CHƯƠNG 1

MỘT SỐ NỘI DUNG CƠ BẢN VỀ KHAI PHÁ DỮ LIỆU

1.1 KHAI PHÁ DỮ LIỆU VÀ PHÁT HIỆN TRI THỨC TRONG CƠ SỞ DỮ LIỆU

Ước tính cứ khoảng 20 tháng lượng thông tin trên thế giới lại tăng gấp đôi. Do đó, lượng dữ liệu mà con người thu thập và lưu trữ được trong các kho dữ liệu hiện nay là rất lớn, nhiều khi vượt quá khả năng quản lý. Thời gian này, người ta bắt đầu đề cập đến khái niệm khủng hoảng phân tích dữ liệu tác nghiệp nhằm cung cấp thông tin với yêu cầu chất lượng cho những người ra quyết định trong các tổ chức tài chính, thương mại, khoa học, ... Đúng như John Naisbett đã cảnh báo "*Chúng ta đang chìm ngập trong dữ liệu mà vẫn đói tri thức*".

Với lượng dữ liệu tăng nhanh và khổng lồ như vậy, rõ ràng các phương pháp phân tích dữ liệu thủ công truyền thống sẽ còn không hiệu quả, gây tốn kém và dễ dẫn đến những kết quả sai lệch. Để có thể khai phá hiệu quả các cơ sở dữ liệu lớn cần phải có những kỹ thuật mới: kỹ thuật khai phá dữ liệu (Data Mining).

Khai phá dữ liệu là một lĩnh vực khoa học mới xuất hiện, nhằm tự động hóa khai thác những thông tin, tri thức hữu ích, tiềm ẩn trong các CSDL lớn cho các tổ chức, doanh nghiệp, ... từ đó thúc đẩy khả năng sản xuất, kinh doanh, cạnh tranh của tổ chức, doanh nghiệp này. Các kết quả nghiên cứu cùng với những ứng dụng thành công trong khám phá tri thức cho thấy khai phá dữ liệu là một lĩnh vực khoa học tiềm năng, mang lại nhiều lợi ích, đồng thời có những ưu thế hơn hẳn so với các công cụ phân tích dữ liệu truyền thống. Hiện nay, khai phá dữ liệu được ứng dụng rộng rãi trong các lĩnh vực như: Phân tích dữ liệu hỗ trợ ra quyết định, điều trị y học, tin-sinh học (Bioinformatics), thương mại, tài chính, bảo hiểm, text mining, web mining ... .

Do sự phát triển nhanh chóng về phương pháp tìm kiếm tri thức và phạm vi áp dụng, đã có nhiều quan điểm khác nhau về khai phá dữ liệu. Tuy nhiên, ở một mức độ trừu tượng nhất định, chúng ta định nghĩa khai phá dữ liệu như sau:

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

Khám phá tri thức trong CSDL (Knowledge Discovery in Databaes – KDD) là mục tiêu chính của khai phá dữ liệu, do vậy hai khái niệm khai phá dữ liệu và KDD được các nhà khoa học xem là tương đương nhau. Thế nhưng, nếu phân chia một cách chi tiết thì khai phá dữ liệu là một bước chính trong quá trình KDD.

Khám phá tri thức trong CSDL là lĩnh vực liên quan đến nhiều ngành như: Tổ chức dữ liệu, xác suất, thống kê, lý thuyết thông tin, học máy, CSDL, thuật toán, trí tuệ nhân tạo, tính toán song song và hiệu năng cao, ... Các kỹ thuật chính áp dụng trong khám phá tri thức phần lớn được thừa kế từ các ngành này.

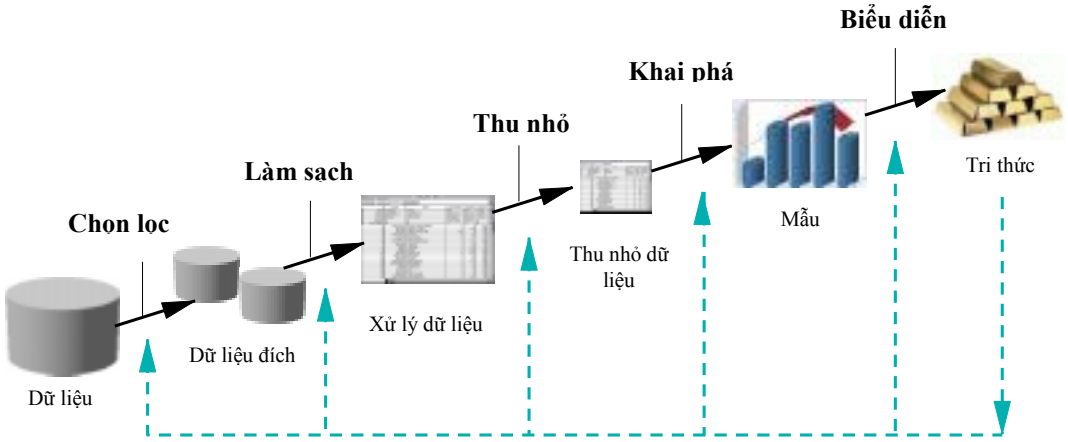
1.2 CÁC BƯỚC KHAI PHÁ TRI THỨC

Quá trình khám phá tri thức có thể phân thành các công đoạn sau:

- *Trích lọc dữ liệu*: Là bước tuyển chọn, tích hợp những tập dữ liệu cần được khai phá từ các tập dữ liệu lớn (databases, data warehouses, data repositories) ban đầu theo một số tiêu chí nhất định.
- *Tiền xử lý dữ liệu*: Là bước làm sạch dữ liệu (xử lý dữ liệu không đầy đủ, dữ liệu nhiễu, dữ liệu không nhất quán, ...), tổng hợp dữ liệu (nén, nhóm dữ liệu, tính toán các đặc trưng tổng hợp, xây dựng các histograms, lấy mẫu, ...), rời rạc hóa dữ liệu (rời rạc hóa dựa vào histograms, entropy, phân khoảng, ...). Sau bước tiền sử lý này, dữ liệu sẽ nhất quán, đầy đủ, được rút gọn và rời rạc hóa.
- *Biến đổi dữ liệu*: Là bước chuẩn hóa và làm mịn dữ liệu nhằm đưa dữ liệu về dạng thuận lợi nhất cho việc áp dụng các kỹ thuật khai phá ở bước sau.
- *Khai phá dữ liệu*: Là bước áp dụng những kỹ thuật phân tích (phần nhiều là các kỹ thuật học máy) nhằm khai thác dữ liệu, trích lọc những mẫu thông tin (information patterns), những mối quan hệ đặc biệt trong dữ liệu. Đây được xem là bước quan trọng và tiêu tốn thời gian nhất của toàn bộ quá trình KDD.
- *Đánh giá và biểu diễn tri thức*: Những mẫu thông tin và mối quan hệ trong dữ liệu đã được phát hiện ở bước khai phá dữ liệu được chuyển sang và biểu diễn ở

dạng gần gũi với người sử dụng như đồ thị, cây, bảng biểu, luật, ... Đồng thời bước này cũng thực hiện việc đánh giá những tri thức khai phá được theo những tiêu chí nhất định.

Hình 1.1 dưới đây mô tả các công đoạn của khai phá dữ liệu:



***Hình* 1.1. Các bước thực hiện quá trình khai phá dữ liệu**

Nếu theo quan điểm của học máy (Machine Learning), thì các kỹ thuật khai phá dữ liệu bao gồm:

- *Học có giám sát* (*Supervised Learning*) : Là quá trình phân lớp các đối tượng trong cơ sở dữ liệu dựa trên một tập các ví dụ huấn luyện về các thông tin về nhãn lớp đã biết.
- *Học không có giám sát* (*Unsupervised Learning*) : Là quá trình phân chia một tập các đối tượng thành các cụm (clusters) tương tự nhau mà không có các ví dụ huấn luyện, không biết trước thông tin về lớp của các đối tượng.
- *Học nửa giám sát* (*Semi-Supervised Learning*) : Là quá trình phân chia một tập các đối tượng thành các lớp dựa trên một *tập các ví dụ huấn luyện* đã biết thông tin về *nhãn lớp* .

1.3 KHAI PHÁ DỮ LIỆU VÀ CÁC LĨNH VỰC LIÊN QUAN

Như đã nói ở trên, khai phá dữ liệu là một lĩnh vực liên quan tới nhiều ngành khoa học khác như: hệ CSDL, thống kê, trực quan hoá… hơn nưa, tuỳ vào cách tiếp cận được sử dụng, khai phá dữ liệu còn có thể áp dụng một số kỹ thuật như mạng nơron, phương

pháp hệ chuyên gia, lý thuyết tập thô, tập mờ, ... So với các phương pháp này, khai phá dữ liệu có một số ưu thế rõ rệt.

- So với phương pháp học máy, khai phá dữ liệu có ưu thế hơn ở chỗ, khai phá dữ liệu có thể sử dụng đối với các CSDL chứa nhiễu, dữ liệu không đầy đủ hoặc biến đổi liên tục. Trong khi đó, phương pháp học máy chủ yếu được áp dụng đối với các CSDL đầy đủ, ít biến động và tập dữ liệu không qua lớn.
- Phương pháp hệ chuyên gia: phương pháp này khác với khai phá dữ liệu ở chỗ các ví dụ của chuyên gia thường ở mức cao hơn nhiều so với các dữ liệu trong CSDL, và chúng thường chỉ bao hàm được các trường hợp quan trọng. Hơn nữa, giá trị và tính hữu ích của các mẫu phát hiện được sẽ được xác nhận bởi các chuyên gia .
- Phương pháp thống kê là một trong những nền tảng lý thuyết của khai phá dữ liệu, nhưng khi so sánh hai phương pháp với nhau có thể thấy các phương pháp thống kê có một số điểm yếu mà chỉ khai phá dữ liệu mới khắc phục được.

Với nhưng ưu điểm trên, khai phá dữ liệu hiện đang được áp dụng một cách rộng rãi trong nhiều lĩnh vực kinh doanh và đời sống khác nhau như: marketing, tài chính, ngân hàng và bảo hiểm, khoa học, y tế, an ninh, internet… Rất nhiều tổ chức và công ty lớn trên thế giới đã áp dụng thành công kỹ thuật khai phá dữ liệu vào các hoạt động sản xuất, kinh doanh của mình và thu được những lợi ích to lớn. Các công ty phần mềm lớn trên thế giới cũng rất quan tâm và chú trọng tới việc nghiên cứu và phát triển kỹ thuật khai phá dữ liệu: Oracle tích hợp các công cụ khai phá dữ liệu vào bộ Oracle9i, IBM đã đi tiên phong trong việc phát triển các ứng dụng khai phá dữ liệu với các ứng dụng như Intelligence Miner…

Các ứng dụng này được chia thành 3 nhóm ứng dụng khác nhau : Phát hiện gian lận (fraud detection), các ứng dụng hỗ trợ tiếp thị và quản lý khách hàng, cuối cùng là các ứng dụng vào phát hiện và xử lý lỗi hệ thống mạng.

Phát hiện gian lận (fraud detection):

Gian lận là một trong những vấn đề nghiêm trọng của các công ty viễn thông, nó có thể làm thất thoát hàng tỷ đồng mỗi năm. Có thể chia ra làm 2 hình thức gian lận khác nhau thường xảy ra đối với các công ty viễn thông : Trường hợp thứ nhất xảy ra khi một khách hàng đăng ký thuê bao với ý định không bao giờ thanh toán khoản chi phí sử dụng dịch vụ. Trường hợp thứ hai liên quan đến một thuê bao hợp lệ nhưng lại có một số hoạt động bất hợp pháp gây ra bởi một người khác. Những ứng dụng này sẽ thực hiện theo thời gian thực bằng cách sử dụng dữ liệu chi tiết cuộc gọi, một khi xuất hiện một cuộc gọi nghi ngờ gian lận, lập tức hệ thống sẽ có hành động ứng xử phù hợp, ví dụ như một cảnh báo xuất hiện hoặc từ chối cuộc gọi nếu biết đó là cuộc gọi gian lận.

Hầu hết các phương thức nhận diện gian lận đều dựa vào việc so sánh hành vi sử dụng điện thoại của khách hàng trước kia với hành vi hiện tại để xác định xem đó là cuộc gọi hợp lệ không.

Các ứng dụng quản lý và chăm sóc khách hàng

Các công ty viễn thông quản lý một khối lượng lớn dữ liệu về thông tin khách hàng và chi tiết cuộc gọi (call detail records). Những thông tin này có thể cho ta nhận diện được những đặc tính của khách hàng và thông qua đó có thể đưa ra các chính sách chăm sóc khách hàng thích hợp dựa trên dự đoán hoặc có những chiến lược tiếp thị hiệu quả.

Một trong các ứng dụng phổ biến của khai phá dữ liệu là phát hiện luật kết hợp giữa các dịch vụ viễn thông khách hàng sử dụng. Hiện nay trên một đường điện thoại khách hàng có thể sử dụng rất nhiều dịch vụ khác nhau, ví dụ như : gọi điện thoại, truy cập internet, tra cứu thông tin từ hộp thư tự động, nhắn tin, gọi 108, .v.v. Dựa trên cơ sở dữ liệu khách hàng, chúng ta có thể khám phá các liên kết trong việc sử dụng các dịch vụ, có thể đưa ra các luật như (khách hàng gọi điện thoai quốc tế) => (truy cập internet), v.v... . Trên cơ sở phân tích được các luật như vậy, các công ty viễn thông có thể điều chỉnh việc bố trí nơi đăng ký các dịch vụ phù hợp, ví dụ điểm đăng ký điện thoại quốc tế nên bố trí gần với điểm đăng ký Internet chẳng hạn.

Một ứng dụng khác phục vụ chiến lược marketing đó là sử dụng kỹ thuật khai phá luật kết hợp của khai phá dữ liệu để tìm ra tập các thành phố, tỉnh nào trong nước thường gọi điện thoại với nhau. Ví dụ, ta có thể tìm ra tập phổ biến (Cần Thơ, HCM, Hà Nội) chẳng hạn. Điều này thật sự hữu dụng trong việc hoạch định chiến lược tiếp thị hoặc xây dựng các vùng cước phù hợp.

Một vấn đề khá phổ biến ở các công ty viễn thông hiện nay là sự thay đổi nhà cung cấp dịch vụ (customer churn), đặc biệt đối với các công ty điện thoại di động. Đây là vấn đề khá nghiêm trọng ảnh hưởng đến tốc độ phát triển thuê bao, cũng như doanh thu của các nhà cung cấp dịch vụ. Thời gian gần đây các nhà cung cấp dịch vụ di động luôn có chính sách khuyến mãi lớn để lôi kéo khách hàng. Điều đó dẫn đến một lượng không nhỏ khách hàng thường xuyên thay đổi nhà cung cấp để hưởng những chính sách khuyến mãi đó. Các kỹ thuật khai phá dữ liệu hiện nay có thể dựa trên dữ liệu tiền sử để tìm ra các quy luật, từ đó có thể tiên đoán trước được khách hàng nào có ý định rời khỏi mạng trước khi họ thực hiện. Sử dụng các kỹ thuật khai phá dữ liệu như xây dựng cây quyết định (decision tree), mạng nơ ron nhân tạo (neural network) trên dữ liệu cước (billing data), dữ liệu chi tiết cuộc gọi (call detail data), dữ liệu khách hàng (customer data) tìm ra các quy luật, nhờ đó ta có thể tiên đoán trước ý định rời khỏi mạng của khách hàng, từ đó công ty viễn thông sẽ có các ứng xử phù hợp nhằm lôi kéo khách hàng.

Cuối cùng, một ứng dụng cũng rất phổ biến đó là phân lớp khách hàng (classifying). Dựa vào kỹ thuật học trên cây quyết định (decision tree learning) xây dựng được từ dữ liệu khách hàng và chi tiết cuộc gọi có thể tìm ra các luật để phân loại khách hàng. Ví dụ ta có thể phân biệt được khách hàng nào thuộc đối tượng kinh doanh hay nhà riêng dựa vào các luật sau:

Luật 1 : Nếu không quá 43% cuộc gọi có thời gian từ 0 đến 10 giây và không đến 13% cuộc gọi vào cuối tuần thì đó là khách hàng kinh doanh.

Luật 2 : Nếu trong 2 tháng có các cuộc gọi đến hầu hết từ 3 mã vùng giống nhau và dưới 56,6% cuộc gọi từ 0-10 giây thì có là khách hàng nhà riêng.

Trên cơ sở tìm được các luật tương tự như vậy, ta dễ dàng phân loại khách hàng, từ đó có chính sách phân khúc thị trường hợp lý.

Các ứng dụng phát hiện và cô lập lỗi trên hệ thống mạng viễn thông (Network fault isolation)

Mạng viễn thông là một cấu trúc cực kỳ phức tạp với nhiều hệ thống phần cứng và phần mềm khác nhau. Phần lớn các thiết bị trên mạng có khả năng tự chuẩn đoán và cho ra thông điệp trạng thái, cảnh báo lỗi (status and alarm message). Với mục tiêu là quản lý hiệu quả và duy trì độ tin cậy của hệ thống mạng, các thông tin cảnh báo phải được phân tích tự động và nhận diện lỗi trước khi nó xuất hiện làm giảm hiệu năng của mạng. Bởi vì số lượng lớn các cảnh báo độc lập và có vẻ như không quan hệ gì với nhau nên vấn đề nhận diện lỗi không ít khó khăn. Kỹ thuật khai phá dữ liệu có vai trò sinh ra các luật giúp hệ thống có thể phát hiện lỗi sớm hơn khi nó xảy ra. Kỹ thuật khám phá mẫu tuần tự (sequential/temporal patterns) của data mining thường được ứng dụng trong lĩnh vực này thông qua việc khai thác cơ sở dữ liệu trạng thái mạng (network status data).

1.4 CÁC KỸ THUẬT ÁP DỤNG TRONG KHÁI PHÁ DỮ LIỆU

Các kỹ thuật khai phá dữ liệu thường được chia làm 2 nhóm chính:

Kỹ thuật mô tả: Bao gồm các kỹ thuật mô tả các tính chất hoặc các đặc tính chung của dữ liệu trong CSDL hiện có. Các kỹ thuật này gồm có: phân cụm (clustering), tóm tắt (summerization), trực quan hóa (visualiztion), phân tích tiến hoá và độ lệch (Evolution and deviation analysis), phân tích luật kết hợp (association rules analysis)… .

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

Với hai đích chính của khai phá dữ liệu là Dự đoán (Prediction) và Mô tả (Description), người ta thường sử dụng các kỹ thuật sau cho khai phá dữ liệu:

- *Phân lớp và dự đoán* (*classification and prediction*) : Là việc xếp các đối tượng vào những lớp đã biết trước. Ví dụ, phân lớp các bệnh nhân, phân lớp các loài thực vật, ... Hướng tiếp cận này thường sử dụng một số kỹ thuật của

học máy như cây quyết định (decision tree), mạng nơ-ron nhân tạo (neural network), ... Phân lớp và dự đoán còn được gọi là học có giám sát.

- *Phân cụm* (*clustering*/*segmentation*) : Là việc xếp các đối tượng theo từng cụm tự nhiên.
- *Luật kết hợp* (*association rules*) : Là việc phát hiện các luật biểu diễn tri thức dưới dạng đơn giản. Ví dụ: "70% nữ giới vào siêu thị mua phấn thì có tới 80% trong số họ cũng mua thêm son".
- *Phân tích hồi quy* (*regression analysis*) : Là việc học một hàm ánh xạ mỗi bộ tập dữ liệu thành một giá trị thực của biến dự đoán. Nhiệm vụ của phân tích hồi quy tương tự như của phân lớp, điểm khác nhau là ở chỗ thuộc tính dự báo là liên tục chứ không phải rời rạc.
- *Phân tích các mẫu theo thời gian* (*sequential/temporal patterns*) : Tương tự như khai phá luật kết hợp nhưng có quan tâm đến tính thứ tự theo thời gian.
- *Mô tả khái niệm* (*concept description and summarization*) : Thiên về mô tả, tổng hợp và tóm tắt các khái niệm. Ví dụ tóm tắt văn bản.

Hiện nay, các kỹ thuật khai phá dữ liệu có thể làm việc với rất nhiều kiểu dữ liệu khác nhau. Một số dạng dữ liệu điển hình là: CSDL quan hệ, CSDL đa chiều (Multidimensional Data Structures), CSDL giao tác, CSDL quan hệ hướng đối tượng, dữ liệu không gian và thời gian, CSDL đa phương tiện, dữ liệu văn bản và web, ...

1.5. CÁC LOẠI DỮ LIỆU CÓ THỂ KHAI PHÁ

Về cơ bản, khai phá dữ liệu có thể ứng dụng cho bất kỳ kho thông tin nào bao gồm:

+ Các cơ sở dữ liệu quan hệ.
+ Kho dữ liệu.
+ Các cơ sở dữ liệu giao tác
+ Các hệ thống cơ sở dữ liệu tiên tiến
+ Các tệp
+ ……

1.6. MỘT SỐ THÁCH THỨC ĐẶT RA CHO VIỆC KHAI PHÁ DỮ LIỆU

- ✓ Số đối tượng trong cơ sở dữ liệu thường rất lớn
- ✓ Số chiều (thuộc tính) của cơ sở dữ liệu lớn
- ✓ Dữ liệu và tri thức luôn thay đổi có thể làm cho các mẫu đã phát hiện không còn phù hợp.
- ✓ Dữ liệu bị thiếu hoặc nhiễu
- ✓ Quan hệ giữa các thuộc tính phức tạp
- ✓ Giao tiếp với người sử dụng và kết hợp với các tri thức đã có.
- ✓ Tích hợp với các hệ thống khác…

CHƯƠNG 2

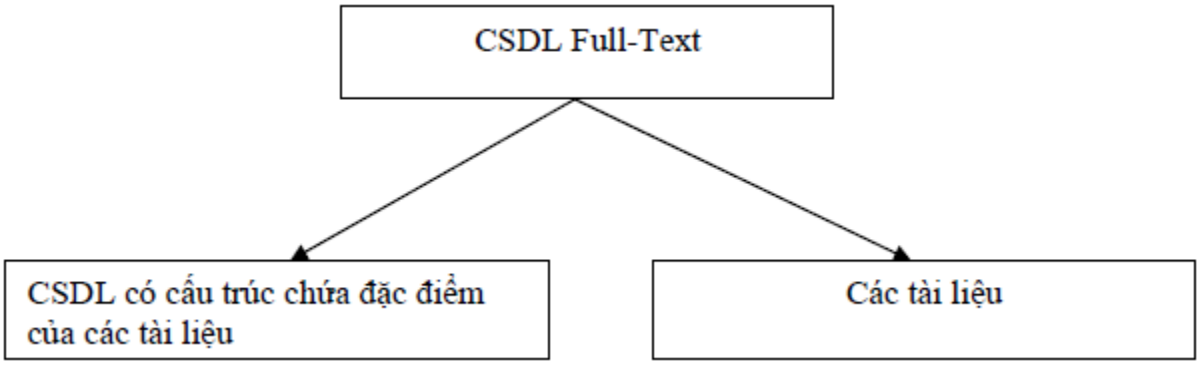
DỮ LIỆU VĂN BẢN VÀ DỮ LIỆU WEB

2.1 CƠ SỞ DỮ LIỆU FULLTEXT

Dữ liệu dạng FullText là một dạng dữ liệu phi cấu trúc với thông tin chỉ gồm các tại liệu dạng Text. Mỗi tài liệu chứa thông tin về một vấn đề nào đó thể hiện qua nội dung của tất cả các từ cấu thành tài liệu đó. Ý nghĩa của mỗi từ trong tài liệu không cố định mà tuỳ thuộc vào từng ngữ cảnh khác nhau sẽ mang ý nghĩa khác nhau. Các từ trong tài liệu được liên kết với nhau theo một ngôn ngữ nào đó.

Trong các dữ liệu hiện nay thì văn bản là một trong những dữ liệu phổ biến nhất, nó có mặt ở khắp mọi nơi và chúng ta thường xuyên bắt gặp do đó các bài toán về xử lý văn bản đã được đặt ra khá lâu và hiện nay vẫn là một trong những vấn đề trong khai phá dữ liệu Text, trong đó có những bài toán đáng chú ý như tìm kiếm văn bản, phân loại văn bản, phân cụm văn bản hoặc trích dẫn văn bản.

CSDL full_text là một dạng CSDL phi cấu trúc mà dữ liệu bao gồm các tài liệu và thuộc tính của tài liệu. Cơ sở dữ liệu Full_Text thường được tổ chức như một tổ hợp của hai thành phần: Một CSDL có cấu trúc thông thường (chứa đặc điểm của các tài liệu) và các tài liệu.



Nội dung của tài liệu được lưu trữ gián tiếp trong CSDL theo nghĩa hệ thống chỉ quản lý địa chỉ lưu trữ nội dung.

Cơ sở dữ liệu dạng Text có thể chia làm hai loại sau:

Dạng không có cấu trúc (unstructured): Những văn bản thông thường mà chúng ta thường đọc hàng ngày được thể hiện dưới dạng tự nhiên của con người và nó không có

một cấu trúc định dạng nào. Ví dụ: tập hợp sách, tạp chí, bài viết được quản lý trong một mạng thư viện điện tử.

Dạng nửa cấu trúc (semi-structured): Những văn bản được tổ chức dưới dạng cấu trúc không chặt chẽ như bản ghi các ký hiệu đánh dấu văn bản và vẫn thể hiện được nội dung chính của văn bản, ví dụ như các dạng HTML, email,...

Tuy nhiên việc phân làm hai loại cũng không thật rõ ràng, trong các hệ phần mềm, người ta thường phải sử dụng các phần kết hợp lại để thành một hệ như trong các hệ tìm tin (Search Engine), hoặc trong bài toán tìm kiếm văn bản (Text Retrieval), một trong những lĩnh vực qua tâm nhất hiện nay. Chẳng hạn trong hệ tìm kiếm như Yahoo, Altavista, Google... đều tổ chức dữ liệu theo các nhóm và thư mục, mỗi nhóm lại có thể có nhiều nhóm con nằm trong đó. Hệ Altavista còn tích hợp thêm chương trình dịch tự động có thể dịch chuyển đổi sang nhiều thứ tiếng khác nhau và cho kết quả khá tốt.

2.2. CƠ SỞ DỮ LIỆU HYPERTEXT

Theo từ điển của Đại học Oxford (Oxford English Dictionary Additions Series) thì Hypertext được định nghĩa như sau: Đó là loại Text không phải đọc theo dạng liên tục đơn, nó có thể được đọc theo các thứ tự khác nhau, đặc biệt là Text và ảnh đồ họa (Graphic) là các dạng có mối liên kết với nhau theo cách mà người đọc có thể không cần đọc một cách liên tục. Ví dụ khi đọc một cuốn sách người đọc không phải đọc lần lượt từng trang từ đầu đến cuối mà có thể nhảy cóc đến các đoạn sau để tham khảo về các vấn đề họ quan tâm.

Như vậy văn bản HyperText bao gồm dạng chữ viết không liên tục, chúng được phân nhánh và cho phép người đọc có thể chọn cách đọc theo ý muốn của mình. Hiểu theo nghĩa thông thường thì HyperText là một tập các trang chữ viết được kết nối với nhau bởi các liên kết và cho phép người đọc có thể đọc theo các cách khác nhau. Như ta đã làm quen nhiều với các trang định dạng HTML, trong các trang có những liên kết trỏ tới từng phần khác nhau của trang đó hoặc trỏ tới trang khác, và người đọc sẽ đọc văn bản dựa vào những liên kết đó.

Bên cạnh đó, HyperText cũng là một dạng văn bản Text đặc biệt nên cũng có thể bao gồm các chữ viết liên tục (là dạng phổ biến nhất của chữ viết). Do không bị hạn chế bởi

tính liên tục trong HyperText, chúng ta có thể tạo ra các dạng trình bày mới, do đó tài liệu sẽ phản ánh tốt hơn nội dung muốn diễn đạt. Hơn nữa người đọc có thể chọn cho mình một cách đọc phù hợp chẳng hạn như đi sâu vào một vấn đề mà họ quan tâm. Sáng kiến tạo ra một tập các văn bản cùng với các con trỏ trỏ tới các văn bản khác để liên kết một tập các văn bản có mối quan hệ với nhau là một cách thực sự hay và rất hữu ích để tổ chức thông tin. Với người viết, cách này cho phép họ có thể thoải mái loại bỏ những băn khoăn về thứ tự trình bày, mà có thể tổ chức vấn đề thành những phần nhỏ, rồi sử dụng kết nối để chỉ ra mối liên hệ giữa các phần nhỏ đó với nhau.

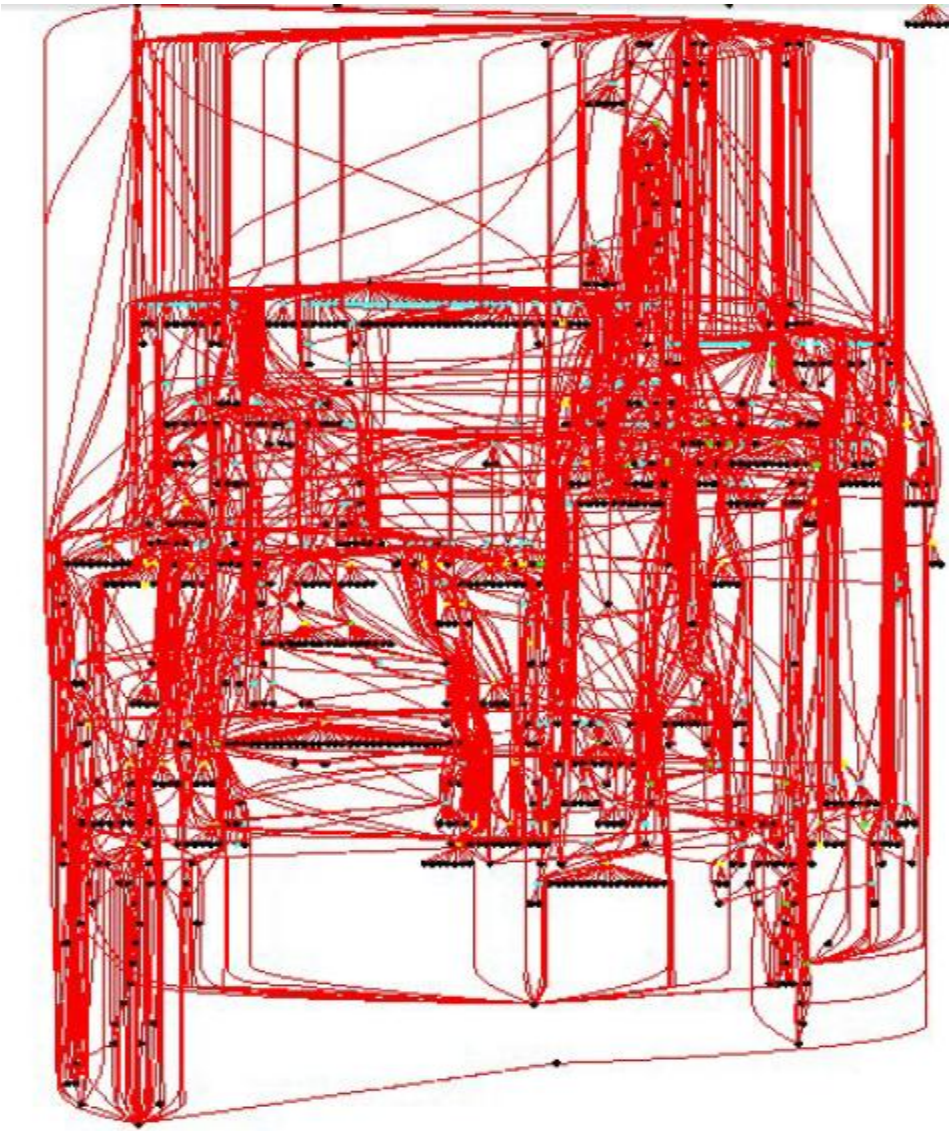
Với người đọc cách này cho phép họ có thể đi tắt trên mạng thông tin và quyết định phần thông tin nào có liên quan đến vấn đề mà họ quan tâm để tiếp tục tìm hiểu. So sánh với cách đọc tuyến tính, tức là đọc lần lượt thì HyperText đã cung cấp cho chúng ta một giao diện để có thể tiếp xúc với nội dung thông tin hiệu quả hơn rất nhiều. Theo khía cạnh của các thuật toán học máy thì HyperText đã cung cấp cho chúng ta cơ hội nhìn ra ngoài phạm vi một tài liệu để phân lớp nó, nghĩa là có tính cả đến các tài liệu có liên kết với nó. Tất nhiên không phải tất cả các tài liệu có liên kết đến nó đều có ích cho việc phân lớp, đặc biệt là khi các siêu liên kết có thể chỉ đến rất nhiều loại các tài liệu khác nhau. Nhưng chắc chắn vẫn còn tồn tại tiềm năng mà con người cần tiếp tục nghiên cứu về việc sử dụng các tài liệu liên kết đến một trang để nâng cao độ chính xác phân lớp trang đó.

Có hai khái niệm về HyperText mà chúng ta cần quan tâm:

Hypertext Document (Tài liệu siêu văn bản): Là một tài liệu văn bản đơn trong hệ thống siêu văn bản. Nếu tưởng tượng hệ thống siêu văn bản là một đồ thị, thì các tài liệu tương ứng với các nút. *Hypertext Link* (Liên kết siêu văn bản): Là một tham chiếu để nối một tài liệu HyperText này với một tài liệu HyperText khác. Các siêu liên kết đóng vai trò như những đường nối trong đồ thị nói trên.

HyperText là loại dữ liệu phổ biến hiện nay, và cũng là loại dữ liệu có nhu cầu tìm kiếm và phân lớp rất lớn. Nó là dữ liệu phổ biến trên mạng thông tin Internet CSDL HyperText với văn bản dạng "nửa cấu trúc" do xuất hiện thêm các "thẻ ": Thẻ cấu trúc (tiêu đề, mở đầu, nội dung), thẻ nhấn trình bày chữ (đậm, nghiêng,…). Nhờ các thẻ này

mà chúng ta có thêm một tiêu chuẩn (so với tài liệu fulltext) để có thể tìm kiếm và phân lớp chúng. Dựa vào các thẻ đã quy định trước chúng ta có thể phân thành các độ ưu tiên khác nhau cho các từ khóa nếu chúng xuất hiện ở những vị trí khác nhau. Ví dụ khi tìm kiếm các tài liệu có nội dung liên quan đến "people " thì chúng ta đưa từ khóa tìm kiếm là "people", và các tài liệu có từ khóa "people" đứng ở tiêu đề thì sẽ gần với yêu cầu tìm kiếm hơn.



Một sơ đồ minh hoạ Hypertext Document như là các nút và các Hypertext Link như là các liên kết giữa chúng.

2.3 SO SÁNH ĐẶC ĐIỂM CỦA DỮ LIỆU FULLTEXT VÀ DỮ LIỆU WEB

Mặc dù trang Web là một dạng đặc biệt của dữ liệu FullText, nhưng có nhiều điểm khác nhau giữa hai loại dữ liệu này. Một số nhận xét sau đây cho thấy sự khác nhau giữa dữ liệu Web và FullText. Sự khác nhau về đặc điểm là nguyên nhân chính dẫn đến sự khác nhau trong khai phá hai loại dữ liệu này (phân lớp, tìm kiếm,…).

Trang web	Văn bản thông thường (Fulltext)
Là dạng văn bản "nửa cấu trúc". Trong nội dung có phần tiêu đề và có các thẻ nhấn mạnh ý nghĩa của từ hoặc cụm từ	Văn bản thường là dạng văn bản "phi cấu trúc". Trong nội dung của nó không có một tiêu chuẩn nào cho ta dựa vào đó để đánh giá
Nội dung của các trang Web thường đươn mô tả ngắn gọn, cô đọng, có các siêu liên kết chỉ ra cho người đọc đến những nơi khác có nội dung liên quan	Nội dung của các văn bản thông thường thường rất chi tiết và đầy đủ
Trong nội dung các trang Web có chứa các siêu liên kết cho phép liên kết các trang có nội dung liên với nhau	Các trng văn bản thông thường không liên kết được đến nội dung của các trang khác

2.4 TEXTMINING VÀ WEBMINING

2.4.1 Các bài toán trong khai phá dữ liệu văn bản

Bài toán 1: Tìm kiếm văn bản

Nội dung: Tìm kiếm văn bản là quá trình tìm kiếm văn bản theo yêu cầu của người dùng. Các yêu cầu được thể hiện dưới dạng các câu hỏi (query), dạng câu hỏi đơn giản nhất là các từ khóa. Có thể hình dung hệ tìm kiếm văn bản sắp xếp văn bản thành hai lớp: Một lớp cho ra những các văn bản thỏa mãn với câu hỏi đưa ra và một lớp không hiển thị những văn bản không được thỏa mãn. Các hệ thống thực tế hiện nay không hiển

thị như vậy mà đưa ra các danh sách văn bản theo độ quan trọng của văn bản tuỳ theo các câu hỏi đưa vào, ví dụ điển hình là các máy tìm tin như Google, Altavista,…

Quá trình thực hiện: quá trình tìm tin được chia thành bốn quá trình chính.

Đánh chỉ số (indexing): Các văn bản ở dạng thô cần được chuyển sang một dạng biểu diễn nào đó để xử lý. Quá trình này còn được gọi là quá trình biểu diễn văn bản, dạng biểu diễn phải có cấu trúc và dễ dàng khi xử lý.

Định dạng câu hỏi: Người dùng phải mô tả những yêu cầu về lấy thông tin cần thiết dưới dạng câu hỏi. Các câu hỏi này phải được biểu diễn dưới dạng phổ biến cho các hệ tìm kiếm như nhập vào các từ khóa cần tìm. Ngoài ra còn có các phương pháp định dạng câu hỏi dưới dạng ngôn ngữ tự nhiên hoặc dưới dạng các ví dụ, đối với các dạng này thì cần có các kỹ thuật xử lý phức tạp hơn. Trong các hệ tìm tin hiện nay thì đại đa số là dùng câu hỏi dưới dạng các từ khóa.

So sánh: Hệ thống phải có sự so sánh rõ ràng và hoàn toàn câu hỏi các câu hỏi của người dùng với các văn bản được lưu trữ trong CSDL. Cuối cùng hệ đưa ra một quyết định phân loại các văn bản có độ liên quan gần với câu hỏi đưa vào và thứ tự của nó. Hệ sẽ hiển thị toàn bộ văn bản hoặc chỉ một phần văn bản.

Phản hồi: Nhiều khi kết quả được trả về ban đầu không thỏa mãn yêu cầu của người dùng, do đó cần phải có quá trình phản hồi để người dùng có thể thay đổi lại hoặc nhập mới các yêu cầu của mình. Mặt khác, người dùng có thể tương tác với các hệ về các văn bản thỏa mãn yêu cầu của mình và hệ có chức năng cập nhật các văn bản đó. Quá trình này được gọi là quá trình phản hồi liên quan (Relevance feeback).

Các công cụ tìm kiếm hiện nay chủ yếu tập trung nhiều vào ba quá trình đầu, còn phần lớn chưa có quá trình phản hồi hay xử lý tương tác người dùng và máy. Quá trình phản hồi hiện nay đang được nghiên cứu rộng rãi và riêng trong quá trình tương tác giao diện người máy đã xuất hiện hướng nghiên cứu là interface agent.

Bài toán 2: phân lớp văn vản (Text categoization)

Nội dung: Phân lớp văn bản được xem như là quá trình gán các văn bản vào một hay nhiều văn bản đã xác định từ trước. Người ta có thể phân lớp các văn bản một cách thủ công, tức là đọc từng văn bản một và gán nó vào một lớp nào đó. Cách này sẽ tốn rất

nhiều thời gian và công sức đối với nhiều văn bản và do đó không khả thi. Do vậy mà phải có các phương pháp phân lớp tự động. Để phân lớp tự động người ta sử dụng các phương pháp học máy trong trí tuệ nhân tạo (Cây quyết định, Bayes, k người láng giềng gần nhất).

+ Một trong những ứng dụng quan trọng nhất của phân lớp văn bản là trong tìm kiếm văn bản. Từ một tập dữ liệu đã phân lớp các văn bản sẽ được đánh chỉ số đối với từng lớp tương ứng. Người dùng có thể xác định chủ đề hoặc phân lớp văn bản mà mình mong muốn tìm kiếm thông qua các câu hỏi.
+ Một ứng dụng khác của phân lớp văn bản là trong lĩnh vực tìm hiểu văn bản. Phân lớp văn bản có thể được sử dụng để lọc các văn bản hoặc một phần các văn bản chứa dữ liệu cần tìm mà không làm mất đi tính phức tạp của ngôn ngữ tự nhiên.
+ Trong phân lớp văn bản, một lớp có thể được gán giá trị đúng sai (True hay False hoặc văn bản thuộc hay không thuộc lớp) hoặc được tính theo mức độ phụ thuộc (văn bản có môt mức độ phụ thuộc vào lớp). Trong trương hợp có nhiều lớp thì phân loại đúng sai sẽ là việc xem một văn bản có thuộc vào một lớp duy nhất nào đó hay không.

Quá trình thực hiện: quá trình phân lớp văn bản. tuân theo các bước sau.

Đánh chỉ số (Indexing): Quá trình đánh chỉ số văn bản cũng giống như trong quá trình đánh chỉ số của tìm kiếm văn bản. Trong phần này thì tốc độ đánh chỉ số đóng vai trò quan trọng vì một số các văn bản mới có thể cần đươc xử lý trong thời gian thực.

Xác định độ phân lớp: Cũng giống như trong tìm kiếm văn bản, phân lớp văn bản yêu cầu quá trình diễn tả việc xác định văn bản đó thuộc lớp nào đó như thế nào, dựa trên cấu trúc biểu diễn của nó. Đối với hệ phân lớp văn bản, chúng ta gọi quá trình này là bộ phân lớp (Categorization hoặc classifier). Nó đóng vai trò như những câu hỏi trong hệ tìm kiếm. Nhưng trong khi những câu hỏi mang tính nhất thời, thì bộ phân loại được sử dụng một cách ổn định và lâu dài cho quá trình phân loại.

So sánh: Trong hầu hết các bộ phân loại, mỗi văn bản đều được yêu cầu gán đúng sai vào một lớp nào đó. Sự khác nhau lớn nhất đối với quá trình so sánh trong hệ tìm

kiếm văn bản là mỗi văn bản chỉ được so sánh với một số lượng các lớp một lần và việc chọn quyết định phù hợp còn phụ thuộc vào mối quan hệ giữa các lớp văn bản.

*Phản hồi (*hoặc *thích nghi):* Quá trình phản hồi đóng vai trò trong hệ phân lớp văn bản. Thứ nhất là khi phân loại thì phải có môt số lượng lớn các văn bản đã được xếp loại bằng tay trước đó, các văn bản này được sử dụng làm mẫu huấn luyện để hỗ trợ xây dựng bộ phân loại. Thứ hai là đối với việc phân loại văn bản này không dễ dàng thay đổi các yêu cầu như trong quá trình phản hồi của tìm kiếm văn bản , người dùng có thể thông tin cho người bảo trì hệ thống về việc xóa bỏ, thêm vào hoặc thay đổi các phân lớp văn bản nào đó mà mình yêu cầu.

Một số bài toán khác: ngoài hai bài toán kể trên, còn có các bài toán sau:

+ Tóm tắt văn bản
+ Phân cụm văn bản
+ Phân cụm các từ mục
+ Phân lớp các từ mục
+ Đánh chỉ mục các từ tiềm năng
+ Trích dẫn văn bản

Trong các bài toán xử lý văn bản đã nêu ở trên, chúng tra thấy vai trò của biểu diễn văn bản rất lớn, đặc biệt trong các bài toán tìm kiếm, phân lớp, phân cụm, trích dẫn.

2.4.2 Khai phá dữ liệu Web

2.4.2.1 Lợi ích của khai phá dữ liệu Web.

Với sự phát triển nhanh chóng của thông tin trên www, KPDL Web đã từng bước trở nên quan trọng hơn trong lĩnh vực KPDL, người ta luôn hy vọng lấy được những tri thức hữu ích thông qua việc tìm kiếm, phân tích, tổng hợp, khai phá Web. Những tri thức hữu ích có thể giúp ta xây dựng nên những Web site hiệu quả để có thể phục vụ cho con người tốt hơn, đặc biệt trong lĩnh vực thương mại điện tử.

Khám phá và phân tích những thông tin hữu ích trên www bằng cách sử dụng kỹ thuật KPDL đã trở thành một hướng quan trọng trong lĩnh vực khám phá tri thức. Khai phá Web bao gồm khai phá cấu trúc Web, khai phá nội dung Web và khai phá các mẫu truy cập Web.

Sự phức tạp trong nội dung của các trang Web khác với các tài liệu văn bản truyền thống. Chúng không đồng nhất về cấu trúc, hơn nữa nguồn thông tin Web thay đổi một cách nhanh chóng, không những về nội dung mà cả về cấu trúc trang. Chẳng hạn như tin tức, thị trường chứng khoán, thông tin quảng cáo, trung tâm dịch vụ mạng,... Tất cả thông tin được thay đổi trên Web theo từng giai đoạn. Các liên kết trang và đường dẫn truy cập cũng luôn thay đổi. Khả năng gia tăng liên tục về số lượng người dùng, sự quan tâm tới Web cũng khác nhau, động cơ người dùng rất đa dạng và phong phú. Vậy làm thế nào để có thể tìm kiếm được thông tin mà người dùng cần? Làm thế nào để có được những trang Web chất lượng cao?....

Những vấn đề này sẽ được thực hiện hiệu quả hơn bằng cách nghiên cứu các kỹ thuật KPDL áp dụng trong môi trường Web. Thứ nhất, ta sẽ quản lý các Web site thật tốt; thứ hai, khai phá những nội dung mà người dùng quan tâm; thứ ba, sẽ thực hiện phân tích các mẫu sử dụng Web.

Dựa vào những vấn đề cơ bản trên, ta có thể có những phương pháp hiệu quả cao để cung cấp những thông tin hữu ích đối với người dùng Web và giúp người dùng sử dụng nguồn tài nguyên Web một cách hiệu quả.

2.4.2.2 Khai phá web

Có nhiều khái niệm khác nhau về khai phá Web, nhưng có thể tổng quát hóa như sau: Khai phá Web là việc sử dụng các kỹ thuật KPDL để tự động hóa quá trình khám phá và trích rút những thông tin hữu ích từ các tài liệu, các dịch vụ và cấu trúc Web. Hay nói cách khác khai phá Web là việc thăm dò những thông tin quan trọng và những mẫu 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 những tri thức, cải tiến việc thiết kế các Web site và phát triển thương mại điện tử tốt hơn. Lĩnh vực này đã thu hút được nhiều nhà khoa học quan tâm. Quá trình khai phá Web có thể chia thành các công việc nhỏ như sau:

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

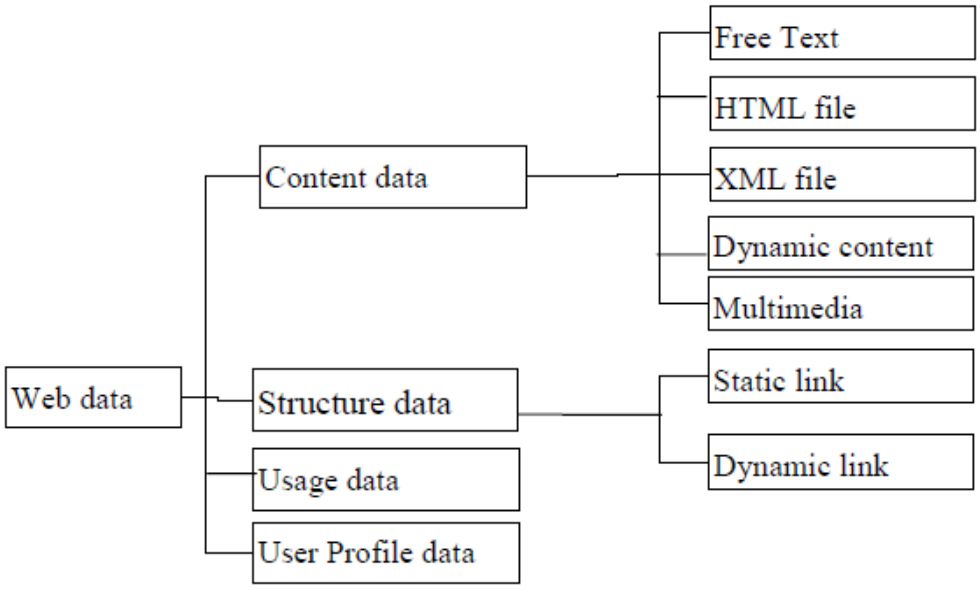
ii. Lựa chọn và tiền xử lý dữ liệu: Lựa chọn và tiền xử lý tự động các loại thông tin từ nguồn tài nguyên Web đã lấy về.

iii. Tổng hợp: Tự động khám phá các mẫu chung tại các Web site riêng lẻ cũng như nhiều Website với nhau.

iiii. Phân tích: Đánh giá, giải thích, biểu diễn các mẫu khai phá được.

2.4.2.3 Các kiểu dữ liệu web

Ta có thể khái quát bằng sơ đồ sau:



Phân loại dữ liệu Web

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

Web logs: Khi người dùng duyệt Web, dịch vụ sẽ phân ra 3 loại dữ liệu đăng nhập: sever logs, error logs, và cookie logs. 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.

2.5 XỬ LÝ DỮ LIỆU VĂN BẢN ỨNG DỤNG TRONG KHAI PHÁ DỮ LIỆU WEB

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

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. Mỗi thành phần của vector được gán một hàm giá trị *f*, nó là một số chỉ mật độ tương ứng của chiều đó trong văn bản. Nếu thay đổi giá trị hàm *f* ta có thể tạo ra nhiều trọng số khác nhau.

Một số vấn đề liên quan đến việc biểu diễn văn bản bằng mô hình 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, ta không phân biệt chữ hoa và chữ thường (nếu chữ hoa thì chuyển về chữ thường).
+ Cắt bỏ từ: 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.

Ngoài ra, để nâng cao chất lượng xử lý, một số công trình nghiên cứu đã đưa ra một số cải tiến thuật toán xem xét đến đặc tính ngữ cảnh của các từ bằng việc sử dụng các cụm từ/văn phạm chứ không chỉ xét các từ riêng lẽ. Những cụm từ này có thể được xác định bằng cách xem xét tần số xuất hiện của cả cụm từ đó trong tài liệu.

Bằng phương pháp biểu diễn không gian vector, ta có thể thấy rõ ràng là 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. Chẳng hạn, số lượng các từ có thể từ 10^3 đến 10^5 đối với các tập văn bản

nhỏ. Vấn đề đặt ra là làm sao để giảm số chiều của vector mà vẫn đảm bảo việc xử lý văn bản đúng và chính xác, đặc biệt là trong môi trường www, ta sẽ xem xét đến một số phương pháp để giảm số chiều của vector.

2.5.1.1 Loại bỏ từ dừng.

Trước hết ta thấy trong ngôn ngữ tự nhiên 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. Do đó, ta có thể loại bỏ những từ đó để giảm số chiều của vector biểu diễn văn bản, những từ như vậy được gọi là những từ dừng.

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

(thống kê của B. Croft, UMass)

Thống kê các từ tần số xuất hiện cao

2.5.1.2 Định luật Zipf

Để giảm số chiều của vector biểu diễn văn bản hơn nữa ta dựa vào một quan sát sau: Nhiều từ trong văn bản xuất hiện rất ít lần, nếu mục tiêu của ta là xác định độ tương tự và sự khác nhau trong toàn bộ tập hợp các văn bản thì các từ xuất hiện một hoặc hai lần (tần số xuất hiện nhỏ) thì ảnh hưởng rất bé đến các văn bản. Tiền đề cho việc lý luận để loại bỏ những từ có tần suất nhỏ được đưa ra bởi Zipf năm 1949. Zipf phát biểu dưới dạng một quan sát nhưng ngay trong thời điểm đó, quan sát đó đã được gọi là định luật Zipf, mặc dù nó thực sự không phải là một định luật mà đúng hơn đó là một hiện tượng xấp xỉ toán học.

Để mô tả định luật Zipf, ta gọi tổng số tần số xuất hiện của từ t trong tài liệu D là f_t. Sau đó sắp xếp tất cả các từ trong tập hợp theo chiều giảm dần của tần số xuất hiện f và gọi thứ hạng của mỗi từ t là r_t.

Định luật Zipf được phát biểu dưới dạng công thức như sau:

$r_t \cdot f_t = K$ (với K là một hằng số).

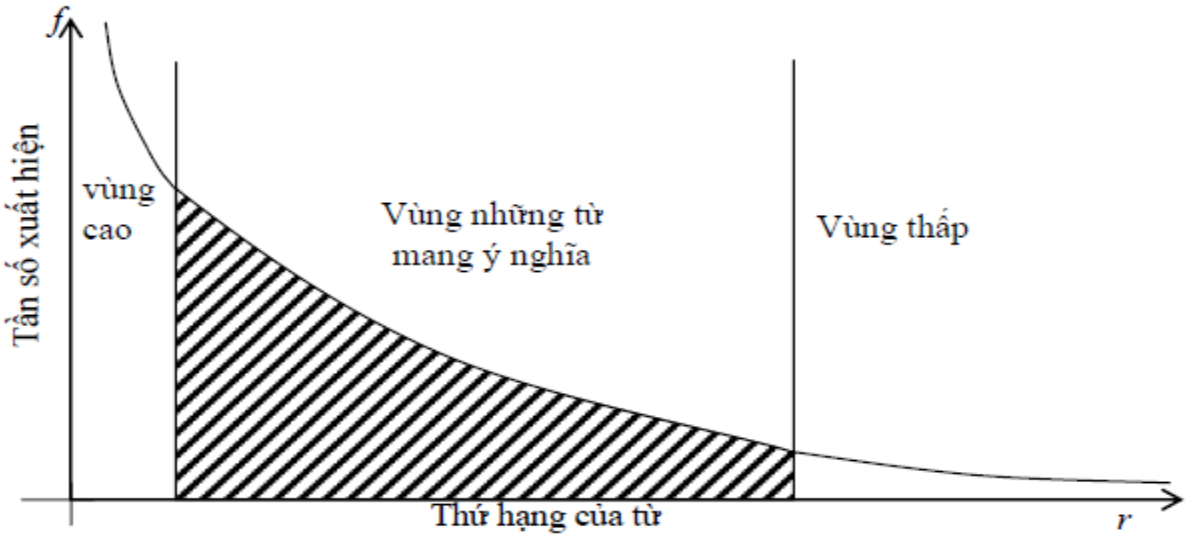
Trong tiếng Anh, người ta thấy rằng hằng số $K = N/10$ trong đó N là số các từ trong văn bản. Ta có thể viết lại định luật Zipf như sau: $r_t = K/f_t$

Giả sử từ t_i được sắp xếp ở vị trí thấp nhất với tần số xuất hiện là b nào đấy và từ t_j cũng được sắp ở vị trí thấp kế tiếp với một tần số xuất hiện là b+1. Ta có thể thu được thứ hạng xấp xỉ của các từ này là $rt_i=K/b$ và $rt_j = K/(b+1)$, trừ 2 biểu thức này cho nhau ta xấp xỉ đối với các từ riêng biệt có tần số xuất hiện là b.

$rt_i - rt_j = K/b - K/(b+1)$

Ta xấp xỉ giá trị của từ trong tập hợp có thứ hạng cao nhất. Một cách tổng quát, một từ chỉ xuất hiện một lần trong tập hợp, ta có $r_{max} = K$.

Xét phân bố của các từ duy nhất xuất hiện b lần trong tập hợp, chia 2 vế cho nhau ta được K/b. Do đó, định luật Zipf cho ta thấy sự phân bố đáng chú ý của các tự riêng biệt trong 1 tập hợp được hình thành bởi các từ xuất hiện ít nhất trong tập hợp.



Lược đồ thống kê tần số của từ theo Định luật Zipf

2.5.2 Các mô hình biểu diễn dữ liệu văn bản.

Trong các bài toán xử lý văn bản, ta thấy rằng vai trò của biểu diễn văn bản rất lớn, đặc biệt trong các bài toán tìm kiếm, phân cụm,…

Theo các nghiên cứu về cách biểu diễn khác nhau trong xử lý văn bản thì cách biểu diễn tốt nhất là bằng các từ riêng biệt được rút ra từ tài liệu gốc và cách biểu diễn này ảnh hưởng tương đối nhỏ đối với kết quả.

Các cách tiếp cận khác nhau sử dụng mô hình toán học khác nhau để tính toán, ở đây ta sẽ trình bày một số mô hình phổ biến và được đăng nhiều trong các bài báo gần đây.

2.5.2.1 Mô hình Boolean

Đây là mô hình biểu diễn vector với hàm *f* nhận giá trị rời rạc với duy nhất hai giá trị đúng/sai (true/false). Hàm *f* tương ứng với thuật ngữ t_i sẽ cho giá trị đúng khi và chỉ khi t_i xuất hiện trong tài liệu đó.

Giả sử rằng có một CSDL gồm *m* văn bản, $D=\{d_1, d_2, ..., d_m\}$. Mỗi văn bản được biểu diễn dưới dạng một vector gồm *n* thuật ngữ $T=\{t_1, t_2,...,t_n\}$. Gọi $W=\{w_{ij}\}$ là ma trận trọng số, w_{ij} là giá trị trọng số của thuật ngữ t_i trong tài liệu d_j.

Mô hình Boolean là mô hình đơn giản nhất, nó được xác định như sau:

$$W_{ij}= \begin{cases} 1 & \text{nếu } t_i \in d_j \\ 0 & \text{nếu } t_i \notin d_j \end{cases}$$

2.5.2.2 Mô hình tần số

Mô hình này xác định giá trị trọng số các phần tử trong ma trận W(w_{ij}) các giá trị là các số dương dựa vào tần số xuất hiện của các từ trong tài liệu hoặc tần số xuất hiện của tài liệu trong CSDL. Có 2 phương pháp phổ biến:

a. Mô hình dựa trên tần số xuất hiện các từ

Trong mô hình dưa trên tần số xuất hiện từ (TF-Term Frequency) giá trị của các từ được tính dựa vào số lần xuất hiện của nó trong tài liệu, gọi tf_{ij} là số lần xuất hiện của từ t_i trong tài liệu d_j, khi đó w_{ij} có thể được tính theo một trong các công thức sau:

- $W_{ij} = tf_{ij}$
- $W_{ij} = 1+\log(tf_{ij})$
- $W_{ij} = \sqrt{tf_{ij}}$

Với mô hình này, trọng số w_{ij} đồng biến với số lần xuất hiện của thuật ngữ t_i trong tài liệu d_j. Khi số lần xuất hiện thuật ngữ t_i trong tài liệu d_j càng lớn thì có nghĩa là d_j càng phụ thuộc nhiều vào thuật ngữ t_i, nói cách khác thuật ngữ t_i mang nhiều thông tin hơn trong tài liệu d_j.

b. *Phương pháp dựa trên tần số văn bản nghịch đảo*

Trong mô hình dưa trên tần số văn bản nghịch đảo (IDF-Inverse Document Frequency) giá trị trọng số của từ được tính bằng công thức sau:

$$W_{ij} = \begin{cases} \log(\frac{n}{h_i}) = \log(n) - \log(h_i) & \text{nếu } t_i \in d_j \\ 0 & \text{nếu ngược lại } (t_i \notin d_j) \end{cases}$$

Trong đó, n là tổng số văn bản trong CSDL, h_i là số văn bản chứa thuật ngữ t_i.

Trọng số w_{ij} trong công thức trên được tính dựa vào độ quan trọng của thuật ngữ t_i trong tài liệu d_j. Nếu t_i xuất hiện càng ít trong các văn bản thì nó càng quan trọng, do đó nếu t_i xuất hiện trong d_j thì trọng số của nó càng lớn, nghĩa là nó càng quan trọng để phân biệt d_j với các tài liệu khác và lượng thông tin của nó càng lớn.

c. *Mô hình kết hợp TF-IDF*

Trong mô hình TF-IDF, mỗi tài liệu d_j được xét đến thể hiện bằng một đặc trưng của (t_1, t_2,.., t_n) với t_i là một từ/cụm từ trong d_j. Thứ tự của t_i dựa trên trọng số của mỗi từ. Các tham số có thể được thêm vào để tối ưu hóa quá trình thực hiện nhóm. Như vậy, thành phần trọng số được xác định bởi công thức sau, nó kết hợp giá trị trọng số *tf* và giá trị trọng số *idf*.

Công thức tính trọng số TF-IDF là:

$$W_{ij} = \begin{cases} tf_{ij} \cdot idf_{ij} = [1 + \log(f_{ij})] \cdot \log(\frac{n}{h_i}) & \text{nếu } t_i \in d_j \\ 0 & \text{nếu ngược lại } (t_i \notin d_j) \end{cases}$$

Trong đó:

tf_{ij} là tần số xuất hiện của t_i trong tài liệu d_j

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

h_i là số các tài liệu mà t_i xuất hiện trong CSDL.

n là tổng số tài liệu trong CSDL.

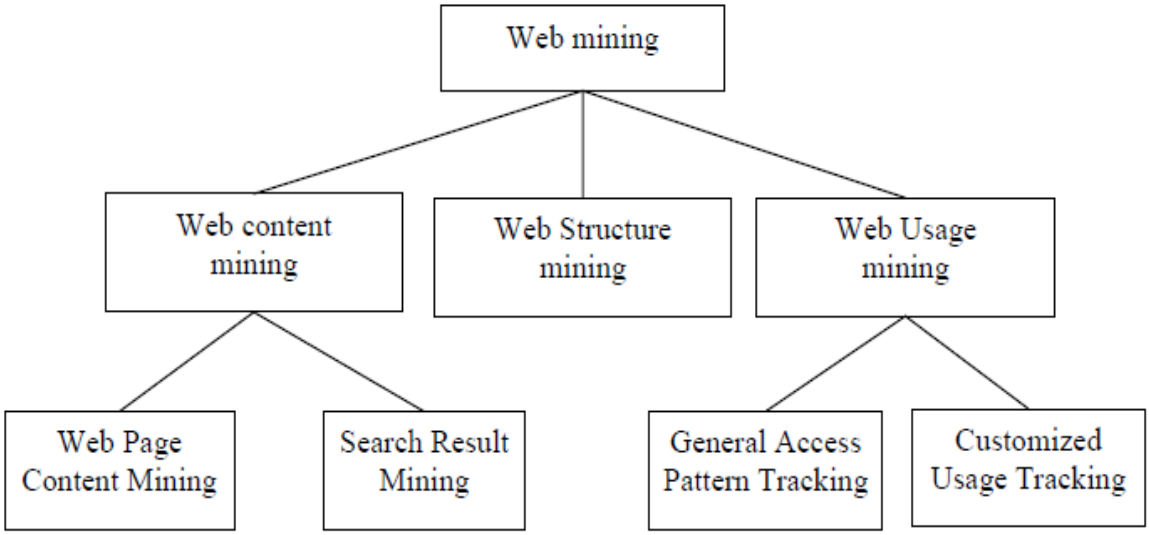
Từ công thức này, ta có thể thấy trọng số của mỗi phần tử là dựa trên nghịch đảo của tần số tài liệu trong CSDL mà t_i và tần số xuất hiện của phần tử này trong tài liệu.

Thông thường ta xây dựng một từ điển từ để lấy đi những từ rất phổ biến và những từ có tần số xuất hiện thấp. Ngoài ra ta phải lựa chọn m (Zemir sử dụng 500) phần tử có trọng số cao nhất như là những từ đặc trưng.

Phương pháp này kết hợp được ưu điểm của cả 2 phương pháp trên. Trọng số w_{ij} được tính bằng tần số xuất hiện của thuật ngữ t_i trong tài liệu d_j và độ "hiếm" của thuật ngữ t_i trong toàn bộ CSDL. Tùy theo ràng buộc cụ thể của bài toán mà ta sử dụng các mô hình biểu diễn văn bản cho phù hợp.

Chương 3
KHAI PHÁ DỮ LIỆU WEB

Tương ứng các kiểu dữ liệu Web, ta có thể phân chia các hướng tiếp cận trong khai phá Web như sau:



Phân loại khai phá Web

3.1 Khai phá nội dung Web

Khai phá nội dung Web tập trung vào việc khám phá một cách tự động nguồn thông tin có giá trị trực tuyến. Không giống như khai phá sử dụng Web và cấu trúc Web, khai phá nội dung Web tập trung vào nội dung của các trang Web, không chỉ đơn thuần là văn bản đơn giản mà còn có thể là dữ liệu đa phương tiện như âm thanh, hình ảnh, phần biến đối dữ liệu và siêu liên kết,....

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 nữa 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 thể được tiếp cận theo 2 cách khác nhau: Tìm kiếm thông tin và KPDL trong CSDL lớn. KPDL đa phương tiện là một phần của khai phá nội dung Web, nó hứa hẹn việc khai thác được các thông tin và tri thức ở mức cao từ nguồn đa phương tiện trực tuyến rộng lớn. KPDL đa phương tiện trên Web gần đây đã thu hút sự quan tâm của nhiều nhà nghiên cứu. Mục đích là làm ra một khung thống nhất đối với việc thể hiện, giải quyết bài toán và huấn luyện dựa vào đa phương tiện. Đây thực sự là một thách thức, lĩnh vực nghiên cứu này vẫn còn là ở thời kỳ sơ khai, nhiều việc đang đợi thực hiện.

Có nhiều cách tiếp cận khác nhau về khai phá nội dung Web, song trong luận văn này sẽ xem xét dưới 2 góc độ: Khai phá kết quả tìm kiếm và khai phá nội dung trang HTML.

3.1.1 Khai phá kết quả tìm kiếm

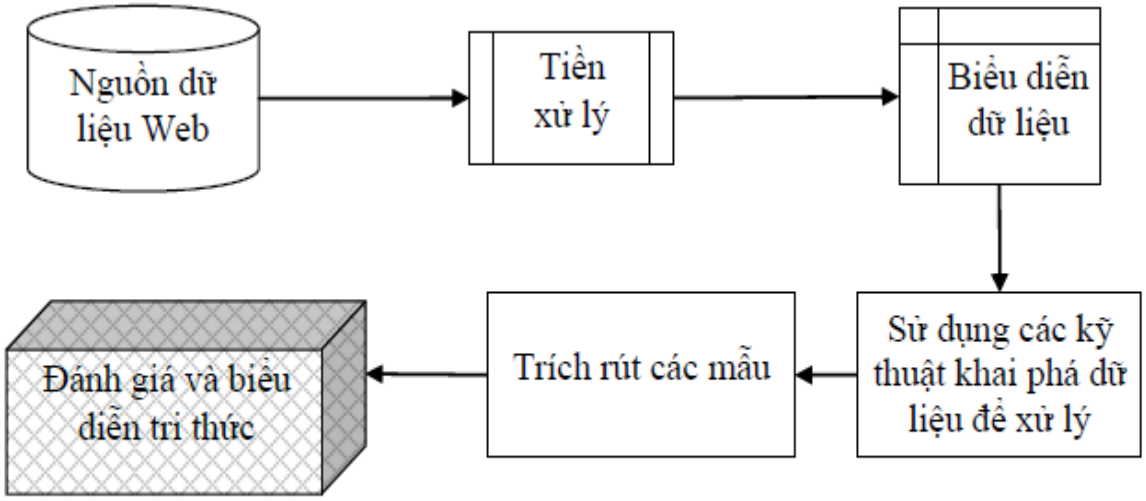
Phân loại tự động tài liệu sử dụng searching engine: Search engine có thể đánh chỉ số tập trung dữ liệu hỗn hợp trên Web. Ví dụ, trước tiên tải về các trang Web từ các Web site. Thứ hai, search engine trích ra những thông tin chỉ mục mô tả từ các trang Web đó để lưu trữ chúng cùng với URL của nó trong search engine. Thứ ba 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 hệ thống 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.

3.1.2. Khai phá văn bản Web

KPVB là một kỹ thuật hỗn hợp. Nó liên quan đến KPDL, xử lý ngôn ngữ tự nhiên, tìm kiếm thông tin, điều khiển tri thức,... KPVB là việc sử dụng kỹ thuật KPDL đối với các tập văn bản để tìm ra tri thức có ý nghĩa tiềm ẩn trong nó. Kiểu đối tượng của nó không chỉ là dữ liệu có cấu trúc mà còn là dữ liệu nữa cấu trúc hoặc không cấu trúc. Kết quả khai phá không chỉ là trạng thái chung của mỗi tài liệu văn bản mà còn là sự phân

loại, phân cụm các tập văn bản phục vụ cho mục đích nào đó. Cấu trúc cơ bản của khai phá thông tin văn bản được thể hiện trong hình dưới đây:



Quá trình khai phá văn bản Web

3.1.2.1. Lựa chọn dữ liệu

Về cơ bản, văn bản cục bộ 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

Ta thường lấy ra những metadata đặc trưng như là một căn cứ và lưu trữ các đặc tính văn bản cơ bản bằng việc sử dụng các quy tắc/ phương pháp để làm rõ dữ liệu. Để có được kết quả khai phá tốt ta cần có dữ liệu rõ ràng, chính xác và xóa bỏ dữ liệu hỗn độn và dư thừa. Trước hết cần hiểu yêu cầu của người dùng và lấy ra mối quan hệ giữa nguồn tri thức được lấy ra từ nguồn tài nguyên. Thứ hai, làm sạch, biến đổi và sắp xếp lại những nguồn tri thức này. Cuối cùng, tập dữ liệu kết quả cuối cùng là bảng 2 chiều. Sau bước tiền xử lý, tập dữ liệu đạt được thường có các đặc điểm như sau:

+ Dữ liệu thống nhất và hỗn hợp cưỡng bức.
+ 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 điễn văn bản

KPVB Web là khai phá các tập tài liệu HTML, là không tự nhiên. 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. Lựa chọn các đặc trưng mà nó chắc chắn trở thành khóa và nó ảnh hưởng trực tiếp đến hiệu quả KPVB.

Phân lớp từ và loại bỏ các từ: Trước hết, chọn lọc các từ có thể mô tả được đặc trưng của tài liệu. Thứ hai, quét tập tài liệu nhiều lần và làm sạch các từ tần số thấp. Cuối cùng ta cũng loại trừ các có tần số cao nhưng vô nghĩa, như các từ trong tiếng Anh: ah, eh, oh, o, the, an, and, of, or,...

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:

- Mỗi từ đặc trưng sẽ nhận được một giá trị trọng số tin cậy bằng việc tính toán hàm trọng số tin cậy. Tần số xuất hiện cao của các từ đặc trưng là khả năng chắc chắn nó sẽ phản ánh đến chủ đề của văn bản, thì ta sẽ gán cho nó một giá trị tin cậy lớn hơn. Hơn nữa, nếu nó là tiêu đề, từ khóa hoặc cụm từ thì chắc chắn nó có giá trị tin cậy lớn hơn. Mỗi từ đặc trưng sẽ được lưu trữ lại để xử lý. Sau đó ta sẽ lựa chọn kích thước của tập các đặc trưng (kích thước phải nhận được từ thực nghiệm).
- Việc rút ra các đặc trưng dựa trên việc phân tích thành phần chính trong phân tích thống kê. Ý tưởng chính của phương pháp này là sử dụng thay thế từ đặc trưng bao hàm của một số ít các từ đặc trưng chính trong phần mô tả để thực hiện giảm bớt số chiều. Hơn nữa, ta cũng sử dụng phương pháp quy nạp thuộc

tính dữ liệu để giảm bớt số chiều vector thông qua việc tổng hợp nhiều dữ liệu thành một mức cao.

3.1.2.5. Khai phá 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 ý nghĩa chính có thể 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.

- 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 $v(y_i)$, $\text{sim}(c_k,y_i)$.

- Lựa chọn tài liệu c_k mà độ tương tự của nó với y_i lớn nhất, như vậy c_k nằm trong phân lớp với y_i, với $\max(\text{sim}(c_k,y_i))$ $i=1,...,m$.

Quá trình được thực hiện lặp lại cho tới khi tất cả các tài liệu đã được phân lớp.

Thuật toán phân lớp K-Nearest Neighbor

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.

- 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 $\text{sim}(c_i,c_j)$ của chúng. Nếu độ tương tự giữa c_i và c_j là lớn nhất, ta sẽ đưa c_i và c_j vào một cụm mới. cuối cùng ta sẽ hình thành được cụm mới $\underline{C}=\{c_1, c_2,..c_{m-1}\}$

- Lặp lại công việc trên cho tới khi chỉ còn 1 phần tử.

Thuật toán phân cụm phân cấp

Toàn bộ quá trình của phương pháp sắp xếp liên kết sẽ tạo nên một cây mà nó phản ánh mối quan hệ lồng 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.

- 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, $\text{sim}(w_i, R_j)$, 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.

Thuật toán phân cụm phân hoạch

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.

3.1.3. Đánh giá chất lượng mẫu

KPDL Web có thể được xem như quá trình của machine learning. Kết quả của machine learning là các mẫu tri thức. Phần quan trọng của machine learning 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

Việc nắm bắt được những đặc tính của người dùng Web là việc rất quan trọng đối với người thiết Web site. Thông qua việc khai phá lịch sử các mẫu truy xuất của người dùng Web, không chỉ thông tin về Web được sử dụng như thế nào mà còn nhiều đặc tính khác như các hành vi của người dùng có thể được xác định. Sự điều hướng đường dẫn người dùng Web mang lại giá trị thông tin về mức độ quan tâm của người dùng đến các WebSite đó.

Dựa trên những tiêu chuẩn khác nhau người dùng Web có thể được phân cụm và các tri thức hữu ích có thể được lấy ra từ các mẫu truy cập Web. Nhiều ứng dụng có thể giúp lấy ra được các tri thức. Ví dụ, văn bản siêu liên kết động được tạo ra giữa các trang Web có thể được đề xuất sau khi khám phá các cụm người dùng Web, thể hiện độ tương tự thông tin. Thông qua việc phát hiện mối quan hệ giữa những người dùng như sở thích, sự quan tâm của người dùng Web ta có thể dự đoán một cách chính xác hơn người sử dụng đang cần gì, tại thời điểm hiện tại có thể dự đoán được kế tiếp họ sẽ truy cập những thông tin và họ cần thông tin gì.

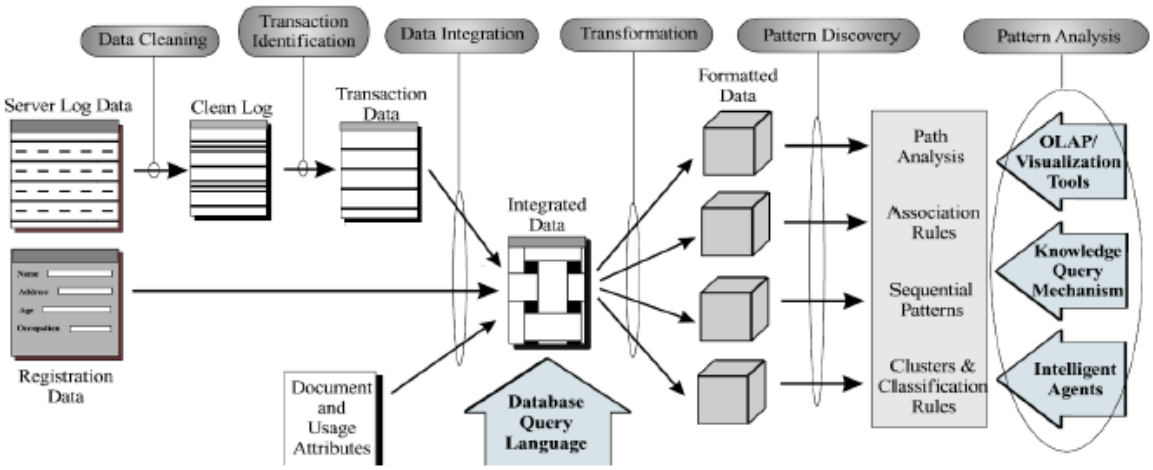
Giả sử rằng tìm được độ tương tự về sự quan tâm giữa những người dùng Web được khám phá từ hiện trạng (profile) của người dùng. Nếu Web site được thiết kết tốt sẽ có nhiều sự tương quan giữa độ tương tự của các chuyển hướng đường dẫn và tương tự giữa sự quan tâm của người dùng.

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 thường 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 Web site 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.

Hiện tại, ta thường sử dụng các công cụ khám phá mẫu và phân tích mẫu. Nó phân tích các hành động người dùng, lọc dữ liệu và khai phá tri thức từ tập dữ liệu bằng cách sử dụng trí tuệ nhân tạo, KPDL, tâm lý học và lý thuyết thông tin. Sau khi tìm ra các mẫu

truy cập ta thường sử dụng các kỹ thuật phân tích tương ứng để hiểu, giải thích và khám phá các mẫu đó. Ví dụ, kỹ thuật xử lý phân tích trực tuyến, tiền phân loại hình thái dữ liệu, phân tích mẫu thói quen sử dụng của người dùng.

Kiến trúc tổng quát của quá trình khai phá theo sử dụng Web như sau:



Kiến trúc tổng quát của khai phá theo sử dụng Web

3.2.1. Ứng dụng của khai phá theo sử dụng Web

- 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.
- Nâng cao chất lượng truyền tải của các dịch vụ thông tin Internet đến người dùng cuối.
- 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.
- 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.
- 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.

3.2.2. 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.

3.2.3. 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 Web site 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 Web site đến lúc họ rời khỏi Web site đó. Những hành động của người dùng trong một Web site 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,...).

a. Chứng thực phiên người dùng

Chứng thực người dùng: Mỗi người dùng với cùng một Client IP được xem là cùng một người.

Chứng thực phiên làm việc: Mỗi phiên làm việc mới được tạo ra khi một địa chỉ mới được tìm thấy hoặc nếu thời gian thăm một trang quá ngưỡng thời gian cho phép (ví dụ 30 phút) đối với mỗi địa chỉ IP.

b. Đăng nhập Web và xác định phiên chuyển hướng người dùng

Dịch vụ file đăng nhập Web: Một file đăng nhập Web là một tập các sự ghi lại những yêu cầu người dùng về các tài liệu trong một Web site, ví dụ:

```
216.239.46.60 - - [04/April/2007:14:56:50 +0200] "GET
/~lpis/curriculum/C+Unix/ Ergastiria/Week-7/filetype.c.txt HTTP/1.0" 304 -
216.239.46.100- - [04/April/2007:14:57:33 +0200]"GET /~oswinds/top.html
HTTP/ 1.0" 200 869
64.68.82.70 - - [04/April/2007:14:58:25 +0200] "GET /~lpis/systems/rdevice/r-
device_examples.html HTTP/1.0" 200 16792
216.239.46.133 - - [04/April/2007:14:58:27 +0200] "GET /~lpis/publications/crc-
chapter1. html HTTP/1.0" 304 -
209.237.238.161 - - [04/April/2007:14:59:11+0200] "GET /robots.txt HTTP/1.0"
404 276
209.237.238.161 - - [04/April/2007:14:59:12 +0200] "GET /teachers/pitas1.html
HTTP/1.0" 404 286
216.239.46.43 - - [04/April/2007:14:59:45 +0200] "GET /~oswinds/publication
```

Nguồn từ: http://www.csd.auth.gr/

Minh họa nội dung logs file

c. Các vấn đề đối với việc xử lý Web log

Thông tin được cung cấp có thể không đầy đủ, không chi tiết.

- Không có thông tin về nội dung các trang đã được thăm.
- Có quá nhiều sự ghi lại các đăng nhập do yêu cầu phục vụ bởi các proxy.
- Sự ghi lại các đăng nhập không đầy đủ do các yêu cầu phục vụ bởi proxy.
- Đặc biệt là việc lọc các mục đăng nhập: Các mục đăng nhập với tên file mở rộng như gif, jpeg, jpg. Các trang yêu cầu tạo ra bởi các tác nhân tự động và các chương trình gián điệp.
- Ước lượng thời gian thăm trang: Thời gian dùng để thăm một trang là một độ đo tốt cho vấn đề xác định mức độ quan tâm của người dùng đối với trang Web đó, nó cung cấp một sự đánh giá ngầm định đối với trang Web đó.
- Khoảng thời gian thăm trang: Đó là khoảng thời gian giữa hai yêu cầu trang khác nhau liên tiếp.
- Quy lui: Nhiều người dùng rời trang bởi họ đã hoàn thành việc tìm kiếm và họ không muốn thời gian lâu để chuyển hướng.

d. Phương pháp chứng thực phiên làm việc và truy cập Web

Chứng thực phiên làm việc: Nhóm các tham chiếu trang của người dùng vào một phiên làm việc dựa trên những phương pháp giải quyết heuristic:

Phương pháp heuristics dựa trên IP và thời gian kết thúc một phiên làm việc (ví dụ 30 phút) được sử dụng để chứng thực phiên người dùng. Đây là phương pháp đơn giản nhất.

Các giao tác nội tại của phiên làm việc có thể nhận được dựa trên mô hình hành vi của người dùng (bao hàm phân loại tham chiếu "nội dung" hoặc "chuyển hướng" đối với mỗi người dùng).

Trọng số được gán cho mỗi trang Web dựa trên một số độ đo đối với sự quan tâm của người dùng (ví dụ khoảng thời gian xem một trang, số lần lui tới trang).

3.2.4. Quá trình khai phá theo sử dụng Web

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

a. 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 đang 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.

b. 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.

Ví dụ:

+ 70% các khách hàng truy cập vào /company/product2 đều xuất phát từ /company thông qua /company/new, /company/products và /company/product1.
+ 80% khách hàng truy cập vào WebSite bắt đầu từ /company/products.
+ 65% khách hàng rời khỏi site sau khi thăm 4 hoặc ít hơn 4 trang.

- 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.

Ví dụ:

+ 40% khách hàng truy cập vào trang Web có đường dẫn /company/product1 cũng truy cập vào /company/product2.
+ 30% khách hàng truy cập vào /company/special đều thông qua /company/product1.

Nó giúp cho việc phát triển chiến lược kinh doanh phù hợp, xây dựng và tổ chức một cách tốt nhất không gian Web của công ty.

-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.

Quá trình thăm của khách hàng được ghi lại trên từng giai đoạn thời gian.

Ví dụ:

+ 30% khách hàng thăm /company/products đã thực hiện tìm kiếm bằng Yahoo với các từ khóa tìm kiếm.
+ 60% khách hàng đặt hàng trực tuyến ở /company/product1 thì cũng đặt hàng trực tuyến ở /company/product4 trong 15 ngày.

-Quy tắc phân loại [22]: 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.

Ví dụ: Khách hàng từ các vị trí địa lý ở một quốc gia hoặc chính phủ thăm site có khuynh hướng bị thu hút ở trang /company/product1 hoặc 50% khách hàng đặt hàng trực tuyến ở /company/product2 đều thuộc nhóm tuổi 20-25 ở Bờ biển Tây.

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

Nó 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.

c. Phân tích đánh giá

Phân tích mô hình: 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.

Ví dụ: Querying: SELECT association-rules(A*B*C*) FROM log.data WHERE (date>= 970101) AND (domain = "edu")AND (support = 1.0) AND (confidence = 90.0)

3.2.5. Ví dụ khai phá theo sử dụng Web

Ví dụ này sử dụng phương pháp khai phá phân lớp và phân cụm, luật kết hợp có thể được dùng để phân tích số lượng người dùng. Sau đó người thiết kế Web có thể đưa ra nhiều dịch vụ khác nhau tại các thời điểm khác nhau theo các quy tắc của người dùng truy cập Web site. Chất lượng dịch vụ tốt sẽ thúc đẩy số lượng người dùng thăm Web site. Quá trình thực hiện như sau:

- Chứng thực người dùng truy cập vào Web site, phân tích những người dùng đặc biệt tìm ra những người dùng quan trọng thông qua mức độ truy cập của họ, thời gian lưu lại trên đó và mức độ yêu thích trang Web.

- Phân tích các chủ đề đặc biệt và chiều sâu nội dung Web. Ví dụ, hoạt động thường ngày của một quốc gia, giới thiệu các tour,... Quan hệ khá tự nhiên giữa người dùng và nội dung Web. Tìm ra những dịch vụ hấp dẫn và tiện lợi với người dùng.

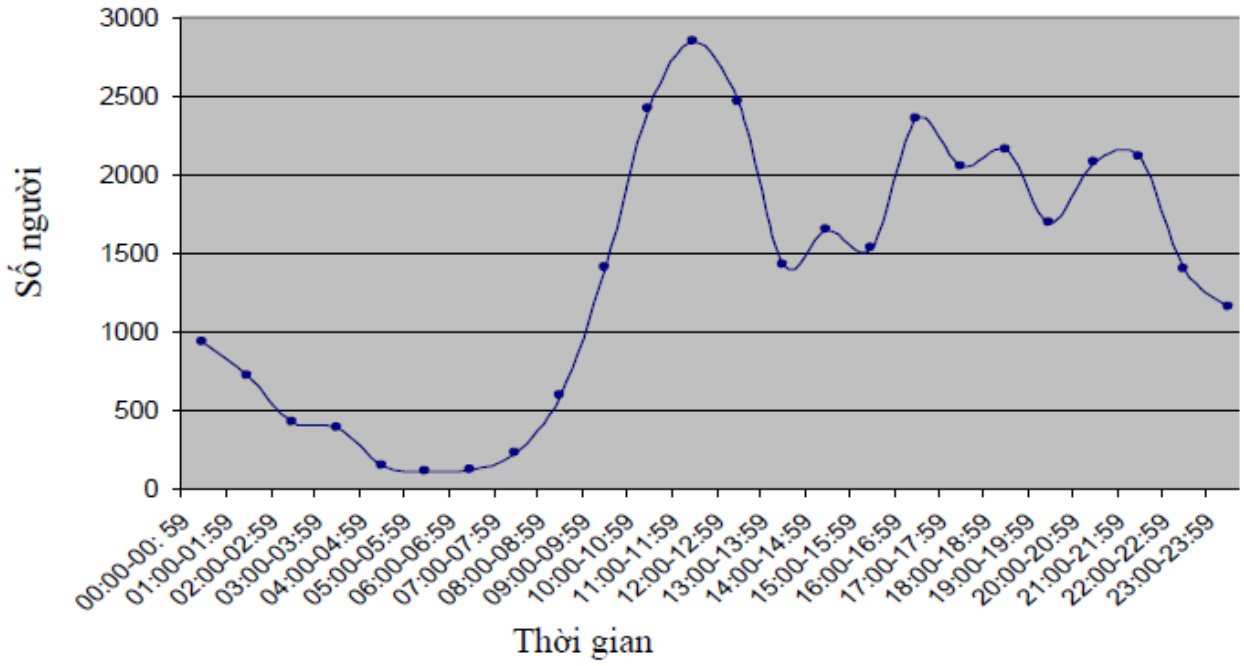
Tùy theo mức độ hiệu quả hoạt động truy cập Web site và điều kiện của việc duyệt Web site ta có thể dự kiến và đánh giá nội dung Web site tốt hơn.

Dựa trên dữ liệu kiểm tra ta xác định mức độ truy xuất của người dùng qua việc phân tích một Web site và phân tích yêu cầu phục vụ thay đổi từng giờ, từng ngày như sau:

Thời gian	Số người truy cập
00:00-00:59	936
01:00-01:59	725
02:00-02:59	433
03:00-03:59	389
04:00-04:59	149
05:00-05:59	118
06:00-06:59	126
07:00-07:59	235
08:00-08:59	599
09:00-09:59	1414
10:00-10:59	2424
11:00-11:59	2846

Thời gian	Số người truy cập
12:00-12:59	2466
13:00-13:59	1432
14:00-14:59	1649
15:00-15:59	1537
16:00-16:59	2361
17:00-17:59	2053
18:00-18:59	2159
19:00-19:59	1694
20:00-20:59	2078
21:00-21:59	2120
22:00-22:59	1400
23:00-23:59	1163

Thống kê số người dùng tại các thời gian khác nhau



Phân tích người dùng truy cập Web

3.3. Khai phá cấu trúc Web

WWW là hệ thống thông tin toàn cầu, bao gồm tất cả các Web site. Mỗi một trang có thể được liên kết đến nhiều trang. Các siêu liên kết thay đổi chứa đựng ngữ nghĩa chung chủ để của trang. Một siêu liên kết trỏ 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.

Sử dụng các phương pháp khai phá người dùng để lấy tri thức hữu ích từ cấu trúc Web, tìm ra những trang Web quan trọng và phát triển kế hoạch để xây dựng các WebSite phù hợp với người dùng.

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

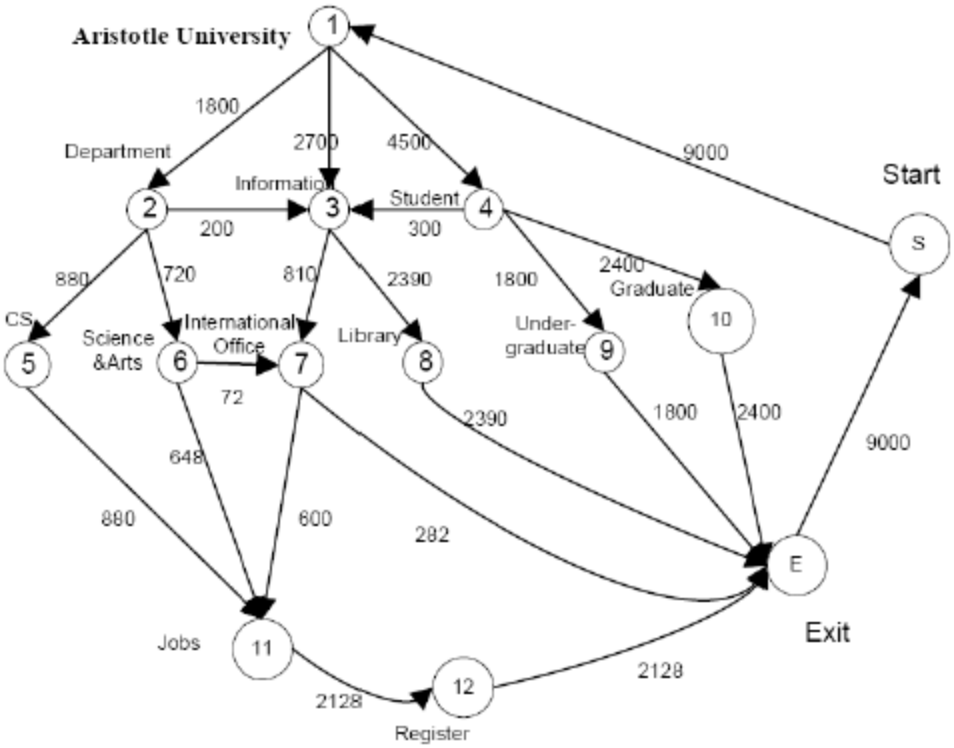
WebSite khác nhau. Nếu trang Web được liên kết trực tiếp với trang Web khác thì ta sẽ muốn phát hiện ra mối quan hệ giữa các trang Web này. Chúng có thể tương tự với nhau về nội dung, có thể thuộc dịch vụ Web giống nhau do đó nó được tạo ra bởi cùng một người. Những nhiệm vụ khác của khai phá cấu trúc Web là khám phá sự phân cấp tự nhiên hoặc mạng lưới của các siêu liên kết trong các Web site của một miền đặc biệt. Điều này có thể giúp tạo ra những luồng thông tin trong Web site 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.

-Việc phân tích liên kết Web được sử dụng cho những mục đích.

+ Sắp thứ tự tài liệu phù hợp với truy vấn của người sử dụng.
+ Quyết định Web nào được đưa vào lựa chọn trong truy vấn.
+ Phân trang.
+ Tìm kiếm những trang liên quan.
+ Tìm kiếm những bản sao của Web.

-Web được xem như là một đồ thị:

+ Đồ thị liên kết: Mỗi nút là một trang, cung có hướng từ *u* đến *v* nếu có một siêu liên kết từ trang Web *u* sang trang Web *v*.



Đồ thi liên kết Web

+ Đồ thị trích dẫn: Mỗi nút cho một trang, không có cung hướng từ *u* đến *v* nếu có một trang thứ ba *w* liên kết cả *u* và *v*.

+ Giả định: Một liên kết từ trang *u* đến trang *v* là một thông báo đến trang *v* bởi trang *u*. Nếu *u* và *v* được kết nối bởi một đường liên kết thì rất có khả năng hai trang Web đó đều có nội dung tương tự nhau.

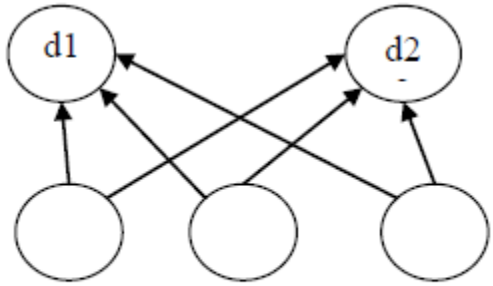
3.3.1. Tiêu chuẩn đánh giá độ tương tự

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: 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.

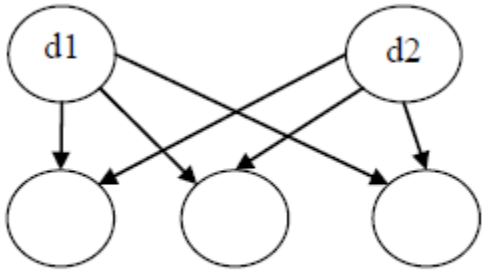
Quan hệ trực tiếp giữa hai trang

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.



Độ tương đồng trích dẫn

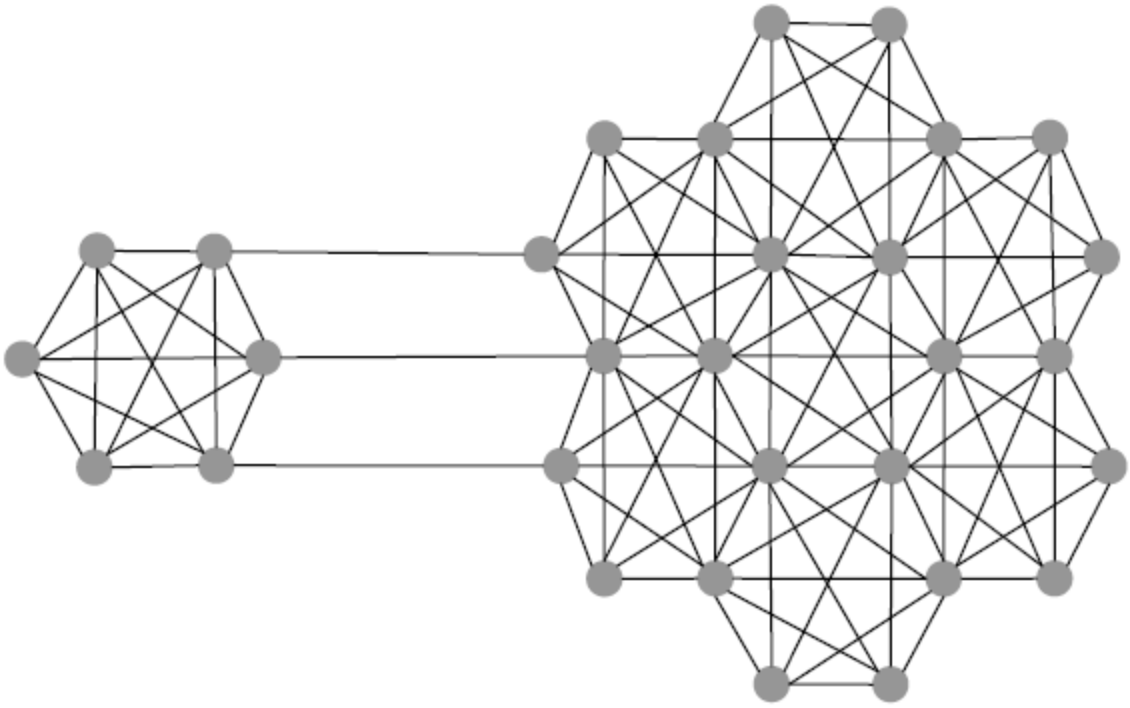
Tương tự chỉ mục: Độ tương tự giữa d_1 và d_2 được đo bằng số trang mà cả d_1 và d_2 đều trỏ tới.



Độ tương tự chỉ mục

3.3.2. Khai phá và quản lý cộng đồng Web

Cộng đồng Web là một nhóm gồm các trang Web chia sẽ chung những vấn đề mà người dùng quan tâm. Các thành viên của cộng đông Web có thể không biết tình trạng tồn tại của mỗi trang (và có thể thậm chí không biết sự tồn tại của cộng đồng). 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.



Cộng đồng Web

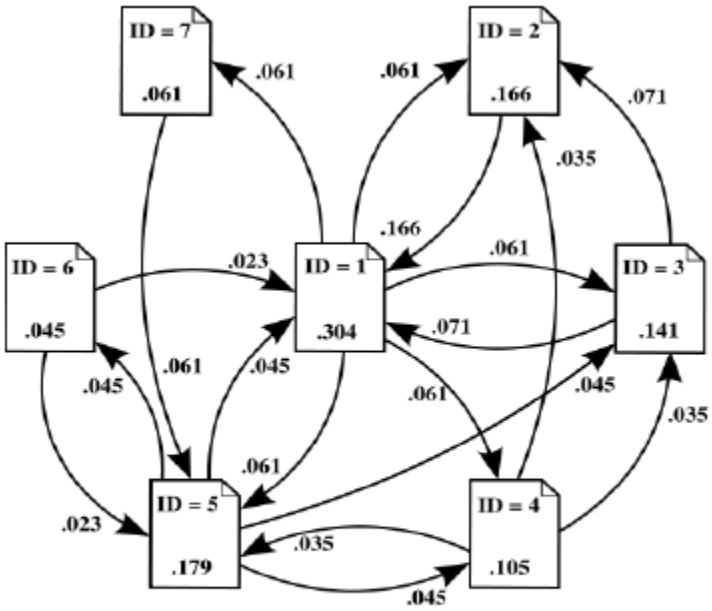
Đặc điểm của cộng đồng Web:

+ 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.

Cộng đồng Web ngày càng được mọi người quan tâm và có nhiều ứng dụng 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. Để trích dẫn ra được các cộng đồng ẩn, ta có thể phân tích đồ thị Web. Có nhiều phương pháp để chứng thực cộng đồng như thuật toán tìm kiếm theo chủ đề HITS, luồng cực đại và nhát cắt cực tiểu, thuật toán PageRank,...

a. Thuật toán PageRank

Google dựa trên thuật toán PageRank [brin98], nó lập chỉ mục các liên kết giữa các Web site và thể hiện một liên kết từ A đến B như là xác nhận của B bởi A. Các liên kết có những giá trị khác nhau. Nếu A có nhiều liên kết tới nó và C có ít các liên kết tới nó thì một liên kết từ A đến B có giá trị hơn một liên kết từ C đến B. Giá trị được xác định như thế được gọi là PageRank của một trang và xác định thứ tự sắp xếp của nó trong các kết quả tìm kiếm (PageRank được sử dụng trong phép cộng để quy ước chỉ số văn bản để tạo ra các kết quả tìm kiếm chính xác cao). Các liên kết có thể được phân tích chính xác và hiệu quả hơn đối với khối lượng chu chuyển hoặc khung nhìn trang và trở thành độ đo của sự thành công và việc biến đổi thứ hạng của các trang.



Kết quả của thuật toán PageRank

PageRank không đơn giản chỉ dựa trên tổng số các liên kết đến. Các tiếp cận cơ bản của PageRank là một tài liệu trong thực tế được xét đến quan trọng hơn là các tài liệu liên kết tới nó, nhưng những liên kết về (tới nó) không bằng nhau về số lượng. Một tài liệu xếp thứ hạng cao trong các phần tử của PageRank nếu như có các tài liệu thứ hạng cao khác liên kết tới nó. Cho nên trong khái niệm PageRank, thứ hạng của một tài liệu được dựa vào thứ hạng cao của các tài liệu liên kết tới nó. Thứ hạng ngược lại của chúng được dựa vào thứ hạng thấp của các tài liệu liên kết tới chúng.

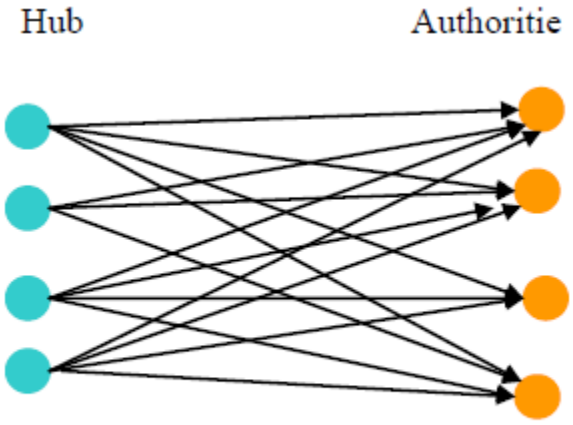
b. Phương pháp phân cụm nhờ thuật toán HITS

Thuật toán HITS (Hypertext-Induced Topic Selection) do Kleinberg đề xuất, là thuật toán phát triển hơn trong việc xếp thứ hạng tài liệu dựa trên thông tin liên kết giữa tập các tài liệu.

Định nghĩa:

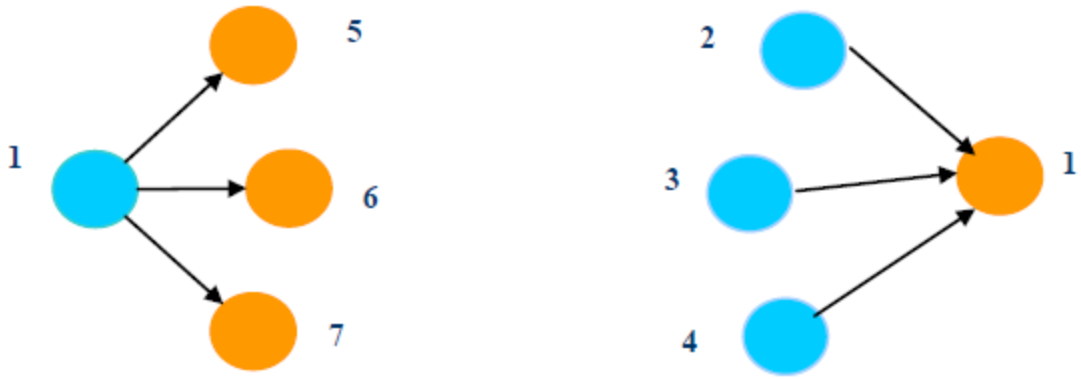
+ Authority: là các trang cung cấp thông tin quan trọng, tin cậy dựa trên các chủ đề đưa ra.
+ Hub: là các trang chứa các liên kết đến authorities
+ Bậc trong: là số các liên kết đến một nút, được dùng để đo độ ủy quyền.
+ Bậc ngoài: là số các liên kết đi ra từ một nút, nó được sử dụng để đo mức độ trung tâm.

Trong đó: Mỗi Hub trỏ đến nhiều Authority, mỗi Authority thì được trỏ đến bởi nhiều Hub. Chúng kết hợp với nhau tạo thành đồ thi phân đôi.



Đồ thị phân đôi của Hub và Authority

Các Authority and hub thể hiện một quan hệ tác động qua lại để tăng cường lực lượng. Nghĩa là một Hub sẽ tốt hơn nếu nó trỏ đến các Authority tốt và ngược lại một Authority sẽ tốt hơn nếu nó được trỏ đến bởi nhiều Hub tốt.



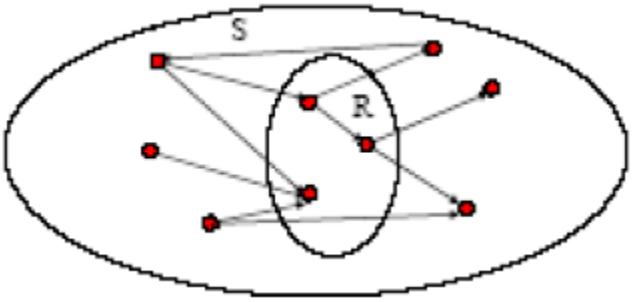
h(1) = a(5) + a(6) + a(7)

a(1) = h(2) + h(3) + h(4)

Sự kết hợp giữa Hub và Authority

Các bước của phương pháp HITS

Bước 1: Xác định một tập cơ bản S, lấy một tập các tài liệu trả về bởi Search Engine chuẩn được gọi là tập gốc R, khởi tạo S tương ứng với R.



Bước 2: Thêm vào S tất cả các trang mà nó được trỏ tới từ bất kỳ trang nào trong R.

Với mỗi trang *p* trong *S*:

Tính giá trị điểm số Authority: a_p (vector a)

Tính giá trị điểm số Hub: h_p (vector h)

Với mỗi nút khởi tạo a_p và h_p là $1/n$ (n là số các trang)

Bước 3. Trong mỗi bước lặp tính giá trị trọng số Authority cho mỗi nút trong S theo công thức:

$$a_p = \sum_{q:\, q \to p} h_q$$

Bước 4. Mỗi bước lặp tính giá trị trọng số Hub đối với mỗi nút trong S theo công thức:

$$h_q = \sum_{q:\, q \to p} a_p$$

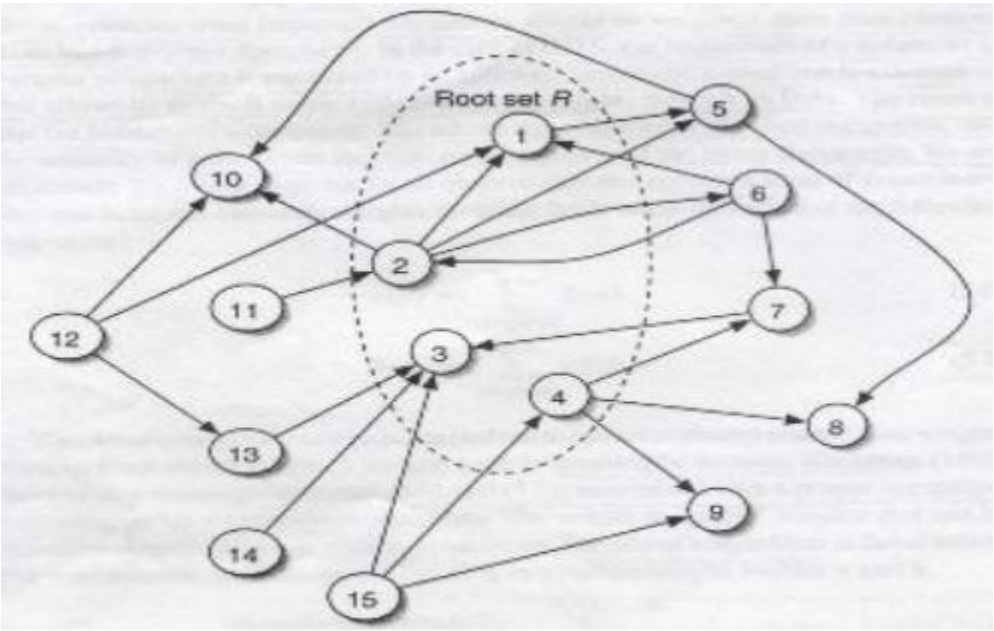
Lưu ý rằng các trọng số Hub được tính toán nhờ vào các trọng số Authority hiện tạo, mà các trọng số Authority này lại được tính toán từ các trọng số của các Hub trước đó.

Bước 5. Sau khi tính xong trọng số mới cho tất cả các nút, các trọng số được chuẩn hóa lại theo công thức:

$$\sum_{p \in S} (a_p)^2 = 1 \quad and \quad \sum_{p \in S} (h_p)^2 = 1$$

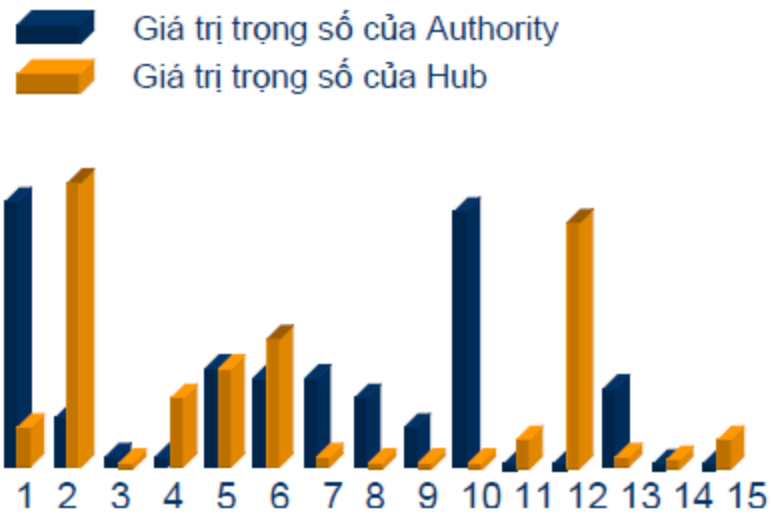
Lặp lại bước 3 cho tới khi các h_p và a_p không đổi.

Ví dụ: Tập gốc R là {1, 2, 3, 4}



Đồ thị Hub-Authority

Kết quả tính được như sau:



Giá trị trọng số các Hub và Authority

KPDL Web là một lĩnh vực nghiên cứu mới, có triển vọng lớn. Các kỹ thuật được áp dụng rộng rãi trên thế giới như KPDL văn bản trên Web, KPDL không gian và thời gian liên tục trên Web. Khai phá Web đối với hệ thống thương mại điện tử, khai phá cấu trúc siêu liên kết Web,... Cho tới nay kỹ thuật KPDL vẫn phải đương đầu với nhiều thử thách lớn trong vấn đề KPDL Web.

Chương 4
HỆ THỐNG TÌM KIẾM VÀ WEB NGỮ NGHĨA

4.1 HỆ THỐNG TÌM KIẾM

4.1.1 Tìm kiếm trên Web

Như đã đề cập ở phần trên, Internet là một kho thông tin khổng lồ và phức tạp. Thông tin trên các trang Web đa dạng về mặt nội dung cũng như hình thức. Tuy nhiên cùng với sự đa dạng và số lượng lớn thông tin như vậy đã nảy sinh vấn đề quá tải thông tin. Cùng với sự thay đổi và phát triển hàng ngày hàng giờ về nội dung cũng như số lượng của các trang Web trên Internet thì vấn đề tìm kiếm thông tin đối với người sử dụng lại ngày càng khó khăn. Đối với mỗi người dùng chỉ một phần rất nhỏ thông tin là có ích, chẳng hạn có người chỉ quan tâm đến trang Thể thao, Văn hóa mà không mấy khi quan tâm đến Kinh tế. Người ta không thể tìm tự kiếm địa chỉ trang Web chứa thông tin mà mình cần, do vậy đòi hỏi cần phải có một trình tiện ích quản lý nội dung của các trang Web và cho phép tìm thấy các địa chỉ trang Web có nội dung giống với yêu cầu của người tìm kiếm.

Định nghĩa: *Máy tìm kiếm (search engine) là một hệ thống phần mềm được xây dựng nhằm tiếp nhận các yêu cầu tìm kiếm của người dùng (thường là một tập các từ khóa), sau đó phân tích yêu cầu này và tìm kiếm thông tin trong cơ sở dữ liệu được tải xuống từ Web và đưa ra kết quả là các trang web có liên quan cho người dùng.* Cụ thể, người dùng gửi một truy vấn, dạng đơn giản nhất là một danh sách các từ khóa, và máy tìm kiếm sẽ làm việc để trả lại một danh sách các trang Web có liên quan hoặc có chứa các từ khóa đó. Phức tạp hơn, thì truy vấn là cả một văn bản hoặc một đoạn văn bản hoặc nội dung tóm tắt của văn bản. Một số máy tìm kiếm điển hình hiện nay: Yahoo, Google, Alvista,...

4.1.2 Cơ chế hoạt động của máy tìm kiếm

Máy tìm kiếm được xem là hệ thống tìm kiếm thông tin điển hình. Hệ thống tìm kiếm thông tin thường tập trung vào việc cải thiện hiệu quả thông tin được lấy ra bằng cách đánh chỉ số dựa trên từ khoá và kỹ thuật cấu trúc lại câu truy vấn. Quá trình xử lý các văn bản dựa trên từ khoá thực hiện việc trích các từ khoá trong văn bản, sử dụng một

từ điển được xây dựng trước, một tập các từ dừng và các quy tắc để chuyển các hình thái của từ về dạng từ gốc. Sau khi các từ được lấy ra, các hệ thống tìm kiếm thường sử dụng phương pháp TF-IDF (hoặc biến thể của nó) để biểu diễn trang Web. Mức độ tương tự đo được giữa một câu truy vấn và một văn bản thường được tính theo độ đo nào đó, trong đó cosin của góc giữa hai vector biểu diễn là một độ đo phổ biến.

Mặc dù trong thực tiễn, mỗi máy tìm kiếm có cách thực thi riêng mà theo đó các thành phần được trình bày như dưới đây có thể được nhập lại hoặc tách ra. Tuy nhiên, những nội dung được trình bày dưới đây mang tính bản chất về hoạt động của các thành phần chức năng trong máy tìm kiếm.

- ***Thành phần crawling*** (Crawler): Đây là thành phần có chức năng thu thập tài nguyên trang Web cho máy tìm kiếm. Thành phần này thực hiện việc duyệt không gian Web, đi dọc theo các liên kết trên các trang Web để thu thập nội dung các trang Web. Crawler nhận tập các địa chỉ URL xuất phát từ dòng xếp hàng các trang Web chưa được thăm (dưới đây gọi là *frontier* theo thuật ngữ tiếng Anh thông dụng của nó) thực hiện tải các trang Web tương ứng về. Trong nhiều trường hợp, thành phần crawling còn bao gồm bộ phân tích cú pháp (parser), bộ điều khiển crawler. Bộ phân tích cú pháp thi hành đối với trang Web, cung cấp các địa chỉ URL chưa được thăm vào dòng xếp hàng. Bộ điều khiển crawler quyết định xem URL nào được duyệt tiếp theo và gửi kết quả cho crawler. Nội dung các trang Web đã được tải về sẽ được store server lưu vào kho trang Web (page repository). Quá trình này được lặp lại cho tới khi đạt tới điều kiện kết thúc.
- ***Thành phần đánh chỉ mục*** (indexer): Đây là thành phần có nhiệm vụ tiếp nhận kết quả phân tích cú pháp trang Web đã được tải về và đánh chỉ mục cho nội dung trang Web. Kết quả của việc đánh chỉ mục sinh ra một tập bảng chỉ mục rất lớn. Nhờ có tập bảng chỉ mục này, máy tìm kiếm nhanh chóng cung cấp được tất cả các địa chỉ URL của các trang Web đáp ứng truy vấn người dùng. Thông thường, bộ tạo chỉ mục tạo ra chỉ mục nội dung (content index) và chỉ mục cấu trúc (structure index). Chỉ mục nội dung chứa thông tin về các từ khoá xuất hiện trong các trang Web. Chỉ mục cấu trúc thể hiện mối liên kết giữa các trang Web, tận

dụng được đặc tính quan trọng của dữ liệu Web là có các liên kết. Nó chính là một dạng đồ thị Web. Cách thức index ngược (invert index) theo từ khoá thường được sử dụng để làm tăng tốc độ tìm kiếm theo từ khoá.

- ***Thành phần phân tích tập*** (Collection Analysis Module): Hoạt động dựa vào đặc trưng của thành phần truy vấn. Chẳng hạn, nếu thành phần truy vấn chỉ đòi hỏi việc tìm kiếm hạn chế trong một số Web site đặc biệt, hoặc giới hạn trong một tên miền, thì công việc sẽ nhanh và hiệu quả hơn. Thành phần này sử dụng thông tin từ hai loại chỉ mục cơ bản (chỉ mục nội dung và chỉ mục cấu trúc) do thành phần đánh chỉ mục cung cấp cùng với thông tin các từ khoá trong trang Web và các thông tin tính hạng để tạo ra các chỉ mục tiện ích.
- ***Thành phần truy vấn*** (query engine): Thành phần này chịu trách nhiệm nhận các yêu cầu tìm kiếm của người sử dụng. Nó thường xuyên truy vấn CSDL, đặc biệt là các bảng chỉ mục để trả về danh sách các tài liệu thoả mãn yêu cầu của người dùng. Do số lượng các trang Web là rất lớn và thông thường người dùng chỉ đưa vào một vài từ khoá trong câu truy vấn nên tập kết quả thường rất lớn. Bộ xếp hạng (ranking) có nhiệm vụ sắp xếp các tài liệu này theo mức độ phù hợp với yêu cầu tìm kiếm để hiển thị kết quả cho người sử dụng. Khi muốn tìm kiếm các trang Web về một chủ đề nào đó, người sử dụng đưa vào một số từ khoá liên quan. Thành phần truy vấn dựa theo các từ khoá này để tìm trong bảng chỉ mục nội dung các địa chỉ URL mà nội dung có chứa từ khoá này. Sau đó, thành phần truy vấn sẽ chuyển các trang Web cho bộ xếp hạng để sắp xếp các kết quả giảm dần về độ liên quan giữa trang Web với truy vấn, rồi hiển thị kết quả cho người sử dụng.

4.2 WEB NGỮ NGHĨA

4.2.1 Giới thiệu Web ngữ nghĩa

Với sự ra đời của Internet và sự phổ biến của Web, số lượng các trang Web ngày càng nhiều đến mức nếu thiếu sự ra đời của các máy tìm kiếm thì người dùng sẽ chẳng khai thác được mấy nội dung của các Web site trên thế giới. Với sự hỗ trợ của các máy tìm kiếm (ví dụ như http://www.google.com hay http://www.yahoo.com), người dùng có thể tìm ra các nội dung mình cần ở một địa chỉ mà mình chưa bao giờ biết đến bằng cách

gõ các từ khoá trong nội dung mình cần tìm hiểu, sau đó một loạt các trang Web liên quan đến nội dung đó sẽ được trả về gần như ngay lập tức. Lúc này người sử dụng chỉ làm công việc đơn giản là duyệt qua các trang Web trả về để tìm ra cái mình cần. Tuy nhiên, hiện tại nội dung các trang Web được viết cho đối tượng là con người, nên các máy tìm kiếm đều "hiểu" nội dung các trang Web ở dưới dạng khá đơn giản, đó là dưới dạng một tập các từ khoá, khi người dùng gõ vào một câu truy vấn dưới dạng một danh sách các từ khoá, hệ thống máy tìm kiếm sẽ tìm tất cả các trang Web chứa các từ khoá đó, sau đó sắp xếp (ranking) theo độ liên quan và trả về cho người dùng. Có những trường hợp kết quả trả về có thể lên đến hàng triệu trang Web, nên người dùng khó mà có thời gian để duyệt toàn bộ nội dung của các trang Web này. Thông thường, người dùng chỉ duyệt khoảng 10 trang kết quả trả về, do đó nếu trang người dùng cần tìm không nằm trong khoảng 10 trang đầu tiên sẽ không được người dùng tìm được.

Từ nhược điểm trên của các máy tìm kiếm, Web ngữ nghĩa (semantic Web) được ra đời như là sự mở rộng của Web, trong đó bên cạnh các thông tin (nội dung) dành cho người dùng, các trang Web còn được bổ sung thêm các thông tin dành cho máy (giúp máy có thể hiểu được "nội dung" của trang Web và xử lý được các thông tin này). Các thông tin dành cho máy này được gọi là siêu dữ liệu (meta–data), tức là dữ liệu dùng để mô tả dữ liệu.

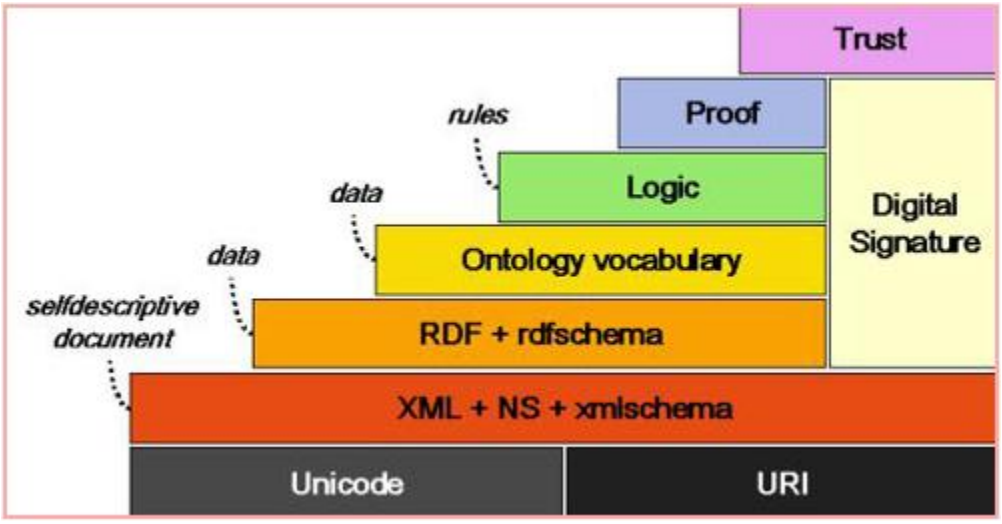
WWW hiện tại là một mạng các tài liệu được diễn tả bằng ngôn ngữ tự nhiên dưới dạng các chuỗi ký tự, nhằm phục vụ đối tượng là con người đọc và hiểu chứ không phải dành cho máy tính. Với Web ngữ nghĩa, bằng việc thêm các siêu dữ liệu và cách tổ chức, biểu diễn thích hợp, các máy tính có thể không những hiểu được "nội dung" của các trang Web mà còn có thể suy diễn ra các tri thức mới, nhằm đáp ứng những yêu cầu của người dùng. Các siêu dữ liệu có thể đơn thuần là các thông tin mô tả các *thực thể tên* (như tên người, tên địa danh,...); *sự kiện* (ví dụ như Olympic thế giới năm 2008); *các con số* (ví dụ như con số biểu thị giá vàng của một ngày cụ thể nào đó, hay tổng thu nhập của một doanh nghiệp trong một năm nào đó); *ngày tháng* (ngày tháng có thể được diễn tả bằng các cách khác nhau trong tài liệu như 20/02/2009, hay "ngày 20 tháng 2 năm 2009"); *tiền tệ* (ví dụ 30$, 40VND); *các tài liệu* (cuốn sách, hay bài báo). Các siêu dữ liệu phức tạp

hơn có thể là các *khái niệm, quan hệ, ràng buộc*,… Tập hợp tất cả các siêu dữ liệu tạo ra một cơ sở tri thức (CSTT) để dựa vào đó có thể suy diễn ra các tri thức mới. Tổ chức World Wide Web Consortium (W3C) (http://www.w3c.org) là người đi tiên phong trong việc xây dựng ra các chuẩn, ngôn ngữ dùng để tổ chức, lưu trữ, thao tác, suy diễn với các cơ sở tri thức phục vụ Web ngữ nghĩa. Và cộng đồng phát triển Web ngữ nghĩa ngày càng thu hút được nhiều người tham gia.

Để so sánh sự khác nhau giữa Web hiện tại và Web ngữ nghĩa, ta có thể so sánh một máy tìm kiếm thông thường và một máy tìm kiếm ngữ nghĩa. Với máy tìm kiếm thông thường, câu truy vấn là một dãy các từ khoá, chẳng hạn như "công ty, viễn thông", máy tìm kiếm sẽ trả về một danh sách các trang Web có chứa các từ khoá trên và sắp xếp (ranking) kết quả theo "mức độ liên quan". Với một máy tìm kiếm ngữ nghĩa, ta có thể đưa một câu truy vấn có "ý nghĩa", chẳng hạn như tìm một "công ty thuộc ngành viễn thông", máy tìm kiếm sẽ trả về danh sách các tài liệu chứa tên các công ty thuộc ngành viễn thông, hay danh sách các công ty thoả mãn điều kiện trên. Điểm khác biệt ở đây là máy tìm kiếm ngữ nghĩa phải "hiểu" được tên các công ty, và "phân loại" được nó thuộc ngành nào, từ đó có thể lọc ra được các công ty thuộc ngành "viễn thông" làm kết quả trả về cho người dùng.

4.2.2 Kiến trúc của Web ngữ nghĩa

Semantic Web là một tập hợp/một chồng (stack) các ngôn ngữ. Tất cả các lớp của Semantic Web được sử dụng để đảm bảo độ an toàn và giá trị thông tin trở nên tốt nhất.



- Lớp **Unicode & URI:** Bảo đảm việc sử dụng tập kí tự quốc tế và cung cấp phương tiện nhằm định danh các đối tượng trong Semantic Web. URI đơn giản chỉ là một định danh Web giống như các chuỗi bắt đầu bằng "http" hay "ftp" mà bạn thường xuyên thấy trên mạng (ví dụ:http://www.cadkas.com)**.** Bất kỳ ai cũng có thể tạo một URI, và có quyền sở hữu chúng. Vì vậy chúng đã hình thành nên một công nghệ nền tảng lý tưởng để xây dựng một hệ thống mạng toàn cầu thông qua đó.
- Lớp **XML** cùng với các định nghĩa về *namespace* (vùng tên gọi) và *schema* (lược đồ) bảo đảm rằng chúng ta có thể tích hợp các định nghĩa Semantic Web với các chuẩn dựa trên XML khác.
- Lớp **RDF [RDF] và RDFSchema [RDFS]:** ta có thể tạo các câu lệnh (*statement)* để mô tả các đối tượng với những từ vựng và định nghĩa của URI, và các đối tượng này có thể được tham chiếu đến bởi những từ vựng và định nghĩa của URI ở trên. Đây cũng là lớp mà chúng ta có thể gán các kiểu (*type*) cho các tài nguyên và liên kết. Và cũng là lớp quan trọng nhất trong kiến trúc Semantic Web.
- Lớp **Ontology**: hỗ trợ sự tiến hóa của từ vựng vì nó có thể định nghĩa mối liên hệ giữa các khái niệm khác nhau. Một Ontology (bản thể luận trong logic) định nghĩa một bộ từ vựng mang tính phổ biến & thông thường, nó cho phép các nhà nghiên cứu chia sẻ thông tin trong một hay nhiều lĩnh vực.
- Lớp **Digital Signature**: được dùng để xác định chủ thể của tài liệu (ví dụ: tác giả hay nhan đề của một loại tài liệu).
- Các lớp **Logic, Proof, Trust:** Lớp logic cho phép viết ra các luật (rule) trong khi lớp proof (thử nghiệm) thi hành các luật và cùng với lớp trust (chấp nhận) đánh giá nhằm quyết định nên hay không nên chấp nhận những vấn đề đã thử nghiệm.