Tìm kiếm bằng sáng chế dựa trên tương đồng ngữ nghĩa

Dương Hớn Minh

Khoa Dược - Trường Đại học Nguyễn Tất Thành
dhminh@ntt.edu.vn

Tóm tắt

Tìm kiếm dựa trên từ khóa là một công cụ tìm kiếm phổ biến, cổ điển và còn nhiều hạn chế. Trong khi đó, tìm kiếm bằng ngữ nghĩa có thể hiểu được chủ đề hay ý nghĩa cụ thể của từng đoạn và câu văn. Hai kĩ thuật chính trong tìm kiếm ngữ nghĩa gồm: tìm kiếm vectơ (vector search) và xử lí ngôn ngữ tự nhiên (Natural Language Processing) cho tài liệu bằng sáng chế tiếng Anh. Nguồn dữ liệu được lấy từ trang web USPTO thuộc về chính phủ Mĩ. Điểm mới của nghiên cứu là tìm được những tài liệu gần nghĩa với tài liệu cho trước, tốc độ tìm kiếm nhanh và chính xác hơn. Kết quả đạt được ban đầu của nghiên cứu tỏ ra hiệu quả so với các phương pháp tìm kiếm đồng nghĩa khác thể hiện ở tốc độ tìm kiếm chỉ tốn 0,3775 giây để tìm ra 10 bằng sáng chế có độ tương đồng cao nhất trong kho dữ liệu gồm 694 bằng sáng chế. Nghiên cứu này đã đưa ra phương pháp tìm kiếm mới để giải quyết vấn đề tìm kiếm bằng sáng chế tương đồng vì tránh đăng kí trùng ý tưởng của tác giả và bảo hộ quyền sở hữu và quyền thương mại.

Nhận 10/02/2023
Được duyệt 10/08/2023
Công bố 01/11/2023

Từ khóa
Tìm kiếm vector,
tìm kiếm bằng sáng chế,
tìm kiếm ngữ nghĩa,
học sâu, chuyển đổi câu

1 Giới thiệu

Tìm kiếm ngữ nghĩa từ lâu đã trở thành một thành phần quan trọng trong hệ thống công nghệ của những tập đoàn công nghệ toàn cầu như Google, Microsoft… cũng như các dịch vụ patent search như Amazon và Netflix [1]. Sự hỗ trợ từ phần cứng và thuật toán mới gần đây của những công nghệ này đã đem đến những ý tưởng mới trong lĩnh vực tìm kiếm. Những công nghệ đã từng ít phổ biến này đang được các tổ chức trong mọi ngành có thể nhanh chóng triển khai và áp dụng vào thực tế để đem lại hiệu quả.

Trong bốn năm gần đây, có sự bùng nổ quan tâm đến tìm kiếm ngữ nghĩa [2]. Những phương pháp mới làm cho việc xây dựng ứng dụng sản phẩm trở nên dễ dàng hơn. Chưa kể, phạm vi ứng dụng trong thực tế vẫn còn rất rộng, chẳng hạn đề xuất các bộ phim trong cùng chủ đề, tương tự nội dung, cho đến gợi ý những tài liệu có sự tương đồng cao về chủ đề nghiên cứu, các công cụ tìm kiếm, tự động sửa, dịch, công cụ đề xuất, ghi lỗi, và phần mềm phát hiện đạo văn. Nhiều công cụ có thể đư c hưởng lợi từ chức năng phân cụm hoặc tìm kiếm tương đồng càng thúc đẩy nhu cầu của tìm kiếm ngữ nghĩa.

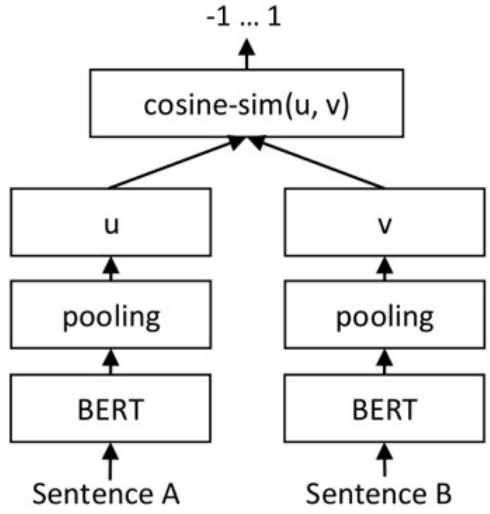
Bên cạnh đó, những mô hình học sâu (Deep Learning) trong việc xử lí ngôn ngữ tự nhiên ngày càng thể hiện rõ hiệu quả vượt trội so với những phương pháp xử lí ngôn ngữ tự nhiên truyền thống. Trước 2017, mạng nơ-ron tái diễn (Recurrent Neural Network-RNN) – với khả năng hiểu ngôn ngữ vẫn còn bị hạn chế và nhiều điểm cần khắc phục.

Kể từ khi giới thiệu mô hình transformer (biến đổi) đầu tiên trong bài báo năm 2017 [2], xử lí ngôn ngữ tự nhiên đã chuyển từ RNN sang các mô hình như BERT (Bidirectional Encoder Representations from Transformers) và GPT (Generative Pre-trained Transformer). Các mô hình mới này có thể trả lời câu hỏi, viết bài nghiên cứu (hiện nay, GPT-3 đã có thể tạo được tài liệu học thuật chuyên sâu từ những từ khóa gợi ý), cho phép tìm kiếm ngữ nghĩa trở nên hiệu quả hơn. Trước khi tìm hiểu sâu vào mô hình chuyển đổi câu (Sentence Transformer) [3, 4], có thể tìm hiểu lí do tại sao việc kết hợp mô hình Transformer lại hiệu quả hơn

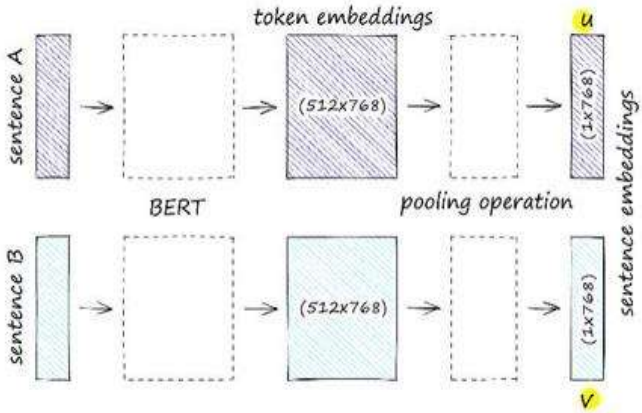
nhiều - và sự khác biệt nằm ở đâu. Mô hình Transformer là bản nâng cấp của mô hình RNN trước đó.

Trong lĩnh vực dịch máy (machine translation), dùng các mạng mã hóa và giải mã, Encoder-Decoder. Mô hình đầu tiên để mã hóa ngôn ngữ gốc thành vectơ ngữ nghĩa và mô hình thứ hai để giải mã điều này vào ngôn ngữ đích. Tuy nhiên, sẽg tạo ra một lượng lớn thông tin qua nhiều bước theo thời gian và cố gắng thu thập tất cả thông qua một kết nối duy nhất. Từ đó phát sinh vấn đề tắc nghẽn thông tin giữa hai mô hình [5].

Để khắc phục những nhược điểm trên, đã có nghiên cứu đề xuất 3 cải tiến nổi bật: Mã hóa vị trí, Tự tập trung, Tập trung đa đầu vào [2]. Tiếp theo đó, mô hình BERT [3] đến từ Google AI nhằm nâng cấp hơn nữa hiệu suất của mô hình transformer bằng cách biểu diễn mã hóa hai chiều được mô tả ở Hình 1. Từ đó sẽ tìm được độ tương đồng giữa hai câu hiệu quả hơn.



Hình 1 Phương pháp tính độ tương đồng của hai câu khi dùng BERT [3]



Hình 2 Sơ đồ mã hóa khi huấn luyện với các cặp câu đầu vào [4].

Nhược điểm của phương pháp này là tốn nhiều thời gian tính toán [3]. Năm 2019, tác giả Iryna và Nils đã đề xuất mô hình Sentence BERT [4] và thư viện sentence transformer. Mô hình này cho phép câu có thể biểu diễn dưới dạng vector, nhờ đó, việc so sánh độ tương đồng bằng phép tính cosine similarity vì tốc độ tính toán nhanh hơn được biểu diễn ở Hình 2.

Bên cạnh đó, phương pháp Get-In Patent Search [6, 7], tác giả đề xuất ý tưởng tách rời các câu, sau đó tách các cụm danh từ trong câu, và so sánh những cụm từ đó với nhau. Chi tiết hơn, khi kiểm tra sự tương đồng giữa 2 cụm danh từ với nhau. Việc đầu tiên là tách cụm từ lớn thành các từ đơn lẻ, rồi tái tạo thành các cụm từ con. Sau đó, sẽ dùng thư viện WordNet [8] để so sánh tất cả bộ đôi cụm từ nhỏ bắt nguồn từ hai cụm ban đầu và lấy kết quả tối ưu nhất. Do phải phân rã, tổ hợp và ghép cặp cũng như tính điểm rất nhiều cụm từ nhỏ với nhau nên dù kết quả khá sát nhưng tốc độ xử lí còn khá chậm. Ngoài ra, còn có một số phương pháp tìm kiếm tương tự dựa vào sự tập trung và phi tập trung trong nghiên cứu [9, 10].

2 Phương pháp nghiên cứu

Với những nghiên cứu ở trên đã chỉ ra nhược điểm của phương pháp cũ trong tìm kiếm và mã hóa ngữ nghĩa. Từ đây, phương pháp trong nghiên cứu này sẽ cải thiện ở những điểm sau:

- Tìm kiếm ngữ nghĩa dựa trên những phần chính của tài liệu thay vì chỉ rút trích thông tin ở phần tóm tắt, như vậy sẽ đảm bảo tính đầy đủ của thông tin lưu trong bằng sáng chế.
- Mã hóa tất cả câu trong bằng sáng chế bằng vector để truy vấn đồng nghĩa được nhanh chóng. Giải thích cho việc này, khi so sánh hai câu dạng văn bản thì tốc độ xử lí rất chậm, trong khi đó, việc so sánh khoảng cách giữa hai vectơ số sẽ nhanh chóng và chính xác hơn nhiều.
- Mô hình học sâu tân tiến nhất được dùng để mã hóa câu, cụm từ thành vectơ ngữ nghĩa. Mô hình học sâu đã được huấn luyện (pretrain model) trên tập dữ liệu bằng sáng chế tương tự nên đảm bảo việc mã hóa ngữ nghĩa của câu sát nghĩa nhất.
- Bộ dữ liệu dùng trong nghiên cứu này là tập dữ liệu công khai, gồm các văn bản bảo hộ bằng sáng chế trong một nhóm bằng phát minh do chính phủ Mĩ công bố tại trang web USPTO [11-13].

Công trình nghiên cứu gồm hai phần chính:

- Chuẩn bị và tiền xử lí cơ sở dữ liệu để chuẩn bị cho việc truy vấn. Tại bước này, tất cả bằng sáng chế sẽ theo quy trình như sau:

• Bước 1: tiền xử lí văn bản, để trích xuất phần Title, Abstract, Claim vì ba phần này chiếm trọn ngữ nghĩa chính trong bằng sáng chế. Những phần như owner, description, related patents chỉ nêu chi tiết quy trình, thông tin cụ thể của vật liệu hay những bằng sáng chế liên quan chứ không nắm giữ ý nghĩa chính.

• Bước 2: sau khi đã tách đư c những đoạn trên, ta sẽ tiến hành tách đoạn thành từng câu đơn (sentence tokenize). Đây là phần chuẩn bị quan trọng cho bước xử lí tiếp theo vì những ngữ nghĩa chính được phân tách ra, giúp nắm bắt đư c những ý nghĩa chính trong từng đoạn.

• Bước 3: mã hóa câu đơn thành vectơ số tương ứng. Việc làm này giúp cho việc truy vấn ngữ nghĩa trở nên nhanh và hiệu quả hơn nhiều so với việc đánh giá độ tương đồng giữa hai câu dạng văn bản.

• Bước 4: đóng gói những vectơ trên vào danh sách vectơ của tài liệu và lưu vào cơ sở dữ liệu. Sau này, khi truy vấn tương đồng, chỉ cần so sánh bộ vectơ truy vấn và danh sách vectơ của từng bằng sáng chế.

- Xử lí truy vấn và tìm kiếm trên cơ sở dữ liệu.

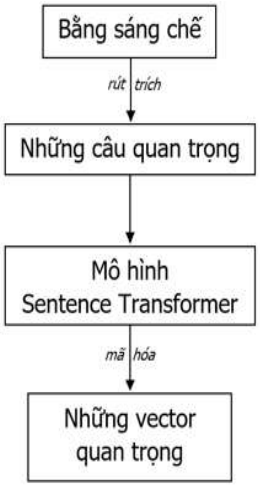
• Bước 1: tách đoạn truy vấn thành các câu đơn. Mục đích của việc này để rút trích ra tất cả ngữ nghĩa của đoạn văn cũng như phục vụ cho bước mã hóa vectơ số.

• Bước 2: mã hóa câu thành vectơ số. Tất cả câu trong truy vấn đều được mã hóa và lưu dưới dạng danh sách vectơ truy vấn.

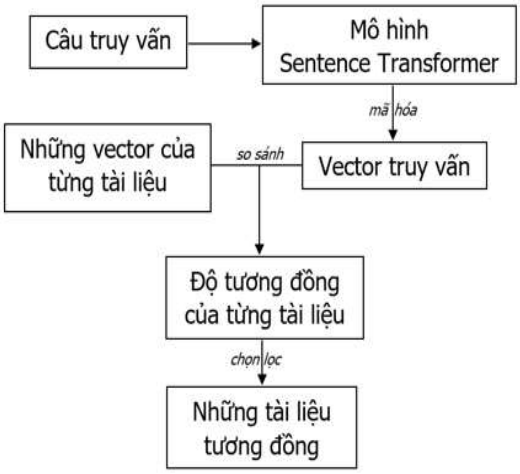
• Bước 3: truy vấn trong cơ sở dữ liệu. Tất cả vector truy vấn sẽ được so sánh với từng bộ vectơ của từng bằng sáng chế. Từ đây tính ra độ tương đồng giữa đoạn truy vấn và từng bằng sáng chế cụ thể.

• Đóng gói kết quả truy vấn vào danh sách, bao gồm điểm tương đồng và ID bằng sáng chế. Cuối cùng, lấy ra 10 bằng sáng chế có điểm tương đồng cao nhất.

Tóm lại, toàn bộ quy trình xử lí và mã hóa văn bản được mô tả trong Hình 3 và Hình 4.



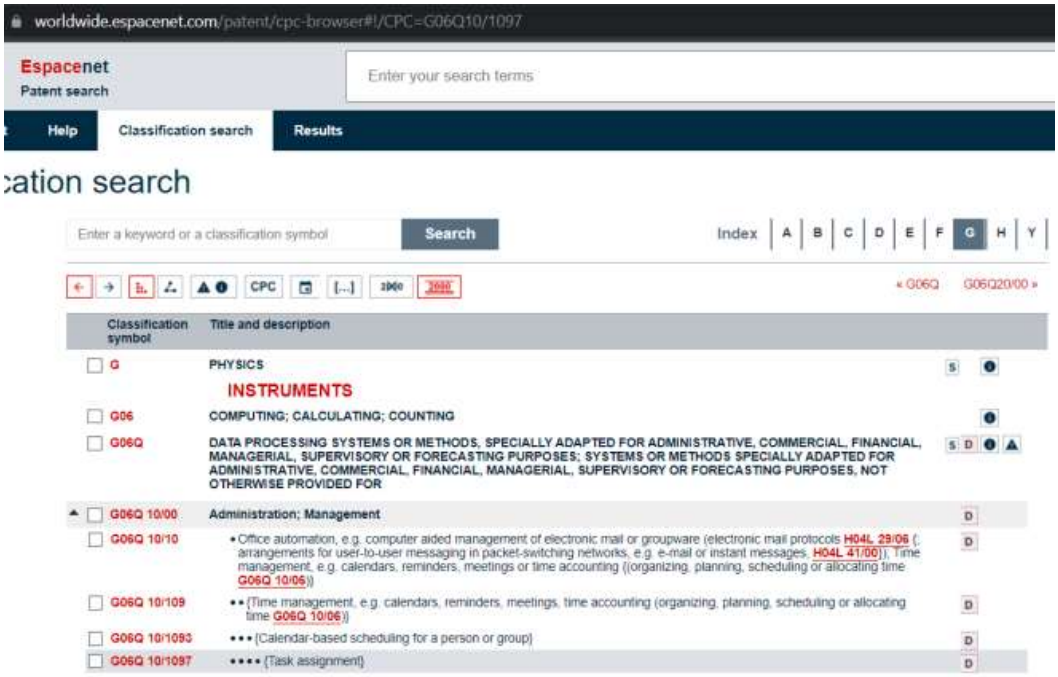
Hình 3 Xử lí bằng sáng chế và tạo bộ dữ liệu



Hình 4 Quá trình truy vấn bằng sáng chế

2.1 Tiền xử lí văn bản

Xây dựng cơ sở dữ liệu gồm các bằng sáng chế trong một nhóm có sẵn từ trang web Espace.net [14] của Châu Âu chuyên lưu trữ các nhóm bằng sáng chế dưới dạng cấu trúc cây minh hoạ ở Hình 5.



Hình 5 Kho dữ liệu bằng sáng chế của châu Âu EspaceNet.

Tuy nhiên, để tải nội dung của tất cả bằng phát minh trong lớp này, phải truy cập trang web USPTO của Mĩ. Trong quá trình tải từng bằng sáng chế, tác giả chỉ giữ lại phần tiêu đề và tóm tắt, đăng kí ý tưởng. Tiếp theo, những câu trong các phần này được tách ra dùng hàm sent_tokenize() trong thư viện nltk.tokenize(). Tiếp theo, những câu trên được mã hóa thành vectơ số tương ứng. Sau cùng, tất cả dữ liệu được đóng gói vào một file json để tiện truy vấn và lưu trữ. Tất cả những bước này và kết quả đầu ra là file json, đư c thể hiện trong Hình 6.

```
Get doc = 692/695
Get doc = 693/695
Get doc = 694/695
Get doc = 695/695
```

```
print('Class dictionary =')
print(cls_dict)
k = cls_dict['Patents']
for i in k:
    print(i)
```

```
Class dictionary =
{'ClassName': 'G06Q10/1097', 'Title': 'Task assignment', 'Description': 'Subject mat
{'DocID': 'UnitedStatesPatent_11270528', 'DocLink': 'https://patft.uspto.gov/netacgi
{'DocID': 'UnitedStatesPatent_11257045', 'DocLink': 'https://patft.uspto.gov/netacgi
{'DocID': 'UnitedStatesPatent_11250531', 'DocLink': 'https://patft.uspto.gov/netacgi
{'DocID': 'UnitedStatesPatent_11250355', 'DocLink': 'https://patft.uspto.gov/netacgi
{'DocID': 'UnitedStatesPatent_11250206', 'DocLink': 'https://patft.uspto.gov/netacgi
```

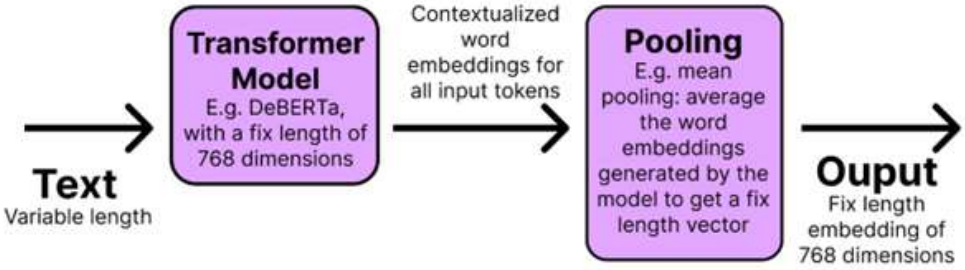
```
# Save to json file to later parts
json_object = json.dumps( cls_dict, indent = 4)

with open("class.json", "w") as outfile:
    outfile.write(json_object)
```

Hình 6 thể hiện kết quả của việc tải bằng sáng chế, xử lí văn bản và đóng gói cơ sở dữ liệu thành file json

2.2 Mô hình SentenceTransformer

Như đã đề cập ở phần trước, mô hình này đảm nhận công việc mã hóa câu dưới dạng vectơ rất chuẩn xác và nhanh chóng. Để đảm bảo tính chính xác và phù hợp giữa câu và vectơ số, việc lựa chọn mô hình được huấn luyện sẵn (pretrain model) cũng rất quan trọng. Vì tập dữ liệu đang làm việc có bản chất là các nghiên cứu khoa học, được đăng kí dưới dạng bằng sáng chế. Do đó, tác

giả chọn mô hình đã được huấn luyện sẵn có tên SPECTER của tác giả Arman Cohan [15] do mô hình được huấn luyện dựa trên tập dữ liệu gồm ba triệu bằng sáng chế đã được công bố. Như vậy, tập dữ liệu dùng trong nghiên cứu này và tập dữ liệu dùng để huấn luyện mô hình có độ tương đồng rất cao vì đều là tập dữ liệu bằng sáng chế. Qua đó, có thể tin tưởng vectơ câu được sinh ra sẽ sát với ý nghĩa cụ thể trong bối cảnh đư c nêu trong bằng sáng chế. Mô hình này được dùng cho việc mã hóa câu ở bước tiền xử lí cơ sở dữ liệu và cả phần mã hóa truy vấn (Hình 7).



Hình 7 Quá trình huấn luyện mô hình SentenceTransformer [3]

2.3 Cấu trúc của cơ sở dữ liệu

Trong nghiên cứu này, nhóm nghiên cứu dùng toàn bộ bằng sáng chế trong nhóm G06Q10/1097 để tiến hành thí nghiệm vì những tính chất sau:

- Số lượng bằng sáng chế không quá nhiều.
- Tập trung nói về chủ đề Giao nhận và quản lí công việc.
- Không quá chuyên sâu về mặt kĩ thuật.

File json chứa cơ sở dữ liệu gồm tên nhóm và các bằng sáng chế trong nhóm này. Cụ thể hơn, nhóm chứa các thuộc tính: ClassName, Title, Description, Patents; trong đó, ClassName là mã số của nhóm, Title là tên của nhóm, Description là đoạn văn mô tả nhóm, Patents là trường chứa tất cả các bằng sáng chế trong nhóm.

Đi sâu vào các thông tin cần lưu của một một bằng sáng chế, có các trường dữ liệu sau: DocID, DocLink, DocTitle, DocAbstract, DocVector. Chi tiết như sau: DocID chứa mã số văn bằng sáng chế, DocLink chứa đường dẫn đến bằng sáng chế trên trang web USPTO, DocTitle chứa tiêu đề của bằng sáng chế, DocAbstract lưu nội dung mô tả văn bằng phát minh, DocVector tập hợp các vector câu được tính từ DocTitle và DocAbstract dùng mô hình Sentence Transformer như thể hiện ở Hình 8.


```
{
    "ClassName": "G06Q10/1097",
    "Title": "Task assignment",
    "Description": "Subject matter drawn to a computerize
    "Patents": [
        {
            "DocID": "UnitedStatesPatent_11270528",
            "DocLink": "https://patft.uspto.gov/netacgi/
            "DocTitle": "Apparatus and method for vehicle
            "DocAbstract": "A vehicle maintenance schedu
            "DocVector": [
                [
                    0.22236116230487823,
                    0.5558313131332397,
                    -0.431037038564682,
```


Hình 8 Cấu trúc file json trong thực tế.

2.4 Phương pháp truy vấn và đánh giá độ tương đồng

Theo bài báo của nhóm nghiên cứu trước [6], cách tính cosine giữa hai vectơ được biểu diễn trong công thức 1. Trong quá trình thực nghiệm, tác giả chọn con số 40% điểm tương đồng cao nhất giữa hai tài liệu để tính điểm tương đồng trung bình. Về mặt ý tưởng, phương pháp [6] sẽ so sánh từng cụm danh từ dạng text với nhau nên quá trình tính toán sẽ phức tạp và tốn rất nhiều thời gian trong quá trình tính ra giá trị lớn nhất giữa các cặp từ với nhau. Trong khi phương pháp đề xuất so sánh bộ vectơ của hai tài liệu với nhau và tìm giá trị trung bình của 40 % điểm tốt nhất sẽ khách quan hơn dù giá trị tối

đa sẽ giảm chút ít nhưng đảm bảo việc đánh giá độ tương đồng ở mức tổng quan các nhóm ý nghĩa giữa hai tài liệu.

$$\cos(\theta) = \frac{A \bullet B}{||A||.||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Công thức 1. Cosine vectơ [6]

Trong đó A và B là hai vectơ đầu vào với n = 768 số thực. θ là góc giữa hai vector trong không gian đa chiều. Phép toán • là tích vô hướng của hai vector, chẳng hạn A(a1, a2) và B(b1, b2), khi đó A • B = a1.b1 + a2.b2. Phép toán |A| tính giá trị độ dài của vector.

Nền tảng cốt lõi của phép đo độ tương đồng dựa trên phép toán đo góc cosine giữa hai vector. Trước hết, cần nắm được tổng quan, cần so sánh hai tài liệu với nhau, từ đó tính được giá trị số từ 0,0 đến 1,0 đại diện cho độ tương đồng, giá trị càng gần 1,0 thì hai tài liệu càng giống nhau. Như vậy, việc tìm những tài liệu tương đồng có thể hiểu là việc so sánh bộ vectơ của tài liệu cho trước với tất cả văn bản khác trong cơ sở dữ liệu. Trong nghiên cứu này, từng cặp vectơ sẽ được so sánh với nhau theo công thức 1. Sau đó, tính điểm tương đồng của hai bộ vectơ dựa trên 40 % điểm cao nhất giữa các cặp vectơ mang ý nghĩa.

3 Kết quả nghiên cứu

Phương pháp đề xuất đã giải quyết được những vấn đề sau: truy vấn tài liệu tương đồng trong cơ sở dữ liệu với điểm tương đồng cao nhất và tốc độ truy vấn nhanh nhờ sử dụng kĩ thuật tìm kiếm vector và mô hình học sâu mới nhất. Tuy nhiên, kết quả về độ tương đồng vẫn chưa có sự cải thiện đáng kể so với hai phương pháp Get-In và GeT. Tổng quan, điểm tương đồng không có nhiều khác biệt giữa ba phương pháp.

Cụ thể, tốc độ truy vấn một văn bản là 0,3775 giây để tìm ra 10 bằng sáng chế tương đồng nhất trong cơ sở dữ liệu có 694 bằng sáng chế (Hình 9). Trong ví dụ này, nhóm sẽ tìm tất cả bằng phát minh tương đồng với tài liệu có ID "*UnitedStatesPatent_11270528*", tên tài liệu "Apparatus and method for vehicle maintenance scheduling and fault monitoring". Điểm tương đồng của 10 văn bản gần nhất được trình bày ở Bảng 1.

Bảng 1 Độ tương đồng của 10 bằng phát minh có điểm tương đồng cao nhất.

ID của bằng sáng chế (UnitedStatesPatent_)	Điểm tương đồng
10497185	0,95806
10423934	0,87347
11182984	0,83806
10733577	0,82903
11107306	0,81667
10665040	0,81156
11080950	0,81081
10825097	0,80698
6580982	0,80006
11080841	0,79855

Đặc biệt, tài liệu giống nhất với bằng phát minh cần tìm lại có cùng tiêu đề, và nội dung rất sát với nhau, đều về cùng chủ đề "Hệ thống quản lí hỏng hóc của xe cộ và những cách xử lí" trong số những kết quả có điểm tương đồng cao nhất.

Cụ thể hơn, kết quả ảnh chụp của bước tìm kiếm bằng sáng chế như Hình 9 và Hình 10.

```
Find_Suitable_Patent(fle, prefer)
```

```
Your chosen patent:
- Patent ID = UnitedStatesPatent_11270528
- Title = Apparatus and method for vehicle maintenance scheduling and fault monitoring
- Abstract = A vehicle maintenance scheduling and fault monitoring apparatus includes a vehicle system maintenance rules generation module
- Link = https://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO2&Sect2=HITOFF&u=%2Fnetahtml%2FPTO%2Fsearch-adv.htm&r=1&p=1&f=G&l=50&d=PTXT&S1
Searching in database, 694 patens....
- Search time = 0.3775 seconds
Most suitable patents:
+ Candiate = 0
- Patent ID = UnitedStatesPatent_10497185
- Similar Score = 0.9580646657882064
- Title = Apparatus and method for vehicle maintenance scheduling and fault monitoring
- Abstract = A vehicle health monitoring system including a rules generation module that receives historical time-stamped vehicle fault dat
```

Hình 9 Kết quả tìm kiếm 10 bằng sáng chế tương đồng nhất so với bằng sáng chế UnitedStatesPatent_11270528

```
+ Candiate = 1
- Patent ID = UnitedStatesPatent_10423934
- Similar Score = 0.8734718937123893
- Title = Automated vehicle diagnostics and maintenance
- Abstract = Systems, methods, and apparatuses described herein are directed to automated vehicle diagnostics and maintenance. For example, v
+ Candiate = 2
- Patent ID = UnitedStatesPatent_11182984
- Similar Score = 0.8380612826117181
- Title = Distributed maintenance system and methods for connected fleet
- Abstract = A system and related methods for management, planning and control of a connected fleet of vehicles. A unique, single integrated
+ Candiate = 3
- Patent ID = UnitedStatesPatent_10733577
- Similar Score = 0.8290313885095447
- Title = Preventive maintenance management system and method for generating maintenance schedule of machine, and cell controller
- Abstract = A preventive maintenance management system and method, and a cell controller, for monitoring preventive maintenance data, calcul
```

Hình 10 Ba kết quả tiếp theo trong kết quả truy vấn

Bảng 2 So sánh với phương pháp Get-In Patent Search [6], GeT-based Ontology search [7] về thời gian tìm kiếm.

Phương pháp	Thời gian (s)
Get-In patent Search	865,9728
GeT patent Search	934,7618
Phương pháp đề xuất	0,3775

So sánh với phương pháp Get-In Patent Search, GeT-based Ontology search về thời gian tìm kiếm ở Bảng 2 thì có thể thấy rõ phương pháp Get-In và GeT có tốc độ tìm kiếm chậm, lí do là:

- Cách làm việc phức tạp: tách từng từ trong cụm lớn và tạo thành những cụm từ nhỏ hơn, sau đó so sánh từng cụm từ nhỏ giữa hai nhóm với nhau.
- Số lượng cụm từ con cần so sánh quá nhiều giữa hai nhóm.
- So sánh bằng chữ và thông qua thư viện của WordNet hay NLTK nên chậm hơn hẳn so với tìm bằng vectơ số.

Ngoài ra, phương pháp Get-In còn có những nhược điểm sau:

- Chưa so sánh được hai câu hoàn chỉnh.
- Chưa mã hóa, tận dụng ưu điểm của vectơ số học khi tính độ tương đồng.

4 Kết luận

Phương pháp đề xuất trong nghiên cứu này đã tận dụng được những ưu điểm của các nghiên cứu trước và cải tiến cho phù hợp với bài toán tìm kiếm bằng phát minh tương đồng. Cải tiến thuật toán, cho phép so sánh ngữ nghĩa cả câu thay vì cụm từ như hiện tại. Ánh xạ câu thành vector với mô hình học sâu hiện đại, đã được huấn luyện với bộ dữ liệu tương tự. Tạo và quản lí cấu trúc dữ liệu phù hợp, nâng cao tốc độ tìm kiếm. Tiền xử lí văn bản theo phương pháp rút trích tất cả cụm từ đảm bảo việc lấy đủ thông tin tổng quát, do đó phương pháp tìm kiếm tương đồng ngữ nghĩa sẽ hiệu quả hơn. Cuối cùng là lựa chọn tham số phù hợp để tính trung bình điểm tốt nhất, tránh việc chỉ chọn điểm số cao nhất, sẽ giảm tính khách quan khi so sánh.

Tương lai phát triển đối với nghiên cứu này, có thể ứng dụng cấu trúc cây để quản lí, lưu trữ dữ liệu hiệu quả. Thay vì phải tìm tất cả bằng sáng chế trong dữ liệu, quá trình tìm kiếm sẽ bắt đầu từ siêu nhóm rồi đi dần đến các nhánh chứa nhóm phù hợp nhất, cuối cùng mới tìm kiếm tất cả tài liệu trong nhóm đó. Thông qua giải pháp này, sẽ nâng cao độ chính xác khi tìm nhóm phù hợp cũng như giảm khối lượng tính toán xuống mức tối thiểu, thay vì phải tìm toàn bộ cơ sở dữ liệu.

Tài liệu tham khảo

1. Google Patent, 2021. https://patents.google.com/

2. Ashish, V., Noam, S., Niki, P., Jakob, U., Jones, L., Aidan, N.G. và Łukasz, K. (2017). Attention Is All You Need. *Neural Information Processing Systems*, 6000-6010. DOI: 10.48550/arXiv.1706.03762

3. Jacob, D., Ming, W.C., Kenton, L. và Kristina, T.(2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv,* 4171-4186. DOI: 10.48550/arXiv.1810.04805

4. Reimers, Nils và Gurevych, Iryna (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Association for Computational Linguistics,* 3982-3992. DOI: 10.48550/arXiv.1908.10084
5. Wang, Kexin, T., Nandan, R., Nils, G., Iryna. (2021). *GPL:* Generative Pseudo Labeling for Unsupervised Domain Adaptation of Dense Retrieval. *arXiv,* 2345-2360. DOI: 10.48550/arXiv.2112.07577
6. Wang, Kexin, R., Nils, G., Iryna (2021). TSDAE: Using Transformer-based Sequential Denoising Auto-Encoder for Unsupervised Sentence Embedding Learning. *Association for Computational Linguistics,* 671-688. DOI: 10.48550/arXiv.2104.06979
7. Cong, P.P., Hong, Q.N., Hoang, M.N., (2017). GeT-IN an Integrated Ontology-Based Approach for Patent Search. *International Journal of Management and Applied Science,* 99-105. DOI: 10.1145/3011141.3011205
8. Vinh, V.X., Hong, Q.N., Khoi, N.T. (2014). GeT-based Ontology Construction for Semantic Disambiguation. *Proceedings of the 16th International Conference on Information Integration and Web-based Applications & Services,* 445-453. DOI: 10.1145/2684200.2684320
9. WordNet, 2021. https://wordnet.princeton.edu/
10. Matthias, K., Patrick, K., Stefan, S., Freddy, L., Abraham, B. (2016). Semantic Web Service Search: A Brief Survey. *Künstliche Intelligenz Journal,* 139-147. DOI: 10.1007/s13218-015-0415-7
11. USPTO Patent Database, 2021. https://ppubs.uspto.gov/pubwebapp/
12. USPTO. *Overview of the U.S. Patent Classification System*, Dec 2012. https://www.uspto.gov/sites/default/files/patents/resources/classification/overview.pdf
13. USPTO. *USPTO Classification & Design*, 2020. https://www.uspto.gov/patents/laws/examination-policy/seven-classification-design-patents
14. Espacenet Patent Search, 2021. https://worldwide.espacenet.com/
15. Arman, C., Sergey, F., Iz, B., Doug, D., Daniel, S.W. (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. *arXiv* (2270-2282). DOI: 10.48550/arXiv.2004.07180

Patent retrieval based on semantic search

Duong Hon Minh
Faculty of Pharmacy - Nguyen Tat Thanh University
dhminh@ntt.edu.vn

Abstract Keyword-based search is a popular, classical, and limited search method. In contrast, semantic search have the ability to understand the specific topic or meaning of each paragraph and sentence.
For the above reasons, a new search method was proposed to solve the problem of finding similar patents because it may prevent the registration of the same ideas and protect the ownership and commercial rights. Two main techniques in semantic search include: vector search and Natural Language Processing for English patent documents. The data source was taken from the USPTO website belonging to the USA government. The new outcome of this research is: to find documents that are close to the meaning of a given document more accurately and rapidly.
The initial results of the present study showed that 10 most similar patents were found in a database of 694 patents within only 0.55692 seconds. This result is much faster than other similar search methods.
Keywords Vector Search, Patent Search, Semantic Search, Deep Learning, Sentence Transformer