Kỷ yếu Hội nghị Quốc gia lần thứ VIII về Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR); Hà Nội, ngày 9-10/7/2015
DOI: 10.15625/vap.2015.000211

TIẾP CẬN MỚI VỀ ĐỐI SÁNH ONTOLOGY

Huỳnh Nhứt Phát[1], Hoàng Hữu Hạnh[1], Phan Công Vinh[2]
[1]Đại học Huế
[2]Trường Đại học Nguyễn Tất Thành TP HCM

huynhnhutphat@yahoo.com, hhhanh@hueuni.edu.vn, pcvinh@ntt.edu.vn

Tóm tắt *– Đối sánh ontology tạo điều kiện trao đổi kiến thức giữa các nguồn dữ liệu đa dạng. Các phương pháp tiếp cận đối sánh ontology sử dụng nhiều độ đo tương đồng cho các thực thể ánh xạ giữa các ontology. Tuy nhiên, nó vẫn còn là một thách thức trong việc xử lý với các thực thể không rõ ràng mà các độ đo đối sánh ontology được sử dụng, tạo ra các kết quả trái ngược nhau về sự tương đồng của các thực thể ánh xạ. Trong bài báo này, chúng tôi trình bày phương pháp tiếp cận mới OARS của chúng tôi, dựa trên các tập thô để đối sánh ontology, nó đạt được mức độ chính xác cao trong các tình huống phát sinh các thực thể không rõ ràng, do những kết quả trái ngược nhau được tạo ra bởi các độ đo tương đồng khác nhau. OARS sử dụng cách tiếp cận tổ hợp có tính toán đến độ đo tương đồng về từ vựng và cấu trúc. OARS thực hiện việc so sánh tốt nhất ở độ đo recall và độ đo precision với một số hệ thống đối sánh của tổ chức Ontology Alignment Evaluation Initiative (OAEI) 2010.*

Từ khóa *– Ontology alignment, Rough sets, semantic matching, semantic interoperability.*

I. GIỚI THIỆU

Các đối sánh ontology tạo thuận lợi để trao đổi kiến thức giữa các nguồn dữ liệu đa dạng. Một ontology là một đặc tả hình thức rõ ràng về các thuật ngữ trong một miền và các quan hệ giữa chúng. Hiện nay số lượng ontology phát triển, phổ biến là các ontology khác nhau với cùng tên miền duy nhất. Người thiết kế ontology có thể nghĩ tới những đối tượng khác biệt trong khi phát triển một ontology tùy thuộc vào nhu cầu ứng dụng của chúng. Việc xác định các đối tượng rõ ràng từ các ontology có sẵn sẽ là điều cần thiết cho phép đạt được kết quả tốt nhất với một tên miền cụ thể của việc chia sẻ kiến thức. Các ontology có thể đa dạng với nhiều hình thức khác nhau bao gồm cả sự đa dạng về thuật ngữ và đa dạng về khái niệm. Các hình thức đa dạng này phải được xử lý với một quá trình đối sánh ontology, nó đóng một vai trò quan trọng về khả năng trao đổi ngữ nghĩa giữa các ứng dụng. Quá trình đối sánh ontology sẽ tạo sự đối sánh giữa các thực thể có liên quan về ngữ nghĩa được xác định trong các ontology không đồng nhất, nó được phát triển với tên miền giống nhau.

Trong những năm gần đây, một số hệ thống đối sánh đã được đề xuất bao gồm các hệ thống tự động, bán tự động và ứng dụng cụ thể được phân tích trong [1] [2]. Sơ đồ các kỹ thuật đối sánh cũng nghiên cứu sâu bởi cộng đồng nghiên cứu khi quá trình đối sánh ontology đòi hỏi việc xác định các tương ứng giữa các thực thể liên quan về ngữ nghĩa. Trong quá trình đối sánh tự động, các thực thể được chọn để ánh xạ khi mức độ tương đồng về ngữ nghĩa được tìm thấy và sẽ loại bỏ các thực thể không tương đồng về ngữ nghĩa. Hầu hết các phương pháp đối sánh ontology so sánh những điểm tương đồng, sử dụng nhiều kỹ thuật cơ bản và các kết quả của các kỹ thuật này được tổng hợp bởi một loạt các chiến lược kết hợp [3]. Sự kết hợp của các kỹ thuật về từ vựng và cấu trúc sẽ cho ra toàn bộ sự tương đồng tốt hơn của một khái niệm được xác định trong một ontology. Mỗi kỹ thuật đối sánh riêng biệt được xử lý như một công cụ đối sánh và các kết quả của tất cả các công cụ đối sánh có thể được tổng hợp với nhiều cách khác nhau để hoàn thiện quá trình liên kết. Những phương pháp kết hợp này có thể sử dụng các kỹ thuật trung bình có trọng số hoặc các phương pháp xác suất để tính toán khả năng có thể xảy ra của một thực thể trong một ontology nguồn là tương đồng với một thực thể trong một ontology đích. Tuy nhiên, vấn đề thực sự phát sinh khi phương pháp tổ hợp với các thực thể không rõ ràng, chúng không hoàn toàn tương đồng bởi vì những kết quả trái ngược nhau được tạo ra bởi các công cụ đối sánh riêng biệt. Vì vậy, việc tìm kiếm các thực thể không rõ ràng và xử lý với các thực thể không rõ ràng là một nhiệm vụ phức tạp so với việc tìm kiếm thực thể chỉ tương đồng hoặc không tương đồng trong quá trình đối sánh ontology. Như vậy, các thực thể không rõ ràng đang trở nên phổ biến hơn khi phần thông tin về một khái niệm có sẵn trong một ontology đem so sánh với phần thông tin có sẵn của cùng khái niệm trong một ontology khác.

Bài báo này trình bày OARS, một cách tiếp cận mới về sự đối sánh ontology để xử lý đối với các thực thể không rõ ràng trong ánh xạ ontology. OARS xây dựng trên các tập Thô để tính toán sự tương đồng của các thực thể ontology trong quá trình liên kết. Trong OARS, đầu tiên các thực thể được đối sánh thông qua ba đối sánh cơ bản chúng dựa trên các cấu trúc, các chuỗi và các ngữ nghĩa tương ứng. Các thực thể đối với các công cụ đối sánh riêng biệt, không thể đạt được một ánh xạ nhất quán về sự tương đồng giữa chúng, sẽ được coi như những thực thể không rõ ràng và được xử lý bởi việc phân loại thành các tập Thô trong OARS. Các thực thể không được ánh xạ tạo ra từ ba đối sánh riêng biệt, được định nghĩa là các thuộc tính của các phần tử tương ứng của các tập Thô. OARS phân loại tập Thô là một tập các phần tử dựa trên các thuộc tính có sẵn và tính toán về việc phân loại các tập Thô để đạt được một quyết định ánh xạ trên các thực thể không rõ ràng.

OARS đã được đánh giá về tính toàn diện bằng cách sử dụng các ontology chuẩn của tổ chức Ontology Alignment Evaluation Initiative (OAEI) 2010, Và nó thực hiện tốt nhất ở khía cạnh độ đo recall khi so sánh với một số

hệ thống tham gia đối sánh của tổ chức OAEI. Ngoài ra, OARS cũng tạo ra một hiệu quả tương đương với độ đo precision.

Điều đáng chú ý là OARS được mở rộng từ hệ thống đối sánh được đề xuất ban đầu bởi [4] và được đánh giá tốt cùng với ba nhóm tập dữ liệu chuẩn. Quan trọng hơn, ý nghĩa của việc sử dụng các tập Thô như là một phương pháp tổng hợp cũng được đánh giá trong bài báo này. Hơn nữa, chúng tôi đã tích hợp OARS vào SemFARM được phát triển trước đây [5], một khung ứng dụng cung cấp một cơ chế tìm kiếm hiệu quả cho việc ghi nhớ và truy hồi file trên các thiết bị di động được kết nối thông qua Bluetooth. Sự tích hợp của OARS cho phép SemFARM sử dụng kiến thức của nhiều ontology khi tìm kiếm một file trên các thiết bị được giới hạn về tài nguyên trong môi trường mạng, nó dẫn đến độ chính xác cao trong việc truy hồi file.

Bài báo này có cấu trúc như sau. Phần I giới thiệu. Phần II các nghiên cứu liên quan về đối sánh ontology. Trong phần III, các độ đo tương đồng và quá trình đối sánh của OARS. Phần IV trình bày sự phân loại các tập Thô, nó xử lý các thực thể không rõ ràng trong ánh xạ ontology. Phần V đánh giá hiệu quả của OARS bằng cách sử dụng các ontology chuẩn của tổ chức OAEI 2010. Phần VI tích hợp OARS vào khung ứng dụng SemFARM để tăng cường việc truy hồi file trên các thiết bị di động, và Phần VII kết luận.

II. CÁC NGHIÊN CỨU LIÊN QUAN VỀ ĐỐI SÁNH ONTOLOGY

Trong những năm gần đây, một nghiên cứu quan trọng được tiến hành để xử lý quá trình đối sánh ontology. Trong phần này, chúng tôi xét các nghiên cứu liên quan về các phương pháp tiếp cận truyền thống, chúng ta không bàn đến vấn đề không rõ ràng trong quá trình ánh xạ và phương pháp tiếp cận mới sẽ quan tâm đến vấn đề này.

A. Các phương pháp tiếp cận truyền thống đối sánh ontology

Các phương pháp nghiên cứu này chủ yếu theo hai cách tiếp cận sau. Một cách tiếp cận sử dụng các công cụ đối sánh riêng lẻ để đối sánh với các thực thể ontology bằng cách so sánh thông tin về nhãn (tên) của chúng với các từ đồng nghĩa tương ứng. Thông thường WordNet được khai thác với cách tiếp cận như vậy. Ví dụ, khả năng tương đồng được thực hiện bởi Rodriguez và Egenhofer [6] dựa trên quá trình đối sánh, trong đó sử dụng bộ từ đồng nghĩa cùng với thông tin khác từ việc xác định ontology. Các đặc điểm khác của từ vựng cũng được khai thác để tìm các mối quan hệ giữa các thực thể như từ khái quát (hypernym), từ khu biệt (hyponym), từ từng phần (meronym) và từ toàn phần (holonym). Công cụ đối sánh riêng biệt dựa trên các hệ thống đối sánh chỉ thực hiện tốt trong việc đối sánh các ontology khi chúng có các cấu trúc bên trong và bên ngoài tương đồng. Việc sử dụng các kỹ thuật đối sánh về cấu trúc, sự so sánh được thực hiện giữa các thực thể dựa trên các đặc điểm về cấu trúc của chúng trong các ontology, dựa trên tập các thuộc tính, miền, các kiểu dữ liệu và số lượng phần tử. GMO là một ví dụ của công cụ đối sánh về cấu trúc trong đó có một tập các cặp được đối sánh ở đầu ra trong quá trình đối sánh và sử dụng đồ thị hai bên (bipartite graphs) để so sánh sự tương đồng về cấu trúc của các ontology khác nhau. Công cụ đối sánh V-Doc đo phạm vi các thuật ngữ của các thực thể tên miền về ý nghĩa của chúng trong mô hình không gian Vector. Tuy nhiên, bất kỳ kỹ thuật đối sánh trong sự tách biệt như GMO hoặc V-Doc vẫn chưa đầy đủ cho kết quả ánh xạ thích hợp. Vì lý do này, chúng tôi đưa ra cách tiếp cận OARS kết hợp chuỗi, ngữ nghĩa và các công cụ đối sánh dựa trên cấu trúc.

Một cách tiếp cận khác là tổng hợp một số công cụ đối sánh riêng lẻ để đối sánh ontology. Ví dụ, RiMOM sử dụng nhiều công cụ đối sánh để tìm ra sự tương đồng về từ vựng và về cấu trúc giữa các thực thể và quyết định mở rộng lý thuyết Bayes để ánh xạ chúng. Các công cụ đối sánh cơ bản được xem là các chiến lược tách biệt so sánh sự phân loại, các ràng buộc, các mô tả, các tên, các thể hiện và tên các đường dẫn trong quá trình ánh xạ. Ở đầu vào, người sử dụng cũng được phép cải thiện các ánh xạ trong quá trình liên kết. Việc nâng cấp phiên bản của RiMOM khai thác hầu hết các kiến thức về ontology có sẵn bằng cách sử dụng chúng thông qua một kỹ thuật lựa chọn và kết hợp tất cả các giá trị tương đồng bởi hàm xích ma, và sau đó khởi tạo một thuật toán cải tiến đối sánh để hoàn thiện quá trình liên kết. Tuy nhiên, các thiết lập thông số trong RiMOM được đánh giá cao phụ thuộc vào các bước tiền xử lý trong đó hai yếu tố giống nhau được so sánh trong các ontology và sau đó các trọng số được gán cho các yếu tố khác nhau để kết hợp các kết quả cuối cùng. Điều này có nghĩa là nếu hai ontology có điểm tương đồng về cấu trúc, giá trị cao hơn sẽ được gán trọng số cho sự tương đồng về cấu trúc trong việc kết hợp các kết quả cuối cùng. Do đó, việc ánh xạ của các thực thể ontology này có sự tương đồng khác sẽ gặp khó khăn vì các thông số giống nhau sẽ được sử dụng cho tất cả các thực thể. Trong OARS, chúng tôi sử dụng sự phân loại các tập Thô cho mỗi thực thể riêng biệt và việc giải quyết ánh xạ được thực hiện trên cơ sở thực thể mà không ảnh hưởng đến quyết định tổng thể của các ánh xạ khác.

Falcon-AO [3] sử dụng sự kết hợp về ngữ nghĩa, cấu trúc và sự phân vùng dựa vào các công cụ đối sánh trong quá trình ánh xạ. Falcon-AO dựa trên nghiên cứu đối sánh của V-Doc, I-Sub [7] và GMO. Falcon-AO cần đến sự kết hợp về tính tương đồng để kết hợp giá trị tương đồng tạo ra bởi mỗi công cụ đối sánh. Một tập các luật kết hợp được sử dụng để giảm tính không đồng nhất về cấu trúc như là một quá trình trước khi ánh xạ. Các kết quả đối sánh được trả về để xác định các mối quan hệ tương đương giữa các lớp và các thuộc tính. Isaac đánh giá về tính hiệu quả của Falcon-AO trong việc sáp nhập từ điển đồng nghĩa, trong đó chủ yếu dựa vào thành phần về từ vựng của nó. Tuy nhiên, việc sử dụng ngữ nghĩa tương đồng, Falcon-AO không phân biệt giữa các thuộc tính kiểu dữ liệu và các thuộc tính đối tượng, trong khi OARS của chúng tôi sử dụng công cụ đối sánh ngữ nghĩa cho tính riêng biệt các lớp và các thuộc tính. Điều này tránh mọi khả năng của việc ánh xạ lớp thực thể của ontology này với thuộc tính thực thể của ontology khác.

ASMOV [7] là một công cụ đối sánh ontology tự động, nó sử dụng cả công cụ đối sánh về cấu trúc và từ vựng để tính toán sự tương đồng cho việc tích hợp ontology. ASMOV tự động hóa quá trình đối sánh bằng cách sử dụng trị trung bình có trọng số của các phép đo về sự tương đồng và nhận được một đối sánh lặp, sau đó nó được kiểm tra sự mâu thuẫn về ngữ nghĩa. Quá trình kiểm tra ngữ nghĩa sẽ xem xét các tương ứng phù hợp và không phù hợp. Nó cần thực hiện nhiều hơn để hoàn thành kết quả ánh xạ và các kết quả thực hiện trung gian giữa việc lặp đi lặp lại được sử dụng để cải thiện các giai đoạn xử lý tiếp theo của liên kết. Tuy nhiên, quá trình kiểm tra không đưa ra các luật hiệu quả cho các đối sánh chưa được kiểm tra.

Thuật toán SOBOM tìm các ràng buộc ở bước đầu tiên và sử dụng Semantic Inductive Similarity Flooding (SISF) để phủ kín sự tương đồng giữa các khái niệm. Sau đó, nó sử dụng các kết quả của SISF để tìm ra các mối quan hệ giữa các liên kết. Thuật toán SOBOM phụ thuộc nhiều về độ đo precision của các ràng buộc được trả về bởi việc đối sánh ngữ nghĩa, tức là việc thực hiện đối sánh tổng thể sẽ bị giảm nếu việc đối sánh mất khái niệm ràng buộc.

AgrMaker sử dụng ba lớp kiến trúc, trong đó bao gồm một số khái niệm và cấu trúc dựa vào các công cụ đối sánh. Nó kết hợp các kết quả bằng cách sử dụng độ đo lân cận đáng tin cậy. AgrMaker chủ yếu tập trung vào việc đưa ra các luật để kết hợp các tập ánh xạ khác nhau chứ không phải xác định việc đối sánh với chính nó. CODI sử dụng logic Markov dựa vào đối sánh theo xác suất mà biến đổi quá trình đối sánh thành một giải pháp tối ưu hóa Maximum-a-Posteriori. Nó kết hợp các độ đo tương đồng về từ vựng với thông tin lược đồ để đối sánh với các thực thể trong quá trình liên kết. Hiệu quả của CODI là phụ thuộc nhiều vào các ánh xạ tiền liên kết.

TaxoMap sẽ đưa vào lời giải thích mô tả các nhãn và lớp con của các ontology với sự đối sánh và sử dụng Partition dựa trên thuật toán Block Matching cho phép việc sử dụng các ánh xạ tương đương được xác định trước để phân vùng các ontology thành các cặp ánh xạ nếu có thể. MapPSO xem đối sánh ontology là một giải pháp tối ưu hóa và sử dụng thuật toán Discrete Particle Swarm Optimization để giải quyết vấn đề. Việc sử dụng phương pháp tiếp cận MapPSO, mọi tính chất được cập nhật và điều chỉnh việc lặp lại cho các tính chất miêu tả tốt nhất trong nhóm. Tuy nhiên, hiệu quả của MapPSO phụ thuộc vào việc lựa chọn các công cụ đối sánh có chất lượng và kết hợp lại.

Các hệ thống nói trên có giá trị nhất định trong việc đối sánh ontology, chúng chỉ xét đến các thực thể rõ ràng trong quá trình liên kết. Trong khi OARS, chúng tôi có xét đến các thực thể không rõ ràng như đề cập trong phần 1.

B. Phương pháp tiếp cận mới để đối sánh ontology

Hiện nay chỉ có một vài hệ thống đối sánh ontology đã đề cập đến sự không rõ ràng trong quá trình ánh xạ. Ví dụ, hệ thống đối sánh được đề xuất bởi [12] trong việc xử lý các thực thể không rõ ràng, nó sử dụng Lý thuyết Dempster-Shafer để tổng hợp các kết quả ánh xạ được tạo ra bởi các công cụ đối sánh riêng lẻ. Lý thuyết Dempster-Shafer cũng được sử dụng trong nghiên cứu để xử lý việc không rõ ràng trong ánh xạ ontology. Sváb và Svátek sử dụng mạng Bayes để mô phỏng các phương pháp ánh xạ và tổng hợp các kết quả ánh xạ. Để cho ra các kết quả ánh xạ thích hợp, các bảng phụ thuộc vào xác suất được xây dựng trong mạng Bayes cần phải đầy đủ thông qua một quá trình nghiên cứu. Pan trình bày Mạng Bayes dựa trên phương pháp xử lý sự không rõ ràng trong ánh xạ ontology. Các ontology nguồn và ontology đích, trước tiên được dịch sang các mạng Bayes. Sau đó, các ánh xạ của các khái niệm (các thực thể) giữa hai ontology được xử lý khi dựa trên khả năng suy luận giữa hai mạng Bayes. Phương pháp này dựa trên giả định mỗi khái niệm được đối sánh với khái niệm tương đương. Garruzzo và Rosaci trình bày một phương pháp với các miêu tả có ngữ nghĩa đồng nhất. Một tập các chú giải được sử dụng cho các miêu tả để giải quyết các thuật ngữ không rõ ràng trong việc trao đổi thông tin. Tuy nhiên, hiệu quả của phương pháp này phụ thuộc vào sự hoàn chỉnh của tập các chú giải. Hơn nữa, phương pháp này đòi hỏi các bước đối sánh về ngữ nghĩa giữa các miêu tả trong việc trao đổi thông tin.

OARS xây dựng dựa trên các tập Thô để xử lý với sự không rõ ràng trong đối sánh ontology. Khác với các phương pháp nói trên dựa vào lý thuyết Dempster Shafer và các mạng Bayes, lý thuyết các tập Thô không cần bất kỳ thông tin ban đầu hoặc thông tin bổ sung về dữ liệu, có nghĩa là lý thuyết các tập Thô là đối tượng trong việc xử lý thông tin như được đề cập bởi Li [9].

III. CÁC TIÊU CHUẨN ĐÁNH GIÁ SỰ TƯƠNG ĐỒNG

Có hai loại chính yếu của tính không đồng nhất là ngữ nghĩa và thuật ngữ. Không đồng nhất về ngữ nghĩa xảy ra do các nguyên nhân khác nhau như sử dụng các tiên đề khác nhau hay sự khác biệt về mô hình cùng khái niệm. Không đồng nhất về thuật ngữ xuất hiện khi sử dụng các từ đồng nghĩa hoặc các tên gọi khác nhau cho cùng một thực thể trong các ontology khác nhau. Để xử lý với hầu hết các loại không đồng nhất về ontology, OARS là cách tiếp cận tổ hợp và sử dụng các công cụ đối sánh về từ vựng và cấu trúc cùng với WordNet, như một tài nguyên mở rộng để tính toán sự tương đồng ngữ nghĩa giữa các thực thể. Có ba công cụ đối sánh riêng lẻ được sử dụng trong OARS dựa trên các kỹ thuật hiện nay. Để đối sánh hai ontology, một ontology nguồn O và một ontology đích O', OARS sử dụng ba công cụ đối sánh để tính toán sự tương đồng giữa các thực thể của O và O':

- Công cụ đối sánh dựa trên Chuỗi được sử dụng để tìm sự tương đồng giữa các lớp và các thực thể được đặt tên.
- Công cụ đối sánh dựa trên WordNet được sử dụng để so sánh sự tương đồng về ngữ nghĩa.
- Công cụ đối sánh dựa trên cấu trúc được sử dụng để so sánh các lớp cha và các lớp con có tính đến các ràng buộc để tìm sự tương đồng của các thuộc tính đối tượng và các thuộc tính dữ liệu của các lớp.

C. Sự tương đồng dựa trên chuỗi

Việc tính toán về tính tương đồng dựa trên chuỗi, các thực thể được xem là các chuỗi không phân biệt cấu trúc của chúng hoặc các thuộc tính liên quan khác. Quá trình chuẩn hóa chuỗi được thực hiện sau khi so sánh tên thực thể. Cả hai chuỗi thực thể được chuyển đổi thành chữ thường và các dấu chấm câu, các dấu gạch ngang và các ký tự trống được loại bỏ. Quá trình chuẩn hóa rất quan trọng trong việc so sánh chuỗi. Ví dụ như, "MasterThesis", "Master-Thesis" và "Master Thesis" được chuẩn hoá thành "masterthesis". Một số kỹ thuật được đề xuất để tính toán sự tương đồng về chuỗi bằng cách sử dụng những đặc trưng của các độ đo. Những kỹ thuật này bao gồm khoảng cách chuỗi con, Levenstein, Jaro-Winkler, Needleman-Wunsch và sự tương đồng n-gram. Một cuộc khảo sát tốt về sự tính toán khoảng cách chuỗi có thể được tìm thấy trong [10].

Stoilos [7] đề xuất chiều dài chuỗi Smoa (String Metric Ontology Alignment) dựa trên tính tương đồng. Smoa tính toán sự tương đồng về chuỗi dựa trên đặc tính chung của các chuỗi cũng như các khác biệt của chúng. Chiều dài Smoa được tính toán bằng cách trừ đi tổng của các khác biệt và tách sự tương đồng từ những điểm chung của các chuỗi. Những điểm chung được tính toán bằng cách sử dụng chiều dài chuỗi con.

Gọi Sim_strng biểu thị sự tương đồng về chuỗi giữa các thực thể e_i và e'_i, thì Sim_strng(e_i, e'_i) có thể được tính toán bằng cách sử dụng phương trình (1).

$$\text{Sim\_strng}(e_i, e'_i) = \text{Smoa}(e_i, e'_i) \quad (1)$$

Để tính toán chiều dài chuỗi con giữa hai chuỗi, một quá trình tìm kiếm và loại bỏ chuỗi con chung lớn nhất và quá trình được tiếp tục lặp lại cho đến khi không còn tìm thấy các chuỗi con chung nữa. Độ dài của các chuỗi con này được tính toán và có thể so sánh với độ dài của các chuỗi. Sự khác biệt được sử dụng trong Smoa là tính với chiều dài của các chuỗi không được đối sánh. Phép đo Smoa được sử dụng trong OARS như một công cụ đối sánh dựa trên chuỗi.

D. Tương đồng về ngữ nghĩa

Tương đồng dựa trên ngữ nghĩa được tính toán và sử dụng các nguồn tài nguyên bên ngoài như các từ điển về ngữ nghĩa, tập từ điển chuyên ngành hoặc cơ sở dữ liệu cụ thể. Như vậy tính tương đồng hữu ích khi các điểm tương đồng dựa trên chuỗi gặp khó khăn giữa các thực thể ontology, đặc biệt là khi các từ đồng nghĩa sử dụng cho cùng khái niệm trong các ontology. Ví dụ, các tên “brochure-tài liệu” và “booklet-tập sách” đề cập đến cùng khái niệm nhưng sự tương đồng dựa trên chuỗi giữa chúng là rất thấp (nó bằng 6, khi sử dụng khoảng cách Levenshtein), chúng phụ thuộc vào khả năng lựa chọn khái niệm ánh xạ. WordNet là một cơ sở dữ liệu về từ vựng nó cung cấp một kho lưu trữ của các mục từ vựng được định nghĩa như là một tập từ vựng về ngữ nghĩa. Trong WordNet, các nghĩa khác nhau của cùng khái niệm được nhóm lại với nhau như bộ từ đồng nghĩa về các danh từ, động từ, tính từ và trạng từ. Bộ từ đồng nghĩa được đối sánh với nhau trong một cấu trúc phân cấp sử dụng các mối quan hệ khác nhau về khái niệm ngữ nghĩa và từ vựng. Ví dụ, các danh từ có mối quan hệ như hypernym (từ khái quát), hyponym (từ khu biệt), holonym (từ toàn phần), meronym (từ từng phần) giữa các từ. Các động từ tương đồng được đối sánh thông qua các mối quan hệ của các thuật ngữ hypernym (từ khái quát), troponym (từ chuyển nghĩa), entailment (từ kế thừa) và coordinate (từ phối hợp). Nếu xét ví dụ về tên hai thực thể “brochure” và “booklet”, chúng sẽ được xem là khái niệm tốt cho việc ánh xạ trong WordNet, trong đó brochure (sách mỏng), folder (tài liệu), leaflet (tờ rơi) và pamphlet (cuốn sách nhỏ) được xác định là các từ đồng nghĩa.

Đối với sự tương đồng về ngữ nghĩa, các độ đo dựa trên ngữ cảnh cũng được sử dụng. Ví dụ, Sahami định nghĩa một hàm mới để đo sự tương đồng về ngữ nghĩa giữa các cặp đoạn văn ngắn bằng cách sử dụng các vector ngữ cảnh. Banerjee đo sự tương quan về ngữ nghĩa của các khái niệm bằng cách sử dụng sự phân cấp của các khái niệm được trình bày trong cơ sở dữ liệu về từ vựng như WordNet. Tương tự, Patwardhan và Pedersen sử dụng thông tin cùng sự kiện, cùng với các định nghĩa WordNet để xây dựng các vector chú thích với mỗi khái niệm tương ứng và được gán điểm số cho mỗi cặp khái niệm bằng cách tính cosin của góc giữa các vectơ chú thích tương ứng của chúng.

OARS sử dụng WordNet để khai thác thông tin được mã hóa theo tên và nhãn của các thực thể ontology. Việc sử dụng WordNet, chúng tôi xét các từ synonyms (từ đồng nghĩa), hyponyms (từ khu biệt), hypernyms (từ khái quát) và antonyms (từ trái nghĩa) của các thực thể.

Gọi

- Sim_lin(w_i, $w_{i'}$) là sự tương đồng về ngữ nghĩa giữa các từ w_i và $w_{i'}$,
- Σ là nguồn tài nguyên bên ngoài (WordNet),
- $s(w_i)$ là tập các từ đồng nghĩa,
- $h(w_i)$ là tập của các từ riêng biệt và các từ khái quát,
- $t(w_i)$ là tập các từ trái nghĩa với $w_{i'}$,

Sự tương đồng về ngữ nghĩa của hai từ w_i và w_i' có thể được tính toán bằng cách sử dụng phương trình (2).

$$Sim\_lin(w_i, w_i') = \begin{cases} 1 \text{ nếu } w_i' \in s(w_i) \\ 0.5 \text{ nếu } w_i' \in h(w_i) \\ 0 \text{ nếu } w_i' \in t(w_i) \end{cases} \quad (2)$$

Các mối quan hệ tương đồng của các từ khu biệt và các từ khái quát được chọn là 0.5 và được tính toán trong việc đối sánh về cấu trúc bằng cách sử dụng phương trình (3), (4), (5) và (6). Đối với những từ là từ đồng nghĩa và trái nghĩa chúng sẽ được coi là tương đồng và không tương đồng tương ứng. Một nhược điểm của việc sử dụng các nguồn tài nguyên như WordNet là một số đối sánh phụ thuộc vào cùng một khái niệm. Để giải quyết vấn đề này, OARS sử dụng ba loại thông tin về cấu trúc trong việc đối sánh giữa các thực thể sẽ được mô tả trong phần sau.

E. Tương đồng về cấu trúc

Thông tin về cấu trúc đóng một vai trò quan trọng trong các tình huống, trong đó tính tương đồng dựa trên ngữ nghĩa và chuỗi giữa hai thực thể ontology đối sánh nhau, được chứng minh là thiếu hoặc không đầy đủ. Ví dụ, Sánchez sử dụng các cấu trúc ontology để cải thiện tính chính xác của mô hình phân loại kiến thức. Trong [13], Sánchez cũng xem xét một số độ đo tương đồng về cấu trúc bao gồm các độ đo tương đồng dựa trên các lớp cha. Các lớp con cũng được xem xét với cấu trúc dựa trên các độ đo tương đồng giữa các ontology [14]. Tương tự, OARS khai thác thông tin về các lớp cha và các lớp con của các ontology để tính toán sự tương đồng về cấu trúc của các thực thể. Quan điểm chính của sự tương đồng về cấu trúc trong OARS được phát biểu như sau:

- Nếu hai lớp từ các ontology khác nhau có các lớp cha tương đồng trong hệ thống phân cấp, thì có khả năng là chúng xác định cùng một khái niệm.
- Nếu hai lớp từ các ontology khác nhau có các lớp con tương đồng trong hệ thống phân cấp, thì có khả năng là chúng xác định cùng một khái niệm.
- Nếu hai lớp từ các ontology khác nhau có các thuộc tính tương đồng, thì có khả năng là chúng xác định cùng khái niệm.
- Nếu hai thực thể có bất kỳ sự kết hợp của hai hoặc cả ba điểm tương đồng nêu trên thì chúng có chung khái niệm tương đồng.

Sự tương đồng về cấu trúc của hai thực thể e_i và e'_i từ hai ontology tương ứng O và O' có tính đến các điểm tương đồng giữa các lớp cha, các lớp con và các thuộc tính của hai thực thể.

Gọi

- Sim_hsp(e_i, e'_i) là sự tương đồng về cấu trúc giữa các lớp cha của các thực thể e_i, và e'_i,
- $K_{sup}(e_i)$ là tập của các lớp cha của thực thể e_i,
- $K_{sup}(e'_i)$ là tập của các lớp cha của thực thể e'_i,
- $|K_{sup}(e_i)|$ là số phần tử của $K_{sup}(e_i)$,
- $|K_{sup}(e'_i)|$ là số phần tử của $K_{sup}(e'_i)$,

Ta có

$$\text{Sim\_hsp}(e_i, e'_i) = \frac{1}{2}\left(\frac{|(K_{sup}(e_i) \cap K_{sup}(e'_i))|}{|K_{sup}(e_i)|} + \frac{|(K_{sup}(e_i) \cap K_{sup}(e'_i))|}{|K_{sup}(e'_i)|}\right) \quad (3)$$

Gọi

- Sim_hsb(e_i, e'_i) có sự tương đồng về cấu trúc giữa các lớp con của các thực thể e_i và e'_i,
- $K_{sub}(e_i)$ là tập các lớp con của thực thể e_i,
- $K_{sub}(e'_i)$ là các tập các lớp con của thực thể e'_i,
- $|K_{sub}(e_i)|$ là số phần tử của $K_{sub}(e_i)$,
- $|K_{sub}(e'_i)|$ là số phần tử của $K_{sub}(e'_i)$,

Ta có

$$\text{Sim\_hsb}(e_i, e'_i) = \frac{1}{2}\left(\frac{|(K_{sub}(e_i) \cap K_{sub}(e'_i))|}{|K_{sub}(e_i)|} + \frac{|(K_{sub}(e_i) \cap K_{sub}(e'_i))|}{|K_{sub}(e'_i)|}\right) \quad (4)$$

Sự tương đồng giữa các thuộc tính của các thực thể cũng đóng một vai trò quan trọng trong việc xác định sự tương đồng tổng thể của hai thực thể trong các ontology khác nhau.

Gọi

- Sim_pr(e_i, e'_i) diễn tả sự tương đồng giữa các thuộc tính của thực thể e_i, và e'_i,
- $Pr(e_i)$ là tập các thuộc tính của thực thể e_i,
- $Pr(e'_i)$ là tập các thuộc tính của thực thể e'_i,
- $|Pr(e_i)|$ là các phần tử của $Pr(e_i)$,
- $|Pr(e'_i)|$ là các phần tử của $Pr(e'_i)$,

Ta có

$$\text{Sim\_pr}(e_i,\ e'_i) = \frac{1}{2}\left(\frac{|(Pr(e_i) \cap Pr(e'_i))|}{|Pr(e_i)|} + \frac{|(Pr(e_i) \cap Pr(e'_i))|}{pr(e'_i)|}\right) \quad (5)$$

Cuối cùng, sự tương đồng về cấu trúc tổng thể Sim_strc(e_i,e'_i) của hai thực thể được tính bằng trung bình cộng của ba đối sánh về cấu trúc sử dụng phương trình (6).

$$\text{Sim\_strc}(e_i, e'_i) = \frac{1}{3}(Sim\_hsp(e_i, e'_i) + Sim\_hsb(e_i, e'_i) + Sim\_pr(e_i, e'_i)) \quad (6)$$

IV. SỬ DỤNG TẬP THÔ KẾT HỢP TÍNH TƯƠNG ĐỒNG

Ví dụ: Chúng ta có bảng sau:

Bảng 1. Bảng thông tin bệnh nhân

Bệnh nhân	Đau đầu	Đau cơ	Sốt	Cúm
P1	Có	Không	Cao	Có
P2	Không	Có	Cao	Có
P3	Có	Có	Rất cao	Có
P4	Không	Có	Bình thường	Không
P5	Có	Không	Cao	Không
P6	Không	Có	Rất cao	Có

Tập đối tượng U = {P1, P2, P3, P4, P5, P6}

Tập thuộc tính Q = {Đau đầu, đau cơ, sốt, cúm}

P $\subseteq$ Q, X $\subseteq$ U và x, y $\in$ U (x, y là hai đối tượng trong tập U)

Quan hệ không thể phân biệt theo P, ký hiệu IND(P) được định nghĩa như sau:

IND(P) = {(x, y) $\in$ U x U: f(x,q) = f(y,q) $\forall$ q$\in$P}

Quan hệ không thể phân biệt là một quan hệ tương đương và chia tập đối tượng U thành một họ các lớp tương đương và ký hiệu là U|IND(P)

Với $\forall$ x $\in$ U, lớp tương đương của x trong quan hệ IND(P) được ký hiệu là I_p

P – xấp xỉ dưới của X, ký hiệu $\underline{P}(X)$:

$\underline{P}(X) = \{x \in U; I_p \subseteq X\}$

P – xấp xỉ trên của X, ký hiệu $\overline{P}(X)$:

$\overline{P}(X) = \{x| I_p \cap X \neq \emptyset\}$

P – biên của X được ký hiệu là Bnp(X) và tính như sau:

$Bnp(X) = \overline{P}(X) - \underline{P}(X)$

$Bn_p(X)$ là tập các phần tử mà sử dụng tập thuộc tính P ta không thể xác định chúng có thuộc vào X hay không.

Định nghĩa tập Thô: Tập X được gọi là tập thô nếu $Bn_p(X)$ là khác rỗng.

Ví dụ: Từ Bảng 1, nếu thuộc tính P = {Cúm} là "Có", thì X = {P1, P2, P3, P6}

Với P = {Đau đầu, sốt}

IND(P) = {(P1, P5)}

U|IND(P) = {{P1, P5}, {P2}, {P3}, {P4}, {P6}}

$\underline{P}(X)$ = {P2, P3, P6}

$\overline{P}(X)$ = {P1, P2, P3, P5, P6}

$Bnp(X) = \{P1, P5\} \neq \emptyset \rightarrow$ Tập thô

Lý thuyết các tập Thô dựa trên mối quan hệ về tính không phân biệt được của các đối tượng liên quan đến thông tin sẵn có, chúng được phân vùng tạo thành các tập gồm các đối tượng tương đồng gọi là các tập cơ bản. Lý thuyết tập thô được xem là phương pháp hữu ích để phân tích các đặc tả về đối tượng. Nó giả định rằng mỗi đối tượng đề cập được kết hợp với số lượng thông tin nhất định, được diễn tả bởi một số thuộc tính. Tập Thô được ứng dụng trong kỹ thuật phát hiện tri thức và khai phá dữ liệu.

Khái niệm về các đối tượng và các thuộc tính của chúng đối với các tập Thô được mở rộng trong OARS, để xử lý với những bất ổn trong suốt quá trình ánh xạ của việc đối sánh ontology, khi các kết quả của các đối sánh riêng biệt không đưa ra dấu hiệu đáng tin cậy về việc các thực thể giống nhau hoặc khác nhau. Việc sử dụng các tập Thô, các kết quả tương đồng của các đối sánh riêng lẻ được xem là các thuộc tính của các phần tử trong việc phân loại, nó được tiếp tục sử dụng để xác định những điểm tương đồng giữa các phần tử dựa trên các giá trị về thuộc tính của chúng.

Gọi

- U là tập các thực thể không được ánh xạ trong ontology đích, U={e_1, e_2, e_3, …, e_n},
- F là tập các nhân tố đối sánh, nó diễn tả sự bao hàm của các kết quả đối sánh riêng lẻ, F={f_1, f_2, f_3},
- X là tập con của U.

Gọi $[x]_F$ biểu thị một tập các thực thể mà các điểm tương đồng giữa chúng có liên quan đến các nhân tố đối sánh. Các phép xấp xỉ dưới và xấp xỉ trên của tập X được định nghĩa như sau.

Gọi

- $\underline{F}(X)$ diễn tả xấp xỉ dưới của tập X đối với F, là tập của các nhân tố đối sánh. Khi đó $\underline{F}(X)$ là một tập các thực thể chắc chắn thuộc về X, được định nghĩa bởi biểu thức (7).
- $\underline{F}(X) = \{x \mid [x]_F \subseteq X\}$ (7)
- $\overline{F}(X)$ diễn tả xấp xỉ trên của tập X đối với F. Khi đó $\overline{F}(X)$ là một tập các thực thể có thể thuộc về X, như được định nghĩa bởi biểu thức (8).

$$\overline{F}(X) = \{x \mid [x]_F \cap X \neq \varnothing\} \quad (8)$$

Tỷ lệ của sự phân loại các tập Thô đối với các xấp xỉ dưới và xấp xỉ trên của tập X có thể được tính bằng cách sử dụng phương trình (9).

$$\alpha_F(X) = \frac{\underline{F}(X)}{\overline{F}(X)} \quad (9)$$

Tỷ lệ này sẽ nằm trong khoảng $0 \leq \alpha_F(X) \leq 1$. Đối với một thực thể được chọn từ ontology nguồn, OARS tính toán sự tương đồng cho từng thực thể không được ánh xạ trong tập đích U. Đối với các thực thể có ba đối sánh riêng lẻ, tức là, Sim_strng(e_i, e_i') và Sim_lin(w_i, w_i') và Sim_strc(e_i, e_i') không tìm thấy các đối sánh chính xác giữa chúng, các kết quả tương đồng được tạo ra bởi các đối sánh này sẽ phân loại theo các tập Thô cho mỗi phần tử không được ánh xạ. Các thực thể trong tập X có thể được xác định đối với F, các thực thể này sẽ được xem xét để ánh xạ khi tỷ lệ của sự phân loại các tập Thô chính xác là 1 dựa vào phương trình (9). Tập F xác định ba nhân tố đối sánh (f_1, f_2, f_3) như sau để gán mức độ tin cậy vào việc phân loại các tập Thô.

- f_1 diễn tả giá trị của Sim_strng(e_i, e'_i), được định nghĩa trong (1).
- f_2 diễn tả giá trị trung bình của Sim_hsp(e_i, e'_i) và Sim_hsb(e_i, e'_i) như được định nghĩa tương ứng trong phương trình (3) và (4).
- f_3 diễn tả giá trị Sim_pr(e_i, e'_i) được định nghĩa trong phương trình (5).

Đối sánh về ngữ nghĩa (Sim_lin(w_i, w_i')) không được xét trong việc tính toán ba nhân tố đối sánh bởi vì nó chỉ tạo ra một giá trị cố định là 0.5 dựa vào phương trình (2) cho các thực thể không rõ ràng trong việc ánh xạ, nó không thích hợp cho việc phân loại các tập Thô. Sự tương đồng của hai thực thể được tính từ bốn khía cạnh, tức là tương đồng về chuỗi, tương đồng về lớp cha, tương đồng về lớp con, và tương đồng về thuộc tính. Mỗi khía cạnh tương đồng được tính toán với trọng số là 25% có nghĩa là f_1 hoặc f_3 với giá trị 25% của tổng tương đồng được diễn tả bởi tập F, trong khi f_2 với giá trị 50% của tổng tương đồng được diễn tả bởi tập F. Để mở rộng tập của các thực thể được phân loại bởi các tập Thô, các giá trị của các nhân tố đối sánh được chuẩn hóa với các giá trị thập phân gần giống nhất trước khi tính toán tỷ lệ chính xác của sự phân loại các tập Thô.

Thuật toán 1 cho thấy mã giả của sự phân loại các tập Thô với các thực thể ánh xạ. Dòng 1 được sử dụng để gán ba nhân tố đối sánh. Dòng 2-6 được sử dụng để chọn các thực thể đối với sự đối sánh dựa trên tỷ lệ chính xác về sự phân loại của các tập Thô. Dòng 7-10 được sử dụng để gán mức độ tin cậy cho các đối tượng được ánh xạ.

Thuật toán 1: Các thực thể ánh xạ sử dụng sự phân loại các tập Thô.

Nhập: E = {e_1, e_2, e_3, …, e_m}, một tập các thực thể không được ánh xạ từ ontology nguồn;

E' = {e'$_1$, e'$_2$, e'$_3$, …, e'$_n$}, một tập các thực thể không được ánh xạ từ ontology đích;

F_1 = F, F = {f_1, f_2, f_3}, một tập các nhân tố đối sánh;

$F_2 \subset$ F, F_2 = {f_1, f_2};

$F_3 \subset$ F, F_3 = {f_2, f_3};

Xuất: đối sánh (e_i, e_j, c) trong đó c là mức độ tin cậy;

1: For k=1 to 3;

```
2:      For i=1 to m;
3:        For j=1 to n;
4:            tính α_F được định nghĩa trong phương trình (9);
5:            If  α_F= 1, then
6:                đối sánh (e_i, e'_j);
7:                If F_k = F_1 then
8:                    c = 1;
9:                Else
10:                   c = 0.75;
11:               Endif
12:           Endif
13:       Endfor
14:     Endfor
15: Endfor
```

Để minh họa thêm việc sử dụng các tập Thô trong OARS xác định những điểm tương đồng giữa các thực thể ontology, chúng tôi đưa ra ví dụ như trong hình 1. Chúng tôi giả định rằng cả hai trường hợp có 5 thực thể chưa được ánh xạ cụ thể là e'_1, e'_2, e'_3, e'_4 và e'_5 trong ontology đích. Ba nhân tố đối sánh đưa ra dựa vào từng thực thể đích sau khi so sánh với thực thể e_i trong ontology nguồn. Trong ví dụ này, chúng tôi chỉ so sánh e_i với e'_1 và e'_4 tương ứng.

Chúng tôi trình bày hai trường hợp riêng biệt cụ thể là Case-1 và Case-2. Case-1 được trình bày để chứng minh sự tính toán tương đồng giữa thực thể nguồn và thực thể đích, nó chỉ dựa trên hai nhân tố f_1 và f_2. Trong khi Case-2 được trình bày để giải thích sự tính toán tương đồng có tính đến ba nhân tố f_1, f_2 và f_3.

Hãy xét Case-1 như trong Hình 1:

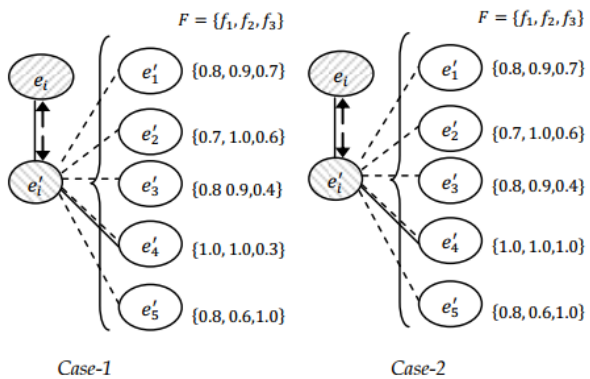
- cho X = $\{e_i, e'_1\}$, với F = $\{f_1, f_2\}$, $\overline{F}(X) = \{\{e_i, e'_4\}, \{e'_1, e'_3\}\}$, và $\underline{F}(X) = \varnothing$, $\alpha_F(X) = 0$, chỉ ra rằng e_i và e'_1 không thể xác định dựa trên các kết quả đã cho và xu hướng không được ánh xạ.
- cho X = $\{e_i, e'_4\}$, với F = $\{f_1, f_2\}$, cả $\overline{F}(X)$ và $\underline{F}(X) = \{e_i, e'_4\}$, và $\alpha_F(X) = 1$ chỉ ra rằng e_i và e'_4 được xem xét để ánh xạ. Giá trị độ tin cậy là 0.75 được gán cho mối quan hệ ánh xạ vì tập F chứa hai nhân tố đối sánh trong trường hợp này.

Hãy xem xét Case-2 như trong Hình 1:

- cho X = $\{e_i, e'_1\}$, với F = $\{f_1, f_2, f_3\}$, $\overline{F}(X) = \{e_i, e'_4\}$, và $\underline{F}(X) = \varnothing$, $\alpha_F(X) = 0$, chỉ ra rằng e_i và e'_1 không thể xác định dựa trên những kết quả đã cho và không được ánh xạ.
- cho X = $\{e_i, e'_4\}$, với F = $\{f_1, f_2, f_3\}$, cả $\overline{F}(X)$ và $\underline{F}(X) = \{e_i, e'_4\}$, $\alpha_F(X) = 1$ chỉ ra e_i và e'_4 có thể được xác định đối với F, và hai thực thể được xem xét để ánh xạ với mức độ tin cậy là 1.

Như đã thảo luận trước đó, sự phân loại các tập Thô với các đối tượng phân loại dựa trên các thuộc tính cụ thể. Tương tự như vậy, trong ví dụ này, sử dụng case-1, các đối tượng (trong trường hợp này, các đối tượng là các thực thể e_i và e'_4) đang xem xét để ánh xạ dựa trên các thuộc tính (trong trường hợp này các thuộc tính là f_1 và f_2).

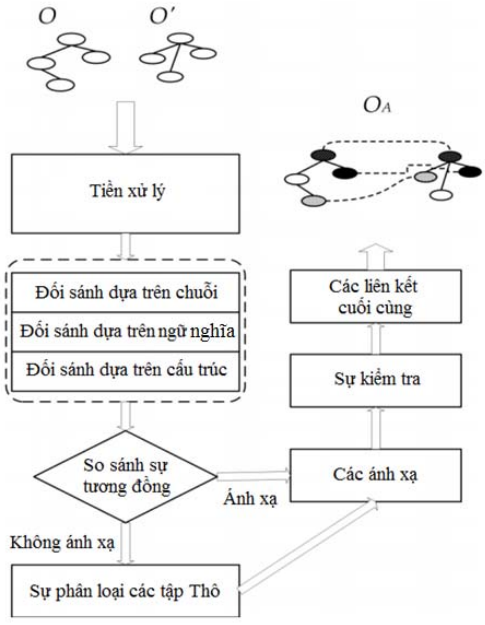
Quá trình đối sánh trong OARS được thể hiện trong Hình 2, nó bắt đầu với tiền xử lý để chuẩn hóa các tên của các thực thể ontology như đã thảo luận trong Phần III mục A. OARS sau đó sử dụng ba đối sánh riêng lẻ để tính toán các giá trị tương đồng của các thực thể giữa các ontology nguồn và ontology đích sử dụng các phương trình (1), (2) và (6). Nếu sự tương đồng chính xác được tìm thấy bởi bất kỳ sự đối sánh riêng lẻ, các thực thể được chọn để ánh xạ và độ tin cậy được gán bằng 1. Đối với các thực thể không rõ ràng, chúng sẽ được đưa vào sự phân loại các tập Thô để tính toán thêm. Sau quá trình ánh xạ, OARS kiểm tra mọi thực thể trong ontology nguồn không được ánh xạ với nhiều hơn một thực thể trong ontology đích và ngược lại. Nếu một ánh xạ được tìm thấy, OARS chọn ánh xạ với độ tin cậy cao hơn trước khi sinh ra các đối sánh sau cùng.



Hình 1. Ví dụ về sự phân loại các tập Thô

V. ĐÁNH GIÁ

Chúng tôi thực hiện OARS sử dụng ngôn ngữ lập trình Java và đối sánh API để ontology nguồn đầu vào và ontology đích tạo ra các kết quả liên kết. Chúng tôi sử dụng các ontology đối sánh chuẩn của OAEI năm 2010 để đánh giá hiệu quả của OARS. Các thử nghiệm chuẩn cung cấp các ontology khác nhau để đánh giá một loạt các tính năng về điểm mạnh và điểm yếu của các đối sánh tồn tại. Các đối sánh tham khảo có sẵn để thử nghiệm, nó được đối sánh thủ công bởi OAEI và được coi như các đối sánh đúng.



Hình 2. Quá trình đối sánh OARS

A. Các tập dữ liệu chuẩn

Các tập dữ liệu chuẩn OAEI 2010 bao gồm một số ontology với các mức độ khác nhau về tính phức tạp. Các ontology này được xây dựng từ một ontology OWL. Ontology cơ sở là test-101 được coi là một ontology tham khảo, chứa 33 lớp được đặt tên, 24 thuộc tính đối tượng, 40 thuộc tính dữ liệu và 76 thực thể, trong đó 20 thực thể được đặt tên trong khi phần còn lại là không có tên.

Những mô tả của các thử nghiệm này được trình bày trong Bảng 2, chủ yếu chứa ba nhóm – các thử nghiệm đơn giản (1xx), các thử nghiệm hệ thống (2xx) và các ontology thực tế (3xx). Nhóm 1xx có 4 ontology với sự thay đổi nhỏ. Các ontology trong các thử nghiệm hệ thống (2xx) được xây dựng để thử nghiệm khả năng của các hệ thống đối sánh khi thông tin cụ thể được loại bỏ từ các ontology. Các thông tin loại bỏ có thể bao gồm các trường hợp sau đây:

- Các lớp được thay thế bằng một số lớp, được mở rộng hoặc thu hẹp.
- Các tên thực thể được thay thế bằng các từ đồng nghĩa, bằng các chuỗi từ với các ngữ nghĩa khác nhau hoặc thậm chí một số chuỗi ngẫu nhiên.
- Các nhận xét ở mức độ khác nhau được dịch với các ngữ nghĩa khác nhau.
- Các thuộc tính bao hàm hoặc các giới hạn của chúng với các lớp được loại trừ.
- Các thể hiện được loại trừ.
- Các phân cấp chỉ định được mở rộng, thu hẹp hoặc loại trừ.

Hơn nữa, các ontology trong nhóm 3xx là các ontology về thế giới thực được cung cấp bởi các tổ chức khác nhau và không thay đổi trong các tập dữ liệu chuẩn.

Bảng 2. Các mô tả của các tập dữ liệu chuẩn

Các tập thử nghiệm	Miêu tả
101-104	Cấu trúc phân cấp giống nhau Tên thực thể giống nhau hoặc hoàn toàn khác nhau
201-210	Cấu trúc phân cấp giống nhau Ngữ nghĩa khác nhau được sử dụng ở một số cấp độ
221-247	Cấu trúc phân cấp khác nhau Nhãn về ngữ nghĩa giống nhau
248-266	Cấu trúc phân cấp và ngữ nghĩa là khác nhau
301-304	Các ontology thế giới thực

B. Các tiêu chuẩn đánh giá

Chúng tôi sử dụng độ đo precision, recall và F-measure để đánh giá độ chính xác của OARS trong đối sánh ontology. Precision và recall được chấp nhận rộng rãi và các độ đo được công nhận trong các lĩnh vực nghiên cứu về việc truy hồi thông tin và đối sánh ontology.

Gọi A_d là tập các đối sánh được sinh ra, A_t là tập đầy đủ của các đối sánh chính xác. Độ đo precision, recall và F-measures được xác định bằng cách sử dụng các phương trình (10), (11) và (12) tương ứng.

$$Prec = \frac{|A_d \cap A_t|}{|A_d|} \tag{10}$$

$$Rec = \frac{|A_d \cap A_t|}{|A_t|} \tag{11}$$

$$F - measure = \frac{2 \times Prec \times Rec}{Prec + Rec} \tag{12}$$

C. Kết quả thực nghiệm

Phần này trình bày đánh giá hiệu quả của OARS trong một số kịch bản. Việc đánh giá về sự kết hợp tính tương đồng cho biết ảnh hưởng của nó đến các kết quả của việc trình bày tổng thể về sự đối sánh ontology. Việc so sánh của OARS với các hệ thống đối sánh hiện có cũng được trình bày trong phần này. Các phân tích quan trọng được trình bày để làm nổi bật những ưu điểm và những hạn chế của OARS. Quá trình đối sánh trong OARS là hoàn toàn tự động và do đó không có người sử dụng tham gia can thiệp vào bất kỳ các thử nghiệm trong quá trình liên kết.

1. Kết hợp tính tương đồng

Để đánh giá hiệu quả hoạt động của OARS một cách toàn diện, chúng tôi đã xây dựng một số kịch bản thử nghiệm bằng cách sử dụng các tập dữ liệu chuẩn và tiêu chuẩn đánh giá được xác định bởi (10), (11). Mục đích chính của các kịch bản thử nghiệm này là để đánh giá:

- Hiệu quả của các đối sánh về tính tương đồng riêng lẻ,
- Hiệu quả của các kết hợp khác nhau về các đối sánh riêng lẻ, và
- Hiệu quả của việc phân loại các tập Thô về các kết quả kết hợp của các đối sánh riêng lẻ.

Chúng tôi thiết kế bốn kịch bản, trong đó mỗi kịch bản sử dụng các kết hợp khác nhau của các đối sánh để tổng hợp các kết quả ánh xạ sau cùng. Với mục đích này, chúng tôi thực hiện bốn thuật toán riêng trong hệ thống liên kết, cụ thể là A1, A2, A3 và A4 như được định nghĩa bởi các biểu thức (13), (14), (15) và (16) tương ứng. Các chi tiết của bốn thuật toán được trình bày dưới đây.

- A1 diễn tả phương pháp trong đó đối sánh ontology có nguồn gốc sử dụng giá trị trung bình của các kết quả trả về bởi các đối sánh về chuỗi và ngữ nghĩa,

$$A1 = (Sim\_strng(e_i, e'_i) + Sim\_lin(w_i, w'_i))/2 \tag{13}$$

- A2 diễn tả phương pháp trong đó đối sánh được bắt nguồn từ việc sử dụng các giá trị trung bình của kết quả trả về bởi các đối sánh về cấu trúc và ngữ nghĩa,

$$A2 = (Sim\_strc(e_i, e'_i) + Sim\_lin(w_i, w'_i))/2 \tag{14}$$

- Tương tự, A3 sử dụng giá trị trung bình của các kết quả được tạo ra bởi các đối sánh dựa trên chuỗi và cấu trúc đối với liên kết,

$$A3 = (Sim\_strng(e_i, e_i') + Sim\_strc(e_i, e_i'))/2 \tag{15}$$

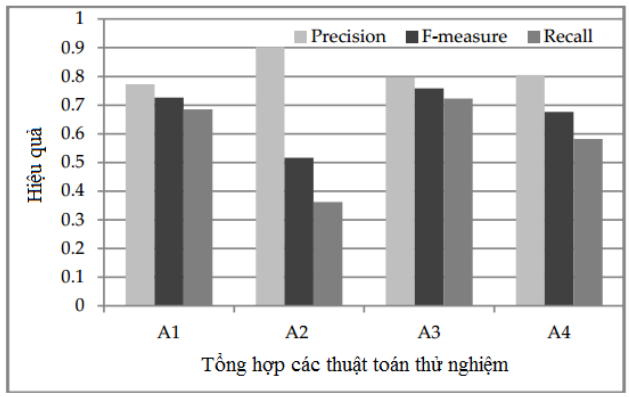
- Cuối cùng, A4 sử dụng giá trị trung bình của các kết quả được tạo ra bởi các đối sánh về chuỗi, ngữ nghĩa và cấu trúc đối với liên kết,

$$A4 = (Sim\_strng(e_i, e'_i) + Sim\_lin(w_i, w'_i) + Sim\_strc(e_i, e'_i))/3 \tag{16}$$

Chúng tôi chọn nhóm 3xx của các ontology thử nghiệm từ các tập dữ liệu chuẩn bởi vì nó chứa các ontology về thế giới thực như mô tả trong Phần V mục A. Hình 3 cho thấy các kết quả so sánh của các phương pháp A1, A2, A3 và A4.

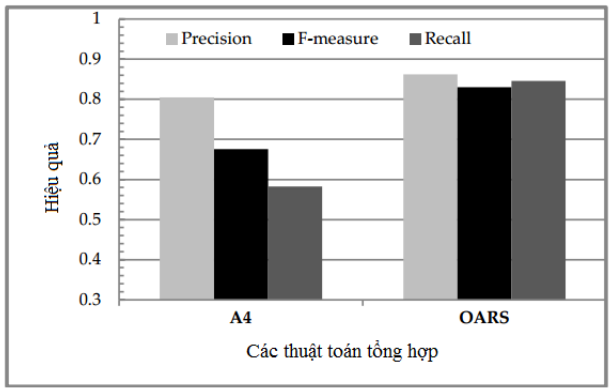
Tập các ontology trong nhóm 3xx có các điểm tương đồng về chuỗi lớn hơn các điểm tương đồng về cấu trúc và ngữ nghĩa trong việc so sánh với ontology tham khảo. Từ Hình 3 nó cũng cho thấy các thuật toán (A1, A3 và A4) sử dụng đối sánh dựa trên chuỗi cho các kết quả với F-measure tốt hơn A2, nó không sử dụng việc đối sánh dựa trên chuỗi. Điều này cũng cho thấy tầm quan trọng của việc đối sánh riêng lẻ trong việc đối sánh các ontology với các tính năng thích hợp. Đối sánh ngữ nghĩa không thực hiện tốt với các ontology trong nhóm 3xx vì nó không thể xử lý một số thực thể với phần tiền tố như "abstract"="hasAbstract", "volume"= "hasVolume" và "copyright"="hasCopyright" bằng cách sử dụng các tập đồng nghĩa WordNet. Như vậy các kết quả làm giảm hiệu quả ánh xạ tổng thể của các đối sánh khác khi giá trị trung bình của tất cả các đối sánh được thực hiện trong sự kết hợp. Trong Hình 3, thuật toán A3 không xét kết quả của việc đối sánh về ngữ nghĩa đem lại một giá trị F-measure tốt hơn so với các thuật toán khác.

Chúng tôi cũng so sánh hiệu quả của OARS với phương pháp A4 sử dụng các ontology của nhóm 3xx. Như thể hiện trong Hình 4, có một cải tiến đáng kể về hiệu quả của OARS khi so sánh với A4 trong ba khía cạnh. Các giá trị precision, recall và F-measure của A4 là 0.805, 0.582 và 0.675 tương ứng trong khi đối với OARS những giá trị này là 0.862, 0.845 và 0.83 tương ứng. Việc cải thiện tổng thể đạt được bởi OARS với F-measure là 22.96% so với phương pháp A4.



Hình 3. Hiệu quả của bốn thuật toán tổng hợp

Các kết quả đánh giá này khẳng định rằng không có đối sánh riêng lẻ đủ đạt được độ chính xác cao trong đối sánh ontology. Quan trọng hơn, các kết quả tổng hợp của các đối sánh riêng lẻ bằng cách lấy giá trị trung bình không những thiếu sót mà còn có thể làm suy giảm hiệu quả ánh xạ tổng thể khi một số đối sánh hiện tại có các giá trị tương đồng thấp.



Hình 4. Sự so sánh các thuật toán tổng hợp

2. Chuẩn hóa trong việc phân loại các tập thô

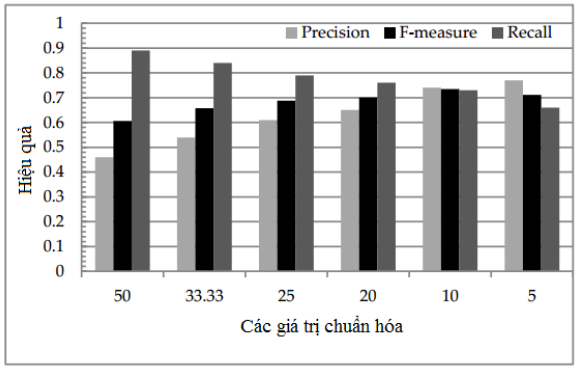
Để chọn giá trị thích hợp nhất đối với các kết quả chuẩn hóa của các đối sánh riêng lẻ cho việc phân loại các tập Thô, chúng tôi thực hiện các thử nghiệm khác nhau, thấy rằng các giá trị chuẩn hóa là 50, 33.33, 25, 20, 10 và 5. Các thử nghiệm này đã thực hiện trên nhóm 2xx của tập dữ liệu chuẩn. Hình 5 cho thấy OARS đạt giá trị recall cao nhất bằng cách sử dụng giá trị chuẩn hóa là 50, nhưng mặt khác, nó cho giá trị precision thấp nhất. Tương tự như vậy, bằng cách sử dụng giá trị chuẩn hóa là 5, OARS đem lại giá trị precision cao nhất, nhưng cho giá trị recall thấp nhất. Chúng tôi sử dụng giá trị chuẩn hóa là 10, OARS khi đó đem lại độ đo F-measure tốt nhất.

3. So sánh OARS với các hệ thống đối sánh hiện nay

Phần này đánh giá OARS trong việc so sánh với một số hệ thống đối sánh tham gia vào tổ chức OAEI 2010 bằng cách sử dụng các tập dữ liệu chuẩn của nhóm 1xx, nhóm 2xx và nhóm 3xx tương ứng.

Nhóm 1xx

Hầu hết các hệ thống đối sánh trong việc so sánh đạt được các kết quả hoàn hảo đối với các ontology trong nhóm 1xx với các giá trị precision và recall. Tuy nhiên, có một ngoại lệ là TaxoMap đạt được giá trị recall thấp là 0.34. Việc thực hiện tốt các hệ thống đối sánh trong các thử nghiệm này chủ yếu là do thực tế các ontology trong nhóm 1xx có các thực thể tương đồng rất cao. Vì không có tính đa dạng về cấu trúc trong số các ontology này, chỉ có các đối sánh dựa trên chuỗi và ngữ nghĩa đã được sử dụng trong OARS đối với đối sánh ontology trong nhóm 1xx.



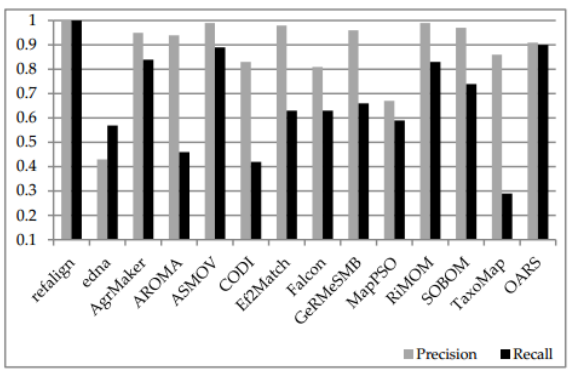
Hình 5. Đánh giá sự chuẩn hóa

Nhóm 2xx

Hầu hết các ontology trong nhóm 2xx được đối sánh phù hợp bởi OARS sử dụng đối sánh về ngữ nghĩa dựa trên WordNet để xử lý với các từ đồng nghĩa (ví dụ trong thử nghiệm 205). Đối sánh dựa vào chuỗi cũng thực hiện tốt trên chuỗi không đồng nhất. Đối sánh về ngữ nghĩa được chứng minh có tính hiệu quả trong các ontology mà các ngữ nghĩa được sử dụng, ví dụ trong các ontology thử nghiệm 201, 202 và 248-266. Hơn nữa, các ontology chỉ thay đổi về cấu trúc cũng được giải quyết thành công trong OARS bởi vì khi thông tin này đã được chặn lại, những điểm tương đồng về ngữ nghĩa hoặc chuỗi vẫn có sẵn trong các ontology. Chúng tôi thấy rằng tác vụ đối sánh đang thách thức nhiều nhất là xử lý với các ontology này, trong đó các thay đổi cả về cấu trúc và các nhãn đã được thực hiện. Trong các thử nghiệm về nhóm 2xx, OARS đạt được giá trị recall tốt nhất trong số các hệ thống đối sánh như thể hiện trong Hình 6, vì khả năng của nó xử lý với các thực thể không rõ ràng trong việc ánh xạ ontology. Điều đáng chú ý là hệ thống đối sánh khác như ASMOV, AgrMaker và RiMOM cũng đạt được các giá trị recall cao là 0.89, 0.83 và 0.84 tương ứng.

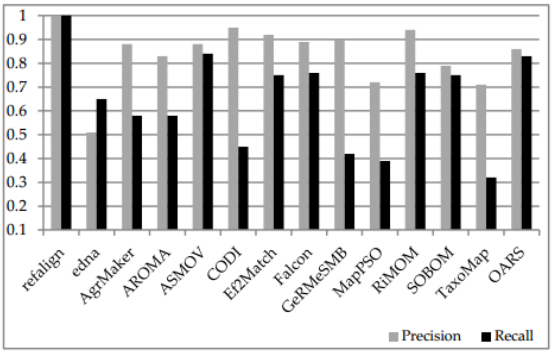
Nhóm 3xx

Có 4 ontology về thế giới thực trong nhóm 3xx có sự kết hợp về tính tối nghĩa tìm thấy trong nhóm dữ liệu 2xx. Với các thử nghiệm trên nhóm dữ liệu 3xx, do có rất ít thông tin về cấu trúc có sẵn trong các ontology này, ví dụ như ontology 302, OARS chủ yếu dựa vào chuỗi và ngữ nghĩa đối sánh trong việc đối sánh các ontology trong nhóm 3xx. Các kết quả thử nghiệm của nhóm này được trình bày trong Hình 7, nó cho thấy ASMOV đem lại kết quả recall tốt nhất, tiếp theo là OARS với giá trị recall là 0.86.



Hình 6. Các kết quả đánh giá với nhóm 2xx

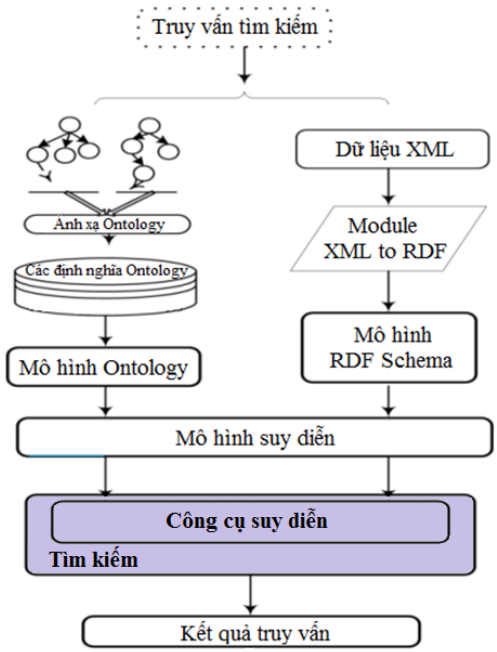
Điều đáng chú ý là hiệu quả của OARS với precision có thể so sánh với các hệ thống đối sánh khác được phản ánh trong cả Hình 6 và Hình 7 tương ứng.



Hình 7. Các kết quả đánh giá về nhóm 3xx

VI. TÍCH HỢP OARS VÀO SEMFARM

Để khai thác các khả năng đối sánh ontology của OARS trong SemFARM, module tìm kiếm được thực hiện trong SemFARM. Quá trình ghi nhớ các file một cách tự động với ba thuộc tính cơ bản và hai tên miền được người sử dụng nhập vào. Dữ liệu tổng hợp được phân tích một cách tự động và lưu trữ vào XML có cấu trúc văn bản được mô tả trong [5]. Hình 8 cho thấy toàn bộ quá trình module tìm kiếm của SemFARM, trong đó các truy vấn file đầu vào được trả lời sau khi sáp nhập hai ontology hiện hành. Khi các ontology OWL tìm thấy trên hệ thống có các truy vấn, trước tiên chúng tạo đối sánh giữa các ontology và các đối sánh này được kết hợp để sử dụng như một ontology đơn. Các đối sánh này được sinh ra khi sáp nhập của hai ontology đầu vào. Sau khi ontology được sáp nhập sẽ nhận được mô hình ontology và được đối sánh với mô hình RDF để tạo thành mô hình suy diễn. Mô hình RDF được tạo ra một cách tự động từ XML bởi mô-đun chuyển đổi từ XML sang RDF như được thể hiện trong Hình 8. Cuối cùng, truy vấn tìm kiếm file được trả lời thông qua mô hình suy diễn.



Hình 8. File truy hồi trong SemFARM

A. Đánh giá file truy hồi trong SemFARM

Một ontology bổ sung được khai thác để đánh giá hiệu quả của SemFARM cho phép bởi OARS. Khái niệm miền của ontology bổ sung được chọn từ một khái niệm con của ontology chính được sử dụng trong việc thực hiện của SemFARM. Mục đích chính là để đánh giá tính hiệu quả của OARS với sự hỗ trợ của đối sánh ontology trong SemFARM.

B. Đánh giá tính hiệu quả của SemFARM

Hai trường hợp sau đây cho việc đánh giá:

- **Case-1: SemFARM không có OARS**

  Một ontology chính được sử dụng để truy hồi các file theo yêu cầu. Trong trường hợp này, module tìm kiếm của SemFARM được sử dụng để trích xuất các thông tin từ ontology chính. Do đó, chỉ một ontology chính được sử dụng trong trường hợp này.

- **Case-2: SemFARM với OARS**

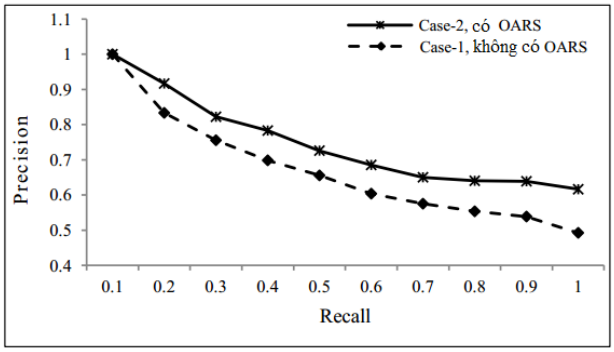
  Hai ontology được sử dụng để truy hồi các file theo yêu cầu. Module tìm kiếm của SemFARM có sử dụng OARS, nó đối sánh ontology chính và ontology phụ. Trong trường hợp này, nhiều kiến thức thu được bằng cách sử dụng hai ontology.

Chúng tôi sử dụng ba thử nghiệm để chứng minh tính hiệu quả việc đối sánh ontology của OARS trong việc truy hồi tập tin. Giá trị precision và recall được tính trung bình với ba kết quả thử nghiệm. Trong mỗi thử nghiệm, số file với các từ khóa khác nhau, được coi là có liên quan đến truy vấn tìm kiếm file. Số file liên quan có các giá trị recall khác nhau. Cần lưu ý rằng một số file với các từ khóa, không được xác định bởi ontology chính. Tuy nhiên, những từ khóa này có thể xác định trong ontology phụ nhưng với một miền khái niệm có giới hạn.

Hơn nữa, các từ truy vấn được sử dụng trong mỗi thử nghiệm khác nhau, được bảo đảm rằng các từ truy vấn này bao gồm các từ khóa xác định trong cả hai ontology, để cho ra cho một kết quả tốt với cả hai trường hợp. Tương tự, các từ truy vấn giống nhau cũng được sử dụng cho cả hai trường hợp trong mỗi thử nghiệm tương ứng.

C. Tính toán Precision và Recall

Việc so sánh tổng thể của hai trường hợp cho thấy một sự cải tiến của Case-2 đối với Case-1 với quan hệ precision so với cùng giá trị của recall như diễn tả trong Hình 9.



Hình 9. Hiệu quả của SemFARM được chấp nhận bởi OARS

Các giá trị precision trung bình của Case-1 và Case-2 là 0.65 và 0.72 tương ứng so với cùng giá trị recall là 0.5. Ta thấy rằng việc giảm các giá trị precision trong Case-2 là ít hơn của Case-1 khi giá trị recall thay đổi từ 0.1 đến 1. Kết quả cho thấy rằng các giá trị precision giảm từ 1 đến 0.49 trong Case-1 và từ 1 đến 0.61 trong Case-2 khi các giá trị recall tương ứng tăng từ 0.1 đến 1. Các giá trị precision cho Case-2 và Case-1 là 0.616 và 0.492 tương ứng khi giá trị recall là 1. Cần lưu ý rằng các giá trị precision là như nhau trong cả hai trường hợp khi giá trị recall là 0.1.

VII. KẾT LUẬN

Trong bài báo này, chúng tôi đã trình bày OARS, một hệ thống đối sánh ontology sử dụng các tập Thô để xử lý các thực thể không rõ ràng trong việc ánh xạ. Việc sử dụng các tập Thô đã được chứng minh tính hiệu quả với các thực thể ánh xạ mà đối sánh riêng lẻ không thể đưa ra quyết định trong việc ánh xạ ontology. Ý nghĩa của việc sử dụng các tập Thô như là một phương pháp tổng hợp để đánh giá và so sánh với một số hệ thống đối sánh hiện có bằng cách sử dụng các tập dữ liệu chuẩn ontology của tổ chức OAEI năm 2010. Các kết quả được đánh giá cao. Tính hiệu quả của OARS trong khung ứng dụng SemFARM cũng được nâng cao.

Hiện nay, chúng tôi đang nghiên cứu quá trình thử nghiệm của OARS để cải thiện tính hiệu quả của nó với giá trị precission mà không làm giảm giá trị recall. Với mục đích này, chúng tôi đang có kế hoạch sử dụng tính tương đồng của thông tin được phân cấp giữa các thực thể ontology. Chúng tôi cũng có kế hoạch tham gia tổ chức OAEI trong tương lai. Chú ý là OARS với hình thức hiện nay không thể đối sánh các ontology với các ngữ nghĩa diễn tả cho các thực thể lớp. Vì vậy, chúng tôi đang có kế hoạch tích hợp một vài từ điển về ngữ nghĩa vào việc đối sánh, cho phép OARS đối sánh các ontology với các ngữ nghĩa khác nhau.

VIII. TÀI LIỆU THAM KHẢO

[1] P. Shvaiko, J. Euzenat, “A Survey of Schema-based Matching Approaches”, Journal on Data Semantics IV, vol. 3730, pp. 146-171, 2005.

[2] N. F. Natalya “Semantic Integration: A Survey of Ontology-based Approaches”, ACM SIGMOD Record, vol. 33, no. 4, pp. 65-70, 2004.

[3] W. Hu, Y. Qu, “Falcon-AO: A Practical Ontology Matching System ”, Journal of Web Semantics, pp. 237-239, vol. 6, no. 3, 2008.

[4] S Jan, M Li, G Al-Sultany, H Al- Raweshidy, “Ontology Alignment using Rough Sets”, in Proc. of the 8th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), pp. 2683-2686, 2011.

[5] S. Jan, M. Li, G. Al-Sultany, Hamed Al-Raweshidy and I.A Shah, “Semantic File Annotation and Retrieval on Mobile Devices”, Mobile Information Systems, vol. 7, no. 2, pp. 107-122, 2011.

[6] M. Rodriguez and M. Egenhofer, “Determining Semantic Similarity among Entity Classes from Different Ontologies”, IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 7, pp. 442-456, 2003.

[7] G. Stoilos, G. Stamou, S. Kollias, “A String Metric for Ontology Alignment”, In proc. of the 4th International Semantic Web Conference, Springer LNCS, vol. 3729, pp. 624-637, 2005.

[8] Y. R. Jean-Mary, E. P. Shironoshita, M. R. Kabuka, “Ontology Matching with Semantic Verification”, Journal of Web Semantics, vol. 7, no.3 , pp. 235-251, 2009.

[9] H. Li, X. Zhou and B. Huang, "Method to Determine α in Rough set Model based on Connection Degree", Journal of Systems Enginee ring and Electronics, vol. 20, no. 1, pp.98–105, 2009.

[10] W. W. Cohen, P. Ravikumar, and S. E. Fienberg, “A Comparison of String Distance Metrics for Name-Matching Tasks”, in Proc. of the Workshop on Information Integration on the Web, pp. 73-78, 2003.

[11] Z. Pawlak, “Rough Sets”, International Journal of Information & Computer Sciences, vol. 11, pp. 341-356. 1982

[12] Y. Wang, W. Liu and D. Bell, “Combining Uncertain Outputs from Multiple Ontology Matchers”, In proc. of the 1st International Conference on Scalable Uncertainty Management, Lecture Notes in Computer Science, Springer, vol. 4772. pp. 201–214, 2007.

[13] D. Sánchez, M. Batet, D. Isern, and A. Valls, " Ontology-based Semantic Similarity: A New Feature-based Approach", Expert Systems with Applications, vol. 39, no. 9, pp. 7718-7728, 2012.

[14] N. Seco, T . Veale, J. Hayes, "An Intrinsic Information Content Metric for Semantic Similarity in WordNet", in Proc. of the 16th Eur opean Conference on Artificial Intelligence (EACI'04), pp. 1089-1090, 2004.

NEW APPROACH FOR ONTOLOGY ALIGNMENT

Huỳnh Nhứt Phát, Hoàng Hữu Hạnh, Phan Công Vinh

ABSTRACT - *Ontology alignment facilitates exchange of knowledge among heterogeneous data sources. Many approaches to ontology alignment use multiple similarity measures for mapping entities between ontologies. However, it remains a key challenge in dealing with uncertain entities for which the employed ontology alignment measures produce conflicting results on similarity of the mapped entities. This paper presents OARS, a Rough sets based new approach to ontology alignment which achieves a high degree of accuracy in situations where uncertainty arises because of the conflicting results generated by different similarity measures. OARS employs a combinational approach and considers both lexical and structural similarity measures. OARS is extensively evaluated with the benchmark ontologies of the Ontology Alignment Evaluation Initiative (OAEI) 2010, and performs best in the aspect of recall and precision in comparison with a number of alignment systems.*