

CHƯƠNG 2

CƠ SỞ LÝ THUYẾT

Chương 2 trình bày cơ sở lý thuyết của đề tài liên quan đến vấn đề truy hồi thông tin (Information Retrieval), các lý thuyết nền tảng về Ontology cùng với các phương pháp và kỹ thuật tính khoảng cách ngữ nghĩa giữa các khái niệm. Đặc biệt, việc nghiên cứu các ontology cho biểu diễn tri thức và biểu diễn ngữ nghĩa, trên cơ sở đó phát triển và xây dựng giải pháp sẽ cho ta giải pháp tốt theo mục tiêu và nhu cầu của ứng dụng thực tế đặt ra.

2.1. VẤN ĐỀ TRUY TÌM THÔNG TIN

2.1.1. Cấu trúc của một hệ thống truy tìm thông tin

Hiện nay, hầu hết các hệ thống tìm kiếm thông tin (Information Retrieval, viết tắt IR) thực chất chỉ là hệ thống truy tìm tài liệu (Document Retrieval), nghĩa là hệ thống sẽ truy tìm những tài liệu (trong số các tài liệu có trong cơ sở dữ liệu lưu trữ) có nội dung liên quan, phù hợp, đáp ứng với nhu cầu thông tin của người dùng, sau đó người dùng sẽ tìm kiếm thông tin họ cần trong các tài liệu liên quan đó. Có hai khái niệm quan trọng luôn đề cập đến đó là tài liệu (document) và câu truy vấn (query). Tài liệu là bất kỳ đối tượng nào mà nó có chứa thông tin, ví dụ như các mẫu văn bản, hình ảnh, âm thanh, video, Tuy nhiên hầu hết các hệ thống IR chỉ đề cập đến các tài liệu là văn bản-text, lý do về sự hạn chế này là vì những khó khăn trong việc biểu diễn các đối tượng không là văn bản.

Một hệ thống IR thường có hai khối chức năng chính, đó là lập chỉ mục và tra cứu hay tìm kiếm. Lập chỉ mục là giai đoạn phân tích tài liệu để rút trích các đơn vị thông tin từ tài liệu và biểu diễn lại tài liệu bởi các đơn vị thông tin đó. Đơn vị thông tin có thể là từ (word), hoặc phức tạp hơn là cụm từ (phrase), khái niệm (concept) và nội dung tài liệu có thể được biểu diễn bởi một cấu trúc đơn giản như danh sách từ (cụm từ)

khóa có đánh trọng số hay một dạng đồ thị giàu ngữ nghĩa hơn. Tra cứu là giai đoạn tìm kiếm trong cơ sở dữ liệu những tài liệu phù hợp với nội dung câu truy vấn. Trong giai đoạn tra cứu, nhu cầu thông tin của người sử dụng được đưa vào hệ thống dưới dạng một câu truy vấn bằng ngôn ngữ tự nhiên hay một dạng thức qui ước nào đó. Câu truy vấn và tập dữ liệu sẽ được phân tích và biểu diễn thành một dạng biểu diễn bên trong. Hệ thống sẽ sử dụng một hàm so khớp (matching function) để so khớp biểu diễn của câu hỏi với các biểu diễn của các tài liệu để đánh giá độ liên quan của các tài liệu với câu truy vấn và trả về các tài liệu có liên quan, được sắp hạng theo độ liên quan với câu truy vấn. Động cơ tìm kiếm có thể tương tác với người dùng thông qua một giao diện (Web chẳng hạn), để có thể hiệu chỉnh dần kết quả trả về cho phù hợp với nhu cầu thông tin của người dùng.

Các hệ thống tìm kiếm thông tin có thể được phân loại như sau:

- Hệ thống tìm kiếm thông tin dựa trên từ khóa: Hệ thống sử dụng một danh sách các từ khóa (keywords) hay thuật ngữ (term) để biểu diễn nội dung tài liệu và câu truy vấn. Tìm kiếm theo từ khóa là tìm kiếm các tài liệu mà những từ trong câu truy vấn xuất hiện nhiều nhất, ngoại trừ stopword (các từ quá thông dụng như mạo từ a, an, the,...), nghĩa là hệ thống giả định nếu một câu hỏi và một tài liệu có chứa một số từ (từ khóa) chung, thì tài liệu là liên quan đến câu hỏi và dĩ nhiên là nếu số từ chung càng nhiều thì độ liên quan càng cao, tài liệu càng được chọn để trả về cho người dùng. Các mô hình tìm kiếm được sử dụng như mô hình Boolean, mô hình không gian vector, các mô hình xác suất, mô hình LSI.

- Hệ thống tìm kiếm thông tin dựa trên khái niệm hay ngữ nghĩa: Nội dung của một đối tượng thông tin được mô tả bởi một tập các khái niệm hay một cấu trúc khái niệm. Để rút trích khái niệm, hệ thống cần sử dụng đến nguồn tri thức về lĩnh vực nhất định nào đó. Hướng tiếp cận chính cho việc nguyên cứu các hệ thống này là sử dụng các kỹ thuật trong xử lý ngôn ngữ tự nhiên và công nghệ ontology.

2.1.2. Hệ thống tìm kiếm thông tin dựa trên khái niệm

Hệ thống tìm kiếm dựa trên khái niệm cũng có chức năng, nguyên lý hoạt động và các bộ phận cấu thành giống như một hệ thống tìm kiếm tổng quát. Tuy nhiên, điểm khác biệt lớn là việc sử dụng khái niệm để lập chỉ mục. Trong bộ lập chỉ mục sẽ có hai nhiệm vụ chính là rút trích toàn bộ các khái niệm có trong cơ sở dữ liệu các tài liệu và lập chỉ mục cho các tài liệu dựa trên các khái niệm này. Cũng giống như bộ truy vấn của hệ tìm kiếm dựa trên từ khóa, bộ truy vấn của hệ thống dựa trên khái niệm có chức năng lấy nội dung câu truy vấn do người dùng nhập vào, sau đó rút trích khái niệm từ câu truy vấn và so trùng với tập chỉ mục đã được lập của các tài liệu để tìm ra các tài liệu có liên quan. Tùy thuộc vào cách lập chỉ mục cho tập khái niệm như thế nào mà sẽ có những cách so trùng câu truy vấn với tập chỉ mục của tài liệu khác nhau, chẳng hạn như nếu bộ lập chỉ mục sử dụng các mô hình truyền thống thì cách bộ truy vấn so trùng các khái niệm cũng giống như trong hệ thống tìm kiếm dựa trên từ khóa, còn nếu một cấu trúc khái niệm biểu diễn tập khái niệm của các tài liệu đã được xây dựng trong quá trình lập chỉ mục, thì cần xây dựng thêm một cấu trúc khái niệm để biểu diễn tập khái niệm của câu truy vấn, sau đó việc tìm kiếm mới có thể được thực hiện dựa trên việc so trùng giữa các cấu trúc khái niệm này.

Các cấu trúc khái niệm có thể tổng quát hoặc cụ thể theo từng lĩnh vực, có thể được tạo thủ công, bán tự động hoặc tự động, chúng có thể khác nhau ở các dạng biểu diễn hoặc ở cách xây dựng mối liên hệ giữa các khái niệm. Các kiểu cấu trúc khái niệm phổ biến: cây khái niệm phân cấp (conceptual taxonomy), nguồn tri thức về lĩnh vực (domain ontology), mạng ngữ nghĩa (semantic linguistic network of concept), các đồ thị khái niệm (conceptual graphs), từ điển từ vựng (thesaurus), mô hình tiên đoán (predictive model) và vector ngữ cảnh (context vector).

Việc xây dựng một hệ thống tìm kiếm dựa trên khái niệm cho đến nay vẫn còn là vấn đề rất khó vì rất nhiều vấn đề vẫn còn khá mới hoặc vẫn chưa có lời giải tối ưu. Ngoài ra, việc xây dựng một cơ sở tri thức cho một lĩnh vực sẽ khó khăn vì tốn nhiều

chi phí xây dựng và duy trì vốn phải có sự can thiệp của con người, đòi hỏi kiến thức của chuyên gia về lĩnh vực và phụ thuộc nhiều vào ngôn ngữ. Đó là lý do khiến các công cụ tìm kiếm theo khái niệm hiện nay chỉ hỗ trợ một lĩnh vực nhất định trong những ứng dụng cụ thể. Mặc dù đã có nhiều công trình nghiên cứu khẳng định hệ thống mà họ xây dựng là một hệ thống tìm kiếm dựa trên khái niệm nhưng vẫn chưa có những đóng góp đáng kể, thực sự không khác nhiều so với một hệ thống tìm kiếm dựa trên từ khóa. Một số công trình nghiên cứu có liên quan gần đây có thể kể đến như:

- Công trình của nhóm tác giả Lê Thị Hoàng Diễm, Jean-Pierre Chevallet và Joo Hwee Lim [10] xây dựng hệ thống tìm kiếm dựa trên khái niệm sử dụng mô hình mạng Bayes, tuy nhiên, cách đánh trọng số cho các mối quan hệ được sử dụng trong mô hình vẫn còn hạn chế.

- Nhóm tác giả Hồ Bảo Quốc, Lê Thúy Ngọc [2] cũng đã tập trung nghiên cứu các vấn đề về tìm kiếm dựa trên khái niệm gồm các phương pháp mở rộng khái niệm, cách tiếp cận lập chỉ mục theo khái niệm và xây dựng thử nghiệm một hệ thống tìm kiếm thông tin y học là CIRS sử dụng nguồn tri thức UMLSMetathesaurus, dùng công cụ MetaMa để rút trích khái niệm tiếng Anh, XIOTA để lập chỉ mục và được thử nghiệm trên bộ dữ liệu ImageCLEFmed của CLEF, tuy nhiên hệ thống lập chỉ mục dựa trên mô hình truyền thống nên không tận dụng được mối liên hệ giữa các khái niệm.

- Nhóm tác giả Đồng Thị Bích Thủy, Nguyễn Phạm Bảo Trâm [3] cũng đã đề xuất một mô hình tìm kiếm dựa trên khái niệm, hướng tới việc xây dựng một hệ thống các dịch vụ hỗ trợ việc tìm kiếm thông tin trong thư viện. Tuy nhiên mô hình này cũng được xây dựng dựa trên các mô hình lý thuyết cổ điển trong lĩnh vực tìm kiếm thông tin đặc biệt là mô hình không gian vector, trong đó có sự cải tiến là biểu diễn tài liệu và câu truy vấn theo các khái niệm dưới dạng vector rồi thực hiện so trùng các vector trong tìm kiếm. Hơn nữa, các khái niệm còn được giả định là hoàn toàn độc lập nhau, nghĩa là ontology ở mức thấp nhất, mối quan hệ giữa các khái niệm không được xem xét đến.

- Một công trình nghiên cứu có liên quan khác là dự án lớn về phát triển một hệ

thống quản lý tri thức và thông tin cho các thực thể có tên ở Việt Nam VN-KIM (dựa theo KIM - Knowledge & Information Management của Ontotext Lab, Bulgaria) [1]. Ontology được xây dựng có khoảng 373 lớp, 114 thuộc tính và khoảng 85000 thực thể về các nhân vật, thành phố, công ty và tổ chức quan trọng và phổ biến có tên ở Việt Nam. Hệ thống sử dụng Sesame để lưu trữ, quản lý Ontology và tri thức, sử dụng công nghệ Lucene để đánh chỉ mục và truy hồi các tài liệu XML đã được chú thích ngữ nghĩa, nhưng theo các thực thể có tên thay vì theo các từ khoá, sử dụng GATE để rút trích thông tin về các thực thể có tên, ứng dụng truy hồi thông tin cho phép trả lời gần đúng và truy vấn bằng đồ thị khái niệm.

- Các tác giả Trương Châu Long [4], Henrik Bulskov Styltsvig [14], Henrik Eriksso [15], Jan Paralic [16] đã áp dụng Ontology cho việc biểu diễn ngữ nghĩa và truy tìm thông tin, dùng Ontology để chuyển việc đánh giá truy vấn dựa trên các từ sang sự đánh giá truy vấn dựa trên khái niệm và dùng tri thức trong các Ontology để so khớp các đối tượng trên ngữ nghĩa cơ bản.

Nhìn chung, các nghiên cứu về tìm kiếm dựa trên khái niệm hiện nay chủ yếu tập trung cải thiện hiệu quả tìm kiếm theo bốn hướng chính [2]:

- Nghiên cứu việc khai thác những nguồn tri thức như WordNet, UMLS, Sensus.
- Nghiên cứu việc mở rộng tài liệu và mở rộng câu truy vấn.
- Nghiên cứu việc sử dụng các kỹ thuật khác để hỗ trợ quá trình tìm kiếm như xử lý ngôn ngữ tự nhiên, fuzzy, khử nhập nhằng, phân loại (clasification),... hay các kỹ thuật để sắp xếp kết quả tìm kiếm (ranking).
- Nghiên cứu cách thức xây dựng, biểu diễn và so trùng các cấu trúc khái niệm, các cách lập chỉ mục khái niệm.

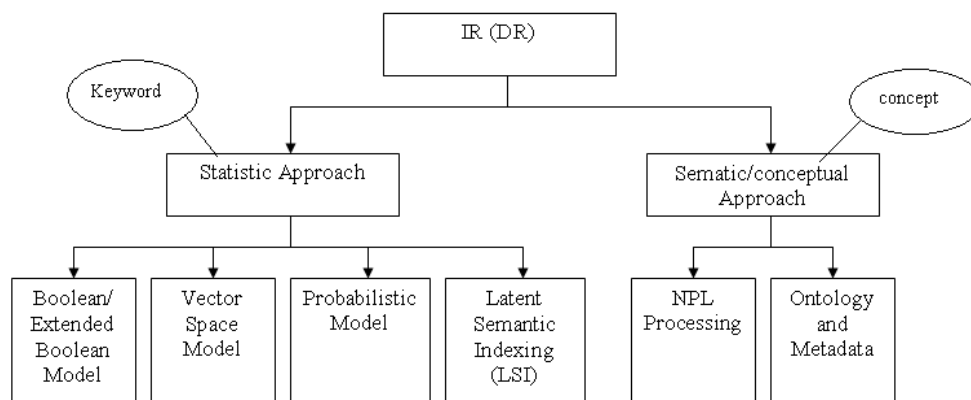
Hele-Mai Haav và Tanel-Lauri Lubi đã làm khảo sát về các công cụ tìm kiếm trên web dựa trên khái niệm [13]. Trong khảo sát này Haav và Lubi cho thấy rằng các công cụ tìm kiếm dựa trên khái niệm chủ yếu vẫn còn là những đề tài nghiên cứu, chưa được thương mại hóa nhiều. Ngoài ra, Haav và Lubi đã liệt kê một số công cụ tìm kiếm,

loại khái niệm cấu trúc, cách biểu diễn cấu trúc, loại mối quan hệ và cách tạo ra cấu trúc khái niệm mà các công cụ đã sử dụng.

Hệ thống tìm kiếm dựa trên khái niệm ngoài áp dụng cho văn bản còn có thể áp dụng cho tìm kiếm hình ảnh và truy vấn thông tin đa ngôn ngữ (Cross language information retrieval – CIRS). Việc tìm kiếm hình ảnh dựa trên khái niệm rất có ý nghĩa. Khi một người dùng tìm kiếm hình ảnh, sẽ chú ý ý nghĩa (nội dung) của hình ảnh đó là gì hơn là hình ảnh đó có màu sắc, hay độ lớn như thế nào. Tuy nhiên, việc lập chỉ mục khái niệm cho hình ảnh khó khăn hơn rất nhiều so với lập chỉ mục văn bản.

2.1.3. Các phương pháp truy hồi thông tin

Nhìn chung, có hai hướng tiếp cận chính cho việc nghiên cứu các hệ thống IR: hướng thống kê và hướng ngữ nghĩa. Trong phương pháp tiếp cận thống kê, các tài liệu kết quả được truy tìm về hoặc được xếp hạng cao là những tài liệu được xem là thích hợp với câu truy vấn nhất theo một số tiêu chí đo lường thống kê, trong khi các phương pháp tiếp cận hướng ngữ nghĩa hay khái niệm lại cố gắng thực hiện việc phân tích cú pháp và ngữ nghĩa, nói cách khác là cố gắng mô phỏng lại các cấp độ hiểu của máy tính về các văn bản theo ngôn ngữ tự nhiên của con người (có thể tham khảo thêm trong các tài liệu [8] và [20]).



Hình 2.1. Các phương pháp truy hồi thông tin

🚦 Truy tìm thông tin theo hướng tiếp cận thống kê

Một số mô hình nổi tiếng được nghiên cứu theo hướng tiếp cận thống kê thuần

túy có thể kể đến là mô hình Boolean, Boolean mở rộng (extended Boolean), Không gian vector (Vector Space), các mô hình xác suất (Probabilistic models). Ý tưởng chính theo hướng tiếp cận này là dùng một danh sách các term xuất hiện trong tài liệu hay câu truy vấn là dạng biểu diễn của nội dung tài liệu và câu truy vấn đó. Term - viết tắt của terminology, nghĩa là thuật ngữ, là một từ hay cụm từ biểu thị một khái niệm khoa học. Khi một phép biểu diễn tài liệu được chọn, chúng ta cần mã hóa chúng trong một dạng thức toán học phù hợp với chương trình máy tính để máy có thể hiểu và xử lý được. Phương pháp đơn giản nhất là mã hóa Boolean.

2.1.3.1. Mô Hình Boolean

Boolean là một mô hình cổ điển và đơn giản nhất được sử dụng trong các hệ thống cũ trước đây. Mô hình Boolean được xây dựng dựa trên lý thuyết tập hợp và đại số Boolean nên đơn giản, dễ hiểu và dễ sử dụng. Với mô hình này, mỗi tài liệu được biểu diễn bởi một vector nhị phân, tức là các vector có các phần tử thuộc $\{0, 1\}$. Term thứ i xuất hiện trong tài liệu d_j thì trọng số $w_{ij} = 1$, ngược lại $w_{ij} = 0$. Các câu truy vấn được đặc tả như một biểu thức Boolean có ngữ nghĩa chính xác, sử dụng ba phép toán cơ bản: not, and, or. Ví dụ, với câu truy vấn “ t_1 AND t_2 ” thì một tài liệu thỏa nhu cầu tìm kiếm nếu và chỉ nếu tài liệu đó chứa cả hai term t_1 và t_2 .

Mô hình Boolean kiểm tra sự xuất hiện của một từ khóa biểu diễn trong một tài liệu hoặc là có hoặc là không. Một truy vấn boolean hoặc là đúng hoặc là sai, tương ứng một tài liệu thỏa hoặc không thỏa hay có liên quan hoặc không liên quan đến nội dung truy vấn. Đây là một hạn chế đáng kể dẫn đến việc không thể sắp hạng kết quả trả về và không thể tìm các tài liệu chỉ liên quan cục bộ hay còn gọi là liên quan một phần với câu truy vấn (ví dụ tài liệu d chỉ có chứa term kB , được xem là không liên quan tới câu truy vấn $q = kA$ AND (kB or kC) bởi vì d không có term kA).

Một số tinh chỉnh trong việc áp dụng mô hình Boolean cổ điển vào các hệ thống IR:

- Thứ nhất, truy vấn có thể được áp dụng cho một thành phần cú pháp đặc biệt của mỗi tài liệu, ví dụ điều kiện boolean có thể được áp dụng cho tiêu đề hoặc phần tóm

tất (abstract) hơn là cho toàn bộ tài liệu.

- Thứ hai, bổ sung thêm một toán tử boolean vào tập hợp ban đầu, ví dụ như toán tử “proximity” dùng để xác định độ gần nhau giữa hai term trong đoạn văn bản. Toán tử này có thể chỉ ra rằng hai term không chỉ cùng xuất hiện trong tài liệu đang xét mà còn cách nhau trong phạm vi n từ ($n = 0$ nghĩa là hai từ đứng liền kề nhau).
- Thứ ba, mô hình boolean cổ điển có thể được xem như là một cách thức thô sơ để biểu diễn những cụm từ và những mối quan hệ đồng nghĩa (gần nghĩa). Ví dụ, t_1 AND t_2 có thể biểu diễn cho một cụm từ gồm 2 term t_1 và t_2 liên kết với nhau hay t_1 OR t_2 có thể biểu diễn cho quan hệ đồng nghĩa giữa 2 term. Thực tế, đã có nhiều hệ thống sử dụng ý tưởng này để xây dựng những điều kiện boolean mở rộng một cách tự động, ví dụ, cho một tập hợp các term truy vấn được cung cấp bởi người dùng, một biểu thức boolean được tạo lập bằng cách dùng các toán tử AND, OR liên kết các term truy vấn với những từ đồng nghĩa tương ứng đã được lưu trữ trước.

2.1.3.2. Mô hình Boolean cải tiến (Advanced Boolean Model)

Thậm chí nếu bổ sung thêm toán tử “proximity” thì điều kiện boolean vẫn là đúng hoặc sai, “tất cả hoặc không có gì” (all – or – nothing) dẫn tới trường hợp là tìm thấy một số lượng lớn tài liệu liên quan hoặc là không có tài liệu nào. Hơn nữa, trong trường hợp câu truy vấn bao gồm nhiều term liên kết với nhau bởi toán tử OR, một tài liệu có chứa tất cả (hay nhiều) term truy vấn cũng không được xem là tốt hơn so với một tài liệu chỉ chứa một term. Tương tự, trong trường hợp với toán tử AND, một tài liệu chứa được gần hết các term vẫn được xem là không phù hợp giống như một tài liệu không chứa term nào. Từ những hạn chế nêu trên, nhiều mô hình boolean mở rộng đã được nghiên cứu phát triển nhằm sắp hạng kết quả trả về. Những mô hình này sử dụng nhiều toán tử boolean mở rộng khác. Ví dụ, một toán tử boolean mở rộng có thể trả về một giá trị cho đối số nằm trong khoảng từ 0 đến 1 (thay vì chỉ là 2 số hoặc 0 hoặc 1) tương ứng với mức độ phù hợp khi so khớp giữa biểu thức logic và tài liệu đang xét (mô hình p – norm là một điển hình).

Ưu điểm của mô hình Boolean:

- Đơn giản, dễ hiểu, dễ cài đặt và sử dụng.
- Mô hình lý thuyết chặt chẽ, rõ ràng.
- Trả về những kết quả chứa chính xác các từ khóa tìm kiếm.

Nhược điểm:

- Đặc tính all – or – nothing, hệ thống chỉ xác định hai trạng thái là tài liệu có liên quan hoặc không liên quan với câu truy vấn nên kết quả trả về hoặc là quá nhiều hoặc không có gì cả. Do đó, hiệu quả truy tìm không cao.
- Mọi quan hệ giữa các term hay thứ tự giữa chúng không được xét đến.
- Không xếp hạng, không xác định được mức độ liên quan giữa tài liệu và câu truy vấn.
- Việc chuyển một câu truy vấn của người dùng sang dạng biểu thức Boolean không đơn giản, người dùng sẽ gặp khó khăn trong việc xây dựng các biểu thức truy vấn boolean.

Nhằm khắc phục những hạn chế trong mô hình Boolean, một mô hình mới đã được đề xuất với ý tưởng chính là xét đến độ tương đồng giữa tài liệu và câu truy vấn thay thế cho việc so khớp chính xác theo cách tiếp cận Boolean.

2.1.3.3. Mô Hình Không Gian Vector(Vector Space Model)

Mô hình không gian vector sẽ biểu diễn mỗi tài liệu văn bản như một tập hợp các term xuất hiện trong toàn bộ tập văn bản và hình thành một không gian mà trong đó mỗi term riêng biệt đóng vai trò là một chiều trong không gian đó, gọi là không gian tài liệu (document space). Người ta gán thêm cho mỗi term một trọng số cục bộ, chỉ có ý nghĩa trong phạm vi tài liệu đang xét. Cùng một term nhưng có thể có trọng số khác nhau trong mỗi tài liệu khác nhau mà nó xuất hiện. Giá trị của mỗi term trong mỗi tài liệu phản ánh mức độ hữu ích, tầm quan trọng của term đó trong việc mô tả nội dung hay chủ đề mà tài liệu đang đề cập tới. Một term có thể mang ý nghĩa lớn trong việc thể hiện nội dung của một tài liệu này nhưng lại kém hiệu quả so với một tài liệu khác và sẽ

có giá trị là 0 nếu như không xuất hiện trong tài liệu đang được xét đến. Các trọng số được gán cho các term trong một tài liệu d có thể được hiểu là tọa độ của d trong không gian tài liệu, nói cách khác, d có thể được biểu diễn như là một điểm (hay vector đi từ gốc tọa độ đến một điểm được định nghĩa là tọa độ của d) trong không gian tài liệu.

Câu truy vấn cũng có thể được cung cấp bởi người sử dụng như là một tập hợp các term đi kèm với các trọng số tương ứng hay được đặc tả dưới dạng ngôn ngữ tự nhiên. Trong trường hợp thứ hai, câu truy vấn sẽ được xử lý như đối với một tài liệu và được chuyển đổi thành tập các term có gán trọng số. Khi đó, câu truy vấn có thể được xem như một tài liệu trong không gian tài liệu.

Một cách hình thức, những tài liệu được biểu diễn trong một không gian tài liệu D có chiều là các đặc trưng $f_i \in F$. Một tài liệu d được biểu diễn như một vector $\vec{d} = (w_{f_1}^d, w_{f_2}^d, \dots, w_{f_n}^d) \in D$ với $w_{f_i}^d$ là trọng số của đặc trưng f_i trong tài liệu d và $n = |F|$. Tương tự, câu truy vấn cũng được biểu diễn trong cùng một không gian tài liệu như một vector $\vec{q} = (w_{f_1}^q, w_{f_2}^q, \dots, w_{f_n}^q) \in D$.

Có nhiều cách tính trọng số được sử dụng, trong đó, phương pháp tính $\text{idf} \times \text{tf}$ được xem là phổ biến và sử dụng rộng rãi nhất. “Term frequency” (tf) là tần số xuất hiện của term trong tài liệu, phản ánh mức độ quan trọng của term trong tài liệu đang xét, ngược lại, “inverse document frequency” (idf) đánh giá mức độ quan trọng của term hay mật độ phân phối của term trong toàn bộ kho tài liệu bằng các xét số tài liệu chứa term đó trên tổng số tài liệu trong kho. Càng có ít tài liệu chứa term đang xét thì giá trị của idf càng lớn và nếu mọi tài liệu đều có chứa term đó thì giá trị của idf sẽ bằng 0. Như vậy, với việc áp dụng $\text{idf} \times \text{tf}$, trọng số được gán tương ứng cho mỗi đặc trưng f của vector \vec{d} được tính bởi công thức sau:

$$w_f^d = \left(\log \frac{N}{N_f}\right) \times tf_f^d = IDF(f) \times tf_f^d$$

trong đó, tf_f^d là tần số xuất hiện của đặc trưng f trong tài liệu d , N là số tài liệu có trong

bộ sưu tập và N_f là số tài liệu mà f xuất hiện.

Sau khi đã biểu diễn tập tài liệu và câu truy vấn thành các vector trong không gian tài liệu, bước tiếp theo là tính toán độ tương quan (giống nhau) giữa chúng bằng cách sử dụng các độ đo sau:

- Inner-product (hoặc *dot-product*): $S_{d,q} = \vec{d} \times \vec{q} = \sum_f w_f^d \times w_f^q$
- Cosin similarity: $S_{d,q} = \cos(\vec{d}, \vec{q}) = \frac{\vec{d} \times \vec{q}}{\|\vec{d}\| \times \|\vec{q}\|} = \frac{\sum_f w_f^d \times w_f^q}{\|\vec{d}\| \times \|\vec{q}\|}$
- Distance metrics: $S_{d,q} = \vec{d} \times \vec{q} = \sqrt[p]{\sum_f (w_f^d \times w_f^q)^p}$
- Hệ số Jaccard: $Jaccard = \frac{n}{N - z}$
- Hệ số Dice: $Dice = \frac{2n}{n_1 + n_2}$

Trong đó: \vec{d} là vector document, \vec{q} là vector truy vấn, n là số term chung của 2 vector d_1 và d_2 , n_1 là số term khác 0 trong d_1 , n_2 là số term khác 0 trong d_2 , N là tổng số term trong không gian vector, z là số term không xuất hiện trong cả d_1 và d_2 ($N - z$ là số term có xuất hiện trong d_1 hoặc d_2 hoặc cả hai)

Ưu điểm của mô hình không gian vector:

- Đơn giản, dễ hiểu, dễ cài đặt.
- Hệ thống đánh trọng số các từ khóa biểu diễn làm tăng hiệu suất tìm kiếm.
- Khắc phục các hạn chế trên mô hình Boolean là tính được mức độ tương đồng giữa một truy vấn và mỗi tài liệu, đại lượng này có thể được dùng để xếp hạng các tài liệu trả về.
- Chiến lược so trùng một phần cho phép trả về các tài liệu phù hợp nhất, thỏa mãn với thông tin truy vấn của người dùng.

Nhược điểm:

- Các từ khóa biểu diễn được xem là độc lập với nhau.
- Số chiều biểu diễn cho tập văn bản có thể rất lớn nên tồn không gian lưu trữ.

2.1.3.4. Mô Hình Xác Suất (Probability Model)

Với câu truy vấn q và tài liệu d_j trong tập hợp các tài liệu, mô hình xác suất cố gắng dự đoán xác suất mà người sử dụng sẽ tìm thấy tài liệu d_j liên quan đến câu truy vấn. Giả định rằng tập tài liệu được chia làm hai phần: ứng với một câu truy vấn q , một tài liệu sẽ có liên quan hay không. Một tài liệu có liên quan đến câu truy vấn hay không khi mà người dùng thích nó (sự kiện L) và ngược lại một tài liệu không liên quan khi không được sự yêu thích của người dùng (sự kiện $\sim L$). Một nguyên tắc xếp hạng được đặt ra như sau:

$$score(d_j) = \frac{P(L | d_j)}{P(\neg L | d_j)}$$

trong đó $P(L | d_j)$ là xác suất tài liệu d_j thích hợp hay liên quan với câu truy vấn q và $P(\neg L | d_j)$ xác suất d_j không thích hợp với q .

Áp dụng chuyển đổi Bayes, ta có thể viết lại các xác suất có điều kiện như sau:

$$score(d_j) = \frac{P(d_j | L)P(L)}{P(d_j | \neg L)P(\neg L)}$$

trong đó, d_j có thể được biểu diễn bởi các thuộc tính hay đặc trưng f_i của nó. Giả định các đặc trưng này là các sự kiện độc lập để đơn giản hóa các tính toán. Đặt A_i là một sự kiện ràng buộc thuộc tính f_i , ta có:

$$score(d_j) = \frac{\prod_i P(A_i | L)P(L)}{\prod_i P(A_i | \neg L)P(\neg L)}$$

Hàm xếp hạng này được chuyển đổi logarit và khi đó các hằng số $P(L)$, $P(\sim L)$ sẽ được loại bỏ, ta được công thức sau:

$$score_{\log}(d_j) = \sum_{A_i \in d_j} \frac{P_i(1 - \bar{P}_i)}{\bar{P}_i(1 - P_i)} = \sum_{A_i \in d_j} weight(A_i)$$

với P_i là xác suất mà thuộc tính A_i xuất hiện trong tài liệu khi nó thích hợp với truy vấn của người dùng và \bar{P}_i là xác suất cho thuộc tính xuất hiện khi tài liệu không thích hợp ($P(A_i | L) = P_i(1 - \bar{P}_i)$).

Ưu điểm của mô hình xác suất:

- Có thể sắp hạng các tài liệu dựa vào xác suất liên quan đến câu truy vấn.
- Mô hình xác suất đạt được nhiều chất lượng về hiệu năng truy tìm hơn so với các mô hình không áp dụng phương pháp xác suất.

Nhược điểm:

- Không thể biểu diễn thông tin ngữ nghĩa về một tài liệu theo công thức xác suất.
- Phương pháp này không lưu ý đến tần suất xuất hiện của các từ khóa biểu diễn trong tài liệu.
- Giả định các từ khóa biểu diễn độc lập nhau.
- Phải chia tập tài liệu được chia thành 2 loại: thích hợp hay không thích hợp.
- Việc tính toán xác suất khá phức tạp và tốn nhiều chi phí.

Một trong những hạn chế lớn của mô hình không gian vector và mô hình xác suất là giả định các term độc lập với nhau, nghĩa là các mối tương quan ngữ nghĩa giữa các term này không được xét đến và do đó không thể so trùng giữa những từ có hình thức thể hiện bên ngoài khác nhau nhưng có nghĩa tương tự nhau. Một nhược điểm khác của mô hình không gian vector là số chiều của không gian tài liệu có thể rất lớn nếu như số lượng các term xuất hiện trong bộ sưu tập các tài liệu là rất lớn. Phần tiếp theo sẽ giới thiệu một kỹ thuật thống kê cố gắng khắc phục những vấn đề nêu trên bằng cách xem xét đến những mối quan hệ giữa các term, theo đó các term cùng biểu diễn một thông tin ngữ nghĩa sẽ được phân nhóm, gom cụm lại với nhau.

2.1.3.5. Latent Semantic Indexing - LSI

Latent Semantic Indexing(*LSI*) là phương pháp tạo chỉ mục tự động dựa trên khái niệm để khắc phục hai hạn chế tồn tại trong mô hình không gian vector chuẩn

(VSM) cũng như các mô hình Boolean và xác suất: *synonymy* và *polysemy*. Với synonymy, nhiều từ có thể được sử dụng để biểu diễn một khái niệm, vì vậy hệ thống không thể trả về những tài liệu liên quan đến câu truy vấn của người dùng khi họ sử dụng những từ trong câu truy vấn đồng nghĩa với những từ trong tài liệu. Với polysemy, một từ có thể có nhiều nghĩa, vì vậy hệ thống có thể trả về những tài liệu không liên quan với những gì mà người dùng mong muốn có được. Điều này thực tế rất thường xảy ra bởi vì các tài liệu được viết bởi rất nhiều tác giả, với cách dùng từ rất khác nhau. Trong LSI, không gian tài liệu được thay thế bởi một không gian tài liệu có chiều thấp hơn gọi là không gian k (k - space) hay không gian LSI, trong đó mỗi chiều là một khái niệm độc lập (nghĩa là không có tương quan với nhau) đại diện cho một nhóm các term cùng biểu diễn cho một thông tin ngữ nghĩa. Mô hình LSI sử dụng chỉ mục khái niệm (conceptual index) được tạo ra bởi phương pháp thống kê thay cho việc sử dụng các từ chỉ mục đơn.

Mô hình LSI dựa trên giả thiết là có các ngữ nghĩa tiềm ẩn (*latent semantic*) trong việc sử dụng từ: có nhiều từ biểu diễn cho một khái niệm và một khái niệm có thể được biểu diễn bởi nhiều từ. Và mô hình này sử dụng phân tích SVD (*Singular Value Decomposition*) ma trận term – document A để phát hiện ra các quan hệ ngữ nghĩa tiềm ẩn đó. Mô hình LSI, mở rộng của mô hình không gian vector, sử dụng phép chiếu trực giao ma trận biểu diễn tập văn bản có hạng r vào không gian k chiều, trong đó $k \ll r$. Việc chọn hệ số k tối ưu cho mô hình LSI vẫn còn là bài toán chưa có lời giải tổng quát. Cho tới hiện tại việc chọn k cho mô hình LSI chỉ thực hiện dựa trên các phương pháp thử nghiệm.

Truy hồi thông tin theo hướng ngữ nghĩa

2.1.3.6. Xử lý ngôn ngữ tự nhiên

Trong các phần trước, chúng ta đã tìm hiểu về các phương pháp truy hồi thông tin theo hướng tiếp cận thống kê. Theo hướng tiếp cận này thì một tài liệu thường được biểu diễn dưới dạng một tập hợp các từ khóa độc lập nhau. Đây được xem là một

phương pháp phổ biến dùng cho việc biểu diễn các tài liệu mà không xét đến hình thái của từ, thứ tự của các từ hay vị trí xuất hiện của từ trong tài liệu cũng như các mối quan hệ ngữ nghĩa giữa chúng, do đó cách biểu diễn này mang mức độ thông tin thấp và nếu nhìn dưới góc nhìn của ngôn ngữ học thì đã không xử lý các biến thể về mặt ngôn ngữ học của các từ như biến thể về hình thái học (morphological variation), biến thể về từ vựng học (lexical variation), biến thể về ngữ nghĩa học (semantical variation) và biến thể về cú pháp học (syntax variation). Biến thể về hình thái học là các dạng khác nhau về mặt cấu trúc (hình dáng, thể hiện bên ngoài) của một từ, ví dụ như các từ *computer*, *computerize*, *computers* là các biến thể về hình thái học của từ *computer*. Hệ thống sẽ cho kết quả không chính xác nếu đối xử với các biến thể này như các từ độc lập nhau. Biến thể về từ vựng học là các từ khác nhau mang cùng một nghĩa, ví dụ *car*, *auto*. Hệ thống sẽ không trả về các tài liệu có chứa từ *auto* mà không chứa từ *car* khi câu hỏi chỉ chứa từ *car*. Biến thể về ngữ nghĩa học là vấn đề một từ đa nghĩa tùy vào ngữ cảnh, ví dụ từ *bank* có nhiều nghĩa như *ngân hàng*, *bờ*, *bãi ngầm*, ... Biến thể về cú pháp học là các kết hợp khác nhau về mặt cú pháp của cùng một nhóm từ sẽ mang các ý nghĩa khác nhau, ví dụ một tài liệu chứa câu ‘near to the river, air pollution is a major problem’ thì không liên quan gì đến ‘river pollution’ cả mặc dù cả hai từ đều có xuất hiện trong tài liệu. Để nâng cao hiệu quả của các hệ tìm kiếm thông tin, người ta phải có các giải thuật để xử lý các biến thể ngôn ngữ học như đã nêu:

Đối với các biến thiên về hình thái học người ta có hai cách để xử lý: cách thứ nhất là mở rộng câu hỏi bằng cách thêm vào câu hỏi tất cả các biến thể hình thái học của tất cả các từ có trong câu hỏi, cách thứ hai là chuẩn hoá các biến thể hình thái học của một từ về một chuẩn chung (stemming), nghĩa là khử các tiền tố và hậu tố thông thường của từ, trả về dạng gốc của mỗi từ. Ví dụ như các từ *computer*, *computed*, *computes*, *computerize* sẽ được chuẩn hoá thành là *compute*. Khi đó, người sử dụng không cần thiết phải đặc tả câu truy vấn của mình theo một hình thái đặc biệt nào của từ mà anh ta tin rằng chúng có thể xuất hiện bên trong tài liệu đang tìm kiếm.

Để xử lý các biến thể về từ vựng học người ta hoặc là mở rộng câu hỏi bằng cách thêm vào câu hỏi tất cả các từ đồng nghĩa có thể có của tất cả các từ trong câu hỏi hoặc là xử lý ở giai đoạn so khớp bằng cách đưa ra các độ đo khoảng cách của các khái niệm. Đối với cách thứ nhất chúng ta cần có một từ điển đồng nghĩa, đối với cách thứ hai chúng ta phải xây dựng một tự điển từ vựng trong đó có định nghĩa khoảng cách giữa các từ.

Biến thể về ngữ nghĩa thường kết hợp chặt chẽ với biến thể về từ vựng học. Để xử lý các biến thể này chúng ta cần một công đoạn xử lý sự đa nghĩa của từ, hiệu năng của hệ thống tìm kiếm sẽ phụ thuộc vào kết quả của giai đoạn xử lý này.

Các kỹ thuật xử lý các biến thể về cú pháp học hay nói cụ thể hơn là xử lý cấu trúc của một cụm từ có thể được chia làm hai loại: kỹ thuật lập chỉ mục dựa vào các cụm từ và kỹ thuật lập chỉ mục là các cấu trúc cây phân tích được từ các mệnh đề. Các kỹ thuật lập chỉ mục dựa trên cụm từ nhằm tăng độ chính xác của hệ thống. Với giả định rằng khi dùng các cụm từ như các chỉ mục thay cho các từ đơn thì độ chính xác sẽ tăng do cụm từ biểu diễn chính xác hơn nội dung của tài liệu. Các hệ thống tìm kiếm dựa trên chỉ mục là các cụm từ ngày càng thu hút nhiều nhóm nghiên cứu và vấn đề làm thế nào để rút trích được các cụm từ một cách tự động từ tài liệu trở thành vấn đề chính trong các hệ này. Các giải pháp rút trích cụm từ thường dựa vào hai cách tiếp cận: tiếp cận dùng thông tin thống kê tần suất đồng xuất hiện hay cách tiếp cận dựa vào tri thức về ngôn ngữ học. Cách tiếp cận thứ hai đòi hỏi phải áp dụng nhiều kỹ thuật của lĩnh vực xử lý ngôn ngữ tự nhiên. Kỹ thuật lập chỉ mục cấu trúc dựa vào các cấu trúc cây có được từ việc phân tích các mệnh đề trong câu của tài liệu và quá trình so khớp là so khớp các cấu trúc của câu hỏi với các cấu trúc của tài liệu. Cách tiếp cận này không thu hút nhiều nhóm nghiên cứu do độ phức tạp của việc phân tích mệnh đề để xây dựng cách cấu trúc cao nhưng lại không tăng được hiệu năng của hệ thống tìm kiếm.

Ngoài ra, để khắc phục những hạn chế trong việc biểu diễn tài liệu từ những mô hình truyền thống, nhiều nghiên cứu khác nhau đã nỗ lực thay đổi cách biểu diễn cho tài

liệu nhằm làm tăng hiệu quả trong biểu diễn và tìm kiếm. Theo đó, một tài liệu vẫn được mô tả bởi các cặp <đặc trưng, trọng số>, tuy nhiên những thành phần đặc trưng cho tài liệu không đơn thuần chỉ là những từ hay cụm từ chính xác xuất hiện trong tài liệu mà đã được thiết kế lại, được chuẩn hóa theo một dạng thức biểu diễn phức tạp và hiệu quả hơn bằng cách sử dụng các kỹ thuật trong xử lý ngôn ngữ tự nhiên. Những nghiên cứu này hướng tới mục tiêu là xây dựng một phép biểu diễn dựa trên các khái niệm hơn là các từ đơn lẻ cũng như cố gắng loại bỏ các vấn đề nhập nhằng trong ngôn ngữ. Một số mô hình nổi tiếng có thể kể đến như:

- **Lemmas:** các đặc trưng của tài liệu được chọn là các hình thái cơ bản của từ như danh từ hay động từ. Như vậy, hệ thống sẽ chuẩn hóa các biến thể về hình thái học của từ về một chuẩn chung và thay thế những từ có trong tài liệu bởi hình thái cơ bản của chúng. Điều này sẽ làm tăng khả năng so khớp giữa những từ có hình thái thể hiện khác nhau nhưng phản ánh cho cùng một khái niệm.
- **Simple n-grams:** một dãy các từ được lựa chọn bằng cách áp dụng kỹ thuật thống kê. Hệ thống tiến hành khảo sát và thống kê các dãy bao gồm n từ liên tiếp tùy ý (n - gram) có trong kho ngữ liệu. Như vậy, mỗi tài liệu sẽ được chia thành những cấu trúc n – gram tương ứng. Những bộ lọc thống kê dựa trên tần số xuất hiện của các n-gram trong kho ngữ liệu được áp dụng để lựa chọn những ứng viên phù hợp nhất làm đặc trưng cho tài liệu.
- **Nouns Phrases:** Những biểu thức chính qui (ví dụ như N^+ là một dãy các danh từ liên kết với nhau theo một qui tắc cú pháp nhất định) dựa trên các từ loại (danh từ, động từ và tính từ) có thể được sử dụng để chọn ra các cụm từ dùng làm đặc trưng cho tài liệu và loại bỏ những kết hợp không khả thi. Cụm từ được chọn bao gồm một từ chính (head) và các phụ ngữ hay từ bổ nghĩa (modifier) đứng trước và sau nó.
- **Các bộ <head, modifier₁, ..., modifier_n>:** Những Bộ phân tích cú pháp (parser) được sử dụng để phát hiện và rút trích ra các quan hệ cú pháp phức tạp như subject-

verb-object từ trong văn bản. Một đặc tính thú vị là những bộ này có thể bao gồm những từ không liên kết nhau, tức là các thành phần có thể là những từ vốn nằm cách nhau trong đoạn văn bản. Việc xây dựng những cụm từ phức hợp này là nhằm cải thiện độ chính xác trong việc so khớp giữa các khái niệm.

- **Semantic concepts:** mỗi từ được thay thế bằng một đại diện cho nghĩa của từ đó. Việc gán nghĩa cho một từ phụ thuộc vào định nghĩa của từ đó có trong từ điển. Có hai cách xác định nghĩa của một từ. Thứ nhất, nghĩa của từ có thể được trình bày, giải thích như trong một mục từ của từ điển giải nghĩa thông thường. Thứ hai, nghĩa của từ có thể được suy ra thông qua những từ khác có cùng nghĩa trong từ điển đồng nghĩa.

Tuy nhiên, cho đến nay thì những kết quả đạt được theo cách tiếp cận này vẫn chưa có sự cải thiện đáng kể so với các phương pháp thống kê kể trên. Nguyên nhân chính là do những mô hình biểu diễn mới cũng chỉ nắm bắt được một phần nhỏ thông tin hơn so với mô hình truyền thống. Hơn nữa, những lỗi xuất hiện trong quá trình rút trích tự động các khái niệm hay trong quá trình xây dựng các mô hình biểu diễn có thể gây nhiễu và làm ảnh hưởng đến tiến trình tìm kiếm.

2.1.3.7. Hướng tiếp cận Ontology

Ontology là bản mô tả tường minh các khái niệm trong một miền ứng dụng nào đó và quan hệ giữa những khái niệm này cùng một số luật logic và suy diễn, cho phép suy luận khái niệm mới từ các khái niệm đã có. Ontology cung cấp từ vựng thống nhất cho việc trao đổi thông tin giữa các ứng dụng. Những tìm hiểu về cơ sở lý thuyết của ontology sẽ được trình bày trong phần 2.2.

2.1.4. Đánh giá một hệ thống tìm kiếm thông tin

Hiệu quả của một hệ truy tìm thông tin có thể được đánh giá theo các tiêu chuẩn sau [2]:

- Để đánh giá **hiệu quả truy tìm của hệ thống**, người ta sử dụng đến hai độ đo cơ bản là độ chính xác (precision) và độ bao phủ (recall). Những độ đo này đo sự thỏa mãn

của người dùng với các tài liệu mà hệ thống tìm thấy. Cho S là tập các tài liệu được tìm thấy (liên quan theo hệ thống). Cho U là tập các tài liệu liên quan theo đánh giá của người dùng. Khi đó, độ chính xác và độ bao phủ sẽ được định nghĩa như sau:

Độ chính xác: là sự tương ứng giữa số tài liệu mà hệ thống tìm thấy có liên quan đến câu truy vấn theo người dùng trên tổng số các tài liệu tìm thấy của hệ thống.

$$\text{Độ chính xác} = \frac{|S \cap U|}{|S|}$$

Độ chính xác 100% nghĩa là tất cả các tài liệu mà hệ thống tìm thấy đều liên quan đến câu truy vấn theo người dùng.

Độ bao phủ: là sự tương quan giữa số tài liệu hệ thống tìm thấy được đánh giá là liên quan theo người dùng trên tổng số các tài liệu có liên quan theo người dùng.

$$\text{Độ bao phủ} = \frac{|S \cap U|}{|U|}$$

Độ bao phủ là 100% có nghĩa là hệ thống tìm thấy tất cả các tài liệu liên quan.

Thông thường, khó đáp ứng được cả hai độ đo này cùng một lúc. Một hệ thống muốn tăng độ chính xác thường sẽ phải giảm độ bao phủ và ngược lại.

- **Hiệu quả thực thi của hệ thống (Execution efficiency)** được đo bởi thời gian thực hiện thủ tục tìm kiếm các văn bản liên quan đến câu truy vấn được cho.
- **Hiệu quả lưu trữ** được đo bởi dung lượng bộ nhớ cần thiết để lưu trữ dữ liệu (cả bộ nhớ ngoài lưu trữ dữ liệu chỉ mục và bộ nhớ RAM khi hệ thống thực thi).

2.2. ONTOLOGY

Công nghệ ontology là một công nghệ được nghiên cứu phát triển mạnh mẽ trong thời gian gần đây. Ontology trở thành một lĩnh vực nghiên cứu phổ biến có mặt trong nhiều lĩnh vực từ xử lý ngôn ngữ tự nhiên, công nghệ tri thức, các hệ thống trao đổi, tích hợp thông tin cho đến biểu diễn và quản lý tri thức. Ontology giúp ta xây dựng mạng lưới ngữ nghĩa, bộ từ điển về các lĩnh vực chuyên môn hỗ trợ trong các ứng dụng, giúp ta mã hóa tri thức lĩnh vực thành một hệ tri thức dùng chung mà máy tính có thể

hiểu được bằng cách phân tách khối tri thức này thành các đối tượng tri thức nhỏ hơn và tìm ra các mối liên hệ giữa chúng. Phần tìm hiểu hiểu tổng quan về ontology dưới đây được tham khảo và có trích dẫn một phần dựa trên tài liệu [7].

2.2.1. Định nghĩa

Trong triết học

Ontology là một thuật ngữ có nguồn gốc từ Triết học diễn tả các thực thể tồn tại trong tự nhiên và các mối quan hệ giữa chúng. Theo cách nhìn của triết học, ontology – bản thể học là “một môn khoa học về nhận thức, cụ thể hơn là một nhánh của siêu hình học về tự nhiên và bản chất của thế giới, nhằm xem xét các vấn đề về sự tồn tại hay không tồn tại của các sự vật”. Theo đó người ta đưa ra khái niệm bộ ba ngữ nghĩa bao gồm *biểu tượng* – *khái niệm* – *sự vật*, đây là mô hình dùng để mô tả hay biểu diễn thế giới thực, *biểu tượng* sẽ gọi lên *khái niệm* và biểu diễn *sự vật* còn *khái niệm* sẽ đề cập tới *sự vật*.

Trong lĩnh vực Trí tuệ nhân tạo

Trong Trí tuệ nhân tạo đã có nhiều cách định nghĩa khác nhau về ontology, một số định nghĩa được xem là kinh điển và được thừa nhận rộng rãi như sau:

- ❖ Gruber (1993) định nghĩa ontology như “một đặc tả tường minh của sự khái niệm hóa trong một lĩnh vực”.
- ❖ Borst (1997) sửa đổi một chút định nghĩa của Gruber, rằng ontology là “sự đặc tả hình thức của sự khái niệm hóa được chia sẻ”. Studer (1998) giải thích hai định nghĩa của Gruber và Borst như sau “Sự khái niệm hóa có nghĩa là mô hình trừu tượng của các sự vật, hiện tượng trên thế giới được xác định qua các khái niệm liên quan của sự vật, hiện tượng đó. Tường minh có nghĩa là các kiểu khái niệm và các ràng buộc giữa chúng là được xác định rõ ràng. Hình thức có nghĩa là ontology phải được hiểu bởi máy tính. Chia sẻ có nghĩa là tri thức trong ontology được kết hợp xây dựng và được chấp nhận bởi một nhóm hoặc một cộng đồng chứ không theo tri thức chủ quan của cá nhân”.

❖ Motta (1999) định nghĩa “ontology là đặc tả một phần của tập hợp các khái niệm được sử dụng hình thức hóa các tri thức của một lĩnh vực cần quan tâm. Vai trò cơ bản của một ontology là nhằm chia sẻ và sử dụng lại tri thức”.

❖ Uschold và Jasper (1999) phát biểu rằng “ontology chứa các định nghĩa và quan hệ giữa các khái niệm, hình thành một cấu trúc lĩnh vực và giới hạn ngữ nghĩa của thuật ngữ trong từ vựng”.

❖ Weiss (1999) định nghĩa “ontology là một đặc tả của các khái niệm và quan hệ trong lĩnh vực quan tâm. Ontology không chỉ là phân cấp các lớp mà còn mô tả các quan hệ”.

❖ Theo định nghĩa của Hendler năm 2001, “ontology là một tập hợp các thuật ngữ tri thức (knowledge term), bao gồm từ vựng, các quan hệ ngữ nghĩa, một số luật suy diễn và logic trong một lĩnh vực đặc thù”.

Nhìn chung, có rất nhiều định nghĩa về ontology, mỗi định nghĩa thể hiện một cách nhìn khác nhau và đi kèm với nó là một phương pháp luận và kỹ thuật xây dựng ontology. Một định nghĩa mang tính tổng hợp và đúng theo định hướng xây dựng ontology của đề tài như sau: “Một ontology xác định một bảng từ vựng chung cho những người cần chia sẻ thông tin trong một lĩnh vực, bao gồm định nghĩa của các khái niệm cơ bản mà máy tính có thể hiểu được trong một lĩnh vực nào đó và sự liên quan giữa chúng”.

2.2.2. Các thành phần của ontology

Ontology được xây dựng thường có các thành phần cơ bản sau:

- Các lớp (class) (tương ứng với các concept – khái niệm): là trung tâm của hầu hết các ontology, mô tả các khái niệm trong miền lĩnh vực. Các lớp thường được tổ chức phân cấp và áp dụng kỹ thuật thừa kế. Một lớp có thể có các lớp con biểu diễn khái niệm cụ thể hơn so với lớp cha.

- Thuộc tính (property hay role, slot): mô tả các đặc tính, đặc trưng, tính chất khác nhau của khái niệm và mỗi thuộc tính đều có giá trị. Thuộc tính được phân biệt với

quan hệ (relation) dựa trên giá trị là một kiểu dữ liệu (string, number, boolean, ...). Một thuộc tính bản thân nó cũng có các thuộc tính con và cũng có các ràng buộc trên nó.

- Quan hệ (relation): biểu diễn các kiểu quan hệ giữa các khái niệm. Các quan hệ nhị phân được sử dụng để biểu diễn thuộc tính. Tuy nhiên, giá trị của quan hệ khác với giá trị của thuộc tính ở chỗ giá trị của quan hệ là một khái niệm.

- Thực thể hay thể hiện (instance): biểu diễn các phần tử riêng biệt của khái niệm, là các thể hiện của lớp. Mỗi thể hiện của lớp biểu diễn một sự cụ thể hóa của khái niệm đó.

- Hàm (function): là một loại thuộc tính hay quan hệ đặc biệt, trong đó, phần tử thứ n là duy nhất đối với $n-1$ phần tử còn lại.

- Tiên đề (Axioms): biểu diễn các phát biểu luôn đúng mà không cần phải chứng minh hay giải thích. Axioms được sử dụng để kiểm chứng sự nhất quán của ontology hoặc cơ sở tri thức. Cả hai thành phần hàm và tiên đề góp phần tạo nên khả năng suy diễn trên ontology.

2.2.3. Phân loại ontology

Về cơ bản có các loại ontology sau:

- Ontology biểu diễn tri thức (Knowledge representation Ontology) nắm giữ các biểu diễn nguyên thủy được dùng để chuẩn hóa tri thức trong một mô hình biểu diễn tri thức. Một trong những ontology thuộc loại này là Frame Ontology của Gruber, ontology này định nghĩa những khái niệm như là frame, slot và các ràng buộc slot cho phép biểu diễn tri thức theo hướng đối tượng hoặc theo frame-based.

- Ontology tổng quát (Generic Ontology) bao gồm từ vựng liên quan tới sự vật, hiện tượng, thời gian, không gian, quan hệ nhân quả ...có ý nghĩa chung chung không chỉ dùng riêng cho một lĩnh vực nào. Ví dụ: WordNet, CYC, ...

- Metadata ontology cung cấp từ vựng dùng để mô tả nội dung của các nguồn thông tin trực tuyến. Ví dụ ontology Dublin Core

- Ontology lĩnh vực (Domain Ontology) là những ontology có thể tái sử dụng

trong một lĩnh vực nào đó, nó cung cấp từ vựng về các khái niệm và các mối quan hệ trong một lĩnh vực. Ví dụ: ontology về y khoa MeSH, GALEN hay ontology về sinh học Gene Ontology, OBO.

- Ontology tác vụ (Task Ontology) cung cấp một tập các thuật ngữ cụ thể cho những tác vụ cụ thể.
- Ontology lĩnh vực - tác vụ (Domain – Task Ontology) là các ontology về tác vụ có thể tái sử dụng trong một lĩnh vực nào đó.
- Ontology ứng dụng (Application Ontology)
- Ontology chỉ mục (Index Ontology)
- Ontology hỏi và trả lời (Tell and Ask Ontology) ...

Các loại metadata ontology, ontology lĩnh vực, ontology ứng dụng nắm giữ tri thức một cách tĩnh nghĩa là độc lập với cách giải quyết vấn đề, trong khi ontology tác vụ, ontology lĩnh vực– tác vụ liên quan đến tri thức giải quyết vấn đề. Tất cả các ontology này có thể kết hợp với nhau để xây dựng lên một ontology mới.

Ngoài ra, cộng đồng nghiên cứu phân biệt các ontology dựa trên độ phức tạp của mô hình biểu diễn ontology.

- Lightweight ontology: chứa các khái niệm, phân cấp khái niệm, mối quan hệ giữa các khái niệm và các thuộc tính mô tả khái niệm.
- Heavyweight ontology: bổ sung vào lightweight ontology các tiên đề, hàm và ràng buộc.

2.2.4. Vai trò của Ontology

Nhu cầu ban đầu cần có ontology là để cung cấp các nguồn thông tin giàu ngữ nghĩa mà máy tính có thể xử lý và thao tác được, đồng thời vẫn có thể dùng ontology để chia sẻ tri thức giữa người với người và với các hệ thống khác. Sự giao tiếp giữa con người với nhau, giữa con người với hệ thống cũng như giữa các hệ thống với nhau cần có sự chia sẻ hiểu biết chung. Thật vậy, mỗi một hệ thống đều có một hệ thống các khái niệm và thuật ngữ riêng, cấu trúc và phương pháp khác nhau hoặc có thể cùng một khái

niệm, cùng một quan hệ nhưng lại được hiểu theo các ngữ cảnh khác nhau hoặc biểu diễn theo các cách khác nhau. Do đó, nếu không có sự hiểu biết chung thì giao tiếp sẽ trở nên nghèo nàn, khó xác định yêu cầu, khó đặc tả hệ thống, khả năng liên kết giữa các hệ thống bị giới hạn, tính tái sử dụng và chia sẻ thấp, cần nhiều chi phí cho việc xây dựng và liên kết các hệ thống. Hơn nữa, việc phát triển các hệ thống thông minh đòi hỏi miền tri thức chung về các sự vật và phân loại chúng càng đóng vai trò then chốt trong hoạt động suy diễn. Do đó, các tri thức này cần phải cho vào một cơ chế thông minh và dễ hiểu, cho phép giảm thiểu tối đa sự nhầm lẫn, xung đột giữa các khái niệm, cung cấp cơ sở ngữ nghĩa tiến tới chia sẻ hiểu biết chung. Ontology chính là một cơ chế như vậy với các chức năng sau:

- Chia sẻ sự hiểu biết chung giữa các ứng dụng và con người, hiểu biết về cấu trúc thông tin giữa con người và các tác tử.
- Cho phép sử dụng lại tri thức. Ví dụ, nếu một nhóm nghiên cứu đã phát triển các ontology, nhóm khác có thể sử dụng lại cho ứng dụng của họ.
- Làm rõ lĩnh vực quan tâm, đưa ra các giả thiết rõ ràng về miền: tạo điều kiện thay đổi khi tri thức về lĩnh vực thay đổi, các đặc tả rõ ràng về miền tri thức sẽ giúp cho người mới dễ tìm hiểu ngữ nghĩa của các từ trong lĩnh vực quan tâm
- Phân tách hay tách rời tri thức lĩnh vực với tri thức xử lý: có thể hình dung 1 tác vụ tạo một tài liệu học tập từ nhiều thành phần theo đặc tả thì độc lập với chương trình ứng dụng làm nhiệm vụ này.
- Phân tích tri thức: Phân tích hình thức của các khái niệm, cần thiết cho việc tái sử dụng và mở rộng ontology. Muốn kế thừa hay sử dụng một ontology ta phải phân tích và tìm hiểu các khái niệm và quan hệ giữa chúng trong ontology đó.

Theo Aldea, các ontology có khả năng:

- Cung cấp một cấu trúc để chú giải nội dung của một tài liệu với thông tin ngữ nghĩa, điều này cho phép trích chọn thông tin thích hợp từ những tài liệu đó.
- Tích hợp thông tin từ nhiều nguồn khác nhau nhờ cung cấp một cấu trúc cho tổ

chức của nó và tạo thuận lợi cho trao đổi dữ liệu, tri thức và các mô hình.

- Đảm bảo sự đồng nhất và chính xác nhờ công thức hóa các ràng buộc nội dung của thông tin.
- Tạo các thư viện của các mô hình có khả năng trao đổi và tái sử dụng.
- Cho phép lập luận, nghĩa là cho phép tiến triển từ xử lý cú pháp đến xử lý ngữ nghĩa và cho phép các hệ thống suy diễn về các đối tượng dựa trên các luật sinh tổng quát.

2.2.5. Các ứng dụng dựa trên Ontology

Hiện nay nhu cầu về ontology ngày càng tăng cao và ontology không những phục vụ cho nhu cầu chia sẻ tri thức đơn thuần mà còn được áp dụng vào nhiều lĩnh vực khác nhau như các hệ thống quản lý tri thức, rút trích thông tin, thương mại điện tử, web ngữ nghĩa, xử lý ngôn ngữ tự nhiên, cơ sở dữ liệu, quản lý thông tin đa ngôn ngữ, khai phá tri thức, học máy, trong công nghệ phần mềm, trong kiến trúc đa tác tử hay trong các hệ thống bảo mật, ... Ontology cung cấp nguồn thông tin giàu ngữ nghĩa giúp cho các hệ thống thực hiện các tác vụ với kết quả tốt hơn.

Ontology được tổ chức W3C đưa vào làm một trong những nền tảng xây dựng Web Ngữ Nghĩa. Web ngữ nghĩa được định nghĩa như là sự mở rộng của Web hiện tại bằng cách thêm vào các mô tả ngữ nghĩa của thông tin dưới dạng mà chương trình máy tính có thể “hiểu” trong đó thông tin được định nghĩa rõ ràng, giúp cho máy tính và con người cộng tác làm việc tốt hơn và do đó các ứng dụng Web có thể xử lý thông tin hiệu quả hơn. Việc phát triển ontology dựa trên mục đích muốn cải thiện việc tìm kiếm trên Web vốn chỉ dựa trên việc duyệt và tìm kiếm theo từ khóa, ontology được dùng để gán nhãn lại các trang web, các web service hay các nguồn dữ liệu khác trên internet nhằm tăng tính hiệu quả trong việc truy xuất, tìm kiếm và khám phá dữ liệu.

Trong tiến trình khai phá dữ liệu hay tích hợp dữ liệu, việc ứng dụng ontology mang lại nhiều lợi thế, chẳng hạn như đối với các hệ thống bao gồm nhiều nguồn cơ sở dữ liệu khác nhau (khác nhau về cách thức lưu trữ và nội dung thông tin), mỗi nguồn dữ

liệu sẽ có một ontology mô tả về nó. Các ontology đó sẽ được hợp nhất vào một ontology chung và khi người dùng đưa ra yêu cầu thì hệ thống sẽ chuyển truy vấn đến nguồn cơ sở dữ liệu tương ứng.

Trong Thương mại điện tử, ontology được sử dụng để mô tả các sản phẩm khác nhau và được ứng dụng vào việc định vị và tìm kiếm sản phẩm tự động với các thông tin có sẵn. Ở đây ontology đóng vai trò chuẩn hóa các nhóm mặt hàng. Ngoài ra, ontology còn có công dụng giúp cho các hệ thống tự động giao tiếp với nhau dễ dàng. Các trang web hoạt động như là cổng thông tin chung, có nhiệm vụ thực hiện các biến đổi trên ontology giữa bên bán và bên mua.

Hiện nay đã có nhiều hệ thống hỗ trợ giáo dục được xây dựng theo cách tiếp cận sử dụng ontology và các công nghệ Web có ngữ nghĩa. Dựa trên các tính năng của hệ thống mà ta có thể phân loại chúng thành ba nhóm chủ yếu sau:

- Các hệ thống chia sẻ tài nguyên giáo dục trực tuyến: GEM - Gateway to Educational Materials (thegateway.org), Connexions (cnx.rice.edu).
- Các mạng chia sẻ ngang hàng về tài nguyên giáo dục: POOL - Portal for Online Objects in Learning , Edutella (www.edutella.org).
- Các hệ thống Elearning dựa trên ontology: PIP - Personalized Instruction Planner (peonto.cityu.edu.hk), TANGRAM (iis.fon.bg.ac.yu/TANGRAM).

Trong các hệ thống hỗ trợ giáo dục, ontology được sử dụng chủ yếu cho 3 mục đích: (i) biểu diễn và lưu trữ tri thức về các lĩnh vực cũng như các đối tượng cần thiết trong ứng dụng; (ii) xây dựng các mô hình tổ chức lưu trữ, biểu diễn ngữ nghĩa, biểu diễn tài liệu, lập chỉ mục cho các tài liệu (iii) xây dựng các chiến lược tìm kiếm theo ngữ nghĩa liên quan đến nội dung tài liệu.

2.2.6. Các hướng tiếp cận xây dựng ontology

Do nhu cầu ontology ngày càng phát triển, nên nhiều phương pháp khác nhau để xây dựng ontology một cách tự động hoặc bán tự động được các tác giả nghiên cứu và phát triển. Các phương pháp này giúp giảm bớt chi phí về thời gian và công sức so với

việc xây dựng các ontology một cách thủ công. Nhưng mặt khác chất lượng của các ontology thu được từ những phương pháp này phụ thuộc khá nhiều tùy vào thuật giải được sử dụng, nguồn dữ liệu mà thuật giải sử dụng, cũng như từng lĩnh vực mà phương pháp được áp dụng vào.

Một trong những hướng xây dựng ontology chính là rút trích ontology từ các nguồn dữ liệu khác nhau. Các phương pháp rút trích ontology sử dụng nhiều cách thức khác nhau từ các phương pháp máy học, xử lý ngôn ngữ tự nhiên cho đến thống kê. Các phương pháp sử dụng việc xử lý ngôn ngữ tự nhiên dựa trên việc phân tích từ vựng, cú pháp của tập hợp các văn bản thuộc về một domain nào đó, từ đó rút trích ra các khái niệm và dựa vào mối quan hệ cú pháp và từ vựng để xây dựng nên mối quan hệ về mặt ngữ nghĩa giữa các khái niệm. Phương pháp rút trích ontology dựa vào việc thống kê sẽ tiến hành thống kê trên các nguồn dữ liệu để rút trích ontology. Các phương pháp sử dụng việc học máy sẽ khai thác các nguồn dữ liệu nhằm rút ra các đặc trưng của dữ liệu, các khuôn mẫu cũng như các tập luật phục vụ cho việc rút trích ontology.

Một trong những hướng tiếp cận đáng quan tâm là rút trích ontology từ dữ liệu web. Các nguồn dữ liệu được dùng trong việc rút trích ontology khá đa dạng, từ dữ liệu dạng văn bản, dữ liệu quan hệ trong các cơ sở dữ liệu quan hệ, cho đến dữ liệu từ web. Trong đó nguồn dữ liệu từ web có lợi thế là nguồn thông tin phong phú, đa dạng và có sẵn trên internet.


Các hệ thống xây dựng ontology có thể sử dụng dữ liệu từ nhiều nguồn khác nhau để xây dựng nên ontology, có thể được phân chia thành các loại sau đây:

- Dữ liệu có cấu trúc: Hệ thống xây dựng lên các ontology dựa vào các dữ liệu có cấu trúc như từ database schema, từ những ontology đã có sẵn, từ những cơ sở tri thức và từ các mạng từ vựng như WordNet.
- Dữ liệu bán cấu trúc: đây cũng là một nguồn khác mà các hệ thống thường sử dụng, bao gồm các từ điển, các văn bản HTML và XML.
- Dữ liệu không có cấu trúc: đây là nguồn dữ liệu khó rút trích tri thức nhất. Các

hệ thống xây dựng ontology phải thực hiện các công đoạn xử lý ngôn ngữ tự nhiên trên các văn bản này để khám phá ra các khái niệm và các quan hệ. Dữ liệu dạng này bao gồm các văn bản viết trên ngôn ngữ tự nhiên hoặc các văn bản lấy từ web.

2.3. CÁC PHƯƠNG PHÁP TÍNH KHOẢNG CÁCH NGỮ NGHĨA GIỮA CÁC KHÁI NIỆM

Có nhiều phương pháp tính độ đo tương đồng ngữ nghĩa hay khoảng cách ngữ nghĩa giữa các khái niệm đã được đề xuất. Dựa vào số lượng tri thức mà hệ thống giả định trước cho việc tính toán các độ đo, người ta phân loại các phương pháp này theo hai hướng tiếp cận chủ yếu như [5]:

 **Hướng tiếp cận dựa trên kho ngữ liệu**, còn gọi là phương pháp nghèo tri thức (knowledge-poor)

Với cách tiếp cận này người ta tiến hành khảo sát và thống kê các mối liên hệ giữa các từ có trong kho ngữ liệu (corpus) để xác định độ đo. Kho ngữ liệu càng lớn thì độ chính xác càng cao. Ý tưởng chính là những từ giống nhau sẽ được sử dụng trong các ngữ cảnh giống nhau và ngược lại ngữ cảnh giống nhau sẽ sử dụng các từ giống nhau và nếu hai từ thường cùng xuất hiện thì chắc chắn tồn tại mối quan hệ ngữ nghĩa giữa chúng. Do đó, tần suất xuất hiện của các từ và phân bố của sự đồng hiện của các từ trong các ngữ cảnh khác nhau sẽ được sử dụng để đánh giá, ước lượng khoảng cách ngữ nghĩa giữa các từ. Các từ sẽ được so sánh với nhau về mặt phân bố ngữ cảnh của chúng. Các từ cùng chia sẻ một số lượng lớn ngữ cảnh thì được xem là giống nhau. Một cách tính độ đo dựa theo hướng tiếp cận này là thực hiện chọn một nhóm các từ làm các từ đặc trưng (có thể bằng kỹ thuật thống kê). Sau đó, ngữ cảnh cục bộ của mỗi từ sẽ sinh ra vector đặc trưng của nó. Khi đó, mỗi từ được đại diện bởi một vector mà mỗi thành phần của vector là số lần mà từ đó đồng xuất hiện với từ khác cho trước có trong một tập ngữ liệu. Cuối cùng, độ giống nhau giữa các từ được tính bằng cách sử dụng phép tính khoảng cách vector. Các ngữ cảnh được phân chia theo hai cách khác

nhau, do đó, tiếp cận này cũng được phân chia làm hai kỹ thuật khác nhau: Kỹ thuật dựa trên các cửa sổ (windows-based techniques) và Kỹ thuật dựa trên cú pháp (syntactic-based techniques).

Cách tiếp cận này không sử dụng tri thức được giả định trước cho việc tính toán, nghĩa là không đòi hỏi phải có trước miền tri thức, không có thông tin ngữ nghĩa hay các tài nguyên tĩnh như từ điển, từ điển đồng nghĩa đi kèm theo, ... Các tiếp cận dựa trên kho ngữ liệu cho phép tự do về mặt tri thức, không phụ thuộc vào miền tri thức đang sử dụng, tuy nhiên mối liên hệ về mặt ngữ nghĩa khác nhau giữa các từ lại không được xét đến do đó giá trị tính toán được không phản ánh chính xác sự khác biệt về khoảng cách ngữ nghĩa vốn có giữa các khái niệm. Hướng tiếp cận dựa trên kho ngữ liệu mặc dù được hỗ trợ bởi các công cụ toán học mạnh mẽ nhưng vẫn có một số thiếu sót khi đụng chạm đến việc xử lý một số khía cạnh sâu hơn của ngôn ngữ. Ví dụ như không tìm được độ tương tự ngữ nghĩa giữa hai khái niệm "picture" và "photograph", nhưng ngược lại điều này có thể xác định dễ dàng khi tiếp cận theo hướng ontology. Ngoài ra, hầu hết các kho ngữ liệu có sẵn chưa được gán nhãn từ loại do đó chỉ có khả năng tìm được độ liên quan giữa các từ và không xác định được độ liên quan giữa các nghĩa của chúng. Hậu quả là các quan hệ giữa các nghĩa của từ có tần suất thấp sẽ không được xem xét trong các phương pháp thống kê. Một vấn đề nghiêm trọng khác là tính thiếu đầy đủ, thậm chí ngay cả trong những kho ngữ liệu lớn như BNC cũng chưa chắc chứa hết các từ ngữ tiếng Anh.

 **Hướng tiếp cận dựa trên ontology**, còn gọi là phương pháp giàu tri thức (knowledge-rich)

Khác với hướng tiếp cận dựa trên kho ngữ liệu, hướng tiếp cận dựa trên ontology sử dụng tất cả các tri thức ngữ nghĩa được định nghĩa trước và còn được gọi là cách tiếp cận dựa trên tài nguyên từ vựng (lexical resource based). Trong cách tiếp cận này, các tài nguyên từ vựng được xây dựng thành một mạng hoặc một đồ thị có hướng. Sự giống nhau giữa các khái niệm sẽ được tính dựa trên các tính chất của các đường nối giữa các

khái niệm có trong đồ thị.

Cách tiếp cận này có thể mắc phải nhiều hạn chế do quá phụ thuộc vào những tài nguyên từ vựng, trong khi những tài nguyên này vốn được xây dựng một cách thủ công bởi con người theo ý kiến chủ quan nên dễ dẫn tới nhiều trường hợp thiếu sót hay dư thừa như lượng từ vựng bị giới hạn, có những từ vựng được định nghĩa trong tài nguyên là không cần thiết hoặc thiếu những từ vựng quan trọng, có ý nghĩa trong miền tri thức đang khảo sát, ... Cho dù người thiết kế có quan tâm hay không đến việc sẽ xây dựng một ontology lớn thì cũng chỉ có hy vọng áp dụng trong những lĩnh vực đặc thù. Đồng thời ontology chỉ xây dựng dựa trên các khái niệm nào sẽ được thể hiện trong lĩnh vực đó. Những thiếu sót này sẽ dẫn đến những “lỗ hổng” và bất cân bằng trong ontology; chúng sẽ dẫn đến những sai lầm to lớn của các hệ thống suy diễn tự động. Ngoài ra, tiêu chuẩn phân loại, phân lớp các từ có thể không rõ ràng, cách phân loại kém và không cung cấp đủ sự phân biệt giữa các từ, hoặc đôi khi lại cung cấp quá chi tiết không cần thiết và trên hết là đòi hỏi nhiều công sức của con người nhằm tạo ra danh sách lớn các từ đồng nghĩa, gần nghĩa, các quan hệ phân cấp hay có liên quan khác một cách thủ công. Và một nhược điểm khác là phụ thuộc vào quan điểm chủ quan trong việc tính toán khoảng cách ngữ nghĩa giữa các từ hay các khái niệm. Tuy nhiên, cách tiếp cận dựa trên các ontology được xem là cách tiếp cận hiện đại và phù hợp nhất cho biểu diễn và xử lý ngữ nghĩa và các tài nguyên tri thức của ontology vẫn là những tài nguyên hết sức có giá trị. Nếu những tài nguyên từ vựng hay các ontology được xây dựng tốt, mô tả được tương đối đầy đủ tri thức của lĩnh vực thì việc sử dụng chúng sẽ làm tăng độ chính xác và khả năng vét cạn trong quá trình tính toán các độ đo ngữ nghĩa cũng như tìm kiếm thông tin. Hơn nữa, các độ đo khoảng cách ngữ nghĩa giữa các từ của cách tiếp cận dựa trên ontology thì đơn giản, trực quan và dễ hiểu.

Hiện nay, cách tiếp cận này được chia thành các hướng:

- **Dựa trên từ điển đơn ngữ**

Một từ điển đơn ngữ sẽ được chuyển thành một mạng bằng cách tạo một nút cho

mỗi đầu mục từ trong từ điển (headword) và liên kết mỗi nút với các nút khác cho tất cả các từ có sử dụng trong định nghĩa của nó. Độ giống nhau giữa các từ được tính bằng sự lan tỏa trên mạng này. Cách tiếp cận này hoạt động kém hiệu quả, tuy nhiên đây là một phương pháp có thể áp dụng dễ dàng cho hầu hết các ngôn ngữ tự nhiên do nó chỉ cần sử dụng từ điển đơn ngữ của ngôn ngữ đó.

Năm 1986, Lesk đã đề xuất phương pháp dùng các từ điển như một tài nguyên để xác định độ tương tự giữa các khái niệm. Theo Leck, các ngữ nghĩa của các khái niệm trong một văn bản cho trước đã được ước đoán dựa vào việc đếm sự chồng lấp giữa các định nghĩa trong từ điển của các ngữ nghĩa đó. Năm 2003, hướng tiếp cận của Lesk đã được Banerjee S. và Pedersen T. mở rộng thêm và đã dùng WordNet như một tài nguyên từ vựng.

▪ **Dựa trên mạng phân cấp ngữ nghĩa:**

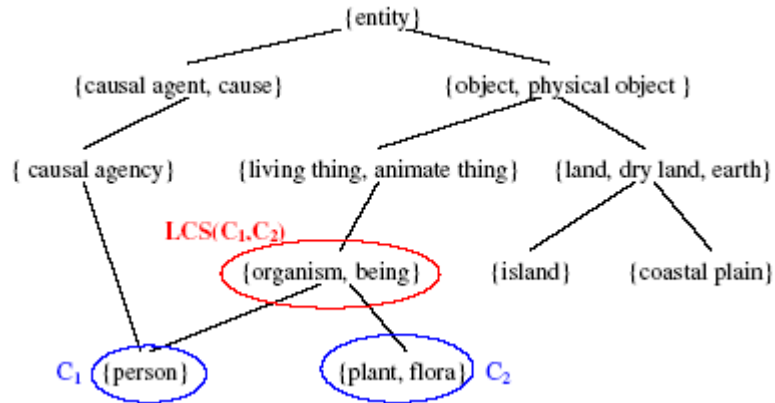
Hầu hết các phương pháp dựa trên mạng phân cấp ngữ nghĩa đều sử dụng WordNet để thực hiện việc nghiên cứu. WordNet là một từ điển điện tử miễn phí chứa một số lượng lớn các danh từ, động từ, tính từ và trạng từ tiếng Anh. WordNet tổ chức các khái niệm có liên quan nhau thành các tập từ đồng nghĩa gọi là synsets. Và giữa các tập đồng nghĩa này có thể mang các mối quan hệ ngữ nghĩa với nhau. Như vậy, ngoài việc cung cấp các nhóm từ đồng nghĩa để biểu diễn khái niệm, WordNet kết nối các khái niệm bởi một tập các quan hệ. Điều này tạo nên một mạng các khái niệm giúp chúng ta có thể xác định các khoảng cách ngữ nghĩa giữa chúng với nhau.

Khoảng cách ngữ nghĩa giữa hai khái niệm được tính dựa trên cách đếm đếm số cạnh hay nút dọc theo con đường ngắn nhất nối giữa các khái niệm. Một số độ đo tương tự ngữ nghĩa giữa hai khái niệm bất kỳ được đề xuất như sau:

Công thức đơn giản nhất là $sim(c_1, c_2) = \frac{1}{dist(c_1, c_2)}$, trong đó $dist(c_1, c_2)$ là số nút

trên đường nối ngắn nhất giữa hai khái niệm c_1 và c_2 . Nếu không có đường nối thì độ đo không xác định. Hạn chế của các phương pháp tính này là trong WordNet có thể sinh ra

các khoảng cách ngữ nghĩa khác nhau giữa hai synset liên kết trực tiếp nhau (nghĩa là có cạnh nối giữa hai synset), có một số liên kết có thể thể hiện một khác biệt lớn về nghĩa trong khi có các liên kết khác chỉ có sự phân biệt rất nhỏ. Đặc biệt các liên kết nằm ở mức cao trong phép phân loại (gần với nút gốc) thể hiện khoảng cách ngữ nghĩa lớn hơn, các liên kết ở mức thấp thể hiện khoảng cách ngữ nghĩa nhỏ hơn. Ví dụ trong mạng phân cấp hình 2.3, khoảng cách ngữ nghĩa giữa synset {object, physical object} với {land, dry land, earth} thì lớn hơn so với {land, dry land, earth} và {island}



Hình 2.2. Ví dụ mạng phân cấp trong WordNet

❖ Độ đo của Sussna

Nhằm khắc phục hạn chế trên, Sussna đã đưa ra một phương pháp tính với ý tưởng là “các khái niệm anh em ở sâu bên dưới trong sự phân loại từ thì gần nghĩa nhau hơn những khái niệm anh em nằm ở trên” (Hai khái niệm c_1 và c_2 trong mạng phân cấp được gọi là anh em nếu như nó có cùng một khái niệm cha chung). Sussna phân tích mỗi cạnh nối hai nút liền kề c_1 và c_2 trong mạng danh từ WordNet tương ứng với hai cạnh có hướng biểu diễn các quan hệ ngược nhau. Mỗi quan hệ như vậy được gán một trọng số có giá trị nằm trong khoảng $[\min_r; \max_r]$. Trọng số của mỗi cạnh có hướng thuộc một quan hệ r xuất phát từ một nút c được xác định bởi một hệ số phụ thuộc vào tổng số cạnh có cùng loại quan hệ r xuất phát từ c :

$$wt(c \rightarrow r) = \frac{\max_r - \min_r}{edges_r(c)}$$

Khi đó, khoảng cách giữa hai nút liên kề c_1 và c_2 được định nghĩa như sau:

$$dist(c_1, c_2) = \frac{wt(c_1 \rightarrow r) + wt(c_2 \rightarrow r')}{2 \times \max\{depth(c_1), depth(c_2)\}}$$

trong đó, r là mối quan hệ giữa c_1 và c_2 và r' là chiều ngược lại, $depth(c)$ là tổng số nút dọc theo con đường ngắn nhất từ c đến nút gốc trong cây phân cấp.

Cuối cùng, khoảng cách ngữ nghĩa giữa hai nút c_i và c_j là tổng khoảng cách giữa các cặp các nút liên kề dọc theo con đường ngắn nhất nối giữa chúng.

Nhược điểm của phương pháp này là khá phức tạp, hiệu quả chúng đem lại không tương xứng với chi phí phải bỏ ra trong quá trình tính toán.

❖ Độ đo của Wu và Palmer

Công thức tính độ giống nhau về ngữ nghĩa giữa hai khái niệm c_1, c_2 trong mạng phân cấp được Wu và Palmer đưa ra như sau:

$$sim_{WP}(c_1, c_2) = \frac{2 \times depth(LCS(c_1, c_2))}{len(c_1, LCS(c_1, c_2)) + len(c_2, LCS(c_1, c_2)) + 2 \times depth(LCS(c_1, c_2))}$$

trong đó $LCS(c_1, c_2)$ là khái niệm chung thấp nhất của hai khái niệm c_1 và c_2 trong cây phân cấp ngữ nghĩa, $depth(c)$ là tổng số nút dọc theo con đường ngắn nhất từ c đến nút gốc và $len(c_i, c_j)$ là tổng số nút dọc theo con đường ngắn nhất từ c_i đến c_j .

❖ Độ đo của Rensik

$$sim_{RWP}(c_1, c_2) = \frac{2 \times depth_{edge}(LCS(c_1, c_2))}{depth_{edge}(c_1) + depth_{edge}(c_2)}$$

trong đó, $depth_{edge}(c)$ là khoảng cách từ c đến nút gốc dùng cách đếm cạnh.

Kết hợp từ hai phương pháp trên, một công thức khác được đề xuất:

$$sim_{RWP}(c_1, c_2) = \frac{2 \times depth_{node}(LCS(c_1, c_2))}{depth_{node}(c_1) + depth_{node}(c_2)}$$

$depth_{node}(c)$ là khoảng cách từ c đến nút gốc dùng cách đếm nút.

❖ Độ đo của Leacock và Chodorow

Cũng tương tự như độ đo của Wu và Palmer, Rensik, phương pháp của

Leacock và Chodorow cũng dựa trên chiều dài của con đường ngắn nhất giữa hai khái niệm trong WordNet, tuy nhiên, công thức được cho ở một dạng khác:

$$sim_{LC}(c_1, c_2) = -\log \frac{len(c_1, c_2)}{2 \times \max_{c \in WordNet} depth(c)}$$

❖ Độ đo của Hirst và St-Onge

Các phương pháp trên chỉ xem xét đến mối quan hệ is-a cho danh từ trong WordNet. Hirst và St-Onge đã đưa ra một độ đo ngữ nghĩa bằng cách xét nhiều mối quan hệ khác trong WordNet và không giới hạn cho danh từ. Ý tưởng chính là hai khái niệm là gần nhau về ngữ nghĩa nếu các tập đồng nghĩa của chúng trong WordNet được nối nhau bởi một con đường không quá dài và không thay đổi hướng quá thường xuyên.

$$rel_{HS}(c_1, c_2) = C - path_length - k \times d$$

trong đó, d là số lần thay đổi hướng trong con đường từ c_1 đến c_2 , C và k là những hằng số. Các hướng có thể là hướng lên, hướng xuống và hướng ngang. Một đường nối hướng lên tương ứng với một sự tổng quát hóa (hypernymy), một đường nối hướng xuống tương ứng với một đặc biệt hóa (hyponymy) và đường nối hướng ngang gồm tất cả các loại còn lại là meronymy, antonymy, holonymy, troponymy, ...



Hướng tiếp cận lai ghép

Đây là phương pháp lai ghép giữa khảo sát dựa trên kho ngữ liệu và các ontology bằng cách dựa trên sự kết hợp cấu trúc phân loại từ vựng với thông tin thống kê có từ kho ngữ liệu để tìm khoảng cách ngữ nghĩa giữa các nút thông qua những tính toán dẫn xuất từ sự thống kê phân bố của dữ liệu có trong kho ngữ liệu. Hướng tiếp cận này sử dụng khái niệm “lượng tin” trong lý thuyết thông tin. Mục tiêu là khắc phục tính không ổn định của các khoảng cách liên kết các khái niệm đã xuất hiện trong hướng tiếp cận dựa trên ontology, bằng cách bổ sung vào các thông số chuẩn hóa của lý thuyết thông tin.

❖ Độ đo của Resnik

Resnik đã kết hợp phương pháp dựa trên kho ngữ liệu và phương pháp dựa trên ontology để đưa ra một độ đo dựa trên một công thức về lượng tin Information Content. Lượng tin là một giá trị được gán cho mỗi khái niệm trong mạng phân cấp dựa trên những tính toán tìm được từ kho ngữ liệu. Ý tưởng chính là sự giống nhau của hai khái niệm là khả năng mà chúng chia sẻ thông tin dùng chung và lượng thông tin chung của hai khái niệm được xác định bởi lượng tin của khái niệm chung thấp nhất trong mạng phân cấp ngữ nghĩa mà bao phủ cả hai khái niệm đó. Công thức tính độ đo được định nghĩa như sau:

$$sim_R(c_1, c_2) = -\log \Pr(LCS(c_1, c_2))$$

trong đó, $\Pr(c)$ là xác suất xuất hiện của khái niệm c trong kho ngữ liệu, được tính theo tần suất xuất hiện của các danh từ được lấy từ kho ngữ liệu Brown Corpus:

$\Pr(c) = \frac{\sum_{w \in W(c)} count(w)}{N}$, trong đó $W(c)$ là tập các danh từ trong kho ngữ liệu mà nghĩa của chúng được bao phủ trong khái niệm c , N là tổng số lượng danh từ có trong kho ngữ liệu mà cũng có trong từ điển WordNet.

Giới hạn của cách tiếp cận này là chỉ xem xét lượng tin của khái niệm chung thấp nhất của cả hai khái niệm cần đo mà không xem xét lượng tin của từng khái niệm cũng như không xem xét chiều dài đường đi giữa hai khái niệm đó, dẫn đến việc nhiều khái niệm có thể có cùng một khái niệm chung thấp nhất và có cùng giá trị cho độ giống nhau giữa chúng.

❖ Độ đo của Jiang và Conrath

Để giải quyết nhược điểm của Resnik, phương pháp của Jiang và Conrath đã đưa vai trò của các cạnh vào công thức tính khoảng cách ngữ nghĩa và sử dụng thông tin thống kê từ kho ngữ liệu để thực hiện việc tính toán. Ý tưởng then chốt của độ đo này là khoảng cách ngữ nghĩa của một liên kết nối một khái niệm c với cha của nó là $\text{par}(c)$ trong mạng phân cấp là lượng tin còn lại của khái niệm c mà không nằm trong $\text{par}(c)$.

Công thức tính độ đo khoảng cách ngữ nghĩa giữa hai khái niệm bất kì c_1 và c_2

trong mạng phân cấp được cho như sau:

$$dist_{JC}(c_1, c_2) = 2 \log \Pr(LCS(c_1, c_2)) - (\log \Pr(c_1) + \log \Pr(c_2))$$

trong đó, $\Pr(c)$ là xác suất xuất hiện của khái niệm c trong kho ngữ liệu, được xác định tương tự như $\Pr(c)$ của Resnik.

❖ Độ đo của Lin

Lin cho rằng tất cả các độ đo trên đều gắn liền với một ứng dụng, miền và một tài nguyên cụ thể. Dựa trên các giả thiết, định nghĩa và công cụ của lý thuyết thông tin, Lin đo sự giống nhau giữa hai đối tượng A và B tổng quát bằng tỉ số giữa lượng tin cần thiết để phát biểu tính chất chung giữa A và B và lượng tin cần thiết để mô tả chúng.

$$sim_L(A, B) = \frac{\log \Pr(comm(A, B))}{\log \Pr(descr(A, B))}$$

trong đó, $comm(A, B)$ là thành phần mô tả thông tin dùng chung giữa A và B , $descr(A, B)$ là thành phần mô tả A và B .

Dựa vào định nghĩa trên, độ giống nhau giữa hai khái niệm c_1 và c_2 trong một mạng phân cấp là một hệ quả của lý thuyết này:

$$sim_L(c_1, c_2) = \frac{2 \times \log \Pr(LCS(c_1, c_2))}{\log \Pr(c_1) + \log \Pr(c_2)}$$

trong đó, $\Pr(c)$ là xác suất xuất hiện của khái niệm c trong kho ngữ liệu, được xác định tương tự như $\Pr(c)$ của Resnik.

Budanitsky và Hist đã thực hiện việc nghiên cứu và tính toán thử nghiệm các độ đo của các công trình nghiên cứu trên WordNet dựa trên một ứng dụng xử lý ngôn ngữ tự nhiên cụ thể và dựa trên sự nhận xét của các chuyên gia về ngôn ngữ. Các kết quả nghiên cứu, so sánh và đánh giá cho thấy rằng độ đo ngữ nghĩa của Jiang-Conrath cho các kết quả tốt nhất, tiếp theo là Lin và Leacock-Chodorow, Resnik và sau đó mới đến Hist – St-Onge.