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; ngoài ra còn phân tích tính ứng dụng của cơ sở dữ liệu đồ thị tri thức Neo4j để xử lý dữ liệu đề xuất.

Tập thể nhóm Neko trên tinh thần cố gắng hết sức nhằm xây dựng thành công một bài báo cáo mang tính hiệu quả trong chia sẻ 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 Database để biểu diễn dữ liệu.

[Slide 6]

Ta thường thấy các công cụ trực tuyến có liên quan nhiều với phân tích ngôn ngữ như công cụ tìm kiếm nổi tiếng Google Search, Microsoft Bing, Baidu hay một số công cụ đề xuất phim của Netflix hay đề xuất bản tin của mạng xã hội Facebook thường phân tích mối tương quan về mặt ngôn ngữ, từ vựng để lọc ra những đề xuất phù hợp nhất cho người dùng của mình. Các công cụ này – phần nào đó – đã ứng dụng kỹ thuật đề xuất dựa trên ngữ nghĩa - Semantics based recommendation system – để cải thiện cho bộ lọc của họ. Nhóm sẽ phân tích và chỉ rõ tính chất của kỹ thuật này thông qua một ứng dụng đã được công bố báo cáo – về một hệ thống đề xuất tin tức - và phần phân tích ứng dụng sẽ được trình bày ở những phần sau của bài thuyết minh – mong thầy và mọi người theo dõi.

[Slide 7]

Thì đầu tiên, phần phân tích về kỹ thuật Semantics-based 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 có khả năng tư vấn và đề xuất các sản phẩm, nội dung liên quan đến nhu cầu của người dùng thông qua các kỹ thuật hiểu biết và xử lý ngôn ngữ tự nhiên. Đây là một ứng dụng quan trọng của Trí tuệ nhân tạo (AI).

Về kỹ thuật, Kỹ thuật đề xuất này sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên và khai thác các mối quan hệ ngữ nghĩa giữa các nội dung để đưa ra nội dung có mức độ tương đồng với sự quan tâm của người dùng.

Về kết quả, các thuật toán máy học nhằm xử lý ngữ nghĩa từ chính thông tin đã thu thập sẽ được thực thi để đưa ra các đề xuất phù hợp nhất với nhu cầu 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 ngữ nghĩa của câu từ được cung cấp. Nó xác định ý định và hàm ý của câu từ 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.
- *Xây dựng mô hình ngữ nghĩa*: Hệ thống sẽ 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 đồng ngữ nghĩa*: hệ thống phải đo và xác định được độ tương đồng về mặt ngữ nghĩa giữa các đối tượng
- *Định nghĩa và biểu diễn ngữ nghĩa*: Phải biểu diễn được ngữ nghĩa của các đối tượng, thuộc tính và quan hệ trong hệ thống đề xuất, từ đây có thể đánh giá được mức độ hiệu quả của một hệ thống đề xuất là như nào.
- *Đề xuất thông tin*: Dựa trên việc phân tích ngữ nghĩa và tri thức, và các điều kiện kể trên thì hệ thống sẽ đề xuất thông tin phù hợp cho người dùng.

Slide 11

Để có thể tạo nên gợi ý đề xuất, các hệ thống sử dụng kỹ thuật đề xuất dựa trên ngữ nghĩa 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 thông qua kỹ thuật *Rút trích thông tin* qua một ví dụ điển hình như sau:

(Phần phân tích này thì nhóm có tham khảo từ https://viblo.asia/p/xu-ly-ngon-ngu-tu-nhien-phan-1-OeVKB8eQlkW)

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 (sử dụng trí tuệ con người vốn có để dịch đoạn văn trên): "*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*."

Và tất nhiên ta sẽ hiểu theo kỹ thuật của học máy vừa được đề cập chứ không phải lả hiểu nôm na rằng ta dịch được như vậy thì nghĩa nó là như vậy.

Slide 12

- *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."*

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 khi có thể hiểu nghĩa của một câu đơn hơn là phải bắt buộc hiểu liền một lúc cả một đoạn văn.

Slide 13

- *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.

Slide 14

- *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

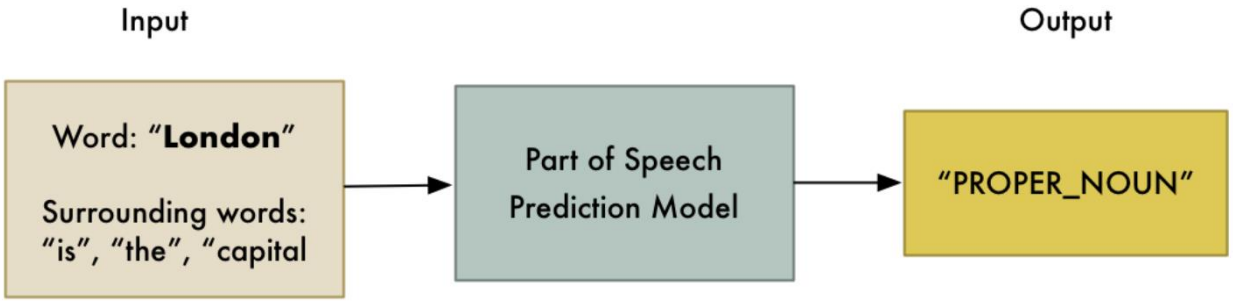
"*Kingdom*": danh từ riêng/tên riêng

".": dấu câu

Hình 1. Phân tách và xác định loại từ cho câu văn cần phân tích

Mô hình 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 2. Phân tích từ "London" trong câu

Như cái ảnh ở mé dưới màn hình – là cái ảnh có 3 cái khung input output gì đó , 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.

Slide 15

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

Việc bổ 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:

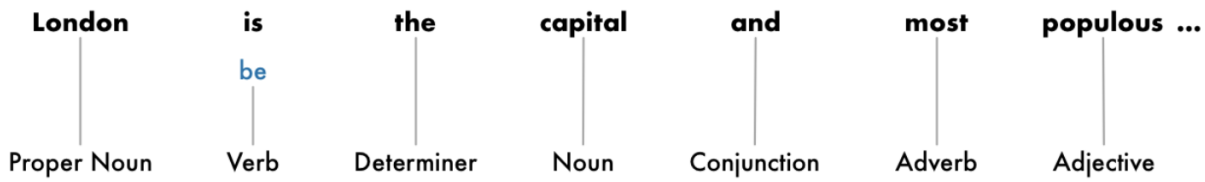
"I had two dogs"

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

"I [have] two [dog]"

Trong phân tích xử lý ngôn ngữ tự nhiên, việc bổ 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 bổ ngữ:



Hình 3. Quá trình bổ ngữ Lemmatization đã đổi động từ "is" trong câu thành động từ gốc "be"

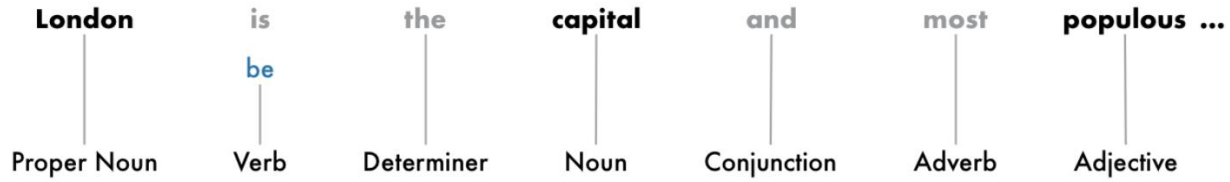
Slide 16

- *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 (những từ đã bị bay màu - bị chuyển sang màu xám mờ mờ trên màn hình):



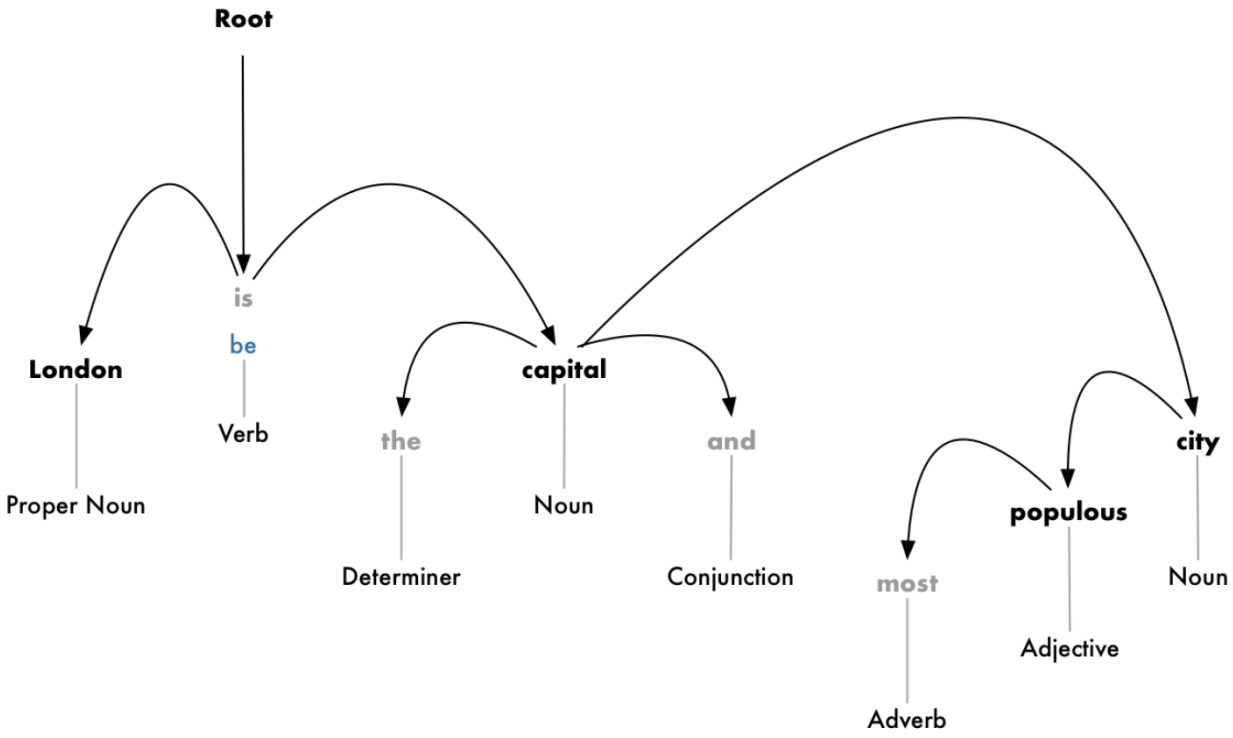
Hình 4. Minh họa sau khi vô hiệu các từ dừng (chữ màu xám)

Slide 17

- *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.

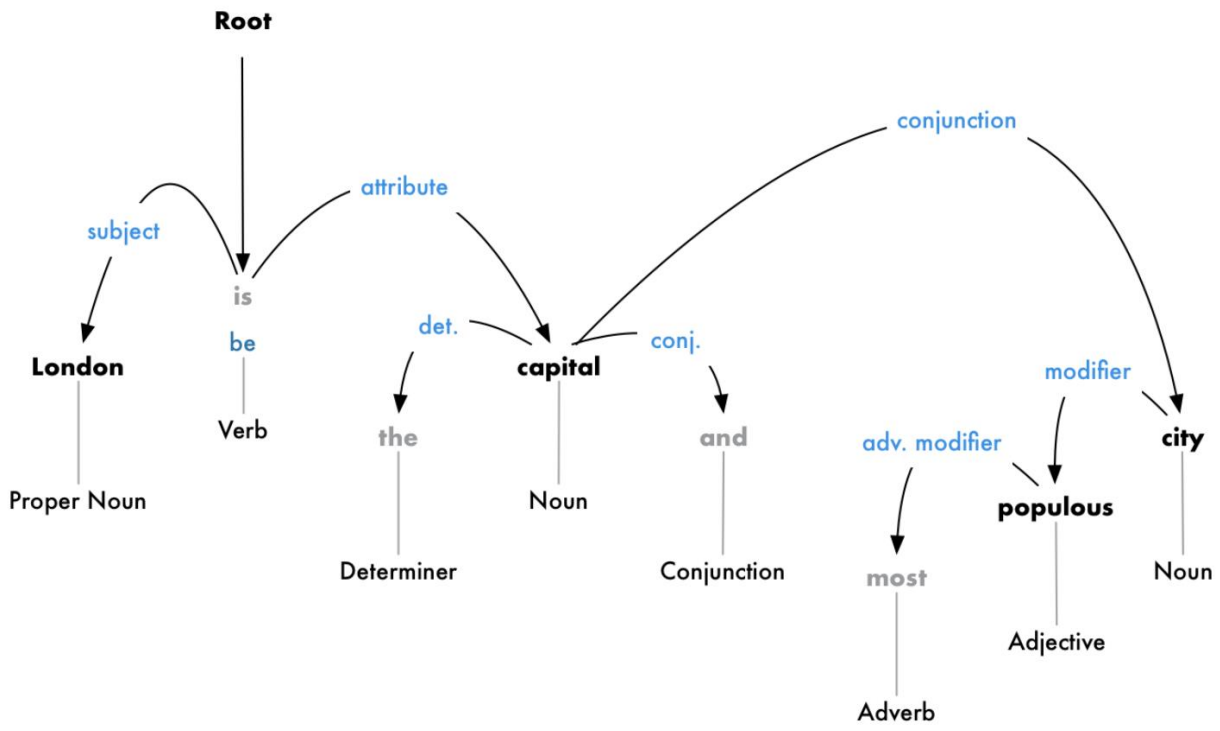
Mục tiêu là xây dựng một mô hình cây có thể gán một từ đơn duy nhất làm root cho mỗi từ trong câu. Từ root của cây này sẽ là động từ "*be*" ("*is*") trong câu. Đây là phần đầu của cây phân tích sẽ trông như thế nào cho câu của chúng ta:



Hình 5. Đặt động từ "be"/"is" làm root cho mô hình cây

Slide 18

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 6. 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.

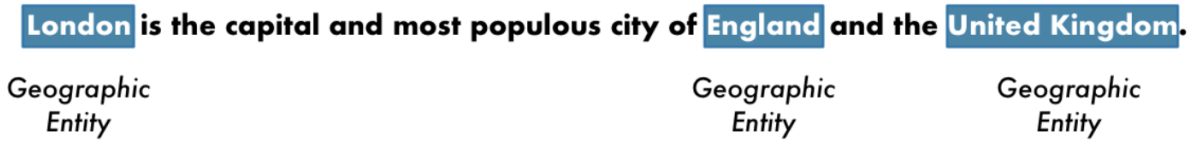
Điều quan trọng cần nhắc lại rằng, nhiều câu trong tiếng Anh là mơ hồ và thực sự khó phân tích. Trong những trường hợp đó, mô hình sẽ đưa ra dự đoán dựa trên phiên bản phân tích cú pháp của câu đó, và có lẽ một số trường hợp sẽ không hoàn hảo và đôi khi mô hình sẽ dự đoán sai. Nhưng theo thời gian, mô 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.

Slide 19

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

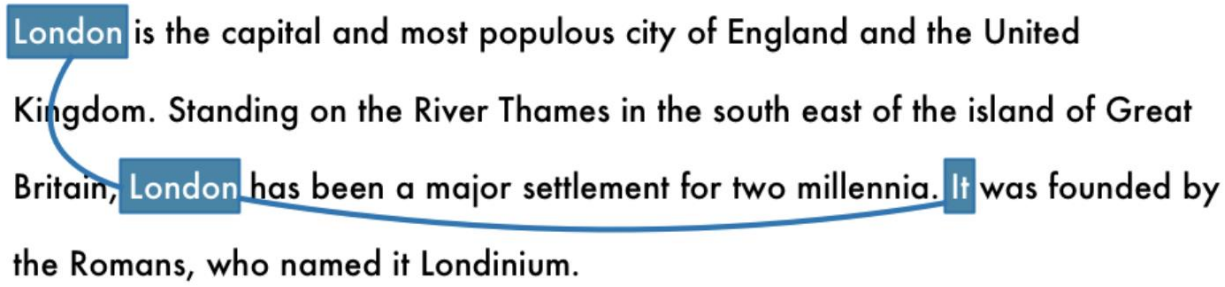
Sau khi đã hoàn thành các bước được coi là khó khăn nhất và cốt lõi nhất của bài toán phân tích, thì điều quan trọng là ta có thể vượt ra ngoài trường ngữ pháp thô và bắt đầu thực sự rút ra ý tưởng và hàn gắn ý nghĩa thực sự của câu văn phân tích.

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 7. 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 8. 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.

Slide 20

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.
- *Xử lý dữ liệu ban đầu*: Dữ liệu được thu thập và chuẩn hóa định dạng và loại bỏ dữ liệu không cần thiết hoặc sai sót.
- *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 được tạo ra.

Slide 21

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

Kỹ thuật tư vấn dựa trên ngữ nghĩa này có một số ưu điểm cải tiến hơn so với các kỹ thuật khuyến nghị truyền thống khác thay vì 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 dựa trên ngữ nghĩa có thể giúp xác định các mục không chỉ giống nhau mà còn bổ sung thêm những nội dung có liên quan mới lạ, có một chút liên quan với nhau, nhằm dẫn đến các kết quả đề xuất được 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 22

- *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 tốt hơn so với các hệ thống đề xuất truyền thống. vì thế trong tương lai nó 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.

Slide 23

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, sẽ là khó khăn cho hệ thống đề xuất nếu dữ liệu ban đầu cho không đầy đủ, không nhất quán hoặc có chất lượng thấp, vô giá trị.
- *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, khó hiểu.
- *Phạm vi hạn chế*: đôi khi hệ thống bị hạn chế khả năng ứng dụng trong các ngữ cảnh nhất định.
- *Mối quan tâm về quyền riêng tư*: có thể gây lo ngại về quyền riêng tư nếu dữ liệu không được xử lý phù hợp.
- *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 24

2.2. Đồ thị tri thức – Knowledge Graph

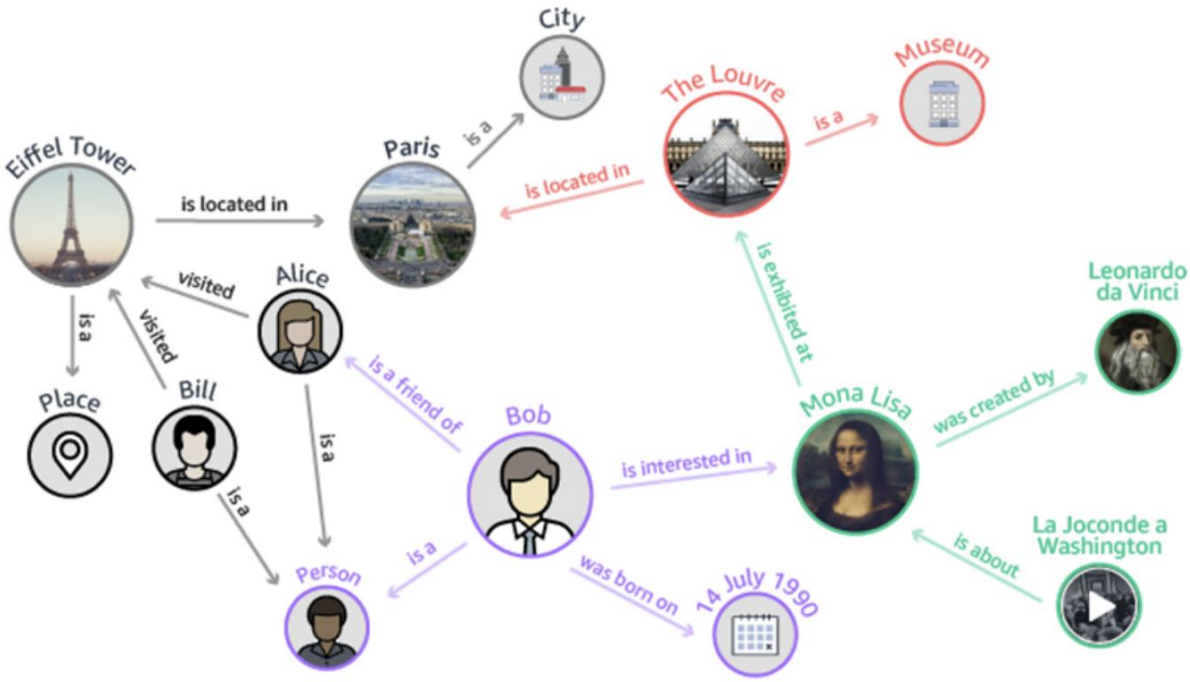
2.2.1. Đồ thị tri thức là gì

Slide 25

Đồ thị tri thức (knowledge graph) là một cấu trúc dữ liệu biểu diễn tri thức dưới dạng đồ thị. 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.

Ta có hình minh họa sau



Hình 9. Một mô hình đồ thị tri thức

Slide 26

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.

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.

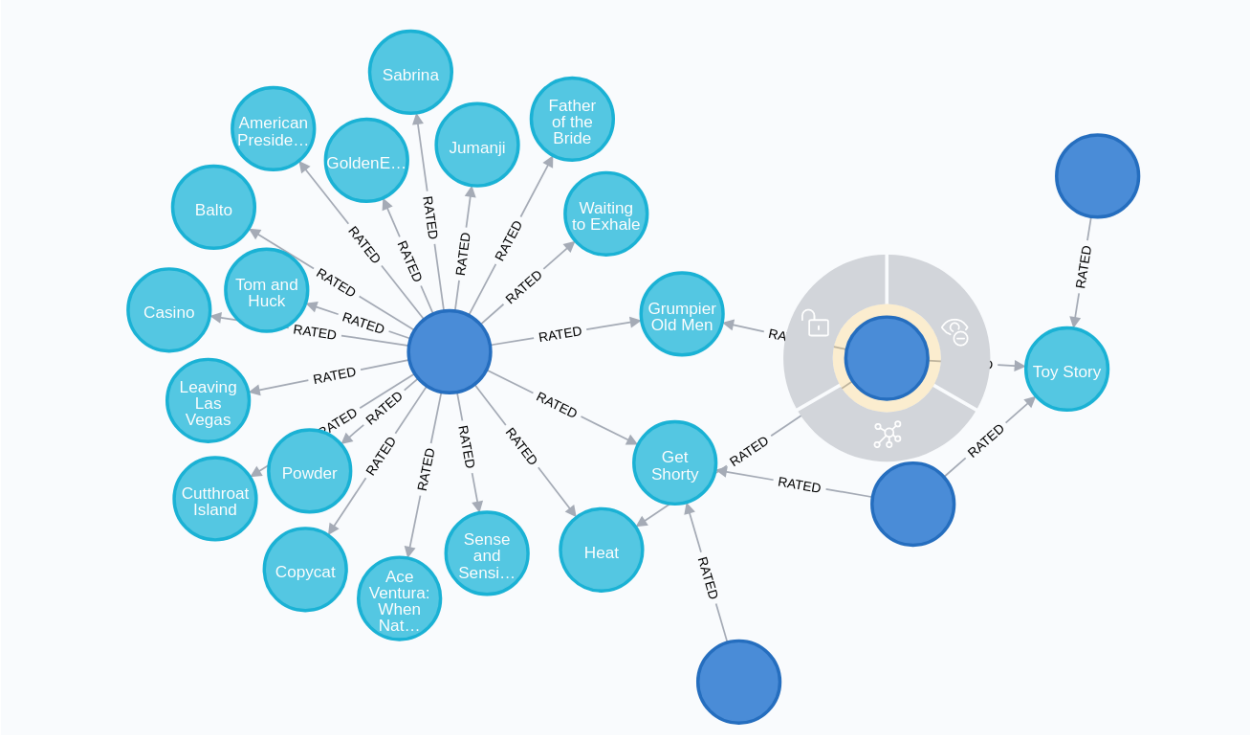
Slide 27

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

- *Mô hình hóa dữ liệu linh hoạt*: 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 đa chiều**, có nghĩa là chúng **có thể xử lý khối lượng dữ liệu lớn và có thể được phân phối tốt, rộng rãi** 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, 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 28

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 29

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

- *Cần định nghĩa một mô hình dữ liệu phù hợp*:
- *Cần phải thu thập và chuẩn hóa dữ liệu*: để đảm bảo tính nhất quán và hiệu quả khi tìm kiếm và truy xuất thông tin.
- *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*: như khi nãy đã phân tích.
- *Tích hợp các công nghệ khác nhau*:
- *Kiểm tra và đánh giá*: tính hiệu quả và độ chính xác

Slide 30

Và 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 31

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

Slide 32

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
- 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.

Thoát PPT mở Neo4j lên demo***

Slide 33

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*: cho phép phân tích và truy vấn dữ liệu **trực quan hơn so với cơ sở dữ liệu quan hệ truyền thống**.
- Thứ hai, *Truy vấn ngữ nghĩa*: **có thể truy vấn dữ liệu dựa trên mối quan hệ về mặt ngữ nghĩa của những từ ngữ, thay vì chỉ xét trên chính các thực thể đó**, **Cho phép thực hiện các truy vấn phức tạp hơn về ngữ cảnh và ý nghĩa** của dữ liệu.

[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"

- Thứ ba, *Khả năng mở rộng*: **Cũng giống như khả năng mở rộng của các hệ cơ sở dữ liệu đồ thị nói chung.**
- Thứ tư, *Khả năng tương tác*: Cơ sở dữ liệu sơ đồ tri thức **hỗ trợ khả năng tương tác và được tích hợp với các kỹ thuật khác, nền tảng khác và hệ thống khác, máy chủ khác** để **tạo ra** một hệ sinh thái dữ liệu **toàn diện hơn**.
- Thứ năm, *Tích hợp dữ liệu*: **dựa trên điều thứ tư vừa kể**, cho phép xác định các mối quan hệ **một cách cặn kẽ hơn** trong các bộ dữ liệu **riêng lẻ**.
- Và cuối cùng là, *Thông tin chi tiết tốt hơn*:

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 *Neo4j* mang lại**.

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

- Thứ nhất, *Quá phức tạp*: **gây khó khăn khi sử dụng và vận hành đối với người dùng không quen thuộc**
- Thứ hai, *Đôi khi hiệu suất truy vấn không đảm bảo cho những hệ thống yêu cầu tốc độ truy xuất dữ liệu phải thật nhanh*:
- 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 tích hợp với các 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 truyền thống.

Slide 34 + 35

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]

Dựa trên bài báo nghiên cứu khoa học đã được xuất bản từ năm 2017 – đặc biệt những người đã cùng nhau nghiên cứu đề tài này là 3 người Việt Nam, đó là Nguyễn Quang Minh, Nguyễn Thành Tâm và Cao Tuấn Dũng.

Thoát PPT để truy cập link bài báo (*https://shs.hal.science/hal-01630538*)

Slide 36

Các trang web tin tức có nhiệm vụ là chúng phải 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 là một vấn đề nghiêm trọng của các hệ thống tổng hợp tin tức là số lượng lớn tin tức được xuất bản hàng ngày, hàng giờ theo nhịp của mọi hoạt động của thế giới, điều này sẽ 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 xây dựng nên 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, không tốn nhiều thời gian cũng như không quá đè nặng về số liệu lý thuyết vì mọi thứ đã được phân tích cặn kẽ trong file word báo cáo của nhóm, nhưng sẽ đảm bảo đủ lượng kiến thức được chia sẻ mà nhóm mang đến cho mọi người.

Đầu tiên là về Cách tiếp cận

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, **cụ thể kết quả thực nghiệm sẽ được phơi bày ở những phần tiếp theo của bài thuyết minh này**.

Slide 37

Thông thường, nhiều hệ đề xuất dựa trên nội dung sử dụng các phương pháp trích xuất thuật ngữ như TF-IDF (Tần suất xuất hiện từ ngữ - tần suất tài liệu nghịch đảo - Term Frequency-Inverse Document Frequency) đây là một phép truy hồi thông tin dựa trên thống kê số học nhằm phản ánh mức độ quan trọng, tần suất xuất hiện và các giá trị khác của những từ ngữ trong câu văn – đoạn văn – và trong văn bản, kết hợp với phép đo độ tương tự cosine để so sánh độ tương tự giữa hai bản tin.

Cụ thể hơn thì 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 đồng giữa hai bản tin tức. Giá trị TF-IDF của từ xuất hiện trong bản tin được tính theo công thức sau: (được hiển thị trên màn hình)

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

với

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

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

Hình 10. 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 38

- Như đã nói, thì nhóm sẽ không đề cập thêm về phần lý thuyết xây dựng nên thực nghiệm vì nó mang tính chất trừu tượng lý thuyết quá nhiều, gây khó hiểu, làm cho bài thuyết minh này bị khô khan mà bản ghi đầy đủ nghiên cứu 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 39

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 40

CHƯƠNG 4. KẾT LUẬN

Slide 41

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.

Cuối cùng, bất chấp những thách thức này, lợi ích của các hệ thống khuyến nghị ngữ nghĩa là rất đáng kể và rất có tiềm năng. 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 42

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 43 - end

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.

Thoát ra, mở youtube lên search ký tự lỗi trong bộ lọc đề xuất

À quên mất, trước khi thật sự kết thúc bài thuyết minh để nhường lại cho các nhóm khác, thì nhóm chúng em có phát hiện ra một sự thật nó khá là thú vị về nền tảng mạng xã hội chia sẻ video mà chúng ta thường hay biết đến, đó là Youtube.

Thì nhóm chúng em không biết có nên gọi đây là một lỗ hổng trong công cụ đề xuất của Youtube hay không, nhưng khi mọi người cố tình search ký tự đặc biệt này, nó giống như dấu ký tự rỗng trong toán học, thì Youtube để hiển thị những kết quả là những video khá quái dị, em sẽ không mở những video này lên vì nó không phù hợp và khá là phản cảm.

Tuy nhiên, theo chúng em có tìm hiểu thì Youtube cũng ứng dụng kỹ thuật đề xuất dựa trên ngữ nghĩa để hiển thị nội dung tìm kiếm, và những ký tự như ký tự rỗng này nó không phải là một từ ngữ, một chữ cái hay một định nghĩa khái niệm gì cả, và chủ nhân của những video quái dị này đã sử dụng những ký tự đặc biệt để vượt qua bộ lọc kiểm duyệt nội dung của Youtube, và Youtube thì luôn kiểm soát rất gắt gao về nội

dung được đăng tải lên nền tảng của họ. Và đây là một lỗ hổng khá hi hữu mà Youtube vẫn đang cố gắng sửa trong tương lai gần, và may mắn là những video này, do không có một khái niệm cụ thể nào trên phần tiêu đề, chi tiết video, hashtag của chúng, vì chỉ là những ký tự vô tri nên chúng cũng nằm ngoài thuật toán đề xuất nội dung trên trang chủ của Youtube. Đó là lý do vì sao mà trên trang Youtube.com ta sẽ không bao giờ thấy được một cách ngẫu nhiên những video như này.

Kết

Nhóm xin kết thúc phần trình bày của mình, cảm ơn thầy và mọi người đã theo dõi và nhóm sẽ tiếp nhận câu hỏi thắc mắc từ phía mọi người.