Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00024

GIẢI PHÁP TÍCH HỢP XỬ LÝ NGỮ NGHĨA VÀO HỆ THỐNG GỢI Ý

Huỳnh Thanh Tài[1], Nguyễn Hữu Hoà[1], Huỳnh Minh Trí[2], Nguyễn Thái Nghe[1]
[1]Khoa Công nghệ Thông tin & Truyền thông, trường Đại học Cần Thơ
[2]Khoa Công nghệ Thông tin, trường Đại học Kiên Giang
httaik21@gmail.com, nhhoa@ctu.edu.vn, hmtri@vnkgu.edu.vn, ntnghe@cit.ctu.edu.vn

TÓM TẮT *— Hệ thống gợi ý thường tạo ra một danh sách các mục tin để gợi ý cho người dùng theo một trong hai cách: lọc dựa trên nội dung (content-based filtering) và lọc cộng tác (collaborative filtering). Lọc dựa trên nội dung là hướng tiếp cận căn cứ vào việc phân tích đặc trưng trên nội dung của các mục tin mà người dùng đã chọn trong quá khứ và hệ thống thực hiện gợi ý cho người dùng những mục tin có đặc trưng nội dung tương tự. Lọc cộng tác là hướng tiếp cận dựa trên nhóm người dùng đã từng chọn những mục tin giống người dùng cần gợi ý để xác định những mục tin cần giới thiệu với người này. Những hướng tiếp cận này chỉ sử dụng dữ liệu có sẵn để xây dựng các mô hình dự đoán. Trên thực tế, tồn tại những hệ thống gợi ý chưa sẵn có hoặc chưa đủ dữ liệu để huấn luyện cho mô hình dự đoán. Điều này cũng là một trong những nguyên nhân làm giảm độ chính xác của các kết quả gợi ý. Trong bài viết này chúng tôi giới thiệu giải pháp tích hợp xử lý ngữ nghĩa vào hệ thống gợi ý. Phương pháp này là sự kết hợp giữa kỹ thuật gợi ý truyền thống và phân tích mối quan hệ ngữ nghĩa của những mục tin trong hệ thống được lưu trữ bằng Ontology. Thông qua mô hình ngữ nghĩa, chúng tôi tiến hành suy diễn dữ liệu nhằm tăng thêm dữ liệu huấn luyện cho các mô hình dự đoán. Thực nghiệm cho thấy với việc tích hợp ngữ nghĩa để suy diễn thêm dữ liệu, các mô hình cho kết quả dự đoán chính xác hơn so với chỉ sử dụng dữ liệu sẵn có.*

Từ khoá *— Hệ thống gợi ý, web ngữ nghĩa, gợi ý dựa trên ngữ nghĩa, dữ liệu suy diễn.*

I. GIỚI THIỆU

Việc sử dụng những hệ thống gợi ý (recommender systems - RSs) hiện nay đã phát triển nhanh chóng và mạnh mẽ. Những kỹ thuật gợi ý giúp con người giải quyết được vấn đề quá tải về thông tin và sự lựa chọn thông tin bằng cách trình bày các mục tin gợi ý phù hợp với sở thích của người dùng. Tuy nhiên, vấn đề của những thuật toán gợi ý hiện nay đang gặp phải là vấn đề thưa thớt dữ liệu và vấn đề mục tin mới thực sự khó đến được với người dùng của hệ thống. Trong đó, phải kể đến việc các hệ thống không có khả năng đưa ra được những gợi ý phù hợp cho đến khi thu thập đủ số lượng xếp hạng từ phía người dùng để huấn luyện cho mô hình dự đoán. Thực tế cho thấy, các hệ thống gợi ý thường sử dụng kết hợp nhiều hơn một mô hình dự đoán nhằm cải thiện độ chính xác, nâng cao tính hiệu quả, cũng như giảm thiểu lỗi cho các mô hình dự đoán xếp hạng. Tuy nhiên, các hướng tiếp cận nêu trên chỉ sử dụng dữ liệu sẵn có để huấn luyện cho mô hình dự đoán. Với những hệ thống chưa sẵn có dữ liệu để huấn luyện, hoặc dữ liệu thưa, người dùng mới thì các mô hình không có khả năng đưa ra những dự đoán chính xác để gợi ý cho người dùng.

Bài viết này đề xuất giải pháp tích hợp ngữ nghĩa vào hệ thống gợi ý. Phương pháp này kết hợp giữa kỹ thuật gợi ý truyền thống và phân tích mối quan hệ ngữ nghĩa của những mục tin trong hệ thống lưu trữ bằng Ontology. Ý tưởng là tận dụng ưu thế của những phương pháp gợi ý truyền thống kết hợp với khai thác mối quan hệ ngữ nghĩa giữa các mục tin nhằm suy diễn thêm dữ liệu để thực hiện dự đoán, giúp hạn chế phần nào tình trạng dữ liệu thưa thớt, trạng thái mục tin mới, người dùng mới và từ đó nâng cao hơn nữa độ chính xác của những mục tin gợi ý cho người dùng. Sau khi xây dựng mô hình, chúng tôi tiến hành thực nghiệm trên các tập dữ liệu chuẩn nhằm xác định tính khả thi của mô hình đề xuất.

II. NHỮNG NGHIÊN CỨU LIÊN QUAN

Framework Hermes[5] là một công cụ cung cấp một cách tiếp cận dựa trên ngữ nghĩa (semantic-based) để lấy các tin tức liên quan trực tiếp hoặc gián tiếp. Các khái niệm mà người dùng có thể thích từ một miền ontology được gọi là cơ sở tri thức. Hermes News Portal (HNP) là một chương trình Java cài đặt Framework Hermes[5], chương trình này cho phép người sử dụng truy vấn các tin tức và xem các thông tin khác. HNP sử dụng thư viện Jena và suy luận dựa trên ngôn ngữ ontology OWL. Đối với truy vấn, HNP sử dụng SPARQL và tSPARQL[5], Đồng thời, có thêm thông số thời gian để thực hiện các truy vấn. Việc phân loại các bài báo được thực hiện bằng cách sử dụng GATE[7] và WordNet[2] ngữ nghĩa của từ vựng. Nghiên cứu sử dụng độ đo IF-IDF và hệ số tương tự Jaccard để tính toán độ tương tự cho các bài báo gợi ý cho người dùng.

Nghiên cứu [6] xây dựng hệ thống gợi ý tích hợp Quickstep[11] là sự kết hợp của AKT Ontology và hệ thống OntoCoPI đã chứng minh một cách tiếp cận mới của các tác giả để giảm "khởi đầu lạnh" của hệ thống. Tác giả chứng minh rằng việc sử dụng một Ontology để khởi động cho hồ sơ người dùng có thể làm giảm đáng kể tác động của hệ thống gợi ý về vấn đề "khởi đầu lạnh". Quickstep là một hệ thống gợi ý lai (hybrid recommender system), nó giải quyết các vấn đề thực tế của việc gợi ý bài báo khoa học trực tuyến để các nhà nghiên cứu có thể tìm thấy chúng. Hành động duyệt web của người dùng sẽ được âm thầm giám sát thông qua một máy chủ (proxy server) ở mỗi lần đăng nhập. Một thuật toán láng giềng gần được dùng để phân loại các URL đã duyệt dựa trên một tập dữ liệu huấn luyện là các bài báo mẫu. Hệ thống lưu trữ những bài báo mới trong một cơ sở dữ liệu trung tâm. Các phản hồi tường minh và các URL đã duyệt là thông tin cơ bản lưu sở thích của mỗi người dùng. Nghiên cứu trước đây của Quickstep sử dụng hồ sơ ban đầu

được xây dựng bằng tay, dựa trên dữ liệu phỏng vấn để đối phó với vấn đề "khởi đầu lạnh" của hệ thống. Tích hợp giữa Quickstep với Ontology AKT sẽ làm tự động hóa quá trình này và cho phép hạn chế tình trạng "khởi đầu lạnh" thực tế hơn với một số lượng lớn người dùng mới.

Ontology có thể được sử dụng để cải thiện hiệu quả tìm kiếm dựa trên nội dung, như trong OntoSeek của Guarino, N., Masolo, C. and Vetere [1999]. Người sử dụng của OntoSeek được điều hướng từ các ontology để xây dựng các truy vấn. Ontology cũng có thể được sử dụng để tự động xây dựng cơ sở tri thức từ các trang web, chẳng hạn như trong Web-KB của Craven, M. DiPasquo, D. Freitag và các đồng sự [1998]. Web-KB sử dụng phương pháp gán nhãn bằng tay cho các khái niệm tên miền và áp dụng các kỹ thuật máy học để phân loại các trang web mới. Trên cơ sở nắm bắt thông tin tự động cũng như sở thích của người sử dụng phục vụ cho việc gợi ý. Những hệ thống liên quan như CiteSeer của Bollacker, K.D., Lawrence, S., và Giles [1998], trong đó các tác giả sử dụng phương pháp tương tự dựa trên nội dung để giúp tìm kiếm các tài liệu nghiên cứu khoa học trong một thư viện kỹ thuật số thông qua độ đo tượng tự Jaccard và TF-IDF gợi ý cho người dùng.

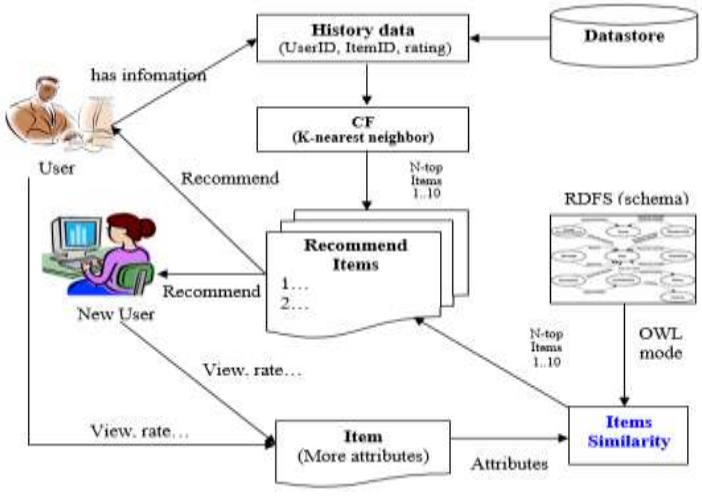
Ánh xạ giữa các Ontology như: Knowledge framework for Indian Medicinal Plants (KIMP): là một dự án quản lý những tri thức về cây thuốc của Ấn Độ được xây dựng từ cấu trúc Ontology thông qua giao diện người dùng và U.S. Medical Subject Heading (MeSH) là một thư viện Y khoa quốc gia nước Mỹ. MeSH được xây dựng nhằm quản lý bộ từ vựng đồng nghĩa trong lĩnh vực y khoa, quản lý một tập hợp của những thuật ngữ theo cấu trúc phân cấp (thứ bậc) và cho phép tìm kiếm những thuật ngữ ở các cấp độ khác nhau, được quản lý bằng Ontology [Vadivu, G. and S. Waheeta Hopper. 2012] [15]. Ánh xạ giữa KIMP và MeSH được thực hiện tự động thông qua hệ số Jaccard và JaroWinkler giúp xác định độ tương đồng cho các thuật ngữ y khoa chuẩn; đồng thời, giúp cải thiện khả năng tái sử dụng và phát hiện ra các mối quan hệ mới giữa các khái niệm.

Nhiều nghiên cứu tập trung vào khai thác ngữ nghĩa để nâng cao chất lượng cho kỹ thuật gợi ý của họ. Hầu hết trong số đó sử dụng phương pháp tương tự ngữ nghĩa (semantic similarity) để nâng cao hiệu suất của phương pháp tiếp cận dựa trên nội dung (CB), tuy vậy cũng có một số hệ thống sử dụng phương pháp lọc cộng tác dựa trên hồ sơ của người dùng lưu trữ trong Ontology. Ví dụ: ePaper[8] là một hệ thống gợi ý các bài báo khoa học, sử dụng các mối quan hệ kế thừa của các khái niệm trong miền để tính toán sự kết hợp giữa các khái niệm mô tả một mục tin và các khái niệm được thu thập từ sở thích của người dùng. Dự án âm nhạc FOAFing[3] là một hệ thống gợi ý nhạc sử dụng tiêu chuẩn vocabulary2 FOAF để thiết lập hồ sơ người dùng và khai thác các mô tả ngữ nghĩa của các bài hát, chủ yếu là các mối quan hệ của nghệ sĩ. FOAF tìm các bài hát tương tự như thói quen nghe nhạc của người dùng để thực hiện gợi ý. Một hệ thống gợi ý sử dụng các phương pháp suy luận ngữ nghĩa trong cả hai giai đoạn của quá trình gợi ý là AVATAR[1] là một hệ thống gợi ý kênh truyền hình sử dụng phương pháp lan truyền ngược (upward-propagation) và phương pháp tương tự ngữ nghĩa.

III. GIẢI PHÁP ĐỀ XUẤT

Trong phạm vi nghiên cứu này, chúng tôi đề xuất giải pháp tích hợp xử lý ngữ nghĩa vào hệ thống gợi ý; kỹ thuật này là sự kết hợp giữa kỹ thuật gợi ý truyền thống (ví dụ: Thực nghiệm chúng tôi sử dụng kỹ thuật lọc cộng tác dựa trên người dùng k-NNs Collaborative Filtering) và phương pháp suy diễn dữ liệu ngữ nghĩa giữa các mục tin lưu trữ bằng Ontology. Ý tưởng là từ các mục tin mà người dùng đang đọc/xem, thông qua mô hình ngữ nghĩa chúng tôi có thể xác định các mục tin tương tự (điều này có ưu điểm là người dùng có thể thậm chí không cần phải thực hiện thao tác đánh giá, xếp hạng hoặc phản hồi trở lại với hệ thống). Bằng việc sinh ra dữ liệu suy diễn từ mô hình ngữ nghĩa sẽ khắc phục được tình trạng thiếu thông tin từ phía người dùng, tình trạng dữ liệu hiếm hoặc tình trạng dữ liệu thưa mà các hệ thống gợi ý đang gặp khó khăn. Từ ý tưởng nêu trên, chúng tôi đề xuất mô hình tích hợp như sau:

a. Mô hình tích hợp



Hình 1. Mô hình tích hợp xử lý ngữ nghĩa vào Hệ thống gợi ý

Mô hình tích hợp xử lý ngữ nghĩa vào hệ thống gợi ý trình bày trong Hình 1. Trong đó, các trường hợp cần xử lý được trình bày dưới đây:

*** Trường hợp 1** (người dùng mới/khách): Người dùng không có thông tin trong hệ thống, tức là không có UserID.

- Bước 1 (gợi ý cơ bản): Hệ thống trả về Ntop mục tin gợi ý cho người dùng theo phương pháp xử lý "truyền thống" có thể là Ntop mục tin người dùng thích nhất, hay bán chạy nhất hoặc mới nhất – tuỳ vào mục tiêu cụ thể của từng ứng dụng.

- Bước 2: Người dùng chọn xem hoặc đánh giá trên 1 mục tin (mục tin ở đây có thể là kết quả của bước 1 hoặc do người dùng tìm được – tuỳ hệ thống thực tế). Hệ thống tiến hành xử lý tìm trong Ontology những mục tin có tương tự về mặt ngữ nghĩa (mục III.d) với mục tin người dùng đang thao tác; từ đó, trả về Ntop mục tin có độ tương đồng cao nhất gợi ý cho người dùng.

*** Trường hợp 2** (người dùng): Người dùng đã có thông tin trong hệ thống, tức là tồn tại UserID và thông tin lịch sử (History data).

- Bước 1: Hệ thống sử dụng giải thuật gợi ý "truyền thống" (như lọc cộng tác dựa trên người dùng k-NNs Collaborative Filtering) để đưa ra gợi ý Ntop mục tin mà người dùng có thể quan tâm.

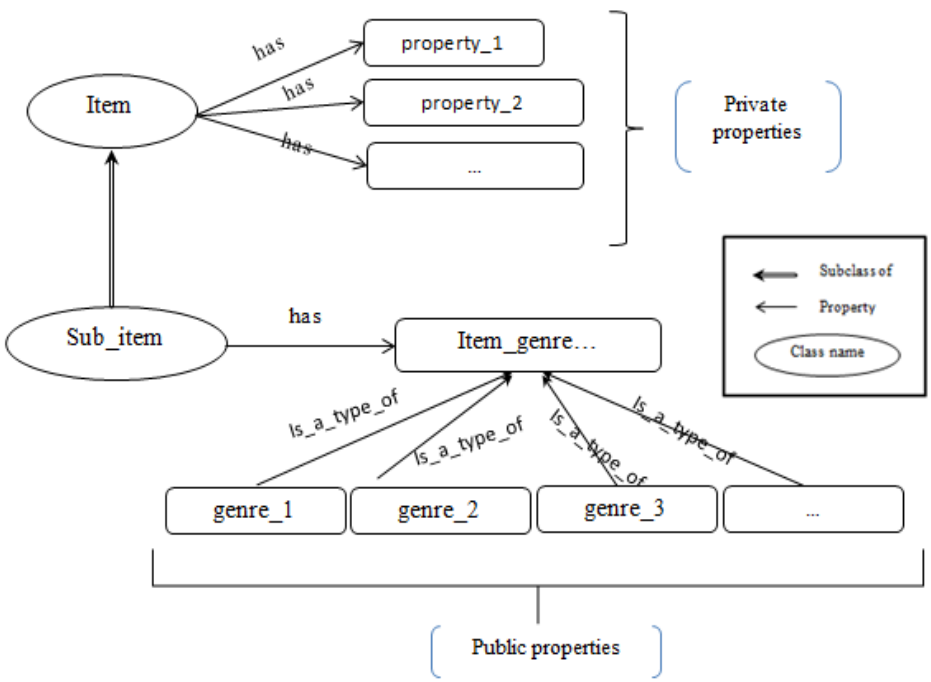
- Bước 2: Người dùng chọn xem hoặc đánh giá trên 1 mục tin (mục tin ở đây có thể là kết quả của bước 1 hoặc do người dùng tìm được – tuỳ hệ thống thực tế). Hệ thống tiến hành quá trình xử lý tìm trong Ontology những mục tin có tương tự về mặt ngữ nghĩa với mục tin người dùng đang thao tác; từ đó trả về Ntop mục tin có độ tương đồng cao nhất. Đồng thời, kết hợp với kết quả bước 1 trả về Ntop mục tin gợi ý cho người dùng tuỳ vào ứng dụng cụ thể.

→ Sự khác biệt chủ yếu giữa 2 trường hợp: Trong trường hợp 1: do hệ thống chưa tồn tại thông tin của người dùng, vì vậy hệ thống sử dụng dự đoán cơ bản để gợi ý ở bước 1 và bước 2 không thực hiện thao tác kết hợp kết quả dự đoán trả về cho người dùng. Trường hợp 2 thì ngược lại, ở bước 1 thực hiện gợi ý theo phương pháp "truyền thống" tuỳ vào hệ thống và bước 2 sẽ thực hiện thao tác kết hợp kết quả của bước 1 trả về cho người dùng.

b. Giải pháp xây dựng Ontology lưu trữ dữ liệu

Để lưu trữ các mục tin phục vụ cho việc xử lý các trường hợp theo mô Hình 1 mà bài viết đề cập bên trên, mô hình Ontology đề nghị như sau:

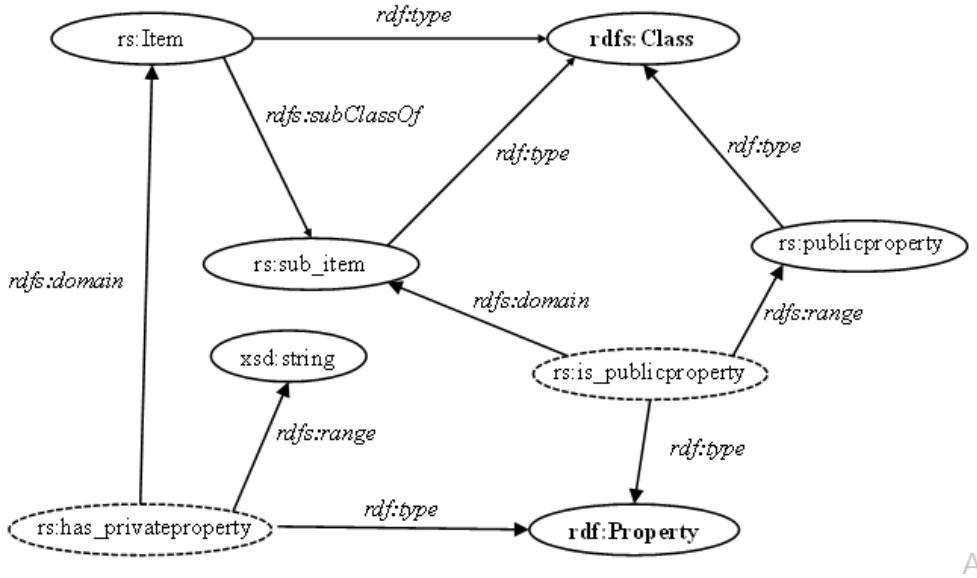
*** Mô hình Ontology:**



Hình 2. Mô hình Ontology

Để lưu trữ các mục tin phục vụ cho việc xử lý cho các trường hợp theo Hình 1. Mô hình Ontology tổng quát đề xuất như Hình 2; Trong đó, mỗi lớp Item có những thuộc tính: property_1, property_2…là những thuộc tính sẽ mang giá trị riêng cho mỗi mục tin (ví dụ: item_id, item_title…) và lớp Sub_item là lớp con (subclass) của lớp Item, Sub_item có các thuộc tính (Item_genre…) là thể loại của mục tin với các giá trị cụ thể (ví dụ: Comedy, Amination…)

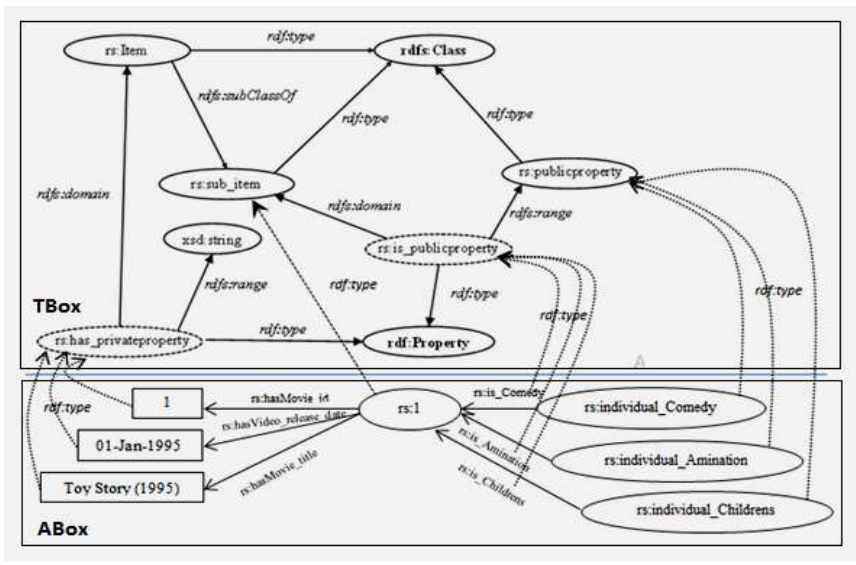
*** Cấu trúc đồ thị RDFS:**



Hình 3. Cấu trúc đồ thị RDFS

Với cấu trúc đồ thị RDFS Hình 3, thì ta nhận thấy: rdfs:Class là lớp của tất cả các lớp, rdfs:subClassOf chuyển thuộc tính của lớp cha rs:Item sang lớp mới rs:sub_item, rdfs:domain chỉ định miền của một thuộc tính, rdfs:range chỉ định phạm vi của thuộc tính, rdf:Property lớp của tất cả các thuộc tính, rdf:type: chỉ định lớp của lớp mới hoặc tài nguyên; Đặc biệt, với mỗi thể hiện của lớp Item ta sẽ có nhiều [privateproperty] và tương tự với mỗi subClass của lớp Item tức là lớp sub_item ta cũng có có nhiều [publicproperty] tuỳ vào từng tập dữ liệu cụ thể khi sử dụng để xây dựng Ontology.

*** Ví dụ minh hoạ:**



Hình 4. Ví dụ minh hoạ

Hình 4 là một ví dụ cụ thể của 1 khái niệm lưu trữ (mẫu tin đầu tiên trong tập dữ liệu MovieLens100k), cụ thể: (Movie_id=**1**, Movie_title=**Toy Story (1995)**, Video_release_date=**01-Jan-1995**, Movie_genre (attributes) = {**Comedy, Amination, Childrens**}.

*** Các bước tiến hành xây dựng Ontology**

Để có thể lưu trữ và truy vấn lại các mục tin đã được người dùng đánh giá, xếp hạng từ các tập dữ liệu đề xuất (mục IV.a), chúng tôi xây dựng Ontology theo cấu trúc Hình 3, các bước thực hiện như sau:

Bước 1. Tạo 1 Mô hình OWLModel rỗng.

Bước 2. Tạo lớp Item kiểu Class, các lớp [Privateproperties] và [Publicproperties]; lưu ý: [Privateproperties] thuộc miền Item và [Publicproperties] là kiểu [Class].

Bước 3. Tạo thể hiện của các lớp [Publicproperties], tức là các thể hiện cụ thể đối tượng của lớp.

Bước 4. Đọc từng mục tin từ dữ liệu truyền vào, mỗi ItemID là 1 mục tin (chính là thể hiện của lớp Item) và subClass là sub_item; mỗi thuộc tính của mục tin chính là các thuộc tính của thể hiện vừa tạo; Lưu ý: mỗi [Publicproperties] là thuộc tính của thể hiện vừa tạo với giá trị là các thể hiện của lớp [Publicproperties] được tạo ở bước 3;

c. Phương pháp chuẩn hoá dữ liệu đầu vào cho mô hình Ontology

Để thực hiện việc lưu trữ dữ liệu mục tin trong cấu trúc Ontology theo Hình 3, cần thiết chuẩn hoá dữ liệu đầu vào cho mô hình này, cụ thể chúng tôi thực nghiệm trên dữ liệu dạng như sau:

[ItemID]|[privateproperties]|[publicproperties]

Và một tập tin lưu trữ cấu trúc thuộc tính chi tiết của dữ liệu mục tin. Ví dụ: cấu trúc thuộc tính của tập dữ liệu MovieLens (u.item) cụ thể như sau (tập dữ liệu này sẽ được giới thiệu chi tiết trong phần thực nghiệm):

movie_id|movie_title|video_release_date|IMDb_URL|unknown|Action|Adventure|…

Từ cấu trúc của dữ liệu ta tiến hành thực hiện các bước trong phần "Giải pháp xây dựng Ontology lưu trữ dữ liệu" (mục III.b).

d. Cách tính độ tương tự giữa các mục tin lưu trữ trong Ontology

Để thực nghiệm chúng tôi sử dụng chỉ số Jaccard để thực hiện đo độ tương tự giữa các mục tin lưu trữ trong cấu trúc Ontology. Chỉ số Jaccard, còn được gọi là hệ số tương tự Jaccard – (Jaccard Similarity), ban đầu được đặt là hệ số de communauté đề xuất bởi Paul Jaccard, là một hệ số thống kê được sử dụng để so sánh sự giống nhau và đa dạng của các bộ mẫu (sample sets). Hệ số Jaccard đo tương đồng giữa các bộ mẫu hữu hạn và được định nghĩa là kích thước của phần giao (intersection) chia cho kích thước của phần hợp (union) của các bộ mẫu, cụ thể:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

(Nếu cả A và B đều rỗng thì ta định nghĩa J(A,B) = 1), Trong đó:

$$0 \leq J(A,B) \leq 1$$

Ví dụ: Xét hai bộ A = {0, 1, 2, 5, 6} và B = {0, 2, 3, 5, 7, 9}. Độ tương tự Jaccard được xác định như sau:

$$J(A,B) = |A \cap B| / |A \cup B| = |\{0, 2, 5\}| / |\{0, 1, 2, 3, 5, 6, 7, 9\}| = 3 / 8 = 0.375$$

Chúng tôi sử dụng hệ số tương tự Jaccard nhằm mục đích lượng hoá độ tương tự giữa các mục tin lưu trữ trong Ontology, kết quả tìm được những mục tin có độ tương tự cao nhất (tức J(A,B) = 1) với mục tin A (người dùng đang xem) trả về cho hệ thống gợi ý. Với những hệ thống cụ thể có thể sử dụng những hệ số tương tự khác, tuỳ tình hình thực tế.

e. Phương pháp tích hợp Ontology vào hệ thống

Để kết hợp Ontology vào hệ thống RS, cần phải truyền vào thuật toán tích hợp Hình 5 các tham số: userID, itemID và tham số ktop mục tin cần trả về gợi ý cho người dùng, thuật toán cụ thể như sau:

```
1: procedure INTEGRATIONRECOMMEND(userID, itemID, ktop)
// Let L[ktop] be return results of the traditional RS for userID
// Let M[ktop] and R[ktop] be return results of Similarity for itemID and Procedure
2: R ← L
3: for k ← 1 to ktop do
4:     R[ktop + k] ← M[k]
5: end for
6: for j ← 1 to 2*ktop – 1 do
7:   biggest ← j
8:   for i ← j + 1 to 2*ktop do
9:        if R[i] == R[biggest] then remove(R, R[i])    // remove(from, what)
10:      if R[i]  > R[biggest] then
11:         biggest ← i
12:   end for
13:   R[j] ↔ R[biggest]
14: end for
15: return R[ktop]
16: end procedure
```

Hình 5. Giải thuật tích hợp Ontology vào hệ thống

Hàm trả về danh sách ktop mục tin có dự đoán lớn nhất để gợi ý cho người dùng có userID cụ thể truyền vào thuật toán, trên cơ sở kết hợp của những mục tin gợi ý cho người dùng này từ mô hình gợi ý truyền thống và những mục tin có độ tương tự với mục tin mà người dùng đang đọc/xem (trong thực nghiệm chúng tôi sử dụng độ tương tự

bằng 1 cho các mục tin lưu trữ trong Ontology thông qua hệ số tương tự Jaccard); Từ danh sách của những mục tin kết hợp từ 2 mô hình này, chúng tôi tiến hành công đoạn loại bỏ những mục tin trùng lắp, đồng thời tiến hành thêm việc sắp xếp lại danh sách kết quả này theo thứ tự giảm dần dựa theo tiêu chí đánh giá (rating) của mục tin; sau công đoạn sắp xếp, thuật toán sẽ trả về ktop mục tin gợi ý cho người dùng (userID).

f. Phương pháp suy diễn dữ liệu thông qua mô hình Ontology

Mục đích của việc suy diễn dữ liệu là nhằm xây dựng nên tập dữ liệu huấn luyện cho mô hình tích hợp mà chúng tôi đề xuất; Với mô hình gợi ý truyền thống, chúng tôi tiến hành công việc đánh giá độ lỗi của giải thuật thông qua 2 độ đo: Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE) [Chai, T., Draxler, R.R.R, 2014] theo phương thức: "3-fold cross validation" [Geisser, Seymour, 1993 và Kohavi, Ron, 1995]; Với mô hình tích hợp, chúng tôi tiến hành công đoạn suy diễn dữ liệu thông qua mô hình Ontology để xây dựng nên tập dữ liệu huấn luyện; Trên cơ sở dữ liệu suy diễn, chúng tôi tiến hành xây dựng lại mô hình đánh giá độ lỗi của giải thuật tích hợp với cùng độ đo và phương thức đã dùng trong đánh giá như mô hình truyền thống, thuật toán suy diễn dữ liệu được thực hiện qua 2 giai đoạn, cụ thể như sau:

- Giai đoạn 1: chuẩn bị

+ Chuẩn bị dữ liệu để suy diễn (dữ liệu huấn luyện của giải thuật gợi ý truyền thống).

+ Sắp xếp lại dữ liệu (ASC/DESC) theo tiêu chí UserID (ví dụ: UserID trong u.data của MovieLens).

+ Chuẩn bị Ontology lưu trữ các mục tin theo cấu trúc (mục III.c). Giai đoạn chuẩn bị dữ liệu hoàn tất, ta tiến hành sang giai đoạn 2.

- Giai đoạn 2: Suy diễn dữ liệu.

```
1: procedure SEMDATA(D^rating, OWLmodel)
      // Let R be return results of Procedure
2: R ← null
3: size ← 0
4: for each item i from D^rating do
5:    M be return results of Similarity for i in OWLModel, even i i with rating is the globalAVG item's
rate for current user.
6:    n ← lengh of M
7:    for k ← 1 to n do
8:           R[size+k] ← M[k]
9:    end for
10:   size ← size + n
11: end for
      // remove dupplicate items on R[size]
12: for i ← 1 to size − 1 do
13:    for j ← i + 1 to size do
14:       if R[i] == R[j] then remove(R, R[j])    // remove(from, what)
15:    end for
16: end for
17: return R
18: end procedure
```

Hình 6. Giải thuật suy diễn dữ liệu

Thuật toán suy diễn dữ liệu Hình 6 thực hiện công việc như sau: Với mỗi mục tin i tương ứng trong tập dữ liệu huấn luyện của mô hình gợi ý truyền thống D^{train} (các mục tin trong D^{train} đã được sắp xếp lại theo tiêu chí userID để truyền vào thuật toán) chúng tôi tiến hành tìm trong Ontology những mục tin tương tự với i (trong thực nghiệm chúng tôi tìm những mục tin có hệ số tương tự Jaccard bằng 1 so với i và giá trị đánh giá là giá trị trung bình đánh giá trên tất cả mục tin - globalAVG) để xây dựng tập dữ liệu huấn luyện cho mô hình tích hợp; sau khi tất cả các mục tin tương tự với i (lưu trữ trong ontology) đã được tìm thấy, chúng tôi tiến hành loại bỏ những mục tin trùng nhau theo tiêu chí userID-itemID là duy nhất. Kết quả thuật toán trả về danh sách những mục tin sau suy diễn theo cấu trúc userID\t itemID\t rating.

Sau công đoạn suy diễn dữ liệu, chúng tôi tiến hành xây dựng lại mô hình huấn luyện trên dữ liệu suy diễn tương tự như phương pháp áp dụng cho mô hình dự đoán truyền thống với cùng tập dữ liệu kiểm tra; Từ đó, chúng tôi thực hiện lại công việc đánh giá cho mô hình tích hợp. Thu thập kết quả đánh giá của mô hình truyền thống và mô hình tích hợp ta có được kết quả thực nghiệm để thực hiện việc so sánh tính khả thi của mô hình đề xuất.

IV. KẾT QUẢ THỰC NGHIỆM

a. Dữ liệu dùng đánh giá

- Tập dữ liệu của hệ thống gợi ý phim MovieLens

Lưu trữ lại địa chỉ: http://grouplens.org/datasets/movielens/ cụ thể: Tập dữ liệu này được sử dụng bởi nhiều nhà nghiên cứu: một số nhà nghiên cứu dùng tập dữ liệu này để kiểm tra những đánh giá kinh nghiệm của bản thân họ; một

số nhà nghiên cứu khác sử dụng tập dữ liệu này để nghiên cứu các hệ thống gợi ý áp dụng những kỹ thuật khác nhau [L. Candillier, K. Jack, F. Fessant, and F. Meyer. 2009]; ngoài ra GroupLens còn cung cấp nhiều phiên bản khác nhau của các bộ dữ liệu, cụ thể: MovieLens 100k, MovieLens 1M và MovieLens 10M.

Dữ liệu sử dụng trong bài viết này chúng tôi đề xuất sử dụng tập MovieLens 100K; tập dữ liệu này có 100.000 đánh giá được thực hiện bởi 943 người dùng trên số lượng 1.682 phim; mỗi người dùng có đánh giá ít nhất 20 phim và đánh giá được gán 1 (tệ) … 5 (tuyệt vời).

*** Cấu trúc của tập dữ liệu Movielens:**

- u.item (1,682 items); [movie id | movie title | release date | video release date | IMDb URL | unknown | Action | Adventure | Animation | Children's | Comedy | Crime | Documentary | Drama | Fantasy | Film-Noir | Horror | Musical | Mystery | Romance | Sci-Fi | Thriller | War | Western].

- u.data (100,000 rate/943 users): [user id | item id | rating | timestamp].

- **Tập dữ liệu MovieTweetings:** tập dữ liệu gợi ý các phim của MovieTweetings: lưu tại địa chỉ (https://github.com/sidooms/MovieTweetings).

MovieTweetings là một bộ dữ liệu bao gồm các đánh giá về phim được thu thập trên website của Twitter, tập dữ liệu này là kết quả của công trình nghiên cứu được chỉ đạo bởi Simon Dooms và các cộng sự của ông tại Đại học Ghent nước Bỉ. Tập dữ liệu được tạo ra nhằm cung cấp thông tin cho hoạt động của hệ thống RecSys (ACM RecSys conferences), để phục vụ công tác nghiên cứu và thử nghiệm MovieTweetings được phân thành những phân đoạn dữ liệu cụ thể: 10k có nghĩa là tập dữ liệu thu thập trước 10.000 đánh giá, 20k nghĩa là dữ liệu thu thập trước 20.000 đánh giá và tương tự đến tập dữ liệu 200k; cụ thể có:

*** Cấu trúc của tập dữ liệu MovieTweetings:**

- Items.dat (3,906 items): [movie_id::movie_title(movie_year)::genre| genre|genre|…]

- Ratings.dat (10,000 rates): [user_id::movie_id::rating::rating_timestamp]

Tập dữ liệu sử dụng trong bài viết này chúng tôi đề xuất sử dụng tập MovieTweetings 10K; với thông tin cơ bản như bảng trên; với đánh giá được gán từ 1…10; Tuy nhiên để có thể sử dụng được tập dữ liệu này cần thiết phải chuẩn hoá cấu trúc tập tin Items.dat làm đầu vào cho mô hình xử lý.

- **Tập dữ liệu Restaurant & Consumer data (RCData):** được cung cấp bởi: Rafael Ponce Medellín, Juan Gabriel González Serna và Blanca Vargas-Govea, bộ môn khoa học máy tính, Trung tâm Nghiên cứu và Phát triển Chiến lược Quốc gia CENIDET (México) lưu trữ tại địa chỉ https://archive.ics.uci.edu/ml/machine-learning-databases/00232/

Tập dữ liệu này được thu thập từ một phần của hệ thống gợi ý Nhà hàng theo đánh giá của khách hàng tại các thành phố trên đất nước México, mục tiêu tạo ra một danh sách Ntop nhà hàng tốt nhất theo bình chọn của khách hàng. Tập dữ liệu này đã được thử nghiệm bằng 2 phương pháp: Lọc cộng tác và ngữ cảnh, cụ thể:

→ Kỹ thuật lọc cộng tác chỉ sử dụng 1 tập tin rating_final.csv.

Tiếp cận theo hướng ngữ cảnh thì sử dụng 8 tập tin còn lại trong tập dữ liệu. Để đánh giá giải thuật, trong phạm vi bài viết này chỉ sử dụng 2 tập tin trong tập dữ liệu, cụ thể:

*** Cấu trúc của tập dữ liệu RCData:**

- rating_final.csv (1,161 rates): [userID, placeID, rating, food_rating, service_rating]

- geoplaces2.csv (130 lines): [placeID, latitude, longitude, the_geom_meter, name, address, city, state, country, fax, zip, alcohol, smoking_area, dress_code, accessibility, price, url, Rambience, franchise, area, other_services]

Các thuộc tính thuộc các miền giá trị cụ thể như sau:

- alcohol [No_Alcohol_Served, Wine_Beer, Full_Bar]

- smoking_area [none, only_at_bar, permitted, section, not_permitted]

- dress_code [informal, casual, formal]

- accessibility [no_accessibility, completely, partially]

- price [medium, low, high]

- rambience [familiar, quiet]

- franchise [t, f]

- area [open, closed]

- other_services [none, internet, variety]

Từ cấu trúc tập tin geoplaces2.csv như trên ta nhận thấy có 09 thuộc tính phụ thuộc miền giá trị; nhận định điều này nhằm mục đích: để có thể sử dụng được mô hình Ontology như Hình 2 trên tập tin dữ liệu này thì cần thiết phải chuẩn hoá các giá trị thuộc miền của các thuộc tính thành các [Publicproperties] của mục tin, tức là các placeID trong tập dữ liệu này và tương tự khi sử dụng các tập dữ liệu khác (nếu có).

b. Các độ đo dùng trong thực nghiệm

Trong thực nghiệm, chúng tôi sử dụng 2 độ đo lỗi thường được các nhà nghiên cứu trong lĩnh vực RSs hay sử dụng để so sánh giải thuật, đó là các độ đo lỗi: Root Mean Squared Error (RMSE) và Mean Absolute Error (MAE); Trong đó, độ đo MAE dùng trong dự đoán mục tin (Item Prediction) và RMSE dùng trong dự đoán xếp hạng (Rating Prediction) [Gunawardana, A và Shani, G, 2009]. Đặc biệt, độ đo lỗi RMSE cũng là độ đo chuẩn được sử dụng trong các kỳ thi của hội nghị kDD Cup về khai thác dữ liệu - khám phá tri thức; Cả 2 độ đo này phù hợp với lĩnh vực nghiên cứu mà chúng tôi đề xuất; Theo nghiên cứu của các tác giả [Chai, T., Draxler, R.R.R, 2014] cho rằng không có phương pháp đánh giá nào là chính xác tuyệt đối trên các giải thuật vì thế các tác giả đề xuất rằng: để đánh giá độ lỗi của giải thuật gợi ý trên dữ liệu dạng offline các nhà nghiên cứu nên sử dụng cả 2 độ đo lỗi RMSE và MAE; Để kết quả đánh giá các mô hình mang tính chính xác và thuyết phục hơn, chúng tôi đề xuất dùng nghi thức kiểm tra "3-fold cross validation" [Geisser, Seymour, 1993 và Kohavi, Ron, 1995] với các độ đo lỗi RMSE và MAE nêu trên, cụ thể như sau:

❖ **RMSE (Root Mean Squared Error)**

RMSE là độ đo phổ biến mà cộng đồng người dùng trong lĩnh vực máy học thường sử dụng, kể cả trong các kỳ thi giải thuật của RSs; Mặc dù có nhiều phương pháp khác nhau mà chúng ta có thể sử dụng để đánh giá giải thuật gợi ý như: F-Meansure, Area Under the ROC curve (AUC),… với mỗi phương pháp đánh giá sẽ thích hợp cho từng lĩnh vực cụ thể (ví dụ: F-Meansure và AUC được dùng trong truy tìm thông tin và phân lớp; MAE dùng trong dự đoán mục tin và RMSE dùng trong dự đoán xếp hạng [Gunawardana, A và Shani, G, 2009]); RMSE được xác định bằng công thức:

$$RMSE = \sqrt{\frac{1}{|D^{test}|}\sum_{u,i,r\in D^{test}}(r_{ui} - \hat{r}_{ui})^2}$$

Trong đó: Dtest $\subseteq U \times I \times R$ là tập dữ liệu kiểm thử; U: tập người dung (user);

I: Tập item; r_{ui}: giá trị thực tế; $\hat{r}_{ui}$: giá trị dự đoán.

❖ **Độ đo MAE (mean absolute error)**

MAE là một đại lượng dùng để đo cường độ trung bình của các sai số trong một tập hợp các dự báo, mà không xem xét hướng của chúng, công thức MAE cụ thể như sau:

$$MAE = \frac{1}{|D^{test}|}\sum_{u,i,r\in D^{test}}\left|(r_{ui} - \hat{r}_{ui})\right|$$

Trong đó, các ký hiệu được mô tả như sử dụng trong RMSE.

c. Kết quả thực nghiệm

Bảng 1. Kết quả đánh giá độ lỗi trên các tập dữ liệu chuẩn

Các tập dữ liệu	Các độ đo đánh giá			
	RMSE error		MAE error	
	OldData	*SemData*	*OldData*	*SemData*
MovieLens 100k	1,1181	1,0635	0,8805	0,8475
MovieTweetings 10k	2,9434	2,0489	2,1949	1,5440
RCData	2,6314	2,3415	2,0809	1,8785

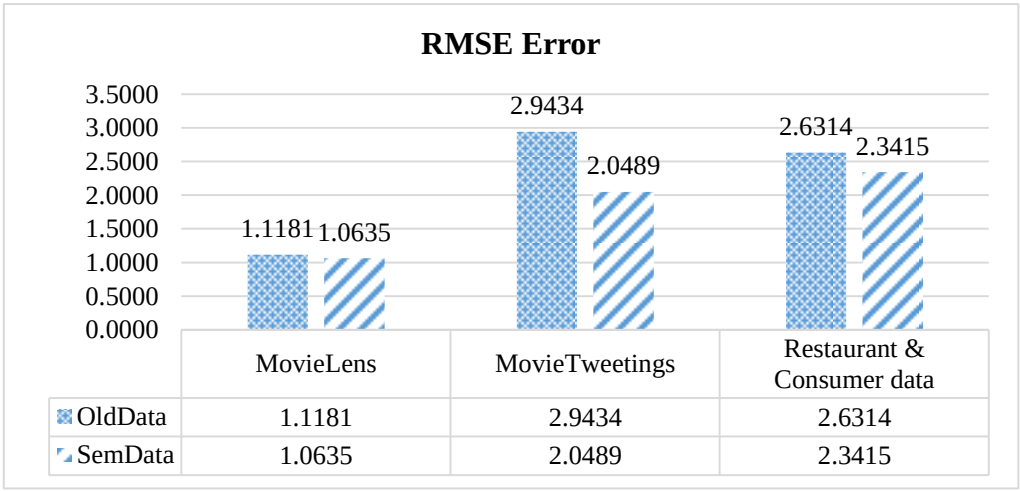
Bảng 1. Trình bày kết quả chạy thực nghiệm trên 3 tập dữ liệu chuẩn bằng giải thuật User-kNN (k=50) với các tập dữ liệu (tập kiểm tra – tập huấn luyện) có số lượng mẫu tin và độ lỗi cụ thể như sau:

*** Tập dữ liệu MovieLens 100k ta có:** tập kiểm tra: 29.942 mẫu tin; OldData: tập huấn luyện: 70.058 mẫu tin; SemData: tập huấn luyện: 897.098 mẫu tin; với cùng tập kiểm tra, độ lỗi RMSE của OldData và SemData tương ứng 1,1181 và 1,0635 và lỗi MAE tương ứng là 0,8805 và 0,8475.

*** Tập dữ liệu MovieTweetings 10k ta có:** tập kiểm tra: 2.569 mẫu tin; OldData: tập huấn luyện: 7.431 mẫu tin; SemData: tập huấn luyện: 307.158 mẫu tin; với cùng tập kiểm tra, độ lỗi RMSE của OldData và SemData tương ứng 2,9434 và 2,0489 và lỗi MAE tương ứng là 2,1949 và 1,5440.

*** Tập dữ liệu RCData ta có:** tập kiểm tra: 336 mẫu tin; OldData: tập huấn luyện: 825 mẫu tin; SemData: tập huấn luyện: 3.421 mẫu tin; với cùng tập kiểm tra, độ lỗi RMSE của OldData và SemData tương ứng 2,6314 và 2,3415 và lỗi MAE tương ứng là 2,0809 và 1,8785.

Để trực quan cho kết quả thực nghiệm trình bày trong Bảng 1, chúng tôi tiến hành xây dựng biểu đồ so sánh kết quả Hình 7 như sau:



Hình 7. So sánh độ lỗi RMSE trên mô hình đề xuất (SemData) và phương pháp truyền thống (OldData)

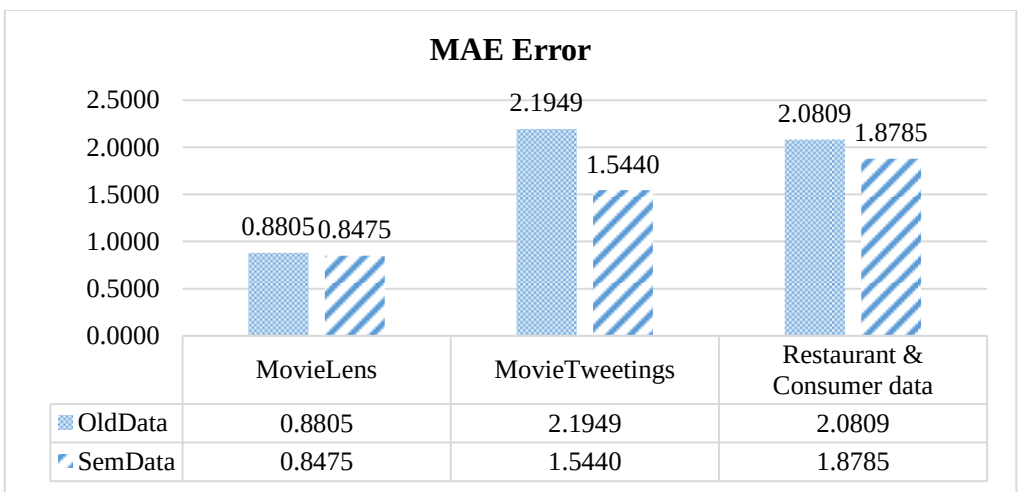
*** Nhận xét:** Qua biểu đồ Hình 7. thể hiện độ lỗi RMSE trên các tập dữ liệu trong 2 trường hợp: OldData (hệ thống khi không kết hợp xử lý ngữ nghĩa) và SemData (hệ thống khi kết hợp xử lý ngữ nghĩa) nhận thấy:

- **Tập dữ liệu MovieLens 100k:** độ lỗi RMSE trên OldData là: 1,1181; và SemData là: 1,0635; → Mô hình tích hợp ngữ nghĩa có độ lỗi RMSE giảm 0,0546 so với mô hình không tích hợp ngữ nghĩa;

- **Tập dữ liệu Movietweeting 10k:** độ lỗi RMSE trên OldData là: 2,9434; và SemData là: 2,0489; → Mô hình tích hợp ngữ nghĩa có độ lỗi RMSE giảm 0,8945 so với mô hình không tích hợp ngữ nghĩa;

- **Tập dữ liệu RCData:** độ lỗi RMSE trên OldData là: 2,6314; và SemData là: 2,3415; → Mô hình tích hợp ngữ nghĩa có độ lỗi RMSE giảm 0,2988 so với mô hình không tích hợp ngữ nghĩa;

⇒ **Tổng quan trên 3 tập dữ liệu:** độ đo lỗi RMSE trên dữ liệu khi kết hợp xử lý ngữ nghĩa (SemData) thấp hơn so với dữ liệu khi không kết hợp xử lý ngữ nghĩa (OldData); Đặc biệt, ta nhận thấy độ lỗi của tập dữ liệu MovieTweeting và RCData khá cao trên cả 2 tập dữ liệu OldData và SemData do nguyên nhân mỗi người dùng đánh giá cho các mục tin trong tập dữ liệu huấn luyện ít, vì thế độ lỗi của mô hình sẽ ở mức cao.



Hình 8. So sánh độ lỗi MAE trên mô hình đề xuất (SemData) và phương pháp truyền thống (OldData)

*** Nhận xét:** Qua biểu đồ Hình 8, thể hiện độ lỗi MAE trên các tập dữ liệu trong 2 trường hợp: OldData (hệ thống khi không kết hợp xử lý ngữ nghĩa) và SemData (hệ thống khi kết hợp xử lý ngữ nghĩa) nhận thấy:

- **Tập dữ liệu MovieLens 100k:** độ lỗi MAE trên OldData là: 0,7188; và SemData là: 0,7823; → Mô hình tích hợp ngữ nghĩa có độ lỗi MAE tăng 0,0635 so với mô hình không tích hợp ngữ nghĩa; Nguyên nhân độ lỗi MAE của mô hình tích hợp tăng cùng nguyên nhân của độ lỗi RMSE trình bày bên trên.

- **Tập dữ liệu Movietweeting 10k:** độ lỗi MAE trên OldData là: 2,0970; và SemData là: 1,5978; → Mô hình tích hợp ngữ nghĩa có độ lỗi MAE giảm 0,4992 so với mô hình không tích hợp ngữ nghĩa;

- **Tập dữ liệu RCData:** độ lỗi MAE trên OldData là: 1,2981; và SemData là: 1,2860; → Mô hình tích hợp ngữ nghĩa có độ lỗi MAE giảm 0,0121 so với mô hình không tích hợp ngữ nghĩa;

⇒ **Tổng quan trên 3 tập dữ liệu**: độ đo lỗi MAE trên dữ liệu khi kết hợp xử lý ngữ nghĩa (SemData) thấp hơn so với dữ liệu khi không kết hợp xử lý ngữ nghĩa (OldData); Đặc biệt, ta nhận thấy độ lỗi của tập dữ liệu MovieTweeting khá cao trên cả 2 tập dữ liệu OldData và SemData cùng nguyên nhân tương tự như RMSE;

→ **Nhận định tổng quan trên kết quả thực nghiệm:** Nhìn chung kết quả chạy thực nghiệm trên các tập dữ liệu khẳng định tính khả thi của mô hình tích hợp ngữ nghĩa vào hệ thống gợi ý mà chúng tôi đề xuất, thực nghiệm trên 3 tập dữ liệu với giải thuật User-kNN (k=50) chứng minh độ lỗi của mô hình gợi ý tích hợp do chúng tôi đề nghị thấp hơn so với mô hình truyền thống (không kết hợp xử lý ngữ nghĩa) kể cả 2 độ đo lỗi RMSE và MAE.

V. KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Nghiên cứu này đề xuất giải pháp tích hợp mô hình ngữ nghĩa vào hệ thống gợi ý. Phương pháp này có thể khai thác các điểm tương đồng ngữ nghĩa khi thực hiện gợi ý cho những tình huống cụ thể, sử dụng kết quả của các kỹ thuật gợi ý hiện tại và tương đồng ngữ nghĩa về mặt nội dung của mục tin lưu trữ bằng Ontology để thực hiện gợi ý cho người dùng sử dụng hệ thống. Từ kết quả thực nghiệm trên các tập dữ liệu chuẩn cho thấy mô hình tích hợp chúng tôi đề xuất có độ lỗi thấp hơn so với phương pháp truyền thống. Việc ứng dụng mô hình tích hợp này vào các hệ thống gợi ý với đối tượng gợi ý cho người dùng là những mục tin có chia sẽ thuộc tính là hoàn toàn khả thi.

Nghiên cứu tiếp theo sẽ ứng dụng giải pháp tích hợp ngữ nghĩa vào hệ thống gợi ý để giải quyết tình trạng "khởi đầu lạnh" của hệ thống gợi ý.

TÀI LIỆU THAM KHẢO

[1]. Blanco-Fernández, Y. et al. 2008. A flexible semantic inference methodology to reason about user preferences in knowledge-based recommender systems. Knowledge-Based Systems 21 (4), 305-320.

[2]. C. Fellbaum, editor. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, 1998.

[3]. Celma, O. and Serra, X. 2008. FOAFing the music: Bridging the semantic gap in music recommendation. Web Semantics: Science, Services and Agents on the World Wide Web 6 (4), 250-256.

[4]. Craven, M. DiPasquo, D. Freitag, D. McCallum, A. Mitchell, T. Nigam K. and Slattery, S. Learning to Extract Symbolic Knowledge from the World Wide Web, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI-98), 1998.

[5]. F. Frasincar, J. Borsje, and L. Levering. A Semantic Web-Based Approach for Building Personalized News Services. International Journal of E-Business Research, 5(3):35–53, 2009.

[6]. Guarino, N., Masolo, C. and Vetere, G. OntoSeek: Content-Based Access to the Web, IEEE Intelligent Systems, Vol. 14,No. 3, May/June 1999.

[7]. H. Cunningham. GATE, a General Architecture for Text Engineering. Computers and the Humanities, 36:223–254, 2002.

[8]. Maidel, V., Shoval, P., Shapira, B., Taieb-Maimon, M. 2008. Evaluation of an ontology-content based filtering method for a personalized newspaper. RecSys'08: Proceedings of the 2008, 91-98.

[9]. Middleton, N. R. Shadbolt, and D. C. D. Roure. Ontological User Profiling in Recommender Systems. ACM Transactions on Information Systems, 22(1):54–88, 2004.

[10]. Middleton, Stuart E., Shadbolt, N.R. and De Roure, D.C. (2004) Ontological User Profiling in Recommender Systems. ACM Transactions on Information Systems (TOIS), 22, (1), 54-88.

[11]. Middleton, S. E., De Roure, D. C., and Shadbolt, N.R.Capturing Knowledge of User Preferences: ontologies on recommender systems, In Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), Oct 2001, Victoria, B. C. Canada.

[12]. Nguyễn Thái Nghe, Nguyễn Hùng Dũng (2014): Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác. Tạp chí Khoa học Trường Đại học Cần Thơ, số 31a (2014), trang 36-51. ISSN: 1859-2333.

[13]. Nguyễn Thái Nghe, Triệu Vĩnh Viêm, Triệu Yến Yến (2013): Xây dựng hệ thống gợi ý phim dựa trên mô hình nhân tố láng giềng. Số chuyên đề: Công nghệ Thông tin (2013): 170-179, Tạp chí Khoa học Trường Đại học Cần Thơ, ISSN: 1859-2333.

[14]. Rocha, C., Schwabe, D., Aragao, M.P.: A hybrid approach for searching in the semantic web. In: Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA (2004) 374–383

[15]. Vadivu, G., and Hopper, W.,"Ontology Mapping of Indian Medicinal Plants with Standardized Medical Terms", Journal of Computer Science, 8(9), pp.1576-1584, 2012.

[16]. MovieTweeting dataset from http://grouplens.org/datasets/movielens/ and MovieTweetings dataset from https://github.com/sidooms/MovieTweetings/

[17]. Restaurant & consumer dataset: https://archive.ics.uci.edu/ml/machine-learning-databases/00232/

A SEMANTIC INTEGRATION APPROACH FOR RECOMMENDER SYSTEMS

Huynh Thanh Tai, Nguyen Huu Hoa, Huynh Minh Tri, Nguyen Thai Nghe

ABSTRACT *— Recommender system (RS) often creates a list of recommendation items to users in one of two ways: content-based filtering and collaborative filtering. Content-based filtering is the approach based on specific analysis on the content of the items that the user has selected in the past and recommend for users of specific items with similar content. Collaborative filtering is an approach that uses groups of users who preferred the same items with the current user. All approaches above use available data to build prediction models. In this article, we introduce an approach to integrate semantic model into recommender systems. This approach can generate semantic data from the proposed Ontology model. Experimental results shows that, by using inferred data from the Ontology, the RS models can improve the prediction results.*

Keywords *— Recommender systems, semantic web, integrated solutions, hybrid systems.*