Truy vấn bài báo Việt Nam

Anonymous ACL submission

Abstract

Nghiên cứu này đề xuất hệ thống truy vấn bài báo tiếng Việt, kết hợp TF-IDF với SBERT-Vietnamese và BM25 với PhoBERT để nâng cao hiệu quả tìm kiếm. TF-IDF và BM25 đảm bảo xử lý từ khóa nhanh, trong khi SBERT và PhoBERT khai thác ngữ nghĩa sâu hơn. Kết quả thực nghiệm trên tập dữ liệu bài báo lớn cho thấy mô hình kết hợp đạt độ chính xác và hiệu suất vượt trội, phù hợp cho các hệ thống tìm kiếm tiếng Việt thực tế.

1 Giới thiệu

Trong thời đại số hóa, khi lượng thông tin trên Internet ngày càng gia tăng, làm thế nào để tìm kiếm các bài báo tiếng Việt một cách nhanh chóng, chính xác và hiệu quả? Với cấu trúc ngôn ngữ phức tạp, đòi hỏi các hệ thống truy vấn thông tin không chỉ dừng lại ở mức xử lý từ khóa mà còn phải nắm bắt được ý nghĩa sâu sắc của văn bản. Liệu các phương pháp truyền thống như TF-IDF và BM25, vốn đơn giản và hiệu quả với từ khóa, có đủ khả năng đáp ứng yêu cầu này? Hay cần đến sự hỗ trợ từ các mô hình hiện đại như SBERT-Vietnamese và PhoBERT, vốn được biết đến với khả năng vượt trội trong xử lý ngữ nghĩa?

Dù các mô hình hiện đại mang lại kết quả tốt hơn, chi phí tính toán cao và sự phức tạp trong triển khai khiến chúng khó áp dụng trực tiếp trong các hệ thống thực tế. Đó chính là lý do nghiên cứu này đề xuất một cách tiếp cận kết hợp, tận dụng ưu điểm của cả hai phương pháp. TF-IDF và BM25 đảm nhiệm vai trò xử lý nhanh và lọc sơ bộ các kết quả truy vấn ban đầu, trong khi SBERT-Vietnamese và PhoBERT được sử dụng để đánh giá lại và sắp xếp các kết quả dựa trên ngữ nghĩa sâu hơn.

Hướng tiếp cận này đã cải thiện độ chính xác tối ưu hóa thời gian phản hồi của hệ thống, mở ra tiềm năng lớn cho các ứng dụng tìm kiếm thông tin tiếng Việt. Bài báo này sẽ trình bày chi tiết phương pháp đề xuất, đánh giá hiệu quả trên tập dữ liệu bài báo tiếng Việt, và so sánh với các phương pháp truyền thống.

2 Bộ dữ liệu

2.1 Tổng quan

Trong đề tài lần này, dữ liệu được thu thập từ 5 trang báo điện tử uy tín tại Việt Nam, bao gồm: Báo Lao Động, Báo Dân Trí, Báo VnExpress, Báo VTC, Báo Đảng Cộng sản Việt Nam. Đối với mỗi trang báo, dữ liệu được khai thác tự động và lưu trữ với 7 đặc trưng chính: title (tiêu đề), abstract (tổng quan), source (nguồn bài báo), link (đường dẫn bài báo), topic (chủ đề), time (thời gian đăng bài) và imglink (đường dẫn chứa ảnh tượng trưng cho bài báo).

Tổng cộng, dữ liệu nhóm thu thập được bao gồm 49,543 bài báo.

	title	abstract	source	link	topic	time	imglink
0	Sửa đổi Luật Điện lực: Yêu cầu xuất phát từ th...	Sau gần 20 năm triển khai thi hành và sửa đổi,...	Báo VTC	https://vtcnews.vn/sua-doi-luat-dien-luc-yeu-c...	Thời sự	22/10/2024 08:52	https://cdn-i.vtcnews.vn/resize/me/upload/2024...
1	Phát huy tối đa sức mạnh Nhân dân để đất nước ...	Theo ông Đỗ Văn Chiến, để đất nước bước vào kỷ...	Báo VTC	https://vtcnews.vn/phat-huy-toi-da-suc-manh-nh...	Thời sự	22/10/2024 07:00	https://cdn-i.vtcnews.vn/resize/me/upload/2024...
2	Xem nghệ nhân tỉ mỉ trùng tu cung điện nơi 13 ...	Các nghệ nhân tỉ mỉ vẽ từng hoa tiết nhỏ nhất,...	Báo VTC	https://vtcnews.vn/xem-nghe-nhan-ti-mi-trung-t...	Thời sự	22/10/2024 06:39	https://cdn-i.vtcnews.vn/resize/me/upload/2024...
3	Tin tai nạn giao thông mới nhất ngày 22/10: Ô ...	Ô tô lao xuống vực khiến một người tử vong, tà...	Báo VTC	https://vtcnews.vn/tin-tai-nan-giao-thong-moi-...	Thời sự	22/10/2024 06:27	https://cdn-i.vtcnews.vn/resize/me/upload/2024...

Figure 1: Một phần của bộ dữ liệu

2.2 Tiền xử lí dữ liệu

Trong tác vụ truy vấn, nhóm lựa chọn hai đặc trưng chính: title và abstract. Đối với mỗi bài báo, dữ liệu dạng văn bản được xây dựng bằng cách kết hợp title và abstract, sau đó trải qua quy trình tiền xử lý gồm 4 bước:

- Lowercasing : Toàn bộ văn bản được chuyển thành chữ thường để đảm bảo tính thống nhất (ví dụ: Bóng đá → bóng đá).

- Word segmentation: Áp dụng phân tách từ để xử lý ngôn ngữ tự nhiên tiếng Việt (ví dụ: bóng đá → bóng_đá).

- Loại bỏ stopword: Xóa bỏ các từ dừng không mang ý nghĩa quan trọng trong truy vấn. (Ví dụ: a lô, a ha, ai, ai ai, ai nấy, ai đó,...)

- Loại bỏ ký tự đặc biệt: Loại bỏ các ký tự đặc biệt để làm sạch văn bản (ví dụ: tp.hcm → tphcm).

Tương tự, câu truy vấn đầu vào cũng được xử lý qua các bước này để đảm bảo tính tương thích và đồng nhất với dữ liệu bài báo, từ đó cải thiện độ chính xác của hệ thống truy vấn.

3 Phương pháp

3.1 Các phương pháp retrieval-based

3.1.1 TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) là một phương pháp đánh giá mức độ quan trọng của một từ trong một văn bản. Nó được tính bằng cách nhân tần suất xuất hiện của từ trong văn bản với tần suất nghịch đảo của từ đó trong toàn bộ tập dữ liệu. Để chuẩn hóa giá trị TF, số lần xuất hiện của từ trong văn bản sẽ được chia cho tổng số từ trong văn bản đó.

Công thức:

$$TF(t,d) = \frac{\text{Số lần từ t xuất hiện trong văn bản d}}{\text{Tổng số từ trong văn bản d}}$$

$$IDF(t,D) = 1 + \log\left(\frac{N}{\text{Số văn bản chứa từ } t}\right)$$

$$TF\text{-}IDF(t,d,D) = TF(t,d) \times IDF(t,D)$$

3.1.2 BM25

BM25 là một phương pháp xếp hạng tựa như tf-idf, là hàm tính thứ hạng được các công cụ tìm kiếm sử dụng để xếp hạng các văn bản theo độ phù hợp với truy vấn nhất định

3.2 Pre-trained language model

3.2.1 Vietnamese-SBERT

Vietnamese_sbert là một mô hình ngôn ngữ đã được pre-trained được sử dụng để biểu diễn dưới dạng vector đặc trưng, được xây dựng dựa trên mô hình Sentence-transformers.

3.2.2 PhoBERT

Đây là một pre-trained được huấn luyện monolingual language, tức là chỉ huấn luyện dành riêng cho tiếng Việt. PhoBERT được train trên khoảng 20GB dữ liệu bao gồm khoảng 1GB Vietnamese Wikipedia corpus và 19GB còn lại lấy từ Vietnamese news corpus. Đây là một lượng dữ liệu khả ổn để train một mô hình như BERT.PhoBERT sử dụng RDRSegmenter của VnCoreNLP để tách từ cho dữ liệu đầu vào trước khi qua BPE encoder. Do tiếp cận theo tư tưởng của RoBERTa, PhoBERT chỉ sử dụng task Masked Language Model để train, bỏ đi task Next Sentence Prediction.

4 Đánh giá mô hình

4.1 nDCG

Trong phương pháp đánh giá mô hình xếp hạng, nDCG/NDCG (Normalized Discounted Cumulative Gain) là một chỉ số được sử dụng rộng rãi để đo lường hiệu quả của các hệ thống xếp hạng, đặc biệt trong bài toán tìm kiếm và gợi ý. Chỉ số này tập trung vào việc đánh giá mức độ liên quan và thứ tự của các kết quả trả về.

Công thức tính nDCG bao gồm hai bước chính:

4.1.1 Tính DCG

$$DCG_k = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}$$

Trong đó:

- k: Số lượng mục trong danh sách xếp hạng.

- rel_i: Mức độ liên quan (relevance) của mục tại vị trí i.

- i: Vị trí của mục trong danh sách.

4.1.2 Tính nDCG

$$nDCG_k = \frac{DCG_k}{iDCG_k}$$

Trong đó:

- $iDCG_k$: Giá trị DCG của danh sách xếp hạng lý tưởng, được tính tương tự DCG_k nhưng dựa trên thứ tự sắp xếp đúng nhất.

4.2 Đánh giá thủ công

Đánh giá thủ công được thực hiện bằng cách kiểm tra trực tiếp các kết quả trả về từ hệ thống dựa trên các tiêu chí cụ thể. Quá trình này bao gồm việc đánh giá mức độ liên quan của các kết quả với truy vấn, so sánh với dữ liệu hoặc kết quả lý tưởng, và

xác định các lỗi hoặc khu vực cần cải thiện. Mặc dù phương pháp này có thể mang lại kết quả chính xác và chi tiết, nhưng nó đòi hỏi nhiều thời gian và công sức.

Trong bài toán này, dựa trên 20 truy vấn và các đánh giá thủ công trên thang điểm năm dựa trên 50 kết quả có điểm tương đồng cosine cao nhất của cả bốn mô hình, điểm nDCG trung bình đạt được cao nhất là 0.778 của mô hình TF-IDF+SBERT, cho thấy mô hình có khả năng xếp hạng khá tốt nhưng vẫn cần cải thiện để đạt hiệu quả tối ưu hơn.

Model	Score
tf-idf	0,63145563
tf-idf+sbert	0,778234073
bm25	0,653780319
bm25+phobert	0,702323336

Figure 2: Điểm nDCG trung bình của các mô hình

5 Tài liệu tham khảo

[1]: Phobert: Dat Quoc Nguyen and Anh Tuan Nguyen. 2020. PhoBERT: Pre-trained language models for Vietnamese. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1037–1042, Online. Association for Computational Linguistics.

[2]: nDCG: Jarvelin, K. and Jaana, K. (2002) Cumulated Gain-Based Evaluation of IR Techniques. ACM Transactions on Information Systems, 20, 422-446.

[3]: huggingface.co/keepitreal/vietnamese-sbert