BỘ GIÁO DỤC VÀ ĐÀO TẠO

TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH

KỊCH BẢN

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

KỊCH BẢN

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

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

[Slide 4]

1.2. Mở đầu

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.

Slide 5

1.3. Road-map báo cáo

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

Slide 6

Ta cùng nhau đi sâu vào kỹ thuật đề xuất dựa trên ngữ nghĩa -Semantics based recommendation system

Slide 7

Phần phân tích về kỹ thuật này gồm có các phần chuyên biệt như sau: Đầu tiên sẽ bóc tách về định nghĩa, sau đó đến phân tích tính chất, đến ưu và nhược điểm của kỹ thuật

Slide 8

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

2.1.1. Semantics Recommendation System là gì?

Về khả năng và khái quát, 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.

Về kỹ thuật, 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.

Về kết quả, 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.

Slide 9

2.1.2. Phân tích hệ đề xuất thông tin dựa trên ngữ nghĩa

Ta cùng nhau phân tích về các yếu tố quan trọng hình thành nên một hệ đề xuất dựa trên ngữ nghĩa và cách thức hoạt động của kỹ thuật này

Slide 10

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 tong 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.

Slide 11

2.1.3. Nguyên lý hoạt động

Ta có các bước cơ bản của một hệ 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.

Slide 12

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.

Slide 13

- *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.

Slide 14

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 mối 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 lí 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.

Slide 15

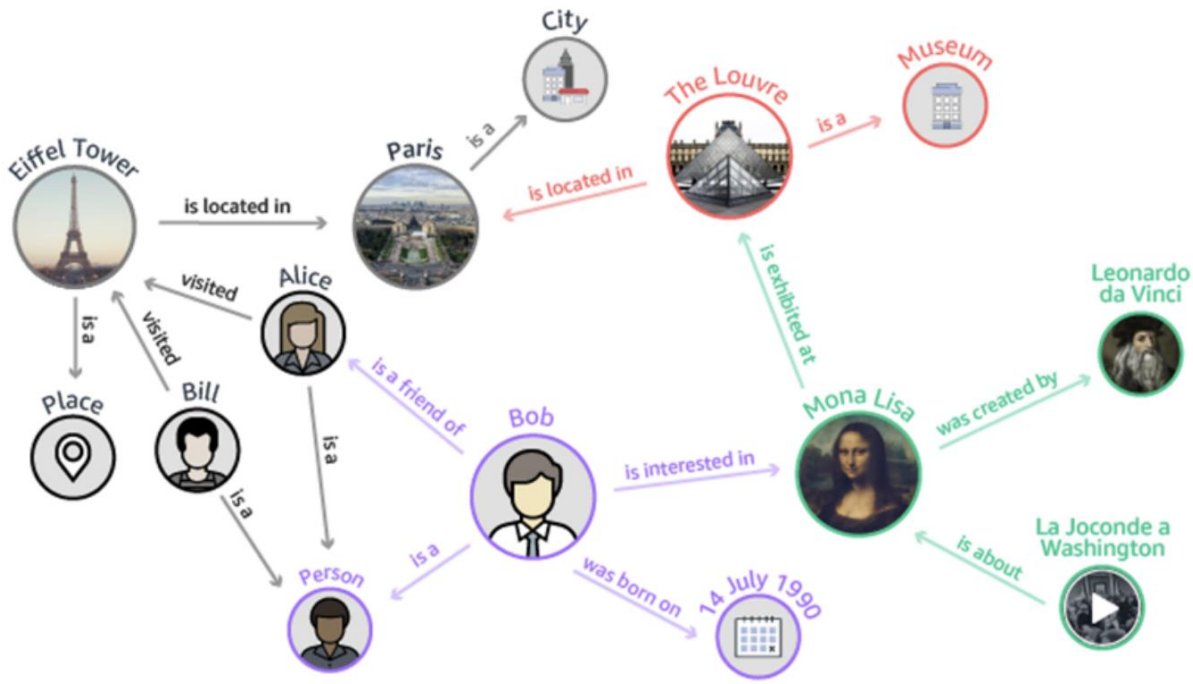
2.2. Đồ thị tri thức – Knowledge Graph

2.2.1. Đồ thị tri thức là gì

Slide 16

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.

Ta có hình minh họa sau



Hình 1. Một mô hình đồ thị tri thức

Slide 17

2.2.2. Cơ sở dữ liệu dạng đồ thị

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.

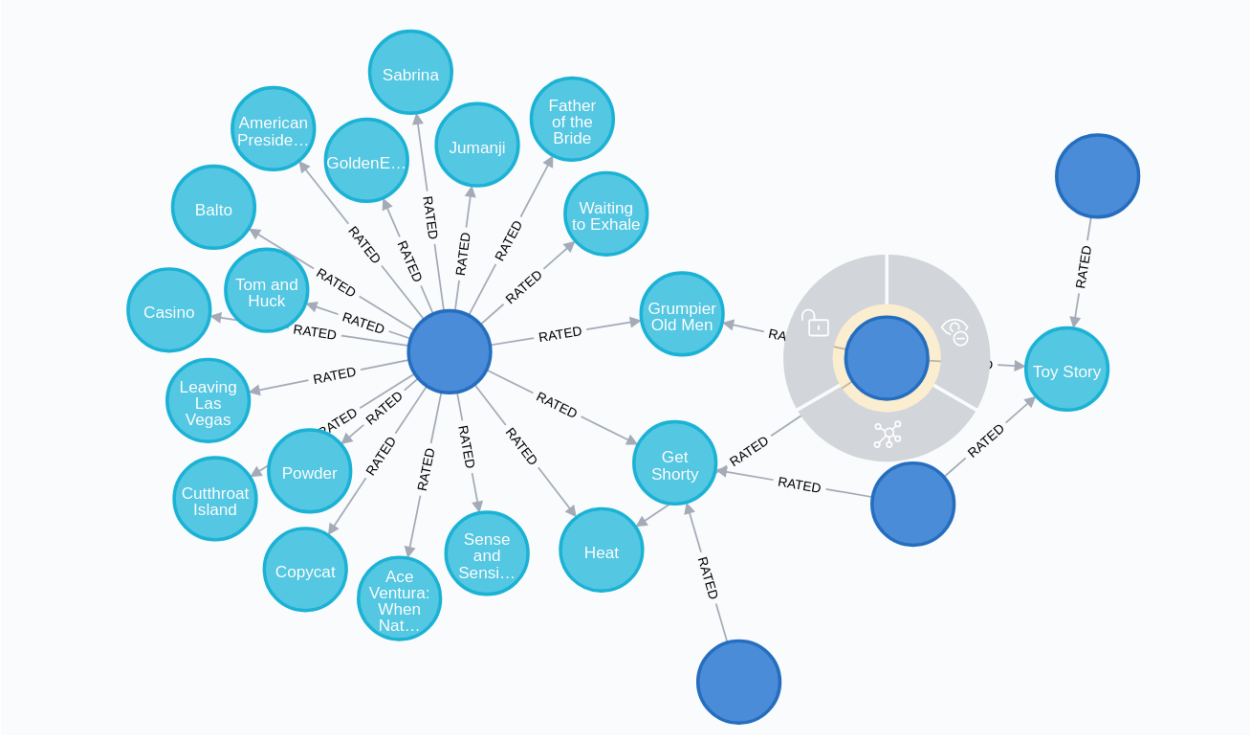
Slide 18

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.

Slide 19

Một ví dụ minh họa cho một chuỗi dữ liệu dạng đồ thị tri thức, có dạng như hình ảnh đang hiển thị trên màn hình



Slide 20

Ta đi đến 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.

Slide 21

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

Slide 22

Ta phân tích về tổng quan về Neo4j và cấu trúc của khái niệm này.

Neo4j là hệ cơ sở dữ liệu đồ thị. Và tất nhiên, 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.

Slide 23

Cụ thể hơn, *Neo4j* 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 JSON[1] và được gắn label để phân biệt loại node phục vụ các thuật toán, các truy vấn

[1] 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"

- Relationships: Là các cạnh trong cơ sở dữ liệu đồ thị, thể hiện mối quan hệ giữa các node, mối 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.

Slide 24

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.

Slide 25 + 26

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

Slide 27

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 các lựa chọn và dự đoán mức độ hữu ích tiềm năng của tin tức đối với mỗi người đọc. Nhóm sẽ cố gắng rút gọn vào những ý chính, nhằm làm cho bài báo cáo này không quá nhàm chán, nhưng cũng đảm bảo đủ lượng kiến thức có thể chia sẻ.

Đầu tiên là về 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.

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.

Slide 28

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) 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:

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

với

$$TF_{ij} = \frac{n_{ij}}{\sum_k n_{kj}}$$

$$IDF_i = log \frac{|D|}{|\{d : t_i \in d\}|}$$

Hình 2. 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 của đề xuất dựa trên ngữ nghĩa

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. Và trong bài nghiên cứu này mong muốn kết hợp hai cách tiếp cận này với mong muốn khắc phục những hạn chế của từng cách tiếp cận, giúp kết quả đề xuất trở nên hiệu quả hơn.

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 các bản thể học[2] và cơ sở tri thức đã được xây dựng.

Slide 29

- Nhóm sẽ không đề cập thêm về phần lý thuyết xây dựng thực nghiệm vì nó mang tính chất trừu tượng quá nhiều, gây khó hiểu, mà bản ghi đầy đủ đã có trong file word báo cáo hoàn chỉnh của nhóm nên mong mọi người có thể tìm đọc để hiểu sâu thêm về vấn đề. Còn bây giờ thì sau khi thực hiện các bước thực nghiệm chuyên biệt, ta có đánh giá về kết quả nghiên cứu như sau

Đánh giá thực nghiệm

[2] 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)

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

Slide 30

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 trên đã 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 mẩu 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.

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

Slide 31

CHƯƠNG 4. KẾT LUẬN

Slide 32

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 kể và 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

Slide 33

Nhóm có tham khảo các nguồn sau

Ngoài ra còn tham khảo các nguồn nhỏ lẻ từ các trang tin như:

https://chat.openai.com/ https://www.researchgate.net/ https://ieeexplore.ieee.org/ https://www.youtube.com/ và hơn thế nữa.

Slide 34

Xin chân thành cảm ơn, dù vẫn còn nhiều thiếu sót nhưng nhóm vẫn cố gắng bằng hết khả năng của mình, đã hoàn thành bài báo cáo này.