

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH**

TIỂU LUẬN

ĐỀ TÀI NGHIÊN CỨU HỆ TƯ VẤN THÔNG TIN

**NGHIÊN CỨU KỸ THUẬT VÀ ỨNG DỤNG ĐỀ
XUẤT THÔNG TIN DỰA TRÊN NGỮ NGHĨA, KẾT
HỢP ĐỒ THỊ TRI THỨC ĐỂ BIỂU DIỄN KẾT QUẢ
DỮ LIỆU ĐỀ XUẤT**

TP. Hồ Chí Minh, 5/2023

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH**

TIỂU LUẬN

ĐỀ TÀI NGHIÊN CỨU HỆ TƯ VẤN THÔNG TIN

**NGHIÊN CỨU KỸ THUẬT VÀ ỨNG DỤNG ĐỀ
XUẤT THÔNG TIN DỰA TRÊN NGŨ NGHĨA, KẾT
HỢP ĐỒ THỊ TRI THỨC ĐỂ BIỂU DIỄN KẾT QUẢ
DỮ LIỆU ĐỀ XUẤT**

Thành viên nhóm:

Trịnh Hoàng Tùng	MSSV: 46.01.104.211
Nguyễn Trịnh Thành	MSSV: 46.01.104.169
Phạm Quốc Anh Quân	MSSV: 46.01.104.146
Hồ Huy Phúc	MSSV: 43.01.104.133

Lớp học phần: 2221COMP131001 – Hệ tư vấn thông tin

Người hướng dẫn: ThS. Trần Thanh Nhã

TP Hồ Chí Minh, 5/2023

MỤC LỤC

MỤC LỤC	3
DANH MỤC HÌNH ẢNH.....	5
LỜI MỞ ĐẦU	7
CHƯƠNG 1. TỔNG QUAN	8
1.1. Đặt vấn đề.....	8
1.2. Mục tiêu cụ thể	10
1.3. Phạm vi đề tài	10
CHƯƠNG 2. CƠ SỞ LÝ THUYẾT	12
2.1. Hệ tư vấn dựa trên ngữ nghĩa - Semantics Recommendation System ...	12
2.1.1. Semantics Recommendation System là gì?	12
2.1.2. Cơ bản về hệ đề xuất thông tin dựa trên ngữ nghĩa	12
2.1.3. Nguyên lý hoạt động	13
2.1.4. Ưu điểm của kỹ thuật tư vấn dựa trên ngữ nghĩa.....	24
2.1.5. Nhược điểm của kỹ thuật tư vấn dựa trên ngữ nghĩa.....	25
2.2. Đồ thị tri thức – Knowledge Graph	26
2.2.1. Đồ thị tri thức là gì	26
2.2.2. Cơ sở dữ liệu dạng đồ thị tri thức	27
2.2.3. Neo4j	29
2.2.4. Ưu điểm của việc ứng dụng cơ sở dữ liệu đồ thị tri thức	31
2.2.5. Nhược điểm của việc ứng dụng cơ sở dữ liệu đồ thị tri thức.....	32

CHƯƠNG 3. TÍNH ỨNG DỤNG.....	33
3.1. Ứng dụng của kỹ thuật đề xuất dựa trên ngữ nghĩa trong các lĩnh vực công nghệ thông tin [0].....	33
3.1.1. Đề xuất người theo dõi dựa trên ngữ nghĩa trên nền tảng mạng xã hội Twitter [1]	33
3.1.2. Đề xuất dựa trên ngữ nghĩa cho hệ thống tổng hợp tin tức thể thao [5]	39
3.2. Kết luận về khả năng ứng dụng.....	48
CHƯƠNG 4. KẾT LUẬN.....	49
TÀI LIỆU THAM KHẢO	51

DANH MỤC HÌNH ẢNH

- Hình 1. Ảnh chụp thống kê các bài báo nghiên cứu về vấn đề Semantics recommender system trên Google Scholar
- Hình 2. Phân tách và xác định loại từ cho câu văn cần phân tích
- Hình 3. Phân tích từ “London” trong câu
- Hình 4. Quá trình bỏ ngữ Lemmatization đã đổi động từ “is” trong câu thành động từ gốc “be”
- Hình 5. Minh họa sau khi vô hiệu các từ dừng (chữ màu xám)
- Hình 6. Đặt động từ “be”/”is” làm root cho mô hình cây
- Hình 7. Mối quan hệ giữa những từ root với những từ khác trong câu
- Hình 8. Nhận dạng thực thể được đặt tên
- Hình 9. Phân tích với các câu còn lại
- Hình 10. Một mô hình đồ thị tri thức
- Hình 11. Ảnh chụp minh họa cho một chuỗi dữ liệu dạng đồ thị tri thức
- Hình 12. Ảnh chụp minh họa một mô hình của Neo4j
- Hình 13. Một ví dụ của đề xuất người theo dõi trên Twitter
- Hình 14. Ảnh chụp minh họa kiến trúc hệ thống đề xuất người theo dõi trên Twitter
- Hình 15. Biểu đồ mạng lưới người theo dõi/người được theo dõi để tìm người dùng ứng viên
- Hình 16. Cách tính độ chính xác ở thứ hạng k
- Hình 17. Độ chính xác của cách tiếp cận dựa trên ngữ nghĩa và cách tiếp cận dựa trên từ vựng trên thí nghiệm quy trình đề xuất

Hình 18. Ảnh chụp cho công thức tính giá trị Term Frequency-Inverse Document Frequency TF-IDF

Hình 19. Công thức xác định trọng lượng chiều dài đường dẫn (Path Length Weight)

Hình 20. Trọng lượng độ dài W của path 1

Hình 21. Trọng lượng độ dài W của path 2

Hình 22. Định nghĩa về trọng số đường dẫn quan hệ của một đường dẫn P

Hình 23. Xác định độ tương đồng ngữ nghĩa dựa trên công thức kết hợp

LỜI MỞ ĐẦU

Nghiên cứu được thực hiện tại Khoa Công nghệ thông tin – Trường Đại học Sư phạm Thành phố Hồ Chí Minh, dưới sự hướng dẫn khoa học của ThS Trần Thanh Nhã.

Trước tiên chúng em xin bày tỏ lòng biết ơn sâu sắc tới thầy ThS Trần Thanh Nhã đã đưa chúng em đến với lĩnh vực nghiên cứu này. Thầy đã tận tình giảng dạy, hướng dẫn chúng em tiếp cận và đạt được những kết quả nhất định trong nghiên cứu của mình. Thầy đã luôn tận tâm động viên, khuyến khích và chỉ dẫn giúp chúng em hoàn thành nghiên cứu này.

Chúng em xin bày tỏ lòng biết ơn tới các Thầy Cô thuộc Khoa Công nghệ thông tin và cán bộ Phòng Khoa học Công nghệ, khoa Công nghệ Thông tin – Trường Đại học Sư Phạm Thành phố Hồ Chí Minh đã tạo mọi điều kiện thuận lợi giúp đỡ chúng em trong quá trình học tập và nghiên cứu.

Sự hướng dẫn của thầy ThS Trần Thanh Nhã đã tận tình hướng dẫn, động viên, cổ vũ của gia đình, bạn bè là nguồn động lực quan trọng để chúng em thực hiện đề tài nghiên cứu. Do kiến thức còn hạn chế, nên đề tài nghiên cứu của chúng em không tránh khỏi những thiếu sót, kính mong sự thông cảm, chỉ bảo của quý Thầy Cô.

Chúng em xin chân thành cảm ơn.

Thay mặt nhóm thực hiện.

Trịnh Hoàng Tùng

CHƯƠNG 1. TỔNG QUAN

1.1. Đặt vấn đề

Hệ tư vấn thông tin (recommendation system) là một bài toán thuộc về lĩnh vực khoa học máy tính – một lĩnh vực xử lý các dữ liệu và đưa ra các kết quả nhằm phục vụ cho một mục đích nào đó. Với chủ đề nghiên cứu về hệ tư vấn dựa trên ngữ nghĩa (Semantics based recommendation system), đây hứa hẹn sẽ là một kỹ thuật đề xuất thông tin với quy mô mới mẻ, thú vị, và là một cách tiếp cận với khoa học máy tính từ cơ bản đến nâng cao.

The screenshot shows the Google Scholar search results for the query "semantic recommender system". The search bar at the top shows the query and a magnifying glass icon. Below the search bar, there are filters for "Bài viết" (Articles) and "Khoảng 178.000 kết quả (0,18 giây)". On the left side, there are filters for "Mọi lúc" (All time), "Từ 2023", "Từ 2022", "Từ 2019", and "Phạm vi tùy chọn...". There are also filters for "Sắp xếp theo mức độ liên quan" (Sort by relevance) and "Sắp xếp theo ngày" (Sort by date). On the right side, there are filters for "Hồ sơ của tôi" (My profile) and "Thư viện của tôi" (My library). The search results are listed in a table with columns for "Mọi lúc", "Từ 2023", "Từ 2022", "Từ 2019", and "Phạm vi tùy chọn...". The first result is "A social-semantic recommender system for advertisements" by F. García-Sánchez, R. Colomo-Palacios, et al., published in Information Processing, 2020, Elsevier. The second result is "A framework of semantic recommender system for e-learning" by S. Fraihat, Q. Shambour, published in Journal of Software, 2015, ammanu.edu.jo. The third result is "Semantic recommender systems. analysis of the state of the topic" by E. Peis, J. M. del Castillo, J. A. Delgado-López, published in Hipertext.net, 2008, researchgate.net. The fourth result is "A semantic recommender system for adaptive learning" by P. Montuschi, F. Lamberti, V. Gatteschi, et al., published in IT, 2015, IEEE Explore. Below the search results, there is a section for "Tìm kiếm có liên quan" (Related searches) with terms like "semantic recommender system adaptive learning", "hybrid recommender systems enhanced semantic layer", "semantic recommender system educational competencies", "enhanced recommender system context recommendation system", and "personalized recommender system".

У СЕРВИСОВ, НА КУЛЬТУРЫ, НА КУЛЬТУРЫ - СЕРВИСЫ, 2010 - Wiley Online Library
... In order to improve the information representation, the **system** makes use of **Semantic Web** technologies in its development, resulting in a so-called **semantic recommender system**. The ...
☆ Lưu 99 Trích dẫn Trích dẫn 76 bài viết Bài viết có liên quan Tất cả 3 phiên bản

A multimedia **semantic recommender system** for cultural heritage applications [PDF] core.ac.uk
M Albanese, A d'Acerno, V Moscatto ... - ... on **semantic** ..., 2011 - IEEE Xplore
... **semantic** multimedia **recommender system** that computes customized recommendations using **semantic** ... We have implemented a **recommender** prototype for browsing the Uffizi Gallery ...
☆ Lưu 99 Trích dẫn Trích dẫn 64 bài viết Bài viết có liên quan Tất cả 13 phiên bản

BizSeeker: a hybrid **semantic recommendation system** for personalized government-to-business e-services [PDF] researchgate.net
J Lu, Q Shambour, Y Xu, Q Lin, G Zhang - Internet Research, 2010 - Emerald.com
... this study focuses on adopting **recommender systems** to provide ... **semantic recommendation** approach which integrates item-based collaborative filtering (CF) and item-based **semantic** ...
☆ Lưu 99 Trích dẫn Trích dẫn 111 bài viết Bài viết có liên quan Tất cả 5 phiên bản

An efficient **semantic recommender** method for arabic text [PDF] researchgate.net
B Hawashin, S Alzubi, T Kanan, A Mansour - The Electronic Library, 2019 - Emerald.com
... This paper aims to propose a new efficient **semantic recommender** method for Arabic ...
Next, a new **semantic recommender system** method for Arabic text is proposed. This method ...
☆ Lưu 99 Trích dẫn Trích dẫn 30 bài viết Bài viết có liên quan Tất cả 5 phiên bản

Hybrid **semantic recommender system** for chemical compounds [HTML] nih.gov
M Barros, A Moitinho, FM Couto - ... Conference on IR Research, ECIR 2020 ..., 2020 - Springer
... The lack of **Recommender Systems** in this particular field presents a challenge for the development of new recommendations models. In this work, we propose a Hybrid **recommender** ...
☆ Lưu 99 Trích dẫn Trích dẫn 8 bài viết Bài viết có liên quan Tất cả 7 phiên bản



Trợ giúp Quyền riêng tư Điều khoản

Google Scholar

semantic recommender system

Bài viết

Trang 2 trong khoảng 178.000 kết quả (0,03 giây)

Hồ sơ của tôi Thư viện của tôi

Mọi lúc
Tứ 2023
Tứ 2022
Tứ 2019
Phạm vi tùy chọn...

Sắp xếp theo mức độ liên quan
Sắp xếp theo ngày

Mọi loại
Bài viết đánh giá

☐ bao gồm bằng sáng chế
☒ bao gồm trích dẫn

☒ Tạo thông báo

A Lesson learned from PMF based approach for **Semantic Recommender System** [PDF] researchgate.net
N Kushwaha, X Sun, B Singh, OP Vyas - ... of Intelligent Information Systems, 2018 - Springer
... **Recommender System**. While most of the existing works incorporate **semantic** web information into **recommendation system** by ... a collaborative filtering based **semantic** dual probabilistic ...
☆ Lưu 99 Trích dẫn Trích dẫn 14 bài viết Bài viết có liên quan Tất cả 8 phiên bản

[TRÍCH DẪN] Trust Based **Recommender System** for **Semantic Web**. [PDF] psu.edu
P Bedi, H Kaur, S Marwaha - IJCAI, 2007
☆ Lưu 99 Trích dẫn Trích dẫn 226 bài viết Bài viết có liên quan Tất cả 10 phiên bản

A **semantic recommender system** based on frequent tag pattern
H Movahedian, MR Khayyambashi - Intelligent Data Analysis, 2015 - content.iospress.com
Social tagging provides an effective way for users to organize, manage, share and search for various kinds of resources. These tagging **systems** have resulted in more and more users ...
☆ Lưu 99 Trích dẫn Trích dẫn 9 bài viết Bài viết có liên quan Tất cả 2 phiên bản

SMARTMUSEUM: A mobile **recommender system** for the Web of Data [PDF] researchgate.net
T Ruotsalo, K Haav, A Stoyanov, S Roche, E Fani - ... of Web Semantics, 2013 - Elsevier
... In summary, mobile **recommender systems** can benefit from **Semantic Web** technologies as a data representation format. **Semantic** reasoning and query expansion are necessary if one ...
☆ Lưu 99 Trích dẫn Trích dẫn 172 bài viết Bài viết có liên quan Tất cả 11 phiên bản

Semantics-aware content-based **recommender systems** [PDF] researchgate.net
M De Gemmis, P Lops, C Musto, F Narducci - ... **Recommender systems** ..., 2015 - Springer
... (NLP) and **Semantic** Technologies, which is one of the most innovative lines of research in **semantic recommender systems** [61]. We roughly classify **semantic** techniques into top-down ...
☆ Lưu 99 Trích dẫn Trích dẫn 292 bài viết Bài viết có liên quan Tất cả 4 phiên bản

A collaborative framework for sensing abnormal heart rate based on a **recommender system: Semantic recommender system for healthcare**

☆ Lưu 99 Trích dẫn Trích dẫn 292 bài viết Bài viết có liên quan Tất cả 4 phiên bản

A collaborative framework for sensing abnormal heart rate based on a **recommender system: Semantic recommender system** for healthcare

G. Guzmán, M. Torres-Ruiz, V. Tambonero... - Journal of Medical and ..., 2018 - Springer

... In this section, the obtained results in the **recommender system** stage are shown. For discussion purposes, two ... The final result of the **recommender system** stage is shown in Fig. 12. ...

☆ Lưu 99 Trích dẫn Trích dẫn 12 bài viết Bài viết có liên quan Tất cả 4 phiên bản

Intelligent services: A **semantic recommender system** for knowledge representation in industry

[PDF] academia.edu

M. Mehrpoor, A. Gjaerde... - ..., and Innovation (ICE), 2014 - IEEE Xplore

... By compiling discussed theoretical basics for the target **semantic recommender system**, this research aims to implement a prototype to provide a real working **system** to enhance the ...

☆ Lưu 99 Trích dẫn Trích dẫn 11 bài viết Bài viết có liên quan Tất cả 3 phiên bản

Semantic recommender system for touristic context based on linked data

[PDF] researchgate.net

L. Cabrera Rivera, L. M. Viches-Bláquez... - ..., Information Systems ..., 2015 - Springer

... In this work, a **recommender system** that exploits **semantic** information based on ... **semantic** touristic route is described. In Fig. 1, we present the core of our **semantic recommender system**...

☆ Lưu 99 Trích dẫn Trích dẫn 10 bài viết Bài viết có liên quan Tất cả 6 phiên bản

Semantic web recommender systems

[PDF] uni-freiburg.de

C. N. Ziegler - Current Trends in Database Technology-EDBT 2004 ..., 2005 - Springer

... **recommender systems** has primarily addressed centralized scenarios and largely ignored open, decentralized **systems** ... integration of **recommender system** facilities for **Semantic Web** ...

☆ Lưu 99 Trích dẫn Trích dẫn 108 bài viết Bài viết có liên quan Tất cả 22 phiên bản

Semantic-enhanced personalized recommender system

[PDF] uvigo.es

R. Q. Wang, F. S. Kong - 2007 International Conference on ..., 2007 - IEEE Xplore

... **systems**, has such well-known limitations as sparsity, scalability and cold-start problem. ...

semantic-enhanced collaborative recommender system is proposed in this paper. The **semantic** ...

☆ Lưu 99 Trích dẫn Trích dẫn 75 bài viết Bài viết có liên quan Tất cả 3 phiên bản



Hình 1. Ảnh chụp thống kê các bài báo nghiên cứu về vấn đề Semantics recommender system trên Google Scholar

1.2. Mục tiêu cụ thể

Ở bài nghiên cứu này, nhóm tập trung vào phân tích, nghiên cứu chủ đề Semantics Recommendation System và chỉ ra một số ứng dụng của kỹ thuật đề xuất này trong thực tế.

Mục tiêu cụ thể của bài báo cáo là chỉ rõ khái niệm, phân tích, ưu điểm, nhược điểm và ứng dụng của kỹ thuật tư vấn dựa trên ngữ nghĩa; phân tích tính ứng dụng của cơ sở dữ liệu đồ thị tri thức để hiển thị kết quả đề xuất của kỹ thuật trên.

Xây dựng thành công một bài báo cáo mang tính hiệu quả trong truyền đạt kiến thức, giá trị nỗ lực tìm tòi, sáng tạo trong tư duy và làm việc nhóm.

1.3. Phạm vi đề tài

Phạm vi đề tài nghiên cứu của tập thể nhóm Neko được đặt ra rõ ràng: Phân tích về cơ sở lý thuyết và áp dụng của kỹ thuật Semantics based Recommendation System, cùng với đồ thị tri thức Knowledge Graph để biểu diễn dữ liệu.

Bài nghiên cứu gồm ba phần chính: nghiên cứu về kỹ thuật Semantics based Recommendation System, nghiên cứu về cơ sở dữ liệu đồ thị và nghiên cứu tính ứng dụng của kỹ thuật Semantics based Recommendation System.

CHƯƠNG 2. CƠ SỞ LÝ THUYẾT

2.1. Hệ tư vấn dựa trên ngữ nghĩa - Semantics Recommendation System

2.1.1. Semantics Recommendation System là gì?

Hệ tư vấn dựa trên ngữ nghĩa (Semantic-based Recommendation System) là một hệ thống máy tính có khả năng tự động tư vấn và đề xuất các sản phẩm, dịch vụ hoặc thông tin liên quan đến nhu cầu của người dùng thông qua sự hiểu biết và sử dụng ngôn ngữ tự nhiên. Đây là một ứng dụng quan trọng của Trí tuệ nhân tạo (AI) và được áp dụng rộng rãi trong các lĩnh vực như thương mại điện tử, du lịch, giáo dục và y tế.

Kỹ thuật đề xuất này là một phương pháp để giới thiệu sản phẩm cho khách hàng dựa trên ý nghĩa hay ngữ nghĩa của sản phẩm thay vì phải phụ thuộc chỉ dựa trên các thông tin tiêu chuẩn như thông tin khách hàng, lịch sử mua hàng hoặc đánh giá sản phẩm.

Phương pháp này sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên và khai thác các mối quan hệ ngữ nghĩa giữa các sản phẩm để đưa ra các sản phẩm tương đồng với sản phẩm mà khách hàng đang quan tâm.

Hệ tư vấn dựa trên ngữ nghĩa hoạt động bằng cách thu thập thông tin từ người dùng, phân tích dữ liệu và xác định ý định của người dùng thông qua sự tương tác với hệ thống. Sau đó, hệ thống sẽ sử dụng các thuật toán máy học nhằm xử lý ngữ nghĩa từ chính thông tin đã thu thập được để đưa ra các đề xuất phù hợp nhất với nhu cầu của người dùng. Hoặc chủ động hơn, kỹ thuật đề xuất này có thể tự phân tích và đưa ra kết quả đề xuất thông qua những ngữ nghĩa trên chính đặc tính của sản phẩm (items), mà không cần tiếp nhận thông tin từ hồ sơ của người dùng. Nhờ đó mà kỹ thuật có thể khởi tạo nên những đề xuất mang tính mới lạ, thu hút người dùng.

2.1.2. Cơ bản về hệ đề xuất thông tin dựa trên ngữ nghĩa

Các yếu tố quan trọng trong hệ đề xuất thông tin dựa trên ngữ nghĩa bao gồm:

- Phân tích ngữ nghĩa: Hệ thống phân tích và hiểu ngữ nghĩa của nhu cầu hoặc câu truy vấn từ người dùng. Nó xác định ý định và mục tiêu của người dùng và tìm hiểu ngữ cảnh để đưa ra đề xuất thông tin phù hợp.

- Trích xuất tri thức: Hệ thống trích xuất tri thức từ nguồn dữ liệu khác nhau như cơ sở dữ liệu, tài liệu hoặc nguồn dữ liệu trực tuyến. Điều này có thể bao gồm trích xuất thông tin cụ thể, quan hệ giữa các khái niệm và bối cảnh liên quan.

- Xây dựng mô hình ngữ nghĩa: Hệ thống xây dựng mô hình ngữ nghĩa để biểu diễn tri thức và thông tin từ nguồn dữ liệu. Điều này giúp hệ thống hiểu được mối quan hệ và ý nghĩa của các đối tượng và thông tin trong tri thức.

- Đo lường độ tương tự ngữ nghĩa: Một thách thức quan trọng khác là đo lường độ tương tự ngữ nghĩa giữa các đối tượng trong hệ thống đề xuất. Cần có các phương pháp đo lường độ tương tự hiệu quả để có thể tìm ra các mối quan hệ ngữ nghĩa giữa các đối tượng và đưa ra đề xuất phù hợp.

- Định nghĩa và biểu diễn ngữ nghĩa: Một thách thức lớn trong kỹ thuật đề xuất thông tin dựa trên ngữ nghĩa là việc định nghĩa, biểu diễn ngữ nghĩa của các đối tượng, thuộc tính và quan hệ trong hệ thống đề xuất. Điều này đòi hỏi sự hiểu biết về khái niệm và cách thức biểu diễn ngữ nghĩa để có thể áp dụng vào quá trình đề xuất

- Đề xuất thông tin: Dựa trên việc phân tích ngữ nghĩa và tri thức, hệ thống đề xuất thông tin phù hợp cho người dùng. Điều này có thể bao gồm đề xuất câu trả lời, đề xuất sản phẩm hoặc dịch vụ, hoặc đề xuất các tài liệu hay nguồn thông tin liên quan.

Hệ đề xuất thông tin dựa trên ngữ nghĩa giúp cải thiện trải nghiệm người dùng bằng cách cung cấp thông tin chính xác, phù hợp và có ý nghĩa dựa trên ngữ cảnh và mục tiêu của người dùng. Nó tận dụng tri thức và ngữ nghĩa để đưa ra các đề xuất thông tin một cách thông minh và hữu ích tới cho người dùng.

2.1.3. Nguyên lý hoạt động

Bản chất Semantic-based Recommendation System hoạt động dựa trên việc khai thác thông tin ngữ nghĩa của dữ liệu để tạo ra các gợi ý sản phẩm phù hợp với người dùng.

Mô hình học máy được sử dụng để xác định sự tương đồng giữa các sản phẩm dựa trên các thuộc tính ngữ nghĩa, ví dụ như đặc tính chung, tên gọi, hoặc miêu tả.

Để thực hiện điều này, các hệ thống sử dụng kỹ thuật đề xuất dựa trên ngữ nghĩa (*semantics-based recommendation system*) thường sử dụng các **kỹ thuật xử lý ngôn ngữ tự nhiên (NLP)** để hiểu văn bản và dữ liệu ngữ nghĩa khác. Cụ thể ta có thể kể đến và phân tích bản chất về kỹ thuật xử lý ngôn ngữ tự nhiên phổ biến nhất đang được ứng dụng trong kỹ thuật đề xuất dựa trên ngữ nghĩa – kỹ thuật *Rút trích thông tin* thông qua một ví dụ điển hình như sau:¹

Ta có đoạn văn cần phân tích:

“London is the capital and most populous city of England and the United Kingdom. Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia. It was founded by the Romans, who named it Londinium.”

Tạm dịch (theo kiểu thông thường – sử dụng trí tuệ con người): *“London là thủ đô và là thành phố đông dân nhất của England và United Kingdom. Đứng trên dòng sông Thamse ở phía đông của đảo Great Britain, London là một khu định cư lớn trong hai thiên niên kỷ. Nó được thành lập bởi người La Mã, những người đã đặt tên cho nó là Londinium.”*

- *Bước 1: Phân đoạn câu văn - Sentence Segmentation*

Ta phân tách đoạn văn ban đầu thành các câu văn riêng biệt, như sau:

1/ *“London is the capital and most populous city of England and the United Kingdom”*

2/ *“Standing on the River Thames in the south east of the island of Great Britain, London has been a major settlement for two millennia.”*

¹ Tham khảo từ <https://viblo.asia/p/xu-ly-ngon-ngu-tu-nhien-phan-1-OeVKB8eQlkW>

3/ *“It was founded by the Romans, who named it Londinium.”*

Sau khi phân tách, chúng ta có thể cho rằng, mỗi câu trong tiếng Anh mang một ý nghĩa đặc biệt và riêng lẻ. Và sẽ dễ dàng hơn để viết một chương trình có thể hiểu một câu đơn hơn là hiểu liền một lúc cả một đoạn văn.

- *Bước 2: Mã hóa các từ - Word Tokenization*

Ta đã chia tài liệu thành các câu văn riêng lẻ, và do đó, cần phải xử lý từng câu văn một. Hãy bắt đầu với câu đầu tiên trong tài liệu:

“London is the capital and most populous city of England and the United Kingdom.”

Bước tiếp theo đó là chia câu văn này thành các từ riêng lẻ, thành các thành phần nhỏ hơn được gọi là các từ (words) hoặc các tokens. Điều này được gọi là các tokenization. Và đây là các kết quả:

“London”, “is”, “the”, “capital”, “and”, “most”, “populous”, “city”, “of”, “England”, “and”, “the”, “United”, “Kingdom”, “.”

Tokenization rất dễ được xác định, đặc biệt là với ngôn ngữ Anh: ta sẽ tách các từ bất cứ khi nào có khoảng cách giữa chúng. Và chúng ta sẽ coi dấu chấm câu là các Token riêng biệt vì dấu chấm câu cũng có mang ý nghĩa riêng của chúng.

- *Bước 3: Dự đoán các thành phần cho mỗi token - Predicting Parts of Speech for Each Token*

Tiếp đến ta sẽ xem xét từng token (tức là từng từ của một câu văn) và cố gắng dự đoán loại từ của token này. Có thể nó là danh từ, động từ, hoặc tính từ,... Biết được vai trò của từng từ/token trong câu, việc đó sẽ giúp ta có thể bắt đầu tìm ra được câu văn đang nói về cái gì.

Ngoài ra, ta có thể làm điều này bằng cách cung cấp từng từ (và một số từ xung quanh nó, để cung cấp ngữ cảnh nhằm dễ hình dung vấn đề) vào một mô hình phân loại một phần của toàn đoạn văn để thực hiện dự đoán từ loại của từ được truyền vào (việc dự đoán một từ thuộc dạng từ nào được gọi là dự đoán một phần của cả đoạn). Sau khi xử lý được toàn bộ câu, chúng ta có thể có kết quả như thế này:

“*London*”: danh từ riêng/tên riêng

“*is*”: động từ

“*the*”: mạo từ

“*capital*”: danh từ

“*and*”: mạo từ

“*most*”: tính từ

“*populous*”: tính từ

“*city*”: danh từ

“*of*”: mạo từ

“*England*”: danh từ riêng/tên riêng

“*and*”: mạo từ

“*the*”: mạo từ

“*United*”: danh từ riêng/tên riêng

“*Kingdom*”: danh từ riêng/tên riêng

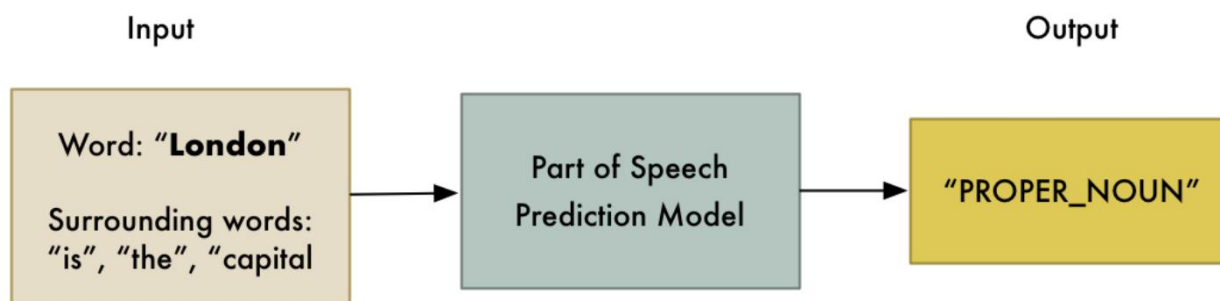
“.”: dấu câu



Hình 2. Phân tách và xác định loại từ cho câu văn cần phân tích

Mô hình “*part of speech for each token*” ban đầu được đào tạo bằng cách cung cấp cho nó hàng triệu câu có sẵn trong từ điển với mỗi từ đã được gắn thẻ và nó có thể tái tạo lại các hành vi đó.

Tuy nhiên, vì mô hình này hoàn toàn dựa trên số liệu thống kê nên nó không thực sự hiểu những từ này có nghĩa giống như cách con người hình dung bằng bộ não của mình. Nó chỉ biết làm thế nào để đoán một phần (tức một từ) của đoạn văn cần phân tích dựa trên các câu và các từ tương tự mà nó đã được cung cấp/đã được biết trước đó.



Hình 3. Phân tích từ “London” trong câu

Như hình 3 ở trên, từ ngữ “London” được phân tách từ câu văn ban đầu, qua các bước xác định và tiên đoán xử lý, từ này được xác định là một “proper noun” – tức danh từ địa phương/tên riêng, cụ thể hơn thì đây là tên của một địa điểm, một thành phố có thực trên thế giới.

Với thông tin đã được xác định ở trên, chúng ta bước đầu lượm nhặt một số ý nghĩa rất cơ bản, rằng các danh từ trong câu bao gồm “London” và “capital”, vì vậy có lẽ câu này có lẽ đang nói về London – một thủ đô của một đất nước nào đó.

- *Bước 4: Bỏ ngữ cho văn bản - Text Lemmatization*

Việc bỏ ngữ (Lemmatization) tức là đưa các từ về định dạng gốc ban đầu, và có thể có một số quy tắc để xử lý các từ mà ta hiếm khi được nhìn thấy trước đây. Ta có ví dụ:

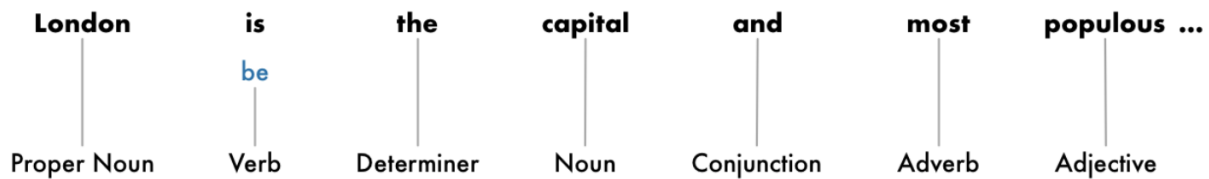
“I had two dogs”

Sau khi thực hiện bỏ ngữ (Lemmatization), ta có câu văn gốc như sau:

“I [have] two [dog]”

Trong phân tích xử lý ngôn ngữ tự nhiên, việc bỏ ngữ này rất hữu ích vì giúp hệ thống biết được dạng cơ bản của mỗi từ để chốt rằng cả hai câu “*I had two dogs*” và “*I have two dog*” đều nói về cùng một khái niệm, cùng một vấn đề.

Đây là những gì mà câu văn được phân tích sẽ trở thành sau khi thực hiện quá trình bỏ ngữ:

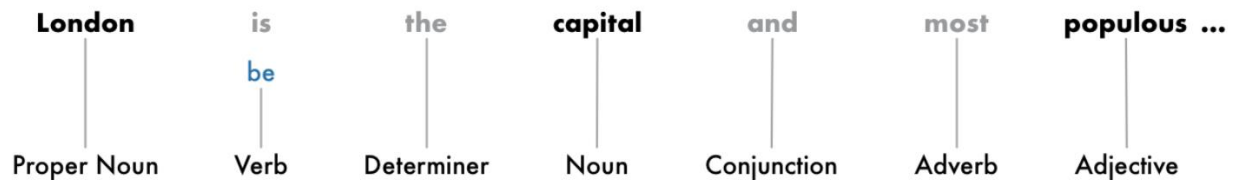


Hình 4. Quá trình bỏ ngữ Lemmatization đã đổi động từ “is” trong câu thành động từ gốc “be”

- *Bước 5: Xác định các từ dừng - Identifying Stop Words*

Các từ dừng (stop words) là những từ không có quá nhiều ý nghĩa trong việc phân biệt ý nghĩa cho nội dung câu. Vì thế mà ta cần phải xác định các từ này nhằm tránh gây nhiễu thông tin. Cụ thể đối với ngôn ngữ Anh, thường xuyên xuất hiện các loại từ nối, mạo từ như “and”, “or”, “the”, “a”, ... Đây chính là những từ dừng và chúng cần được loại bỏ khỏi thành phần phân tích

Ở đây, nguyên câu văn được phân tích trông như thế nào khi các từ dừng đã được vô hiệu hóa (chuyển sang màu xám):

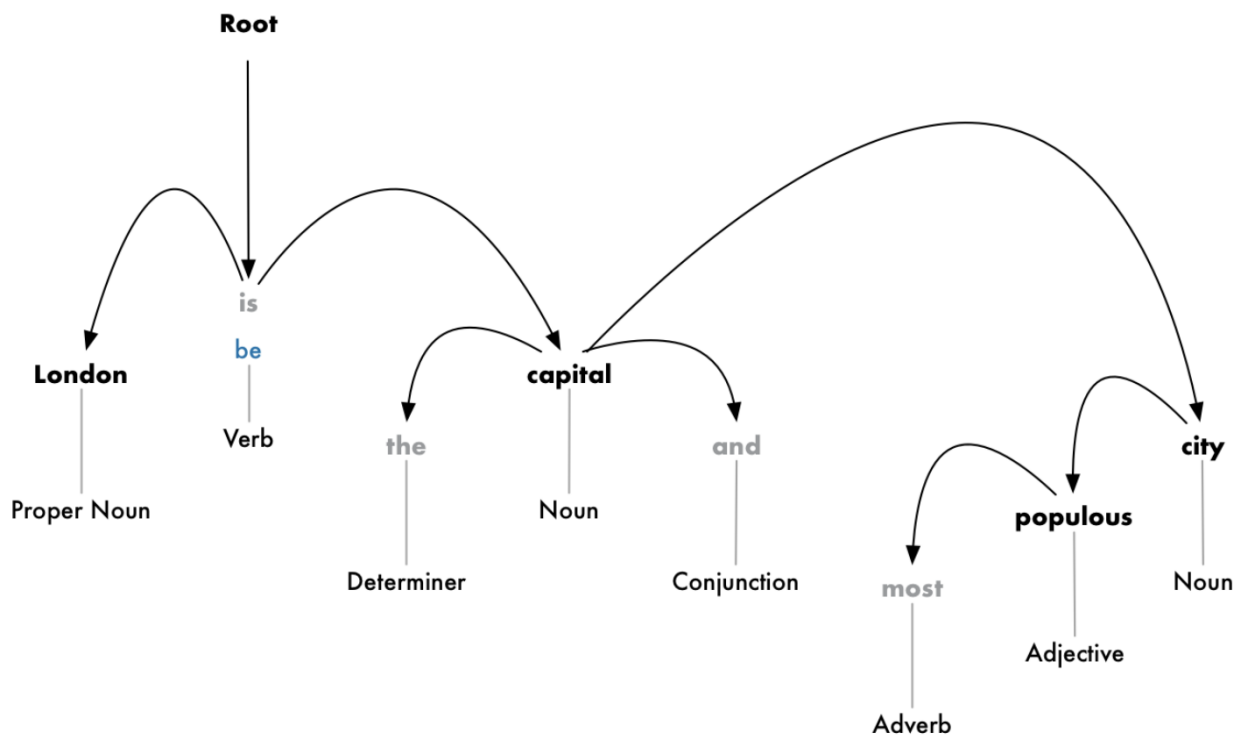


Hình 5. Minh họa sau khi vô hiệu các từ dừng (chữ màu xám)

- *Bước 6: Phân tích sự phụ thuộc về cú pháp - Dependency Parsing*

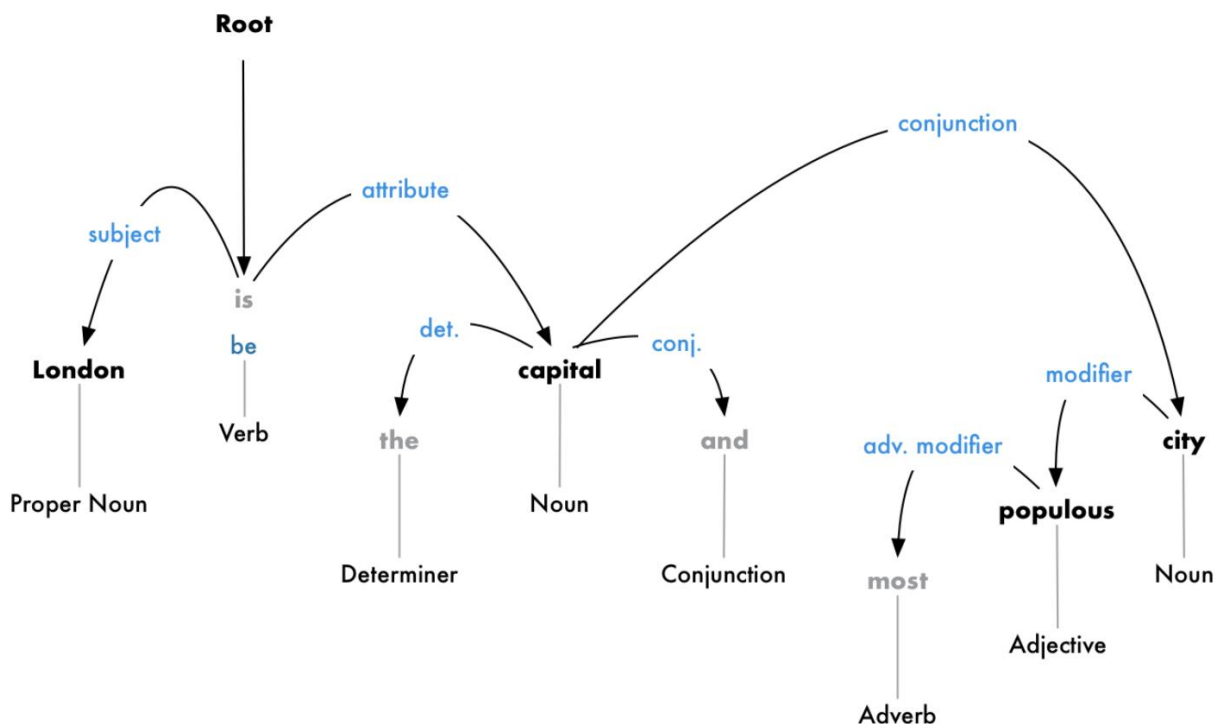
Bước tiếp theo đó là tìm hiểu xem làm thế nào tất cả các từ trong câu có thể liên quan đến nhau. Hay nói cách khác là tìm hiểu xem các từ trong câu được phân tích - chúng liên quan đến nhau như thế nào. Điều này được gọi là quá trình phân tích phụ thuộc về cú pháp.

Mục tiêu là xây dựng một mô hình cây có thể gán một từ đơn duy nhất làm root cho mỗi từ trong câu. Từ root của cây này sẽ là động từ “*be*” (“*is*”) trong câu. Đây là phần đầu của cây phân tích sẽ trông như thế nào cho câu của chúng ta:



Hình 6. Đặt động từ “be”/”is” làm root cho mô hình cây

Nhưng không dừng lại ở đó, ta vẫn có thể thực hiện thêm một bước nữa. Tức ngoài việc xác định từ root, chúng ta có thể dự đoán được loại mối liên hệ, mối liên quan tồn tại giữa những từ trong câu với từ root đó.



Hình 7. Mối quan hệ giữa những từ root với những từ khác trong câu

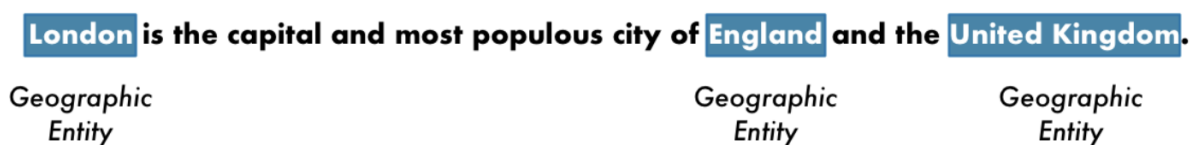
Cây phân tích cú pháp này cho chúng ta thấy chủ đề của câu là danh từ "London" và nó có có quan hệ "be" với "capital". Cuối cùng, chúng ta cũng biết một điều hữu ích rằng London là một thủ đô! Và nếu chúng ta đi theo cây phân tích hoàn chỉnh cho câu (ngoài những gì đã được hiển thị), chúng ta thậm chí còn có thể phát hiện ra rằng London là thủ đô của United Kingdom.

Điều quan trọng cần nhắc lại rằng, nhiều câu trong tiếng Anh là mơ hồ và thực sự khó phân tích. Trong những trường hợp đó, mô hình sẽ đưa ra dự đoán dựa trên phiên bản phân tích cú pháp của câu đó, và có lẽ một số trường hợp sẽ không hoàn hảo và đôi khi mô hình sẽ dự đoán sai. Nhưng theo thời gian, mô hình phân tích xử lý ngôn ngữ tự nhiên của chúng ta sẽ tiếp tục trở nên tốt hơn trong việc phân tích văn bản một cách hợp lý.

- *Bước 7: Nhận dạng thực thể được đặt tên – Named Entity Recognition*

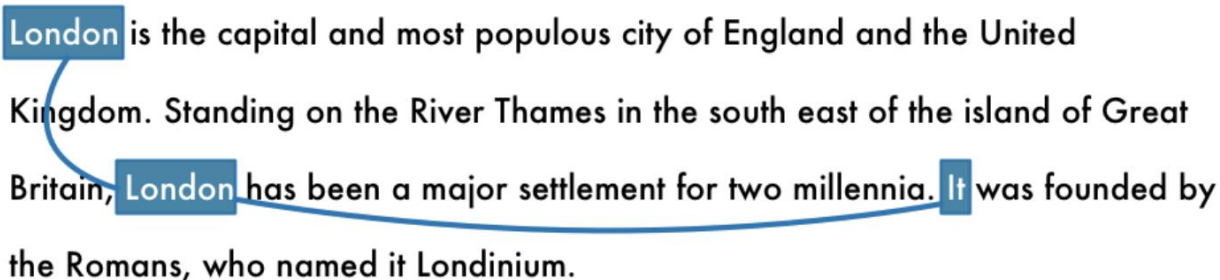
Sau khi đã hoàn thành các bước được coi là khó khăn nhất và cốt lõi nhất của bài toán phân tích, thì điều quan trọng là ta có thể vượt ra ngoài trường ngữ pháp thô và bắt đầu thực sự rút ra ý tưởng và hàn gắn ý nghĩa thực sự của câu văn phân tích.

Một số danh từ này trình bày những thứ có thật trên thế giới. Ví dụ: "*London*", "*England*" hay "*United Kingdom*" đều đại diện cho các địa điểm thực tế trên bản đồ thế giới. Với thông tin đó, ta có thể tự động trích xuất danh sách các địa điểm trong thế giới được đề cập trong tài liệu.



Hình 8. Nhận dạng thực thể được đặt tên

Và cuối cùng, áp dụng tương tự kỹ thuật với các câu còn lại của đoạn văn cần phân tích, hệ thống sẽ thu được ý nghĩa của đoạn:



Hình 9. Phân tích với các câu còn lại

Ngoài ra, còn một vấn đề nằm trong đoạn văn này là các từ ngữ tránh lặp từ như "*It*". Là một người đọc câu này, bạn có thể dễ dàng hiểu rằng "*it*" đại diện cho "*London*", "*It*" (nó) ở đây là London. Mục tiêu của giải pháp là tìm ra phép ánh xạ tương tự này bằng cách theo dõi các đại từ qua các câu nhằm diễn giải cho những từ ngữ tránh lặp từ

như này. Ta/hệ thống đang muốn tìm ra tất cả các từ “*it*” đang đề cập đến cùng một thực thể nào.

Ở đây, kết quả của việc sử dụng “*it*” đều ám chỉ đến một thực thể đầu tiên của đoạn văn – đó là đại diện cho từ “*London*”.

Bằng cách hiểu ngữ nghĩa, hệ thống có khả năng đưa ra đề xuất chính xác hơn và phù hợp hơn với ý nghĩa thực sự của người dùng. Điều này có thể cải thiện trải nghiệm người dùng, tăng cường khả năng tìm kiếm và khuyến nghị, và mang lại lợi ích kinh doanh cho các tổ chức.

Các bước cơ bản trong Semantic-based Recommendation System bao gồm:

- Phân tích ngữ nghĩa của dữ liệu: Hệ thống sử dụng các phương pháp khai phá dữ liệu để phân tích và hiểu nội dung của dữ liệu được cung cấp, bao gồm cả các ý nghĩa đồng nghĩa và liên quan giữa các thuật ngữ khác nhau.
- Tiền xử lý dữ liệu: Dữ liệu được thu thập và tiền xử lý để chuẩn hóa định dạng và loại bỏ dữ liệu không cần thiết hoặc sai sót.
- Xác định mục tiêu và sở thích của người dùng: Hệ thống thu thập thông tin về người dùng, bao gồm lịch sử tìm kiếm, đánh giá và các mục tiêu quan tâm, từ đó đưa ra các gợi ý phù hợp với sở thích của họ.
- So sánh và lọc dữ liệu: Hệ thống so sánh các thuật ngữ, ý nghĩa và sở thích của người dùng với dữ liệu được cung cấp để lọc và đưa ra những gợi ý phù hợp nhất.
- Đánh giá hiệu suất: Để đảm bảo độ chính xác của mô hình, các thước đo hiệu suất như độ chính xác, độ phủ, và độ lặp lại được sử dụng để đánh giá mô hình.
- Đưa ra gợi ý: Khi mô hình đã được xây dựng và đánh giá hiệu suất, các gợi ý sản phẩm được tạo ra bằng cách tìm kiếm các sản phẩm tương đồng với sản phẩm mà người dùng đang xem hoặc đã mua trước đó. Các sản phẩm tương đồng này sẽ được sắp xếp theo thứ tự giảm dần của độ tương đồng với sản phẩm người dùng đang xem.

2.1.4. Ưu điểm của kỹ thuật tư vấn dựa trên ngữ nghĩa

Kỹ thuật tư vấn dựa trên ngữ nghĩa có một số ưu điểm cải thiện hơn so với các kỹ thuật khuyến nghị truyền thống khác dựa trên hồ sơ lịch sử người dùng, các ưu điểm có thể kể bao gồm:

- Cải thiện độ chính xác: Các hệ thống đề xuất dựa trên ngữ nghĩa sử dụng phân tích ngữ nghĩa để xác định mối quan hệ và điểm tương đồng giữa các mục hoặc khái niệm khác nhau, từ đó có thể đưa ra các đề xuất phù hợp và chính xác hơn..
- Trải nghiệm người dùng tốt hơn: Bằng cách cung cấp các đề xuất chính xác và phù hợp hơn, các hệ thống đề xuất dựa trên ngữ nghĩa có thể cải thiện trải nghiệm người dùng tổng thể và tăng sự hài lòng của người dùng.
- Tăng tính đa dạng: Các hệ thống đề xuất truyền thống có xu hướng đề xuất các mục tương tự với các mục mà người dùng đã tương tác. Các hệ thống đề xuất dựa trên ngữ nghĩa có thể giúp xác định các mục không chỉ giống nhau mà còn bổ sung hoặc có liên quan với nhau, dẫn đến các đề xuất đa dạng hơn.
- Cá nhân hóa nâng cao: Các hệ thống đề xuất dựa trên ngữ nghĩa có thể phân tích hành vi và sở thích của người dùng để cung cấp các đề xuất được cá nhân hóa hơn phù hợp với từng người dùng.
- Khả năng mở rộng tốt hơn: Các hệ thống đề xuất dựa trên ngữ nghĩa có thể phân tích lượng lớn dữ liệu và xác định các mẫu và mối quan hệ có thể không rõ ràng với các hệ thống đề xuất truyền thống. Điều này có thể dẫn đến các hệ thống đề xuất hiệu quả và có thể mở rộng hơn, có thể xử lý các tập dữ liệu lớn hơn và hành vi người dùng phức tạp hơn.
- Đáp ứng theo sở thích người dùng: Hệ thống dựa trên ngữ nghĩa có thể thích nghi và học từ phản hồi của người dùng, cho phép đưa ra những đề xuất cá nhân hóa phù hợp theo thời gian. Khi hệ thống thu thập thêm dữ liệu và tinh chỉnh hiểu biết ngữ nghĩa của mình, nó có thể cung cấp những đề xuất ngày càng chính xác phù hợp với sở thích thay đổi của người dùng.

2.1.5. Nhược điểm của kỹ thuật tư vấn dựa trên ngữ nghĩa

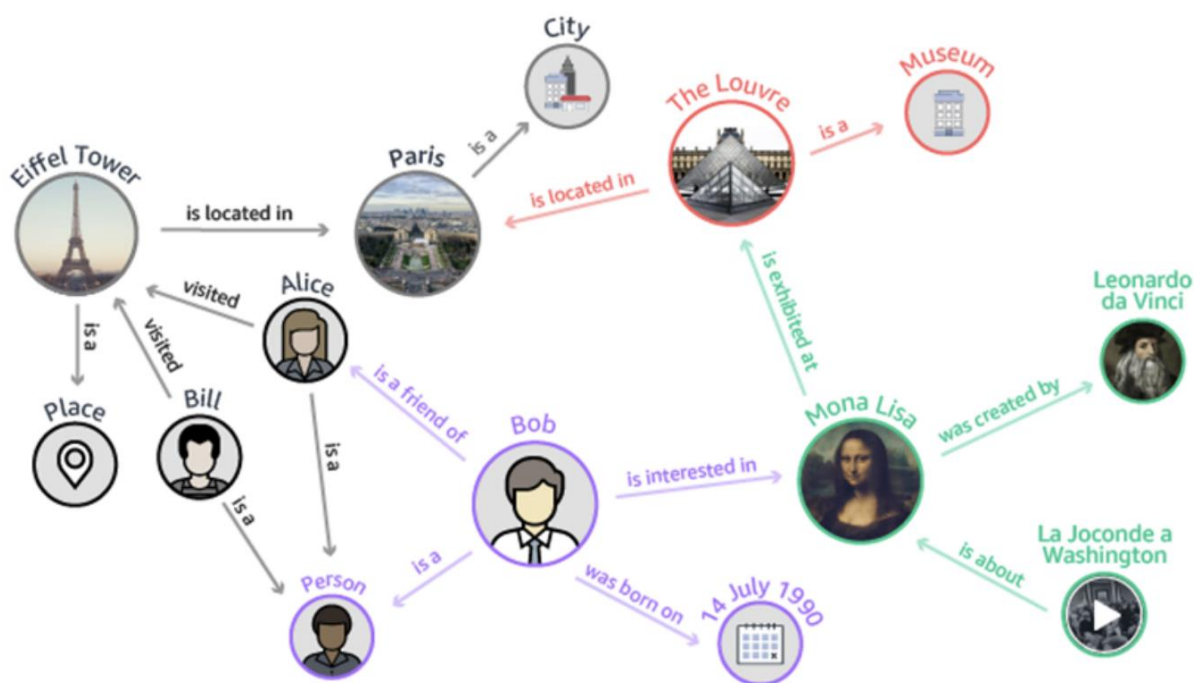
Mặc dù các hệ thống khuyến nghị sử dụng kỹ thuật đề xuất dựa trên ngữ nghĩa có nhiều ưu điểm vượt trội kể trên, nhưng chúng cũng có một số nhược điểm tiềm ẩn, bao gồm:

- Chi phí tính toán cao: Các hệ thống đề xuất dựa trên ngữ nghĩa dựa trên các thuật toán phức tạp đòi hỏi tài nguyên tính toán đáng kể, đây có thể là một yếu tố hạn chế đối với các tổ chức nhỏ hơn hoặc những tổ chức có nguồn lực hạn chế.
- Phụ thuộc vào dữ liệu chính xác: Các hệ thống đề xuất dựa trên ngữ nghĩa chủ yếu dựa vào dữ liệu chính xác, đây có thể là một thách thức nếu dữ liệu manh mún cho một vấn đề ban đầu không đầy đủ, không nhất quán hoặc có chất lượng thấp.
- Khó diễn giải kết quả: Do các hệ thống đề xuất dựa trên ngữ nghĩa sử dụng các thuật toán phức tạp nên kết quả mà chúng tạo ra có thể khó diễn giải hoặc giải thích, điều này có thể khiến các tổ chức khó hiểu cách hệ thống đưa ra đề xuất. Đây là lý do mà đôi khi xuất hiện những kết quả đề xuất khó hiểu được xử lý từ kỹ thuật này.
- Phạm vi hạn chế: Các hệ thống đề xuất dựa trên ngữ nghĩa thường được thiết kế để hoạt động với các loại dữ liệu cụ thể hoặc trong các miền cụ thể, điều này có thể hạn chế khả năng ứng dụng của chúng trong các ngữ cảnh nhất định.
- Mối quan tâm về quyền riêng tư: Các hệ thống đề xuất dựa trên ngữ nghĩa dựa trên lượng lớn dữ liệu người dùng, điều này có thể gây lo ngại về quyền riêng tư nếu dữ liệu không được xử lý phù hợp. Các tổ chức phải cẩn thận để đảm bảo rằng dữ liệu người dùng được bảo vệ và sử dụng một cách có đạo đức.
- Khó khăn trong việc mở rộng và đa ngôn ngữ: Mở rộng hệ thống để hỗ trợ nhiều lĩnh vực và đa ngôn ngữ có thể gặp khó khăn. Mỗi ngôn ngữ có những đặc điểm và thực tế ngữ nghĩa riêng, do đó việc áp dụng hệ thống cho nhiều ngôn ngữ đòi hỏi sự nghiên cứu và điều chỉnh kỹ lưỡng, phù hợp với ngữ cảnh, ngữ pháp của từng loại ngôn ngữ.

2.2. Đồ thị tri thức – Knowledge Graph

2.2.1. Đồ thị tri thức là gì

Đồ thị tri thức (knowledge graph) là một cấu trúc dữ liệu biểu diễn tri thức và các mối quan hệ giữa các khái niệm khác nhau dưới dạng đồ thị (graph). Nó bao gồm một tập hợp các nút (nodes) biểu diễn các thực thể (entities) như người, địa điểm, sản phẩm, sự kiện và các thuộc tính (attributes) của chúng, cũng như các mối quan hệ (relationships) giữa các thực thể. Điều này giúp người dùng truy cập dữ liệu một cách nhanh chóng và dễ dàng hơn, vì các thông tin được tổ chức theo cấu trúc rõ ràng và có liên kết chặt chẽ với nhau.



Hình 10. Một mô hình đồ thị tri thức

Đồ thị tri thức là công cụ quan trọng trong lĩnh vực trí tuệ nhân tạo (AI) và học máy (machine learning), được sử dụng để cải thiện các ứng dụng như tìm kiếm web, trả lời câu hỏi tự động, phân tích ngôn ngữ tự nhiên và hỗ trợ ra quyết định. Như trong chủ đề nghiên

cứu này, nhóm sẽ tập trung nghiên cứu về tính ứng dụng của đồ thị tri thức đối với kỹ thuật đề xuất dựa trên ngữ nghĩa.

2.2.2. Cơ sở dữ liệu dạng đồ thị tri thức

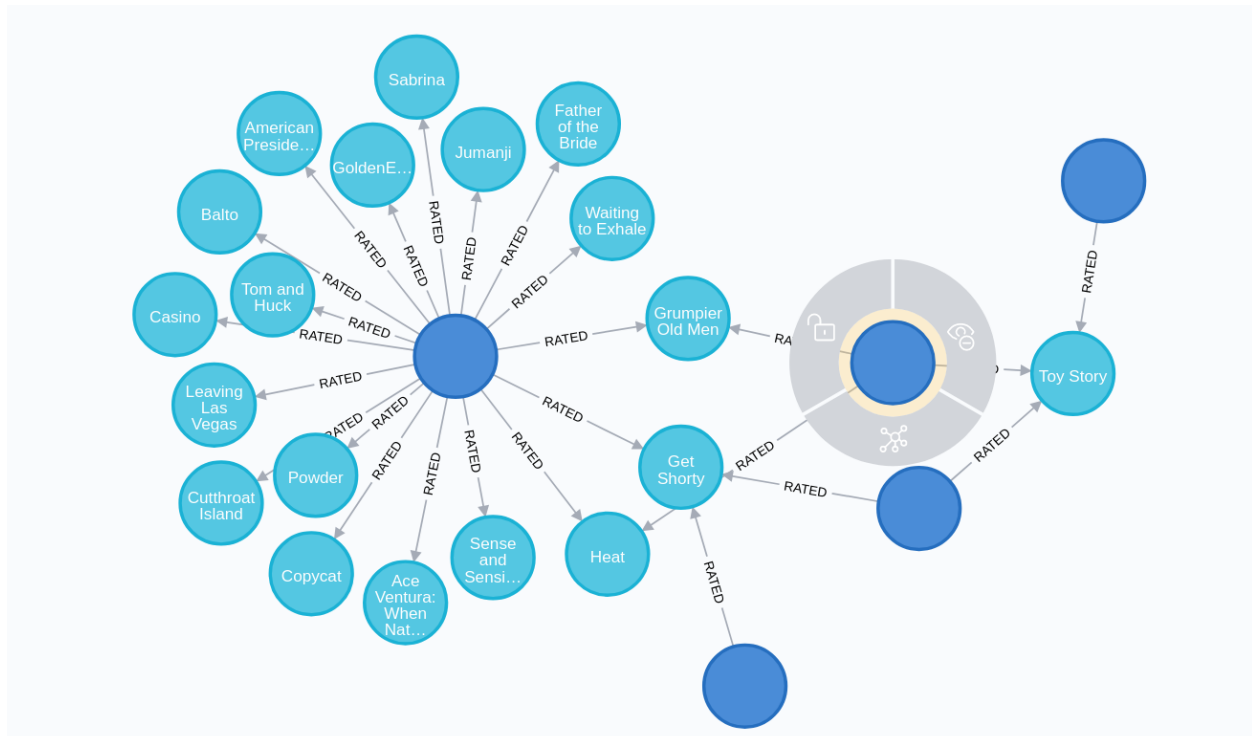
Cơ sở dữ liệu đồ thị tri thức là một loại cơ sở dữ liệu lưu trữ và quản lý dữ liệu bằng mô hình dữ liệu đồ thị. Trong cơ sở dữ liệu sơ đồ tri thức, dữ liệu được biểu diễn dưới dạng các nút và cạnh, trong đó các nút biểu thị các thực thể (chẳng hạn như người, địa điểm hoặc sự vật) và các cạnh biểu thị mối quan hệ giữa các thực thể đó.

Cơ sở dữ liệu sơ đồ tri thức được thiết kế để hỗ trợ các cấu trúc dữ liệu phức tạp, được kết nối với nhau và cho phép phân tích và truy vấn dữ liệu phức tạp. Chúng thường được sử dụng trong các ứng dụng như công cụ tìm kiếm, hệ thống đề xuất và hệ thống quản lý tri thức.

Một số chức năng chính của cơ sở dữ liệu đồ thị tri thức bao gồm:

- Mô hình hóa dữ liệu linh hoạt: Cơ sở dữ liệu sơ đồ tri thức hỗ trợ mô hình hóa dữ liệu linh hoạt, có nghĩa là dữ liệu có thể được tổ chức theo cách phản ánh mối quan hệ giữa các thực thể khác nhau.
- Truy vấn ngữ nghĩa: Cơ sở dữ liệu sơ đồ tri thức hỗ trợ truy vấn ngữ nghĩa, nghĩa là người dùng có thể truy vấn dữ liệu dựa trên mối quan hệ giữa các thực thể, thay vì chỉ trên chính các thực thể đó.
- Khả năng mở rộng: Cơ sở dữ liệu sơ đồ tri thức được thiết kế để mở rộng theo chiều ngang, có nghĩa là chúng có thể xử lý khối lượng dữ liệu lớn và có thể được phân phối trên nhiều máy chủ.
- Khả năng tương tác: Cơ sở dữ liệu sơ đồ tri thức hỗ trợ khả năng tương tác với các hệ thống và nguồn dữ liệu khác, có nghĩa là chúng có thể được tích hợp với các cơ sở dữ liệu và hệ thống khác để tạo ra một hệ sinh thái dữ liệu toàn diện hơn.

- Tích hợp dữ liệu: Cơ sở dữ liệu sơ đồ tri thức cho phép tích hợp dữ liệu trên các nguồn dữ liệu khác nhau, có nghĩa là dữ liệu từ nhiều nguồn có thể được kết hợp và phân tích cùng nhau.



Hình 11. Ảnh chụp minh họa cho một chuỗi dữ liệu dạng đồ thị tri thức

Đồ thị tri thức là một công cụ mạnh mẽ cho phép sắp xếp và trực quan hóa thông tin phức tạp theo cách dễ hiểu hơn. Một trong những ưu điểm lớn nhất của sơ đồ tri thức là nó cho phép ta nhìn thấy mối quan hệ giữa các mẫu thông tin khác nhau. Điều này đặc biệt hữu ích khi xử lý một lượng lớn dữ liệu, vì nó nhanh chóng xác định các kết nối có thể khó nhìn thấy. Bằng cách sử dụng biểu đồ tri thức, ta có thể đưa ra quyết định sáng suốt hơn, nâng cao hiểu biết của mình về các chủ đề phức tạp và cuối cùng trở nên hiệu quả và hiệu quả hơn trong công việc.

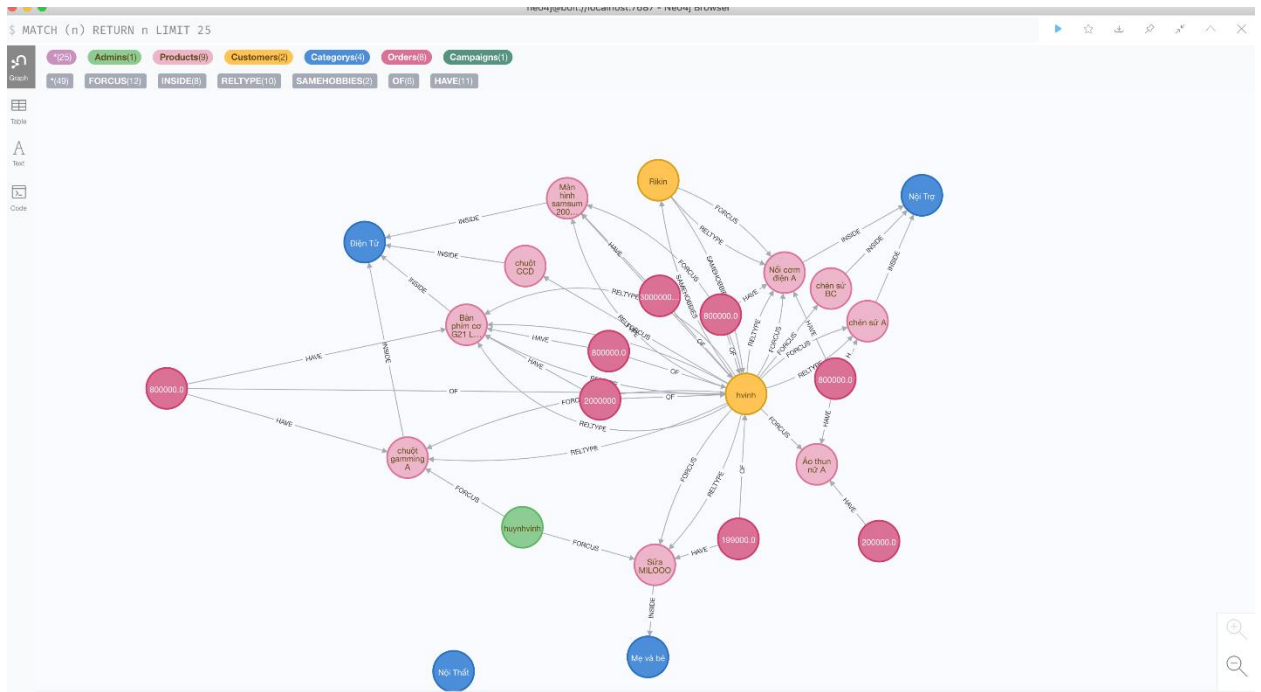
Một số yêu cầu cơ bản khi ứng dụng xây dựng một đồ thị tri thức:

- Định nghĩa một mô hình dữ liệu phù hợp: Cần định nghĩa một mô hình dữ liệu phù hợp để đại diện cho các khái niệm và mối quan hệ giữa chúng. Điều này đòi hỏi có sự hiểu biết rõ về các khái niệm và quan hệ liên quan đến lĩnh vực cần xây dựng đồ thị tri thức, cụ thể ở đây đang là về khả năng hiển thị đề xuất dựa trên ngữ nghĩa.
- Thu thập và chuẩn hóa dữ liệu: Cần thu thập dữ liệu từ nhiều nguồn khác nhau và chuẩn hóa dữ liệu để đảm bảo tính nhất quán và hiệu quả khi tìm kiếm và truy xuất thông tin.
- Xây dựng cấu trúc đồ thị: Sau khi thu thập dữ liệu, cần xây dựng cấu trúc đồ thị phù hợp để lưu trữ và quản lý dữ liệu đề xuất.
- Áp dụng các kỹ thuật phân tích ngôn ngữ tự nhiên: Các kỹ thuật phân tích ngôn ngữ tự nhiên, bao gồm trích xuất thông tin và tóm tắt nội dung, được sử dụng để phân tích và tổ chức dữ liệu trong knowledge graph.
- Tích hợp các công nghệ khác nhau: Để tăng cường tính hiệu quả và khả năng ứng dụng, knowledge graph có thể được tích hợp với các công nghệ khác như trí tuệ nhân tạo, machine learning, và các công nghệ xử lý ngôn ngữ tự nhiên khác.
- Kiểm tra và đánh giá: Khi hoàn thành xây dựng knowledge graph, cần kiểm tra và đánh giá tính hiệu quả và độ chính xác của nó để đảm bảo rằng nó phù hợp với mục đích sử dụng.

Và các tính năng kể trên được ứng dụng để biểu diễn kết quả đề xuất của kỹ thuật tư vấn dựa trên ngữ nghĩa. Một hệ quản trị cơ sở dữ liệu dạng đồ thị nổi tiếng và được sử dụng rộng rãi nhất là *Neo4j*.

2.2.3. Neo4j

Neo4j là cơ sở dữ liệu đồ thị. Dữ liệu sẽ được tổ chức dưới dạng đồ thị, mỗi đối tượng dữ liệu sẽ được lưu thành một nút (node) trong đồ thị và thường những nút này sẽ được gắn nhãn (label) để phân biệt các loại node với nhau. Mối tương quan giữa các node sẽ là các cạnh (relationships) thể hiện mối quan hệ giữa các đối tượng. Như hình minh họa phía dưới ta có thể dễ dàng hình dung ra ứng dụng của *Neo4j*



Hình 12. Ảnh chụp minh họa một mô hình của Neo4j

Với Neo4j, một hệ quản trị cơ sở dữ liệu dạng đồ thị phổ biến với các cấu trúc cơ bản của một đồ thị tri thức, gồm các thành phần như sau:

- Node: Một nút trong cơ sở dữ liệu đồ thị có thể lưu thông tin trên một node dưới dạng JSON² và được gắn label để phân biệt loại node phục vụ các thuật toán, các truy vấn
- Relationships: Là các cạnh trong cơ sở dữ liệu đồ thị, thể hiện mối quan hệ giữa các node, mỗi quan hệ này có thể gắn thêm giá trị (dạng JSON) trên các cạnh này, các cạnh này rất quan trọng trong việc truy vấn dữ liệu, sử dụng thuật toán.

Trong đó các dữ liệu về đối tượng (sản phẩm, người dùng...) được thể hiện bằng một node trong đồ thị tri thức.

² JSON: viết tắt của JavaScript Object Notation, là một kiểu dữ liệu mở trong JavaScript. Kiểu dữ liệu này bao gồm chủ yếu là text, có thể đọc được theo dạng cặp "thuộc tính - giá trị"

2.2.4. Ưu điểm của việc ứng dụng cơ sở dữ liệu đồ thị tri thức

Các hệ cơ sở dữ liệu sơ đồ tri thức nói chung, và *Neo4j* nói riêng, có một số ưu điểm tiêu biểu, bao gồm:

- Thứ nhất, Mô hình hóa dữ liệu linh hoạt: Cơ sở dữ liệu sơ đồ tri thức hỗ trợ mô hình hóa dữ liệu linh hoạt, có nghĩa là dữ liệu có thể được tổ chức theo cách phản ánh mối quan hệ giữa các thực thể khác nhau. Điều này cho phép phân tích và truy vấn dữ liệu tinh vi hơn so với cơ sở dữ liệu quan hệ truyền thống.
- Thứ hai, Truy vấn ngữ nghĩa: Cơ sở dữ liệu sơ đồ tri thức hỗ trợ truy vấn ngữ nghĩa, nghĩa là người dùng có thể truy vấn dữ liệu dựa trên mối quan hệ giữa các thực thể, thay vì chỉ trên chính các thực thể đó. Điều này cho phép thực hiện các truy vấn phức tạp hơn có tính đến ngữ cảnh và ý nghĩa của dữ liệu.
- Thứ ba, Khả năng mở rộng: Cơ sở dữ liệu sơ đồ tri thức được thiết kế để mở rộng theo chiều ngang, có nghĩa là chúng có thể xử lý khối lượng dữ liệu lớn và có thể được phân phối trên nhiều máy chủ. Điều này cho phép hiệu suất cao và khả năng mở rộng, ngay cả đối với các bộ dữ liệu rất lớn.
- Thứ tư, Khả năng tương tác: Cơ sở dữ liệu sơ đồ tri thức hỗ trợ khả năng tương tác với các hệ thống và nguồn dữ liệu khác, có nghĩa là chúng có thể được tích hợp với các cơ sở dữ liệu và hệ thống khác để tạo ra một hệ sinh thái dữ liệu toàn diện hơn. Điều này giúp truy cập và phân tích dữ liệu từ nhiều nguồn dễ dàng hơn.
- Thứ năm, Tích hợp dữ liệu: Cơ sở dữ liệu sơ đồ tri thức cho phép tích hợp dữ liệu trên các nguồn dữ liệu khác nhau, có nghĩa là dữ liệu từ nhiều nguồn có thể được kết hợp và phân tích cùng nhau. Điều này cho phép phân tích toàn diện hơn và có thể giúp xác định các mối quan hệ và mẫu có thể không hiển thị trong các bộ dữ liệu riêng lẻ.
- Thứ sáu, Thông tin chi tiết tốt hơn: Cơ sở dữ liệu sơ đồ tri thức cho phép người dùng có được thông tin chi tiết mới và khám phá các mối quan hệ mới giữa các điểm dữ liệu. Bằng cách lập mô hình dữ liệu dưới dạng biểu đồ, cơ sở dữ liệu sơ đồ tri thức cho phép người dùng xem các thực thể khác nhau được kết nối với nhau như thế

nào, điều này có thể tiết lộ các mẫu và thông tin chi tiết mới có thể không rõ ràng trong cơ sở dữ liệu truyền thống.

Và hệ cơ sở dữ liệu do *Neo4j* cung cấp cũng đã đem đến những lợi ích kể trên, chủ yếu tập trung đem lại khả năng mô hình hóa dữ liệu tuyệt vời hơn, trực quan hơn so với phương pháp thể hiện dữ liệu truyền thống trước kia. Tuy nhiên, vẫn còn đó tồn tại những nhược điểm đến từ cơ sở dữ liệu đồ thị nói chung, và từ *Neo4j* nói riêng mang lại.

2.2.5. Nhược điểm của việc ứng dụng cơ sở dữ liệu đồ thị tri thức

Mặc dù có nhiều lợi ích khi sử dụng cơ sở dữ liệu đồ thị tri thức, nhưng cũng có một số nhược điểm tiềm ẩn cần xem xét. Chúng bao gồm:

- Thứ nhất, *Quá phức tạp*: Cơ sở dữ liệu đồ thị tri thức có thể quá phức tạp để thiết lập và sử dụng, đặc biệt đối với người dùng không quen thuộc với cơ sở dữ liệu đồ thị hoặc ngôn ngữ truy vấn mà cơ sở dữ liệu sử dụng.

- Thứ hai, *Hiệu suất truy vấn*: Mặc dù cơ sở dữ liệu sơ đồ tri thức cho phép truy vấn phức tạp hơn so với cơ sở dữ liệu truyền thống, nhưng hiệu suất truy vấn có thể chậm hơn đối với các tập dữ liệu rất lớn. Đây có thể là một vấn đề đối với các ứng dụng yêu cầu thời gian phản hồi truy vấn nhanh mà cơ sở dữ liệu đồ thị lại không đáp ứng được.

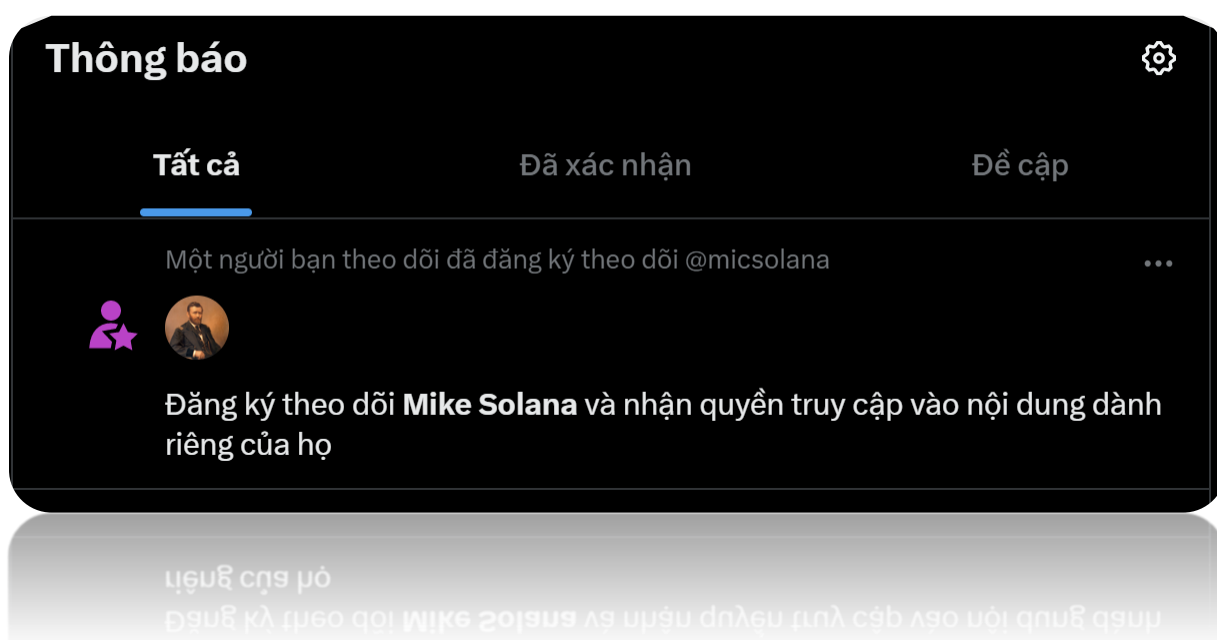
- Thứ ba, *Thiếu tiêu chuẩn*: Hiện tại không có tiêu chuẩn cho cơ sở dữ liệu đồ thị tri thức, điều đó có nghĩa là có thể có sự khác biệt đáng kể trong cách các cơ sở dữ liệu khác nhau triển khai mô hình đồ thị và ngôn ngữ truy vấn. Điều này có thể gây khó khăn cho việc di chuyển dữ liệu giữa các hệ thống khác nhau hoặc tích hợp với các cơ sở dữ liệu hoặc công cụ khác.

- Thứ tư, *Hệ sinh thái hạn chế*: Mặc dù hệ sinh thái các công cụ và thư viện để làm việc với cơ sở dữ liệu đồ thị tri thức đang phát triển, nhưng nó vẫn còn tương đối hạn chế so với các loại cơ sở dữ liệu khác. Điều này có thể gây khó khăn hơn trong việc tìm kiếm các nguồn lực và kiến thức chuyên môn cần thiết để làm việc với các cơ sở dữ liệu này một cách hiệu quả.

CHƯƠNG 3. TÍNH ỨNG DỤNG

3.1. Ứng dụng của kỹ thuật đề xuất dựa trên ngữ nghĩa trong các lĩnh vực công nghệ thông tin [0]

3.1.1. Đề xuất người theo dõi dựa trên ngữ nghĩa trên nền tảng mạng xã hội Twitter [1]



Hình 13. Một ví dụ của đề xuất người theo dõi trên Twitter

Giới thiệu: [2][3]

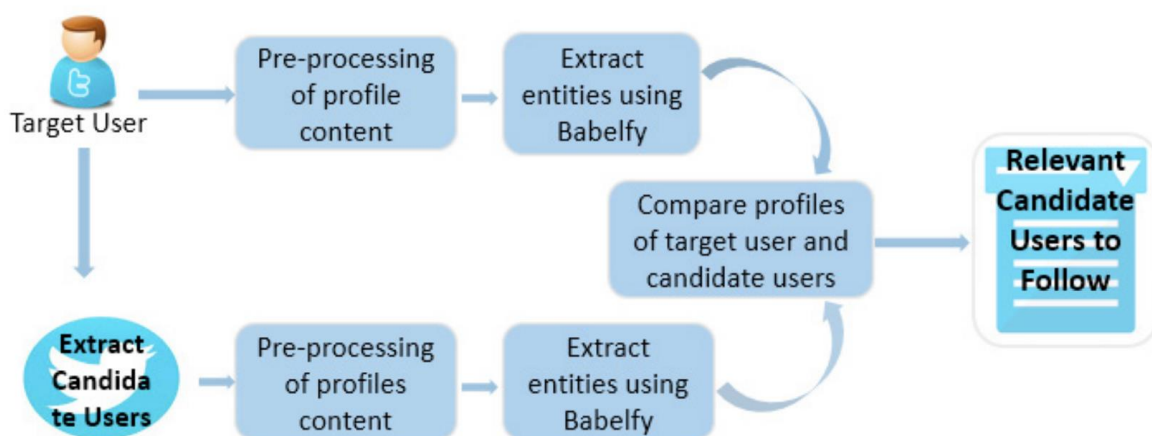
- Các người dùng Twitter sử dụng khái niệm người theo dõi/người được theo dõi để thông báo và được thông báo về tất cả các hoạt động gần đây của người dùng có sở thích và sở thích tương tự. Hơn nữa, việc tìm kiếm những người dùng có liên quan để theo dõi trở thành một nhiệm vụ quan trọng do tốc độ tăng trưởng nhanh chóng của mạng Twitter và số lượng người dùng đăng ký hàng ngày rất lớn. Vì vậy, nhu cầu về một hệ thống hỗ trợ người dùng trong công việc đó là hết sức cần thiết.
- Thật vậy, các nghiên cứu gần đây sử dụng phân tích từ vựng để ứng dụng cho vấn đề được đặt ra ở trên. Vì thế để đề xuất một người theo dõi hệ thống đề xuất dựa

trên phân tích ngữ nghĩa của nội dung hồ sơ người dùng bằng cách tận dụng cấu trúc liên kết người theo dõi/người được theo dõi.

Cách tiếp cận:

Theo bài báo khoa học *Semantic-based Followee Recommendations on Twitter Network* được đăng tải trên *ScienceDirect*, hệ thống đề xuất người theo dõi trong Twitter nhằm mục đích giúp người dùng tìm kiếm thông tin và bao gồm việc xác định người dùng nguồn đăng các tweet quan trọng cho người dùng mục tiêu, để những người dùng mục tiêu có thể theo dõi những người dùng nguồn đó và nhận thông tin theo thời gian thực từ họ. Trong tổng thể, hệ thống gợi ý bao gồm ba bước chính:

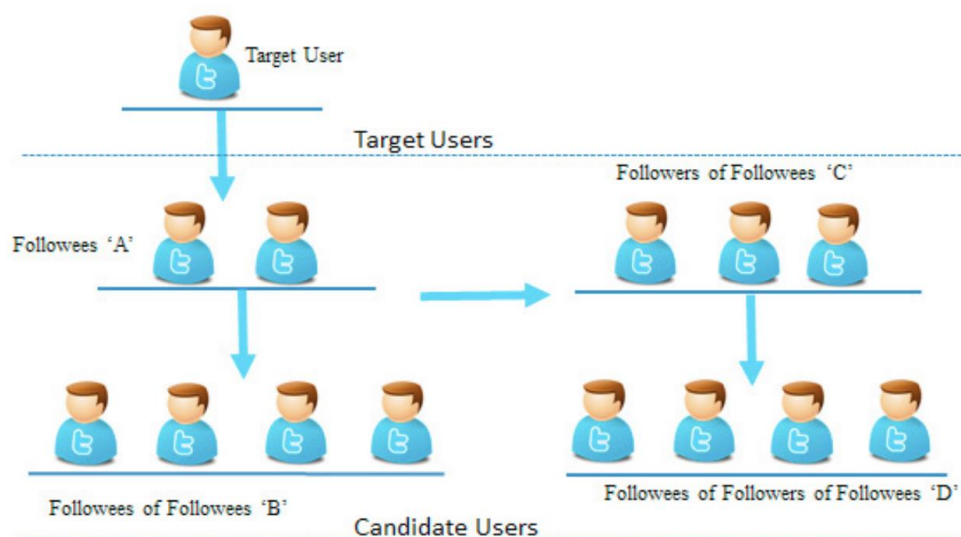
- Đầu tiên, nó trích xuất biểu đồ mạng lưới người theo dõi của người dùng mục tiêu để chọn người dùng ứng cử viên được đánh giá để giới thiệu.
- Thứ hai, hệ thống áp dụng quy trình chuẩn hóa nội dung hồ sơ, sau đó sử dụng chiến lược lập hồ sơ để suy ra mục tiêu và sở thích của người dùng ứng viên.
- Bước cuối cùng thực hiện so sánh ngữ nghĩa giữa hồ sơ mục tiêu và các hồ sơ khác để đề xuất danh sách người dùng nguồn mà người dùng mục tiêu theo dõi.



Hình 14. Ảnh chụp minh họa kiến trúc hệ thống đề xuất người theo dõi trên Twitter
Tìm kiếm ứng viên dựa trên cấu trúc liên kết:

Trước hết, trong hệ thống đề xuất người dùng Twitter, cần xác định một nhóm người dùng ứng viên “ SUC ” để đề xuất cho người dùng mục tiêu “ U_T ”. Ở đây ta tận dụng cấu trúc liên kết mạng của số lượng người theo dõi/người được theo dõi để đạt được nhiệm vụ này. Thật vậy, ta sử dụng giả thuyết rằng các ứng viên có thể được thu thập từ những người đang được theo dõi U_T , là mục tiêu của người dùng. *Hình 6* bên dưới minh họa một ví dụ về việc xây dựng một nhóm người dùng ứng viên “ SUC ” bằng cách thực hiện các bước sau:

- Tìm kiếm danh sách người dùng mục tiêu theo dõi: A.
- Trích xuất danh sách những người theo dõi của mỗi người dùng từ A và tạo một liên kết (B) của tất cả những người theo dõi này.
- Trích xuất danh sách những người theo dõi của mỗi người dùng từ A và tạo một liên kết (C) của tất cả những người theo dõi này.
- Đối với mỗi người dùng trong C, nhận được những người theo dõi của anh ấy/cô ấy. D là sự kết hợp của tất cả những người theo dõi này.
- Tạo một nhóm người dùng ứng cử viên “ SUC ” bao gồm sự kết hợp của B, C và D, loại trừ những người theo dõi người dùng mục tiêu hiện tại.



Hình 15. Biểu đồ mạng lưới người theo dõi/người được theo dõi để tìm người dùng ứng viên

Đề xuất người theo dõi dựa trên ngữ nghĩa:

Sau khi xây dựng một nhóm người dùng ứng viên “SU_c” thì áp dụng chiến lược lập hồ sơ để tìm sở thích của người dùng ứng viên và người dùng mục tiêu. Chiến lược này phân tích ngữ nghĩa văn bản của các tweet và re-tweet của anh ấy/cô ấy. Ta có thể tin rằng giải pháp thay thế tốt nhất để xây dựng hồ sơ người dùng trên Twitter là phân tích các tweet/re-tweet của chính họ, bởi vì các bài đăng của người dùng chủ yếu là về những điều mà họ quan tâm.

Đối với người dùng mục tiêu U_T, ta lưu ý các tweet (U_T) tập hợp tất cả các tweet và re-tweet của anh ấy/cô ấy, trong khi: tweet (U_T) = {t₁, t₂, ..., t_n}.

Vì vậy, hồ sơ của người dùng là một vector trong đó các thực thể được tính trọng số theo tần suất xuất hiện của chúng trong văn bản tweet (U_T). Để trích xuất các thực thể đó từ các tweet và liên kết chúng với cơ sở tri thức (BabelNet³), sử dụng Babelfy, một hệ thống phân biệt định hướng ngữ nghĩa đa ngôn ngữ (Word Sense Disambiguation – WSD [4]) và liên kết thực thể thống nhất, đa ngôn ngữ dựa trên BabelNet giúp định hướng và truy xuất văn bản tweet được viết bằng các ngôn ngữ khác nhau, đồng thời tạo ra dữ liệu được liên kết đa ngôn ngữ làm đầu ra. Babelfy xử lý tốt cả văn bản dài, chẳng hạn như văn bản của các WSD và cả các câu ngắn. Độ tương tự giữa các vector (hồ sơ người dùng) được tính toán bằng phép đo độ tương tự cosine.[4]

Đánh giá thực nghiệm

- Mô tả tập dữ liệu:

Tập dữ liệu được sử dụng trong nghiên cứu là mối quan hệ biểu đồ xã hội của người theo dõi/người được theo dõi giữa 400.000 người dùng và 20 tweet được xuất bản gần đây nhất của họ, tương ứng đạt tổng số 8.000.000 tweet. Nghiên

³ <https://www.babelnet.org/>

cứu đã chọn ngẫu nhiên 200 người dùng có hơn 100 người theo dõi làm người dùng mục tiêu. *API Twitter4J*⁴ được sử dụng để thu thập người dùng, biểu đồ xã hội của người theo dõi/người được theo dõi và các tweet của họ.

- Phương pháp:

Trong suốt quá trình thử nghiệm, nhóm người theo dõi của mỗi người dùng được chia thành hai phần. Phần đầu tiên chứa 70% số người theo dõi để đào tạo và 30% số người theo dõi để thử nghiệm.

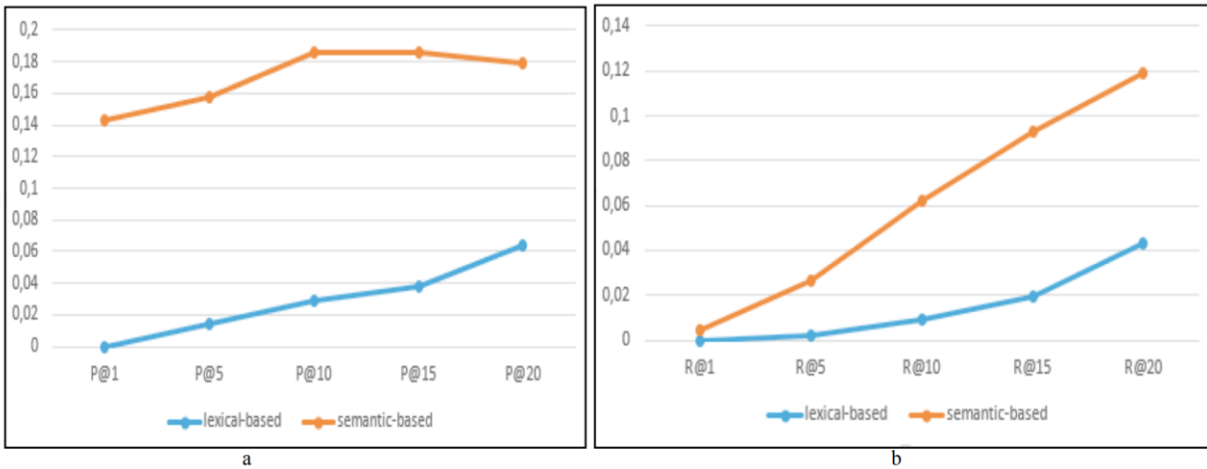
Chất lượng của danh sách đề xuất N người theo dõi hàng đầu được tạo ra được đánh giá bằng cách tính toán độ chính xác ở thứ hạng 1, 5, 10, 15 và 20, độ chính xác ở thứ hạng k được định nghĩa là tỷ lệ người theo dõi được đề xuất có liên quan, tức là nằm trong nhóm thử nghiệm của người dùng mục tiêu.

$$\text{Độ chính xác} = \frac{(\text{Relevant_Candidates_Recommended trong } \text{top-}k)}{(k_Candidates_Recommended)}$$

Hình 16. Cách tính độ chính xác ở thứ hạng k

- Kết quả:

⁴ <https://twitter4j.org/>



Hình 17. Độ chính xác của cách tiếp cận dựa trên ngữ nghĩa và cách tiếp cận dựa trên từ vựng trên thí nghiệm quy trình đề xuất

Ở hai ảnh 9a và 9b, là kết quả hiển thị hiệu suất, độ chính xác và khả năng thu hồi ở thứ hạng 1, 5, 10, 15 và 20 bằng cách sử dụng chiến lược đề xuất dựa trên ngữ nghĩa bằng cách xem xét phương pháp tiền xử lý đã được áp dụng, so với hiệu suất của kỹ thuật đề xuất dựa trên từ vựng.

Theo các kết quả thử nghiệm này, quan sát thấy rằng cách tiếp cận bao gồm phân tích hồ sơ người dùng về mặt ngữ nghĩa có kết quả rõ ràng so với cách tiếp cận dựa trên từ vựng được trình bày trên đường cơ sở của biểu đồ hình ảnh kết quả, do đó chứng tỏ rằng việc xử lý lỗi hồng ngữ nghĩa và ngữ cảnh là cao hơn trong nội dung các bài tweet phù hợp hơn với chất lượng đề xuất những người dùng có cùng chí hướng.

Kết luận

Bài nghiên cứu đã chỉ ra một hệ thống giới thiệu người theo dõi trên nền tảng xã hội Twitter dựa trên ngữ nghĩa. Trên thực tế, phương pháp này hoạt động bằng cách tận dụng cấu trúc liên kết người theo dõi/người đang được theo dõi để tìm người dùng ứng viên được đề xuất, xử lý từng hồ sơ người dùng theo ngữ nghĩa và cuối cùng đề xuất người dùng phù hợp nhất bằng cách đối sánh hồ sơ của họ. Vì mục đích đó, nghiên cứu

đã sử dụng Babelfy từ BabelNet, cho phép cải thiện chất lượng của hệ thống đề xuất của mình. Hơn nữa, nghiên cứu đã so sánh cách tiếp cận dựa trên ngữ nghĩa với cách tiếp cận dựa trên từ vựng được đề xuất trên đường cơ sở của mô hình kết quả. Và quan trọng không kém, sẽ sử dụng các bước tiền xử lý tinh vi hơn để đối phó với xác suất xảy ra sự khan hiếm ngữ cảnh trong các chủ đề được tweet.

3.1.2. Đề xuất dựa trên ngữ nghĩa cho hệ thống tổng hợp tin tức thể thao [5]

Giới thiệu

Công cụ tổng hợp tin tức là các trang web thu thập tin tức từ nhiều nguồn khác nhau và cung cấp một cái nhìn tổng hợp về các sự kiện đang diễn ra trên khắp thế giới. Thật không may, một vấn đề nghiêm trọng của các hệ thống tổng hợp tin tức là số lượng lớn tin tức được xuất bản hàng ngày cản trở người đọc khi họ muốn tìm những tin tức liên quan đến sở thích cụ thể của họ. Một giải pháp khả thi cho vấn đề này là sử dụng các hệ thống gợi ý vì chúng có thể duyệt qua không gian của các lựa chọn và dự đoán mức độ hữu ích tiềm năng của tin tức đối với mỗi người đọc.

Cách tiếp cận

Các công trình nghiên cứu gần đây về đo lường độ tương tự tin tức dựa trên hai cách tiếp cận nổi bật: đề xuất độ tương tự dựa trên nội dung và đề xuất độ tương tự dựa trên ngữ nghĩa.

Theo cách tiếp cận dựa trên nội dung, độ tương đồng của tin tức được tính toán dựa trên thống kê từ vựng xuất hiện trong nội dung tin tức và hầu hết các tin tức được đề xuất đều chỉ tập trung vào một chủ đề mà tin tức đó có chủ đích hướng tới. Ngược lại, trong cách tiếp cận dựa trên ngữ nghĩa, độ tương đồng của tin tức thường dựa trên cơ sở tri thức sẵn có để khai thác ngữ nghĩa.

Do đó, tin tức được đề xuất dựa trên ngữ nghĩa sẽ có khả năng mở rộng đối tượng hơn so với cách tiếp cận dựa trên nội dung.

Cách tiếp cận của bài nghiên cứu là một phương pháp kết hợp giữa đề xuất dựa trên nội dung và đề xuất dựa trên ngữ nghĩa. Nói một cách cụ thể, sự tương đồng nhau của các bài tin tức là sự kết hợp tuyến tính giữa sự giống nhau dựa trên nội dung và sự giống nhau dựa trên ngữ nghĩa. Kết quả thực nghiệm cho thấy sự kết hợp này mang lại kết quả gợi ý tin tức hiệu quả hơn so với việc sử dụng riêng lẻ từng biện pháp.

Thông thường, nhiều hệ đề xuất dựa trên nội dung sử dụng các phương pháp trích xuất thuật ngữ như TF-IDF (Tần số tài liệu nghịch đảo tần suất dữ liệu - Term Frequency-Inverse Document Frequency [6]) kết hợp với phép đo độ tương tự cosine để so sánh độ tương tự giữa hai bản tin. TF-IDF được sử dụng để đo tầm quan trọng của một từ trong bản tin dựa trên tần suất xuất hiện của từ đó trong toàn bộ tập dữ liệu của bản tin. Sau khi tính toán giá trị TF-IDF cho mỗi từ trong bản tin, số liệu này được kết hợp với thước đo Cosine để tính toán độ tương tự giữa hai bản tin tức. Giá trị TF-IDF của từ xuất hiện trong bản tin được tính theo công thức sau:

$$TF-IDF_{ij} = TF_{ij} \times IDF_i$$

$$\text{với} \begin{cases} TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \\ IDF_i = \log \frac{|D|}{|\{d:t_i \in d\}|} \end{cases}$$

Hình 18. Ảnh chụp cho công thức tính giá trị Term Frequency-Inverse Document Frequency TF-IDF [6]

Trong đó: n_{ij} là số lần xuất hiện của từ i trong tài liệu j

$|D|$ là tổng số tài liệu trong tập dữ liệu.

Mức độ tương đồng giữa các mục tin tức

Như đã đề cập, có hai cách tiếp cận chính trong việc tính toán độ tương tự giữa các văn bản tin tức là dựa trên nội dung và dựa trên ngữ nghĩa. Mỗi cách tiếp cận đều có ưu điểm và nhược điểm riêng. Nghiên cứu mong muốn kết hợp hai cách tiếp cận này bằng cách kết hợp phép đo tương đồng dựa trên nội dung và phép đo tương đồng dựa trên ngữ nghĩa với mong muốn khắc phục những hạn chế của từng cách tiếp cận, giúp kết quả đề xuất trở nên hiệu quả hơn.

Mức độ tương đồng của đề xuất dựa trên ngữ nghĩa

Để tính toán độ tương đồng ngữ nghĩa, nghiên cứu đã khai thác mối quan hệ ngữ nghĩa lẫn nhau giữa các thành phần trong các bản tin. Các quan hệ này được xác định dựa trên bản thể học⁵ và cơ sở tri thức đã được xây dựng. Sau đó trích xuất và phân tích các thành phần trong tin gồm: thực thể, loại thực thể và chú thích ngữ nghĩa.

Cụ thể, để khai thác mối quan hệ giữa các thực thể để tính toán độ tương tự giữa các mục tin tức, ta mở rộng phương pháp *Ranked Semantic Recommendation 2* đã được phê duyệt bởi *Frasincar et al.* [7]. Trong phương pháp này, nhóm tác giả cũng sử dụng bản thể học và cơ sở tri thức để khai thác các mối quan hệ giữa các thực thể. Tuy nhiên, phương pháp vẫn còn một số hạn chế như:

- Nó chỉ xét quan hệ trực tiếp giữa các chủ thể mà không xét quan hệ gián tiếp
- Không xét đến tầm quan trọng của các thực thể khi chúng xuất hiện ở nhiều vị trí khác nhau trong văn bản tin tức (như tiêu đề, mô tả...)

Để khắc phục những hạn chế trên, nghiên cứu trình bày phương pháp tính trọng số quan hệ giữa các thực thể dựa trên bản thể học và cơ sở tri thức. Ngoài ra, còn kết hợp

⁵ Bản thể học là các nền tảng có cấu trúc cho việc tổ chức thông tin được áp dụng trong các lĩnh vực như trí tuệ nhân tạo, web ngữ nghĩa (semantic web)

phương pháp thống kê về sự xuất hiện đồng thời của các thực thể trong cùng một tin để xác định trọng số quan hệ giữa các thực thể được trình bày. Cuối cùng, trình bày phương pháp sử dụng trọng số quan hệ giữa các thực thể trong việc xác định độ tương đồng ngữ nghĩa giữa các mẫu tin.

Trọng số quan hệ giữa các thực thể dựa trên bản thể học và cơ sở tri thức:

- Aleman-Meza và cộng sự đã trình bày các phương pháp tính toán xếp hạng của *Semantic Association based on Semantic Path* giữa hai thực thể nhằm xác định trọng số quan hệ giữa các thực thể [8]. Cụ thể, họ định nghĩa như sau: nếu hai thực thể e_1 và e_n có thể được kết nối với nhau bằng một hoặc nhiều dãy $e_1, P_1, e_2, P_2, e_3, P_3, \dots, e_{n-1}, P_{n-1}, e_n$ trong đồ thị RDF; ở đây e_i , với $1 \leq i \leq n$, là các thực thể và P_j , với $1 \leq j \leq n$ là các quan hệ trong bản thể học thì ta nói giữa e_1 và e_n tồn tại quan hệ ngữ nghĩa. Ta nói: Dãy $e_1, P_1, e_2, P_2, e_3, P_3, \dots, e_{n-1}, P_{n-1}, e_n$ là một đường dẫn ngữ nghĩa (*Semantic Path*).

- Ví dụ: trong cơ sở tri thức, chúng ta có:

$\langle \text{Lionel-Messi} \rangle \langle \text{playFor} \rangle \langle \text{Barcelona-FC} \rangle$

$\langle \text{Luis-Suarez} \rangle \langle \text{playFor} \rangle \langle \text{Barcelona-FC} \rangle$

Khi đó, tồn tại đường dẫn ngữ nghĩa giữa hai thực thể Lionel Messi và Luis Suarez như sau:

$\langle \text{Lionel-Messi} \rangle \rightarrow \langle \text{playFor} \rangle \rightarrow \langle \text{Barcelona-FC} \rangle \leftarrow \langle \text{playFor} \rangle \leftarrow \langle \text{Luis-Suarez} \rangle$

- Dựa trên các thuộc tính của đường dẫn ngữ nghĩa, từ đó xác định một giá trị xếp hạng đường dẫn để hiển thị trọng số quan hệ giữa hai thực thể ở cả hai đầu của đường dẫn. Bởi vì có thể có nhiều đường dẫn ngữ nghĩa giữa hai thực thể, ta quy ước lấy giá trị thứ hạng đường dẫn cao nhất để biểu thị trọng số quan hệ. Theo nghiên cứu của Aleman-Meza và cộng sự [8] đã sử dụng bốn đặc điểm của một đường dẫn ngữ nghĩa để tính thứ hạng đường dẫn, chúng gồm có:

- Subsumption Weight: dựa vào cấu trúc của bản thể học để xác định trọng số thành phần cho từng thành phần (thực thể) trong đường dẫn, từ đó tính ra trọng số cho cả đường dẫn.
- Path Length Weight: dựa trên độ dài của đường dẫn.
- Context Weight: dựa trên việc xác định mỗi thành phần của đường đi thuộc vùng nào trong bản thể học. Mỗi vùng trong bản thể có một trọng số riêng tùy thuộc vào sở thích của người dùng.
- Trust Weight: dựa trên trọng số của các thuộc tính trong bản thể học.

- Áp dụng trong đề xuất tin tức, nghiên cứu nhận thấy Path Length Weight và Trust Weight là hai trọng số có ý nghĩa và phù hợp. Vì lý do này mà chỉ sử dụng hai trọng số này để xác định thứ hạng đường dẫn của đường dẫn ngữ nghĩa.

- Với *Path Length Weight*: Độ dài của một đường ngữ nghĩa $e_1, P_1, e_2, P_2, e_3, P_3, \dots, e_{n-1}, P_{n-1}, e_n$ là số lượng thực thể và quan hệ trong đường dẫn (ngoại trừ e_1 và e_n). Ta có thể thấy rằng, khi hai thực thể còn quan hệ gián tiếp với nhau, qua đó càng có nhiều thực thể và quan hệ thì tính tương đồng giữa hai thực thể đó càng thấp. Do đó, thứ hạng đường dẫn của một đường dẫn ngữ nghĩa phải tỷ lệ nghịch với độ dài của đường dẫn đó.

- Trọng lượng chiều dài đường dẫn được xác định trong như sau:

$$W_{length} = \frac{1}{length_{path}}$$

Hình 19. Công thức xác định trọng lượng chiều dài đường dẫn (*Path Length Weight*)

Trong đó: $length_{path}$ là độ dài của đường dẫn ngữ nghĩa (*length of semantics path*)

W_{path} là trọng số của *Path Length Weight*

- Ví dụ: ta có hai đường dẫn ngữ nghĩa như sau:

Path 1: $\langle \text{Lionel Messi} \rangle \rightarrow \langle \text{playFor} \rangle \rightarrow \langle \text{Barcelona-FC} \rangle \rightarrow \langle \text{competeIn} \rangle \rightarrow \langle \text{La-Liga} \rangle \leftarrow \langle \text{competeIn} \rangle \leftarrow \langle \text{Real-Madrid} \rangle \leftarrow \langle \text{playFor} \rangle \leftarrow \langle \text{Karim Benzema} \rangle$

Path 2: $\langle \text{Lionel-Messi} \rangle \rightarrow \langle \text{playFor} \rangle \rightarrow \langle \text{Barcelona-FC} \rangle \leftarrow \langle \text{playFor} \rangle \leftarrow \langle \text{Luis-Suarez} \rangle$

Path 1 có độ dài bằng 7, ta thu được:

$$W_{length}(P_1) = \frac{1}{length_{path}} = \frac{1}{7}$$

Hình 20. Trọng lượng độ dài W của path 1

Path 2 có độ dài bằng 3, ta thu được:

$$W_{length}(P_2) = \frac{1}{length_{path}} = \frac{1}{3}$$

Hình 21. Trọng lượng độ dài W của path 2

Từ đó, ta có thể thấy sự tương đồng giữa Lionel Messi và Luis Suarez cao hơn so với Lionel Messi và Karim Benzema.

- Với *Path Relation Weight*: Có nhiều quan hệ khác nhau được định nghĩa trong bản thể học. Mỗi quan hệ đại diện cho một ý nghĩa khác nhau do đó cũng đại diện cho một trọng số quan hệ khác nhau giữa các thực thể. Một số quan hệ thể hiện sự liên kết chặt chẽ, một số quan hệ khác thể hiện sự liên kết lỏng lẻo. Ví dụ: có hai bộ ba trong cơ sở tri thức như bên dưới:

$\langle Luis\ Enrique \rangle \langle managerOf \rangle \langle Barcelona-FC \rangle$

$\langle Luis\ Suarez \rangle \langle playFor \rangle \langle Barcelona-FC \rangle$

- Ở đây tồn tại hai quan hệ là quan hệ $\langle managerOf \rangle$ và quan hệ $\langle playFor \rangle$. Ta có thể thấy rằng, quan hệ $\langle managerOf \rangle$ thể hiện sự gần gũi hơn quan hệ $\langle playFor \rangle$, vì xét trên ngữ nghĩa thì mỗi đội chỉ có một người quản lý duy nhất tại một thời điểm nhất định. Tuy nhiên, đội bóng sẽ có rất nhiều người chơi (cầu thủ). Do đó, chỉ định trọng số của $\langle managerOf \rangle$ cao hơn $\langle playFor \rangle$. Và vì lý do này, từ bộ ba trên, có thể kết luận $\langle Barcelona-FC \rangle$ có độ tương đồng với $\langle Luis-Enrique \rangle$ cao hơn so với $\langle Luis\ Suarez \rangle$. Trọng số của các quan hệ nằm trong khoảng $(0, 1]$. *Path Relation Weight* của một đường dẫn tổng thể P được định nghĩa như sau:

$$W_{predicate} = \prod_{p \in path} w_p$$

Hình 22. Định nghĩa về trọng số đường dẫn quan hệ của một đường dẫn P

Thuật toán đề xuất tin tức với sự tương đồng kết hợp

Để kết hợp độ tương đồng ngữ nghĩa $Similarity_{semantic}$ với độ tương đồng nội dung $Similarity_{TF-IDF}$ của hai bản tin, ta sử dụng cặp trọng số $\gamma_{semantic}$ và $\gamma_{content}$. Ta xác định công thức kết hợp như sau:

$$Similarity_{combined}(A, B) = Similarity_{semantic}(A, B) \times \gamma_{semantic} + Similarity_{TF-IDF} \times \gamma_{content}$$

Hình 23. Xác định độ tương đồng ngữ nghĩa dựa trên công thức kết hợp

Từ đó, xác định thuật toán đề xuất tin tức như sau:

Dữ liệu vào: Xác định mục tiêu tin tức A và đặt N ứng cử cho tin tức C .

Dữ liệu ra: tập tin K có độ tương đồng ngữ nghĩa cao nhất với A .

- Bước 1: Xác định đối tượng đặt tên, chú thích ngữ nghĩa cho tin A

và ứng cử mẫu tin tức trong bộ C .

- Bước 2: Xây dựng tập từ ngữ có trọng số TF-IDF cao nhất cho mẫu tin A và các mẫu tin trong tập C .
- Bước 3: Với mỗi tin C_i trong tập C , thực hiện các bước con sau:
 - Bước 3.1: Tính $Similarity_{based-entity}(A, C_i)$
 - Bước 3.2: Tính $Similarity_{based-annotation}(A, C_i)$
 - Bước 3.3: Tính $Similarity_{based-type}(A, C_i)$
 - Bước 3.4: Tính $Similarity_{semantic} A, C_i$ dựa trên kết quả của các bước 3.1, 3.2 và 3.3.
 - Bước 3.5: Tính $Similarity_{TF-IDF}(A, C_i)$
 - Bước 3.6: Tính $Similarity_{combined} A, C_i$ dựa trên kết quả của bước 3.4 và 3.5.
- Bước 4: Sắp xếp các tin C_i giảm dần theo giá trị $Similarity_{combined}(A, C_i)$.
- Bước 5: Lấy k tin ở đầu danh sách đã sắp xếp ở Bước 4 để giới thiệu cho mẫu tin A .

Đánh giá thực nghiệm

Sau khi chạy ba phương pháp riêng biệt cho một tập A chứa 100 mẫu tin tức theo kịch bản thử nghiệm, nghiên cứu thu được kết quả của từng phương pháp như trong Bảng 1:

Bảng 1. Kết quả khuyến nghị tin tức trong các trường hợp [5]

Phương pháp	Mức độ chính xác
Chỉ có content-based (đề xuất dựa trên nội dung)	82.2%
Chỉ có semantics-based (đề xuất dựa trên ngữ nghĩa)	75.8%
Kết hợp cả hai phương pháp trên	85.6%

Từ bảng 1 chỉ ra rằng, đối với tập dữ liệu thử nghiệm A chứa 100 mục tin tức, phương pháp đề xuất dựa trên ngữ nghĩa không chính xác bằng phương pháp đề xuất dựa trên nội dung. Trong khi đó, nếu kết hợp cả hai phương pháp kể trên thì sẽ mang lại hiệu quả tốt nhất. Điều này có thể được giải thích như sau:

- Khi chỉ sử dụng tương đồng ngữ nghĩa (semantic-based approach) thì chủ yếu phụ thuộc vào các thực thể trong các mẫu tin. Do đó, trong một số trường hợp, thuật toán đề xuất các mục tin chính xác về các thực thể có liên quan nhưng chủ đề hoàn toàn khác.

- Theo cách tiếp cận dựa trên nội dung, chủ đề của tin đề xuất thường khá gần với tin mục tiêu. Tuy nhiên, phương pháp này không có khả năng mở rộng chủ đề. Nếu chúng ta có hai tin tức về câu lạc bộ Barcelona, trong đó tin đầu tiên là về lối chơi của Câu lạc bộ và tin thứ hai là về việc chuyển nhượng cầu thủ của Câu lạc bộ, cách tiếp cận dựa trên nội dung sẽ xác định rằng sự giống nhau của các tin này là thấp.

- Khi kết hợp giữa tương đồng nội dung và tương đồng ngữ nghĩa, tin được khuyến nghị sẽ khắc phục được hạn chế của từng phương pháp riêng biệt, dẫn đến hiệu quả khuyến nghị cao hơn.

Kết luận

Trong nghiên cứu này đã trình bày phương pháp gợi ý dựa trên sự kết hợp giữa sự giống nhau về nội dung và sự giống nhau về ngữ nghĩa của các mẫu tin tức. Độ đo dựa trên ngữ nghĩa được tính toán dựa trên mối quan hệ ngữ nghĩa giữa các đối tượng. Nó cho phép gợi ý không chỉ dừng lại ở gợi ý các tin cùng chủ đề hoặc các tin xoay quanh một

đối tượng chính của tin mục tiêu, mà còn có thể gợi ý các tin của các đối tượng khác mà các đối tượng này có quan hệ ngữ nghĩa với các đối tượng khác. Tuy nhiên, độ đo tương tự chủ yếu tập trung vào thực thể chứ chưa xét đến ngữ cảnh được đề cập trong mẫu tin. Phương pháp đo lường dựa trên nội dung sẽ khắc phục được nhược điểm của phương pháp đo lường dựa trên ngữ nghĩa bằng cách trích xuất từ tin tức những từ có giá trị TF-IDF cao nhất và những từ này được đặc trưng cho ngữ cảnh chính được đề cập trong tin tức.

Kết quả thực nghiệm cho thấy, sự kết hợp giữa hai phương pháp tương đồng giúp phát huy hiệu quả của cả hai phương pháp và khắc phục nhược điểm của nhau, cuối cùng làm tăng khuyến nghị tốt hơn. Tuy nhiên, phương pháp đề xuất vẫn còn một số hạn chế như sự phụ thuộc của nó vào tính đầy đủ của cơ sở tri thức và bản thể học. Việc xác định các trọng số sao cho việc kết hợp các biện pháp đạt được hiệu quả cao nhất cũng là một bài toán nan giải của phương pháp.

3.2. Kết luận về khả năng ứng dụng

Chung quy lại, qua các nghiên cứu thực tế trên đã nêu ra rằng, đề xuất theo ngữ nghĩa có tính ứng dụng quan trọng trong việc tạo ra các đề xuất cho môi trường tương tác ảo, giúp dễ dàng đề xuất các vấn đề mà người dùng mong muốn hay quan tâm. Giúp quản lý dễ dàng truy xuất các dữ liệu có liên quan nhau để gợi ý cho người dùng khác qua đó tạo ra nhiều đề xuất.

CHƯƠNG 4. KẾT LUẬN

Tóm lại, các hệ thống khuyến nghị ngữ nghĩa đang nhanh chóng trở thành một công cụ thiết yếu để tổ chức và hiểu ý nghĩa của dữ liệu phức tạp. Bằng cách sử dụng biểu đồ tri thức và các phương pháp khác, các hệ thống này cung cấp những hiểu biết sâu sắc và mối quan hệ có giá trị trong dữ liệu, đặc biệt là đi đầu trong lĩnh vực trí tuệ nhân tạo.

Một trong những lợi ích đáng kể nhất của các hệ thống đề xuất ngữ nghĩa là khả năng đề xuất thông tin mới trên nhiều nguồn khác nhau một cách hoàn hảo. Bằng cách tận dụng các công nghệ đề xuất bằng ngữ nghĩa, các hệ thống này có thể xác định các mẫu và mối quan hệ giữa các bộ dữ liệu có thể không rõ ràng khi sử dụng các công cụ phân tích dữ liệu hoặc tìm kiếm truyền thống. Khả năng tích hợp này dẫn đến những hiểu biết toàn diện hơn và hiểu rõ hơn về dữ liệu phức tạp, cho phép đưa ra kết quả đề xuất nhanh hơn và sáng suốt hơn.

Mặc dù có những ưu điểm của các hệ thống khuyến nghị ngữ nghĩa, nhưng vẫn còn một số thách thức liên quan đến việc sử dụng chúng. Một trong những thách thức quan trọng nhất là sự phức tạp của công nghệ liên quan đến việc xây dựng và quản lý các hệ thống này. Tập thể nhóm vẫn không có nhiều kinh nghiệm trong lĩnh vực này nên sẽ gặp khó khăn trong việc thiết lập và duy trì thành công hệ thống.

Một thách thức khác là thiếu tiêu chuẩn hóa trên cơ sở dữ liệu đồ thị tri thức. Là một công nghệ tương đối mới, vẫn chưa có cách tiêu chuẩn để xây dựng hoặc truy vấn các cơ sở dữ liệu này. Việc thiếu tiêu chuẩn hóa này có thể gây khó khăn cho việc chia sẻ dữ liệu giữa các cơ sở dữ liệu đồ thị tri thức khác nhau hoặc tích hợp chúng với các hệ thống khác.

Cuối cùng, tùy thuộc vào kích thước và độ phức tạp của dữ liệu được lưu trữ, cơ sở dữ liệu đồ thị tri thức có thể gặp sự cố về hiệu suất. Điều này có thể đặc biệt đúng nếu dữ liệu liên tục thay đổi hoặc nếu có nhiều mối quan hệ giữa các thực thể khác nhau.

Bất chấp những thách thức này, lợi ích của các hệ thống khuyến nghị ngữ nghĩa là rất đáng để nghiên cứu sâu rộng. Khi các lĩnh vực liên quan đến ngữ nghĩa tiếp tục phát triển, chúng ta có thể kỳ vọng sẽ thấy các hệ thống gợi ý tinh vi và mạnh mẽ hơn xuất hiện trong tương lai.

TÀI LIỆU THAM KHẢO

0. Fatih Gedikli, Dietmar Jannach: Recommender Systems, Semantic-Based -
Department of Computer Science, TU Dortmund, Dortmund, Germany
1. Brahim DIB, Fahd KALLOUBI, El Habib NFAOUI, Abdelhak BOULAALAM:
Semantic-based Followee Recommendations on Twitter Network - The First International
Conference On Intelligent Computing in Data Sciences
2. W. J, -P. L. E, L. J and H. Q, «TwitterRank: finding topic-sensitive influential
twitterers,» Proceedings of the 3rd ACM International Conference on Web Search and
Data Mining (WSDM'10), New York, NY, USA, p. 261–270, 2010.
3. Y. Y, T. T and K. H, «Twitter user ranking based on user-tweet graph analysis,»
Web Information Systems Engineering, Lecture Notes in Computer Science, vol. 6488, p.
240–253, 2010.
4. N. Roberto, «Word sense disambiguation: A survey,» ACM Computing Surveys
(CSUR), 2009
5. Quang-Minh Nguyen, Thanh-Tam Nguyen, Tuan-Dung Cao: Semantic-based
Recommendation Method For Sport News Aggregation System, Hanoi (2017)
6. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval.
Information Processing and Management: an International Journal archive. Volume 24
Issue 5, 1988, pp. 513-523.
7. Frasincar, F., IJntema, W., Goossen, F., Hogenboom, F.: Ontology-based news
recommendation. Proceedings of the 2010 EDBT/ICDT Workshops, Lausanne,
Switzerland, March 22-26, (2010)
8. Aleman-Meza, B., Halaschek, C., Arpinar, I.B., Sheth, A.: Context-Aware
Semantic Association Ranking. In Proceedings of the Semantic Web and Database
Workshop, Berlin, pp. 33-50

9. Viblo.asia - Xử lý ngôn ngữ tự nhiên - <https://viblo.asia/p/xu-ly-ngon-ngu-tu-nhien-phan-1-OeVKB8eQlkW>