớp dữ liệu mục tiêu với những lớp dữ liệu đối chiếu khác. Lớp dữ liệu mục tiêu và các lớp đối chiếu là do người dùng chỉ ra và tương ứng với các đối tượng dữ liệu nhận được nhờ truy vấn.

Phân tích sự kết hợp: Phân tích sự kết hợp là khám phá luật kết hợp thể hiện mối quan hệ giữa các thuộc tính giá trị mà ta nhận biết được nhờ tần suất xuất hiện cùng nhau của chúng.

Phân lớp và dự báo: Phân lớp là quá trình tìm kiếm một tập các mô hình hoặc chức năng mà nó mô tả và phân biệt nó với các lớp hoặc khái niệm khác. Các mô hình này nhằm mục đích dự báo về lớp của một số đối tượng. Việc xây dựng mô hình dựa trên sự phân tích một tập các dữ liệu được huấn luyện có nhiều dạng thể hiện mô hình như luật phân lớp (IF-THEN), cây quyết định, công thức toán học hay mạng nơron,... Sự phân lớp được sử dụng để dự đoán nhãn lớp của các đối tượng trong dữ liệu. Tuy nhiên trong nhiều ứng dụng, người ta mong muốn dự đoán những giá trị khuyết thiếu

nào đó. Thông thường đó là trường hợp dự đoán các giá trị của dữ liệu kiểu số. Trước khi phân lớp và dự báo, có thể cần thực hiện phân tích thích hợp để xác định và loại bỏ các thuộc tính không tham gia vào quá trình phân lớp và dự báo.

Phân cụm: Không giống như phân lớp và dự báo, phân cụm phân tích các đối tượng dữ liệu khi chưa biết nhãn của lớp. Nhìn chung, nhãn lớp không tồn tại trong suốt quá trình huấn luyện dữ liệu, nó phân cụm có thể được sử dụng để đưa ra nhãn của lớp. Sự phân cụm thực hiện nhóm các đối tượng dữ liệu theo nguyên tắc: Các đối tượng trong cùng một nhóm thì giống nhau hơn các đối tượng khác nhóm. Mỗi cụm được tạo thành có thể được xem như một lớp các đối tượng mà các luật được lấy ra từ đó. Dạng của cụm được hình thành theo một cấu trúc phân cấp của các lớp mà mỗi lớp là một nhóm các sự kiện tương tự nhau.

Phân tích các đối tượng ngoài cuộc: Là các đối tượng không tuân theo mô hình dữ liệu trong CSDL. Hầu hết các phương pháp KPDL đều coi các đối tượng ngoài cuộc là nhiễu và loại bỏ chúng. Tuy nhiên trong một số ứng dụng, chẳng hạn như phát hiện nhiễu, thì sự kiện hiếm khi xảy ra lại được chú ý hơn những gì thường xuyên gặp phải. Sự phân tích dữ liệu ngoài cuộc được coi như là sự khai phá các đối tượng ngoài cuộc. Một số phương pháp được sử dụng để phát hiện đối tượng ngoài cuộc: sử dụng các test mang tính thống kê trên cơ sở một phân phối dữ liệu hay một mô hình xác suất cho dữ liệu, dùng các độ đo khoảng cách mà theo đó các đối tượng có một khoảng cách đáng kể đến cụm bất kì khác được coi là đối tượng ngoài cuộc, dùng các phương pháp dựa trên độ lệch để kiểm tra sự khác nhau trong những đặc trưng chính của các nhóm đối tượng.

Phân tích sự tiến hoá: Phân tích sự tiến hoá thực hiện việc mô tả và mô hình hoá các quy luật hay khuynh hướng của những đối tượng mà hành vi của chúng thay đổi theo thời gian. Phân tích sự tiến hoá có thể bao gồm cả đặc trưng hoá, phân biệt, tìm luật kết hợp, phân lớp hay PCDL liên quan đến thời gian, phân tích dữ liệu theo chuỗi thời gian, so sánh mẫu theo chu kỳ và phân tích dữ liệu dựa trên độ tương tự.

1.1.6 Ứng dụng của khai phá dữ liệu

KPDL là lĩnh vực được ứng dụng vào nhiều lĩnh vực hiện nay như:

Kinh doanh - thương mại:

- Xác định thói quen mua hàng của khách hàng
- Dự doán chu kỳ kinh doanh sản phẩm
- Liên hệ giữa khách hàng và yếu tố khác

- Xác định khách hàng tiềm năng, đối tượng có khả năng trở thành khách hàng
- Dự đoán hiệu quả của một đợt quảng cáo, tiếp thị

Thương mại điện tử:

- Phân tích hoạt động duyệt Web để phân tích sở thích của khách hàng

Ngân hàng:

- Dự đoán các dấu hiệu của một cuộc giao dịnh trái luật
- Xác dịnh khách hàng sẽ cộng tác lâu dài
- Dự đoán rủi do của các khoản cho vay
- Xác định nhân tố đẫn đến vỡ nợ vay
- Liên hệ các chỉ số tài chính đến hoạt động ngân hàng

Bảo hiểm:

- Loại khách hàng có rủi do cao, gian lận
- Xác định khách hàng tiềm năng
- Xác định các đối tượng sẽ trở thành khác hàng

Viễn thông:

- Nhận biết các dấu hiệu của cuộc gian lận dịch vụ
- Xu thế phát triển khách hàng, đối tượng, khu vực cần phát triển

Y tế:

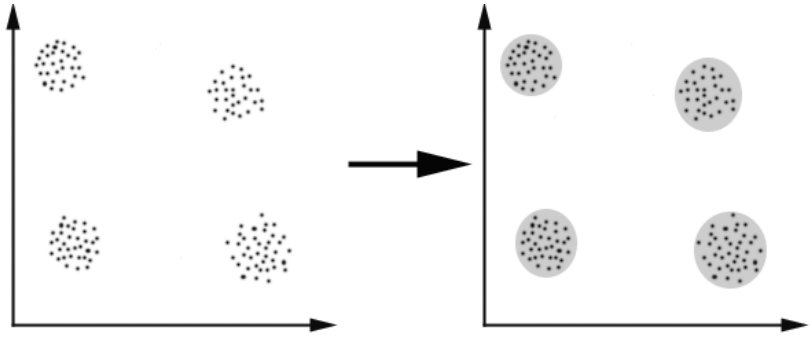
- Chuẩn đoán bệnh qua các triệu chứng
- Liên hệ giữa các loại bệnh
- Dự doán hiệu quả của một cuộc phẫu thuật, điều trị

1.2 Phương pháp phân cụm dữ liệu

1.2.1 Giới thiệu về kỹ thuật phân cụm

PCDL là quá trình phân chia một tập dữ liệu ban đầu thành các cụm dữ liệu sao cho các đối tượng trong một cụm đó "tương tự" với nhau. PCDL là một kỹ thuật trong KPDL, nhằm tìm kiếm, phát hiện các cụm, các mẫu dữ liệu tự nhiên, tiềm ẩn, quan trọng trong tập dữ liệu lớn từ đó cung cấp thông tin, tri thức hữu ích cho việc ra quyết định. Mục đích chính của PCDL nhằm khám phá cấu trúc của mẫu dữ liệu để thành lập các nhóm dữ liệu từ tập dữ liệu lớn, theo đó nó cho phép người ta đi sâu vào phân tích

và nghiên cứu cho từng cụm dữ liệu này nhằm khám phá và tìm kiếm các thông tin tiềm ẩn, hữu ích phục vụ cho việc ra quyết định. Ví dụ: "Nhóm khách hàng có khả năng trả nợ cao"… Như vậy, PCDL xử là một phương pháp lý thông tin quan trọng và phổ biến, nó nhằm khám phá mối liên hệ giữa các mẫu dữ liệu bằng cách tổ chức chúng thành các cụm [1].



Hình 1-2: Mô phỏng sự phân cụm

Trong hình trên sau khi phân cụm thì những phần tử tương tự nhau thì được sắp xếp vào một cụm và ngược lại, hay là những phần tử có chung một định nghĩa hoặc xấp xỉ về khái niệm cho trước cũng được xếp vào một cụm. Một số vấn đề thường gặp trong PCDL là dữ liệu "nhiễu" và "phần tử ngoại lai". "Nhiễu" có thể là các đối tượng dữ liệu không chính xác hoặc các đối tượng dữ liệu 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. "Phần tử ngoại lai" là những phần tử có sự khác biệt đáng kể đối với những phần tử còn lại. Có nhiều cách xác định phần tử ngoại lai, như xác định theo khoảng cách: Sử dụng hàm đo khoảng cách giữa các phần tử trong tập dữ liệu, các phần tử ngoại lai là các phần tử cách khá xa so với các phần tử còn lại. Xác định theo thống kê: Xác định các mô hình phân phối thống kê mà các phần tử phải tuân theo, "phần tử ngoại lai" là những phần tử không tuân theo các quy luật này. Xác định theo độ khác biệt: Xác định những đặc trưng cơ bản của các cụm, "phần tử ngoại lai" sẽ có đặc trưng khác biệt lớn với những phần tử còn lại.

Những vấn đề cần giải quyết khi PCDL:

- Biểu diễn dữ liệu
- Xây dựng hàm tính độ tương tự
- Xây dựng các tiêu chuẩn phân cụm
- Xây dựng mô hình cho cấu trúc cụm dữ liệu
- Xây dựng thuật toán phân cụm và xác lập các điều kiện khởi tạo
- Xây dựng các thủ tục biểu diễn và đánh giá kết quả phân cụm

1.2.2 Ứng dụng của phân cụm dữ liệu

PCDL là một trong những công cụ chính của KPDL được ứng dụng trong nhiều lĩnh vực như thương mại và khoa học. Các kỹ thuật PCDL đã được áp dụng cho một số ứng dụng điển hình trong các lĩnh vực sau [5]:

Thương mại: PCDL có thể giúp các thương nhân khám phá ra các nhóm khách hàng quan trọng có các đặc trưng tương đồng nhau và đặc tả họ từ các mẫu mua bán trong CSDL khách hàng, phát hiện và dự đoán các giao dịch gian lận.

Sinh học: PCDL được sử dụng để phân cụm các loại sinh vật, phân loại các Gen với chức năng tương đồng và thu được các cấu trúc trong các mẫu, phát hiện và dự đoán các biến dị.

Lập quy hoạch đô thị: Nhận dạng các nhóm nhà theo kiểu và vị trí địa lý,… nhằm cung cấp thông tin cho quy hoạch đô thị.

Địa lý: Phân lớp các động vật, thực vật và đưa ra đặc trưng của chúng theo vị trí địa lý.

Khai phá Web: PCDL có thể khám phá các nhóm tài liệu quan trọng, có nhiều ý nghĩa trong môi trường Web. Các lớp tài liệu này trợ giúp cho việc khám phá tri thức từ dữ liệu Web, khám phá ra các mẫu truy cập của khách hàng đặc biệt hay khám phá ra cộng đồng Web,…

1.2.3 Các yêu cầu đối với kỹ thuật phân cụm dữ liệu

Việc xây dựng, lựa chọn một thuật toán phân cụm là bước then chốt cho việc giải quyết vấn đề phân cụm, sự lựa chọn này phụ thuộc vào đặc tính dữ liệu cần phân cụm, mục đích của ứng dụng thực tế hoặc xác định độ ưu tiên giữa chất lượng của các cụm hay tốc độ thực hiện thuật toán,…

Những yêu cầu để phát triển thuật toán PCDL [5]:

Có khả năng mở rộng: 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 định danh, hạng mục,... và thích nghi với kiểu dữ liệu hỗn hợp.

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 dữ liệu 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 ảnh hưởng rất lớn 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 KPDL 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, dễ cài đặt và khả thi.

1.2.4 Các kiểu dữ liệu và độ đo tương tự

Trong PCDL, các đối tượng dữ liệu cần phân tích có thể là con người, nhà cửa, tiền lương, các thực thể phần mềm,… Các đối tượng này thường được diễn tả dưới dạng các thuộc tính của nó. Việc phân loại các kiểu thuộc tính khác nhau là một vấn đề cần giải quyết đối với hầu hết các tập dữ liệu nhằm cung cấp các phương tiện thuận lợi để nhận dạng sự khác nhau của các phần tử dữ liệu. Có hai cách phân lớp dựa trên hai đặc trưng của dữ liệu là: kích thước miền và hệ đo [2].

1.2.4.1 Phân loại kiểu dữ liệu dựa trên kích thước miền

Thuộc tính liên tục: Nếu miền giá trị của nó là vô hạn không đếm được, nghĩa là giữa hai giá trị tồn tại vô số giá trị khác. Thí dụ như các thuộc tính về màu, nhiệt độ hoặc cường độ âm thanh.

- *Thuộc tính rời rạc:* Nếu miền giá trị của nó là tập hữu hạn hoặc đếm được. Thí dụ như các thuộc tính về số serial của một cuốn sách, số thành viên trong một gia đình,…

1.2.4.2 Phân loại 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 hóa 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 # y hoặc x = y.

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 # y hoặc x = y hoặc x>y hoặc x<y.

Thuộc tính khoảng (Interval Scale): 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.

Thuộc tính tỉ lệ (Ratio Scale): là thuộc tính khoảng nhưng được xác định một cách tương đối so với điểm mốc, 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),* thuộc tính khoảng và thuộc tính tỉ lệ được gọi là *thuộc tính số (Numeric).*

1.2.4.3 Khái niệm và phép đo độ tương tự, phi tương tự

Để phân cụm, người ta phải đi tìm cách thích hợp để xác định "khoảng cách" giữa các đối tượng, hay là phép đo tương tự dữ liệu. Đây là các hàm để đo sự giống nhau giữa các cặp đối tượng dữ liệu, thông thường các hàm này hoặc là để tính độ tương tự (Similar) hoặc là tính độ phi tương tự (Dissimilar) giữa các đối tượng dữ liệu.

Tất cả các độ đo dưới đây được xác định trong không gian metric. Một không gian metric là một tập trong đó có xác định các "khoảng cách" giữa từng cặp phần tử, với những tính chất thông thường của khoảng cách hình học. Nghĩa là, một tập X (các

phần tử của nó có thể là những đối tượng bất kỳ) các đối tượng dữ liệu trong cơ sở dữ liệu D như đã đề cập ở trên được gọi là một không gian metric nếu:

Với mỗi cặp phần tử x, y thuộc X đều có xác định, theo một quy tắc nào đó, một số thực δ(x, y), được gọi là khoảng cách giữa x và y.

Quy tắc trên thoả mãn hệ tính chất sau:

- δ(x, y) > 0 nếu x ≠ y ;
- δ(x, y)=0 nếu x =y;
- δ(x, y) = δ(y, x) với mọi x, y; (iv) δ(x, y) ≤ δ(x, z)+δ(z, y).

Hàm δ(x, y) được gọi là một metric của không gian. Các phần tử của X được gọi là các điểm của không gian này.

Mỗi phép đo độ tương tự sẽ phù hợp với mỗi kiểu dữ liệu khác nhau[5].

Thuộc tính khoảng:

Sau khi chuẩn hoá, độ đo phi tương tự của hai đối tượng dữ liệu *x*, *y* được xác định bằng các metric như sau:

Khoảng cách Minskowski: $d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|^q)^{\frac{1}{q}}$, với q là số nguyên dương.

Khoảng cách Euclidean: $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$, (trường hợp đặc biệt của khoảng cách Minskowski trong trường hợp q =2).

Khoảng cách Manhattan: $d(x,y) = \sum_{i=1}^{n}|x_i - y_i|$, (trường hợp đặc biệt của khoảng cách *Minskowski* trong trường hợp q=1).

Khoảng cách cực đại: $d = (x,y) = Max_{i=1}^{n}|x_i - y_i|$, đây là trường hợp của khoảng cách *Minskowski* trong trường hợp $q \to \infty$.

Thuộc tính nhị phân:

Trước hết ta có xây dựng bảng tham số sau:

	y:1	y:0	
x:1	α	β	$\alpha + \beta$
y:1	γ	δ	$\gamma + \delta$
	$\alpha + \gamma$	$\beta + \delta$	τ

Bảng 1-1: Bảng tham số thuộc tính nhị phân

Trong đó: $\tau = \alpha + \beta + \gamma + \delta$. x,y là các đối tượng có thuộc tính đều là nhị phân.

α là tổng số các thuộc tính có giá trị là 1 trong cả hai đối tượng x, y.

β là tổng số các giá trị thuộc tính có giá trị là 1 trong *x* và 0 trong y.

γ là tổng số các giá trị thuộc tính có giá trị là 0 trong *x* và 1 trong y.

δ là tổng số các giá trị thuộc tính có giá trị là 0 trong *x* và y.

Các phép đo độ tương tự đối với dữ liệu thuộc tính nhị phân được định nghĩa như sau:

Hệ số đối sánh đơn giản: $d(x,y) = \frac{\alpha+\delta}{\delta}$, ở đây cả hai đối tượng *x* và *y* có vai trò như nhau, nghĩa là chúng đối xứng và có cùng trọng số.

Hệ số Jacard:$d(x,y) = \frac{\alpha}{\alpha+\beta+\gamma}$, tham số này bỏ qua số các đối sánh giữa 0-0. Công thức tính này được sử dụng trong trường hợp mà trọng số của các thuộc tính có giá trị 1 của đối tượng dữ liệu có giá trị cao hơn nhiều so với các thuộc tính có giá trị 0, như vậy các thuộc tính nhị phân ở đây là không đối xứng.

Thuộc tính định danh:

Độ đo phi tương tự giữa hai đối tượng *x* và *y* được định nghĩa như sau:

$d(x,y) = \frac{p-m}{p}$, trong đó *m* là số thuộc tính đối sánh tương ứng trùng nhau và *p* là tổng số các thuộc tính.

Thuộc tính có thứ tự:

Phép đo độ phi tương tự giữa các đối tượng dữ liệu với thuộc tính thứ tự được thực hiện như sau, ở đây ta giả sử i là thuộc tính thứ tự có M_i giá trị (M_i kích thước miền giá trị):

Các trạng thái M_i được sắp thứ tự như sau: [1…M_i], ta có thể thay thế mỗi giá trị của thuộc tính bằng giá trị cùng loại r_i, với r_i {1,…,M_i}. Mỗi một thuộc tính thứ tự có các miền giá trị khác nhau, vì vậy ta chuyển đổi chúng về cùng miền giá trị [0,1] bằng cách thực hiện phép biến đổi sau cho mỗi thuộc tính:

$z_i^{(j)} = \frac{r_i^{(j)}-1}{M_i-1}$, với i=1,..,M_i.

Sử dụng công thức tính độ phi tương tự của thuộc tính khoảng đối với các giá trị $z_i^{(j)}$, đây cũng chính là độ phi tương tự của thuộc tính có thứ tự.

Thuộc tính tỉ lệ:

Có nhiều cách khác nhau để tính độ tương tự giữa các thuộc tính tỉ lệ. Một trong những số đó là sử dụng công thức tính logarit cho mỗi thuộc tính x_i, thí dụ $q_i = \log(x_i)$, lúc này q_i đóng vai trò như thuộc tính khoảng. Phép biến đổi logarit này thích hợp trong trường hợp các giá trị của thuộc tính là số mũ.

Trong thực tế, khi tính độ đo tương tự dữ liệu, người ta chỉ xem xét một phần các thuộc tính đặc trưng đối với các kiểu dữ liệu hoặc đánh trọng số cho cho tất cả các thuộc tính dữ liệu. Trong một số trường hợp, người ta loại bỏ đơn vị đo của các thuộc tính dữ liệu bằng cách chuẩn hoá chúng hoặc gán trọng số cho mỗi thuộc tính giá trị trung bình, độ lệch chuẩn. Các trọng số này có thể sử dụng trong các độ đo khoảng cách trên, thí dụ với mỗi thuộc tính dữ liệu đã được gán trọng số tương ứng wi ($1 \leq i \leq k$), độ tương tự dữ liệu được xác định như sau:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} w_i (x_i - y_i)^2}.$$

Tóm lại, tuỳ từng trường hợp dữ liệu cụ thể mà người ta sử dụng các mô hình tính độ tương tự khác nhau. Việc xác định độ tương tự dữ liệu thích hợp, chính xác, đảm bảo khách quan là rất quan trọng và giúp xây dựng thuật toán PCDL có hiệu quả cao trong việc đảm bảo chất lượng cũng như chi phí tính toán của thuật toán.

1.3 Khai phá dữ liệu Web

Khai phá dữ liệu Web là việc sử dụng các kỹ thuật KPDL để tự động hóa quá trình phát hiện và trích chọn những thông tin hữu ích từ các tài liệu, các thông tin dịch vụ, hồ sơ sử dụng và cấu trúc Website. Hay nói cách khác khai phá Web là việc thăm dò những thông tin quan trọng và những mẫu dữ liệu tiềm năng từ nội dung Web, từ thông tin truy cập Web, từ liên kết trang và từ nguồn tài nguyên thương mại điện tử bằng việc sử dụng các kỹ thuật KPDL, nó có thể giúp con người rút ra tri thức, cải tiến việc thiết kế các Website và phát triển thương mại điện tử tốt hơn [1].

Quá trình khai phá Web:

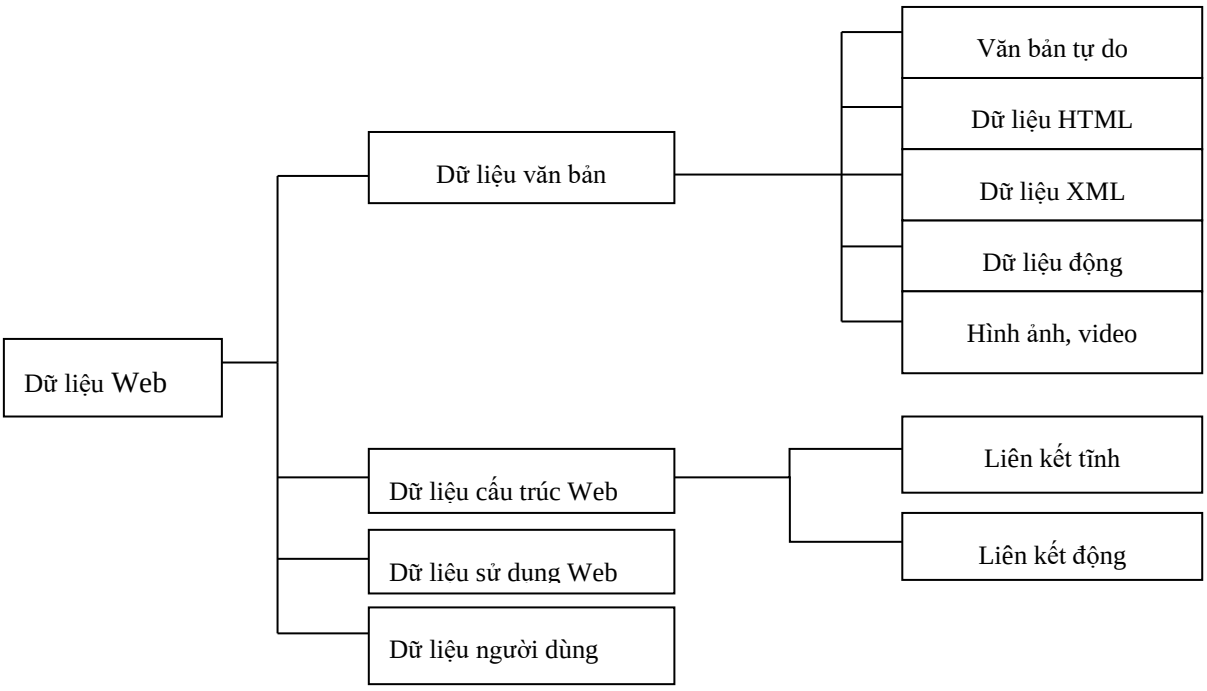
Tìm kiếm nguồn tài nguyên: Thực hiện tìm kiếm và lấy các tài liệu Web phục vụ cho việc khai phá.

Lựa chọn và tiền xử lý dữ liệu: Lựa chọn và tiền xử lý tự động các loại thông tin từ nguồn tài nguyên Web đã lấy về.

Tổng hợp: Tự động khám phá các mẫu chung tại các Website riêng lẽ cũng như nhiều Website với nhau.

1.3.1 Các kiểu dữ liệu Web

Sơ đồ phân loại dữ liệu Web :



Hình 1-3: Phân loại dữ liệu Web

Các đối tượng của khai phá Web bao gồm[4] *:* Server logs, Web pages, Web hyperlink structures, dữ liệu thị trường trực tuyến và các thông tin khác.

Web logs(dữ liệu đăng nhập Web): Khi người dùng duyệt Web, dịch vụ sẽ phân ra 3 loại dữ liệu đăng nhập: sever logs(dữ liệu đăng nhập trên server), error logs(dữ liệu đăng nhập lỗi), và cookie logs(thông số của từng người dùng truy cập Wepsite). Thông qua việc phân tích các tài liệu đăng nhập này ta có thể khám phá ra những thông tin truy cập.

Web pages: Hầu hết các phương pháp KPDL Web được sử dụng trong Web pages là theo chuẩn HTML.

Web hyperlink structure: Các trang Web được liên kết với nhau bằng các siêu liên kết, điều này rất quan trọng để khai phá thông tin. Do các siêu liên kết Web là nguồn tài nguyên rất xác thực.

Dữ liệu thị trường trực tuyến: Như lưu trữ thông tin thương mại điện tử trong các site thương mại điện tử.

Các thông tin khác: Chủ yếu bao gồm các đăng ký người dùng, nó có thể giúp cho việc khai phá tốt hơn.

1.3.2 Xử lý dữ liệu văn bản ứng dụng trong khai phá dữ liệu Web

1.3.2.1 Dữ liệu văn bản

Văn bản là loại tài liệu phổ biến, được sử dụng trong mọi hoạt động của con người, đặc biệt trong môi trường truyền truyền thông số và trên Internet. Do vậy, các bài toán xử lý loại dữ liệu này đã được đặt ra từ rất sớm và hiện nay nó vẫn là vấn đề rất được nhiều nhà nghiên cứu quan tâm, một trong những bài toán đó là tìm kiếm và trích dẫn văn bản, biểu diễn và phân loại văn bản,….

CSDL văn bản có thể chia làm 2 loại chính [6]***:***

Dạng không có cấu trúc: Đây là những tài liệu văn bản thông thường mà ta dùng hằng ngày, thường xuất hiện trên các sách, báo, internet,… đây là dạng dữ liệu của ngôn ngữ tự nhiên của con người và nó không theo một khuôn mẫu định sẵn nào cả.

Dạng nửa cấu trúc: Đây là những văn bản được tổ chức dưới dạng cấu trúc lỏng, nhưng vẫn thể hiện nội dung chính của văn bản, như văn bản HTML, Email,..

1.3.3 Một số vấn đề trong xử lý dữ liệu văn bản

Trong việc sử lý các dữ liệu văn bản thì mỗi văn bản được biểu diễn bằng một vector Boolean hoặc vector số. Những vector này được xét trong một không gian đa chiều, trong đó mỗi chiều tương ứng với một từ mục riêng biệt trong tập văn bản.

- ***Một số lưu ý khi biểu điễn văn bản bằng không gian vector:***
- *Không gian vector:* là một tập hợp bao gồm các từ.
- *Từ:* là một chuỗi các ký tự (chữ cái và chữ số). Ngoại trừ các khoảng trống (space, tab), ký tự xuống dòng, dấu câu (như dấu chấm, phẩy, chấm phẩy, dấu cảm,...). Mặt khác, để đơn giản trong quá trình xử lý, người ta không phân biệt chữ hoa và chữ thường (nếu chữ hoa thì chuyển về chữ thường).
- *Gộp từ đồng nghĩa:* Trong nhiều ngôn ngữ, nhiều từ có cùng từ gốc hoặc là biến thể của từ gốc sang một từ khác. Việc sử dụng từ gốc làm giảm đáng kể số lượng các từ trong văn bản (giảm số chiều của không gian), nhưng việc cắt bỏ các từ lại rất khó trong việc hiểu văn bản.
- *Loại bỏ từ:* Trong phương pháp biểu diễn dữ liệu văn bản bằng không gian vector, thì chiều của một vector sẽ rất lớn bởi số chiều của nó được xác định bằng số lượng các từ khác nhau trong tập hợp từ. Vì vậy, vấn đề dặt ra là làm sao để giảm số chiều của vector mà vẫn đảm bảo việc xử lý

văn bản đúng và chính xác. Để giải quyết vấn đề này người ta đưa ra một số phương pháp là: loại bỏ từ dừng và áp dụng định luật Zipf

1.3.3.1 Loại bỏ từ dừng

Trong ngôn ngữ văn bản hằng ngày có nhiều từ chỉ dùng để biểu diễn cấu trúc câu chứ không biểu đạt nội dung của nó. Như các giới từ, từ nối,... những từ như vậy xuất hiện nhiều trong các văn bản mà không liên quan gì tới chủ đề hoặc nội dung của văn bản, những từ như vậy được gọi là những từ dừng. vậy nên, ta có thể loại bỏ từ dừng để giảm số chiều của vector trong biểu diễn văn bản.

Sau đây là ví dụ về tần số xuất hiện cao của một số từ (tiếng Anh) trong 336,310 tài liệu gồm tổng cộng 125.720.891 từ, 508.209 từ riêng biệt.

Frequent Word	Number of Occurrences	Percentage of Total
The	7,398,934	5.9
of	3,893,790	3.1
to	3,364,653	2.7
and	3,320,687	2.6
in	2,311,785	1.8
is	1,559,147	1.2
for	1,313,561	1.0
The	1,144,860	0.9
that	1,066,503	0.8
said	1,027,713	0.8

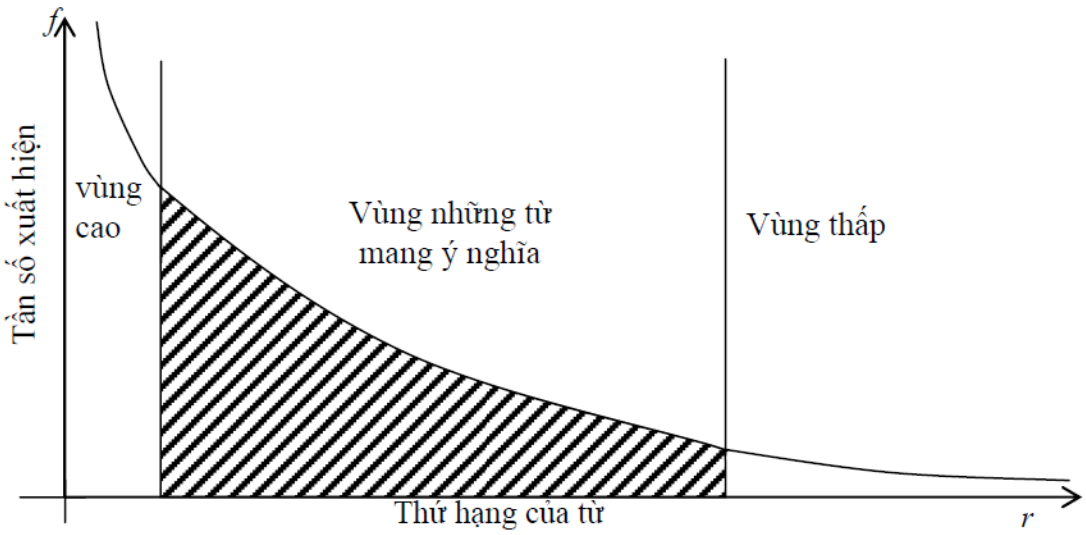
Bảng 1-2: Thống kê các tần số xuất hiện cao

(Thống kê của B. Croft, UMass)

1.3.3.2 Định luật Zipf

Định luật được đưa ra bởi Zipf năm 1949 được hiểu là: Trong văn bản có một số từ có tần số xuất hiện thấp thì ảnh hưởng đến ngữ nghĩa và lượng thông tin có trong văn bản, không cần thiết cho quá trình xử lý, cho nên ta có thể loại bỏ chúng để giảm số chiều của vector biểu diễn văn bản.

Năm 1958 Luhn đề xuất những từ "phổ biến" và "hiếm" và không cần thiết cho quá trình xử lý như sau:



Hình 1-4: Đồ thị thống kê tần số của từ theo định luật Zipf

1.4 Tiểu kết chương 1

Chương 1 trình bày những kiến thức cơ bản về khai phá dữ liệu và khám phá tri thức trong CSDL, các kỹ thuật phân cum trong khai phá dữ liệu, những chức năng chính, ứng dụng của nó trong xã hội,...

Chương này cũng trình bày một hướng nghiên cứu và ứng dụng trong khai phá dữ liệu là phân cụm dữ liệu, gồm tổng quan về kỹ thuật phân cụm, các ứng dụng của phân cụm, các yêu cầu đối với kỹ thuật phân cụm, các kiểu dữ liệu và độ đo tương tự,...

CHƯƠNG 2: MỘT SỐ KỸ THUẬT PHÂN CỤM DỮ LIỆU

Trong PCDL các thuật toán PCDL phải hướng tới hai mục tiêu là: Chất lượng của các cụm khám phá được và tốc độ thực hiện của thuật toán. Tuy nhiên, mỗi loại thuật toán PCDL có thể được áp dụng cho từng loại dữ liệu khác nhau [5].

2.1 Thuật toán k-means

Thuật toán phân cụm k-means do MacQueen 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 ban đầu gồm n đối tượng trong không gian d chiều Xi =(x_{i1}, x_{i2},…,x_{id}), i=(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 trọg tâm của cụm C_i, D là khoảng cách giữa hai đối tượng.

Trọng tâm của một cụm là một vector, trong đó giá trị của mỗi phần tử của nó là trung bình cộng các thành phần tương ứng của các đối tượng vector dữ liệu trong cụm đang xét. Tham số đầu vào của thuật toán là số cụm k, tập CSDL gồm n phần tử 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 Euclidean, 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 được chứng minh là hội tụ và có độ phức tạp tính toán là: $O((n \cdot k \cdot d) \cdot \tau \cdot T^{flop})$. Trong đó: n là số đối tượng dữ liệu, k là số cụm dữ liệu, d là số chiều, τ là số vòng lặp, T^{flop} là thời gian để thực hiện một phép tính cơ sở như phép tính nhân, chia,…

Thuật toán k-means bao gồm các bước cơ bản như sau:

INPUT: Một CSDL gồm n đối tượng và số các cụm k.

OUTPUT: Các cụm C_i (i=1,..,k) sao cho hàm tiêu chuẩn E đạt giá trị tối thiểu.

Bước 1: Khởi tạo

Chọn k đối tượng m_j (j=1...k) là trọng tâm ban đầu của k cụm từ tập dữ liệu (việc lựa chọn này có thể là ngẫu nhiên hoặc theo kinh nghiệm).

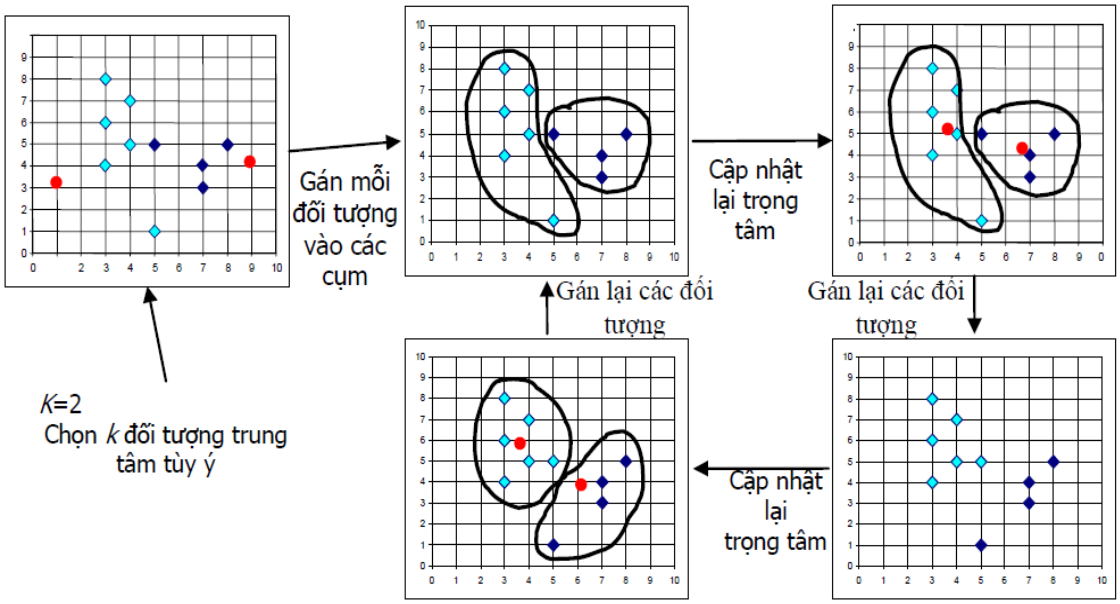
Bước 2: Tính toán khoảng cách

Đối với mỗi đối tượng x_i =(x_{i1}, x_{i2}, …,x_{id}), tính toán khoảng cách từ nó tới mỗi trọng tâm m_j (j=1,..,k), sau đó tìm trọng tâm gần nhất đối với mỗi đối tượng.

Bước 3: Cập nhật lại trọng tâm

Đối với mỗi *j=1,..,k,* cập nhật trọng tâm cụm mj bằng cách xác định trung bình cộng của các vector 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.



Hình 2-1: Hình dạng cụm dữ liệu được khám phá bởi k-means

Nhận xét: Do k-means phân tích phân cụm đơn giản nên có thể áp dụng đối với tập dữ liệu lớn. Nhưng k-means còn nhiều mặt hạn chế như: k-means chỉ áp dụng với dạng dữ liệu có thuộc tính số và có dạng hình cầu, k-means còn rất nhạy cảm với nhiễu và các phần tử ngoại lai trong dữ liệu. Ngoài ra, chất lượng PCDL của thuật toán k-means phụ thuộc nhiều vào các tham số đầu vào như: số cụm k và k trọng tâm khởi tạo ban đầu. Trong trường hợp, các trọng tâm khởi tạo ban đầu mà quá lệch so với các trọng tâm cụm tự nhiên thì kết quả phân cụm của k-means là rất thấp, cụm dữ liệu được khám phá rất lệch so với các cụm trong thực tế.

Hiện nay có rất nhiều thuật toán kế thừa tư tưởng của thuật toán k-means để KPDL ma có CSDL rất lớn như: k-medoid, PAM, CLARA, CLARANS, …

2.2 Thuật toán PAM

Thuật toán PAM (Partitioning Around Medoids) được Kaufman và Rousseeuw đề xuất 1987, là thuật toán mở rộng của thuật toán k-means, nhằm có khả năng xử lý hiệu quả đối với dữ liệu nhiễu hoặc các phần tử ngoại lai. Thay vì sử dụng các trọng tâm như k-means, PAM sử dụng các đối tượng medoid để biểu diễn cho các cụm dữ liệu, một đối tượng medoid là đối tượng thuộc cụm đó. Vì vậy, các đối tượng medoid ít bị ảnh hưởng của các đối tượng ở rất xa trung tâm, trong khi đó các trọng tâm của thuật toán k-means lại rất bị tác động bởi các điểm xa trung tâm này. Ban đầu, PAM chọn k đối tượng medoid và phân phối các đối tượng còn lại vào các cụm với các đối tượng medoid đại diện tương ứng sao cho chúng tương tự với đối tượng medoid trong cụm nhất. Sau mỗi bước thực hiện, PAM cố gắng hoán chuyển giữa đối tượng medoid Om và một đối tượng Op không phải là medoid, miễn là sự hoán chuyển này nhằm cải tiến chất lượng của phân cụm, quá trình này kết thúc khi chất lượng phân cụm không thay đổi. Chất lượng phân cụm được đánh giá thông qua hàm tiêu chuẩn, chất lượng phân cụm tốt nhất khi hàm tiêu chuẩn đạt giá trị tối thiểu [1].

Để quyết định hoán chuyển hai đối tượng O_m và O_p hay không, thuật toán PAM sử dụng giá trị tổng chi phí hoán chuyển C_{jmp} làm căn cứ:

O_m: Là đối tượng medoid hiện thời cần được thay thế

O_p: Là đối tượng medoid mới thay thế cho O_m

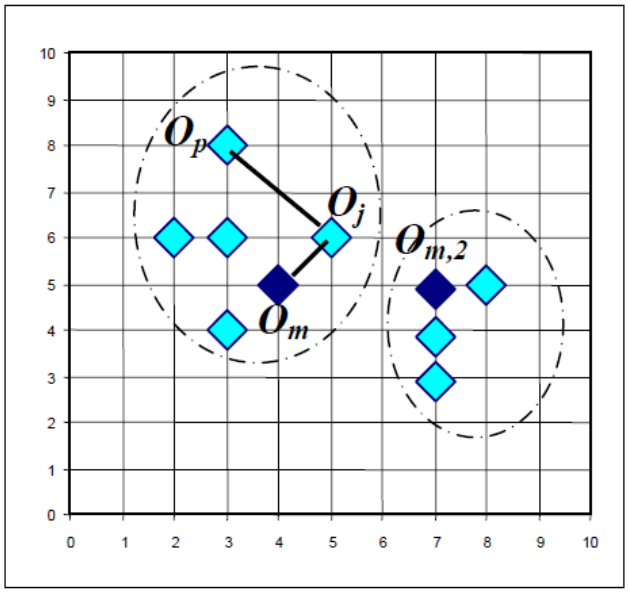
O_j: Là đối tượng dữ liệu (không phải là medoid) có thể được di chuyển sang cụm khác.

$O_{m,2}$: Là đối tượng medoid hiện thời khác với O_m mà gần đối tượng O_j nhất.

Sau đây là 4 trường hợp tính C_{jmp} để làm căn cứ hoán chuyển hai đối tượng medoid:

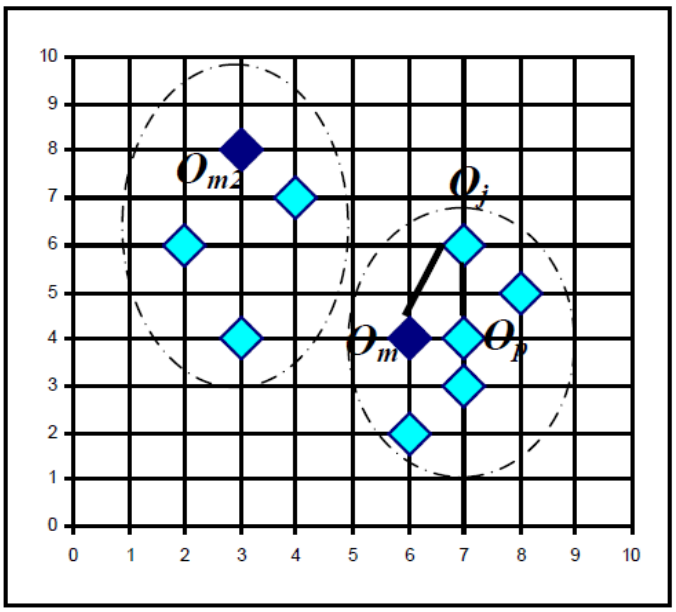
Trường hợp 1: Giả sử Oj hiện thời thuộc về cụm có đại diện là Om và Oj tương tự với $O_{m,2}$hơn O_p (d(O_j, O_p)≥d(O_j, $O_{m,2}$)). Trong khi đó, $O_{m,2}$ là đối tượng medoid tương tự xếp thứ 2 tới O_j trong số các medoid. Trong trường hợp này, ta thay thế O_m bởi đối tượng medoid mới O_p và O_j sẽ thuộc về cụm có đối tượng đại diện là $O_{m,2}$. Vì vậy, giá trị hoán chuyển C_{jmp} được xác định như sau: $C_{jmp} = \mathrm{d}(Oj, O_{m,2}) - \mathrm{d}(O_j, O_m)$. Giá trị C_{jmp} là không âm.

.



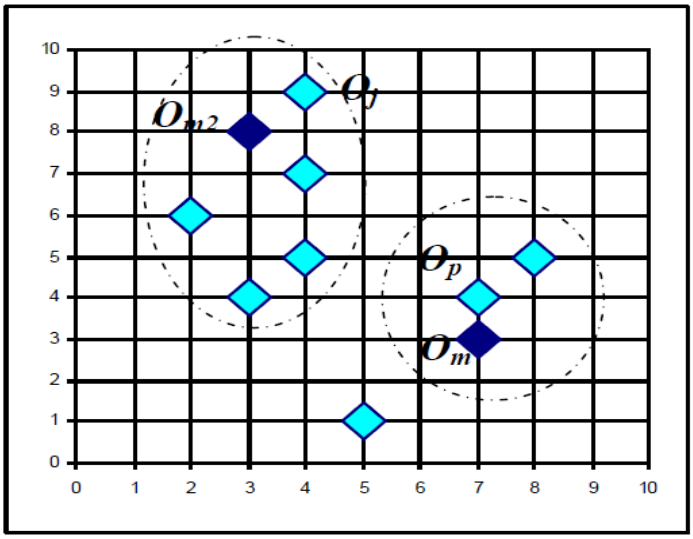
Hình 2-2: C_{jmp} = d(O_j, $O_{m,2}$) – d(O_j, O_m) Cjmp không âm

Trường hợp 2: O_j hiện thời thuộc về cụm có đại diện là O_m, nhưng O_j ít tương tự với $O_{m,2}$ so với O_p *(d(O_j,O_p)< d(O_j,$O_{m,2}$)).* Nếu thay thế O_m bởi O_p thì O_j sẽ thuộc về cụm có đại diện là O_p. Vì vậy, giá trị C_{jmp} được xác định như sau: $C_{jmp} = d(O_j, O_p) - d(O_j, O_m)$. C_{jmp} ở đây có thể là âm hoặc dương.



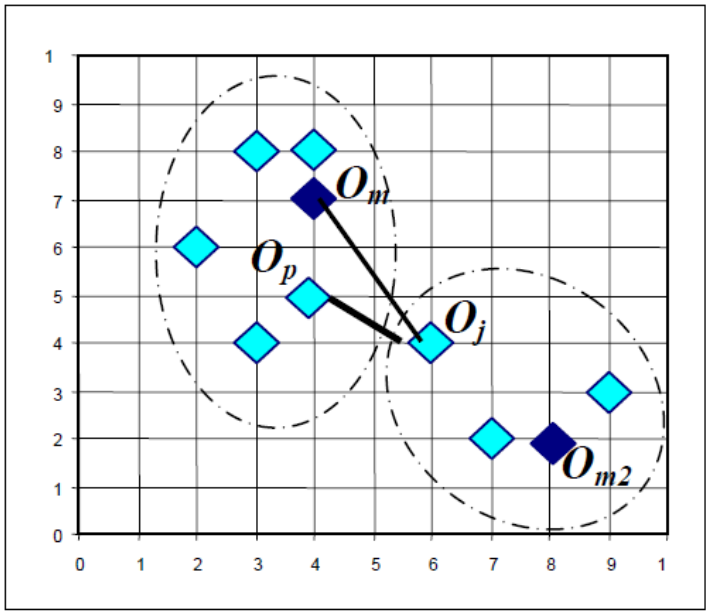
Hình 2-3 : $C_{jmp} = d(O_j, O_p) - d(O_j, O_m)$ có thể âm hoặc dương.

Trường hợp 3: Giả sử O_j hiện thời không thuộc về cụm có đối tượng đại diện là O_m mà thuộc về cụm có đại diện là $O_{m,2}$. Mặt khác, giả sử O_j tương tự với $O_{m,2}$ hơn so với O_p, nếu O_m được thay thế bởi O_p thì O_j vẫn sẽ ở lại trong cụm có đại diện là Om,2. Do đó: C_{jmp}= 0.



Hình 2-4 Trường hợp C_{jmp}= 0

Trường hợp 4: O_j hiện thời thuộc về cụm có đại diện là $O_{m,2}$ nhưng O_j ít tương tự tới $O_{m,2}$ hơn so với O_p. Vì vậy, nếu ta thay thế O_m bởi O_p thì O_j sẽ chuyển từ cụm $O_{m,2}$ sang cụm O_p. Do đó, giá trị hoán chuyển C_{jmp} được xác định là: C_{jmp}= (O_j,O_p)-$d(O_j, O_{m,2})$. C_{jmp} ở đây luôn âm.



Hình 2-5: Trường hợp C_{jmp}= (O_j,O_p)- $d(O_j, O_{m,2})$. C_{jmp} luôn âm

Hình 2 1 Trường hợp Cjmp= (Oj,Op)- d(Oj, Om,2) luôn âm

- Kết hợp cả bốn trường hợp trên, tổng giá trị hoán chuyển Om bằng Op được xác định như sau: $T_{Cmp} = \sum_j C_{imp}$

Thuật toán PAM bao gồm các bước sau:

INPUT: Tập dữ liệu có n phần tử, số cụm k.

OUTPUT: k cụm dữ liệu sao cho chất lượng phân hoạch là tốt nhất.

Bước 1: Chọn k đối tượng medoid bất kỳ.

Bước 2: Tính T_{Cmp} cho tất cả các cặp đối tượng O_m, O_p. Trong đó O_m là đối tượng medoid và O_p là đối tượng không phải là modoid.

Bước 3: Với mỗi cặp đối tượng O_m và O_p. Tính min(O_m), min(O_p), T_{Cmp}. Nếu T_{Cmp} là âm, thay thế O_m bởi O_p và quay lại bước 2. Nếu T_{Cmp} dương, chuyển sang bước 4.

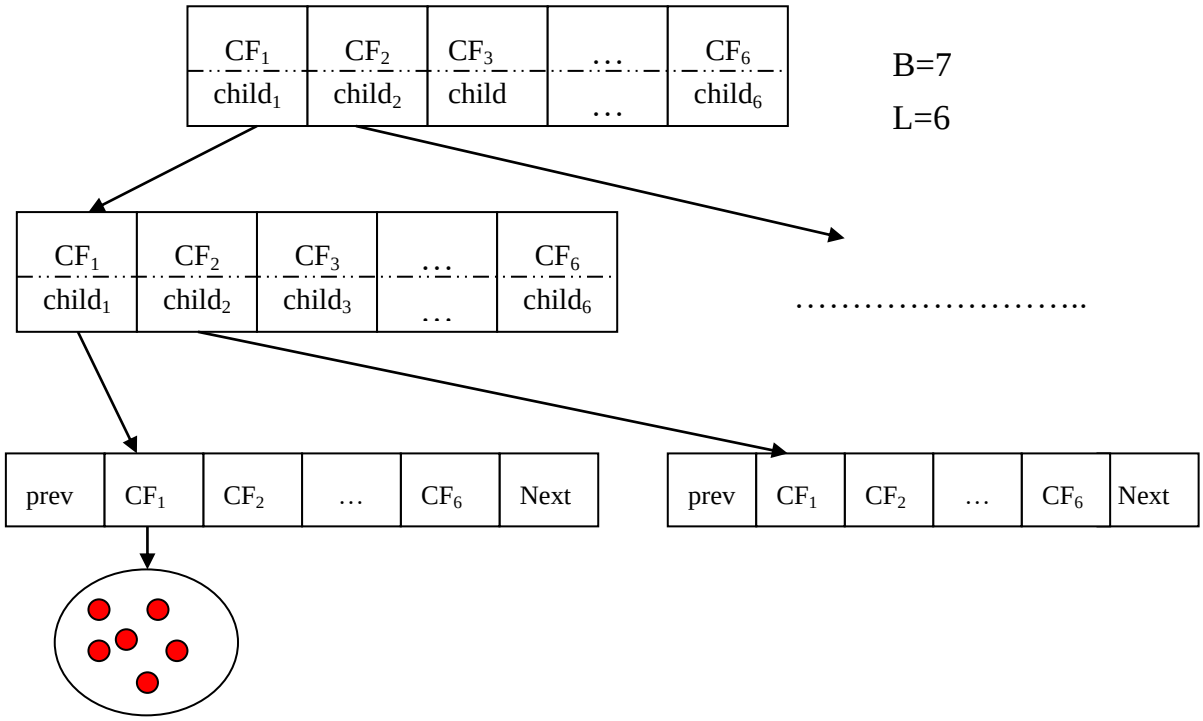
Bước 4: Với mỗi đối tượng không phải là medoid, xác định đối tượng medoid tương tự với nó nhất đồng thời gán nhãn cụm cho chúng.

Độ phức tạp tính toán của PAM là $O(i \cdot k \cdot (n-k)^2)$, i là số vòng lặp. Vì, PAM phải duyệt tất cả $k(n-k)$ cặp O_m, O_p và việc tính toán T_{Cmp} yêu cầu kiểm tra $n-k$ đối tượng. Vậy nên, thuật toán PAM kém hiệu quả về thời gian tính toán khi giá trị của k và n là lớn.

Nhận xét: Ngoài việc kế thừa được ưu điểm của thuật toán k-means ra thì PAM còn khắc phục được việc sử lý dữ liệu nhiễu và các phần tử ngoại lai. Mặt khác, thời gian tính toán của PAM kém hiệu quả khi CSDL lớn và nhiều medoid.

2.3 Thuật toán BIRCH

BIRCH *(Balanced Iterative Reducing and Clustering Using Hierarchies)* do Tian Zhang, amakrishnan và Livny đề xuất năm 1996, là thuật toán phân cụm phân cấp sử dụng chiến lược 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 cụm dữ liệu, BIRCH chỉ lưu một bộ ba (n, LS, SS), với 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 các giá trị thuộc tính của các đối tượng trong cụm. Các bộ ba này được gọi là các đặc trưng của cụm CF=(n, LS, SS) (Cluster Features - CF) và được lưu giữ trong một cây được gọi là cây CF. Hình sau đây biểu thị một ví dụ về cây CF. Chúng ta thấy rằng, tất cả các nút trong của cây lưu tổng các đặc trưng cụm CF của nút con, trong khi đó các nút lá lưu trữ các đặc trưng của các cụm dữ liệu.



Hình 2-6: Cây CF được tạo bởi BIRCH

Cây CF là cây cân bằng, nhằm để lưu trữ các đặc trưng của cụm. Cây CF chứa các nút trong và nút lá. Nút trong lưu giữ tổng các đặc trưng cụm của các nút con của nó. Một cây CF được đặc trưng bởi hai tham số:

- *Yếu tố nhánh (B)*: Nhằm xác định số tối đa các nút con của mỗi nút trong của cây.

- *Ngưỡng (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.

Thuật toán BIRCH:

INPUT: CSDL gồm n đối tượng, ngưỡng T

OUTPUT: k cụm dữ liệu

***Bước 1*:** Duyệt tất cả các đối tượng trong CSDL và xây dựng một 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.

Bước 2: Nếu cây CF hiện thời không có đủ bộ nhớ trong thì tiến hành xây dựng một cây CF nhỏ hơn bằng cách điều khiển bởi tham số *T* (vì tăng T sẽ làm 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.

Bước 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.

Bước 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

Nhận xét:

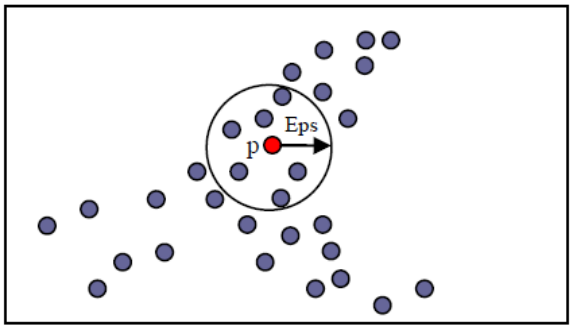
Ưu điểm: Thuật toán BIRCH có tốc độ thực hiện PCDL nhanh và có thể áp dụng đối với tập dữ liệu lớn, BIRCH đặc biệt hiệu quả khi áp dụng với tập dữ liệu tăng trưởng theo thời gian. BIRCH có độ phức tạp là $O(n)$, vì vậy, *BIRCH* có tốc độ xử lý rất nhanh trong PCDL.

Nhược điểm: BIRCH khám phá các cụm có chất lượng được không được tốt. BIRCH chỉ phù hợp với dữ liệu số, phụ thuộc vào thứ tự của dữ liệu, ngưỡng T có ảnh hưởng rất lớn tới cụm và BIRCH không phù hợp với dữ liệu đa chiều.

2.4 Thuật toán DBSCAN

Thuật toán DBSCAN (Density Based Spatial Clustering of Applications with Noise) do Martin Ester và các tác giả khác đề xuất là thuật toán phân cụm dựa trên mật độ, hiệu quả với cơ sở dữ liệu lớn, có khả năng xử lý nhiễu. Ý tưởng chính của thuật toán là vùng lân cận mỗi đối tượng trong một cụm có số đối tượng lớn hơn ngưỡng tối thiểu. Hình dạng vùng lân cận phụ thuộc vào hàm khoảng cách giữa các đối tượng (nếu sử dụng khoảng cách Manhattan trong không gian 2 chiều thì vùng lân cận có hình chữ nhật, nếu sử dụng khoảng cách Eucler trong không gian 2 chiều thì vùng lân cận có hình tròn) [3].

Định nghĩa 1: Các lân cận của một điểm P với ngưỡng Eps, ký hiệu $N_{Eps(p)}$ được xác định như sau: $N_{Eps(p)} = \{q \in D \mid$ khoảng cách $Dist(p,q) \leq Eps\}$, D là tập dữ liệu cho trước.



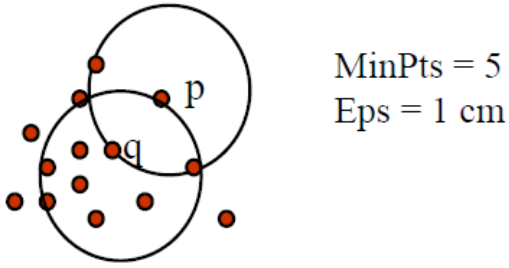
Hình 2-7: Lân cận của một điểm p với ngưỡng Eps

Một điểm p muốn nằm trong một cụm C nào đó thì $N_{Eps(p)}$ phải có tối thiểu *MinPts* điểm. Như vậy, chỉ những điểm thực sự nằm trong cụm mới thoả mãn điều kiện là điểm thuộc vào cụm. Những điểm nằm ở biên của cụm thì không thoả mãn điều kiện đó, bởi vì thông thường thì lân cận với ngưỡng Eps của điểm biên thì bé hơn lân cận với ngưỡng cũng Eps của điểm nhân.

Định nghĩa 2 (*Mật độ - đến được trực tiếp (Directly Density – reachable)):* Một điểm *p* được gọi là *mật độ-đến được trực tiếp* từ điểm *q* với ngưỡng *Eps* và *MinPts* trong tập đối tượng *D* nếu:

1) $p \in N_{Eps(q)}$ Với $N_{Eps(q)}$ là tập con của *D*

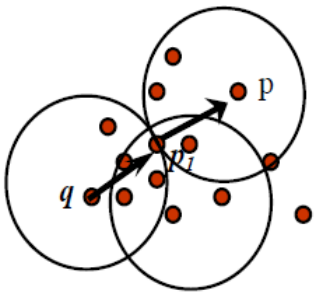
2) $||N_{Eps(q)}|| \geq MinPts$ Điều kiện đối tượng nhân.



Hình 2-8: Mật độ-đến được trực tiếp

- Nếu *p, q* đều là đối tượng nhân, quan hệ *mật độ - đến được trực tiếp (Directly Density -reachable)* là đối xứng nghĩa là *p* tới được trực tiếp theo mật độ từ *q* và ngược lại.
- Nếu trong *p, q* có một đối tượng nhân, một đối tượng biên như hình trên thì chỉ đối tượng biên có tới được trực tiếp theo mật độ từ đối tượng nhân mà không có chiều ngược lại (bất đối xứng).

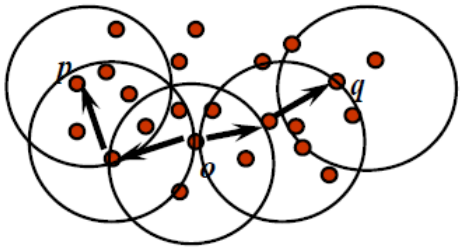
Định nghĩa 3 Mật độ - đến được (Density - Reachable): Một điểm p được gọi là *mật độ- đến được* từ một điểm q với hai tham số Eps và MinPts nếu tồn tại một dãy $p = p_1, p_2, ..., p_n = q$ sao cho p_{i+1} là *mật độ- đến được trực tiếp từ* p_i *(i=1,n-1).*



Hình 2-9: Mật độ - đến được

Hai điểm p,q đến được với nhau, vì *p,q* không là điểm nhân và tồn tại điểm nhân p_1 trong cụm mà hai điểm *p,q* có thể đến được nó.

Định nghĩa 4 Mật độ - liên thông (Density - Connected): Đối tượng *p* là *mật độ-liên thông* với điểm *q* theo hai tham số *Eps* với *MinPts* nếu như có một đối tượng *o* mà cả hai đối tượng *p, q* điều là *mật độ- đến được o* theo tham số *Eps* và *MinPts*.



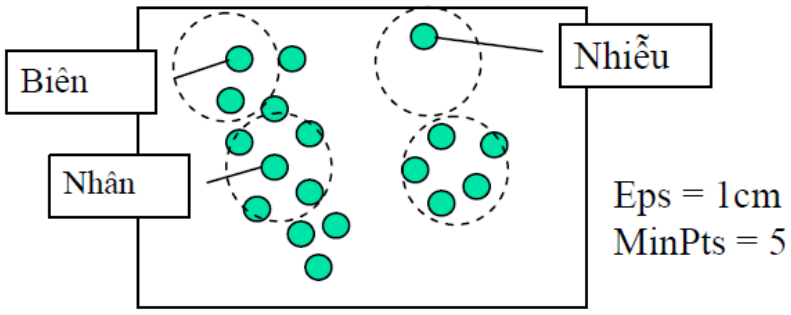
Hình 2-10: Mật độ- liên thông

Định nghĩa 5: Cho CSDL *D*, cụm *C* thỏa mãn *Eps* và *MinPts* sẽ được gọi là *tập con khác rỗng* của *D* nếu thỏa mãn hai điều kiện sau:

1) Cực đại: Với ∀ *p,q* ∈ *D*, nếu *p* ∈ *C* và *q* là *mật độ- đến được p* theo *Eps* và *MinPts* thì *q* ∈ *C*.

2) Với ∀ *p,q* ∈ *C, p* là *mật độ-liên thông* với *q* theo *Eps* và *MinPts*.

Mọi đối tượng không thuộc cụm nào cả thì gọi là nhiễu.



Hình 2-11: Các đối tượng nhiễu

Có hai bổ đề trong thuật toán DBSCAN:

Bổ đề 1: Giả sử p là một đối tượng trong *D*, trong đó $||N_{Eps(p)}|| \geq MinPts$, tập *O = {o|o ∈ D và o là mật độ-đến được từ p theo Eps và MinPts}* là một cụm theo *Eps* và *MinPts*. Như vậy, cụm *C* không hoàn toàn là duy nhất, tuy nhiên, mỗi một điểm trong C đến được mật độ từ bất cứ một điểm nhân nào của C, vì vậy C chứa đúng một số điểm liên thông với điểm nhân tuỳ ý.

Bổ đề 2: Giả sử *C* là một cụm theo *Eps* và *MinPts*, *p* là một điểm bất kỳ trong C với $||N_{Eps(p)}|| \geq MinPts$. Khi đó *C* trùng với tập *O = {o|o□D và o là mật độ-đến được từ p theo Eps và MinPts}.*

Các bước của thuật toán DBSCAN như sau:

INPUT: *Eps* và *MinPts*

OUTPUT: Các cụm dữ liệu sau khi được khám phá

Bước 1: Chọn một đối tượng p tuỳ ý

Bước 2: Lấy tất cả các *đối tượng mật độ - đến được* từ p với *Eps* và *MinPts*.

Bước 3: Nếu p là điểm nhân thì tạo ra một cụm theo *Eps* và *MinPts*.

Bước 4: Nếu p là một điểm biên, không có điểm nào là *mật độ - đến được* từ p và DBSCAN sẽ đi thăm điểm tiếp theo của tập dữ liệu.

Bước 5: Quá trình tiếp tục cho đến khi tất cả các đối tượng được xử lý.

Khi sử dụng *Eps* và *MinPts* toàn cục thì DBSCAN có thể gộp hai cụm thành một cụm nếu mật độ của chúng gần bằng nhau.

Điểm mạnh của thuật toán DBSCAN: DBSCAN có thể khám phá các cụm có hình dáng bất kỳ, ít bị ảnh hưởng bởi thứ tự của các đối tượng dữ liệu đầu vào. Thuận lợi cho việc cập nhật dữ liệu, vì đối tượng được chèn vào chỉ tác động đến một láng giềng xác định.

2.5 Tiểu kết chương 2

Chương này trình bày một số thuật toán phân cụm dữ liệu phổ biến như: k-means, PAM, DBSCAN.

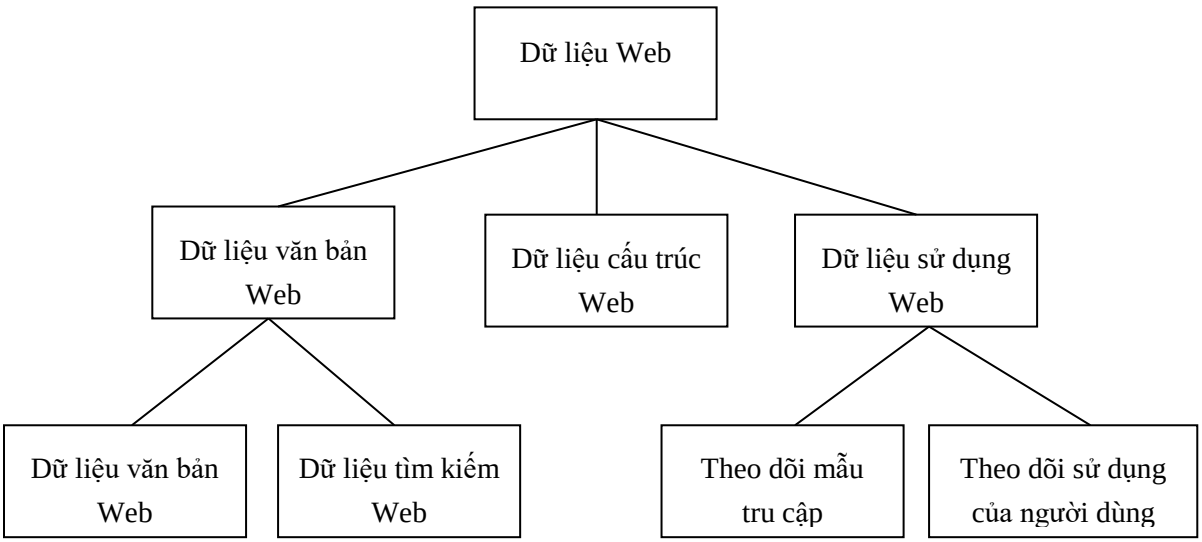
Thuật toán k-means, PAM dựa trên ý tưởng ban đầu tạo ra *k* cụm bất kỳ, sau đó lặp lại nhiều lần để phân bố lại các đối tượng dữ liệu giữa các cụm nhằm cải thiện chất lượng phân cụm.

Thuật toán BIRCH dựa trên ý tưởng cây phân cấp để phân cụm dữ liệu theo *top-down* hoặc *botton-up*

Thuật toán DBSCAN căn cứ vào hàm mật độ của các đối tượng dữ liệu để xác định cụm cho các đối tượng.

CHƯƠNG 3: KHAI PHÁ DỮ LIỆU WEB

Có ba hướng tiếp cận chính trong khai phá Web là Web content, Web structure, Web usage.



Hình 3-1: Phân loại khai phá Web

3.1 Khai phá nội dung Web

Khai phá nội dung Web tập trung vào việc nội dung của các trang Web,bao gồm: văn bản, âm thanh, hình ảnh,… Trong lĩnh vực khai phá Web, khai phá nội dung Web được xem xét như là kỹ thuật KPDL đối với CSDL quan hệ, bởi nó có thể phát hiện ra các kiểu tương tự của tri thức từ kho dữ liệu không cấu trúc trong các tài liệu Web. Nhiều tài liệu Web là nửa cấu trúc (như HTML) hoặc dữ liệu có cấu trúc (như dữ liệu trong các bảng hoặc CSDL tạo ra các trang HTML) nhưng phần đa dữ liệu văn bản là không cấu trúc. Đặc điểm không cấu trúc của dữ liệu đặt ra cho việc khai phá nội dung Web những nhiệm vụ phức tạp và thách thức.

Khai phá nội dung Web có nhiều cách tiếp cận, nhưng trong khuôn khổ đồ án sẽ xem xét dưới góc độ: *Khai phá kết quả tìm kiếm* và *khai phá nội dung trang HTML*.

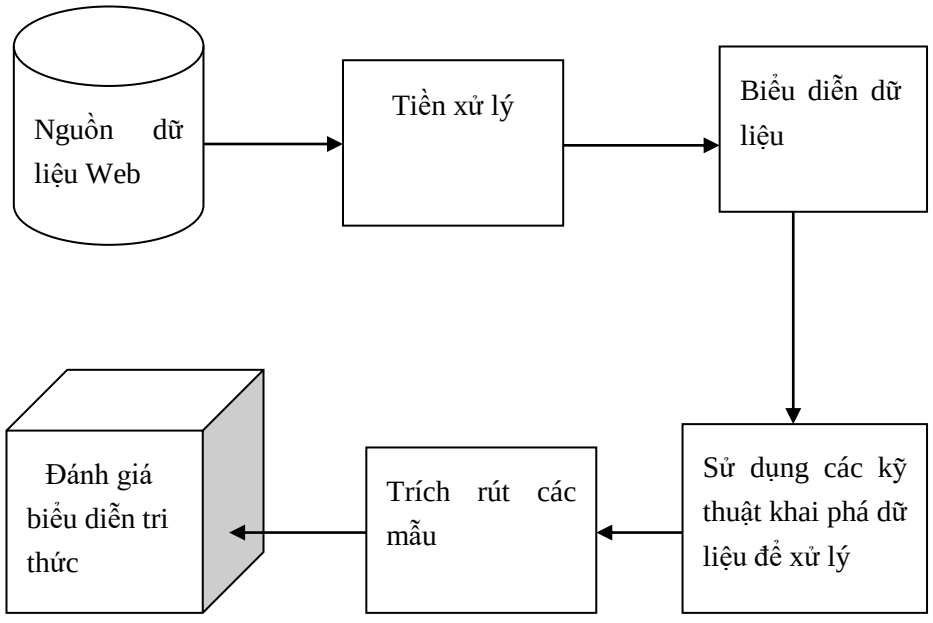
3.1.1 Khai phá kết quả tìm kiếm

Hiện nay, việc sử dụng công cụ Web Searching Engine để phân loại tự động tài liệu Web đang rất phát triển, do Web Searching Engine có thể đánh trọng số cho các trang Web. Việc đánh trọng số trang Web gồm các bước sau: Đầu tiên, tải về các dữ liệu Web từ các Website. Tiếp theo, Web Search Engine trích chọn ra những thông tin chỉ mục mô tả từ các trang Web đó để lưu trữ chúng cùng với URL của nó trong Web Search Engine. Cuối cùng, sử dụng các phương pháp KPDL để phân lớp tự động và tạo điều kiện thuận tiện cho hệ thống phân loại trang Web và được tổ chức bằng cấu trúc siêu liên kết.

Trực quan hoá kết quả tìm kiếm: Trong quá trình phân loại, có nhiều tài liệu thông tin không liên quan nhau. Nếu ta có thể phân tích và phân cụm kết quả tìm kiếm, thì hiệu quả tìm kiếm sẽ được cải thiện tốt hơn, nghĩa là các tài liệu "tương tự" nhau về mặt nội dung thì đưa chúng vào cùng nhóm, các tài liệu "phi tương tự" thì đưa chúng vào các nhóm khác nhau. Hay, ta có thể phân cụm các tài liệu theo một vài tiêu trí để người dùng có thể tìm kiếm theo từng nhóm.

3.1.2 Khai phá văn bản Web

KPVB là việc sử dụng kỹ thuật KPDL đối với các tập văn bản để tìm ra tri thức tiềm ẩn trong đó. Vì vậy, KPVB liên quan đến các kỹ thuật: KPDL, xử lý ngôn ngữ tự nhiên, tìm kiếm thông tin. Các đối tượng của KPVB là: Dữ liệu có cấu trúc, không cấu trúc và nửa cấu trúc. Kết quả của KPVB là: trạng thái trung của mỗi tài liệu, phân loại tài liệu để phục vụ cho mục đích nào đó. Quá trình của việc KPVB như sau [LizhenLiu]:



Hình 3-2: Quá trình khai phá văn bản Web

3.1.2.1 Nguồn dữ liệu

Nguồn dữ liệu Web là văn bản cục bộ trên các trang Web được định dạng, tích hợp thành các tài liệu theo mong muốn để khai phá và phân phối trong nhiều dịch vụ Web bằng việc sử dụng kỹ thuật truy xuất thông tin.

3.1.2.2 Tiền xử lý dữ liệu

Tiền xử lý là quá trình sử dụng các quy tắc hoặc phương pháp để làm rõ dữ liệu. Tức là: Làm cho dữ liệu trở lên rõ ràng, chính xác và xóa bỏ dữ liệu hỗn độn, dư thừa, để phục vụ cho quá trình khai phá. Các bước cơ bản của quá trình tiền sử lý: Đầu tiên, tìm hiểu yêu cầu của người dùng và tìm ra mối quan hệ giữa các tri thức từ đối tượng dữ liệu. Tiếp theo, làm sạch, chuẩn hóa và sắp xếp lại những tri thức này. Kết quả cuối cùng là bảng hai chiều có các đặc trưng sau [LizhenLiu]:

- Dữ liệu thống nhất.
- Làm sạch dữ liệu không liên quan, nhiễu và dữ liệu rỗng. Dữ liệu không bị mất mát và không bị lặp.
- Giảm bớt số chiều và làm tăng hiệu quả việc phát hiện tri thức bằng việc chuyển đổi, quy nạp, cưỡng bức dữ liệu,...
- Làm sạch các thuộc tính không liên quan để giảm bớt số chiều của dữ liệu.

3.1.2.3 Biểu diễn văn bản

KPVB Web là khai phá các tập tài liệu HTML. Do đó ta sẽ phải biến đổi và biểu diễn dữ liệu thích hợp cho quá trình xử lý. Ta có thể xử lý và lưu trữ chúng trong mảng 2 chiều mà dữ liệu đó có thể phản ánh đặc trưng của tài liệu. Người ta thường dùng mô hình TF-IDF để vector hóa dữ liệu. Nhưng có một vấn đề quan trọng là việc biểu diễn này sẽ dẫn đến số chiều vector khá lớn. Vì thế,ta cần lựa chọn các đặc trưng mà nó chắc chắn trở thành khóa và nó ảnh hưởng trực tiếp đến hiệu quả KPVB.

Loại bỏ từ dừng và phân lớp: Đầu tiên, ta áp dụng định luật *Zipf* để loại bỏ các từ có tần số xuất hiện quá cao (ví dụ: và, hoặc, hay,…), nhưng mang ít thông tin và loại bỏ các từ có tần số xuất hiện quá ít để loại bỏ dữ liệu nhiễu. Tiếp theo, ta chọn lọc các từ còn lại và phân lớp cho chúng để mô tả các đặc trưng của tài liệu.

3.1.2.4 Trích rút các từ đặc trưng

Rút ra các đặc trưng là một phương pháp, nó có thể giải quyết số chiều vector đặc trưng lớn được mang lại bởi kỹ thuật KPVB. Việc rút ra các đặc trưng dựa trên hàm trọng số:

Phương pháp tính trọng số tần xuất logarit (TF:term-frequency):

TF của một từ *t* trong tài liệu d được tính như sau:

$$w_{t,d} = \begin{cases} 1 + log(tf_{t,d}) & Nếu\ tf_{t,d} > 0 \\ 0 & Nếu\ tf_{t,d} = 0 \end{cases}$$

$tf_{t,d}$: Tần số từ *t* trong tài liệu *d*

Điểm cho một cặp tài *liệu-truy vấn (document-query*) được tính bằng tổng của các trọng số của từ t trong tài liệu d và truy vấn $q = \sum_{t \in q \cap d} (1 + log(tft, d))$. Điểm sẽ bằng 0 nếu như từ truy vấn *(query terms)* không xuất hiện trong tài liệu.

Phương pháp tính trọng số nghịch đảo văn bản:

Trong việc tính trọng số tần xuất (TF), thì những từ có tần số xuất hiện cao thì có điểm càng cao. Nhưng những từ hiếm lại mang nhiều ý nghĩa hơn, cho nên ta cần một đánh giá khác cho những từ hiếm.

Trọng số idf_t của từ *t:* $idf_t = \log_{10}\left(\frac{N}{df_t}\right)$

N: Tổng số tài liệu

df_t: Tổng số tài liệu mà *d* xuất hiện

Lưu ý: idf không có ảnh hưởng trên việc xếp hạng tài liệu (với 1 từ khóa), nó chỉ giúp phân loại tài liệu. *idf* chỉ có ảnh hưởng lên sự xếp hạng các tài liệu nếu như truy vấn có ít nhất 2 từ.

3.1.2.5 Khai phá dữ liệu văn bản

Sau khi tập hợp, lựa chọn và trích ra tập văn bản hình thành nên các đặc trưng cơ bản, nó sẽ là cơ sở để KPDL. Từ đó ta có thể thực hiện trích, phân loại, phân cụm, phân tích và dự đoán.

a. Trích rút văn bản

Việc trích rút văn bản là để đưa ra các từ có ý nghĩa chính để mô tả tóm tắt tài liệu văn bản trong quá trình tổng hợp. Sau đó, người dùng có thể hiểu ý nghĩa chính của văn bản nhưng không cần thiết phải duyệt toàn bộ văn bản. Đây là phương pháp

đặc biệt được sử dụng trong searching engine, thường cần để đưa ra văn bản trích dẫn. Nhiều searching engines luôn đưa ra những câu dự đoán trong quá trình tìm kiếm và trả về kết quả, cách tốt nhất để thu được ý nghĩa chính của một văn bản hoặc tập văn bản chủ yếu bằng việc sử dụng nhiều thuật toán khác nhau. Theo đó, hiệu quả tìm kiếm sẽ tốt hơn và phù hợp với sự lựa chọn kết quả tìm kiếm của người dùng.

b. Phân lớp văn bản

Trước hết, nhiều tài liệu được phân lớp tự động một cách nhanh chóng và hiệu quả cao. Thứ hai, mỗi lớp văn bản được đưa vào một chủ đề phù hợp. Do đó nó thích hợp với việc tìm và duyệt qua các tài liệu Web của người sử dụng.Ta thường sử dụng phương pháp phân lớp Navie Bayesian và "K-láng giềng gần nhất" (K-Nearest Neighbor) để khai phá thông tin văn bản. Trong phân lớp văn bản, đầu tiên là phân loại tài liệu. Thứ hai, xác định đặc trưng thông qua số lượng các đặc trưng của tập tài liệu huấn luyện. Cuối cùng, tính toán kiểm tra phân lớp tài liệu và độ tương tự của tài liệu phân lớp bằng thuật toán nào đó. Khi đó các tài liệu có độ tương tự cao với nhau thì nằm trong cùng một phân lớp. Độ tương tự sẽ được đo bằng hàm đánh giá xác định trước. Nếu ít tài liệu tương tự nhau thì đưa nó về 0. Nếu nó không giống với sự lựa chọn của phân lớp xác định trước thì xem như không phù hợp. Sau đó, ta phải chọn lại phân lớp. Trong việc lựa chọn có 2 giai đoạn: *Huấn luyện* và *phân lớp.*

Thuật toán phân lớp K-Nearest Neighbor :

- Lựa chọn trước đặc trưng phân lớp, $Y=\{y_1, y_2,..., y_m\}$
- Tập tài liệu huấn luyện cục bộ, $X=\{x_1, x_2,...x_n\}$, $v(x_j)\}$ là vector đặc trưng của x_j.
- Mỗi $v(y_i)$ trong Y được xác định bằng $v(x_j)$ thông qua việc huấn luyện $v(x_j)$ trong X.
- Tập tài liệu kiểm tra, $C=\{c_1, c_2,...,c_p\}$, c_k trong C là một tài liệu phân lớp mong đợi, công việc của ta là tính toán độ tương tự giữa $v(c_k)$ và $v(y_i)$, $sim(c_k,y_i)$.
- Lựa chọn tài liệu c_k mà độ tương tự của nó với yi lớn nhất, như vậy c_k nằm trong phân lớp với y_i, với $max(sim(c_k,y_i)), (i=1,...,m)$.
- Quá trình được thực hiện lặp lại cho tới khi tất cả các tài liệu đã được phân lớp.

c. Phân cụm văn bản

Chủ đề phân loại không cần xác định trước. Nhưng ta phải phân loại các tài liệu vào nhiều cụm. Trong cùng một cụm, thì tất cả độ tương tự của các tài liệu yêu cầu cao hơn, ngược lại ngoài cụm thì độ tương tự thấp hơn. Như là một quy tắc, quan hệ các

cụm tài liệu được truy vấn bởi người dùng là "gần nhau". Do đó, nếu ta sử dụng trạng thái trong vùng hiển thị kết quả searching engine bởi nhiều người dùng thì nó được giảm bớt rất nhiều. Hơn nữa, nếu phân loại cụm rất lớn thì ta sẽ phân loại lại nó cho tới khi người dùng được đáp ứng với phạm vi tìm kiếm nhỏ hơn. Phương pháp sắp xếp liên kết và phương pháp phân cấp thường được sử dụng trong phân cụm văn bản.

Thuật toán phân cụm phân cấp:

- Trong tập tài liệu xác định, $W=\{w_1, w_2, ..,w_m\}$, mỗi tài liệu w_i là một cụm c_i, tập cụm C là $C=\{c_1, c_2, ...c_m\}$.
- Chọn ngẫu nhiên 2 cụm c_i và c_j, tính độ tương tự $sim(c_i,c_j)$ của chúng. Nếu độ tương tự giữa c_i và c_j là lớn nhất, ta sẽ đưa ci và cj vào một cụm mới. cuối cùng ta sẽ hình thành được cụm mới $C=\{c_1, c_2,..c_{m-1}\}$.
- Lặp lại công việc trên cho tới khi chỉ còn 1 phân tử.

Toàn bộ quá trình của phương pháp sắp xếp liên kết sẽ tạo nên một cây mà nó phản ánh mối quan hệ với nhau về độ tương tự giữa các tài liệu. Phương pháp có tính chính xác cao. Nhưng tốc độ của nó rất chậm bởi việc phải so sánh độ tương tự trong tất cả các cụm. Nếu tập tài liệu lớn thì phương pháp này không khả thi.

Thuật toán phân cụm phân hoạch

- Trước hết ta sẽ chia tập tài liệu thành các cụm khởi đầu thông qua việc tối ưu hóa hàm đánh giá theo một nguyên tắc nào đó, $R=\{R_1, R_2,...,R_n\}$, với n phải được xác định trước.
- Với mỗi tài liệu trong tập tài liệu W, $W=\{w_1, w_2,..,w_m\}$, tính toán độ tương tự của nó tới R_j ban đầu, $sim(w_i, R_j)$, sau đó lựa chọn tài liệu tương tự lớn nhất, đưa nó vào cụm R_j.
- Lặp lại các công việc trên cho tới khi tất cả các tài liệu đã đưa vào trong các cụm xác định.

Phương pháp này có các đặc điểm là kết quả phân cụm ổn định và nhanh chóng. Nhưng ta phải xác định trước các phần tử khởi đầu và số lượng của nó, mà chúng sẽ ảnh hưởng trực tiếp đến hiệu quả phân cụm.

d. Phân tích và dự đoán xu hướng

Thông qua việc phân tích các tài liệu Web, ta có thể nhận được quan hệ phân phối của các dữ liệu đặc biệt trong từng giai đoạn của nó và có thể dự đoán được tương lai phát triển.

e. Đánh giá chất lượng mẫu

KPDL Web có thể được xem như quá trình của học máy. Kết quả của học máy là các mẫu tri thức. Phần quan trọng của học máy là đánh giá kết quả các mẫu. Ta thường phân lớp các tập tài liệu vào tập huấn luyện và tập kiểm tra. Sau đó lặp lại việc học và kiểm thử trong tập huấn luyện và tập kiểm tra. Cuối cùng, chất lượng trung bình được dùng để đánh giá chất lượng mô hình.

3.2 Khai phá theo sử dụng Web

Khai phá theo sử dụng Web là việc dựa vào các mẫu truy xuất của người dùng, để thiết kế các trang Web theo các lĩnh vực khác nhau, phù hợp nhất với phần lớn người dùng internet. Ngoài ra, ta có thể khai phá các thông tin khác như: Xác định hành vi của người dùng, điều tra sự quan tâm của người dùng đến trang Web đó.

Khai phá theo sử dụng Web là khai phá truy cập Web (Web log) để khám phá các mẫu người dùng truy nhập vào Website. Thông qua việc phân tích và khảo sát những quy tắc trong việc ghi nhận lại quá trình truy cập Web ta có thể chứng thực khách hàng trong thương mại điện tử, nâng cao chất lượng dịch vụ thông tin trên Internet đến người dùng, nâng cao hiệu suất của các hệ thống phục vụ Web. Thêm vào đó, để tự phát triển các Website bằng việc huấn luyện từ các mẫu truy xuất của người dùng. Phân tích quá trình đăng nhập Web của người dùng cũng có thể giúp cho việc xây dựng các dịch vụ Web theo yêu cầu đối với từng người dùng riêng lẽ được tốt hơn.

Ứng dụng của khai phá theo sử dụng Web

- Cải tiến hiệu suất hệ thống phục vụ của các máy chủ Web.
- Cá nhân dịch vụ Web thông quan việc phân tích các đặc tính cá nhân người dùng.
- Tìm ra những khách hàng tiềm năng trong thương mại điện tử.
- Chính phủ điện tử (e-Gov), giáo dục điện tử (e-Learning).
- Xác định những quảng cáo tiềm năng.
- Cải tiến thiết kế Web thông qua việc phân tích thói quen duyệt Web và phân tích các mẫu nội dung trang quy cập của người dùng.
- Phát hiện gian lận và xâm nhập bất hợp lệ trong dịch vụ thương mại điện tử và các dịch vụ Web khác.
- Nâng cao chất lượng truyền tải của các dịch vụ thông tin Internet đến người dùng.
- Thông qua việc phân tích chuỗi truy cập của người dùng để có thể dự báo những hành vi của người dùng trong quá trình tìm kiếm thông tin.

3.2.1 Các kỹ thuật được sử dụng trong khai phá theo sử dụng Web

Luật kết hợp: Để tìm ra những trang Web thường được truy cập cùng nhau của người dùng những lựa chọn cùng nhau của khách hàng trong thương mại điện tử.

Kỹ thuật phân cụm: Phân cụm người dùng dựa trên các mẫu duyệt để tìm ra sự liên quan giữa những người dùng Web và các hành vi của họ.

3.2.1.1 Những vấn đề trong khai khá theo sử dụng Web.

Khai phá theo cách dùng Web có 2 việc: Trước tiên, Web log cần được làm sạch, định nghĩa, tích hợp và biến đổi. Dựa vào đó để phân tích và khai phá.

Những vấn đề tồn tại:

- Cấu trúc vật lý các Website khác nhau từ những mẫu người dùng truy xuất.
- Rất khó có thể tìm ra những người dùng, các phiên làm việc, các giao tác.

Vấn đề chứng thực phiên người dùng và truy cập Web:

Các phiên chuyển hướng của người dùng: Nhóm các hành động được thực hiện bởi người dùng từ lúc họ truy cập vào Website đến lúc họ rời khỏi Website đó. Những hành động của người dùng trong một Website được ghi và lưu trữ lại trong một file đăng nhập (log file) (file đăng nhập chứa địa chỉ IP của máy khách, ngày, thời gian từ khi yêu cầu được tiếp nhận, các đối tượng yêu cầu và nhiều thông tin khác như các giao thức của yêu cầu, kích thước đối tượng,...).

3.2.2 Quá trình khai phá theo sử dụng Web

Khai phá sử dụng Web có 3 pha [6]*:* Tiền xử lý, khai phá và phân tích đánh giá, biểu diễn dữ liệu.

3.2.2.1 Tiền xử lý dữ liệu

Chứng thực người dùng, chứng thực hoạt động truy nhập, đường dẫn đầy đủ, chứng thực giao tác, tích hợp dữ liệu và biến đổi dữ liệu. Trong pha này, các thông tin về đăng nhập Web có thể được biến đổi thành các mẫu giao tác thích hợp cho việc xử lý sau này trong các lĩnh vực khác nhau. Trong giai đoạn này gồm cả việc loại bỏ các file có phần mở rộng là gif, jpg,... Bổ sung hoặc xóa bỏ các dữ liệu khuyết thiếu như cache cục bộ, dịch vụ proxy. Xử lý thông tin trong các Cookie, thông tin đăng ký người dùng kết hợp với IP, tên trình duyệt và các thông tin lưu tạm. Chứng thực giao tác: Chứng thực các phiên người dùng, các giao tác.

3.2.2.2 Khai phá dữ liệu

Sử dụng các phương pháp KPDL trong các lĩnh vực khác nhau như luật kết hợp, phân tích, thống kê, phân tích đường dẫn, phân lớp và phân cụm để khám phá ra các mẫu người dùng.

Phân tích đường dẫn: Hầu hết các các đường dẫn thường được thăm được bố trí theo đồ thị vật lý của trang Web. Mỗi nút là một trang, mỗi cạnh là đường liên kết giữa các trang đó. Thông qua việc phân tích đường dẫn trong quá trình truy cập của người dùng ta có thể biết được mối quan hệ trong việc truy cập của người giữa các đường dẫn liên quan.

Luật kết hợp: Sự tương quan giữa các tham chiếu đến các file khác nhau có trên dịch vụ nhờ việc sử dụng luật kết hợp.

Chuỗi các mẫu: Các mẫu thu được giữa các giao tác và chuỗi thời gian. Thể hiện một tập các phần tử được theo sau bởi phân tử khác trong thứ tự thời gian lưu hành tập giao tác.

Quy tắc phân loại [6]: Profile của các phần tử thuộc một nhóm riêng biệt theo các thuộc tính chung. Ví dụ như thông tin cá nhân hoặc các mẫu truy cập. Profile có thể sử dụng để phân loại các phần tử dữ liệu mới được thêm vào CSDL.

Phân tích phân cụm: Nhóm các khách hàng lại cùng nhau hoặc các phần tử dữ liệu có các đặc tính tương tự nhau.

Khai phá theo sử dụng Web giúp cho việc phát triển và thực hiện các chiến lược tiếp thị khách hàng cả về trực tuyến hoặc không trực tuyến như việc trả lời tự động cho các khách hàng thuộc nhóm chắc chắn, nó tạo ra sự thay đổi linh động một Website riêng biệt đối với mỗi khách hàng.

3.2.2.3 Phân tích đánh giá

Phân tích mô hình [6]: Thống kê, tìm kiếm tri thức và tác nhân thông minh. Phân tích tính khả thi, truy vấn dữ liệu hướng tới sự tiêu dùng của con người. Trực quan hóa: Trực quan Web sử dụng lược đồ đường dẫn Web và đưa ra đồ thị có hướng OLAP.

3.3 Khai phá cấu trúc Web

WWW là mạng lưới thông tin toàn cầu, bao gồm tất cả các Website. Các trang Web có thể được liên kết với nhau. Các siêu liên kết chứa đựng ngữ nghĩa mô tả chủ

để của trang. Một siêu liên kết dẫn tới một trang Web khác có thể được xem như là một chứng thực của trang Web đó. Do đó, nó rất có ích trong việc sử dụng những thông tin ngữ nghĩa để lấy được thông tin quan trọng thông qua phân tích liên kết giữa các trang Web.

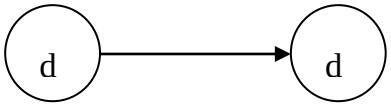
Mục tiêu của khai phá cấu trúc Web là để phát hiện thông tin cấu trúc về Web. Nếu như khai phá nội dung Web chủ yếu tập trung vào cấu trúc bên trong tài liệu thì khai phá cấu trúc Web cố gắng để phát hiện cấu trúc liên kết của các siêu liên kết ở mức trong của tài liệu. Dựa trên mô hình hình học của các siêu liên kết, khai phá cấu trúc Web sẽ phân loại các trang Web, tạo ra thông tin như độ tương tự và mối quan hệ giữa các Website khác nhau. Nếu trang Web được liên kết trực tiếp với trang Web khác thì ta sẽ muốn phát hiện ra mối quan hệ giữa các trang Web này. Chúng có thể tương tự với nhau về nội dung, có thể thuộc dịch vụ Web giống nhau do đó nó được tạo ra bởi cùng một người. Những nhiệm vụ khác của khai phá cấu trúc Web là khám phá sự phân cấp tự nhiên hoặc mạng lưới của các siêu liên kết trong các Website của một miền đặc biệt. Điều này có thể giúp tạo ra những luồng thông tin trong Website mà nó có thể đại diện cho nhiều miền đặc biệt. Vì thế việc xử lý truy vấn sẽ trở nên dễ dàng hơn và hiệu quả hơn.

3.3.1 Tiêu chuẩn đánh giá độ tương tự

Khám phá ra một nhóm các trang Web giống nhau để khai phá, ta phải chỉ ra sự giống nhau của hai nút theo một tiêu chuẩn nào đó.

Tiêu chuẩn 1(Quan hệ trực tiếp):

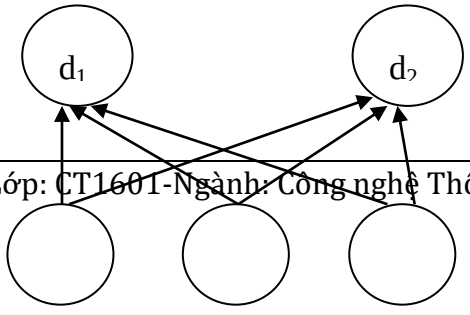
Đối với mỗi trang Web d_1 và d_2. Ta nói d_1 và d_2 quan hệ với nhau nếu có một liên kết từ d_1 đến d_2 hoặc từ d_2 đến d_1.



Hình 3-3: Quan hệ trực tiếp giữa 2 trang

Tiêu chuẩn 2(Đồng trích dẫn):

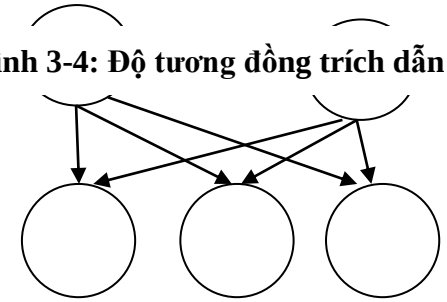
Độ tương tự giữa d_1 và d_2 được đo bởi số trang dẫn tới cả d_1 và d_2.



Tiêu chuẩn 3(Tương tự chỉ mục):

Độ tương tự giữa d_1 và d_2 được đo bằng số trang mà cả d_1 và d_2 đều trở tới.

Hình 3-4: Độ tương đồng trích dẫn



Hình 3-5: Độ tương tự chỉ mục

3.3.2 Khai phá và quản lý cộng đồng Web

Một cộng đồng Web là tập hợp các trang Web mà dữ liệu của nó liên quan đến một lĩnh vực. Nhận biết được các cộng đồng Web, hiểu được sự phát triển và những đặc trưng của các cộng đồng Web là rất quan trọng. Việc xác định và hiểu các cộng đồng trên Web có thể được xem như việc khai phá và quản lý Web.

Đặc điểm của cộng đồng Web:

- Các trang Web trong cùng một cộng đồng sẽ "tương tự" với nhau hơn các trang Web ngoài cộng đồng.
- Mỗi cộng đồng Web sẽ tạo thành một cụm các trang Web.
- Các cộng đồng Web được xác định một cách rõ ràng, tất cả mọi người đều biết, như các nguồn tài nguyên được liệt kê bởi Yahoo.
- Cộng đồng Web được xác định hoàn chỉnh: Chúng là những cộng đồng bất ngờ xuất hiện.

Khai phá cộng đồng Web ngày càng được quan tâm và được ứng dụng nhiều trong thực tiễn. Vì vậy, việc nghiên cứu các phương pháp khám phá cộng đồng là rất có ý nghĩa to lớn trong thực tiễn. Ngoài ra, việc phân tích đồ thị Web có tác dụng lớn trong việc tìm kiếm ra các cộng đồng ẩn. Có nhiều phương pháp chứng thực cộng đồng Web như: thuật toán tìm kiếm theo chủ đề HITS, luồng cực đại và nhát cắt cực tiểu, thuật toán PageRank,...

3.4 Áp dụng thuật toán trong tìm kiếm và phân cụm tài liệu Web

Hiện nay, do sự phát triển của các công cụ Web Search Engine đã giúp người dùng dễ dàng hơn trong việc tìm kiếm thông tin trên Web. Tuy nhiên, không phải lúc nào các công cụ Web Search Engine đều cho ra những kết quả dúng với nhu cầu của người dùng. Vậy nên, ta có thể nhóm các kết quả tìm được thành các nhóm theo từng chủ đề, sau đó người dùng có thể tìm kiếm những thông tin theo chủ đề mà họ cần. Điều này giúp cho người dùng thực hiện việc tìm kiếm nhanh và hiệu quả hơn. Trong đồ án này ta sẽ tìm hiểu về việc sử dụng kỹ thuật phân cụm tài liệu Web, dựa trên kho dữ liệu đã được tìm kiếm và lưu trữ.

3.4.1 Tìm hiểu kỹ thuật phân cụm tài liệu Web

Ngày nay, có rất nhiều phương pháp đánh giá độ quan trọng của một trang Web như: PageRank, HITS, … Tuy vậy, các phương pháp đánh giá này chủ yếu đều dựa vào các liên kết trang để xác định trọng số cho trang. Ta có thể tiếp cận cách đánh giá mức độ quan trọng theo một hướng khác là dựa vào nội dung của các tài liệu để xác định trọng số, nếu các tài liệu "gần nhau" về nội dung thì sẽ có mức độ quan trọng tương đương và sẽ thuộc về cùng một nhóm.

Giả sử cho tập S gồm các trang web, hãy tìm trong tập S các trang chứa nội dung câu hỏi truy vấn ta được tập R. Sử dụng thuật toán phân cụm dữ liệu để phân tập R thành k cụm (k xác định) sao cho các phần tử trong cụm là tương tự nhau nhất, các phần tử ở các cụm khác nhau thì phi tương tự với nhau. Từ tập S-R, chúng ta đưa các phần tử này vào một trong k cụm đã được thiết lập ở trên. Những phần tử nào tương tự với trọng tâm của cụm (theo một ngưỡng xác định nào đó) thì đưa vào cụm này, những phần tử không thỏa mãn xem như không phù hợp với truy vấn và loại bỏ nó khỏi tập kết quả. Kế tiếp, chúng ta đánh trọng số cho các cụm và các trang trong tập kết quả theo thuật toán sau:

INPUT: tập dữ liệu D chứa các trang gồm k cụm và k trọng tâm

OUTPUT: trọng số của các trang

BEGIN

Mỗi cụm dữ liệu thứ *m* và trọng tâm C_m ta gán một trọng số ts_m. Với các trọng tâm C_i ,C_j bất kỳ ta luôn có $ts_i > ts_j$ nếu t_i tương tự với truy vấn hơn t_j.

Với mỗi trang *p* trong cụm *m* ta xác định trọng số trang pw_m. Với mỗi pw_i, pw_j bất kỳ, ta luôn có $pw_1 > pw_2$ nếu pw_1 gần trọng tâm hơn pw_2.

END

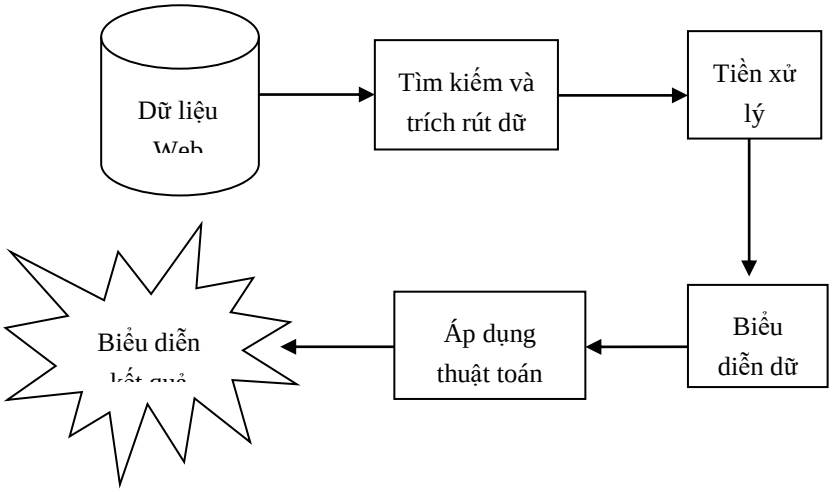
Như vậy, theo cách tiếp cận này ta sẽ giải quyết được các vấn đề sau:

- Kết quả tìm kiếm sẽ được phân thành các cụm theo các chủ đề khác nhau, tùy vào yêu cầu cụ thể người dùng sẽ xác định chủ đề mà họ cần.
- Quá trình tìm kiếm và xác định trọng số cho các trang chủ yếu tập trung vào nội dung của trang hơn là dựa vào các liên kết trang.
- Giải quyết được vấn đề từ/cụm từ đồng nghĩa trong câu truy vấn của người dùng.
- Có thể kết hợp phương pháp phân cụm trong lĩnh vực khai phá dữ liệu với các phương pháp tìm kiếm đã có.

3.4.2 Quá trình tìm kiếm và phân cụm tài liệu

Về cơ bản, quá trình phân cụm kết quả tìm kiếm sẽ diễn ra theo các bước được thể hiện như sau :

- Tìm kiếm các trang Web từ các Website thỏa mãn nội dung truy vấn.
- Trích rút thông tin mô tả từ các trang và lưu trữ nó cùng với các URL tương ứng.
- Sử dụng kỹ thuật phân cụm dữ liệu để phân cụm tự động các trang Web thành các cụm, sao cho các trang trong cụm "tương tự" về nội dung với nhau hơn các trang ngoài cụm.



Hình 3-6: Các bước phân cụm kết quả tìm kiếm trên Web

3.4.2.1 Tìm kiếm dữ liệu trên Web

Nhiệm vụ chủ yếu của giai đoạn này là dựa vào tập từ khóa tìm kiếm để tìm kiếm và trả về tập gồm toàn văn tài liệu, tiêu đề, mô tả tóm tắt, URL,… tương ứng với các trang đó. Nhằm nâng cao tốc độ xử lý, ta tiến hành tìm kiếm và lưu trữ các tài liệu này trong kho dữ liệu để sử dụng cho quá trình tìm kiếm (tương tự như các Web Search Engine Yahoo, Google,…). Mỗi phần tử gồm toàn văn tài liệu, tiêu đề, đoạn mô tả nội dung, URL,…

3.4.2.2 Tiền xử lý dữ liệu

Quá trình làm sạch dữ liệu và chuyển dịch các tài liệu thành các dạng biểu diễn dữ liệu thích hợp. Giai đoạn này bao gồm các công việc như sau: Chuẩn hóa văn bản, xóa bỏ các từ dừng, kết hợp các từ có cùng từ gốc, số hóa và biểu diễn văn bản,..

a. Chuẩn hóa văn bản

Đây là giai đoạn chuyển văn bản thô về dạng văn bản sao cho việc xử lý sau này được dễ dàng, đơn giản, thuật tiện, chính xác so với việc xử lý trực tiếp trên văn bản thô mà ảnh hưởng ít đến kết quả xử lý.

Bao gồm:

- Xóa các thẻ HTML và các loại thẻ khác để trích ra các từ/cụm từ.
- Chuyển các ký tự hoa thành các ký tự thường.
- Xóa bỏ các dấu câu, xoá các ký tự trắng dư thừa,...

b. Xóa bỏ các từ dừng

Trong văn bản có những từ mang quá ít thông tin, không có nhiều tác dụng trong quá trình xử lý, những từ có tần số xuất hiện thấp, những từ xuất hiện với tần số lớn nhưng không quan trọng cho quá trình xử lý đều được loại bỏ. Theo một số nghiên cứu gần đây cho thấy việc loại bỏ các từ dùng có thể giảm bởi được khoảng 20-30% tổng số từ trong văn bản. Có rất nhiều từ xuất hiện với tần số lớn nhưng nó không hữu ích cho quá trình phân cụm dữ liệu. Ví dụ trong tiếng Anh các từ như a, an, the, of, and, to, on, by,... trong tiếng Việt như các từ "thì", "mà", "là", "và", "hoặc",... Những từ xuất hiện với tần số quá lớn cũng sẽ được loại bỏ.

Để đơn giản trong ứng dụng thực tế, ta có thể tổ chức thành một danh sách các từ dừng, sử dụng định luật Zipf để xóa bỏ các từ có tần số xuất hiện quá thấp hoặc quá cao.

3.4.2.3 Xây dựng từ điển

Trong quá trình vector hóa văn bản thì, xây dựng từ điển là quá trình rất quan trọng. Từ điển gồm một bảng các từ và chỉ số của nó sau khi được sắp xếp theo thứ tự.

3.4.2.4 Tách từ, số hóa văn bản và biểu diễn tài liệu

Tách từ là quá trình tìm kiếm các từ và thay thế nó bởi chỉ số của từ đó trong từ điển. Một số mô hình tính trong số của từ là: TF, IDF, TF-IDF,… Ở đây ta có thể sử dụng một trong các mô hình toán học TF-IDF, để biểu diễn văn bản.

Chúng ta sử dụng mảng *W* (trọng số) hai chiều có kích thước *m* x *n*, với *n* là số các tài liệu, *m* là số các từ trong từ điển (số chiều), hàng thứ *j* là một vector biểu diễn tài liệu thứ *j* trong cơ sở dữ liệu, cột thứ *i* là thuật ngữ thứ *i* trong từ điển. W_{ij} là giá trị trọng số của từ *i* đối với tài liệu *j*.

Giai đoạn này thực hiện thống kê tần số từ t_i xuất hiện trong tài liệu d_j và số các tài liệu chứa t_i. Từ đó xây dựng bảng trọng số của ma trận *W* theo công thức sau:

$$Wij = \begin{cases} tf_{ij} \cdot idf_{ij} = [1 + \log(tf_{ij})] \cdot \log_{10}\left(\frac{n}{h_i}\right) & (1) \\ 0 & (2) \end{cases}$$

(1): Nếu $t_i \in d_j$

(2): Ngược lại

Trong đó:

Tf_{ij} : là tần số xuất hiện của từ t_i trong tài liệu d_j

idf_{ij} : là nghịch đảo tần số xuất hiện của từ t_i trong tài liệu d_j

h_i : là số các tài liệu mà từ t_i xuất hiện

n: là tổng số tài liệu

3.4.2.5 Phân cụm tài liệu

Sau khi đã tìm kiếm, trích rút dữ liệu và tiền xử lý và biểu diễn văn bản chúng ta sử dụng kỹ thuật phân cụm để phân cụm tài liệu.

INPUT: Tập gồm n tài liệu và k cụm.

OUTPUT: Các cụm C_i *(i=1,..,k)* sao cho hàm tiêu chuẩn đạt giá trị cực tiểu.

BEGIN

Bước 1: Khởi tạo ngẫu nhiên k vector làm đối tượng trọng tâm của k cụm.

Bước 2: Với mỗi tài liệu d_j xác định độ tương tự của nó đối với trọng tâm của mỗi cụm theo một trong các độ đo tương tự thường dùng (Euclidean, Manhattan). Xác định trọng tâm tương tự nhất cho mỗi tài liệu và đưa tài liệu vào cụm đó.

Bước 3: Cập nhận lại các đối tượng trọng tâm. Đối với mỗi cụm ta xác định lại trọng tâm bằng cách xác định trung bình cộng của các vector tài liệu trong cụm đó.

Bước 4: Lặp lại bước 2 và 3 cho đến khi trong tâm không thay đổi.

END.

Để xác định trọng tâm của cụm tài liệu: Xét một cụm văn bản c, trong đó trọng tâm C của cụm c được tính nhờ vào vector tổng D $(D = \sum_{d \in c} d)$ của các văn bản trong cụm c: $C=\frac{D}{|c|}$

|c|: là số phần tử thuộc tập tài liệu c.

Trong kỹ thuật phân cụm, trọng tâm của các cụm được sử dụng để làm đại diện cho các cụm tài liệu.

Vấn đề tính toán độ tương tự giữa 2 cụm tài liệu: Giả sử ta có 2 cụm c_1, c_2, khi đó độ tương tự giữa 2 cụm tài liệu được tính bằng mức độ "gần nhau" giữa 2 vector trọng tâm C_1, C_2: Sim(c_1,c_2)= sim(C_1,C_2). Ở đây, ta hiểu rằng c_1 và c_2 cũng có thể chỉ gồm một tài liệu vì khi đó có thể coi một cụm chỉ gồm 1 phần tử.

Trong thuật toán k-means, chất lượng phân cụm được đánh giá thông quan hàm tiêu chuẩn: $E = \sum_{i=1}^{k} \sum_{x \in C_i} D^2(x - m_i)$, trong đó x là các vector biểu diễn tài liệu, m_i là các trọng tâm của các cụm, k là số cụm, Ci là cụm thứ i.

- Độ phức tạp của thuật toán k-means là $O((n \cdot k \cdot d) \cdot r)$. Trong đó, n là số đối tượng dữ liệu, k là số cụm dữ liệu, d là số chiều, r là số vòng lặp.

3.5 Thực nghiệm

Sử dụng phần mềm *Orange data mining* để thực hiện phân cụm dữ liệu.

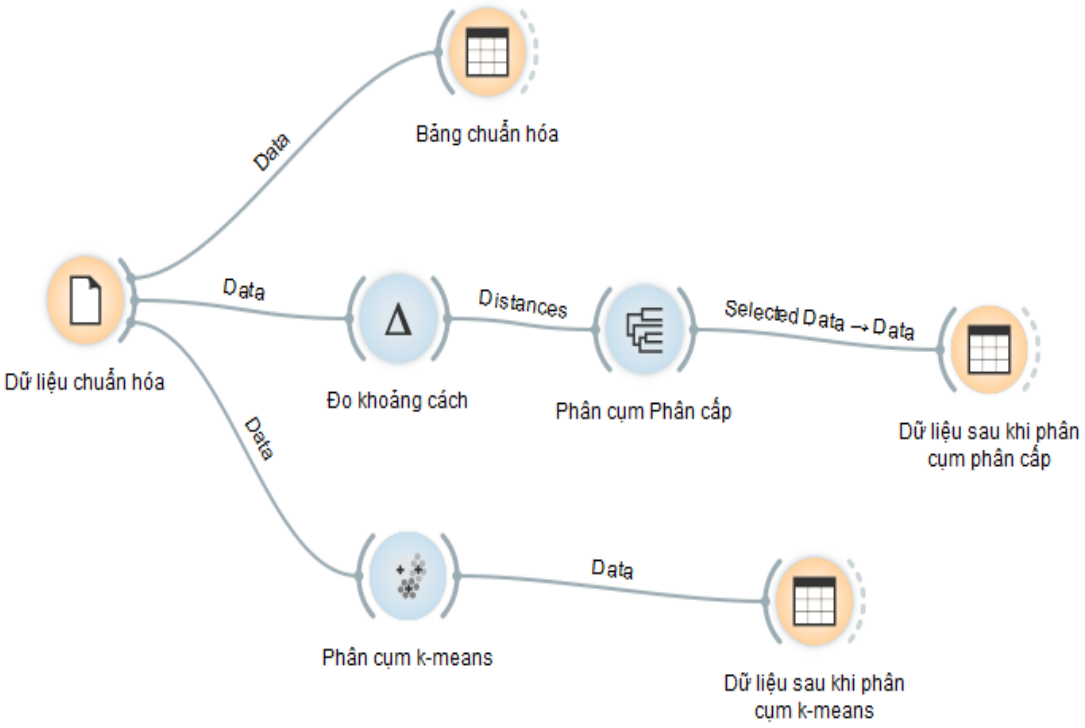
Input:

- Dữ liệu ban đầu gồm 100 file text có tên file từ file001-file100. Sau đó. chuẩn hóa dữ liệu ban đầu theo các cụm: Kinh tế, Chính trị, Khoa học, Công nghệ, Giáo dục, giải trí, Y tế (sử dụng file cvs trong excel để lưu trữ).

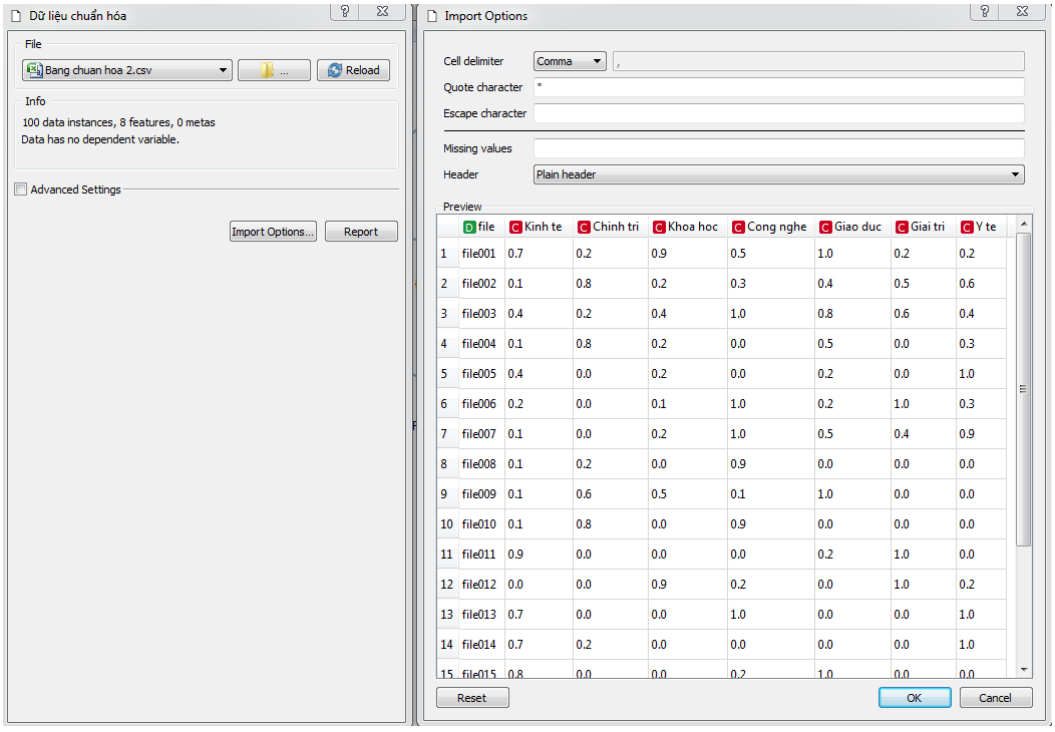
Output:

- Dữ liệu đã được phân cụm bằng *Orange*.

Quá trình phân cụm dữ liệu bằng *Orange*:

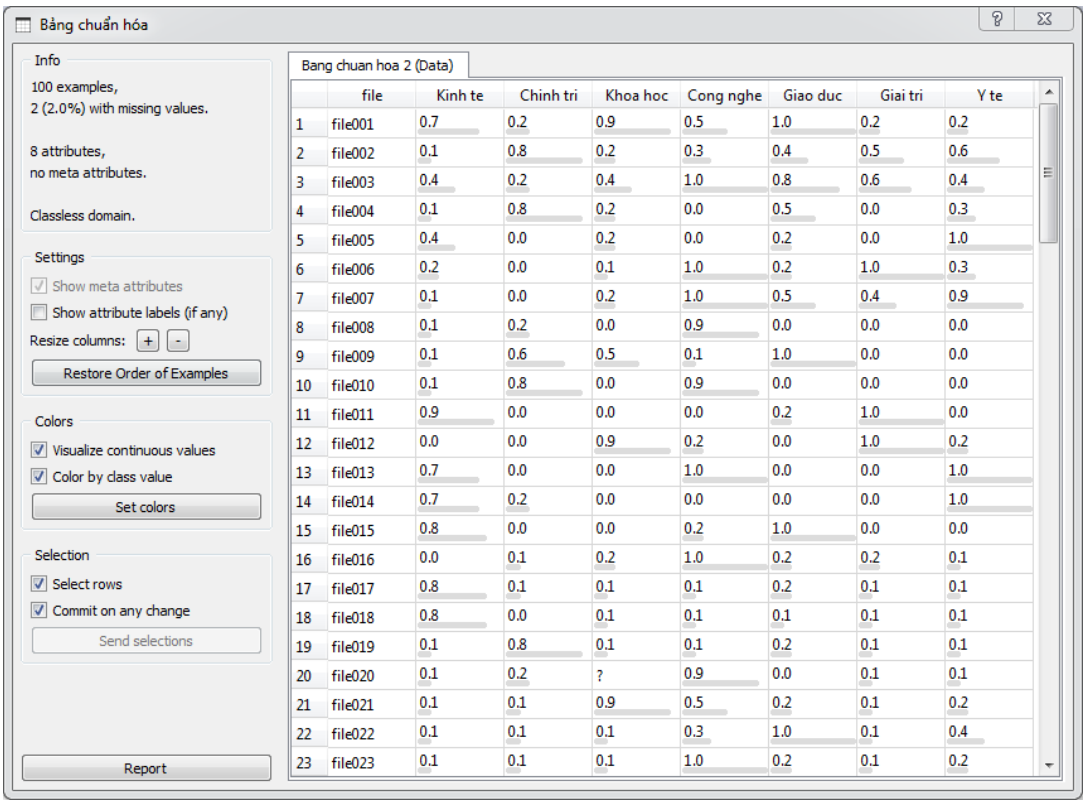


Hình 3-7: Mô hình phân cụm dữ liệu trên Orange



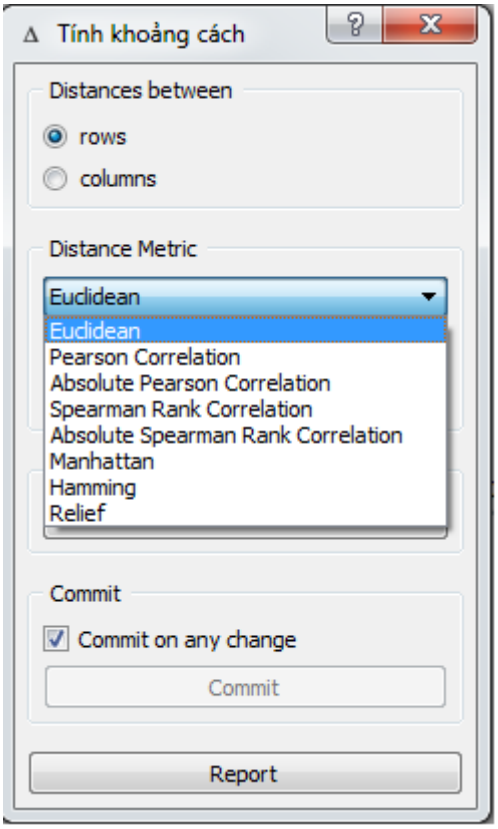
	file	Kinh te	Chinh tri	Khoa hoc	Cong nghe	Giao duc	Giai tri	Y te
1	file001	0.7	0.2	0.9	0.5	1.0	0.2	0.2
2	file002	0.1	0.8	0.2	0.3	0.4	0.5	0.6
3	file003	0.4	0.2	0.4	1.0	0.8	0.6	0.4
4	file004	0.1	0.8	0.2	0.0	0.5	0.0	0.3
5	file005	0.4	0.0	0.2	0.0	0.2	0.0	1.0
6	file006	0.2	0.0	0.1	1.0	0.2	1.0	0.3
7	file007	0.1	0.0	0.2	1.0	0.5	0.4	0.9
8	file008	0.1	0.2	0.0	0.9	0.0	0.0	0.0
9	file009	0.1	0.6	0.5	0.1	1.0	0.0	0.0
10	file010	0.1	0.8	0.0	0.9	0.0	0.0	0.0
11	file011	0.9	0.0	0.0	0.0	0.2	1.0	0.0
12	file012	0.0	0.0	0.9	0.2	0.0	1.0	0.2
13	file013	0.7	0.0	0.0	1.0	0.0	0.0	1.0
14	file014	0.7	0.2	0.0	0.0	0.0	0.0	1.0
15	file015	0.8	0.0	0.0	0.2	1.0	0.0	0.0

Hình 3-8: Đưa dữ liệu chuẩn hóa và mô hình

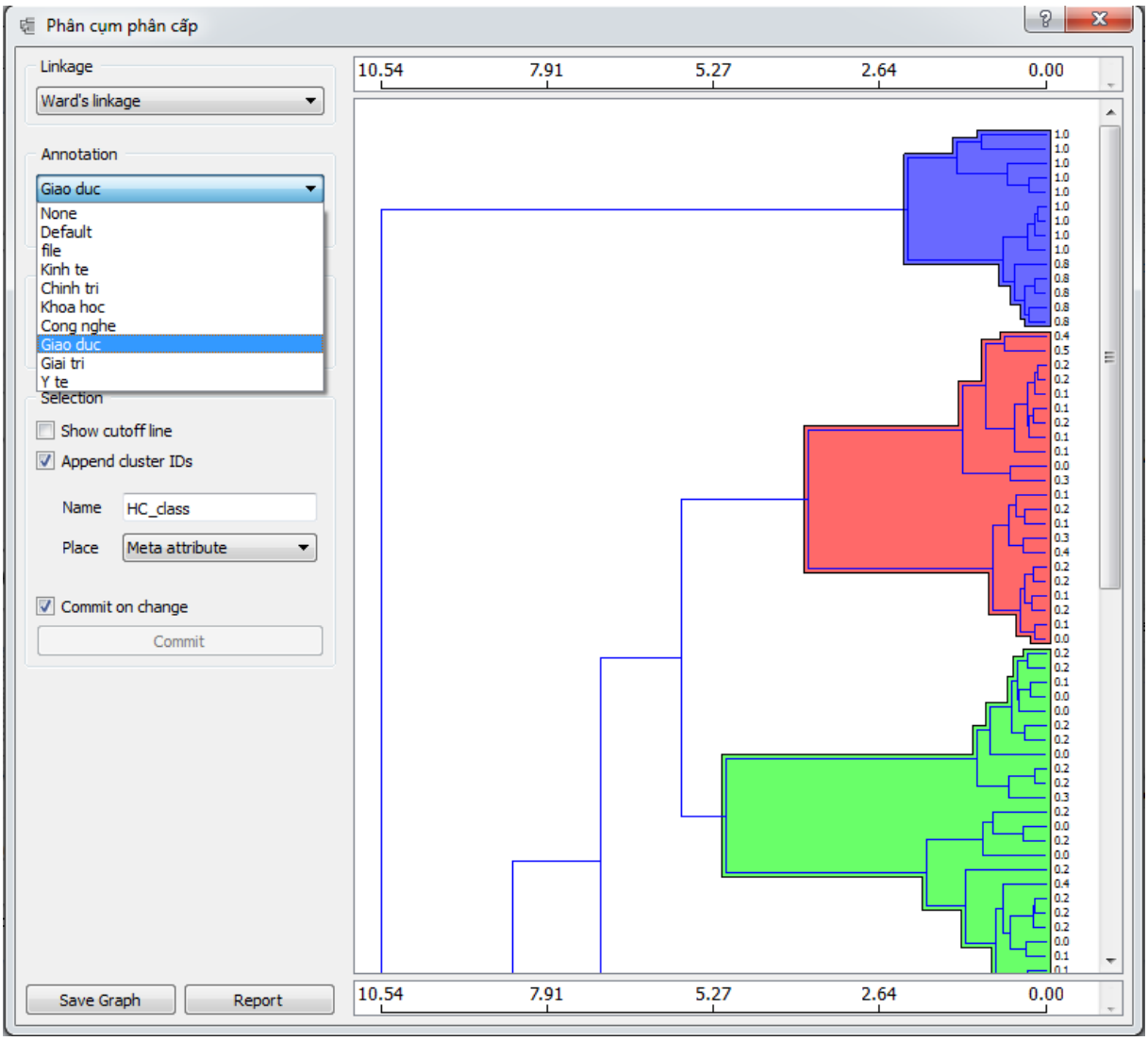


	file	Kinh te	Chinh tri	Khoa hoc	Cong nghe	Giao duc	Giai tri	Y te
1	file001	0.7	0.2	0.9	0.5	1.0	0.2	0.2
2	file002	0.1	0.8	0.2	0.3	0.4	0.5	0.6
3	file003	0.4	0.2	0.4	1.0	0.8	0.6	0.4
4	file004	0.1	0.8	0.2	0.0	0.5	0.0	0.3
5	file005	0.4	0.0	0.2	0.0	0.2	0.0	1.0
6	file006	0.2	0.0	0.1	1.0	0.2	1.0	0.3
7	file007	0.1	0.0	0.2	1.0	0.5	0.4	0.9
8	file008	0.1	0.2	0.0	0.9	0.0	0.0	0.0
9	file009	0.1	0.6	0.5	0.1	1.0	0.0	0.0
10	file010	0.1	0.8	0.0	0.9	0.0	0.0	0.0
11	file011	0.9	0.0	0.0	0.0	0.2	1.0	0.0
12	file012	0.0	0.0	0.9	0.2	0.0	1.0	0.2
13	file013	0.7	0.0	0.0	1.0	0.0	0.0	1.0
14	file014	0.7	0.2	0.0	0.0	0.0	0.0	1.0
15	file015	0.8	0.0	0.0	0.2	1.0	0.0	0.0
16	file016	0.0	0.1	0.2	1.0	0.2	0.2	0.1
17	file017	0.8	0.1	0.1	0.1	0.2	0.1	0.1
18	file018	0.8	0.0	0.1	0.1	0.1	0.1	0.1
19	file019	0.1	0.8	0.1	0.1	0.2	0.1	0.1
20	file020	0.1	0.2	?	0.9	0.0	0.1	0.1
21	file021	0.1	0.1	0.9	0.5	0.2	0.1	0.2
22	file022	0.1	0.1	0.1	0.3	1.0	0.1	0.4
23	file023	0.1	0.1	0.1	1.0	0.2	0.1	0.2

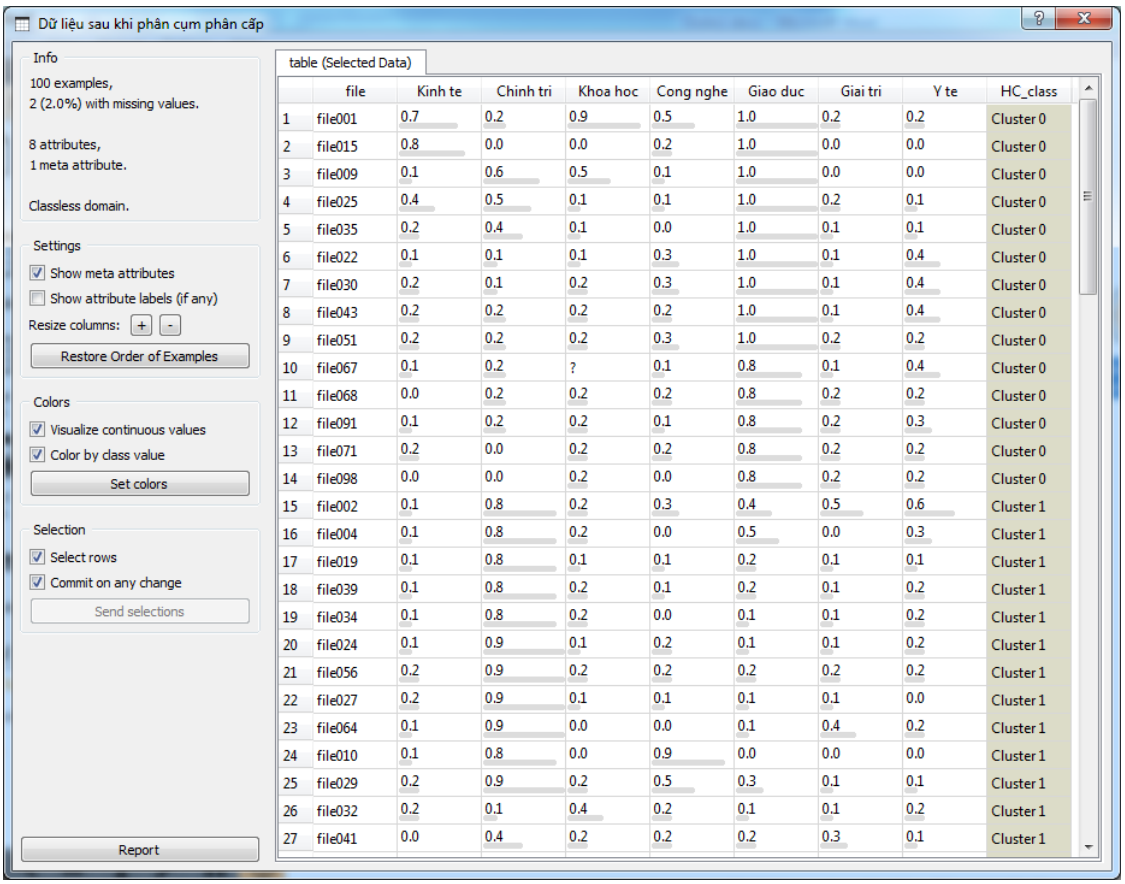
Hình 3-9: Bảng chuẩn hóa



Hình 3-10: Do khoảng cách bằng Euclidean

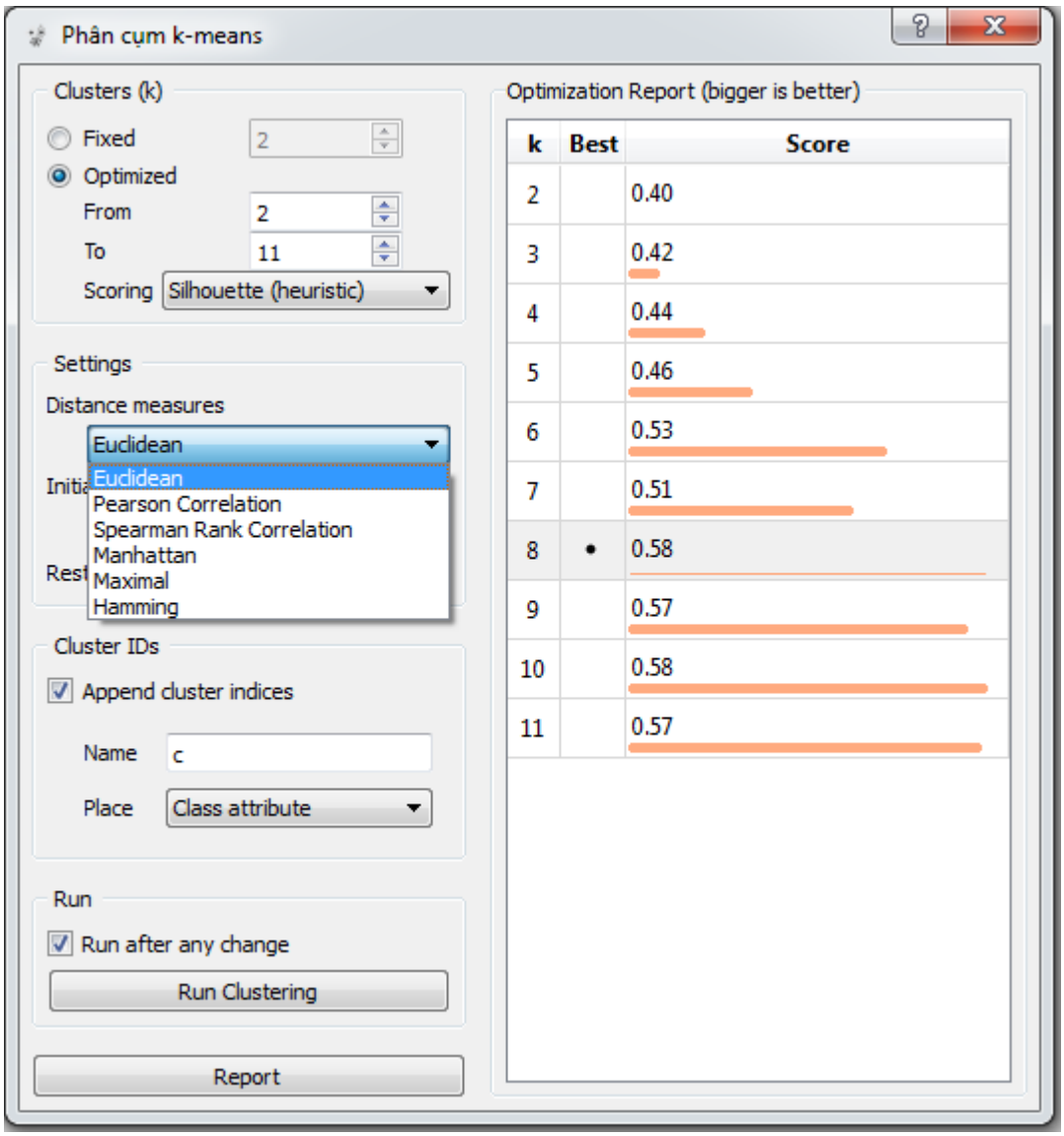


Hình 3-11: Phân cụm dữ liệu theo phương pháp phân cụm phân cấp



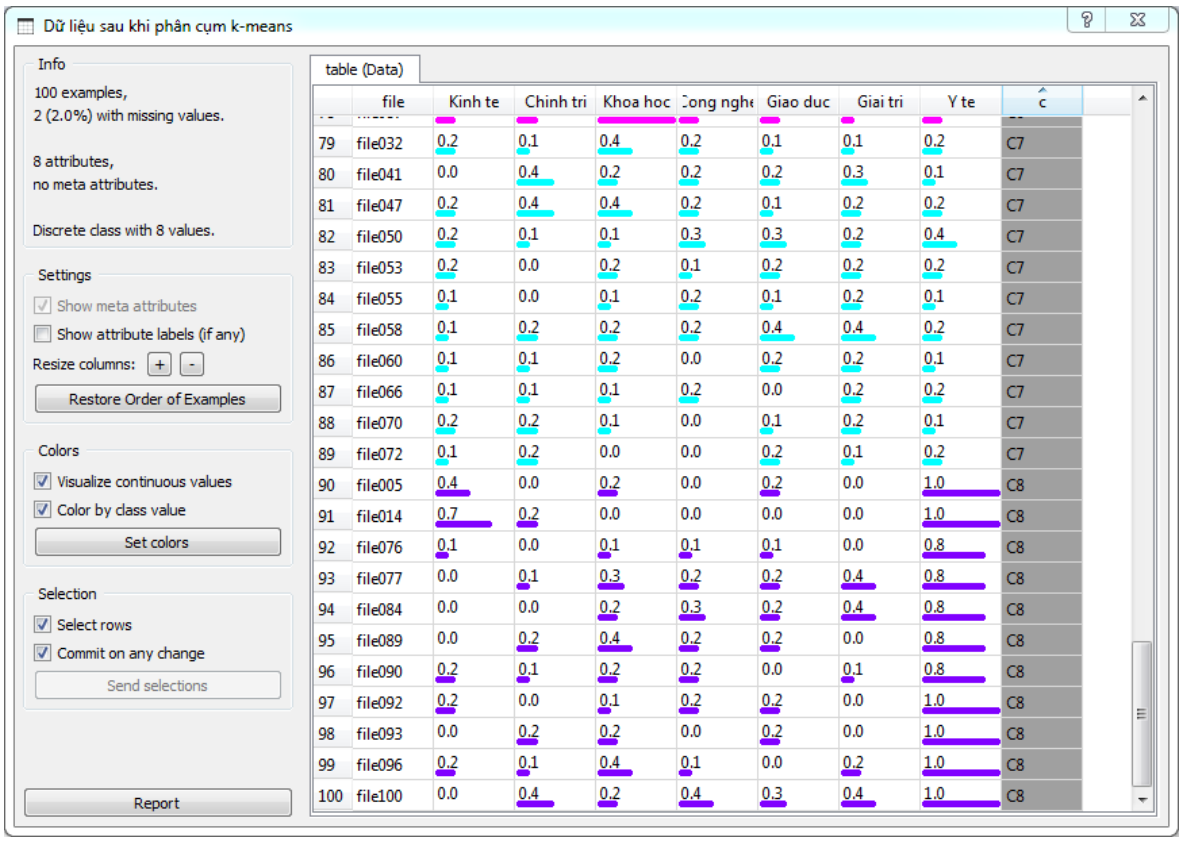
	file	Kinh te	Chinh tri	Khoa hoc	Cong nghe	Giao duc	Giai tri	Y te	HC_class
1	file001	0.7	0.2	0.9	0.5	1.0	0.2	0.2	Cluster 0
2	file015	0.8	0.0	0.0	0.2	1.0	0.0	0.0	Cluster 0
3	file009	0.1	0.6	0.5	0.1	1.0	0.0	0.0	Cluster 0
4	file025	0.4	0.5	0.1	0.1	1.0	0.2	0.1	Cluster 0
5	file035	0.2	0.4	0.1	0.0	1.0	0.1	0.1	Cluster 0
6	file022	0.1	0.1	0.1	0.3	1.0	0.1	0.4	Cluster 0
7	file030	0.2	0.1	0.2	0.3	1.0	0.1	0.4	Cluster 0
8	file043	0.2	0.2	0.2	0.2	1.0	0.1	0.4	Cluster 0
9	file051	0.2	0.2	0.2	0.3	1.0	0.2	0.2	Cluster 0
10	file067	0.1	0.2	?	0.1	0.8	0.1	0.4	Cluster 0
11	file068	0.0	0.2	0.2	0.2	0.8	0.2	0.2	Cluster 0
12	file091	0.1	0.2	0.2	0.1	0.8	0.2	0.3	Cluster 0
13	file071	0.2	0.0	0.2	0.2	0.8	0.2	0.2	Cluster 0
14	file098	0.0	0.0	0.2	0.0	0.8	0.2	0.2	Cluster 0
15	file002	0.1	0.8	0.2	0.3	0.4	0.5	0.6	Cluster 1
16	file004	0.1	0.8	0.2	0.0	0.5	0.0	0.3	Cluster 1
17	file019	0.1	0.8	0.1	0.1	0.2	0.1	0.1	Cluster 1
18	file039	0.1	0.8	0.2	0.1	0.2	0.1	0.2	Cluster 1
19	file034	0.1	0.8	0.2	0.0	0.1	0.1	0.2	Cluster 1
20	file024	0.1	0.9	0.1	0.2	0.1	0.1	0.2	Cluster 1
21	file056	0.2	0.9	0.2	0.2	0.2	0.2	0.2	Cluster 1
22	file027	0.2	0.9	0.1	0.1	0.1	0.1	0.0	Cluster 1
23	file064	0.1	0.9	0.0	0.0	0.1	0.4	0.2	Cluster 1
24	file010	0.1	0.8	0.0	0.9	0.0	0.0	0.0	Cluster 1
25	file029	0.2	0.9	0.2	0.5	0.3	0.1	0.1	Cluster 1
26	file032	0.2	0.1	0.4	0.2	0.1	0.1	0.2	Cluster 1
27	file041	0.0	0.4	0.2	0.2	0.2	0.3	0.1	Cluster 1

Hình 3-12: Dữ liệu sau khi phân cụm phân cấp



Hình 3-13: Phân cụm bằng k-means

- Đo khoảng cách bằng Euclidean, cho thấy 8 cụm thì thuật toán là tối ưu nhất.



Hình 3-14: Biểu diễn dữ liệu sau khi phân cụm k-means

3.6 Tiểu kết chương 3

Chương này tác giả đã trình bày một số hướng tiếp cận trong khai phá Web như khai phá dữ liệu toàn văn của tài liệu Web, khai phá cấu trúc Web, khai phá sử dụng Web và một số thuật toán đang được áp dụng trong khai phá Web.

Phần này trình bày quá trình phân cụm dữ liệu bằng phương pháp phân cụm phân cấp và phân cụm theo thuật toán k-means, bằng phần mềm mô hình hóa *Orange Data Mining*.

Kết luận

Tìm hiểu nội dung đề tài giúp em có thêm kiến thức về khai phá dữ liệu Web và phương pháp áp dụng các kỹ thuật phân cụm trong khai phá dữ liệu Web.

Đồ án này tập chung vào việc tìm hiểu về: Khai phá tri thức, phân cụm dữ liệu, khai phá Web và áp dụng các kỹ thuật phân cụm trong khai phá dữ liệu Web. Ngoài ra, ta cần xây dựng một trương trình thực nghiệm phục vụ cho việc tìm kiếm,lưu trữ để phân cụm tài liệu Web để phục vụ cho quá trình tìm kiếm.

Nhưng trong khuôn khổ đồ án tốt nghiệp, em chưa kịp xây dựng trương trình, mà thay vào đó em lập một mô hình phân cụm dữ liệu Web bằng phần mền Orange. Trong đó, công việc tìm kiếm, lưu trữ và chuẩn hóa dữ liệu sẽ được làm thủ công và quá trình phân cụm dữ liệu sẽ được giải quyết bằng phần mền Orange.

Hướng phát triển: xây dựng hoàn thiện phần mềm đầy đủ chức năng: tìm kiếm, lư trữ, phân cụm dữ liệu phục vụ cho việc tìm kiếm.

Tài liệu tham khảo

Tài liệu tiếng việt

[1] Hoàng Văn Dũng, "Khai phá dữ liệu Web bằng kỹ thuật phân cụm", Đồ án thạc sĩ, Trường Đại học Sư phạm Hà Nội,2007.

[2] Cao Chính Nghĩa, "*Một số vấn đề về phân cụm dữ liệu*", Luận văn thạc sĩ, Trường Đại học Công nghệ, ĐH Quốc gia Hà Nội, 2006.

[3] Hoàng Hải Xanh, "*Về các kỹ thuật phân cụm dữ liệu trong data mining*", luận văn thạc sĩ, Trường ĐH Quốc Gia Hà Nội, 2005.

Tài liệu tiếng anh

[4] Bing Liu, *Web mining*, Springer, 2007.

[5] Ho Tu Bao,*Knowledge Discovery and Data Mining*, 2000.

[6] Khoo Khyou Bun, "*Topic Trend Detection and Mining in World Wide Web*", A thesis for the degree of PhD, Japan, 2004.