BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

……..……..……..……..

ĐINH NGỌC THI

PHÁT TRIỂN CÁC MÔ HÌNH HỌC SÂU KẾT HỢP CẤU TRÚC ĐỒ THỊ VÀ PHÂN TÍCH NGỮ NGHĨA CHO BÀI TOÁN KHUYẾN NGHỊ TRÍCH DẪN

LUẬN ÁN TIẾN SĨ MÁY TÍNH

Hà Nội - 2025

BỘ GIÁO DỤC
VÀ ĐÀO TẠO

VIỆN HÀN LÂM KHOA HỌC
VÀ CÔNG NGHỆ VIỆT NAM

HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ

……..……..……..

ĐINH NGỌC THI

PHÁT TRIỂN CÁC MÔ HÌNH HỌC SÂU KẾT HỢP CẤU TRÚC ĐỒ THỊ VÀ PHÂN TÍCH NGỮ NGHĨA CHO BÀI TOÁN KHUYẾN NGHỊ TRÍCH DẪN

LUẬN ÁN TIẾN SĨ MÁY TÍNH

Ngành: Khoa học máy tính
Mã số: 9 48 01 01

Xác nhận của Học viện Khoa học và Công nghệ	**Người hướng dẫn 1** (*Ký, ghi rõ họ tên*)	**Người hướng dẫn 2** (*Ký, ghi rõ họ tên*)
	PGS.TS. Võ Đình Bảy	PGS.TS. Nguyễn Long Giang

Hà Nội - 2025

LỜI CAM ĐOAN

Tác giả xin cam đoan luận án "Phát triển các mô hình học sâu kết hợp cấu trúc đồ thị và phân tích ngữ nghĩa cho bài toán khuyến nghị trích dẫn" là công trình nghiên cứu của chính mình dưới sự hướng dẫn khoa học của hai Thầy giáo PGS.TS. Võ Đình Bảy và PGS.TS. Nguyễn Long Giang. Luận án sử dụng thông tin trích dẫn từ nhiều nguồn tham khảo khác nhau và các thông tin trích dẫn được ghi rõ ràng nguồn gốc. Các kết quả nghiên cứu của tác giả công bố chung với các tác giả khác đã được sự nhất trí của đồng tác giả khi đưa vào luận án. Các số liệu, kết quả được trình bày trong luận án là hoàn toàn trung thực và chưa từng được công bố trong bất kỳ một công trình nào khác ngoài các công trình công bố của tác giả. Luận án được hoàn thành trong thời gian tác giả làm nghiên cứu sinh tại Học viện Khoa học và Công nghệ, Viện Hàn lâm Khoa học và Công nghệ Việt Nam.

Hà nội, ngày 05 tháng 03 năm 2025

Tác giả luận án

Đinh Ngọc Thi

LỜI CẢM ƠN

Luận án này được hoàn thành với sự nỗ lực không ngừng của bản thân tác giả và sự giúp đỡ tận tâm hết mình từ các Thầy giáo hướng dẫn, bạn đồng hành và người thân trong suốt những năm tháng học tập và nghiên cứu tại Viện Công nghệ thông tin (Viện Hàn lâm Khoa học và Công nghệ Việt Nam).

Đầu tiên, tác giả xin bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới hai Thầy giáo hướng dẫn PGS.TS. Võ Đình Bảy và PGS.TS Nguyễn Long Giang. Sự tận tình chỉ bảo, hướng dẫn và động viên của các Thầy dành cho tác giả trong suốt thời gian thực hiện làm nghiên cứu sinh là không thể nào kể hết được. Tác giả cũng xin bày tỏ lòng biết ơn sâu sắc đối với sự giúp đỡ, chia sẻ và đồng hành của TS. Phạm Thế Anh Phú trong suốt quá trình nghiên cứu vừa qua. Tác giả cũng xin gửi lời cảm ơn tới các Thầy, Cô giáo và các Cán bộ của Viện Công nghệ thông tin, Ban Lãnh đạo, phòng Đào tạo, các phòng chức năng của Học viện Khoa học và Công nghệ (Viện Hàn lâm Khoa học và Công nghệ Việt Nam) đã nhiệt tình giúp đỡ và tạo ra môi trường nghiên cứu thuận lợi để tác giả có thể hoàn thành công trình nghiên cứu của mình.

Đặc biệt tác giả xin kính dâng tặng luận án này như một lời tri ân và tưởng nhớ đến Bố, lúc sinh thời đã luôn muốn tác giả đi theo con đường nghiên cứu khoa học. Tác giả cũng xin được bày tỏ lòng biết ơn sâu sắc tới Mẹ, vợ con và các thành viên trong gia đình, những người đã luôn khuyến khích, động viện và truyền cảm hứng cho tác giả trong suốt quá trình nghiên cứu.

Tác giả xin trân trọng cảm ơn!

Hà nội, ngày 05 tháng 03 năm 2025

Tác giả luận án

Đinh Ngọc Thi

MỤC LỤC

DANH MỤC CÁC KÝ HIỆU, CÁC CHỮ KÝ VIẾT TẮT

Tiếng Anh	Viết tắt	Diễn giải
Association for Computational Linguistics	ACL	Hiệp hội ngôn ngữ học tính toán
ACL Anthology Reference Corpus	ACL-ARC	Tuyển tập bài báo của hội nghị ACL
ACL Anthology Network	AAN	Mạng tuyển tập ACL
Bidirectional Encoder Representations from Transformers	BERT	Biểu diễn mã hóa 2 chiều từ bộ biến đổi
Bidirectional Long-Short Term Memory	BiLSTM	Bộ nhớ hai chiều
Convolutional Neural Networks	CNN	Mạng nơ-ron tích chập
Citation recommendation	CR	Khuyến nghị trích dẫn
Decoder		Bộ giải mã
Dual Local Citation Recommendation	DualLCR	Khuyến nghị trích dẫn cục bộ kép
Encoder		Bộ mã hóa
Feedforward Neural Networks	FNN	Mạng nơ-ron truyền thẳng
Gated Recurrent Units	GRU	Đơn vị hồi quy có cổng
Graph Convolutional Networks	GCN	Mạng tích chập đồ thị
Graph Neural Networks	GNN	Mạng nơ-ron đồ thị
Mean Average Precision	MAP	Độ chính xác trung bình
Mean Reciprocal Rank	MRR	Xếp hạng đối ứng trung bình
Neural Citation Networks	NCN	Mạng nơ-ron trích dẫn
Recurrent Highway Networks	RHN	Mạng hồi quy
Recurrent Neural Networks	RNN	Mạng nơ-ron hồi quy
Time-Delay Neural Networks	TDNN	Mạng nơ-ron trễ
Variational Graph Auto-Encoders	VGAE	Bộ mã hóa đồ thị tự động điều chỉnh

DANH MỤC BẢNG

DANH MỤC HÌNH VẼ, ĐỒ THỊ

MỞ ĐẦU

Trong thời đại của khoa học và công nghệ, việc viết bài báo khoa học là một hoạt động không thể thiếu đối với các nhà nghiên cứu, học giả, và sinh viên. Công bố các bài khoa học không chỉ là cách để truyền đạt kiến thức, kết quả, và ý tưởng của quá trình nghiên cứu, mà còn là cách để giao lưu, hợp tác, và đóng góp cho cộng đồng khoa học. Tuy nhiên, việc viết bài báo khoa học cũng đòi hỏi nhiều kỹ năng và nỗ lực, trong đó một trong những kỹ năng quan trọng nhất là trích dẫn tài liệu tham khảo.

Trích dẫn tài liệu tham khảo là việc ghi nhận nguồn gốc của các thông tin, ý tưởng hoặc dữ liệu của người khác mà được sử dụng trong bài viết hoặc công trình của mình. Trích dẫn tài liệu tham khảo có nhiều lợi ích, như kiểm tra tính chính xác của các thông tin, trở thành một nhà nghiên cứu có cách viết văn phong khoa học tốt hơn, thể hiện kiến thức khoa học, xây dựng uy tín như một học giả chính trực, và cho phép xác minh tính liêm chính công trình của mình. Đồng thời, việc trích dẫn cũng là cách thể hiện sự tôn trọng và công nhận công lao của những các tác giả trước đó. Tuy nhiên, việc trích dẫn tài liệu tham khảo cũng gặp nhiều thách thức, như tìm kiếm và lựa chọn các tài liệu tham khảo phù hợp, tuân thủ các quy tắc trích dẫn và tránh đạo văn hoặc sao chép.

Với sự gia tăng vượt bậc về số lượng các bài báo khoa học được xuất bản hàng năm trên nhiều lĩnh vực khác nhau, việc xác định các bài viết phù hợp với một chủ đề cụ thể và có giá trị trích dẫn đã trở thành một nhiệm vụ ngày càng phức tạp, đòi hỏi nhiều nỗ lực và thời gian từ các nhà nghiên cứu [1] [2] [3] [4]. Để giải quyết vấn đề này và hướng tới mục tiêu giảm bớt gánh nặng cho các nhà khoa học, việc nghiên cứu và phát triển các mô hình khuyến nghị trích dẫn tự động trở thành một yêu cầu cấp thiết. Vấn đề nghiên cứu này được gọi là bài toán khuyến nghị trích dẫn (*citation recommendation*) [5]. Đặc biệt, bài toán khuyến nghị trích dẫn này đã thu hút nhiều sự quan tâm gần đây [6] [7] [8] [9] vì nó hứa hẹn sẽ nâng cao chất lượng quá trình nghiên cứu khoa học.

1. Tính cấp thiết của đề tài luận án

Hiện nay, số lượng các bài báo khoa học được công bố đang gia tăng với tốc độ chưa từng có, dẫn đến tình trạng quá tải thông tin trong lĩnh vực hàn lâm. Điều này đặt ra những thách thức lớn cho các nhà nghiên cứu, đặc biệt là những người trẻ và thiếu

kinh nghiệm, trong việc xác định các tài liệu có liên quan và đảm bảo chất lượng để trích dẫn cho nghiên cứu của mình. Trước thực trạng này, các hệ thống khuyến nghị trích dẫn tự động được kỳ vọng sẽ giảm bớt gánh nặng bằng cách cung cấp các đề xuất phù hợp, hỗ trợ các nhà nghiên cứu định hướng hiệu quả trong khối lượng thông tin khổng lồ.

Tuy nhiên các cách tiếp cận hiện tại cho bài toán khuyến nghị trích dẫn vẫn tồn tại những hạn chế nhất định. Hạn chế thứ nhất liên quan đến việc mô hình khuyến nghị chưa được cung cấp đủ thông tin của các bài báo khoa học. Một trong những nghiên cứu được công bố sớm cho bài toán khuyến nghị trích dẫn là của Ebesu và Fang [10] công bố vào năm 2017 và nhóm của Färber [11] nâng cao kết quả này vào năm 2020. Họ đề xuất một kiến trúc sử dụng bộ mã hóa-giải mã (*encoder-decoder*) linh hoạt được gọi là mạng nơ-ron trích dẫn (*Neural Citation Networks - NCN*). Mặc dù mô hình này thu được kết quả tốt hơn so với các mô hình cùng thời trên các bộ dữ liệu RefSeer và arXiv CS, tuy nhiên nó vẫn còn những hạn chế như là chưa sử dụng toàn bộ các thông tin của bài báo (tiêu đề, tác giả, năm xuất bản, nơi công bố.v.v…) để đưa vào huấn luyện cho mô hình.

Hạn chế thứ hai liên quan đến việc các mô hình vẫn chưa sử dụng những thành tựu mới nhất của các nghiên cứu về học sâu. Ví dụ như các mô hình khuyến nghị kép DualLCR [12] hay DualLCR-design [13] do nhóm Medić và Šnajder công bố lần lượt vào các năm 2020 và 2022 vẫn sử dụng bộ nhớ hai chiều (*Bidirectional Long-Short Term Memory, BiLSTM*) [14]. Hoặc như mô hình BERT-GCN của nhóm nghiên cứu Jeong [15] vẫn chưa sử dụng kết quả mới nhất về xử lý ngôn ngữ tự nhiên hoặc đồ thị liên kết trích dẫn trong các bài báo khoa học.

Hạn chế thứ ba liên quan đến việc các mô hình khuyến nghị trích dẫn hiện tại đang tập trung vào ngữ cảnh trích dẫn và nội dung của bài báo ứng viên [16] [17], mà chưa quan tâm đúng mức đến siêu dữ liệu của bài báo, như là tên tác giả, năm và địa điểm công bố của bài báo, trong khi những yếu tố này có ảnh hưởng đáng kể đến xu hướng trích dẫn của các nhà khoa học. Bởi vì họ thường ưu tiên trích dẫn các công trình của những tác giả có uy tín, các công bố gần đây, hoặc các bài viết được đăng tải trên các tạp chí và hội nghị có danh tiếng trong lĩnh vực nghiên cứu của họ.

Việc góp phần giải quyết các hạn chế nêu trên là nội dung chính được trình bày trong luận án này.

2. Mục tiêu của luận án

Mục tiêu chung của luận án: Áp dụng các thành tựu tiên tiến của các mô hình học sâu để phát triển mô hình khuyến nghị trích dẫn mới. Bên cạnh đó, đề xuất các giải pháp nâng cao hiệu năng cho các mô hình khuyến nghị trích dẫn hiện nay.

Mục tiêu cụ thể: Đề xuất một số phương pháp nhằm nâng cao hiệu suất của các mô hình khuyến nghị trích dẫn hiện nay:

- Theo hướng tiếp cận lọc nội dung, đề xuất các giải pháp để nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn (*Neural Citation Networks - NCN*).
- Theo hướng tiếp cận kết hợp lọc nội dung và lọc cộng tác, đề xuất một mô hình mới tên là RHN-DualLCR, trong đó bao gồm các giải pháp để nâng cao hiệu suất cho mô hình khuyến nghị trích dẫn kép.
- Theo hướng tiếp cận kết hợp lọc nội dung và lọc dựa vào đồ thị, đề xuất mô hình khuyến nghị trích dẫn mới tên là SciBERT-GraphSAGE bằng cách kết hợp các thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên và đồ thị biểu diễn các liên kết trích dẫn.

3. Đối tượng nghiên cứu của luận án:

Luận án tập trung nghiên cứu và tìm hiểu một số đối tượng liên quan đến áp dụng học sâu cho bài toán khuyến nghị trích dẫn:

- Một số mô hình học sâu tiên tiến đã có hiện nay cho bài toán khuyến nghị trích dẫn.
- Những cải tiến của các mô hình học sâu, đặc biệt là các thành tựu nổi bật trong xử lý ngôn ngữ tự nhiên cũng như các phương pháp biểu diễn các dạng dữ liệu của bài báo khoa học.
- Một số chỉ số đánh giá hiệu suất cũng như các bộ dữ liệu thường dùng trong các mô hình khuyến nghị trích dẫn tiên tiến hiện nay.

4. Phạm vi nghiên cứu của luận án

- Nghiên cứu và đưa ra các giải pháp để nâng cao hiệu suất cho mô hình mạng nơ-ron khuyến nghị trích dẫn NCN.
- Nghiên cứu và đưa ra các giải pháp để nâng cao hiệu suất cho mô hình khuyến nghị trích dẫn cục bộ kép DualLCR (*Dual Local Citation Recommendation*).
- Nghiên cứu cách kết hợp các thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên và đồ thị biểu diễn các liên kết trích dẫn để xây dựng một mô hình khuyến nghị trích dẫn mới.
- Tập dữ liệu thử nghiệm cho các mô hình trong luận án chỉ bao gồm các bài báo được viết bằng tiếng Anh. Các bài báo viết bằng các ngôn ngữ khác không thuộc phạm vi nghiên cứu của luận án. Ngoài ra, các vấn đề như trích dẫn ảo hay trích dẫn sai cũng không thuộc phạm vi nghiên cứu của luận án này.

5. Phương pháp nghiên cứu

Nghiên cứu lý thuyết: Nghiên cứu và phân tích các kết quả đạt được từ các hệ thống khuyến nghị trích dẫn hiện nay, đánh giá ưu điểm và hạn chế của các hệ thống này, đồng thời đề xuất các phương pháp cải tiến nhằm nâng cao hiệu năng và độ chính xác của kết quả khuyến nghị thông qua việc ứng dụng các kỹ thuật và mô hình học sâu. Ngoài ra, tiến hành khảo sát các chỉ số đánh giá hiệu năng và các bộ dữ liệu phổ biến được sử dụng trong lĩnh vực mô hình khuyến nghị trích dẫn.

Nghiên cứu thực nghiệm: Thực hiện cài đặt và triển khai mã nguồn trên các bộ dữ liệu phổ biến trong môi trường thực nghiệm, nhằm đo lường và đánh giá hiệu quả của các kết quả đạt được từ các mô hình khuyến nghị trích dẫn được đề xuất trong luận án.

6. Các đóng góp của luận án

Với mục tiêu góp phần nâng cao hiệu năng của các mô hình khuyến nghị trích dẫn tiến tiến hiện nay, luận án đã có những đóng góp như sau:

- Theo hướng tiếp cận lọc nội dung, đề xuất các giải pháp để nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn NCN [10] [11] [CT1].

- Theo hướng tiếp cận kết hợp lọc nội dung và lọc cộng tác, đề xuất một mô hình mới tên là RHN-DualLCR, trong đó bao gồm các giải pháp để nâng cao hiệu suất cho mô hình khuyến nghị trích dẫn cục bộ kép DualLCR đã được công bố bởi Medić và Šnajder [12] [13] [CT2, CT4].
- Theo hướng tiếp cận kết hợp lọc nội dung và lọc đồ thị, đề xuất xây dựng mô hình khuyến nghị trích dẫn mới tên là SciBERT-GraphSAGE bằng cách kết hợp 2 thành tựu tiên tiến hơn trong xử lý ngôn ngữ tự nhiên SciBERT [18] và đồ thị biểu diễn các liên kết trích dẫn GraphSAGE [19] [CT3, CT5].

7. Bố cục của luận án

Luận án được thiết kế bao gồm phần mở đầu và bốn chương nội dung. Thêm vào đó là phần kết luận, danh mục các công trình của tác giả và danh mục các tài liệu tham khảo. Cụ thể, nội dung của các phần được tóm tắt như sau:

Chương 1 sẽ trình bày tổng quan về bài toán khuyến nghị trích dẫn và một số các kiến thức nền tảng cần thiết để thuận tiện cho việc hiểu phương pháp được đề xuất ở các chương phía sau. Các nghiên cứu liên quan về bài toán khuyến nghị trích dẫn cũng được trình bày cụ thể và những nghiên cứu đó được phân theo từng nhóm mô hình học sâu: lọc cộng tác, lọc nội dung, lọc dựa vào đồ thị và các phương pháp kết hợp. Trên cơ sở đó, luận án phân tích một số các hạn chế của các cách tiếp cận hiện tại và nêu rõ mục tiêu để giải quyết các hạn chế đó.

Chương 2, chương 3 và chương 4 là các đóng góp chính của luận án. Mỗi chương được thiết kế là một phương pháp đưa ra để nâng cao hiệu quả của các mô hình khuyến nghị tiên tiến. Các đề xuất của luận án này có liên quan chặt chẽ với nhau và bao phủ lên các hướng nghiên cứu về khuyến nghị trích dẫn hiện nay. Cụ thể, chương 2 tập trung vào hướng nghiên cứu lọc nội dung, trong đó đề xuất các giải pháp để nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn NCN bằng cách thêm tiêu đề của bài báo vào mô hình và sử dụng thành tựu xử lý ngôn ngữ tự nhiên mới nhất.

Chương 3 tập trung vào hướng nghiên cứu kết hợp lọc nội dung và lọc cộng tác, trong đó trình bày chi tiết một phương pháp đề xuất để cải thiện hiệu quả của mô hình kép DualLCR cho bài toán khuyến nghị trích dẫn. Phương pháp đề xuất bao gồm 2 phần

chính: thay thế cơ chế nhúng văn bản hiện tại bằng SciBERT, một thành tựu xử lý ngôn ngữ trong các bài báo học thuật, và thay thế bộ nhớ hai chiều BiLSTM [14] trong mô hình hiện tại bằng mạng hồi quy (*Recurrent Highway Networks - RHN*) [20].

Chương 4 tập trung vào hướng nghiên cứu kết hợp lọc nội dung và lọc dựa vào đồ thị, trong đó trình bày chi tiết phương pháp xây dựng một mô hình khuyến nghị trích dẫn mới bằng cách kết hợp các thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên SciBERT và đồ thị biểu diễn các liên kết trích dẫn GraphSAGE.

Cuối cùng, phần kết luận nêu các đóng góp chính của tác giả trong luận án và những hướng nghiên cứu tương lai được mở ra từ luận án này.

CHƯƠNG 1. TỔNG QUAN NGHIÊN CỨU VÀ KIẾN THỨC NỀN TẢNG

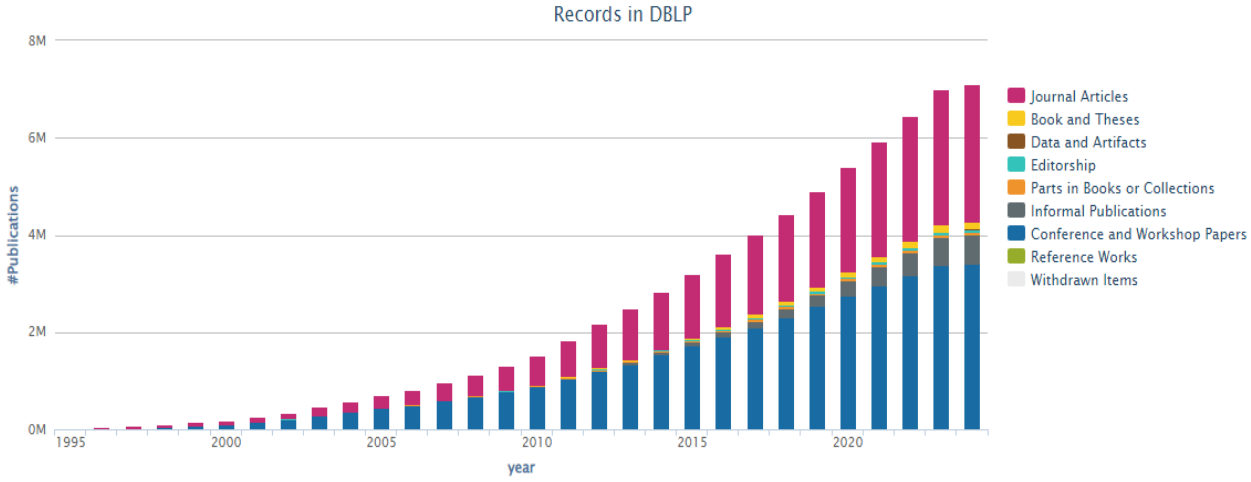
Nội dung trong chương này sẽ bao gồm việc giới thiệu chi tiết về bài toán khuyến nghị trích dẫn cũng như các nghiên cứu liên quan hiện nay để giải quyết cho bài toán này. Chương này cũng trình bày một số các kiến thức nền tảng liên quan được sử dụng cho các phương pháp được đề xuất ở các chương phía sau.

1.1. Giới thiệu

Có thể nói rằng các bài báo hay ấn phẩm khoa học được coi là một trong những nguồn tài nguyên nghiên cứu quan trọng trong cuộc sống chúng ta ngày nay. Bài báo khoa học được công bố đầu tiên có niên đại từ năm 1673 khi Hiệp hội khoa học Hoàng gia Anh[1] xuất bản tập đầu tiên của tạp chí triết học của Hiệp hội Hoàng gia [21]. Kể từ đó, các nhà khoa học đã xuất bản hàng triệu bài báo mô tả các ý tưởng và phát hiện trong các nghiên cứu của họ. Các bài báo không chỉ được sử dụng để truyền đạt kết quả nghiên cứu giữa các nhà khoa học đồng nghiệp mà còn là nguồn tài nguyên chính để học hỏi và theo dõi tiến trình phát triển trong từng lĩnh vực cụ thể. Mặc dù có rất nhiều nguồn thông tin khoa học đa dạng hiện có trên mạng Internet – các bài báo khoa học vẫn đóng vai trò chính trong việc lưu trữ và phổ biến kiến thức của nhân loại và có thể sẽ tiếp tục làm như vậy trong tương lai có thể thấy trước. Ngoài ra, những tiến bộ công nghệ và sự phát triển như vũ bão của Internet đã đẩy nhanh hoạt động nghiên cứu và công bố khoa học lên mức chưa từng có. Trong khi lợi ích cho xã hội là không thể phủ nhận, nhược điểm của khoa học hiện đại là số lượng bài báo khoa học được xuất bản gần đây ngày càng tăng đến mức các nhà khoa học khó theo kịp các kết quả nghiên cứu đã được công bố mới nhất trong lĩnh vực mà mình quan tâm. Một nghiên cứu gần đây báo cáo rằng số lượng bài báo khoa học và kỹ thuật được xuất bản từ năm 2004 đến năm 2014 tăng trưởng với tốc độ trung bình hàng năm là xấp xỉ 10%, đạt gần 4.1 triệu vào năm 2018 [22]. Ví dụ, chỉ riêng trong lĩnh vực khoa học máy tính, hơn 100,000 bài báo mới được xuất bản hàng năm và số bài báo được xuất bản vào năm 2010 nhiều gấp ba lần so với năm 2000 [1]. Xu hướng tương tự có thể được quan sát thấy trong các ngành khác [23]. Thêm một ví

[1] https://royalsocietypublishing.org

dụ khác, trong thư viện trực tuyến y tế số PubMed[2], số lượng ấn phẩm năm 2014 (514 nghìn) nhiều hơn gấp ba lần số lượng xuất bản năm 1990 (137 nghìn) và hơn 100 lần số lượng xuất bản năm 1950 (4 nghìn) [24]. Hình 1.1 cho thấy sự phát triển bùng nổ của các ấn phẩm khoa học được lập chỉ mục trong DBLP[3] từ năm 1995 đến nay. Tốc độ tăng trưởng như vậy gây áp lực lên các nhà khoa học trong việc phải chọn lọc kỹ lưỡng hơn các bài báo họ muốn đọc, vì việc đọc tất cả các ấn phẩm liên quan trở nên không khả thi. Một nghiên cứu được công bố năm 2014 [25] cho thấy vào năm 2012, các nhà khoa học từ các trường đại học Mỹ và Úc ước tính rằng họ đọc trung bình 22 bài báo mỗi tháng, con số này khớp với con số được báo cáo trong một nghiên cứu trước đó từ năm 2005 [6]. Những phát hiện này cho thấy các nhà khoa học có thể đã đạt đến mức giới hạn hiệu suất về số lượng bài báo mà họ có thể xử lý trong một khoảng thời gian nhất định.



Hình 1.1. Sự phát triển của các ấn phẩm khoa học được lập chỉ mục trong DBLP[3] từ năm 1995 đến năm 2024 (hình ảnh lấy từ DBLP[3])

Sự phát triển của công nghệ hiện đại đã góp phần vào sự gia tăng vượt bậc trong xuất bản khoa học, nhưng nó cũng có thể đưa ra các biện pháp khắc phục trình trạng quá tải cho các nhà khoa học. Đặc biệt, với các công cụ chuyên dụng cho tìm kiếm và lập chỉ mục các cơ sở dữ liệu khoa học lớn, có thể tạo điều kiện thuận lợi cho các nhà khoa học truy cập vào các bài báo đã xuất bản và giúp họ tận dụng tốt nhất thời gian để theo dõi tiến bộ trong lĩnh vực của mình. Một số công cụ tìm kiếm như vậy đang được sử dụng

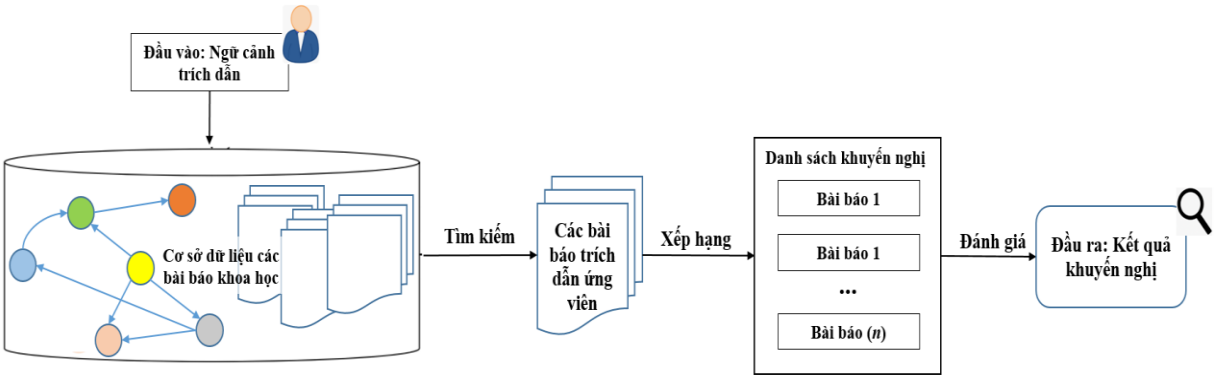
[2] https://www.ncbi.nlm.nih.gov/pubmed
[3] https://dblp.org/statistics/recordsindblp.html

rộng rãi ngày nay, bao gồm cả Google Scholar[4], Microsoft Academic[5] và Semantic Scholar[6]. Các hệ thống trực tuyến này lập chỉ mục các phần thông tin khác nhau được trích xuất từ văn bản bài viết hoặc siêu dữ liệu của nó, bao gồm từ khóa, dữ liệu tác giả, dữ liệu xuất bản và trích dẫn, giúp người dùng không chỉ truy xuất các bài viết phù hợp nhất cho truy vấn của họ mà còn có thể điều hướng theo ngữ nghĩa toàn bộ bộ sưu tập các bài viết cũng như giới thiệu các bài viết để đọc. Với những tiến bộ đáng kinh ngạc gần đây về trí tuệ nhân tạo và học máy, đặc biệt là với các thành quả trong xử lý ngôn ngữ tự nhiên (*Natural Language Processing, NLP*), các công cụ tìm kiếm đã bắt đầu kết hợp các kỹ thuật phức tạp hơn để xử lý ngữ nghĩa của các bài báo khoa học. Ví dụ như Semantic Scholar hiện sử dụng mô hình học máy để xác định những trích dẫn nào trong một bài báo khoa học nhất định có ảnh hưởng nhất cho bài viết đó [26], một tính năng có thể giúp các nhà khoa học tìm thấy các bài viết liên quan dễ dàng hơn. Tuy nhiên những hệ thống tìm kiếm trực tuyến này vẫn tồn tại nhưng hạn chế nhất định. Đó là trong nhiều trường hợp người sử dụng đưa vào hệ thống tìm kiếm từ khóa truy vấn quá rộng. Vì vậy, kết quả trả về có thể là những bài báo ít liên quan đến những gì mà nhà nghiên cứu thực sự cần. Hoặc trong một số trường hợp, từ khóa truy vấn quá hẹp nên chỉ có thể tìm thấy các bài báo liên quan một chút, hoặc tệ hơn, không có gì cả. Các nhà nghiên cứu lại phải dành nhiều thời gian và công sức để tìm kiếm các thông tin liên quan bài viết để trích dẫn trong bài viết của họ. Ngoài ra, công việc này còn yêu cầu các kỹ năng như đưa ra các từ khóa phù hợp để truy vấn và khả năng để đánh giá thủ công mức độ phù hợp với bối cảnh trích dẫn cụ thể của mỗi tài liệu được công cụ tìm kiếm đưa ra.

Một trong các cách để cải thiện tình trạng này là đề xuất giải pháp cho các hệ thống khuyến nghị trích dẫn tự động, và do đó chủ đề này đã nhận được sự quan tâm rộng rãi từ cộng đồng các nhà nghiên cứu. Từ năm 2002 McNee và cộng sự [5] đã giới thiệu bài toán khuyến nghị trích dẫn cho các tài liệu nghiên cứu. Khuyến nghị trích dẫn là nhiệm vụ tự động xác định từ một tập hợp các bài báo khoa học ra một hoặc một vài bài báo có thể hoặc lẽ ra phải được trích dẫn trong một bài báo khác hoặc trong một bản

[4] https://scholar.google.com
[5] https://academic.microsoft.com
[6] https://www.semanticscholar.org

thảo chưa được xuất bản. Việc giải quyết nhiệm vụ này có tiềm năng trực tiếp nâng cao chất lượng các công bố khoa học, vì nó có thể đảm bảo rằng tất cả các công việc có liên quan trước đó đã được xác định và bối cảnh hóa một cách phù hợp. Về mặt hình thức, một mô hình khuyến nghị trích dẫn có thể được định nghĩa như sau (Hình 1.2 [9]): đầu vào là truy vấn của người dùng (q) và cơ sở dữ liệu các ấn phẩm khoa học đã xuất bản $D(d_1, d_2, \dots, d_n)$, hệ thống khuyến nghị trích dẫn trước hết là mô hình hóa truy vấn của người dùng và tập hợp tài liệu. Sau đó, nó tính toán độ tương thích trích dẫn của một bài báo cụ thể cho truy vấn nhất định đối với tất cả các cặp truy vấn và bài báo, tức là $R(q, d_1), \dots, R(q, d_n)$, trong đó biểu thị bằng (R) ước tính dựa trên các thuật toán đề xuất. Tiếp đến, hệ thống sẽ tiếp tục đưa ra danh sách trích dẫn ứng viên cuối cùng $C(c_1, c_2, \dots, c_k)$ được xếp hạng từ giá trị lớn nhất đến nhỏ nhất về mặt tiện ích dự đoán hoặc các yếu tố, cơ chế khuyến nghị bổ sung. Cuối cùng đầu ra của hệ thống là danh sách tham chiếu thực đã được xếp hạng theo từng truy vấn của người dùng.



Hình 1.2. Sơ đồ luồng của mô hình khuyến nghị trích dẫn [9]

Phần sau đây sẽ trình bày chi tiết và rõ ràng một số khái niệm quan trọng liên quan đến bài toán khuyến nghị trích dẫn, được sử dụng xuyên suốt luận án này.

1.1.1. Trích dẫn và tài liệu tham khảo

Một "*trích dẫn*" (*citation*) là mối liên kết giữa một tài liệu trích dẫn và một tài liệu được trích dẫn tại một vị trí cụ thể trong tài liệu trích dẫn. Vị trí này được gọi là "*điểm đánh dấu trích dẫn*" (*citation marker*), thường biểu thị bằng ký hiệu như [1], [2], v.v…như ở Hình 1.3. Phần văn bản cần được hỗ trợ bởi trích dẫn đó được gọi là "*ngữ cảnh trích dẫn*" (*citation context*). Ngữ cảnh trích dẫn này là rất quan trọng, vì nó cung

cấp thông tin cho thấy tại sao một tài liệu được trích dẫn và cách nó đóng góp vào nội dung hiện tại.

Trong khi hai thuật ngữ "*trích dẫn*" và "*tài liệu tham khảo*" (*reference*) đôi khi được sử dụng thay thế nhau, cần lưu ý rằng chúng có sự khác biệt rõ rệt. "*Trích dẫn*" là liên kết trực tiếp xuất hiện trong văn bản, tại các điểm đánh dấu như [] để chỉ đến tài liệu cụ thể, trong khi "*tài liệu tham khảo*" là danh sách tài liệu đầy đủ, thường xuất hiện ở phần cuối của bài báo hoặc tài liệu khoa học.

1.1.2. Liên kết trích dẫn của các bài báo khoa học

Liên kết trích dẫn của các bài báo khoa học (*citation link*) là mối quan hệ giữa hai hoặc nhiều bài báo khoa học, trong đó một bài báo (bài trích dẫn) tham chiếu đến một bài báo khác (bài được trích dẫn) như là một nguồn tham khảo hoặc bằng chứng để hỗ trợ cho lập luận, ý tưởng, hoặc kết quả nghiên cứu của nó. Đây là một phần quan trọng của việc công bố khoa học vì nó tạo ra một mạng lưới các nghiên cứu liên kết với nhau, cho phép các nhà nghiên cứu theo dõi và xây dựng dựa trên kiến thức hiện có. Liên kết trích dẫn được biểu thị bằng mũi tên liên kết giữa 2 bài báo như ở Hình 1.3.

1.1.3. Ngữ cảnh trích dẫn và biểu diễn trừu tượng

Ngữ cảnh trích dẫn có thể được chuyển đổi thành một biểu diễn trừu tượng để sử dụng trong các hệ thống khuyến nghị, nhằm hỗ trợ tìm kiếm và đối sánh các tài liệu phù hợp. Ví dụ, trong các mô hình khuyến nghị trích dẫn, ngữ cảnh trích dẫn có thể được biểu diễn dưới dạng vectơ nhúng (*embedding vector*) hoặc thông qua các mô hình chuyển đổi (*translation models*). Các biểu diễn này giúp máy tính có thể phân tích, đối sánh và khuyến nghị tài liệu một cách chính xác hơn bằng cách dựa vào nội dung ngữ nghĩa thay vì chỉ dựa vào từ khóa cụ thể.

1.1.4. Bài báo ứng viên trích dẫn

Trong bài toán khuyến nghị trích dẫn, các "*tài liệu có thể trích dẫn*" (*citeable documents*) hay còn gọi là "*ứng viên trích dẫn*" (*citation candidates*) là các tài liệu có khả năng được khuyến nghị để bổ sung vào danh sách tài liệu hiện tại. Các tài liệu này thường là bài báo khoa học, nhưng cũng có thể là sách hoặc các tài liệu học thuật khác.

Tất cả tài liệu ứng viên này cần được hệ thống phân tích dựa trên ngữ cảnh trích dẫn đã được biểu diễn trừu tượng để tìm ra tài liệu phù hợp nhất.

1.1.5. Người dùng của hệ thống khuyến nghị trích dẫn

Người dùng của hệ thống khuyến nghị trích dẫn trong môi trường học thuật thường là các nhà nghiên cứu (*researchers*), học giả (*scholars*), hay nhà khoa học (*scientists*). Họ là những người sử dụng hệ thống để tìm kiếm các tài liệu hỗ trợ cho bài viết của mình. Trong luận án này, các thuật ngữ "*người dùng*" (*users*) và "*nhà nghiên cứu*" sẽ được sử dụng thay thế cho nhau trong ngữ cảnh mô tả người dùng của hệ thống.

1.1.6. Nội dung và siêu dữ liệu của bài báo khoa học

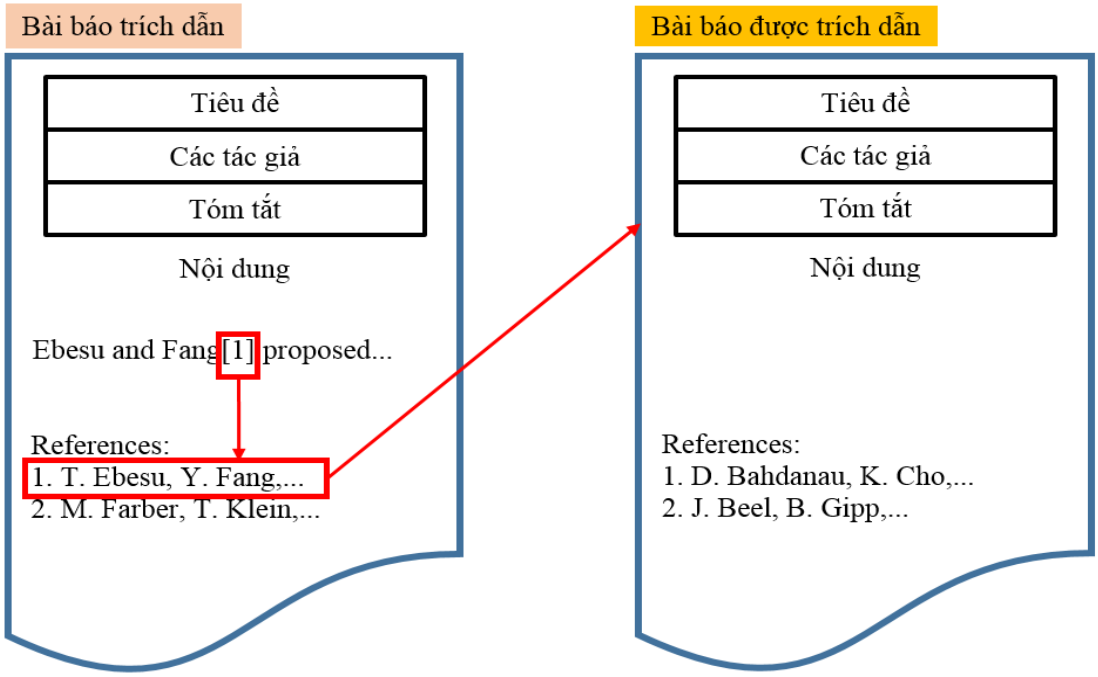
Một tài liệu học thuật bao gồm cả nội dung và siêu dữ liệu. Nội dung của một bài báo khoa học là phần chính, nơi mà tác giả trình bày chi tiết các kết quả nghiên cứu, thí nghiệm hoặc phân tích của họ. Đây là nơi tác giả cung cấp các dữ liệu, lập luận, và bằng chứng để chứng minh cho những kết luận mà họ đưa ra. Siêu dữ liệu của một bài báo khoa học thường bao gồm tên tác giả, năm xuất bản, địa điểm công bố (ví dụ như tạp chí hay hội nghị).v.v... Đây là các thông tin quan trọng giúp các mô hình khuyến nghị không chỉ xem xét nội dung mà còn phân tích các yếu tố khác để đưa ra quyết định trích dẫn hợp lý.

1.1.7. Khuyến nghị trích dẫn cục bộ và khuyến nghị trích dẫn toàn cục

Có hai loại khuyến nghị trích dẫn phổ biến trong các hệ thống hiện đại [6][27]:

Khuyến nghị trích dẫn cục bộ (*local citation recommendation*): Đây là phương pháp khuyến nghị sử dụng một đoạn nhỏ của tài liệu làm ngữ cảnh trích dẫn, chẳng hạn như một câu hoặc một đoạn văn xung quanh điểm đánh dấu trích dẫn (*citation marker*). Loại khuyến nghị này thường được gọi là "*nhận biết ngữ cảnh*", vì nó chỉ dựa vào ngữ cảnh nhỏ tại một điểm cụ thể trong văn bản.

Khuyến nghị trích dẫn toàn cục (*global citation recommendation*): Đây là phương pháp không phụ thuộc vào ngữ cảnh cụ thể mà dựa vào toàn bộ nội dung bài báo hoặc phần tóm tắt để khuyến nghị tài liệu trích dẫn. Đây là loại khuyến nghị "*không nhận biết ngữ cảnh*", vì nó sử dụng toàn bộ tài liệu hoặc phần tóm tắt làm đầu vào cho mô hình.

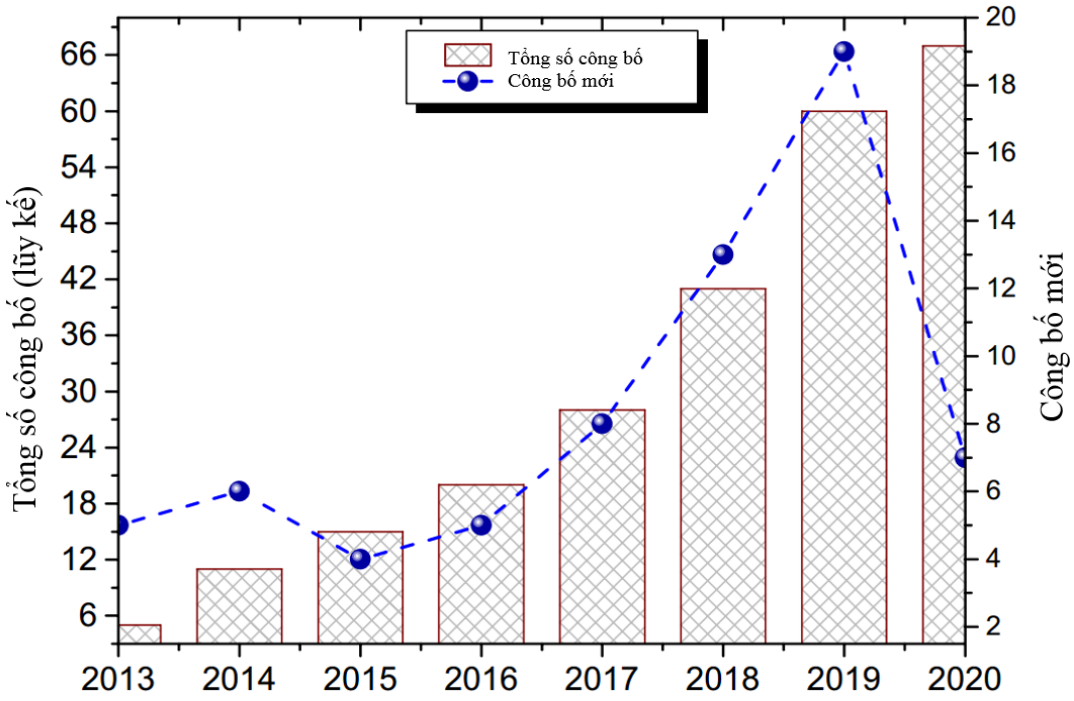


Hình 1.3. Minh hoạ về trích dẫn trong một bài báo khoa học

Trong những năm gần đây, đã có một số lượng lớn các nghiên cứu cung cấp các phương pháp khuyến nghị trích dẫn các bài báo đưa ra những giải pháp liên quan đến một chủ đề nghiên cứu cụ thể. Trong phần tiếp theo sẽ trình bày các cách tiếp cận được công bố trong những năm gần đây để giải quyết bài toán khuyến nghị trích dẫn.

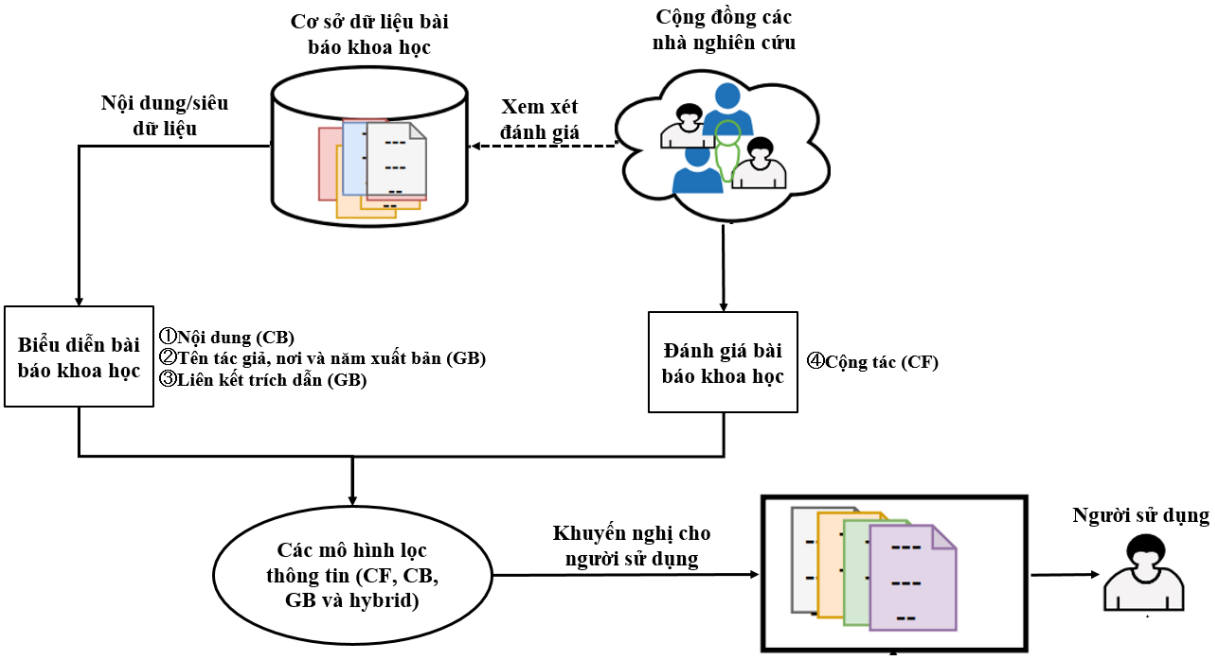
1.2. Tổng quan các nghiên cứu liên quan

Theo khảo sát của Ali và các cộng sự [8] lấy từ 2 cơ sở dữ liệu là Google Scholar và Web of Science (Hình 1.4), thì số lượng các mô hình học sâu cho bài toán khuyến nghị được tăng đều từ năm 2013 cho đến năm 2020, và tính đến năm 2020 là có khoảng 67 mô hình học sâu cho bài toán khuyến nghị trích dẫn đã được công bố.



Hình 1.4. Các mô hình khuyến nghị trích dẫn được công bố hàng năm [8]

Về bản chất, các mô hình khuyến nghị trích dẫn đều lọc thông tin của bài báo và sự ưu tiên của người dùng để đưa ra khuyến nghị phù hợp như đã biểu diễn ở Hình 1.5. Bởi vì cơ sở dữ liệu bài báo khoa học có rất nhiều thông tin, cho nên với mỗi bài báo khoa học, các mô hình khuyến nghị trích dẫn thông thường sẽ có 4 loại thông tin cần quan tâm là: (1) nội dung của bài báo; (2) tên tác giả, năm và nơi công bố (thông tin này gọi là siêu dữ liệu); (3) liên kết trích dẫn của bài báo; (4) đánh giá bài báo trong cộng đồng học thuật. Các khảo sát hay nghiên cứu gần đây [8] [27] đã phân loại các mô hình khuyến nghị trích dẫn dựa các thông tin của bài báo mà mô hình sử dụng: lọc cộng tác (*collaborative filtering, CF*), lọc nội dung (*content-based filtering, CB*), lọc dựa vào đồ thị (*graph-based filtering, GB*) và mô hình kết hợp (*hybrid*). Bởi vì chủ đề nghiên cứu của luận án bao trùm lên các phương pháp này, nên trong phần này sẽ trình bày các thành tựu nghiên cứu gần đây theo các hướng tiếp cận này.



Hình 1.5. Mô hình khuyến nghị trích dẫn trong đó thông tin bài báo và hồ sơ người dùng được khai thác bằng các phương pháp lọc thông tin khác nhau [8] [27]

1.2.1. Mô hình lọc cộng tác

Trong các mô hình lọc cộng tác, đánh giá hoặc xếp hạng từ bạn bè và đồng nghiệp của người dùng được khai thác để dự đoán tài liệu trích dẫn. Dữ liệu về xếp hạng và sở thích của người dùng giúp xác định những cá nhân có mối quan tâm tương đồng. Mô hình sau đó đề xuất các bài báo dựa trên sở thích của những người dùng có xu hướng lựa chọn tương tự. Ma trận xếp hạng được sử dụng để dự đoán các mục phù hợp cho những người dùng mới. Trong một mô hình lọc cộng tác điển hình có danh sách (m) người dùng $U = \{u_1, u_2, u_3, \ldots, u_m\}$ và (n) bài báo $P = \{p_1, p_2, p_3, \ldots, p_n\}$ [28]. Mỗi người dùng có một danh sách các bài viết được xếp hạng rõ ràng hoặc ngầm định thể hiện sự quan tâm của họ. Bằng cách này, ma trận xếp hạng của người dùng sẽ được tạo ra để biểu diễn sở thích của người dùng. Để đưa ra dự đoán cho các bài báo chưa được nhìn thấy, mô hình sử dụng các kỹ thuật như lọc cộng tác dựa trên người dùng hoặc hay dựa trên các bài báo. Theo cách này, mô hình CF tìm "*hàng xóm gần nhất*" cho người dùng mới và đề xuất các bài báo bằng cách khai thác xếp hạng của hàng xóm gần nhất của họ.

Vì thông tin xếp hạng đánh giá các bài báo của cộng đồng nghiên cứu là khan hiếm hơn so với các lĩnh vực khác (ví dụ phim ảnh hoặc nhà hàng), do đó không có nhiều

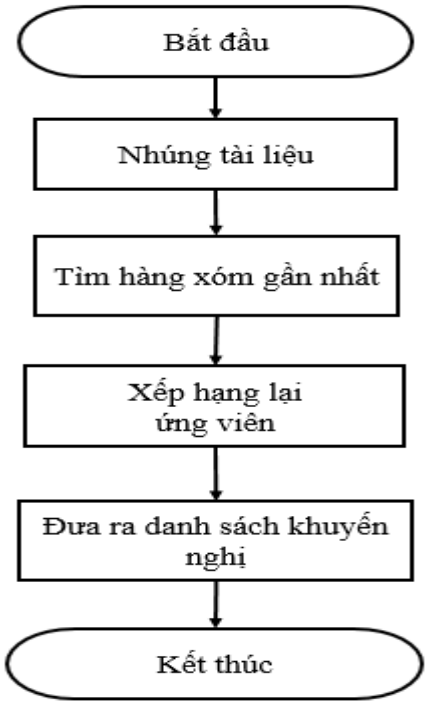
kết quả nghiên cứu đi theo phương pháp lọc cộng tác. Một trong những nghiên cứu đó là của nhóm Liu và cộng sự [29] phát triển một mô hình khuyến nghị trích dẫn bằng cách sử dụng phương pháp lọc cộng tác dựa trên trích dẫn tương đồng. Hai bài báo được coi là có mối liên quan với nhau nếu chúng cùng trích dẫn một bài báo chung. Xem xét cơ sở lý luận này, một kỹ thuật khai thác kết hợp được sử dụng để xác định các tài liệu trích dẫn cùng xảy ra. Mô hình này biểu diễn mỗi bài báo trích dẫn dựa trên các bài báo trích dẫn khác. Sau đó, các bài báo được biểu diễn theo theo cặp được so sánh để tính toán sự tương đồng giữa các bài báo được trích dẫn. Cuối cùng, các trích dẫn của các bài báo trích dẫn tương tự như bài báo mục tiêu được sử dụng để dự đoán số lần trích dẫn của bài báo mục tiêu. Tương tự, nhóm của Bansal [30] công bố một mô hình tạo ra các dự đoán bài báo để trích dẫn trong nhiệm vụ lọc cộng tác bằng cách sử dụng các đơn vị hồi quy có cổng (*Gated Recurrent Units, GRU*) cho thông tin văn bản của các tài liệu nghiên cứu. Sugiyama và Kan [31] đã xác định các bài báo trích dẫn tiềm năng bằng cách khai thác ma trận trích dẫn bài báo để hình thành một vùng lân cận. Mô hình này đã sử dụng hệ số tương quan Pearson (*Pearson correlation coefficient*) để tính toán độ tương tự giữa vectơ của ngữ cảnh trích dẫn và bài báo trích dẫn mục tiêu. Galke và các cộng sự [32] sử dụng ma trận xếp hạng $X \in \{0, 1\}^{m \times n}$ trong đó bản thân các bài báo được coi là người sử dụng hơn là tác giả của chúng. Lý do đưa ra đề xuất này là vì trên thực tế, một tác giả có thể tham gia viết nhiều bài báo viết thuộc các lĩnh vực nghiên cứu khác nhau và tất cả các tác giả cộng tác của một bài báo đều phải nhận được những khuyến nghị trích dẫn tương tự.

Các mô hình lọc cộng tác có thể tạo ra kết quả chất lượng, tuy nhiên, các mô hình này gặp phải các vấn đề về tỉ lệ xếp hạng sớm, thưa thớt, quyền riêng tư và tấn công *shilling* [33] [34]. Ngoài ra, các mô hình này đều đã không khai thác nội dung hay thông tin phụ trợ của bài báo, cho nên các bài báo mới công bố và chưa được trích dẫn hầu như sẽ không được khuyến nghị [35].

1.2.2. Mô hình lọc nội dung

Khi tìm các bài báo hoặc tài liệu trích dẫn thì yếu tố quan trọng nhất là nội dung của tài liệu có phù hợp hay không. Mô hình lọc nội dung bao gồm việc đề xuất các bài

báo học thuật để trích dẫn bằng cách phân tích nội dung của một tài liệu hay truy vấn và tìm các tài liệu tương tự. Dưới đây là tổng quan ngắn gọn về cách thức hoạt động của mô hình này (Hình 1.6).



Hình 1.6. Sơ đồ khối của mô hình lọc nội dung

(1) Nhúng tài liệu: Mô hình lọc nội dung bắt đầu bằng cách nhúng tài liệu truy vấn vào không gian vectơ. Điều này thường được thực hiện bằng cách sử dụng các mô hình như Doc2vec [36], stuct2vec [37] hay Hyperdoc2vec [38] để chuyển đổi văn bản thành vectơ số biểu diễn cho nội dung của bài báo.

(2) Tìm hàng xóm gần nhất: Sau khi tài liệu được nhúng, hệ thống sẽ xác định các hàng xóm gần nhất của nó trong không gian vectơ. Những hàng xóm này là những tài liệu khác có nội dung tương tự và được coi là các trích dẫn tiềm năng.

(3) Xếp hạng lại ứng viên: Các ứng viên (hàng xóm gần nhất) sau đó được xếp hạng lại bằng cách sử dụng mô hình phân biệt đối xử. Mô hình này được huấn luyện để phân biệt giữa các trích dẫn được quan sát (trích dẫn thực tế) và những trích dẫn không được quan sát (không được trích dẫn).

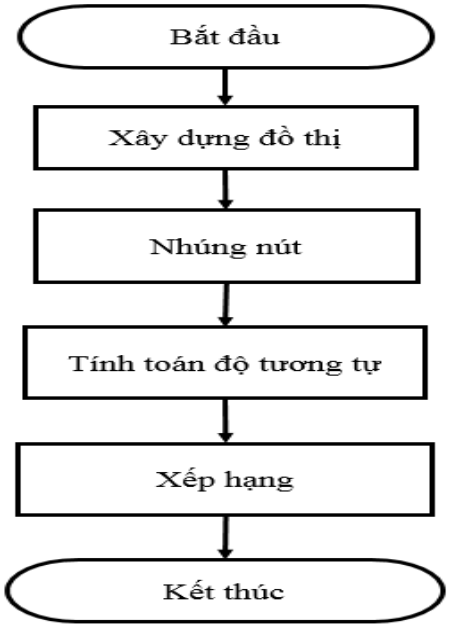
(4) Khuyến nghị: Các bài viết được xếp hạng hàng đầu từ quy trình này sau đó sẽ được đề xuất làm trích dẫn tiềm năng cho tài liệu truy vấn.

Mô hình lọc nội dung tập trung hoàn toàn vào nội dung của bài báo mà không yêu cầu siêu dữ liệu như tên tác giả hay năm công bố. Nó được thiết kế để cải thiện mức độ liên quan của các trích dẫn được đề xuất và có thể đặc biệt hữu ích khi siêu dữ liệu không đầy đủ hoặc không có. Nhóm của Kong [39] đã phát triển một hệ thống khuyến nghị trích dẫn tên là VOPRec (*vector representation learning of paper*). Khái niệm cơ bản của VOPRec là thực tế trong mạng trích dẫn, các bài báo khoa học được biểu diễn bằng cách sử dụng vectơ. Khi đó, sự giống nhau của các bài báo khoa học có thể được đánh giá thông qua sự tương đồng của các vectơ. Yang và Ma [40] đã xây dựng một thuật toán nhúng văn bản mới là DocCit2Vec, cùng với khái niệm mới về "bối cảnh cấu trúc (*structural context*)", để giải quyết các vấn đề là các tài liệu được khuyến nghị có thể đã được người dùng biết hoặc chỉ liên quan đến xung quang bối cảnh chứ không phải các ý tưởng khác được thảo luận trong bản thảo bài báo. Huang và các cộng sự [41] đã đề xuất một phương pháp "*chuyển*" các tài liệu nghiên cứu thành tài liệu tham khảo. Bằng cách coi các trích dẫn và ngữ cảnh của chúng từ các bài báo hiện có là dữ liệu song song được viết bằng hai "*ngôn ngữ*" khác nhau, họ áp dụng mô hình "*dịch thuật*" để tạo mối quan hệ giữa hai "*từ vựng*" này. Họ tiếp tục mở rộng nghiên cứu [42] bằng cách bổ sung thêm ngữ nghĩa của ngữ cảnh trích dẫn và các tài liệu được trích dẫn. Cuối cùng, dựa trên khoảng cách ngữ nghĩa trong không gian vectơ sẽ đưa ra khuyến nghị trích dẫn phù hợp. Ebesu và Fang [10], Färber và cộng sự [11] đã đề xuất mô hình mạng trích dẫn nơ-ron NCN mà trong đó sử dụng mạng nơ-ron trễ (*Time-Delay Neural Networks, TDNN*) và mạng nơ-ron hồi quy (*Recurrent Neural Networks, RNN*) để xây dựng bộ mã hóa-giải mã nhằm trích xuất thông tin từ bối cảnh trích dẫn và các bài báo ứng viên trích dẫn. Sau đó mô hình này sử dụng hàm phân hạng nổi tiếng Okapi BM25 [43] để sắp xếp lại danh sách các bài báo ứng viên này. Nhóm của Färber [44] tiếp tục đề xuất một phương pháp tiếp cận di truyền để giải quyết vấn đề khuyến nghị trích dẫn. Các tác giả đã kết hợp phương pháp nhúng và phương pháp truy xuất thông tin bằng cách sử dụng trọng số điểm thích nghi (*fitness score*) để đưa ra các khuyến nghị trích dẫn. Nhóm nghiên cứu của Gu

[16] đưa ra một hệ thống khuyến nghị trích dẫn hai giai đoạn. Trong giai đoạn tìm nạp trước, mô hình sử dụng hệ thống truy xuất bài báo trên nhúng văn bản, trong đó bộ mã hóa văn bản *siamese* tính toán mỗi vectơ cho mỗi bài báo trong cơ sở dữ liệu. Truy vấn của người sử dụng sau đó được ánh xạ vào cùng một không gian nhúng để truy xuất (*K*) lân cận gần nhất của vectơ truy vấn. Để mã hóa các truy vấn và bài báo có độ dài khác nhau theo cách hiệu quả về bộ nhớ, mô hình sử dụng cơ chế tự chú ý [45]. Trong giai đoạn sắp xếp lại, họ đã sử dụng SciBERT [18] để xếp hạng lại các bài báo trích dẫn ứng viên đã được truy xuất ở giai đoạn tìm nạp trước.

1.2.3. Mô hình lọc dựa vào đồ thị

Lọc dựa vào đồ thị là một cách tiếp cận phổ biến với một số ứng dụng trong nhiều lĩnh vực, chẳng hạn như phân tích mạng các xu hướng trong nội dung xã hội [46]. Trong vài năm qua, nhiều công trình về khuyến nghị dựa trên mạng nơ-ron đồ thị (*Graph Neural Networks*, *GNN*) đã được công bố [47] [48]. Ưu điểm của GNN là nó cung cấp các công cụ mạnh mẽ và có hệ thống để khám phá các mối quan hệ đa dạng đã được chứng minh là có lợi cho các hệ thống khuyến nghị. Mô hình lọc dựa vào đồ thị tận dụng cấu trúc của mạng liên kết trích dẫn để khuyến nghị các bài báo liên quan. Mô hình này sẽ thực hiện các bước như giải thích Hình 1.7.



Hình 1.7. Sơ đồ khối của mô hình lọc dựa vào đồ thị

(1) Xây dựng đồ thị: Một mạng trích dẫn được xây dựng trong đó các nút đại diện cho các bài báo và các cạnh đại diện cho các liên kết trích dẫn giữa chúng.

(2) Nhúng nút: Để nắm bắt thông tin cấu trúc của mạng liên kết trích dẫn, các bài báo được nhúng vào không gian vectơ bằng cách sử dụng các mạng nơ-ron đồ thị như mạng tích chập đồ thị (*Graph Convolutional Networks, GCN*) [49], GraphSAGE [19], LightGCN [50], HGN [51], GAT [52], GGNN [53] hay các mạng không đồng nhất [54] [55] [56] [57].

(3) Tính toán độ tương tự: Độ tương tự giữa các vectơ nhúng của bài báo được tính toán để xác định các trích dẫn tiềm năng.

(4) Xếp hạng: Các bài báo được xếp hạng dựa trên điểm tương đồng của chúng và các bài báo được xếp hạng cao nhất được đề xuất làm trích dẫn.

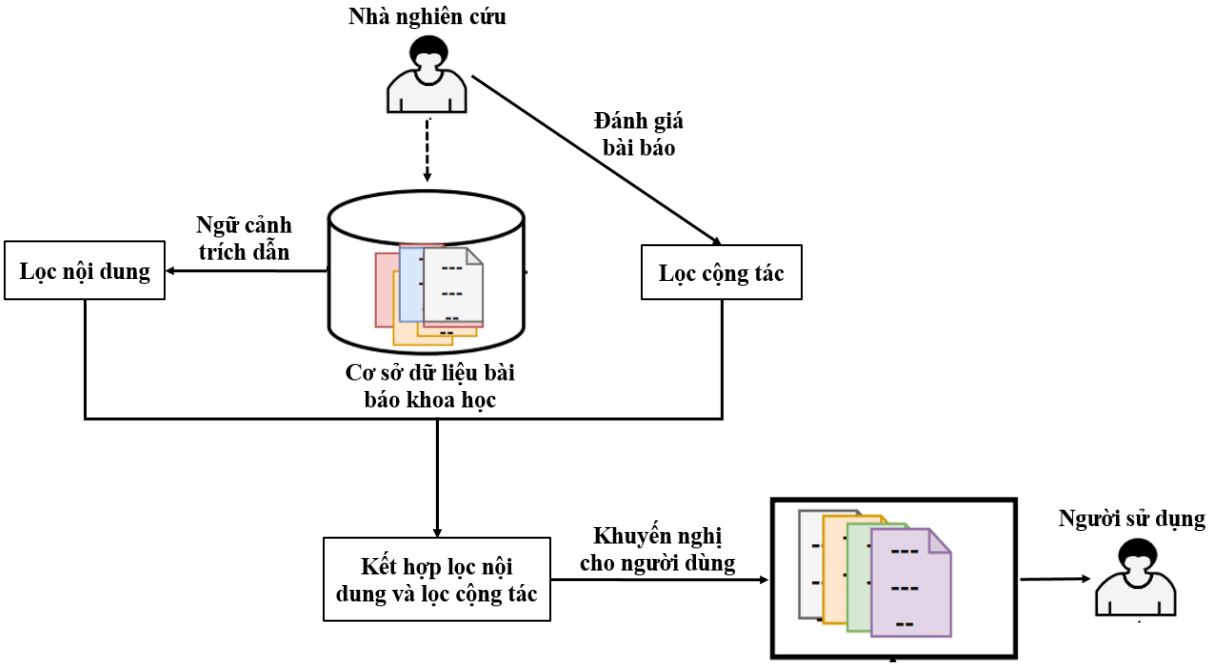
Theo cách tiếp cận này, Chen và cộng sự [58] đã sử dụng các mạng thông tin không đồng nhất bao gồm mạng bài báo-bài báo, bài báo-tác giả và bài báo-thuật ngữ. Nhóm của Yang [59] huấn luyện mô hình mạng thông tin học thuật không đồng nhất chứa các nút bài báo, tác giả và địa điểm công bố. Nhóm của Cai [60] đề xuất biểu diễn mạng thông tin học thuật (*bibliographic network representation*) để lưu trữ thông tin của bài báo, chẳng hạn như thông tin tác giả, tóm tắt bài báo và nơi xuất bản của bài báo để tạo ra các biểu diễn vectơ chiều thấp của các đối tượng là các bài báo. Sau đó, mô hình này sẽ tính toán độ tương đồng của bài viết và tác giả đại diện để tạo ra danh sách khuyến nghị trích dẫn. Nhóm nghiên cứu của Ali [61] phát triển một mô hình khuyến nghị có trọng số được gọi là PR-HNE (*Paper Recommendation based on Heterogeneous Network Embedding*). Mô hình này biểu diễn các thông tin về trạng thái trích dẫn, tình trạng cộng tác của các tác giả, địa điểm công bố, thông tin được gán nhãn và mức độ liên quan theo chủ đề để tạo ra các khuyến nghị được cá nhân hóa. Wang và cộng sự [62] xây dựng một mô hình khuyến nghị trích dẫn sử dụng đồ thị biểu diễn 3 thông tin của bài báo: địa điểm công bố, ảnh hưởng chủ đề nghiên cứu và mức độ ưu tiên của cộng đồng học giả. Chen và cộng sự [63] đã công bố một mô hình mạng thông tin không đồng nhất để giải quyết bài toán khuyến nghị trích dẫn. Nhóm nghiên cứu xây dựng một mạng thông tin không đồng nhất có trọng số bao gồm hai loại đỉnh (bài báo và tác giả) kết hợp

với bốn loại mối quan hệ (liên kết ngữ nghĩa, trích dẫn, cách viết và đồng tác giả). Danh sách các bài báo khuyến nghị trích dẫn được tạo ra thông qua sự kết hợp tuyến tính giữa những điểm tương đồng đa phương thức. Guo và cộng sự [64] đã đề xuất mô hình CSCR (*Content-Sensitive for Citation Recommendation*) để thêm các liên kết giữa hai bài báo nếu chúng có nội dung tương tự nhau, điều này có thể làm giảm bớt vấn đề thưa thớt dữ liệu nhưng không làm tăng kích thước của ma trận kề. Họ đã sử dụng doc2vec [36] để tìm các bài báo tương tự nhằm xây dựng mạng liên kết trích dẫn và sử dụng DeepWalk [65] để huấn luyện cách biểu diễn nút từ mạng trích dẫn mở rộng. Pornprasit và cộng sự [66] phát triển thuật toán nhúng mạng trích dẫn mới, ConvCN, để thể hiện mối quan hệ trích dẫn giữa các bài báo. Sau đó, họ đã đề xuất nâng cao các thuật toán khuyến nghị trích dẫn dựa trên đồ thị hiện có bằng cách kết hợp với ConvCN để cải thiện hiệu quả khuyến nghị.

Phương pháp đề xuất trích dẫn dựa trên mạng sử dụng các mối quan hệ liên kết khác nhau để xây dựng các mô hình mạng nhằm tìm hiểu cách biểu diễn ngữ nghĩa của các bài báo, từ đó có thể tính toán tầm quan trọng của từng bài báo trong mạng thông qua các chỉ số dựa trên mạng khác nhau. Nó chuyển đổi vấn đề đề xuất trích dẫn thành vấn đề dự đoán liên kết trong mạng trích dẫn hoặc mạng đồng trích dẫn. Tuy nhiên, phương pháp này tập trung vào các mối quan hệ trích dẫn đến từ các bộ dữ liệu chung để xác nhận những cải tiến của thuật toán và quá chú trọng đến việc đề xuất bài báo trích dẫn. Ngoài ra, nội dung văn bản của cả bài báo trích dẫn và bài báo được trích dẫn, đặc biệt là nội dung của bài báo được trích dẫn, hầu hết đều bị bỏ qua [67].

1.2.4. Mô hình kết hợp

Từ những năm 2020 đã có các công bố đề xuất các mô hình kết hợp các phương pháp lọc (cộng tác, nội dung và đồ thị) để giảm bớt các vấn đề cố hữu của các phương pháp lọc riêng lẻ và mang lại kết quả cải thiện đáng kể. Phần này khảo sát các mô hình tiếp cận áp dụng kết hợp từ hai phương pháp khuyến nghị nói trên. Ví dụ, có thể thấy trong Hình 1.8 [8] [27] rằng phương pháp lọc cộng tác và lọc nội dung được kết hợp để tạo ra một mô hình lai ghép có thể khắc phục các vấn đề thiếu thông tin hay thưa thớt các đánh giá mà mỗi mô hình này gặp phải khi áp dụng độc lập.



Hình 1.8. Mô hình kết hợp lọc nội dung và lọc cộng tác [8] [27]

Một trong những nghiên cứu đáng chú ý gần đây theo hướng này là của nhóm Jeong [15]. Mô hình BERT-GCN mà họ đã công bố là một mô hình kết hợp giữa phương pháp là lọc nội dung và lọc dựa vào đồ thị: sử dụng BERT [68] để biểu diễn dữ liệu văn bản (ngữ cảnh trích dẫn, tiêu đề, tóm tắt của bài báo) và mạng tích chập đồ thị GCN [49] để biểu diễn siêu dữ liệu của bài báo (tên tác giả, năm xuất bản, nơi công bố). Medić và Šnajder trong hai nghiên cứu [13] [12] đã đưa ra mô hình DualLCR bao gồm 2 mô đun độc lập là mô đun ngữ nghĩa (*semantic module*) tập trung vào nội dung của bài báo trích dẫn và được trích dẫn và mô đun thông tin học thuật (*bibliographic module*) quan tâm đến các yếu tố khác của bài báo như là tên tác giả, số lượng trích dẫn hàng năm và tổng số trích dẫn lũy kế. Wang và nhóm cộng sự [69] đề xuất mô hình đề xuất trích dẫn theo ngữ cảnh dựa trên mạng bộ nhớ đầu cuối. Mô hình này biểu diễn các bài báo và bối cảnh trích dẫn tương ứng dựa trên bộ nhớ hai chiều BiLSTM. Đặc biệt, mô hình này cũng tích hợp thông tin tác giả và mối quan hệ trích dẫn vào trong vectơ phân tán biểu diễn bối cảnh và bài báo trích dẫn. Sau đó tính toán mức độ liên quan liên tục giữa chúng dựa trên mạng bộ nhớ nhiều lớp tính toán. Nhóm nghiên cứu của Cai [70] đã đưa ra mô hình GLNNR (*Global-Local Neighborhoods based Network Representation*) trong đó tích hợp cấu trúc mạng (là các liên kết trích dẫn) và thông tin phi cấu trúc (tên tác giả, thời

gian và địa điểm công bố) để thu được biểu diễn nút tốt hơn trong mạng nơ-ron đồ thị trích dẫn. Nhóm nghiên cứu của Färber [71] giới thiệu C-Rex[7], một hệ thống trực tuyến đưa ra khuyến nghị trích dẫn theo ngữ cảnh dựa sử dụng mạng nơ-ron trích dẫn kết hợp với hàng triệu ấn phẩm từ Microsoft Academic Graph[8]. Đây là một trong những hệ thống khuyến nghị trích dẫn theo ngữ cảnh trực tuyến đầu tiên và cũng là hệ thống đầu tiên không chỉ kết hợp phương pháp đề xuất học sâu mà còn kèm theo cả các giải thích để giúp người dùng hiểu rõ hơn lý do tại sao bài báo được khuyến nghị. Zhang và Ma đã đề xuất 2 khuyến nghị trích dẫn tên là MP-BERT4CR [72] và MP-BERT4REC [73]. Cả 2 mô hình này đều sử dụng trình xử lý ngôn ngữ tự nhiên BERT kết hợp với các mục tiêu của bộ ba đa phù hợp (*Multi-Positive Triplet: anchor, positive và negative*) để khuyến nghị các trích dẫn phù hợp cho các ngữ cảnh truy vấn. Ali và các cộng sự [74] công bố một mô hình nhúng mạng được gọi là GCR-GAN (*Global Citation Recommendation employing Generative Adversarial Network*). Mô hình nay khai thác mạng thông tin học thuật không đồng nhất (*Heterogeneous Bibliographic Network, HBN*) để tạo ra các đề xuất trích dẫn được cá nhân hóa. Đặc biệt, mô hình này khai thác các mối quan hệ ngữ nghĩa giữa các đối tượng trong HBN và mô tả sự kết nối trong cấu trúc mạng thông qua việc xây dựng phép nhúng bài báo khoa học bằng Citation-informed Transformers (*SPECTER*). Đồng thời, mạng bộ mã hóa tự động khử nhiễu (*Denoising Auto-encoder networks*) được sử dụng để huấn luyện các biểu diễn đồ thị, đảm bảo tính bảo toàn ngữ nghĩa.

Nhóm của Wang [75] đã phát triển mô hình DeepCite kết hợp giữa BERT, CNN và BiLSTM. Đầu tiên, BERT được sử dụng để trích xuất các vectơ biểu diễn ngữ nghĩa trong văn bản, sau đó CNN và BiLSTM được sử dụng để thu được thông tin cục bộ và thông tin trình tự của ngữ cảnh trong câu và các vectơ văn bản được so khớp để tạo ra các tập ứng viên để trích dẫn. Mô hình RecCite của nhóm Yadav [76] khuyến nghị các ấn phẩm nghiên cứu bằng cách áp dụng kỹ thuật kết hợp giữa mức độ phù hợp của nội dung với mức độ ”*phổ biến*” của bài báo. Mức độ phổ biến là điểm tích lũy đạt được

[7] http://c-rex.org
[8] https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

bằng cách kết hợp một số yếu tố tiềm ẩn bao gồm liên quan đến tác giả (ví dụ: chỉ số *h*), liên quan đến bài báo (ví dụ: số lượng trích dẫn) và liên quan đến tạp chí (ví dụ: hệ số tác động). Nhóm của Xie [77] phát triển mô hình khuyến nghị kết hợp nhúng văn bản và ảnh hưởng (*Joint Text and Influence Embedding, JTIE*) để xem xét cả chất lượng bài báo và nội dung liên quan. Mô hình này được huấn luyện với các bài báo dựa trên các yếu tố cốt lõi bao gồm nội dung, tác giả và địa điểm công bố. Chất lượng của một bài báo mới được đánh giá dựa trên thẩm quyền của tác giả và danh tiếng của địa điểm công bố. Nhóm của Chang [78] giới thiệu CiteSee, một công cụ đọc bài báo tận dụng các hoạt động công bố, đọc và lưu của người dùng để cung cấp các khuyến nghị trực quan được cá nhân hóa. Đầu tiên, CiteSee kết nối bài viết hiện tại với các bối cảnh quen thuộc bằng cách hiển thị các trích dẫn đã biết mà người dùng đã trích dẫn hoặc mở. Tiếp theo, CiteSee giúp người dùng ưu tiên bằng cách đánh dấu các trích dẫn có liên quan nhưng chưa biết dựa trên lịch sử lưu và đọc của họ. Nhóm của Li [79] phát triển mô hình HNTA (*Heterogeneous Network and Temporal Attributes*) khuyến nghị bài viết học thuật dựa trên sự kết hợp giữa mạng không đồng nhất và các thuộc tính thời gian. HNTA bao gồm một mạng lưới không đồng nhất các loại thực thể khác nhau để tính toán độ tương tự giữa hai bài báo, sau đó thuộc tính thời gian được đưa vào như mối quan tâm nghiên cứu của các học giả, được chia thành mối quan tâm tức thời và mối quan tâm liên tục để tính toán sự giống nhau giữa các học giả và bài báo.

1.3. Một số hạn chế của các mô hình khuyến nghị trích dẫn hiện nay

Thông qua việc giới thiệu các nghiên cứu liên quan gần đây cho bài toán khuyến nghị trích dẫn trong mục 1.2, có thể thấy được rằng có nhiều cách tiếp cận khác nhau để giải quyết cho bài toán này, tuy nhiên những phương pháp đó vẫn tồn tại những hạn chế nhất định. Mục tiêu nghiên cứu của luận án là góp phần giải quyết các hạn chế đó.

1.3.1. Hạn chế các mô hình lọc nội dung

Các mô hình khuyến nghị trích dẫn lọc nội dung xem xét nội dung văn bản được thể hiện bằng từ khóa, tiêu đề và tóm tắt của các bài báo trích dẫn và áp dụng các phương pháp trình bày văn bản để thể hiện ngữ nghĩa của những nội dung này, có thể đề xuất tài liệu tương tự và mới nhất bằng cách tính toán độ tương tự của văn bản. Tuy nhiên, cách

tiếp cận này không thể phân biệt được tầm quan trọng của các tài liệu có nội dung tương tự và không xem xét nội dung văn bản được trích dẫn trong các bài báo trích dẫn có thể đưa ra mô tả chi tiết hơn về động cơ trích dẫn [8]. Thêm vào đó, các mô hình này không tận dụng thông tin phụ trợ, điều này có thể giúp tạo ra kết quả mạnh mẽ hơn và được cá nhân hóa tốt hơn [33] [80].

Một trong những mô hình tiêu biểu cho bài toán khuyến nghị trích dẫn theo hướng lọc nội dung là nghiên cứu của Ebesu và Fang [10] công bố vào năm 2017 và nhóm của Färber [11] cải tiến mô hình này vào năm 2020. Cả hai nhóm này đề xuất một kiến trúc bộ mã hóa-giải mã linh hoạt được gọi là mạng nơ-ron trích dẫn NCN để biểu diễn tốt nhất của bối cảnh trích dẫn với mạng nơ-ron trễ TDNN, được tăng cường hơn nữa với cơ chế chú ý và mạng tác giả (*authors network*). Bộ giải mã mạng nơ-ron hồi quy (*recurrent neural network decoder*) tham khảo cách biểu diễn dữ liệu này khi xác định bài báo tối ưu để khuyến nghị dựa trên tiêu đề của nó. Mặc dù mô hình này thu được kết quả tốt hơn so với các mô hình cùng thời trên các bộ dữ liệu RefSeer và arXiv CS, tuy nhiên nó vẫn còn những hạn chế như là chưa sử dụng toàn bộ các thông tin của bài báo (tiêu đề, tác giả, năm xuất bản, nơi công bố.v.v…) để đưa vào huấn luyện cho mô hình. Ngoài ra, mô hình NCN này cũng chưa sử dụng các thành tựu mới nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên để áp dụng nhúng từ (*word embedding*) cho nội dung của các bài báo khoa học.

1.3.2. Hạn chế của các mô hình kết hợp lọc nội dung và lọc cộng tác

Phương pháp lọc cộng tác được kết hợp với phương pháp lọc nội dung nhằm tận dụng siêu dữ liệu hoặc thông tin phụ trợ cho bài báo. Phương pháp kết hợp này đòi hỏi cần phải có các mô hình biểu diễn thông tin hiệu quả cho các loại dữ liệu khác nhau trong mô hình. Tuy nhiên các mô hình kết hợp hiện nay có thể vẫn chưa sử dụng đến các thành tựu tiên tiến hơn của biểu diễn dữ liệu văn bản hoặc thông tin học thuật [8] [9]. Điển hình cho hạn chế này là hai nghiên cứu của Medić và Šnajder [13] [12]. Khuyến nghị trích dẫn cục bộ nhằm mục đích tìm kiếm các bài viết có liên quan đến bối cảnh trích dẫn nhất định. Trong khi hầu hết các phương pháp khuyến nghị trích dẫn cục bộ trước đây biểu diễn ngữ cảnh chỉ sử dụng văn bản xung quanh trích dẫn, thì trong hai nghiên

cứu của Medić và Šnajder [13] [12] đã đề xuất mô hình DualLCR tăng cường trình bày ngữ cảnh bằng thông tin tổng thể bằng cách đưa thêm cả tiêu đề và phần tóm tắt của bài báo trích dẫn vào phần vectơ biểu diễn ngữ cảnh. Mô hình DualLCR này gồm 2 mô đun độc lập là mô đun ngữ nghĩa (*semantic module*) tập trung vào quan hệ ngữ nghĩa trong trích dẫn và mô đun thông tin học thuật (*bibliographic module*) quan tâm đến các yếu tố khác của bài báo như là tên tác giả và số lượng trích dẫn của bài báo. Tuy nhiên trong mô hình này, mô đun ngữ nghĩa vẫn dùng bộ nhớ hai chiều BiLSTM [14]. Với các thành tựu nghiên cứu trong lĩnh vực học sâu hiện nay thì đã có nhiều cải tiến cho BiLSTM mà có thể áp dụng được. Ngoài ra, mô đun ngữ nghĩa cũng chưa áp dụng những kết quả mới nhất trong lĩnh vực xử lý ngôn ngữ tự nhiên. Đó là những điểm hoàn toàn có thể cải tiến cho mô hình DualLCR này.

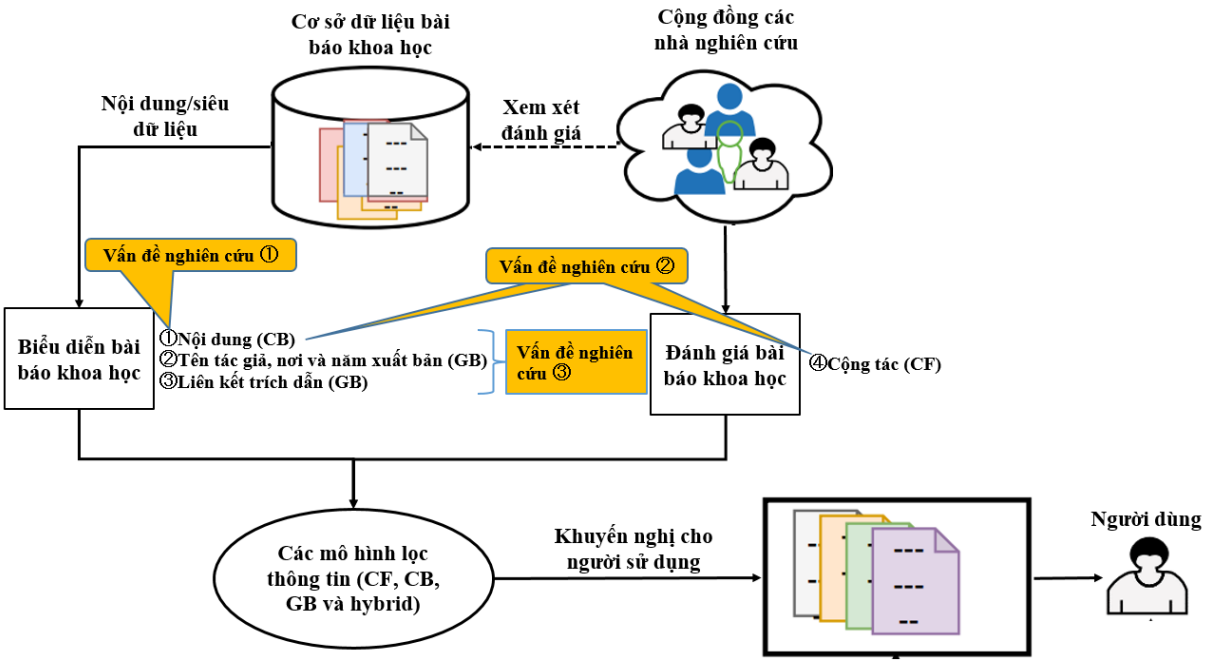
1.3.3. Hạn chế các mô hình kết hợp lọc nội dung và lọc dựa vào đồ thị

Để trích dẫn một bài báo trong nghiên cứu của mình thì ngoài điều kiện phù hợp về nội dung nghiên cứu, các thông tin khác của bài báo như liên kết trích dẫn, tên tuổi của tác giả, nơi công bố và năm xuất bản cũng là các yếu tố cần được cân nhắc. Do đó, các mô hình kết hợp giữa phân tích ngữ nghĩa của nội dung bài báo và mạng liên kết đồ thị trích dẫn cho bài toán khuyến nghị trích dẫn cũng là một trong các hướng nghiên cứu hứa hẹn hiện nay. Tuy nhiên, cũng giống như mô hình kết hợp lọc nội dung kết hợp với lọc cộng tác, các mô hình kết hợp lọc nội dung và lọc dựa vào đồ thị cũng sẽ có những hạn chế với các phương pháp để biểu diễn dữ liệu của bài báo. Ngoài ra, vì cơ sở dữ liệu của bài báo có thể chứa đến hàng triệu bài báo, các mô hình sẽ gặp khó khăn về chi phí tính toán với các mô hình đồ thị liên kết trích dẫn [8] [74]. Một trong những nghiên cứu đáng chú ý gần đây theo hướng này là của nhóm Jeong [15]. Mô hình BERT-GCN mà họ đã công bố là một mô hình dựa trên học sâu là sự kết hợp giữa mô hình BERT [68] để biểu diễn dữ liệu văn bản (ngữ cảnh trích dẫn, tiêu đề, tóm tắt của bài báo) và mạng tích chập đồ thị GCN [49] để biểu diễn siêu dữ liệu của bài báo (tên tác giả, năm xuất bản, nơi công bố). Tuy nhiên mô hình này vẫn có hạn chế nhất định. Mô hình xử lý ngôn ngữ tự nhiên BERT cũng như mạng tích chập đồ thị GCN sau này đều đã có những được

bước cải tiến lớn về hiệu năng, cho phép áp dụng để cải thiện hiệu suất của mô hình khuyến nghị trích dẫn.

1.4. Đặt vấn đề nghiên cứu

Trong lĩnh vực khuyến nghị trích dẫn, mặc dù các mô hình lọc nội dung, lọc cộng tác, lọc dựa trên đồ thị và các mô hình kết hợp mặc dù đã có những thành tựu đáng kể nhưng vẫn tồn tại nhiều hạn chế như đã phân tích ở phần bên trên. Vấn đề nghiên cứu chính của luận án này là làm thế nào để góp phần khắc phục những hạn chế đó và phát triển các mô hình khuyến nghị trích dẫn mạnh mẽ hơn. Để đạt được mục tiêu này, luận án sẽ tập trung nghiên cứu và giải quyết 3 vấn đề như sau:



Hình 1.9. Hình ảnh về 3 vấn đề nghiên cứu của luận án trong bài toán khuyến nghị trích dẫn

1.4.1. Vấn đề nghiên cứu 1

Góp phần giải quyết các hạn chế của các mô hình tiếp cận theo hướng lọc nội dung, luận án đề xuất các giải pháp để nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn NCN. Vấn đề nghiên cứu 1 này sẽ được trình bày chi tiết ở chương 2.

1.4.2. Vấn đề nghiên cứu 2

Góp phần giải quyết các hạn chế của các mô hình tiếp cận theo hướng kết hợp lọc nội dung và lọc cộng tác, luận án đề xuất một mô hình mới tên là RHN-DualLCR, trong đó bao gồm các giải pháp để nâng cao hiệu suất cho mô hình khuyến nghị trích dẫn kép DualLCR. Vấn đề nghiên cứu 2 này sẽ được trình bày chi tiết ở chương 3.

1.4.3. Vấn đề nghiên cứu 3

Góp phần giải quyết các hạn chế của các mô hình tiếp cận theo hướng kết hợp lọc nội dung và lọc dựa vào đồ thị, luận án đề xuất mô hình khuyến nghị trích dẫn mới tên là SciBERT-GraphSAGE bằng cách kết hợp các thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên và đồ thị biểu diễn các liên kết trích dẫn. Vấn đề nghiên cứu 3 này sẽ được trình bày chi tiết ở chương 4.

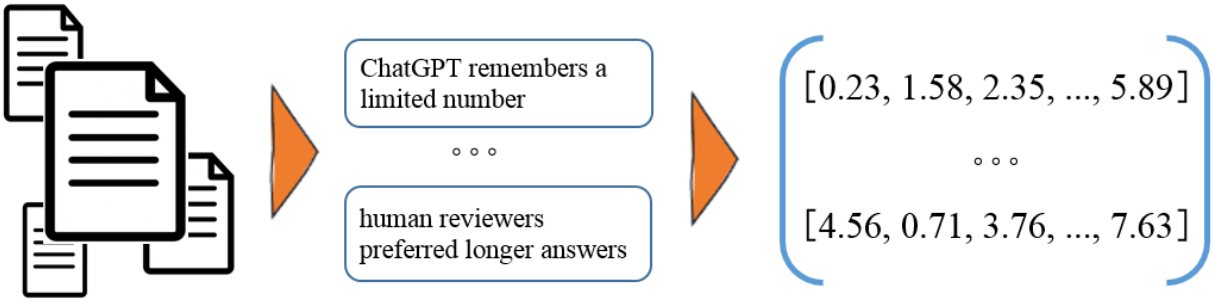
Như vậy có thể thấy rằng 3 mục tiêu nghiên cứu của luận án này có liên kết chặt chẽ với nhau, và bao phủ lên các vấn đề còn tồn tại hiện nay của bài toán khuyến nghị trích dẫn. Hình 1.9 đã minh họa rõ hơn về mối liên kết chặt chẽ của nội dung các chương trong luận án này.

1.5. Các lý thuyết nền tảng

Một số kiến thức nền tảng cần thiết cho việc hiểu phương pháp đề xuất sẽ được trình bày trong phần mục này. Những kiến thức này bao gồm: phép biến đổi nhúng văn bản, họ các mô hình nơ-ron hồi quy, độ liên quan của tài liệu trong các hệ thống truy xuất thông tin và hàm mất mát.

1.5.1. Phép biến đổi nhúng văn bản (*document embedding*)

Trong lĩnh vực học máy hay trí tuệ nhân tạo, phép biến đổi nhúng văn bản (*document embedding*) là một cách thể hiện một đoạn văn bản, chẳng hạn như một câu, một đoạn văn hoặc một văn bản hoàn chỉnh dưới dạng vectơ số. Vectơ nắm bắt ý nghĩa ngữ nghĩa của văn bản, do đó nếu các các văn bản có nội dung tương tự thì sẽ có vectơ biểu diễn tương tự. Điều này cho phép so sánh, tìm kiếm, phân cụm và phân tích văn bản dựa trên ý nghĩa của chúng. Hình 1.6 minh họa mục tiêu của phép đổi nhúng văn bản hay được sử dụng trong xử lý ngôn ngữ tự nhiên.



Hình 1.10. Minh họa phép biến đổi nhúng văn bản

Có nhiều cách để thực hiện phép biến đổi nhúng văn bản, nhưng chúng có thể được phân chia thành hai loại: dựa trên từ (*word-based*) và dựa trên văn bản (*document-based*). Các phương pháp dựa trên từ sử dụng các phần nhúng từ, là các vectơ biểu thị ý nghĩa của các từ riêng lẻ và kết hợp chúng để có được phần nhúng văn bản. Các phương pháp dựa trên văn bản sử dụng các mô hình đã được huấn luyện để mã hóa toàn bộ văn bản cùng một lúc và tạo ra văn bản được nhúng trực tiếp.

Một số ví dụ về phương pháp biến đổi nhúng dựa trên từ là:

· Tính trung bình: Đây là phương pháp đơn giản nhất, trong đó chỉ lấy mức trung bình của các từ nhúng cho mỗi từ trong tài liệu. Phương pháp này nhanh chóng và dễ dàng nhưng lại bỏ qua thứ tự và cấu trúc của từ và có thể làm mất một số thông tin quan trọng.

· Trọng số tần suất nghịch đảo (*term frequency-inverse document frequency, TF-IDF*): Đây là một biến thể của phương pháp tính trung bình, trong đó trọng số được tính cho mỗi từ được nhúng theo điểm tần suất nghịch đảo của từ trong văn bản đó. Điểm này phản ánh tầm quan trọng của một từ trong một văn bản và trong một tập hợp văn bản. Phương pháp này mang lại nhiều trọng số hơn cho những từ hiếm và có nhiều thông tin, đồng thời ít trọng số hơn cho những từ phổ biến và không liên quan.

· Trọng số tần số nghịch đảo mượt mà (*Smooth Inverse Frequency, SIF*): Đây là một biến thể khác của phương pháp tính trung bình, trong đó trọng số cho mỗi từ được nhúng được tính bằng xác suất nghịch đảo của nó trong kho dữ liệu tham chiếu. Phương pháp này mang lại nhiều trọng số hơn cho các từ dành riêng cho tài liệu và ít trọng lượng hơn cho các từ phổ biến trong ngôn ngữ. Phương pháp này cũng loại bỏ thành phần chính

đầu tiên của các từ nhúng, được cho là nắm bắt một số thông tin phổ biến và không có ý nghĩa.

Một số ví dụ về phương pháp biến đổi nhúng dựa trên văn bản là:

· Doc2Vec [36]: Đây là mô hình tìm hiểu cách nhúng tài liệu bằng cách dự đoán các từ trong văn bản dựa trên cách nhúng của nó hoặc ngược lại. Mô hình này dựa trên mô hình Word2Vec, học cách nhúng từ bằng cách dự đoán các từ xung quanh trong câu khi có từ nhúng hoặc ngược lại. Mô hình này có thể nắm bắt được bối cảnh và sự mạch lạc của văn bản, nhưng nó đòi hỏi một kho tài liệu lớn và đa dạng để huấn luyện.

· Bộ chuyển đổi câu: Đây là mô hình tìm hiểu cách nhúng tài liệu bằng cách mã hóa toàn bộ tài liệu bằng cách sử dụng mạng nơ-ron dựa trên bộ biến đổi (*transformers*), chẳng hạn như BERT [68] hay SciBERT [18]. Các mạng nơ-ron này được huấn luyện trước về số lượng lớn văn bản bằng cách sử dụng các tác vụ hiểu ngôn ngữ khác nhau, chẳng hạn như mô hình hóa ngôn ngữ ẩn và dự đoán câu tiếp theo. Mô hình này có thể nắm bắt được các đặc điểm ngữ nghĩa và cú pháp của tài liệu nhưng đòi hỏi nhiều tài nguyên tính toán và tinh chỉnh để đạt được kết quả tốt.

Quay lại với mục đích chính của việc sử dụng phép nhúng văn bản là đưa các văn bản có nội dung khác nhau về cùng một không gian số sao cho những văn bản có nội dung tương tự thì giá trị nhúng của chúng cũng nằm gần nhau trong không gian. Trong các mô hình khuyến nghị trích dẫn hiện nay, thông thường khoảng cách này được tính bằng hàm tương tự cosine như sau:

$$similar\_cosine(e_1, e_2) = \frac{e_1^T e_2}{||e_1||||e_2||} \quad (1.1)$$

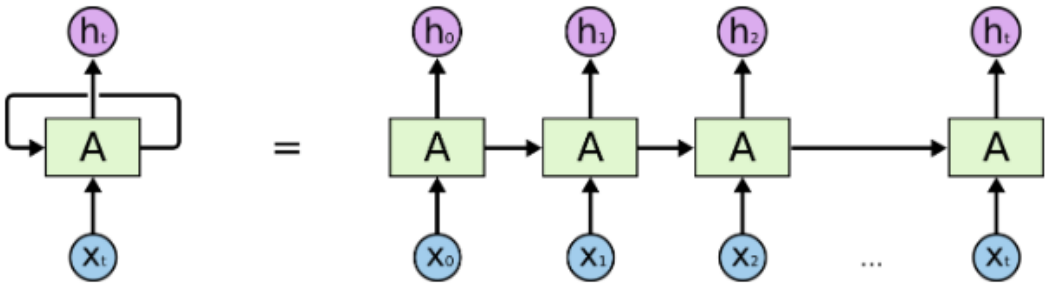
ở đây $||e|| = \sqrt{\sum_{i=1}^{d} e_i^2}$ là độ lớn của vectơ $(e) \in \mathbb{R}^d$. Góc giữa hai vectơ càng nhỏ (hai văn bản càng tương tự nhau) thì độ tương tự cosine càng cao. Độ tương tự cosine nhỏ nhất bằng -1 nếu hai vectơ này trái dấu nhau.

Trong luận án này, phép nhúng văn bản sẽ được sử dụng trong các mô hình được đề xuất ở các chương 2, chương 3 và chương 4.

1.5.2. Họ các mô hình nơ-ron hồi quy

Nghiên cứu về mô hình ngôn ngữ ngày càng được chú trọng và đã thu được những thành quả vượt trội trong những năm gần đây. Khả năng của các mô hình ngôn ngữ trong việc nhận dạng, tóm tắt, dịch, dự đoán và tạo văn bản cũng như các nội dung khác một cách tự nhiên cho phép ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau. Tuy nhiên, dữ liệu văn bản (*text-based data*), hay được gọi là dữ liệu tuần tự (*sequential data*), rất khó mô hình hóa do độ dài thay đổi của nó. Mạng nơ-ron truyền thẳng (*Feedforward Neural Network, FNN*) hoạt động tốt với đầu vào có kích thước cố định nhưng không tính đến cấu trúc không cố định. Mặt khác, mạng nơ-ron tích chập (*Convolutional Neural Networks, CNN*) có thể xử lý các chuỗi (*sequence*) dữ liệu dài nhưng bị hạn chế bởi thực tế là chúng không bảo toàn thứ tự tuần tự của chuỗi trong suốt quá trình mô hình hóa. Họ mô hình mạng nơ-ron hồi quy thường hoạt động tốt hơn khi lập mô hình dữ liệu tuần tự bằng cách sử dụng đầu ra từ lần gặp trước làm đầu vào cho lần lặp sau. Trong phần này sẽ trình bày về ba loại mô hình trong họ mô hình nơ-ron hồi quy: mạng nơ-ron hồi quy RNN, bộ nhớ dài-ngắn (*Long-Short Term Memory*, *LSTM*) và BiLSTM, đồng thời cung cấp các ví dụ trong lĩnh vực xử lý ngôn ngữ tự nhiên.

Trong lý thuyết về ngôn ngữ, ngữ nghĩa của một câu được tạo thành từ mối liên kết của những từ trong câu theo một cấu trúc ngữ pháp. Nếu xét từng từ một đứng riêng lẻ thì không thể hiểu được nội dụng của toàn bộ câu, nhưng dựa trên những từ xung quanh có thể hiểu được trọn vẹn một câu nói. Như vậy cần phải có một kiến trúc đặc biệt hơn cho các mạng nơ-ron biểu diễn ngôn ngữ nhằm mục đích liên kết các từ liền trước với các từ ở hiện tại để tạo ra mối liên hệ xâu chuỗi. Mạng nơ-ron hồi quy RNN đã được thiết kế đặc biệt để giải quyết yêu cầu này:



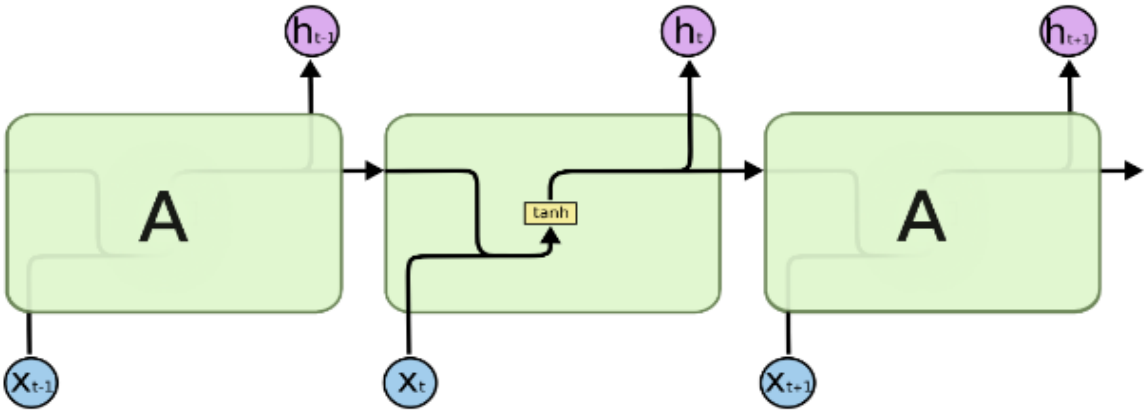
Hình 1.11. Mạng nơ-ron hồi quy RNN với vòng lặp

Hình 1.11 biểu diễn kiến trúc của một mạng nơ-ron truy hồi. Trong kiến trúc này mạng nơ-ron sử dụng một đầu vào là một vectơ ($\boldsymbol{x_t}$) và trả ra đầu ra là một giá trị ẩn ($\boldsymbol{h_t}$). Đầu vào được nối với một thân mạng nơ-ron ($\boldsymbol{A}$) có tính chất hồi quy và thân ($\boldsymbol{A}$) này được đấu tới đầu ra ($\boldsymbol{h_t}$). Vòng lặp ($\boldsymbol{A}$) ở thân mạng nơ-ron là điểm mấu chốt trong nguyên lý hoạt động của mạng nơ-ron truy hồi. Đây là chuỗi sao chép nhiều lần của cùng một kiến trúc nhằm cho phép các thành phần có thể kết nối liền mạch với nhau theo mô hình chuỗi. Đầu ra của vòng lặp trước chính là đầu vào của vòng lặp sau.

Một trong những đặc điểm giúp phân biệt RNN với các cấu trúc mạng nơ-ron khác là về mặt lý thuyết, chúng có thể kết nối thông tin trước đó với tác vụ hiện tại. Trong trường hợp dữ liệu văn bản, sử dụng các từ trước đó để tiết lộ thông tin về từ hiện tại trong một câu hoàn chỉnh. Tuy nhiên, thật không may trong thực tế, RNN không phải lúc nào cũng làm tốt công việc kết nối thông tin, đặc biệt khi khoảng cách càng lớn. Nếu mô hình đang cố gắng dự đoán từ cuối cùng trong "*những đám mây trên bầu trời*", thì không cần thêm ngữ cảnh nào nữa, mô hình cũng có thể rút ra từ năm từ trước đó rằng "*bầu trời*" là từ tiếp theo. Trong những trường hợp như vậy, khi khoảng cách giữa thông tin liên quan và nơi cần thông tin đó là nhỏ, RNN có thể học cách sử dụng thông tin trong quá khứ. Ngược lại, khi mô hình thử đoán từ cuối cùng trong văn bản "*Tôi lớn lên ở Pháp...tôi nói trôi chảy tiếng Pháp*". Thông tin gần đây cho thấy từ tiếp theo có lẽ là tên của một ngôn ngữ, nhưng nếu muốn thu hẹp ngôn ngữ nào, chúng ta cần bối cảnh của *Pháp* từ xa hơn. Hoàn toàn có thể xảy ra trường hợp khoảng cách giữa thông tin liên quan và điểm cần thiết sẽ trở nên rất lớn. Tuy nhiên, RNN thường không thực hiện tốt công việc lập mô hình trong tình huống như vậy. Do đó mô hình LSTM [14] đã được đề xuất để giải quyết vấn đề này.

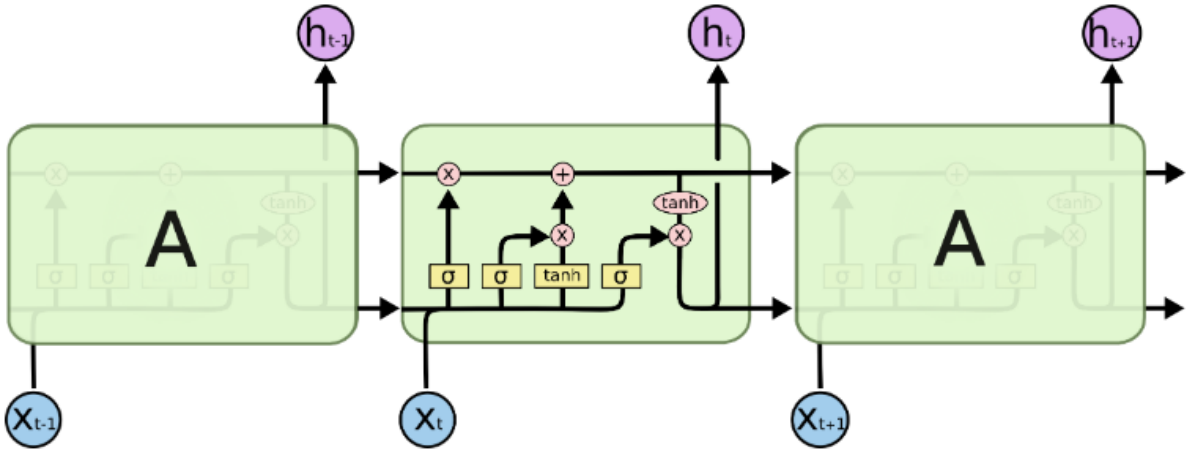
LSTM là một loại RNN đặc biệt, có khả năng học các phụ thuộc dài hạn. Chúng hoạt động rất hiệu quả trong nhiều vấn đề khác nhau và hiện được sử dụng rộng rãi. LSTM được thiết kế rõ ràng để cải tiến vấn đề phụ thuộc lâu dài. Kết nối thông tin trong thời gian dài gần như là hành vi mặc định của chúng. LSTM, giống như RNN, cũng có cấu trúc giống như chuỗi, nhưng mô-đun lặp lại có cấu trúc khác, phức tạp hơn nhiều. Thay vì có một lớp mạng nơ-ron duy nhất, có bốn lớp mạng tương tác với nhau.

Một nút trong mạng RNN tiêu chuẩn sẽ có kiến trúc rất đơn giản chẳng hạn như đối với kiến trúc gồm một tầng ẩn là hàm *tanh* như Hình 1.12.



Hình 1.12. Sự lặp lại cấu trúc mô-đun trong mạng RNN chứa một tầng ẩn

LSTM cũng có một chuỗi dạng như thế nhưng phần kiến trúc lặp lại có cấu trúc khác biệt hơn. Thay vì chỉ có một tầng đơn, chúng có tới 4 tầng ẩn (3 *sigmoid* và 1 *tanh*) tương tác với nhau theo một cấu trúc đặc biệt.

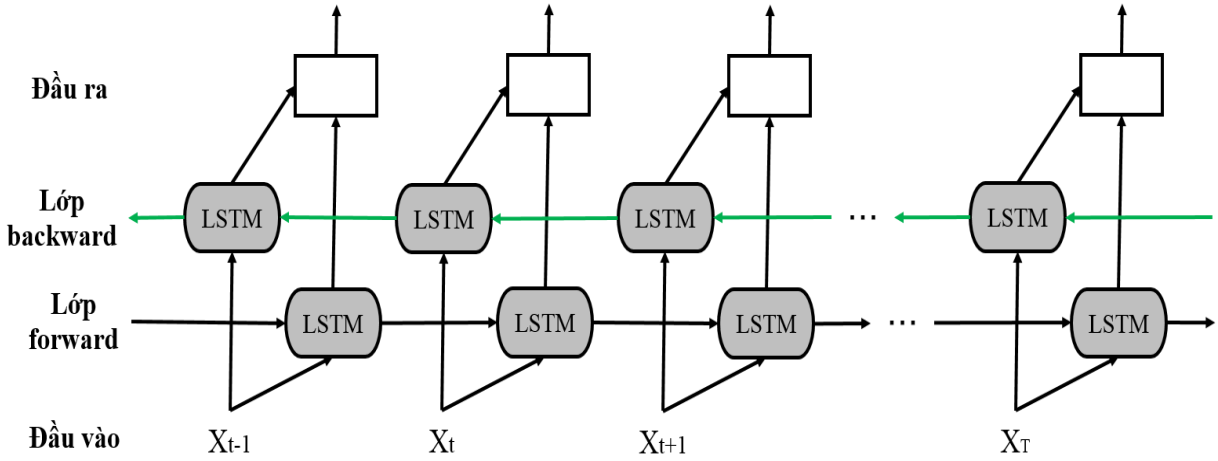


Hình 1.13. Sự lặp lại cấu trúc mô-đun trong mạng LSTM chứa bốn tầng ẩn (3 sigmod và 1 tanh) tương tác

Trong sơ đồ tính toán ở Hình 1.13, mỗi một phép tính sẽ triển khai trên một vectơ. Trong đó hình tròn màu hồng biểu diễn một toán tử đối với vectơ như phép cộng vectơ hay phép nhân vô hướng các vectơ. Màu vàng thể hiện hàm kích hoạt mà mạng nơ-ron sử dụng để học trong tầng ẩn, thông thường là hai hàm phi tuyến *sigmoid* và *tanh*. Kí hiệu hai đường thẳng nhập vào thể hiện phép chập kết quả trong khi kí hiệu hai đường

thẳng rẽ nhánh thể hiện cho nội dung vectơ trước đó được sao chép để đi tới một phần khác của mạng nơ-ron.

BiLSTM (LSTM hai chiều, Hình 1.14) là mạng nơ-ron hồi quy được sử dụng chủ yếu trong xử lý ngôn ngữ tự nhiên. Không giống như LSTM tiêu chuẩn, BiLSTM có đầu vào chạy theo cả hai luồng và có khả năng sử dụng thông tin từ cả hai phía, điều này khiến nó trở thành một công cụ mạnh mẽ để mô hình hóa sự phụ thuộc tuần tự giữa các từ và cụm từ theo cả hai hướng của văn bản. BiLSTM thêm một lớp LSTM nữa, đảo ngược hướng của luồng thông tin. Điều đó có nghĩa là chuỗi đầu vào sẽ quay ngược lại trong lớp LSTM bổ sung, sau đó tổng hợp các đầu ra từ cả hai lớp LSTM theo một số cách, chẳng hạn như trung bình, tổng, nhân hoặc ghép (*concatenation*).



Hình 1.14. Kiến trúc bộ nhớ 2 chiều BiLSTM

Kiểu kiến trúc này có nhiều ưu điểm trong các vấn đề trong thế giới thực, đặc biệt là trong xử lý ngôn ngữ tự nhiên. Điểm đáng chú ý chính là mọi thành phần của chuỗi đầu vào đều có thông tin từ cả quá khứ và hiện tại. Như đã nói, BiLSTM có thể tạo ra kết quả đầu ra có ý nghĩa hơn, đặc biệt là trong trường hợp xây dựng mô hình ngôn ngữ, vì các từ trong khối văn bản thường được kết nối theo cả hai cách - với các từ trước đó và các từ trong tương lai. Ví dụ, trong câu "*Apple is something that*…", từ Apple có thể nói về quả táo như một loại trái cây hoặc về công ty Apple. LSTM truyền thống sẽ không thể biết ý nghĩa của Apple vì nó không biết bối cảnh trong tương lai. Ngược lại, rất có thể ở hai câu sau: "*Apple is something that competitors simply cannot reproduce.*" và "*Apple is something that I like to eat.*", BiLSTM có thể làm rất tốt việc phân biệt quả táo với công ty công nghệ Apple, sử dụng thông tin từ bối cảnh tương lai của nó. Vì vậy,

chúng ta có thể thấy rõ rằng mô hình BiLSTM có lợi trong nhiều nhiệm vụ xử lý ngôn ngữ tự nhiên, chẳng hạn như phân loại câu, dịch thuật và nhận dạng thực thể. Ngoài ra, nó còn tìm thấy các ứng dụng của nó trong nhận dạng giọng nói, dự đoán cấu trúc protein, nhận dạng chữ viết tay và các lĩnh vực tương tự.

Tuy nhiên, điều đáng nói là BiLSTM là mô hình chậm hơn nhiều và cần nhiều thời gian huấn luyện hơn so với LSTM một chiều. Do đó, để giảm bớt gánh nặng tính toán, thông thường người ta chỉ triển khai nó khi thực sự cần thiết, chẳng hạn như trong trường hợp mô hình LSTM một chiều không hoạt động như mong đợi.

1.5.3. Độ liên quan của tài liệu trong các hệ thống truy xuất thông tin

Như đã mô tả ở Hình 1.2, mô hình khuyến nghị trích dẫn về bản chất là một mô hình truy xuất thông tin. Trong các hệ thống truy xuất thông tin, mức độ liên quan là thước đo mức độ một tài liệu được truy xuất hoặc một bộ tài liệu đáp ứng nhu cầu thông tin của người dùng. Nói cách khác, mức độ liên quan cho biết mức độ hữu ích hoặc thỏa mãn của một tài liệu trong việc trả lời truy vấn của người dùng. Mức độ liên quan có thể bị ảnh hưởng bởi nhiều yếu tố, chẳng hạn như nội dung, chủ đề, chất lượng, tính kịp thời, thẩm quyền và tính mới của tài liệu. Những người dùng khác nhau có thể có những tiêu chí khác nhau để đánh giá mức độ liên quan của một tài liệu, tùy thuộc vào nền tảng, sở thích và mục tiêu của họ. Vì vậy, sự liên quan thường mang tính chủ quan và năng động hơn là khách quan và tĩnh tại. Trong các hệ thống khuyến nghị trích dẫn, mức độ liên quan của tài liệu được đánh giá qua độ phù hợp về nội dung, tác giả, nơi xuất bản và quan hệ trích dẫn trước đó của nó.

Có nhiều cách khác nhau để đánh giá mức độ liên quan của một tài liệu hoặc một bộ tài liệu. Một phương pháp phổ biến là sử dụng độ chính xác MAP, MRR và và chỉ số Recall@K, là các số liệu so sánh số lượng tài liệu liên quan được hệ thống truy xuất với tổng số tài liệu liên quan trong bộ sưu tập. Độ chính xác là tỷ lệ tài liệu liên quan được truy xuất trên tổng số tài liệu được truy xuất, trong khi Recall@K là tỷ lệ tài liệu liên quan được truy xuất trên tổng số tài liệu liên quan. Độ chính xác cao có nghĩa là hầu hết các tài liệu được truy xuất đều có liên quan, trong khi Recall@K cao có nghĩa là hầu hết các tài liệu liên quan đều được truy xuất. Một phương pháp khác là sử dụng xếp hạng,

đó là quá trình sắp xếp thứ tự các tài liệu được truy xuất theo điểm phù hợp ước tính của chúng. Thứ hạng của một tài liệu càng cao thì nó càng được mong đợi là có liên quan. Xếp hạng có thể dựa trên nhiều yếu tố khác nhau, chẳng hạn như tần suất, vị trí, mức độ gần gũi và tầm quan trọng của các thuật ngữ truy vấn trong tài liệu, cũng như mức độ phổ biến, quyền hạn và tính đa dạng của tài liệu.

Cách tính toán độ chính xác MAP, MRR, chỉ số Recall@K và cách xếp hạng sẽ được trình bày chi tiết ở các chương 2, chương 3 và chương 4.

1.5.4. Hàm mất mát bộ ba (*triplet loss function*)

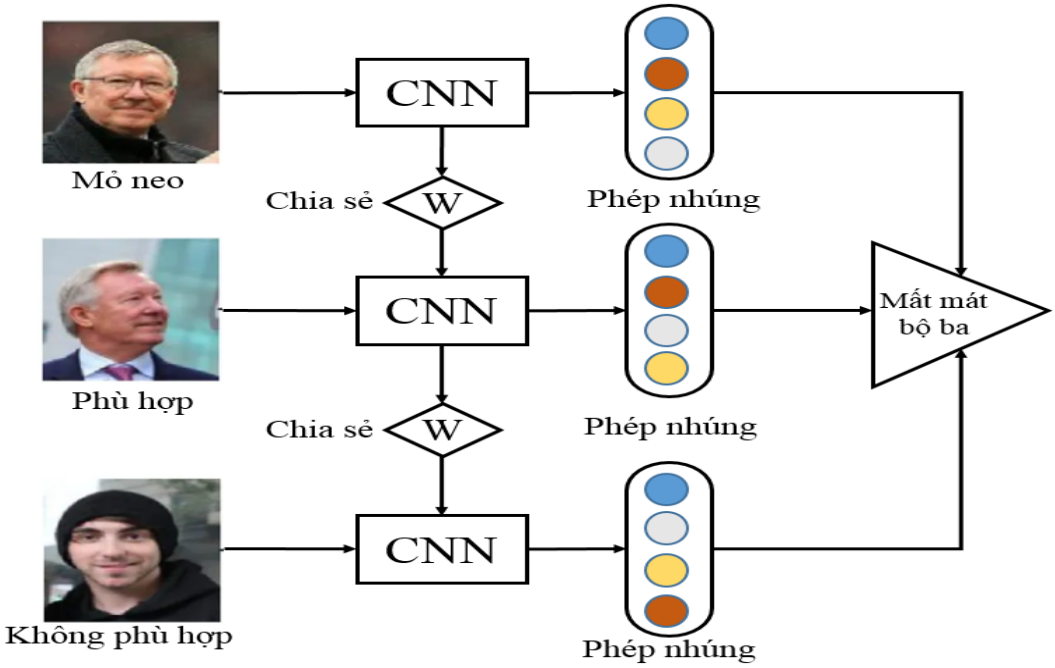
Hàm mất mát là một cách đo lường mức độ của một mô hình học máy dự đoán kết quả hoặc giá trị thực của một đầu vào nhất định. Ví dụ: nếu có một mô hình dự đoán giá một ngôi nhà dựa trên các đặc điểm của nó, hàm mất mát sẽ cho biết dự đoán đó khác bao nhiêu so với giá thực tế.

Hàm mất mát rất quan trọng vì hai lý do:

· Hàm mất mát giúp đánh giá hiệu suất và độ chính xác của mô hình. Mức tổn thất thấp hơn có nghĩa là sự phù hợp tốt hơn giữa dự đoán và thực tế.

· Hàm mất mát điều hướng quá trình huấn luyện mô hình học sâu. Bằng cách giảm thiểu hàm mất mát, có thể điều chỉnh các tham số để cải tiến dự đoán của mô hình.

Có nhiều loại hàm mất mát như là Mean Squared Error (MSE), Mean Absolute Error (MAE), Binary Cross-Entropy (BCE) hay Categorical Cross-Entropy (CCE), tùy thuộc vào dữ liệu đầu vào và nhiệm vụ của mô hình học sâu. Tuy nhiên trong các mô hình khuyến nghị trích dẫn ở các chương 2, chương 3 và chương 4, luận án sử dụng hàm mất mát bộ ba. Hàm mất mát bộ ba là một hàm đo lường mức độ mô hình học máy học cách so sánh và xếp hạng các dữ liệu đầu vào khác nhau dựa trên mức độ giống nhau của chúng. Ví dụ nếu có một mô hình nhận dạng khuôn mặt, hàm mất mát bộ ba sẽ giúp mô hình học sâu biết cách phân biệt giữa những người khác nhau và xác định cùng một người trên các hình ảnh khác nhau.



Hình 1.15. Ý nghĩa của hàm mất mát bộ ba

Hàm mất mát bộ ba hoạt động bằng cách sử dụng bộ ba dữ liệu, bao gồm:

· Mẫu đầu vào mỏ neo (*anchor*), là điểm tham chiếu để so sánh.

· Mẫu đầu vào phù hợp (*positive*), có cùng nhãn hoặc lớp với mỏ neo.

· Mẫu đầu vào không phù hợp (*negative*), có nhãn hoặc lớp khác với mỏ neo

Mục tiêu của hàm mất mát bộ ba là làm cho mô hình học các phần nhúng (chính là các biểu diễn số của dữ liệu đầu vào) sao cho khoảng cách giữa mỏ neo và mẫu phù hợp nhỏ hơn khoảng cách giữa mỏ neo và mẫu không phù hợp. Bằng cách này, mô hình học sâu có thể học cách xếp hạng dữ liệu đầu vào dựa trên mức độ giống nhau của chúng với điểm mỏ neo.

Hàm mất mát bộ ba có thể được định nghĩa như sau:

$$L(a, p, n) = \max(0, ||f(a) - f(p)|| - ||f(a) - f(n)|| + m) \quad (1.2)$$

Trong đó:

· (a) là mẫu đầu vào mỏ neo

· (p) là mẫu đầu vào phù hợp

· (n) là mẫu đầu vào không phù hợp

· ||d|| là thước đo khoảng cách, chẳng hạn như khoảng cách Euclide hoặc hàm tương tự cosine

· (m) là một tham số kiểm soát khoảng cách tối thiểu giữa các mẫu phù hợp và mẫu không phù hợp

· f là một phép biến đổi nhúng

Hàm mất mát bộ ba cố gắng giảm thiểu giá trị của $\boldsymbol{L}(a, p, n)$ cho mỗi bộ ba dữ liệu, có nghĩa là mô hình cố gắng tạo khoảng cách giữa điểm neo và mẫu phù hợp càng nhỏ càng tốt và khoảng cách giữa điểm neo và mẫu không phù hợp càng lớn càng tốt, đồng thời duy trì mức chênh lệch (m).

1.6. Kết luận chương 1

Trong chương 1, luận án đã giới thiệu chi tiết bài toán khuyến nghị trích dẫn và các nghiên cứu liên quan hiện nay để giải quyết cho bài toán này. Các hướng tiếp chính hiện nay để giải quyết bài toán khuyến nghị trích dẫn bao gồm: lọc nội dung, lọc cộng tác, lọc dựa trên đồ thị và phương pháp kết hợp.

Trên cơ sở phân tích một số phương pháp tiếp cận hiện tại, có thể thấy một số hạn chế trong các mô hình khuyến nghị trích dẫn như sau:

· Thứ nhất là với các mô hình lọc nội dung, các cách tiếp cận dựa trên mô hình mạng nơ-ron trích dẫn vẫn chưa khai thác hết các thông tin từ các bài báo học thuật.

· Thứ hai là một số mô hình kép xử lý cả thông tin ngữ cảnh và thông tin học thuật của bài báo thì vẫn chưa sử dụng hết các thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên và các mô hình học sâu.

· Thứ ba là các mô hình sử dụng thêm các siêu dữ liệu của bài báo để thiết lập mạng liên kết trích dẫn thì vẫn chưa dành sự chú ý thỏa đáng cho các kết quả mới nhất của các đồ thị liên kết trích dẫn gần đây.

Vì vậy, vấn đề cần nghiên cứu của luận án là đề xuất các mô hình hoặc cách tiếp cận mới hiệu quả để giải quyết cho 3 hạn chế đã đề cập. Thêm vào đó, chương 1 cũng giới thiệu một số các kiến thức nền tảng cần thiết để phục vụ cho việc xây dựng các phương pháp đề xuất trong các chương tiếp theo. Các kiến thức nền tảng bao gồm: phép

biến đổi nhúng văn bản, bộ nhớ hai chiều BiLSTM, độ liên quan của tài liệu trong hệ thống truy xuất thông tin và hàm mất mát.

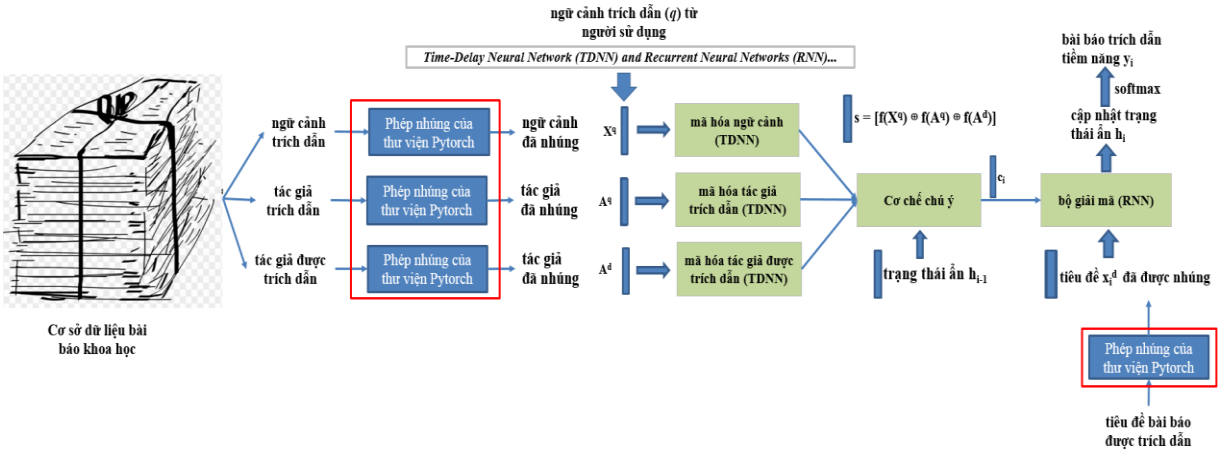
CHƯƠNG 2. MÔ HÌNH ENHANCLED-NCN BỔ SUNG THÔNG TIN TIÊU ĐỀ VÀ SỬ DỤNG PHÉP NHÚNG BERT

Khi trích dẫn một bài báo nào đó cho công trình nghiên cứu của mình, thì yếu tố đầu tiên mà các nhà khoa học quan tâm là nội dung của bài báo dự định để trích dẫn đó có phù hợp với ngữ cảnh trích dẫn hay không? Phương pháp khuyến nghị trích dẫn dựa trên nội dung xem xét văn bản được thể hiện bằng từ khóa, tiêu đề và tóm tắt của các bài báo trích dẫn và áp dụng các phương pháp biểu diễn văn bản để thể hiện ngữ nghĩa của những nội dung này, từ đó có thể khuyến nghị tài liệu tương tự và mới nhất bằng cách tính toán độ tương tự của văn bản [81]. Tuy nhiên các mô hình theo hướng này hiện nay vẫn còn đó các hạn chế về cách nhúng dữ liệu văn bản vào không gian vectơ hoặc vẫn chưa sử dụng hết thông tin quan trọng của bài báo. Một trong những nghiên cứu nổi bật đi theo hướng lọc nội dung là mô hình NCN của hai nhóm nghiên cứu Ebesu [10] và Färber [11]. Chương 2 trình bày chi tiết về đề xuất cải tiến mô hình NCN này bằng cách bổ sung thêm thông tin của bài báo và sử dụng phép nhúng văn bản BERT.

Các đề xuất và kết quả nghiên cứu của chương này được công bố tại công trình [CT1] trong phần "Danh mục các công trình đã công bố của tác giả".

2.1. Phân tích vấn đề tồn tại của mô hình NCN

Mô hình mạng trích dẫn nơ-ron (*Neural Citation Netwwork, NCN*) là một trong những mô hình nổi tiếng cho bài toán khuyến nghị trích dẫn. Mô hình NCN được công bố đầu tiên vào năm 2017 bởi nhóm nghiên cứu của Ebesu và Fang [10], và sau đó được xây dựng lại bởi nhóm của Färber [11] vào năm 2020. Như Hình 2.1, mô hình NCN sử dụng kiến trúc mã hóa-giải mã kết hợp với cơ chế chú ý (*attention*) để đưa ra danh sách khuyến nghị là các bài báo có thể trích dẫn dựa trên ngữ cảnh của một đoạn văn bản. Đầu vào của mô hình NCN bao gồm ngữ cảnh trích dẫn (bao gồm cả từ cơ sở dữ liệu bài báo khoa học và từ người sử dụng), tên tác giả trích dẫn và tác giả được trích dẫn.



Hình 2.1. Kiến trúc của mô hình NCN

Mô hình NCN sử dụng dữ liệu ngữ cảnh trích dẫn (câu hoặc đoạn chứa dấu chấm hỏi trích dẫn) làm đầu vào, sau đó sử dụng các bộ mã hóa để tạo ra các biểu diễn đặc trưng của ngữ cảnh và các tác giả liên quan. Bộ giải mã RNN sau đó sử dụng các biểu diễn này để tính toán mức độ liên quan của các tài liệu có thể được trích dẫn. Cuối cùng, cơ chế chú ý giúp mô hình tập trung vào các phần thông tin quan trọng trong ngữ cảnh và tiêu đề tài liệu để khuyến nghị trích dẫn phù hợp nhất.

2.1.1. Bộ mã hóa

Bộ mã hóa được xây dựng như một phần của mô hình NCN để chuyển đổi ngữ cảnh trích dẫn và tên tác giả được trích dẫn/được trích dẫn thành các thang đo đặc trưng chứa thông tin quan trọng về ngữ cảnh và tác giả tương ứng. Bộ mã hóa này bao gồm 2 thành phần là mã hóa ngữ cảnh trích dẫn và mã hóa tác giả trích dẫn/được trích dẫn.

Mã hóa ngữ cảnh trích dẫn được xây dựng bằng cách sử dụng mạng nơ-ron trễ (TDNN) được giới thiệu bởi nhóm nghiên cứu Collobert [82]. Bộ mã hóa này giúp chuyển ngữ cảnh trích dẫn thô thành tensor đặc trưng, chứa các thông tin quan trọng về ngữ cảnh. Nó bao gồm một lớp tích chập, một lớp *pooling*, và một lớp kết nối đầy đủ.

Bộ mã hóa tác giả được sử dụng để mã hóa tên tác giả của bài báo trích dẫn và bài báo được trích dẫn. Bộ mã hóa này có kiến trúc giống với bộ mã hóa ngữ cảnh và được áp dụng nhiều lần với các kích thước bộ lọc khác nhau trong lớp tích chập. Kết quả của việc áp dụng bộ mã hóa ngữ cảnh và mã hóa tác giả được biểu diễn dưới dạng:

$$\mathbf{s} = [f(\boldsymbol{X}^q) \oplus f(\boldsymbol{A}^q) \oplus f(\boldsymbol{A}^d)] \quad (2.1)$$

trong đó ($\boldsymbol{X}^q$) biểu diễn cho ngữ cảnh trích dẫn, ($\boldsymbol{A}^q$) là tên tác giả bài báo trích dẫn, ($\boldsymbol{A}^d$) là tên tác giả bài báo được trích dẫn, f là phép biến đổi nhúng. Ký hiệu $\oplus$ là phép nối (*concatenation*) các vectơ biểu diễn ngữ cảnh trích dẫn, tên tác giả bài báo trích dẫn và bài báo được trích dẫn. Sau khi kết hợp ba thành phần này lại, sẽ có một vectơ ($\boldsymbol{s}$) tổng hợp, đại diện cho cả nội dung của ngữ cảnh trích dẫn và thông tin về các tác giả để đưa vào cơ chế chú ý, giúp mô hình đưa ra khuyến nghị trích dẫn phù hợp.

2.1.2. Bộ giải mã

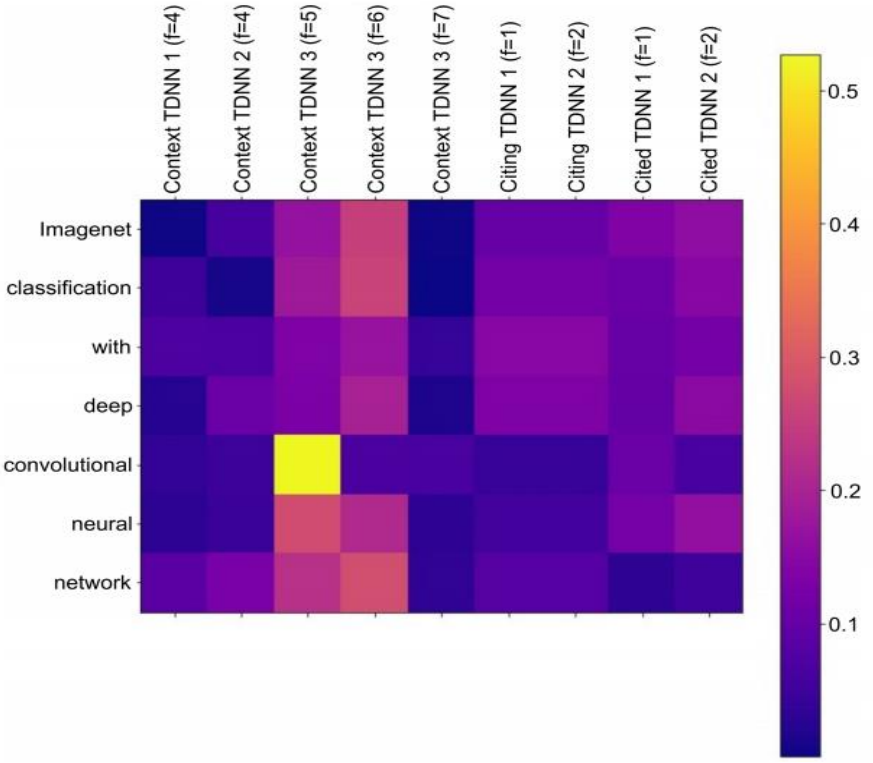
Bộ giải mã của NCN là mạng nơ-ron hồi quy RNN sử dụng đơn vị hồi quy có cổng GRU [83] làm cơ chế kiểm soát (*gating mechanism*) cũng như cơ chế chú ý [84]. Nó được áp dụng cho tiêu đề của mọi tài liệu có thể được sử dụng làm trích dẫn cho ngữ cảnh truy vấn. Mục đích của bộ giải mã là tạo ra điểm số cho mỗi tài liệu trong cơ sở dữ liệu, từ đó xác định mức độ phù hợp của tài liệu với ngữ cảnh trích dẫn. Bộ giải mã áp dụng GRU để tạo ra điểm số cho tiêu đề của mỗi tài liệu có thể được trích dẫn.

2.1.3. Cơ chế chú ý

NCN sử dụng cơ chế chú ý được giới thiệu ban đầu bởi nhóm của Bahdanau [84]. Cơ chế chú ý giúp gán trọng số cho các mã hóa khác nhau (mã hóa ngữ cảnh và tác giả) dựa trên độ liên quan của chúng với từ trước đó trong quá trình giải mã. Với cơ chế chú ý này, các mã hóa (s_j) bắt nguồn từ bộ mã hóa ngữ cảnh và tác giả được gán cho các trọng số phụ thuộc vào đầu ra (h_{i-1}) của bộ giải mã cho từ đứng trước (i). Kết quả là một vectơ ngữ cảnh (c_i) được tạo thành từ tổng có trọng số của đầu ra bộ mã hóa (s_j) theo mức độ liên quan của chúng. Cơ chế chú ý được sử dụng để nhấn mạnh vào các mã hóa đặc biệt quan trọng đối với bước thời gian hiện tại. Cơ chế chú ý được xây dựng dưới dạng mạng nơ-ron truyền thẳng FNN kết thúc bằng lớp hồi quy *softmax* để chuyển đổi vectơ chú ý (a_{ij}) thành điểm chú ý (α_{ij}). Những điều này cho thấy tầm quan trọng của đầu ra bộ mã hóa (s_j) đối với từ thứ (i) trong tiêu đề của bài báo hiện đang được giải mã.

Để minh họa cho cơ chế chú ý, trong hình 2.2 nhóm của Färber [11] đã minh họa ma trận điểm (α_{ij}) cho câu mục tiêu "*Imagenet classification with deep convolutional neural networks*" của các tác giả nổi tiếng "*Alex Krizhevsky, Ilya Sutskever, Geoffrey E.*

Hinton" sau khi chuỗi được mã hóa và được xử lý trước. Bối cảnh của ví dụ này được đặt thành "*Neural networks are really cool, especially if they are convolutional*" với các tác giả "*Chuck Norris, Bruce Lee*". Ví dụ này đã trực quan hóa cách mà bộ giải mã ít chú trọng đến các tác giả trích dẫn so với bối cảnh và các tác giả được trích dẫn. Vectơ ngữ cảnh (c_i) được xác định cho mỗi từ *i* trong tiêu đề của mỗi bài báo.



Hình 2.2. Ví dụ minh họa trọng số của cơ chế chú ý [11]

2.1.4. Thảo luận về các vấn đề còn tồn tại của mô hình NCN

Mặc dù là một mô hình khuyến nghị trích dẫn nổi tiếng (cho đến nay đã có hơn 170 công trình nghiên cứu có trích dẫn về mô hình này), tuy nhiên mô hình NCN hiện nay vẫn có những điểm hạn chế như sau:

(1) Dữ liệu của bài báo dưới dạng văn bản (*textual*) cần phải thực hiện phép biến đổi nhúng (*embedding*) trước khi đưa vào bộ mã hóa. Tuy nhiên mô hình NCN[9] hiện tại đang sử dụng hàm torch.nn.Embedding[10] của thư viện PyTorch để thực hiện biến đổi nhúng văn bản. Trong thư viện PyTorch, torch.nn.Embedding là một hàm được sử dụng

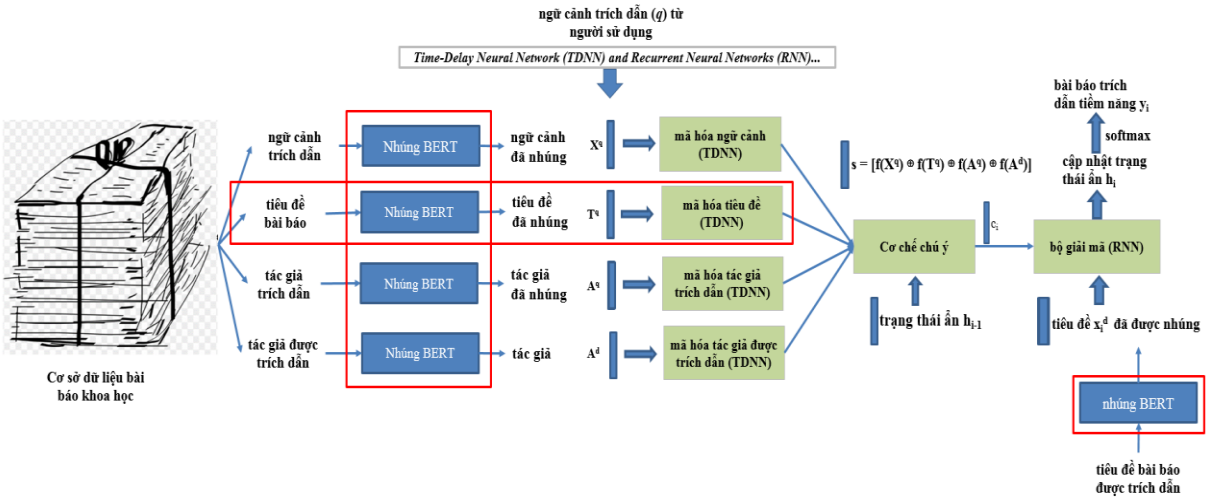
[9] https://github.com/timoklein/neural_citation
[10] https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html

để biểu diễn các phần nhúng, là các biểu diễn vectơ dày đặc của các đối tượng rời rạc. Hàm này thường được sử dụng trong các tác vụ xử lý ngôn ngữ tự nhiên để ánh xạ các biến phân loại, chẳng hạn như từ hoặc chỉ mục, tới các không gian vectơ liên tục. Ý tưởng cơ bản là tìm hiểu cách biểu diễn chiều thấp cho từng danh mục trong dữ liệu văn bản.

(2) Chưa sử dụng hết thông tin văn bản của bài báo. Cụ thể, tiêu đề của mỗi bài báo là thông tin quan trọng, vì nó là thông tin cô đọng nhất về nội dung của cả bài báo. Tuy nhiên như đã thấy ở Hình 2.1 bên trên, trong kiến trúc của mô hình NCN hiện tại, tiêu đề của bài báo trích dẫn vẫn chưa được đưa vào mô hình NCN để thực hiện mã hóa. Hạn chế này làm ảnh hưởng đáng kể đến hiệu suất của mô hình NCN hiện nay.

2.2. Cải tiến mô hình NCN

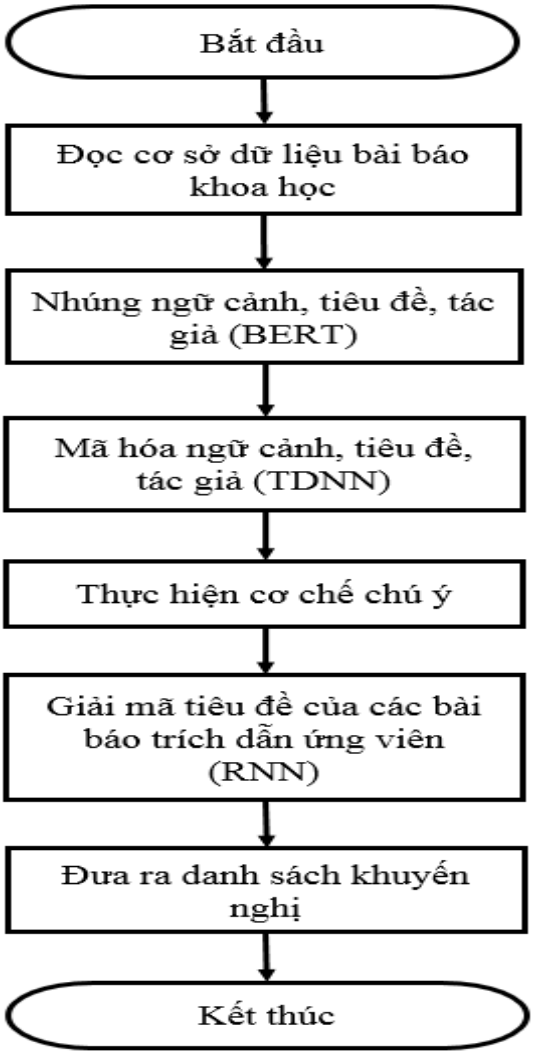
Dựa trên những phân tích về hạn chế của mô hình NCN hiện tại, luận án đã áp dụng hai cải tiến sau đây để nâng cao hiệu suất của mô hình NCN: (1) Thay phép nhúng torch.nn.Embedding bằng phép nhúng của BERT [68] (là thành tựu về xử lý ngôn ngữ tự nhiên mới hơn) và (2) Đưa cả tiêu đề của bài báo trích dẫn vào mô hình để thực hiện mã hóa. Phần cải tiến của mô hình NCN được khoanh đỏ trong Hình 2.3. Mô hình NCN đã cải tiến được đặt tên là Enhanced-NCN.



Hình 2.3. Kiến trúc của mô hình Enhanced-NCN

Trong mô hình đã cải tiến Enhanced-NCN, dữ liệu bao gồm ngữ cảnh trích dẫn, tiêu đề và thông tin tác giả sẽ được xử lý tuần tự qua các khối thực hiện nhúng, mã hóa,

thực hiện cơ chế chú ý và giải mã. Sơ đồ khối xử lý tuần tự của mô hình Enhanced-NCN được thể hiện ở Hình 2.4.



Hình 2.4. Sơ đồ khối xử lý tuần tự của mô hình Enhanced-NCN

2.2.1. Phép nhúng BERT

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) là một kỹ thuật học máy dựa trên các bộ biến đổi được dùng cho việc huấn luyện trước và xử lý ngôn ngữ tự nhiên được phát triển vào năm 2019 bởi nhóm của Jacob đến từ Google [68]. BERT là một mô hình ngôn ngữ mạnh mẽ có thể tạo các phần nhúng theo ngữ cảnh cho các từ và câu được kết xuất từ dữ liệu văn bản. Phần nhúng là các vectơ chiều thấp (*low-dimensional vector*) để nắm bắt ý nghĩa và mối quan hệ của từ và câu theo cách mà các mô hình hoặc tác vụ khác có thể sử dụng.

Phép nhúng BERT bao gồm ba phần: nhúng mã thông báo (*token*), nhúng phân đoạn (*segment*) và nhúng vị trí. Ba phần này được kết hợp lại với nhau để tạo thành phần nhúng cuối cùng cho mỗi từ trong văn bản đầu vào. Chi tiết hơn từng phần được giải thích như sau:

· Nhúng mã thông báo: Đây là các phần nhúng đại diện cho mỗi từ hoặc từ phụ trong văn bản đầu vào. BERT sử dụng WordPiece[11] để chia các từ thành các đơn vị nhỏ hơn, chẳng hạn như "*playing*" ⇒ "*play*" + "*##ing*". Điều này làm giảm kích thước từ vựng và cho phép BERT xử lý các từ hiếm hoặc chưa biết. BERT có vốn từ vựng cố định là 30,522 mã thông báo và mỗi mã thông báo có vectơ nhúng tương ứng có kích thước 768 (đối với BERT base) hoặc 1,024 (đối với BERT large). Trong nghiên cứu ở chương này đã dùng BERT large.

· Phần nhúng phân đoạn: Đây là phần nhúng cho biết một từ thuộc câu đầu tiên hay câu thứ hai trong văn bản đầu vào. BERT có thể lấy hai câu làm đầu vào, được phân tách bằng mã thông báo đặc biệt [SEP]. Điều này hữu ích cho các tác vụ yêu cầu so sánh hoặc liên hệ hai câu với nhau, chẳng hạn như suy luận ngôn ngữ tự nhiên hoặc trả lời câu hỏi. BERT có hai phần nhúng phân đoạn, một phần cho mỗi câu và mỗi phần nhúng phân đoạn có cùng kích thước với phần nhúng mã thông báo.

· Phần nhúng vị trí: Đây là phần nhúng mã hóa vị trí của mỗi từ trong văn bản đầu vào. BERT có thể xử lý văn bản đầu vào lên tới 512 mã thông báo và mỗi vị trí có một vectơ nhúng tương ứng có cùng kích thước với phần nhúng mã thông báo. Việc nhúng vị trí rất quan trọng để BERT nắm bắt được thứ tự và cấu trúc của văn bản đầu vào, vì BERT không sử dụng các lớp hồi quy (*recurrent layer*) hoặc lớp tích chập.

Kết quả nhúng cuối cùng của mỗi từ có được bằng cách thêm nhúng mã thông báo, nhúng phân đoạn và nhúng vị trí. Sau đó, các phần nhúng cuối cùng được đưa vào một chồng các lớp biến đổi, tiếp tục xử lý và tinh chỉnh các phần nhúng dựa trên ngữ cảnh và sự chú ý của văn bản đầu vào. Đầu ra của các lớp biến đổi có thể được sử dụng để nhúng toàn bộ câu hoặc nhúng từng từ trong câu.

[11] https://huggingface.co/learn/nlp-course/chapter6/6?fw=pt

2.2.2. Thêm tiêu đề bài báo vào mô hình

Có mối tương quan chặt chẽ giữa tiêu đề bài báo và ngữ cảnh của bài báo. Tiêu đề bài báo cung cấp một bản tóm tắt ngắn gọn và thông tin của bài viết, trong khi ngữ cảnh của bài báo, bao gồm cả ngữ cảnh trích dẫn (một đoạn văn ngắn xung quanh vị trí trích dẫn), giúp làm rõ sự liên quan của trích dẫn trong nội dung. Mặc dù tiêu đề của bài báo trích dẫn là yếu tố quan trọng, chứa thông tin liên quan đến truy vấn của người dùng để hệ thống khuyến nghị có thể xác định kết quả đề xuất trích dẫn, nhưng trong mô hình NCN của cả nhóm Ebesu và Yi Fang [10], và sau đó được cải tiến bởi nhóm của Färber [11] chỉ sử dụng bối cảnh trích dẫn, tác giả trích dẫn và tác giả được trích dẫn, nhưng bỏ qua tiêu đề trích dẫn khi thực hiện phép nhúng trước khi đưa vào bộ mã hóa. Để làm tăng thêm hiệu suất của mô hình NCN, luận án đã tích hợp khả năng xử lý của bộ mã hóa cho tiêu đề của bài báo trích dẫn vào mô hình Enhanced-NCN. Mô hình Enhanced-NCN sử dụng cả hai yếu tố này để tạo ra một biểu diễn phong phú và chính xác hơn cho việc khuyến nghị trích dẫn. Bằng cách tích hợp tiêu đề bài báo vào quá trình huấn luyện, mô hình nâng cao cơ chế chú ý, cho phép tập trung tốt hơn vào các đặc điểm quan trọng từ cả tiêu đề và ngữ cảnh. Sau khi thêm tiêu đề bài báo vào mô hình Enhanced-NCN để huấn luyện, cơ chế chú ý có sự thay đổi, trở nên mạnh mẽ và giàu thông tin hơn. Điều này là do mô hình bây giờ không chỉ sử dụng ngữ cảnh trích dẫn mà còn cả tiêu đề của bài báo, chứa những thông tin quan trọng có thể liên quan trực tiếp đến truy vấn của người dùng. Cụ thể, cơ chế chú ý học cách tập trung không chỉ vào ngữ cảnh trích dẫn và tên tác giả mà còn vào tiêu đề của bài báo, từ đó làm phong phú thêm việc biểu diễn thông tin của các bài báo. Việc kết hợp có trọng số giữa các yếu tố (ngữ cảnh, tiêu đề và tác giả) cung cấp một biểu diễn phong phú và chính xác hơn cho việc khuyến nghị trích dẫn. Lúc này, biểu diễn cuối cùng của văn bản là kết quả của việc áp dụng mã hóa ngữ cảnh, mã hóa tiêu đề và bộ mã hóa tác giả được ký hiệu là:

$$\boldsymbol{s} = [f(\boldsymbol{X}^q) \oplus f(\boldsymbol{T}^q) \oplus f(\boldsymbol{A}^q) \oplus f(\boldsymbol{A}^d)] \qquad (2.2)$$

trong đó ($\boldsymbol{T}^q$) biểu diễn tiêu đề của bài báo được thêm vào mô hình. Ký hiệu $\oplus$ là phép nối (*concatenation*) các vectơ biểu diễn ngữ cảnh trích dẫn, tiêu đề bài báo và tên tác giả

bài báo trích dẫn và bài báo được trích dẫn. Sau khi kết hợp bốn thành phần này lại, sẽ có một vectơ (***s***) tổng hợp, đại diện cho cả nội dung của ngữ cảnh trích dẫn, tiêu đề bài báo và thông tin về các tác giả để đưa vào cơ chế chú ý, giúp mô hình đưa ra khuyến nghị trích dẫn phù hợp hơn so với mô hình NCN ban đầu.

2.3. Tiến thành thực nghiệm với mô hình Enhanced-NCN

Nội dung nghiên cứu của chương 2 này tập trung vào việc cải tiến mô hình NCN tiên tiến nhất, do đó phần này mô tả chi tiết về cách thức tiến hành thực nghiệm với mô hình đã Enhanced-NCN để chứng minh được hiệu quả của những cải tiến đã áp dụng cho mô hình này.

2.3.1. Cài đặt mô hình Enhanced-NCN

Ban đầu, mô hình NCN được xây dựng bằng mã nguồn mở TensorFlow[12] phiên bản r.0.11 bởi nhóm nhiên cứu Ebesu [10]. Sau đó, nhóm Färber [11] đã cấu trúc lại mô hình NCN bằng cách sử dụng mã nguồn mở PyTorch. Mô hình Enhanced-NCN của luận án được cải tiến từ mã nguồn của mô hình NCN[13] trong bài báo của nhóm Färber [11] bằng cách thêm mô hình BERT, một trong những mô hình tốt nhất để xử lý ngôn ngữ tự nhiên hiện nay, đồng thời tạo thêm bộ mã hóa cho cho tiêu đề của bài báo rồi đưa chúng vào mô hình đã cải tiến Enhanced-NCN. Luận án đã sử dụng Python phiên bản 3.8.5 và PyTorch phiên bản 1.7.1 để xây dựng mô hình Enhanced-NCN. Đối với mô hình BERT, luận án đã sử dụng BertTokenizer và BertModel từ thư viện transformers của Python. Mô hình Enhanced-NCN cũng đã sử dụng thư viện torchtext[14] để chuyển đổi tập dữ liệu sang định dạng phù hợp cho PyTorch và tạo điều kiện thuận lợi cho các bước tiền xử lý. Hơn nữa, luận án đã sử dụng thư viện SpaCy[15] kết hợp với torchtext để mã hóa tập dữ liệu. Sau khi bổ nghĩa hóa dữ liệu và xóa từ dừng (*stopword*) trong văn bản bằng cách sử dụng kết hợp SpaCy và tập từ dừng nltk[16], luận án đã số hóa tập dữ liệu bằng cách sử dụng tập từ vựng có kích thước 30,522 mã thông báo của BERT cho ngữ cảnh trích dẫn,

[12] https://www.tensorflow.org/
[13] https://github.com/timoklein/neural_citation
[14] https://pytorch.org/text/stable/index.html
[15] https://spacy.io/
[16] https://www.nltk.org/

tiêu đề của bài báo và tác giả trích dẫn/được trích dẫn. Để tạo điều kiện thuận lợi cho việc truyền bá các lô (*batch*) qua mạng NCN, luận án đã sử dụng kỹ thuật thu thập dữ liệu (*bucketing technique*) mà Ebesu và Fang [10] cũng đã sử dụng. Giống như Ebesu và Fang [10], luận án tiếp tục sử dụng chức năng xếp hạng Okapi BM25[17] trong phần giải mã của mô hình Enhanced-NCN để chọn trước tiêu đề cho ngữ cảnh trích dẫn nhất định. Mô hình Enhanced-NCN được chạy trên môi trường Linux 3.10.0-x86-64, bộ xử lý NVIDIA GPU H100.

2.3.2. Mô tả về bộ dữ liệu thực nghiệm

Theo công bố của nhóm nghiên cứu Ebesu [10], họ thực nghiệm mô hình NCN ban đầu với hai bộ dữ liệu là RefSeer và arXiv CS. Tuy nhiên, trong bài báo của nhóm Färber [11] cho biết họ không thể xây dựng lại bộ dữ liệu RefSeer đã được sử dụng trong nghiên cứu của nhóm Ebesu [10]. Do luận án này tập trung cải tiến mô hình NCN của nhóm Färber [11], cho nên phần thực nghiệm luận án chỉ so sánh kết quả được cải tiến của mô hình Enhanced-NCN trên bộ dữ liệu arXiv CS. Các thông tin từ bộ dữ liệu arXiv CS để sử dụng cho mô hình Enhanced-NCN là ngữ cảnh trích dẫn (bao gồm cả từ cơ sở dữ liệu bài báo khoa học và từ người sử dụng), tên tác giả trích dẫn và tên tác giả được trích dẫn.

Tập dữ liệu arXiv CS nguyên bản bao gồm 1.7 triệu bài báo của arXiv[18] từ các lĩnh vực khoa học máy tính, bao gồm nhiều trường chứa thông tin và chủ đề khác nhau. Tập dữ liệu này chứa siêu dữ liệu của từng bài báo, chẳng hạn như tiêu đề, tác giả, tóm tắt, danh mục và tài liệu tham khảo. Bộ dữ liệu có thể được sử dụng cho nhiều ứng dụng khác nhau, chẳng hạn như phân tích xu hướng, khuyến nghị trích dẫn, dự đoán danh mục, xây dựng biểu đồ tri thức và tìm kiếm ngữ nghĩa. Bộ dữ liệu bao gồm các bài viết từ tháng 1 năm 1993 đến tháng 4 năm 2021 và được cập nhật hàng tháng. Tập dữ liệu có sẵn ở định dạng json và có thể được tải xuống từ Hugging Face[19]. Bộ dữ liệu arXivCS thu gọn dùng để kiểm tra hiệu suất của mô hình Enhanced-NCN có 502,355 bản ghi có

[17] https://pypi.org/project/rank-bm25/
[18] https://arxiv.org/
[19] https://huggingface.co/datasets/arxiv_dataset

định dạng bao gồm bốn thông tin: ngữ cảnh trích dẫn, tác giả trích dẫn, tên bài báo và tác giả được trích dẫn. Luận án đã cắt bỏ bối cảnh trích dẫn và tiêu đề trích dẫn với độ dài tương ứng là 100 và 30 từ để đạt được sự cân bằng giữa hiệu suất mô hình và thời gian huấn luyện. Không giống như Färber và cộng sự [11], luận án đã đưa tiêu đề của bài báo vào để triển khai bộ mã hóa tiêu đề, từ đó cải thiện đáng kể hiệu suất của hệ thống. Để huấn luyện và đánh giá mô hình, tập dữ liệu arXiv CS được chia thành 80% để huấn luyện, 10% để xác nhận (*validation*) và 10% để kiểm tra.

2.3.3. Phương pháp đánh giá mô hình

Hầu hết các nghiên cứu về bài toán khuyến nghị trích dẫn đều sử dụng các tiêu chí đánh giá nổi tiếng, chẳng hạn như mức tăng tích lũy chiết khấu chuẩn hóa (*Normalized Discounted Cumulative Gain, NDCG*), xếp hạng đối ứng trung bình MRR, độ chính xác trung bình MAP, Recall@K và Hits@K để đánh giá hiệu suất của một mô hình khuyến nghị trích dẫn. Bởi vì cả 2 mô hình NCN và Enhanced-NCN đều tiếp cận theo hướng lọc nội dung để tìm các bài báo phù hợp về ngữ nghĩa với ngữ cảnh trích dẫn mà không quan tâm đến thứ hạng của các bài báo trong danh sách khuyến nghị, nên chỉ cần sử dụng tiêu chí Recall@K là phù hợp. Thêm một lý do nữa là để so sánh, do nghiên cứu của nhóm Färber [11] cũng chỉ sử dụng tiêu chí Recall@10 để đánh giá hiệu suất của mô hình NCN, nên trong nghiên cứu này, luận án cũng chỉ sử dụng tiêu chí Recall@10 để đánh giá hiệu suất của mô hình đã cải tiến Enhanced-NCN.

Recall@K được định nghĩa là tỷ lệ giữa số lượng mục có liên quan được truy xuất trong kết quả top-K trên tổng số mục có liên quan trong toàn bộ tập dữ liệu. Công thức tính Recall@K được trình bày trong công thức (2.3).

$$Recall@K = \frac{Số\ lượng\ các\ mục\ có\ liên\ quan\ truy\ xuất\ được}{Tổng\ số\ mục\ liên\ quan\ có\ trong\ tập\ dữ\ liệu} \qquad (2.3)$$

Ví dụ: nếu người dùng tìm kiếm "*sách học máy*" và có 100 cuốn sách có liên quan trong tập dữ liệu, đồng thời hệ thống trả về (K) = 10 cuốn sách trong kết quả top-K, trong đó có 4 cuốn có liên quan, thì Recall@10 là 4/100 = 0.04. Điều này có nghĩa là hệ thống đã truy xuất 4% trong tổng số sách có liên quan cho truy vấn. Recall@K cao hơn có nghĩa là hệ thống hiệu quả hơn trong việc tìm kiếm các mục phù hợp cho người dùng.

Tuy nhiên, Recall@K không xem xét thứ tự hoặc chất lượng của kết quả tìm kiếm trả về, mà điều này cũng có thể ảnh hưởng đến trải nghiệm của người dùng. Do đó, Recall@K thường được sử dụng cùng với các số liệu khác, chẳng hạn như Precision@K, NDCG hoặc MRR để đưa ra đánh giá toàn diện hơn về các hệ thống tìm kiếm dữ liệu.

2.4. Đánh giá kết quả thực nghiệm và thảo luận

Để đảm bảo sự công bằng và tính khoa học khi so sánh, luận án đã chạy mã nguồn của cả 2 mô hình NCN và Enhanced-NCN trên cùng một môi trường thực nghiệm cũng như một tập dữ liệu arXiv CS. Cả 2 mô hình đều được điều chỉnh các tham số như ở Bảng 2.1 để tìm ra giá trị tham số tốt nhất cho mỗi mô hình. Để tìm ra tham số tốt nhất cho mô hình Enhanced-NCN, luận án đã thay đổi 4 siêu tham số là tỉ lệ phân chia dữ liệu (*split data*), số lượng lớp (*number of layers*), số lần lặp huấn luyện (*epochs*) và kích thước nhúng (*embedding size*) trong mô hình để so sánh kết quả với nhóm của Färber [11]. Như được hiển thị trong Bảng 2.1, luận án đã điều chỉnh các giá trị của các tham số này để tìm ra kết quả tốt nhất cho mô hình đề xuất Enhanced-NCN. Bối cảnh trích dẫn là nơi chứa nhiều thông tin nhất để có thể đưa ra các khuyến nghị trích dẫn. Tiêu đề của bài báo được trích dẫn thường chứa các thông tin liên quan đến nội dung được trích dẫn. Đó cũng là điều đầu tiên mà các nhà nghiên cứu nhìn thấy khi tìm kiếm tài liệu để trích dẫn cho nội dung của họ.

Bảng 2.1. So sánh kết quả mô hình NCN của nhóm Färber [11] và mô hình Enhanced-NCN

Tên mô hình	Phân chia dữ liệu	Số lớp	Epochs	Kích thước nhúng	Recall@10
NCN của nhóm Färber [11]	[0.8, 0.1, 0.1]	1	20	128	**0.0801**
	[0.8, 0.1, 0.1]	1	20	164	0.0663
	[0.8, 0.1, 0.1]	1	20	196	0.0527
	[0.8, 0.1, 0.1]	1	20	256	0.0413
	[0.8, 0.1, 0.1]	2	20	128	**0.1074**
	[0.8, 0.1, 0.1]	3	20	128	0.0867
Enhanced-NCN	[0.8, 0.1, 0.1]	1	20	128	0.0723
	[0.8, 0.1, 0.1]	1	20	164	**0.0921**
	[0.8, 0.1, 0.1]	1	20	196	0.0853
	[0.8, 0.1, 0.1]	1	20	256	0.0763
	[0.8, 0.1, 0.1]	2	20	164	**0.1285**
	[0.8, 0.1, 0.1]	3	20	164	0.1115

Kết quả so sánh cho thấy giữa mô hình NCN gốc của nhóm Färber [11] và mô hình được cải tiến Enhanced-NCN cho thấy sự cải thiện rõ rệt của mô hình cải tiến trong việc truy xuất các mục tiêu tại ngưỡng Recall@10. Dựa vào Bảng 2.1, có thể thấy rằng ở nhiều cấu hình khác nhau, Enhanced-NCN luôn đạt được hiệu suất cao hơn so với mô hình NCN ban đầu, đặc biệt là khi xét các yếu tố như số lớp, kích thước nhúng và số epochs. Mô hình Enhanced-NCN đã cho thấy sự cải thiện đáng kể về hiệu suất so với mô hình NCN ban đầu của nhóm Färber. Sự cải tiến này đặc biệt thể hiện rõ ở giá trị Recall@10, với giá trị cao nhất là 0.125 so với 0.107 của mô hình gốc. Ngoài việc cải thiện mô hình NCN ban đầu bằng thêm tiêu đề hay thay thế phép nhúng, thì yếu tố quan trọng góp phần vào sự cải tiến này bao gồm việc điều chỉnh kích thước nhúng phù hợp, tăng số lớp và tối ưu hóa kiến trúc. Những kết quả này cho thấy rằng mô hình Enhanced-NCN có tiềm năng ứng dụng trong các nhiệm vụ truy vấn thông tin chính xác cao.

2.4.1. Phân chia dữ liệu

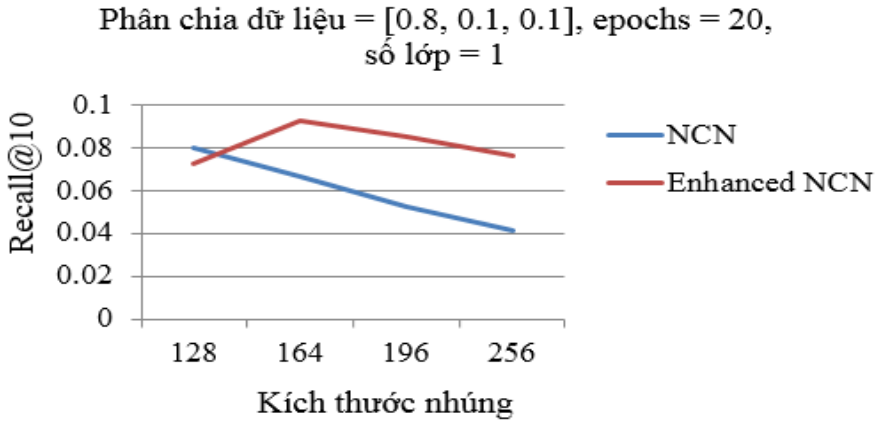
Cả hai mô hình đều được huấn luyện trên các tập dữ liệu được phân chia theo tỷ lệ huấn luyện 80%, xác nhận 10% và kiểm tra 10%. Điều này đảm bảo rằng sự so sánh

giữa hai mô hình là công bằng, vì cả hai đều làm việc trên cùng một tập dữ liệu. Sự khác biệt trong kết quả chủ yếu đến từ kiến trúc và cách thức huấn luyện của hai mô hình, không phải từ việc thay đổi tập dữ liệu.

2.4.2. Kích thước nhúng

Kích thước nhúng là yếu tố quan trọng ảnh hưởng đến khả năng học của các mô hình dựa trên biểu diễn (*representation learning*). Cả hai mô hình đều được thử nghiệm với các kích thước nhúng là 128, 196, và 256 (đối với NCN gốc) và 128, 164, và 256 (đối với Enhanced-NCN). Kết quả cho thấy, khi tăng kích thước nhúng từ 128 lên 164, Recall@10 của mô hình Enhanced-NCN tăng lên một cách đáng kể, đạt giá trị cao nhất 0.125 khi sử dụng 3 lớp và kích thước nhúng 164. Kết quả thực nghiệm cho thấy rằng kích thước nhúng 164 là điểm cân bằng tốt nhất cho mô hình Enhanced-NCN, cung cấp đủ thông tin biểu diễn mà không gây ra quá tải tính toán.

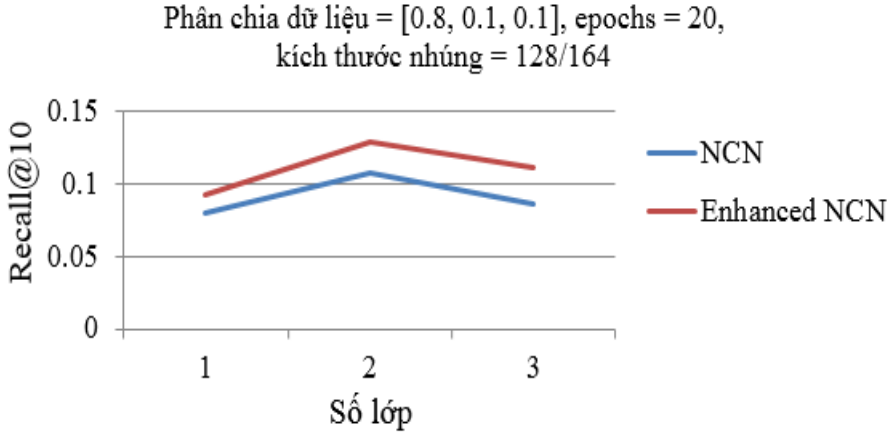
Ngược lại, với mô hình NCN ban đầu, khi tăng kích thước nhúng từ 128 lên 196, giá trị Recall@10 lại có xu hướng giảm. Điều này có thể là do mô hình NCN không khai thác hiệu quả thông tin từ các vectơ nhúng lớn hơn, dẫn đến sự giảm hiệu suất khi kích thước nhúng quá lớn. Kết quả so sánh khi điều chỉnh về kích thước nhúng được minh họa trong Hình 2.5.



Hình 2.5. So sánh kết quả Recall@10 của mô hình Enhanced-NCN và mô hình NCN của nhóm Färber [11], được điều chỉnh về kích thước nhúng

2.4.3. Số lớp

Trong bảng kết quả, cả hai mô hình đều được thử nghiệm với nhiều số lớp khác nhau: 1 lớp, 2 lớp, và 3 lớp. Khi số lớp tăng, có thể thấy rằng Recall@10 của Enhanced-NCN tăng dần, đạt giá trị cao nhất với 3 lớp và kích thước nhúng là 164, với Recall@10 đạt 0.125. Trong khi đó, mô hình NCN ban đầu của nhóm Färber [11] chỉ đạt được Recall@10 cao nhất là 0.107 với cùng cấu hình. Điều này chứng tỏ việc tăng số lớp giúp mô hình học được nhiều biểu diễn hơn và mô hình cải tiến Enhanced-NCN khai thác điều này tốt hơn so với mô hình gốc. Kết quả này được minh họa rõ ràng trong Hình 2.6.



Hình 2.6. So sánh kết quả Recall@10 của mô hình Enhanced-NCN và mô hình NCN của nhóm Färber [11], được điều chỉnh về số lượng lớp

2.4.4. Số epochs

Số epochs cho cả hai mô hình NCN và Enhanced-NCN được giữ cố định ở mức 20, giúp tránh việc mô hình bị quá khớp (*overfitting*) hoặc chưa khớp (*underfitting*). Việc giữ nguyên số epochs cũng giúp nghiên cứu của luận án có thể tập trung vào việc so sánh giữa các yếu tố khác như số lớp và kích thước nhúng, từ đó có được cái nhìn rõ ràng hơn về những yếu tố ảnh hưởng đến kết quả.

2.4.5. So sánh tổng quát

Nhìn chung, kết quả cho thấy mô hình Enhanced-NCN đã đạt được những cải thiện rõ ràng về hiệu suất so với mô hình NCN ban đầu của nhóm Färber. Giá trị Recall@10 của Enhanced-NCN ở các cấu hình khác nhau đều cao hơn NCN ban đầu,

đặc biệt là với cấu hình 3 lớp và kích thước nhúng 164. Những cải tiến này có thể được giải thích bởi các yếu tố sau:

(1) Tối ưu hóa kiến trúc: Enhanced-NCN có thể đã được tối ưu hóa về cách thức kết nối giữa các lớp và các tham số học, giúp mô hình học được nhiều thông tin hơn từ dữ liệu.

(2) Hiệu quả từ việc điều chỉnh kích thước nhúng: Mô hình Enhanced-NCN cho thấy sự phù hợp tốt hơn với kích thước nhúng trung bình (164) so với các kích thước nhúng lớn hơn, giúp nó tận dụng tốt hơn thông tin từ dữ liệu mà không bị quá tải.

(3) Sử dụng hợp lý số lớp: Mô hình Enhanced-NCN cho thấy rằng việc tăng số lớp từ 1 lên 3 lớp có lợi cho việc cải thiện hiệu suất. Điều này chứng tỏ mô hình cải tiến đã được thiết kế để tận dụng tối đa các thông tin biểu diễn từ việc tăng độ sâu của mạng.

2.5. Thực hiện tinh chỉnh các tham số của mô hình Enhanced-NCN

Trong mô hình đã cải tiến Enhanced-NCN, có 5 siêu tham số là tỉ lệ phân chia tập dữ liệu, số lớp, số lần lặp khi huấn luyện, kích thước nhúng và kích thước ẩn. Luận án đã tinh chỉnh các siêu tham số này để cung cấp thông tin chi tiết về sự ảnh hưởng của những tham số này đến hiệu suất của mô hình Enhanced-NCN. Từ kết quả chạy thực nghiệm ở Bảng 2.2, có thể thấy rằng khi ngữ cảnh trích dẫn được tiền xử lý bằng mô hình BERT và thông tin tiêu đề của bài báo được thêm vào mô hình Enhanced-NCN để nâng cao nên phong phú hơn về mặt ngữ nghĩa, cần phải điều chỉnh kích thước nhúng từ 128 lên 164. Tuy nhiên, việc tăng kích thước nhúng lên hơn 164 (ví dụ như 196 hay 232) cũng không mang lại kết quả tối ưu. Khi thử điều chỉnh kích thước của tham số trạng thái ẩn và nhận thấy rằng giá trị 128 mang lại hiệu suất tối ưu cho mô hình Enhanced-NCN. Cả NCN và Enhanced-NCN đều đạt được kết quả tốt nhất khi tập dữ liệu được phân chia thành 80% để huấn luyện, 10% để xác thực và 10% để đánh giá, cũng như số lớp lặp lại là 2. Khi thay đổi tỷ lệ phân chia dữ liệu hoặc số lượng của các lớp lặp lại thì đều không nhận được kết quả tốt như vậy cho cả hai mô hình. Ngoài ra, từ kết quả thực nghiệm cũng nhận thấy mô hình sẽ cho kết quả tốt nhất khi thực hiện với số bước lặp

trong huấn luyện là 20 như đã đặt trong bài báo gốc của nhóm Färber [11]. Việc tăng thêm số lần huấn luyện không góp phần nâng cao hiệu suất của mô hình.

Bảng 2.2. Kết quả chạy thực nghiệm khi điều chỉnh tham số của mô hình Enhanced-NCN

Tinh chỉnh tham số	Kích thước ẩn	Phân chia dữ liệu	Số lượng lớp	Epochs	Kích thước nhúng	Recall @10
Kích thước nhúng	128	[0.8, 0.1, 0.1]	1	20	128	0.0723
	128	[0.8, 0.1, 0.1]	1	20	164	**0.0921**
	128	[0.8, 0.1, 0.1]	1	20	196	0.0853
	128	[0.8, 0.1, 0.1]	1	20	256	0.0763
Số lớp trong mô hình	128	[0.8, 0.1, 0.1]	2	20	164	**0.1285**
	128	[0.8, 0.1, 0.1]	3	20	164	0.1115
Kích thước ẩn	164	[0.8, 0.1, 0.1]	2	20	164	0.0687
Phân chia dữ liệu	128	[0.7, 0.15, 0.15]	2	20	164	0.0968
	128	[0.6, 0.2, 0.2]	2	20	164	0.0870
Số lần lặp để huấn luyện	128	[0.8, 0.1, 0.1]	2	30	164	0.0823

2.6. Kết luận chương 2

Chương 2 tập trung vào các mô hình khuyến nghị trích dẫn sử dụng phương pháp lọc nội dung. Nghiên cứu của chương 2 đã cải tiến mô hình tiêu biểu theo hướng này, đó là mạng trích dẫn nơ-ron (NCN) của nhóm của Ebesu và Fang [10] và nhóm Färber [11]. Mô hình đã cải tiến Enhanced-NCN được tích hợp mô hình BERT để tiền xử lý văn bản của bài báo và thêm bộ mã hóa tiêu đề bài báo để có thể sử dụng như là dữ liệu đầu vào cho mô hình khuyến nghị trích dẫn. Mô hình Enhanced-NCN đã được thực nghiệm đánh giá hiệu suất với bộ dữ liệu arXiv CS và thu được kết quả tốt hơn đáng kể khi so sánh với công bố của Färber [11] khi sử dụng cùng chỉ số đánh giá Recall@10. Ngoài ra,

nghiên cứu của chương 2 này cũng đã cung cấp thông tin chi tiết về cách các tham số khác nhau có thể ảnh hưởng đến hiệu suất mô hình của Enhanced-NCN và cách sử dụng những thông tin chi tiết này để làm cho mô hình hiệu quả hơn trong tương lai.

Kết quả nghiên cứu của chương 2 này được công bố trong công trình [CT1] phần "Danh mục các công trình đã công bố của tác giả".

Bởi vì Enhanced-NCN là mô hình khuyến nghị trích dẫn tiếp cận theo hướng lọc nội dung và các dữ liệu đầu vào cho mô hình này chỉ là dạng văn bản, cho nên với các bộ dữ liệu chỉ chứa dữ liệu văn bản mà không chứa các siêu dữ liệu hoặc thông tin học thuật của bài báo thì mô hình Enhanced-NCN là một lựa chọn hợp lý. Mặc dù nội dung của bài báo ứng viên có phù hợp với ngữ cảnh trích dẫn hay không là yếu tố đầu tiên được xem xét đến, tuy nhiên các thông tin khác cũng đóng vai trò đáng kể như là đánh giá của cộng đồng nghiên cứu về bài báo (thể hiện qua số lượt trích dẫn), năm xuất bản và nơi xuất bản (các nhà nghiên cứu thường có xu hướng chọn lựa những công bố mới hơn và được xuất bản ở những tạp chí hay hội nghị có uy tín hơn). Đó cũng chính là nội dung nghiên cứu tiếp theo ở các chương 3 và chương 4.

CHƯƠNG 3. MÔ HÌNH RHN-DUALLCR SỬ DỤNG MẠNG HỒI QUY RHN VÀ PHÉP NHÚNG SCIBERT

Các hệ thống khuyến nghị trích dẫn luôn muốn hướng đến việc cung cấp một danh sách các bài báo có nội dung phù hợp và có chất lượng tốt cho người dùng. Do đó bên cạnh văn bản của bài báo đã được sử dụng như đã nghiên cứu ở chương 2, thì chất lượng của bài báo cũng là yếu tố cần phải đưa vào mô hình khuyến nghị. Đánh giá của cộng đồng nghiên cứu cho chất lượng một bài báo có thể được thể hiện bằng nhiều yếu tố, trong đó số lượt trích dẫn là một định lượng quan trọng. Nội dung nghiên cứu của chương 3 này sẽ tập trung vào các mô hình sử dụng kết hợp cả nội dung văn bản và đánh giá của người dùng, trong đó trọng tâm là trình bày chi tiết về mô hình khuyến nghị trích dẫn RHN-DualLCR từ việc cải tiến mô hình DualLCR của nhóm nghiên cứu Medić và Šnajder [12] [13] bằng cách sử dụng mạng hồi quy (*Recurrent Highway Networks*, *RHN*) và phép nhúng văn bản SciBERT.

Các đề xuất và kết quả nghiên cứu của chương này được công bố tại công trình [CT2] và [CT4] trong phần "Danh mục các công trình đã công bố của tác giả".

3.1. Phân tích vấn đề tồn tại của mô hình DualLCR

Kết hợp cả thông tin ngữ nghĩa và thông tin học thuật của bài báo là một hướng nghiên cứu được chú ý trong bài toán khuyến nghị trích dẫn [8] [81]. Trong số này, có nghiên cứu đáng chú ý về mô hình khuyến nghị trích dẫn cục bộ kép (DualLCR - **D**ual **L**ocal **C**itation **R**ecommendation) cho bài toán khuyến nghị trích dẫn được công bố vào năm 2020 bởi nhóm nghiên cứu Medić và Šnajder [12] [13]. Trong khi hầu hết các phương pháp trước đây chỉ sử dụng đoạn văn bản xung quanh vị trí trích dẫn để biểu diễn ngữ cảnh [15] [85] [86] [44] [16], thì trong nghiên cứu của họ, Medić và Šnajder [12] đã đề xuất tích hợp một cách biểu diễn ngữ cảnh bao gồm một phần thông tin toàn cục, chẳng hạn như tiêu đề và tóm tắt của bài báo được trích dẫn. Ngoài ra, trong mô hình DualLCR này còn có thêm một mô-đun thông tin học thuật để đánh giá chất lượng của bài báo dựa trên số lượng trích dẫn bài báo đó trong cộng đồng nghiên cứu.

Để khuyến nghị trích dẫn cho một truy vấn ngữ cảnh, đầu vào của mô hình DualLCR yêu cầu năm thông tin như sau từ cơ sở dữ liệu các bài báo khoa học: (1) ngữ cảnh trích dẫn dạng văn bản; (2) tiêu đề và tóm tắt của bài báo có chứa trích dẫn (gọi là bài báo trích dẫn); (3) tiêu đề và tóm tắt của bài báo đang được xem xét trích dẫn (được gọi là bài báo ứng viên); (4) danh sách tác giả của bài báo được trích dẫn và (5) tần suất trích dẫn mà bài báo ứng viên nhận được trong vòng (y) năm qua, cùng với tổng số trích dẫn của nó. Đầu ra của mô hình DualLCR là tổng điểm khuyến nghị cho biết bài báo ứng viên có phù hợp để được trích dẫn trong ngữ cảnh đầu vào hay không.

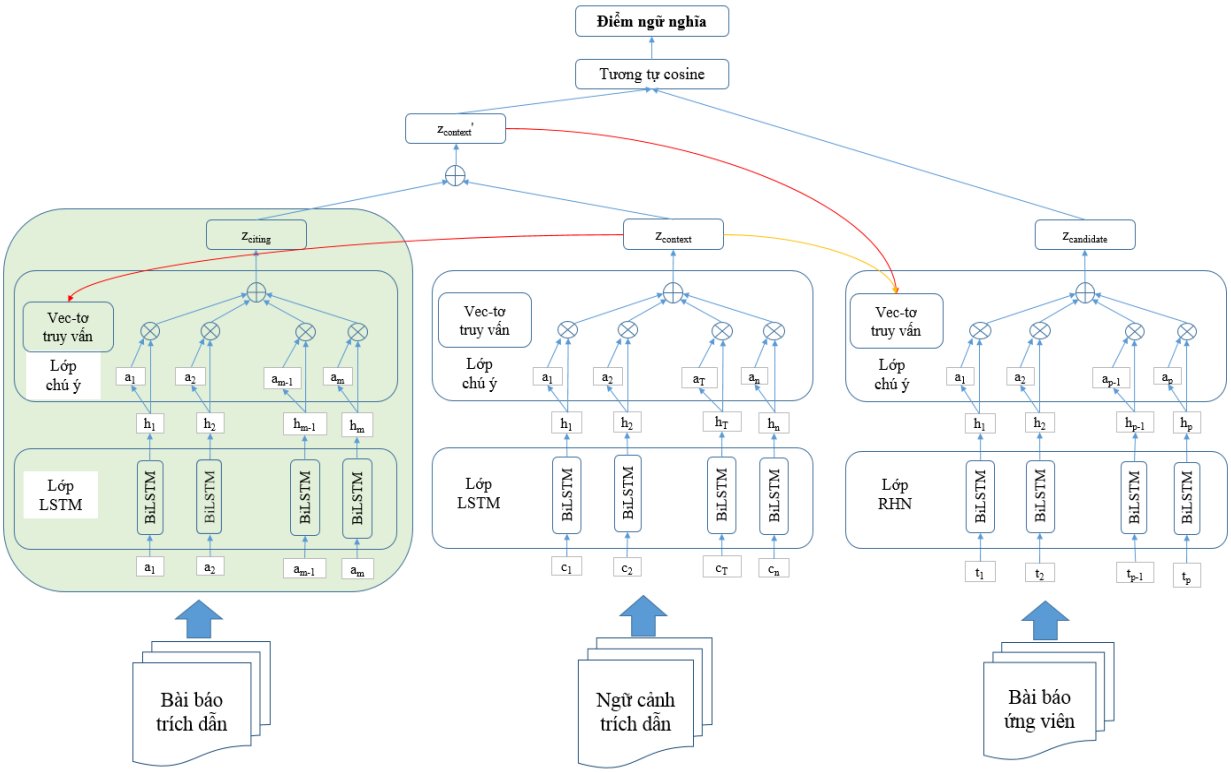
Gọi DualLCR là mô hình khuyến nghị trích dẫn kép bởi vì nó bao gồm hai mô-đun: mô-đun ngữ nghĩa và mô-đun thông tin học thuật. Kiến trúc của hai mô-đun này được thể hiện ở Hình 3.1 và 3.2. Điểm khuyến nghị trích dẫn cuối cùng là tổng có trọng số của các điểm do hai mô-đun này tạo ra. Trực giác đằng sau tổng điểm có trọng số là tùy thuộc vào ngữ cảnh, các tác giả có thể ưu tiên trích dẫn các bài viết có ảnh hưởng trong cộng đồng nghiên cứu (tức là các bài báo có điểm thông tin học thuật cao) hoặc các bài viết liên quan đến chi tiết cụ thể trong nghiên cứu của họ (ví dụ: một bài báo mà công trình của họ dựa trên đó hoặc một phương pháp đã có mà họ sử dụng). Theo trực giác, trong trường hợp trước, mô hình khuyến nghị nên đặt trọng số nhiều hơn vào điểm thông tin học thuật, trong khi ở trường hợp sau, dự kiến sẽ có trọng số cao hơn về điểm ngữ nghĩa.

3.1.1. Mô-đun ngữ nghĩa

Mô-đun ngữ nghĩa của mô hình DualLCR tính điểm khuyến nghị dựa trên sự tương đồng ngữ nghĩa giữa ngữ cảnh trích dẫn của bài viết và nội dung của các bài báo ứng viên (những là những bài báo có thể được trích dẫn). Cụ thể, mô-đun này một mạng nơ-ron LSTM hai chiều (BiLSTM) để biểu diễn cả ngữ cảnh trích dẫn, bài báo trích dẫn và thông tin toàn cục của bài báo đó (bao gồm tiêu đề và tóm tắt). Các đoạn văn bản được xử lý qua các lớp BiLSTM và lớp chú ý (*attention layer*), nhằm tìm ra các từ khóa liên quan trong ngữ cảnh trích dẫn. Các bước thực hiện chính bao gồm: (1) Tách và mã hóa các từ trong ngữ cảnh trích dẫn, tiêu đề và tóm tắt của bài báo. (2) Sử dụng lớp BiLSTM để tạo ra các biểu diễn ngữ nghĩa cho các đoạn văn bản này. (3) Sử dụng một

lớp chú ý để tập trung vào các từ trong ngữ cảnh trích dẫn có ý nghĩa quan trọng nhất. (4) Sau đó, mô hình tính toán độ tương đồng cosine giữa biểu diễn ngữ cảnh và biểu diễn bài báo ứng viên, từ đó cho ra điểm khuyến nghị dựa trên ngữ nghĩa.

Tương tự như nhóm nghiên cứu của Dai [17], mô-đun ngữ nghĩa cũng sử dụng bộ nhớ hai chiều BiLSTM [14] để biểu diễn bối cảnh trích dẫn, tiêu đề và tóm tắt của cả bài báo trích dẫn và bài báo ứng viên để trích dẫn. Văn bản trước khi đưa vào mô-đun ngữ nghĩa được phân chia thành các câu bằng thư viện SpaCy. Trong mỗi ngữ cảnh trích dẫn, trích dẫn mục tiêu và các trích dẫn khác được đánh dấu bằng các thẻ giữ chỗ TARGETCIT và OTHERCIT tương ứng. Tất cả ba đầu vào dạng văn bản đều được truyền qua hai lớp giống nhau là lớp BiLSTM và lớp chú ý.



Hình 3.1. Cấu trúc của mô-đun ngữ nghĩa trong mô hình DualLCR [12]

Gọi (n) là tổng số từ của một đoạn văn bản đầu vào, ký hiệu là: $s = (t_1,...,t_n)$. Mỗi từ (t_i) được ánh xạ tới vectơ nhúng d_e chiều $x_i \in \mathbb{R}^{d_e}$ để tạo ra một chuỗi có dạng như $x = (x_1, \ldots, x_n)$ bằng cách sử dụng phép nhúng đã được huấn luyện trước AI2 trong nghiên cứu của Bhagavatula [87]. Sau đó, chuỗi đã nhúng (x) được chuyển qua lớp BiLSTM với kích thước trạng thái ẩn là (d_h), trong đó đầu ra ($\boldsymbol{h}_i$) ở mỗi bước (i) được

hình thành bằng cách nối các trạng thái ẩn tiến và lùi: $\boldsymbol{h_i} = [\rightarrow \boldsymbol{h}_i; \leftarrow \boldsymbol{h}_i], \boldsymbol{h}_i \in \mathbb{R}^{2d_h}$. Các trạng thái ẩn của chuỗi đầu vào ($\boldsymbol{s}$) được chuyển qua lớp chú ý [84] để tạo ra chuỗi nhúng cuối cùng ($\mathbf{z}_s$). Cho vectơ truy vấn đầu vào ($\boldsymbol{q}$) và vectơ trạng thái ẩn ($\boldsymbol{h}_i$), điểm chú ý (*attention score*) cho mỗi bước (i) được tính toán theo công thức (3.1).

$$a_i = \boldsymbol{v} \cdot tanh(W \cdot [\boldsymbol{q}; \boldsymbol{h}_i]) \quad (3.1)$$

trong đó ($\boldsymbol{v}$) và (W) là tham số của mô hình DualLCR. Điểm chú ý được chuẩn hóa, áp dụng cho các trạng thái ẩn tương ứng và được tính tổng để tạo ra chuỗi nhúng (z_s) cuối cùng. Các vectơ truy vấn (q) khác nhau sẽ được sử dụng tùy thuộc vào loại đầu vào.

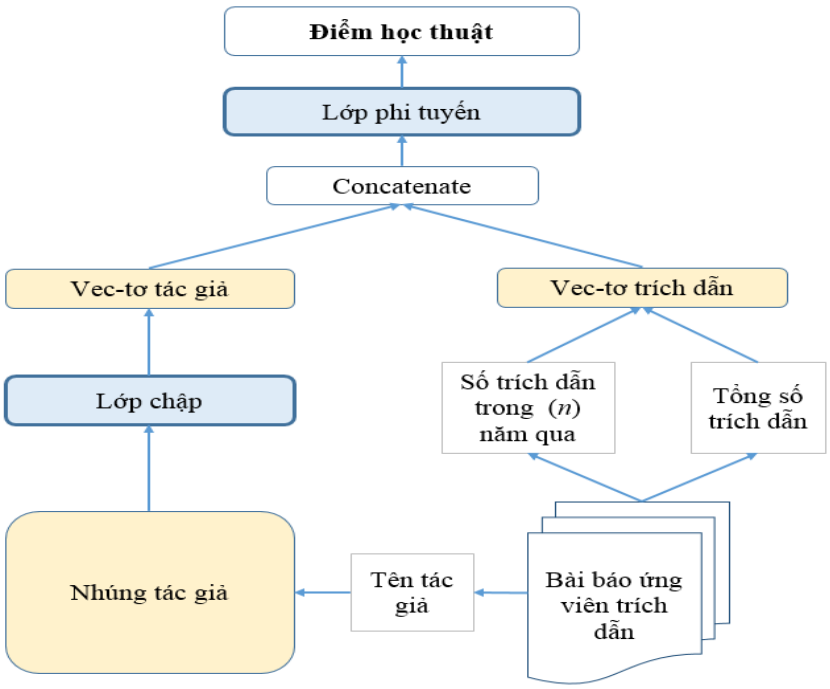
Đối với ngữ cảnh trích dẫn, vectơ truy vấn (q) là trạng thái ẩn tương ứng với vị trí của trích dẫn ($\boldsymbol{h_T}$), nhằm tập trung biểu diễn ngữ cảnh vào trích dẫn cụ thể đang được dự đoán. Đối với văn bản của bài báo trích dẫn, vectơ truy vấn (q) là biểu diễn cuối cùng của ngữ cảnh ($\mathbf{z}_{\text{context}}$), trong khi đối với văn bản của bài báo ứng viên trích dẫn, vectơ truy vấn (q) là tổng của biểu diễn bài báo trích dẫn và biểu diễn ngữ cảnh ($\mathbf{z}_{\text{citing}} + \mathbf{z}_{\text{context}}$). Điều này cho phép mô-đun ngữ nghĩa tập trung vào thông tin cụ thể của ngữ cảnh trong cả bài báo trích dẫn và bài báo được trích dẫn. Cụ thể hơn, bằng cách sử dụng trạng thái ẩn của vị trí trích dẫn làm vectơ truy vấn để tính toán điểm chú ý trên các từ trong ngữ cảnh trích dẫn, mô-đun này sẽ tập trung vào các từ trong ngữ cảnh có liên quan đến việc tạo ra biểu diễn cho vị trí trích dẫn. Tương tự, bằng cách sử dụng biểu diễn ngữ cảnh trích dẫn làm truy vấn cho văn bản của bài báo trích dẫn (hoặc bài báo ứng viên), mô-đun ngữ nghĩa sẽ tập trung vào các từ trong văn bản có liên quan đến ngữ cảnh trích dẫn, vì văn bản của bài báo mô tả nhiều khía cạnh khác nhau và không phải tất cả đều có liên quan như nhau đến ngữ cảnh hiện tại.

Với ngữ cảnh trích dẫn (c) và bài báo ứng viên (p), hàm tính điểm ngữ nghĩa $s_{\text{sem}}(c, p)$ được định nghĩa là độ tương đồng cosine giữa biến đổi nhúng của ngữ cảnh nâng cao và biến đổi nhúng của bài báo ứng viên.

3.1.2. Mô-đun thông tin học thuật

Khi bối cảnh ngữ nghĩa chấp nhận một số trích dẫn, Medić và Šnajder [12] cho rằng các tác giả nói chung sẽ muốn trích dẫn các bài báo nổi tiếng trong cộng đồng nghiên cứu. Đây chính là nhiệm vụ của mô-đun thông tin học thuật. Mô-đun này lấy tên

tác giả và số lượng trích dẫn của bài báo trích dẫn ứng viên (*p*) làm đầu vào và tạo ra một điểm thông tin học thuật duy nhất.



Hình 3.2. Cấu trúc của mô-đun thông tin học thuật trong mô hình DualLCR [12]

Cấu trúc của mô-đun học thuật được thể hiện ở Hình 3.2. Tương tự như cách làm của Ebesu và Fang [10], mô hình DualLCR thể hiện tên tác giả bài báo dưới dạng biến đổi nhúng. Danh sách tên tác giả của bài báo (a) = (a_1, . . ., a_m) trước tiên được chuyển thành chuỗi các biến đổi nhúng của tên tác giả ($\boldsymbol{a_e}$) = (a_{e1}, ..., a_{em}). Tiếp theo, chuỗi ($\boldsymbol{a_e}$) được chuyển qua một lớp tích chập, sau đó là phép biến đổi tổng hợp tối đa và phi tuyến (*max-pooling and non-linear transformation*) để cuối cùng tạo ra biến đổi nhúng cho danh sách tên tác giả, rồi biến đổi nhúng này được kết hợp với tổng số trích dẫn và số lượng trích dẫn của bài viết trong (y) năm qua. Cuối cùng, toàn bộ vectơ được truyền qua lớp phi tuyến tính để tạo ra điểm thông tin học thuật $s_{bib}(p)$.

3.1.3. Điểm khuyến nghị cuối cùng

Điểm khuyến nghị cuối cùng được tính dựa trên sự kết hợp của hai thành phần chính: điểm ngữ nghĩa và điểm thông tin học thuật. Điểm ngữ nghĩa được tính từ mô-đun ngữ nghĩa. Đây là điểm số dựa trên sự tương đồng ngữ nghĩa giữa ngữ cảnh của trích dẫn và nội dung của bài báo ứng viên (bài có khả năng được trích dẫn). Điểm ngữ nghĩa

giúp mô hình DualLCR tập trung vào các từ trong ngữ cảnh liên quan đến trích dẫn và bài báo. Ngược lại, điểm thông tin học thuật dựa trên thông tin học thuật của bài báo ứng viên, bao gồm số lượng trích dẫn của bài báo và các tác giả của bài báo đó.

Điểm khuyến nghị cuối cùng $s_{fin}(c, p)$ là tổng trọng số của điểm ngữ nghĩa $s_{sem}(c, p)$ và điểm thông tin học thuật $s_{bib}(p)$. Các trọng số này được xác định dựa trên biểu diễn ngữ cảnh ($\mathbf{z}_{context}$) của bài báo trích dẫn thông qua một lớp phi tuyến. Ý tưởng là nếu ngữ cảnh trích dẫn cụ thể liên quan nhiều hơn đến ngữ nghĩa của bài báo, trọng số sẽ nghiêng về điểm ngữ nghĩa, ngược lại nếu bài báo nổi tiếng trong cộng đồng, trọng số sẽ nghiêng về điểm thư mục. Điểm này giúp mô hình DualLCR đưa ra các khuyến nghị trích dẫn tối ưu dựa trên cả nội dung ngữ nghĩa và mức độ phổ biến của bài báo trong cộng đồng nghiên cứu.

3.1.4. Hàm mất mát (*loss function*)

Mô hình DualLCR sử dụng hàm mất mát bộ ba để tối đa hóa điểm khuyến nghị cho các cặp bài báo và ngữ cảnh trích dẫn đúng, đồng thời giảm thiểu điểm đề xuất cho các cặp trích dẫn sai. Tập huấn luyện chứa bộ ba (c, p_+, p_-), trong đó (c) là ngữ cảnh, (p_+) bài báo được trích dẫn và (p_-) bài báo không được trích dẫn trong ngữ cảnh. Khi lấy mẫu các bài báo, mô hình DualLCR áp dụng tính năng lọc theo thời gian (*timebased filtering*) để chọn các trường hợp phủ định trong bộ ba, tức là mô hình chỉ xem xét các bài viết được xuất bản trước bài viết trích dẫn. Hàm mất mát bộ ba với hàm tính điểm cho trước $s(c, p_*)$ được tính dựa trên công thức (3.2).

$$\boldsymbol{L}_s = \max(0, s(c, p_-) - s(c, p_+) + m) \qquad (3.2)$$

trong đó $s(c, p_*)$ là điểm khuyến nghị cho bài báo (p_*) và (m) là mức chênh lệch được sử dụng để nâng cao sự khác biệt giữa các điểm. Mô hình DualLCR sử dụng tổng của ba hàm mất mát làm hàm mất mát tổng thể, $\boldsymbol{L} = \boldsymbol{L}_{sem} + \boldsymbol{L}_{bib} + \boldsymbol{L}_{fin}$, trong đó $\boldsymbol{L}_{sem}$, $\boldsymbol{L}_{bib}$ và $\boldsymbol{L}_{fin}$ tương ứng với điểm ngữ nghĩa, thông tin học thuật và khuyến nghị cuối cùng.

3.1.5. Thảo luận về các vấn đề còn tồn tại của mô hình DualLCR

Trong khi hầu hết các mô hình khuyến nghị trích dẫn trước đây khi biểu diễn ngữ cảnh thì chỉ sử dụng văn bản xung quanh vị trí trích dẫn, thì mô hình DualLCR [12] đã

tăng cường biểu diễn ngữ cảnh bằng thông tin tổng thể của bài báo. Cụ thể, mô hình DualLCR đã đưa tiêu đề và phần tóm tắt của bài báo trích dẫn vào phần biểu diễn ngữ cảnh, do đó, so với các mô hình hiện tại thì mô hình DualLCR đã được nâng lên đáng kể về mặt hiệu suất. Tuy nhiên, mô hình này vẫn còn những nhược điểm như sau:

(1) Bộ nhớ hai chiều BiLSTM:

Mô hình DualLCR vẫn đang sử dụng bộ nhớ hai chiều BiLSTM [14] để biểu diễn ngữ cảnh tăng cường. BiLSTM thường được sử dụng cho các tác vụ xử lý ngôn ngữ tự nhiên như phân loại văn bản (*text classification*) hay dịch máy, tuy nhiên BiLSTM cũng có một số hạn chế như:

· BiLSTM có thể gặp phải vấn đề biến mất hoặc bùng nổ gradient, xảy ra khi gradient trở nên quá nhỏ hoặc quá lớn trong quá trình truyền ngược (*backpropagation*), làm cho mạng khó huấn luyện

· BiLSTM có thể tốn kém về mặt tính toán và tốn nhiều bộ nhớ vì nó yêu cầu hai lớp LSTM cho mỗi hướng và một lớp móc nối (*concatenation layer*) để hợp nhất các đầu ra.

· BiLSTM có thể nhạy cảm với nhiễu và các giá trị ngoại lệ trong dữ liệu vì nó dựa trên giả định rằng chuỗi đầu vào trơn tru và nhất quán, tuy nhiên văn bản của bài báo khoa học thì phải không như vậy.

· BiLSTM có thể gặp khó khăn trong việc lập mô hình phụ thuộc lâu dài vì thông tin từ các phần tử ở xa có thể bị loãng hoặc bị lãng quên theo thời gian.

(2) Phép nhúng ngữ cảnh AI2:

Mô hình DualLCR đang sử dụng phép nhúng được sử dụng trong nghiên cứu của nhóm Bhagavatula [87] để nhúng ngữ cảnh trích dẫn cũng như thông tin (bao gồm tiêu đề và phần tóm tắt) của bài báo trích dẫn và được trích dẫn. Đây là phép nhúng AI2 (*AI2 embedding*) được công bố từ năm 2017 bởi nhóm nghiên cứu đến từ Viện trí tuệ nhân tạo Allen[20] (*Allen Institute for Artificial Intelligence, AI2*). Phép nhúng AI2 dựa trên ký tự và có thể nắm bắt các đặc điểm phức tạp của việc sử dụng từ (như cú pháp và ngữ

[20] https://allenai.org/

nghĩa) cũng như cách sử dụng này khác nhau tùy theo ngữ cảnh ngôn ngữ. Không giống như cách nhúng từ truyền thống, phép nhúng AI2 tạo ra các cách nhúng có chức năng của toàn bộ câu đầu vào, cung cấp sự hiểu biết phong phú hơn về nghĩa của từ. Tuy nhiên phép nhúng AI2 này vẫn bộc lộ những hạn chế như sau so với phép nhúng SciBERT:

· Hạn chế về ngữ cảnh: Các phần nhúng AI2 được huấn luyện trên các ngữ liệu cụ thể, điều này có thể hạn chế tính hiệu quả của chúng trong các lĩnh vực hoặc bối cảnh không được thể hiện rõ trong dữ liệu huấn luyện. Ví dụ: các phần nhúng được huấn luyện trên văn bản web thông dụng có thể không hoạt động tốt trên các văn bản khoa học chuyên ngành.

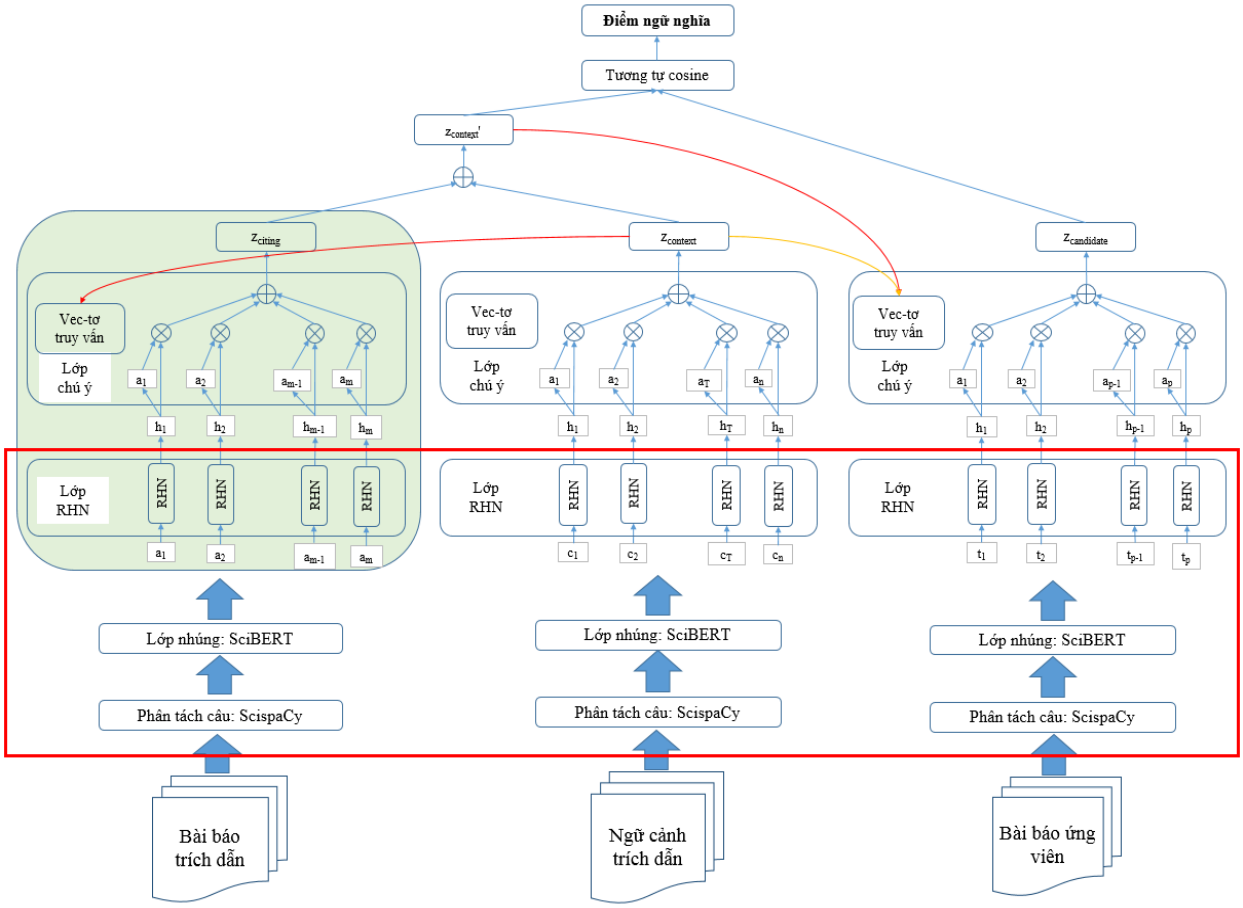
· Biểu diễn tĩnh: Nếu phần nhúng AI2 thuộc loại truyền thống, không theo ngữ cảnh (như BERT hoặc SciBERT), thì mỗi từ có một cách biểu diễn cố định bất kể ngữ cảnh của nó. Điều này có thể gây khó khăn cho các từ đa nghĩa (từ có nhiều nghĩa) vì sắc thái của ý nghĩa trong các bối cảnh khác nhau bị mất đi.

· Xu hướng: Các phần nhúng được huấn luyện trước có thể phản ánh và duy trì các thành kiến có trong dữ liệu huấn luyện. Đây là một vấn đề thiết yếu khi các phần nhúng được sử dụng trong quá trình ra quyết định hoặc trong bối cảnh mà sự công bằng và vô tư là rất quan trọng.

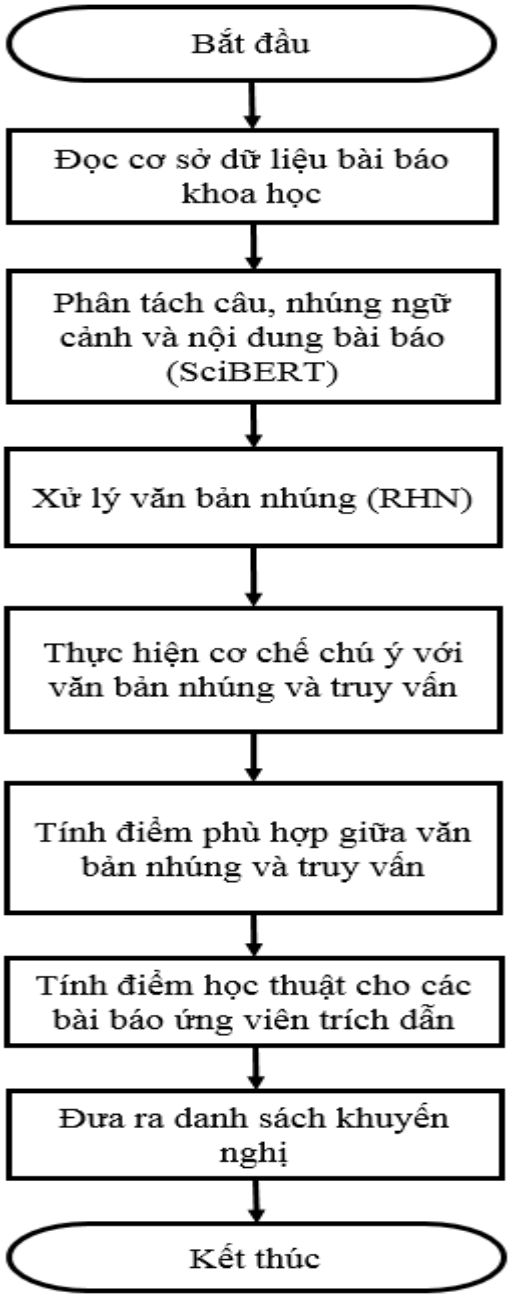
· Từ ngoài từ vựng: Các từ nhúng được huấn luyện trước gặp khó khăn với những từ không được nhìn thấy trong quá trình huấn luyện. Điều này đặc biệt có vấn đề trong các lĩnh vực như khoa học và công nghệ, nơi các thuật ngữ mới liên tục được đặt ra.

3.2. Mô hình RHN-DualLCR

Dựa trên những phân tích về hạn chế của mô hình DualLCR [12] hiện tại, luận án đã áp dụng hai phương pháp sau đây để nâng cao hiệu suất của mô hình DualLCR: (1) thay thế bộ nhớ hai chiều BiLSTM bằng mạng hồi quy RHN [20] và (2) thay thế phép nhúng AI2 từ công bố của nhóm nghiên cứu Bhagavatula [87] bằng phép nhúng SciBERT [18]. Phần thay đổi của mô hình DualLCR này được khoanh đỏ trong Hình 3.3. Mô hình DualLCR đã cải tiến được đặt tên là RHN-DualLCR.



Hình 3.3. Cấu trúc của mô-đun ngữ cảnh trong mô hình RHN-DualLCR



Hình 3.4. Sơ đồ khối xử lý tuần tự của mô hình RHN-DualLCR

3.2.1. Mạng hồi quy RHN

Với các mô hình huấn luyện, quá trình chuyển đổi phi tuyến từ bước này sang bước khác trong các nhiệm vụ xử lý tuần tự rất phức tạp, điều này khiến cho việc huấn luyện các mạng nơ-ron hồi quy với các hàm chuyển tiếp "*sâu*" trở thành thách thức ngay cả với các bộ nhớ hai chiều BiLSTM. Nhằm giải quyết vấn đề này, nhóm nghiên cứu

Zilly [20] đã giới thiệu một nghiên cứu lý thuyết mới về các mạng hồi quy sử dụng lý thuyết vòng tròn Geršgorin [88]. Nghiên cứu này đã giúp hiểu thêm về mô hình hóa và tối ưu hóa, cũng như cải thiện hiệu suất của mô hình BiLSTM. Dựa trên phân tích này, họ đã công bố mô hình mạng hồi quy RHN, là một mô hình mở rộng của BiLSTM để cho phép chuyển đổi từng bước sâu hơn. RHN được coi là một mô hình mạnh mẽ được tạo ra để tận dụng độ sâu ngày càng tăng trong quá trình chuyển đổi lặp lại trong khi vẫn duy trì các đặc điểm huấn luyện dễ dàng của BiLSTM. Kiến trúc đề xuất đã được đánh giá trong các thực nghiệm mô hình hóa ngôn ngữ khác nhau để chứng minh hiệu suất và tính hiệu quả. Ví dụ, trong một nghiên cứu sử dụng kho ngữ liệu Penn Treebank [89], chỉ cần tăng độ sâu chuyển tiếp từ 1 lên 10, độ phức tạp ở cấp độ từ của mô hình đã giảm từ 90.6 lên 65.4 trong khi vẫn giữ nguyên số tham số. Hơn nữa, khi đánh giá trên các bộ dữ liệu Wikipedia lớn hơn [90] về dự đoán ký tự (text8 và enwik8), RHN đã đánh bại tất cả các mô hình trước đó, đạt được entropy 1.27 bit cho mỗi ký tự. Theo cách tiếp cận của luận án, RHN được áp dụng trong mô hình khuyến nghị trích dẫn sẽ đạt được sự biểu diễn theo ngữ cảnh tuần tự chuyên sâu hơn về mối quan hệ giữa các bài báo mục tiêu và các trích dẫn ứng viên. Do đó, áp dụng RHN được coi là một hướng đi đầy hứa hẹn để nâng cao hiệu quả của các hệ thống khuyến nghị trích dẫn hiện có.

Vì mô hình RHN-DualLCR được xây dựng trên mô hình mạng hồi quy RHN thay thế cho BiLSTM [14] của mô hình DualLCR ban đầu, luận án sẽ so sánh độ phức tạp của mô hình RHN với BiLSTM. Độ phức tạp của mô hình RHN phụ thuộc vào một số yếu tố, chẳng hạn như số lượng nút trong đồ thị, số lớp trong quá trình chuyển đổi hồi quy, số lượng đơn vị ẩn trong mỗi lớp và số lượng tham số trong mỗi lớp. Tính toán chính của mô hình RHN là hàm chuyển đổi hồi quy, được định nghĩa là:

$$\mathbf{h}_i^{(l)} = \delta(\mathbf{W}_h^{(l)} \mathbf{x}_i + \mathbf{b}_h^{(l)}) \tag{3.3}$$

$$\mathbf{t}_i^{(l)} = \delta(\mathbf{W}_t^{(l)} \mathbf{x}_i + \mathbf{b}_t^{(l)}) \tag{3.4}$$

$$\mathbf{c}_i^{(l)} = \delta(\mathbf{W}_c^{(l)} \mathbf{x}_i + \mathbf{b}_c^{(l)}) \tag{3.5}$$

$$\mathbf{x}_i^{(l+1)} = \mathbf{t}_i^{(l)} \odot \mathbf{h}_i^{(l)} + (1 - \mathbf{t}_i^{(l)}) \odot \mathbf{c}_i^{(l)} \tag{3.6}$$

trong đó $\boldsymbol{x}_i$ là vectơ đầu vào của nút thứ (i), $\boldsymbol{x}_i^{(l+1)}$ là vectơ đầu ra của nút thứ (i) sau lớp thứ l, $\boldsymbol{h}_i^{(l)}$ là vectơ trạng thái ẩn của nút thứ (i) ở lớp thứ (l), $\boldsymbol{t}_i^{(l)}$ là vectơ cổng chuyển đổi (*transform gate vector*) của nút thứ (i) ở lớp thứ (l), $\boldsymbol{c}_i^{(l)}$ là vectơ cổng mang (*carry gate vector*) của nút thứ (i) ở lớp thứ (l), $\boldsymbol{W}_h^{(l)}$, $\boldsymbol{W}_t^{(l)}$, $\boldsymbol{W}_c^{(l)}$ là các ma trận trọng số và $\boldsymbol{b}_h^{(l)}$, $\boldsymbol{b}_t^{(l)}$, $\boldsymbol{b}_c^{(l)}$ là các vectơ điều chỉnh sai lệch (*bias vector*).

Giả sử vectơ đầu vào $\boldsymbol{x}_i$ có d chiều, vectơ trạng thái ẩn $\boldsymbol{h}_i^{(l)}$ có (h) chiều và số lớp trong quá trình chuyển đổi hồi quy là ($\boldsymbol{L}$), độ phức tạp tính toán (*time complexity*) và độ phức tạp lưu trữ (*space complexity*) của mô hình RHN được tính như sau:

Độ phức tạp về thời gian của thuật toán tỷ lệ thuận với số phép toán cần thiết để tính vectơ đầu ra $\boldsymbol{x}_i^{(l+1)}$ từ vectơ đầu vào $\boldsymbol{x}_i$. Mỗi lớp bao gồm bốn phép nhân vectơ ma trận, bốn phép cộng vectơ và bốn kích hoạt phi tuyến theo phần tử. Mỗi phép nhân vectơ ma trận yêu cầu các phép toán $O(dh)$, mỗi phép cộng vectơ yêu cầu các phép toán $O(h)$ và mỗi kích hoạt phi tuyến theo phần tử yêu cầu các phép toán $O(h)$. Do đó, mỗi lớp trong mô hình RHN yêu cầu thao tác $O(dh + 3h)$. Vì có ($\boldsymbol{L}$) lớp nên tổng độ phức tạp về thời gian là $O(L(dh + 3h))$ [20]. Theo phương pháp tính toán tương tự, độ phức tạp về thời gian của BiLSTM có ($\boldsymbol{L}$) lớp sẽ là $O(L(4h(d + 3 + h)))$.

Độ phức tạp lưu trữ tỷ lệ thuận với số lượng tham số cần thiết để lưu trữ ma trận trọng số và vectơ điều chỉnh sai lệch. Mỗi ma trận trọng số có số chiều $h \times d$ và mỗi vectơ điều chỉnh sai lệch có số chiều h. Do đó, mỗi lớp yêu cầu tham số $O(hd + h)$. Vì có ($\boldsymbol{L}$) lớp nên tổng độ phức tạp lưu trữ sẽ là $O(L(hd + h))$.

3.2.2. Tiếp cận học mô hình biểu diễn các bài báo khoa học bằng SciBERT

Phép nhúng SciBERT được công bố bởi nhóm nghiên cứu Beltagy [18] là một biến thể chuyên biệt của BERT với mục tiêu huấn luyện đặc biệt cho văn bản khoa học. SciBERT được huấn luyện trên một lượng lớn các bài báo khoa học, tài liệu nghiên cứu và nội dung học thuật khác. Mục tiêu của SciBERT là nắm bắt ngôn ngữ và cấu trúc độc đáo có trong tài liệu khoa học, giúp nó trở nên hiệu quả hơn đối với các nhiệm vụ liên quan đến phân tích văn bản khoa học. SciBERT đã được chứng minh là có hiệu quả trong các nhiệm vụ xử lý ngôn ngữ tự nhiên trong văn bản khoa học khác nhau và đã góp phần tạo nên những tiến bộ trong việc trích xuất thông tin có giá trị từ tài liệu khoa học.

SciBERT được huấn luyện trên bộ 1.14 triệu tài liệu nghiên cứu được chọn ngẫu nhiên từ Semantic Scholar [26]. Các bài báo này bao gồm 18% bài báo về khoa học máy tính và 82% bài báo từ lĩnh vực y sinh rộng hơn và toàn bộ nội dung của các bài báo, thay vì chỉ là phần tóm tắt, được sử dụng để huấn luyện. Kho văn bản này có trung bình 154 câu trên mỗi bài báo, tương đương với 2,769 mã thông báo, nên kích thước của toàn bộ kho văn bản là 3.17 tỷ mã thông báo. Con số này có thể so sánh với kích thước của kho dữ liệu được sử dụng để huấn luyện BERT là 3.3 tỷ mã thông báo. Thử nghiệm của nhóm nghiên cứu cho thấy với bộ dữ liệu của các bài báo khoa học như ACL-ARC hay RefSeer, kết quả của SciBERT tốt hơn nhiều so với BERT [68]. Trong nghiên cứu được trình bày ở chương 3 này, mô hình SciBERT đã được áp dụng để nhúng bối cảnh trích dẫn, tiêu đề và tóm tắt của cả các bài báo được trích dẫn và trích dẫn trước khi đưa nó vào mạng hồi quy RHN của mô-đun ngữ nghĩa. Hiệu suất ấn tượng của mô hình SciBERT trong xử lý ngôn ngữ tự nhiên của các bài báo khoa học được kỳ vọng sẽ nâng cao hiệu quả của mô hình DualCLR hiện tại đối với vấn đề khuyến nghị trích dẫn.

3.3. Tiến hành thực nghiệm với mô hình RHN-DualLCR

Nội dung của phần này trình bày chi tiết về cách thức xây dựng mô hình RHN-DualLCR, cũng như mô tả về các bộ dữ liệu và các chỉ số được dùng để đánh giá hiệu suất của mô hình.

3.3.1. Cài đặt mô hình RHN-DualLCR

Luận án đã cài đặt lại mã nguồn[21] của mô hình khuyến nghị trích dẫn DualLCR từ bài báo của nhóm nghiên cứu Medić và Šnajder [12] bằng cách thêm SciBERT [18], mô hình tốt nhất hiện nay để xử lý ngôn ngữ tự nhiên của các bài báo khoa học, cũng như sử dụng RHN [20] để thay thế lớp BiLSTM hiện tại. Luận án đã sử dụng Python phiên bản 3.8.5 và PyTorch phiên bản 1.7.1 để xây dựng mô hình nâng cao. Đối với mô hình SciBERT được huấn luyện trước, mô hình RHN-DualLCR đã sử dụng AutoTokenizer và AutoModel từ thư viện transformers của Python. Luận án phân chia

[21] https://github.com/zoranmedic/DualLCR

các câu trong bối cảnh trích dẫn và tóm tắt của bài viết bằng cách sử dụng ScispaCy[22] [91] đã được tối ưu hóa cho văn bản khoa học. Sau khi được phân tách, các câu văn trong các bài báo thực hiện nhúng theo mô hình SciBERT. Mô hình RHN-DualLCR được chạy trên môi trường Linux 3.10.0-x86-64, bộ xử lý NVIDIA GPU H100.

Để đánh giá sự cải tiến các kỹ thuật khác nhau khi áp dụng cho bài toán khuyến nghị trích dẫn, cả 2 mô hình DualLCR và RHN-DualLCR đều có 3 biến thể được thể hiện như ở Bảng 3.1.

Bảng 3.1. Các biến thể của mô hình DualLCR và RHN-DualLCR

Tên biến thể	**Sử dụng cả điểm ngữ cảnh và thông tin học thuật**	**Bao gồm các thông tin toàn cục**	**Sử dụng trọng số khi tính điểm cuối cùng**
DualCon-ws	Có	Không	Có
DualEnh-s	Có	Có	Không
DualEnh-ws	Có	Có	Có

Luận án so sánh kết quả thực nghiệm với 3 biến thể DualCon-ws, DualEnh-s và DualEnh-ws vì 3 biến thể này cho kết quả tốt nhất với mô hình DualLCR [12]. Biến thể DualCon-ws chỉ sử dụng ngữ cảnh trích dẫn mà không sử dụng các thông tin bổ sung từ bài báo (tiêu đề và tóm tắt) để huấn luyện mô hình. Biến thể DualEnh-s không sử dụng trọng số giữa 2 mô-đun ngữ nghĩa và mô-đun thông tin học thuật (có nghĩa là coi sự phù hợp với ngữ cảnh trích dẫn và độ nổi tiếng của bài báo là như nhau). Biến thể DualEnh-ws sử dụng cả thông tin bổ sung và trọng số phù hợp ngữ cảnh và điểm thông tin học thuật. Để thực thi ba biến thể của mô hình này là DualCon-ws, DualEnh-s và DualEnh-ws, việc cài đặt tham số như trong Bảng 3.2. Mô hình RHN-DualLCR thu được kết quả tốt nhất khi sử dụng các tham số sau: kích thước ẩn = 200, độ sâu hồi quy = 6 và số lớp = 2. Chi tiết về các siêu tham số khác, lọc trước và huấn luyện mô hình được trình bày trong phần phụ lục của bài báo gốc [12].

[22] https://github.com/allenai/SciSpaCy

Bảng 3.2. Thiết lập tham số cho mô hình RHN-DualLCR

Tên biến thể	Thiết lập tham số		
	-dual	**-global_info**	**-weighted_sum**
DualCon-ws	Có	Không	Có
DualEnh-s	Có	Có	Không
DualEnh-ws	Có	Có	Có

3.3.2. Mô tả về bộ dữ liệu thực nghiệm

Để so sánh hiệu năng thì mô hình RHN-DualLCR cũng được đánh giá trên hai tập dữ liệu RefSeer và ACL-ARC như đã sử dụng trong bài báo gốc của Medić và Šnajder [12]. Cả hai bộ dữ liệu này cũng thường được sử dụng để tính toán hiệu suất của các hệ thống khuyến nghị trích dẫn được công bố gần đây [10] [15] [16]. Theo chiến lược xác thực chéo (*cross-validation strategy*), khi thực nghiệm đã giữ nguyên các tiêu chuẩn khi so sánh với thành tích của hai bài báo gốc và các mô hình tiên tiến khác. Cũng giống như việc thực hiện thử nghiệm hai bài báo gốc và các mô hình tiên tiến sử dụng hai tiêu chuẩn đánh giá này, đồng thời để đảm bảo tính công bằng khi thực hiện so sánh hiệu suất, luận án vẫn giữ nguyên tỉ lệ các tập con huấn luyện/xác nhận/kiểm tra của tập dữ liệu ACL-ARC và RefSeer. Bảng 3.3 cho thấy số liệu thống kê của hai bộ dữ liệu này.

Bảng 3.3. Thống kê tập dữ liệu theo số lượng bối cảnh trích dẫn và bài báo [12]

Tập dữ liệu	Huấn luyện	Xác thực	Kiểm tra	Bài báo
ACL-ARC	30,390	9,381	9,585	19,711
RefSeer	3,521,582	124,551	126,021	624,957

Bộ dữ liệu RefSeer[23] bao gồm các bài báo và bối cảnh trích dẫn từ các lĩnh vực khoa học và kỹ thuật khác nhau. Ngữ cảnh trích dẫn là các đoạn văn bản trích dẫn dài 200 ký tự trước và sau vị trí trích dẫn trong bài báo. Tập dữ liệu này đã được lọc để loại bỏ khoảng 230,000 bài viết có tiêu đề và tóm tắt dưới 100 ký tự, được cho là do lỗi phân tích cú pháp. Bởi vì năm xuất bản bị thiếu trong phần lớn các bài báo, luận án chỉ xem

[23] https://github.com/chbrown/refseer

xét tổng số trích dẫn trong mô-đun thông tin khoa học mà không thực hiện bất kỳ bộ lọc thời gian nào đối với hàm mất mát bộ ba.

Bộ dữ liệu ACL-ARC (*Association for Computational Linguistics* (*ACL*) - *Anthology Reference Corpus*) bao gồm các bài báo đã được xuất bản tại các địa điểm và hội nghị của ACL. Tương tự như các nghiên cứu đã có, luận án đã tạo ra một phiên bản mã hóa ngữ cảnh được xử lý bằng bộ mã hóa SciBERT, với ngữ cảnh trích dẫn được đặt là 200 và 600 ký tự trước và sau trích dẫn. Để phân tích mức độ ảnh hưởng của các kích thước ngữ cảnh khác nhau đến hiệu suất của mô hình, luận án sử dụng hai biến thể của tập dữ liệu: ACL-600 (trong đó kích thước ngữ cảnh trích dẫn là 600 ký tự) và ACL-200 (trong đó kích thước ngữ cảnh trích dẫn là 200 ký tự, tương đương với RefSeer). Vì năm xuất bản được bao gồm trong định danh của mỗi bài báo nên thông tin này được sử dụng khi xây dựng hàm mất mát bộ ba cho các tính năng đếm trích dẫn và lọc theo thời gian. Bộ dữ liệu này được tách thành ba tập con dựa trên thời gian công bố của bài báo: tập huấn luyện chứa bối cảnh từ năm 2009 đến 2013, tập xác thực chứa bối cảnh từ năm 2014 và tập kiểm tra chứa bối cảnh từ năm 2015 trở đi.

3.3.3. Phương pháp đánh giá mô hình

Trong nghiên cứu của Medić và Šnajder [12] đang sử dụng 2 chỉ số đánh giá hiệu suất của mô hình khuyến nghị trích dẫn là Recall@K và xếp hạng đối ứng trung bình MRR, do đó để so sánh hiệu suất của mô hình đã cải tiến RHN-DualLCR cũng sẽ được đánh giá dựa trên 2 tiêu chí này.

Xếp hạng đối ứng trung bình MRR là một tiêu chí thường được sử dụng trong các hệ thống truy xuất thông tin và công cụ tìm kiếm để đánh giá tính hiệu quả của các thuật toán xếp hạng. MRR tập trung vào vị trí của kết quả đúng đầu tiên trong danh sách khuyến nghị. Nó đo lường khả năng mô hình đặt kết quả đúng ở những vị trí cao nhất, tức là đo độ ưu tiên của kết quả chính xác nhất. MRR quan trọng trong bối cảnh người dùng chỉ quan tâm đến những kết quả đứng đầu, như trường hợp họ chỉ cần trích dẫn một vài bài báo quan trọng nhất. Bởi vì cả 2 mô hình DualLCR và RHN-DualLCR đều quan tâm đến độ ưu tiên của các bài báo khi đưa ra danh sách khuyến nghị cho người dùng, do đó việc sử dụng tiêu chí MRR để đánh giá hiệu suất của mô hình là cần thiết.

MRR được tính bằng trung bình của các nghịch đảo thứ hạng của các kết quả tìm kiếm đối với một truy vấn. Nghịch đảo thứ hạng đối ứng của một câu trả lời truy vấn là nghịch đảo nhân với thứ hạng của câu trả lời đúng đầu tiên: 1 cho vị trí thứ nhất, 1⁄2 cho vị trí thứ hai, 1⁄3 cho vị trí thứ ba, .v.v… Thứ hạng đối ứng trung bình là trung bình của các thứ hạng đối ứng của kết quả đối với một mẫu truy vấn (*Q*). Công thức của tính toán của MRR được định nghĩa như sau:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank(i)} \tag{3.7}$$

trong đó thứ hạng *rank*(*i*) là vị trí thứ hạng của tài liệu liên quan đầu tiên cho truy vấn đầu vào thứ (*i*).

Ví dụ: giả sử có ba truy vấn mẫu sau đây cho một hệ thống cố gắng dịch các danh từ tiếng Anh sang số nhiều của chúng. Trong mỗi trường hợp, hệ thống sẽ đưa ra ba lần dự đoán, trong đó dự đoán đầu tiên là dự đoán mà hệ thống cho là có nhiều khả năng đúng nhất:

Bảng 3.4. Ví dụ để tính toán chỉ số MRR

Truy vấn	**Kết quả truy vấn**	**Kết quả chính xác**	**Xếp hạng**	**Thứ hạng đối ứng**
cat	catten, cati, **cats**	cats	3	1/3
torus	torii, **tori**, toruses	tori	2	1/2
virus	**viruses**, virii, viri	viruses	1	1

Với ba kết quả truy vấn này, có thể tính MRR = (1/3 + 1/2 + 1)/3 = 11/18 hoặc khoảng 0,61.

3.4. Đánh giá kết quả thực nghiệm và thảo luận

Nội dung nghiên cứu trong Chương 3 này tập trung vào việc nâng cao mô hình đề xuất trích dẫn tiên tiến nhất DualLCR của nhóm Medić và Šnajder [12], và do đó luận án đánh giá hiệu quả thực nghiệm của mô hình đã cải tiến RHN-DualLCR với mô hình

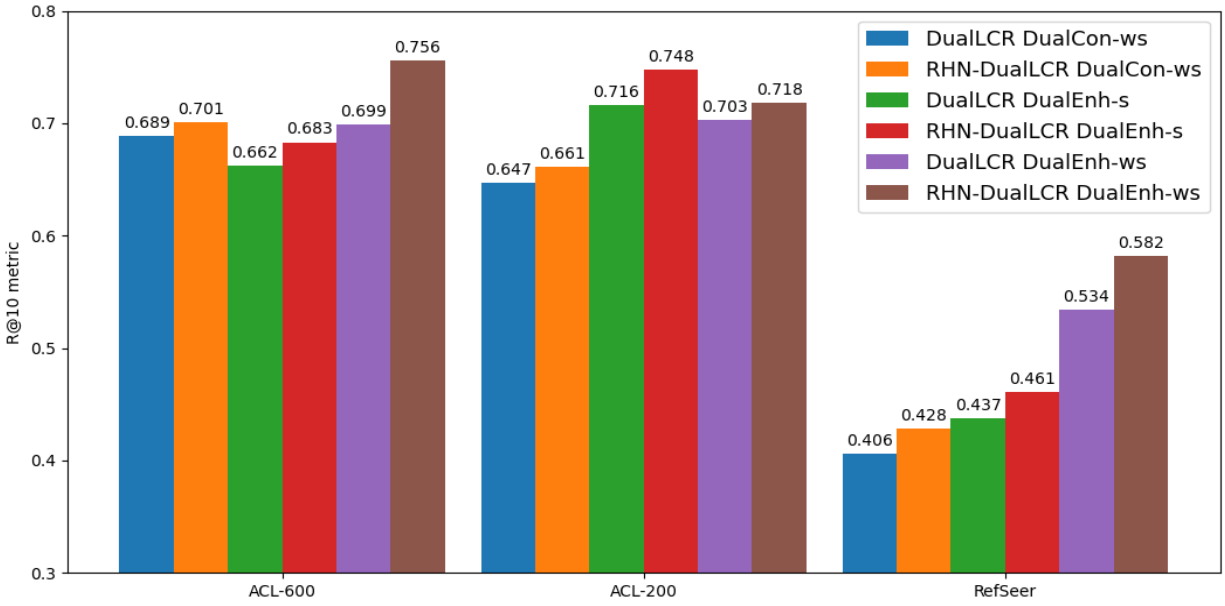
DualLCR ban đầu. Tương tự như Medić và Šnajder [12], luận án đã đánh giá hiệu suất khuyến nghị trích dẫn bằng cách sử dụng các chỉ số đánh giá tiêu chuẩn xếp hạng đối ứng trung bình MRR và Recall@K (viết tắt là R@K trong các bảng so sánh). Cả 2 tiêu chí này đều đã được giải thích ở phần 2.3.3 của chương 2 và 3.3.3 của chương 3. Luận án không thực hiện thực nghiệm lại các kết quả của mô hình DualLCR mà lấy các kết quả đã được các tác giả công bố trong nghiên cứu của họ [12]. Để đảm bảo công bằng khi so sánh, cơ chế huấn luyện và các giá trị tham số của 2 mô hình RHN-DualLCR và DualLCR là như nhau và đã được mô tả ở mục 3.3.1.

Theo kết quả được báo cáo bởi Medić và Šnajder [12], DualCon-ws đạt kết quả tốt nhất cho tiêu chí MRR với bộ dữ liệu ACL-600. Hơn nữa, với tập dữ liệu ACL-600 này, DualEnh-ws thu được kết quả tốt nhất với tiêu chí Recall@10. Với bộ dữ liệu ACL-200, DualEnh-s đạt kết quả tốt nhất với tiêu chí Recall@10 trong khi DualEnh-ws nhận được điểm MRR tốt nhất. Đối với tập dữ liệu RefSeer, DualEnh-ws có kết quả tốt nhất với cả chỉ số đánh giá Recall@10 và MRR. Kết quả thực nghiệm của bài báo này cho chứng minh rằng việc làm phong phú khi biểu diễn ngữ cảnh trích dẫn bằng cách thêm thông tin toàn cục là có lợi khi ngữ cảnh trích dẫn ngắn hơn, nhưng quá trình làm phong phú này không cần thiết khi chúng dài hơn, vì các ngữ cảnh dài hơn đã cung cấp đủ thông tin cho khuyến nghị trích dẫn.

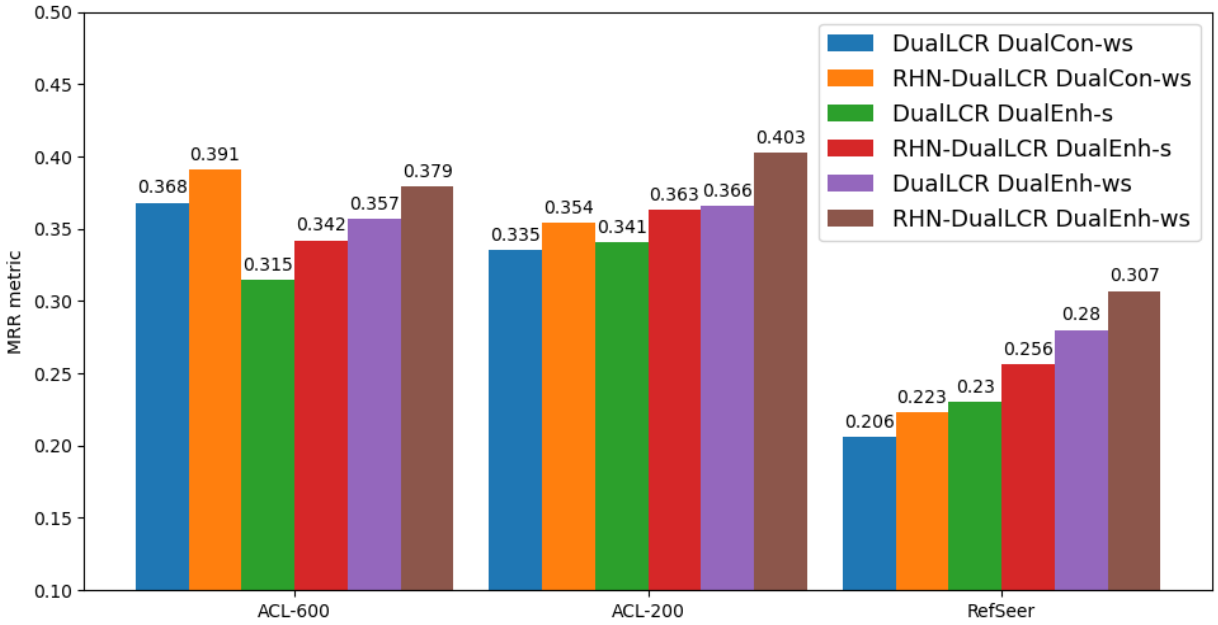
Bảng 3.5. So sánh kết quả từ Medić và Šnajder [12] và mô hình nâng cao RHN-DualLCR

Mô hình		**ACL-600**		**ACL-200**		**RefSeer**	
		R@10	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**
DualLCR [12]	**DualCon-ws**	0.689	**0.368**	0.647	0.335	0.406	0.206
	DualEnh-s	0.662	0.315	**0.716**	0.341	0.437	0.230
	DualEnh-ws	**0.699**	0.357	0.703	**0.366**	**0.534**	**0.280**
RHN-DualLCR	**DualCon-ws**	0.701	**0.391**	0.661	0.354	0.428	0.223
	DualEnh-s	0.683	0.342	**0.748**	0.363	0.461	0.256
	DualEnh-ws	**0.756**	0.379	0.718	**0.403**	**0.582**	**0.307**

Luận án đã so sánh kết quả của mô hình RHN-DualLCR với ba biến thể mô hình hoạt động tốt nhất của Medić và Šnajder [12]: DualCon-ws, DualEnh-s và DualEnh-ws. Nói chung, như được hiển thị trong Bảng 3.5, trên cả ba bộ dữ liệu ACL-600, ACL-200 và RefSeer và với cả tiêu chí đánh giá tiêu chuẩn Recall@10 và MRR, mô hình cải tiến RHN-DualLCR cho điểm cao hơn với cả ba biến thể. Đặc biệt với kết quả tốt nhất của DualCon-ws, DualEnh-s và DualEnh-ws, mô hình đề xuất trích dẫn nâng cao RHN-DualLCR vẫn cho kết quả được cải thiện đáng kể (như được in đậm và đỏ trong Bảng 3.5 ở trên). Điều này chứng tỏ phương pháp đã đề xuất khi sử dụng ScispaCy thay vì spaCy để phân tách các câu trong dữ liệu đầu vào văn bản, sử dụng mô hình SciBERT cho lớp nhúng và thay thế BiLSTM bằng RHN đã tạo ra sự cải thiện rõ rệt và hiệu quả của mô hình hiện tại. Những thành tựu này được minh họa trong Hình 3.5 và 3.6.



Hình 3.5. So sánh hiệu suất của RHN-DualLCR và DualLCR [12] với 3 biến thể trên 3 bộ dữ liệu (ACL-200, ACL-600 và RefSeer) và chỉ số đánh giá Recall@10



Hình 3.6. So sánh hiệu suất của RHN-DualLCR và DualLCR [12] với 3 biến thể trên 3 bộ dữ liệu (ACL-200, ACL-600 và RefSeer) và chỉ số đánh giá MRR

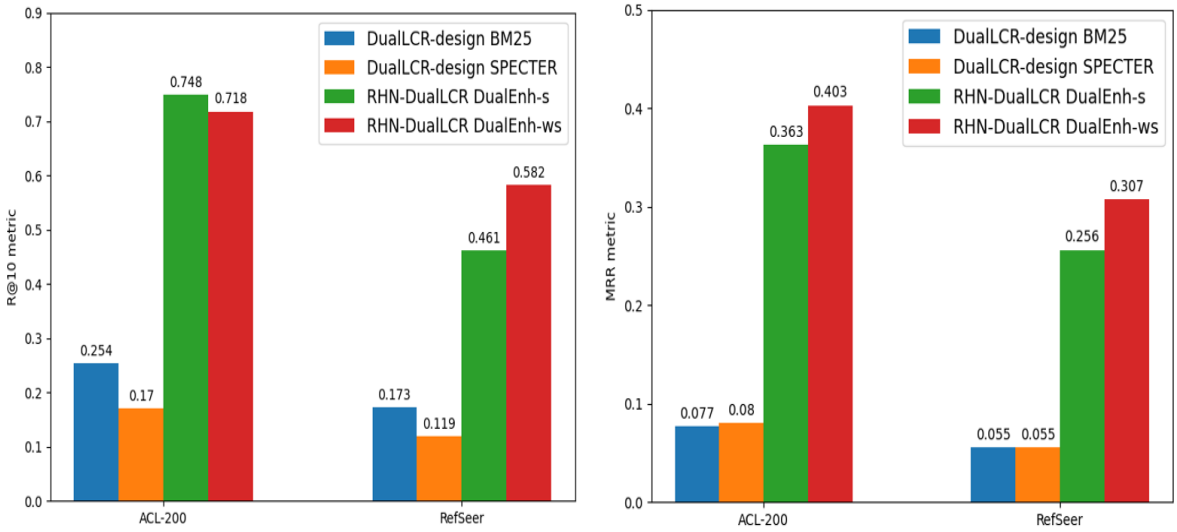
Medić và Šnajder đã tiếp tục thực hiện một nghiên cứu thực nghiệm [13] về tác động của ba lựa chọn thiết kế trong mô hình do chính họ công bố trước đó [12], vì vậy luận án cũng so sánh thêm kết quả thu được từ mô hình RHN-DualLCR với kết quả từ nghiên cứu thực nghiệm của họ để làm nổi bật thêm những đóng góp của mình. Medić và Šnajder đã tiến hành thực nghiệm về ba lựa chọn thiết kế: tham số của mô hình lọc trước, chế độ huấn luyện, chiến lược lấy mẫu phủ định trên hai bộ dữ liệu thường được sử dụng ACL-200 và RefSeer, cho nên luận án so sánh lần lượt mô hình RHN-DualLCR với các lựa chọn thiết kế trên bộ dữ liệu này. Với 2 bộ dữ liệu ACL-200 và RefSeer, RHN-DualLCR đạt thành tích tốt nhất với 2 biến thể DualEnh-s và DualEnh-ws nên luận án chỉ đem 2 biến thể này ra để so sánh. Kết quả so sánh được thể hiện lần lượt ở Bảng 3.6, 3.7 và 3.8, tương ứng với các hình Hình 3.7, 3.8 và 3.9.

Bảng 3.6. So sánh hiệu suất của model lọc trước BM25 và SPECTER [13] với RHN-DualLCR

Mô hình		**ACL-200**		**RefSeer**	
		R@10	**MRR**	**R@10**	**MRR**
DualLCR-design [13]	**BM25**	0.254	0.077	0.173	0.055
	SPECTER	0.170	0.080	0.119	0.055

RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	**0.403**	**0.582**	**0.307**

Trên các bộ dữ liệu ACL-200 và RefSeer, Medić và Šnajder đã nâng cao hiệu quả của hai mô hình lọc trước, đó là BM25, mô hình truy xuất thông tin thông thường và SPECTER, mô hình dựa trên học sâu [92]. Tuy nhiên, kết quả ở Bảng 3.6 cho thấy với hai tiêu chí Recall@10 và MRR, mô hình RHN-DualLCR vẫn cho kết quả vượt trội. Những kết quả này được minh họa rõ hơn trong Hình 3.7.

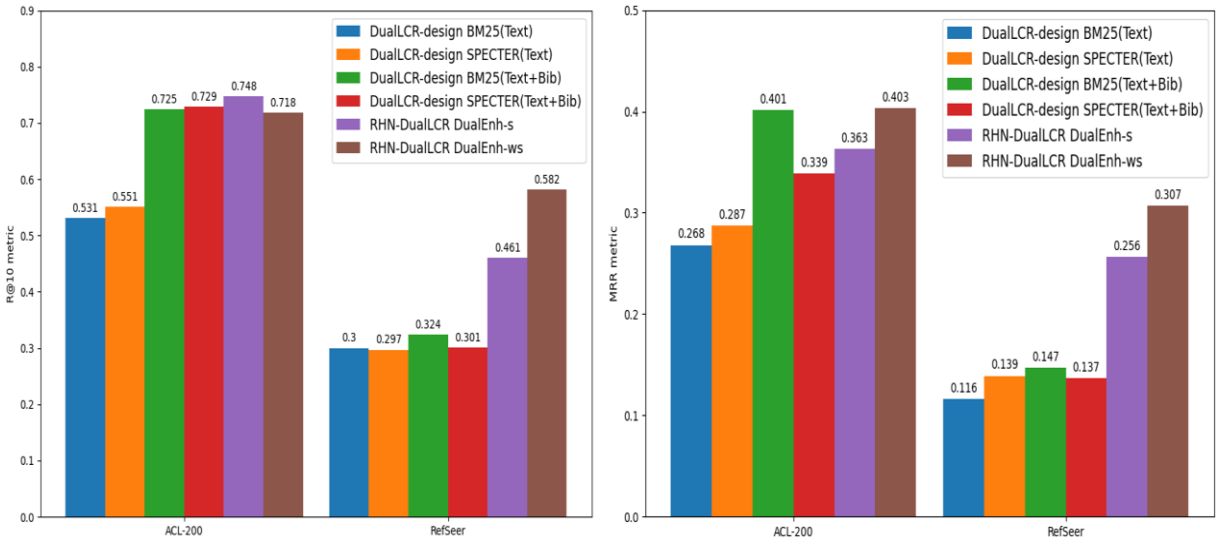


Hình 3.7. So sánh hiệu suất của model lọc trước BM25 và SPECTER [13] với RHN-DualLCR

Bảng 3.7. So sánh hiệu suất của mô hình sắp xếp lại Text+Bib cho các chế độ huấn luyện DualLCR-design[13] với RHN-DualLCR

Mô hình		**ACL-200**		**RefSeer**	
		R@10	**MRR**	**R@10**	**MRR**
DualLCR-design [13]	**BM25(Text)**	0.531	0.268	0.300	0.116
	SPECTER(Text)	0.551	0.287	0.297	0.139
	BM25(Text+Bib)	0.725	0.401	0.324	0.147
	SPECTER(Text+Bib)	0.729	0.339	0.301	0.137
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	**0.403**	**0.582**	**0.307**

Kết quả từ Bảng 3.7 cho thấy, với kết quả thực nghiệm về lựa chọn thiết kế mô hình sắp xếp lại Text+Bib cho các chế độ huấn luyện DualLCR-design trên bộ dữ liệu ACL-200, các biến thể BM25(Text+Bib) và SPECTER(Text+Bib) cho kết quả sát sao với RHN-DualLCR trên 2 tiêu chí Recall@10 (0.729 và 0.748) và MRR (0.401 và 0.403). Tuy nhiên, với bộ dữ liệu RefSeer, hiệu suất của RHN-DualLCR vẫn vượt trội. Điều này được thể hiện rõ ràng trong Hình 3.8.



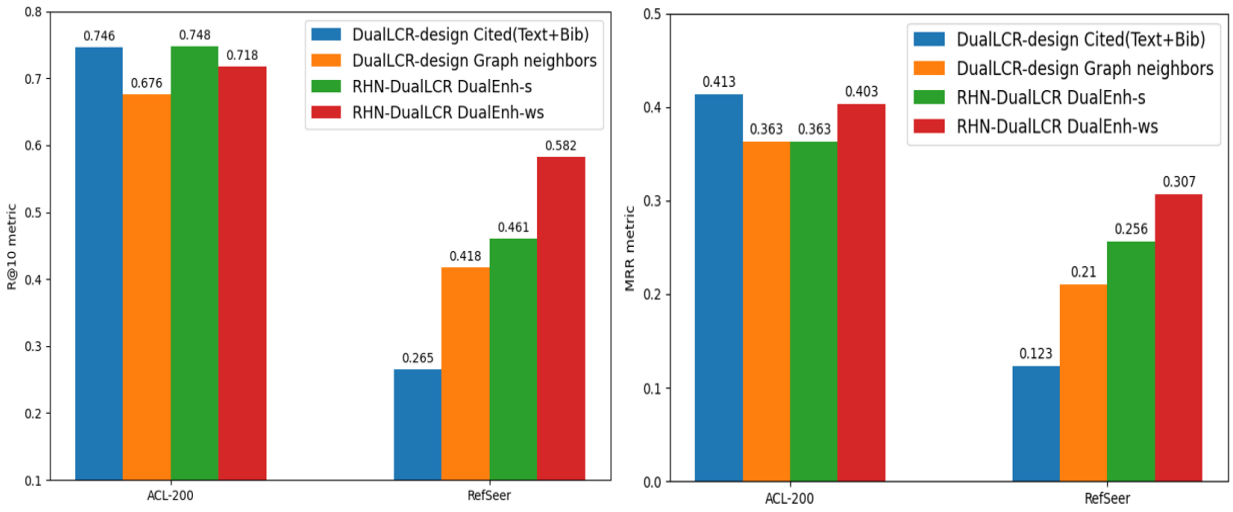
Hình 3.8. So sánh hiệu suất của mô hình sắp xếp lại Text+Bib cho các chế độ huấn luyện DualLCR-design [13] với RHN-DualLCR

Bảng 3.8. So sánh hiệu suất của các chiến lược lấy mẫu phủ định của thiết kế DualLCR [13] với RHN-DualLCR

Mô hình		**ACL-200**		**RefSeer**	
		R@10	**MRR**	**R@10**	**MRR**
DualLCR-design [13]	**Cited(Text+Bib)**	0.746	**0.413**	0.265	0.123
	Graph neighbors	0.676	0.363	0.418	0.210
RHN-DualLCR	**DualEnh-s**	**0.748**	0.363	0.461	0.256
	DualEnh-ws	0.718	0.403	**0.582**	**0.307**

Trong kết quả thực nghiệm của ở công bố của họ [13] với chiến lược lấy mẫu phủ định lựa chọn thiết kế, Medić và Šnajder đã thực hiện trên nhiều biến thể. Tuy nhiên, để

so sánh, luận án chọn hai biến thể Cited(Text+Bib) và Graph neighbors vì chúng có kết quả tốt nhất trên cả hai bộ dữ liệu ACL-200 và RefSeer. Từ Bảng 3.8, có thể thấy rằng với tiêu chí Recall@10 cho tập dữ liệu ACL-200, biến thể Cited(Text+Bib) cho kết quả gần với kết quả tốt nhất của RHN-DualLCR (0.746 và 0.748). Thậm chí ngay cả với tiêu chí MRR, biến thể sử dụng đồ thị lân cận cho kết quả tốt hơn của RHN-DualLCR (0.413 và 0.403). Nhưng với bộ dữ liệu RefSeer, hiệu suất của mô hình RHN-DualLCR vẫn cho kết quả tốt hơn. Những kết quả này được minh họa trong Hình 3.9.



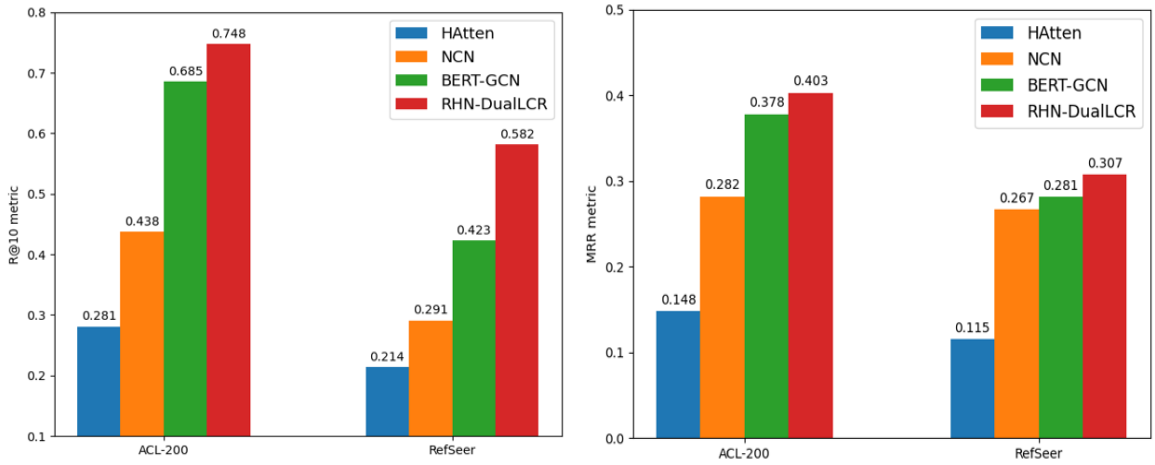
Hình 3.9. So sánh hiệu suất của các chiến lược lấy mẫu phủ định của thiết kế DualLCR [13] với RHN-DualLCR

Để chứng minh rõ hơn hiệu suất của mô hình RHN-DualLCR, ngoài việc so sánh với kết quả của hai mô hình DualLCR [12] và DualLCR-design [13] do nhóm Medić và Šnajder công bố, luận án tiếp tục so sánh với 3 mô hình tiên tiến nhất hiện nay cho bài toán khuyến nghị trích dẫn. Đó là các mô hình: (1) HAtten [16] bao gồm hai giai đoạn: tìm nạp trước và giai đoạn sắp xếp lại; (2) mô hình mạng nơ-ron trích dẫn NCN được đề xuất bởi Ebesu và Fang [10] và cải tiến mới của nhóm Färber [11]; (3) BERT-GCN [15] kết hợp BERT cho xử lý văn bản và mạng tích chập đồ thị GCN [49] cho bộ mã hóa siêu dữ liệu của các bài báo. Kết quả của việc so sánh hiệu suất các mô hình được trình bày ở Bảng 3.9.

Bảng 3.9. So sánh kết quả từ 3 mô hình khuyến nghị trích dẫn tiên tiến với mô hình RHN-DualLCR

Mô hình	ACL-200		RefSeer	
	R@10	MRR	R@10	MRR
HAtten [16]	0.281	0.148	0.214	0.115
NCN [10] [11]	0.438	0.282	0.291	0.267
BERT-GCN [15]	0.685	0.378	0.423	0.281
RHN-DualLCR	**0.748**	**0.403**	**0.582**	**0.307**

Trong công bố của mình [15], các tác giả của mô hình BERT-GCN đã không thực hiện thực nghiệm với hai bộ dữ liệu là ACL-200 và RefSeer, vì vậy luận án đã triển khai mô hình BERT-GCN[24] trong môi trường của mình và tiến hành thực nghiệm với hai bộ dữ liệu này. Kết quả từ Bảng 3.9 cho thấy, với cả bộ dữ liệu ACL-200 và RefSeer, được đánh giá trên cả hai tiêu chuẩn Recall@10 và MRR, mô hình cải tiến RHN-DualLCR vẫn cho kết quả tốt hơn 3 mô hình tiên tiến còn lại. Kết quả này có thể được thấy rõ hơn trên Hình 3.10.



Hình 3.10. So sánh kết quả từ 3 mô hình khuyến nghị trích dẫn tiên tiến với mô hình RHN-DualLCR

3.5. Thực hiện điều chỉnh các tham số của mô hình RHN-DualLCR

Mô hình RHN-DualLCR có ba siêu tham số, đó là kích thước ẩn, độ sâu hồi quy và số lớp. Luận án đã thực hiện tối ưu hóa các siêu tham số này để hiểu rõ hơn về cách

[24] https://github.com/TeamLab/bert-gcn-for-paper-citation

chúng tác động đến hiệu suất của mô hình RHN-DualLCR. Do biến thể DualEnh-ws thu được kết quả tốt nhất nên biến thể này được lấy ra để thực hiện các thực nghiệm liên quan đến điều chỉnh tham số của mô hình. Kết quả thực nghiệm về điều chỉnh tham số thể hiện trong Bảng 3.10, 3.11 và 3.12.

Bảng 3.10. Kết quả của điều chỉnh tham số kích thước lớp ẩn

Kích thước lớp ẩn	**ACL-600**		**ACL-200**		**RefSeer**	
	R@10	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**
100	0.724	0.336	0.695	0.389	0.524	0.283
200	**0.756**	**0.379**	**0.718**	**0.403**	**0.582**	**0.307**
300	0.713	0.351	0.702	0.378	0.556	0.285

Như kết quả được trình bày trong Bảng 3.10, mô hình RHN-DualLCR thu được kết quả tốt nhất khi kích thước lớp ẩn = 200. Tăng hoặc giảm giá trị cài đặt của kích thước lớp ẩn (100 hoặc 300) đều không đạt được kết quả tốt như vậy.

Bảng 3.11. Kết quả của điều chỉnh tham số độ sâu hồi quy

Độ sâu hồi quy	**ACL-600**		**ACL-200**		**RefSeer**	
	R@10	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**
4	0.738	0.359	0.685	0.382	0.524	0.293
6	**0.756**	**0.379**	**0.718**	**0.403**	**0.582**	**0.307**
8	0.722	0.341	0.701	0.387	0.551	0.291

Độ sâu hồi quy là độ sâu tùy ý trong một lớp của mô hình RHN. Dựa trên kết quả nghiên cứu của nhóm nghiên cứu Zilly [20], luận án đã chọn giá trị cài đặt tốt nhất của độ sâu lặp lại là 6. Các giá trị cài đặt khác (4 hoặc 8) không góp phần nâng cao hơn được nữa thành tích của mô hình RHN-DualLCR.

Bảng 3.12. Kết quả của điều chỉnh siêu tham số số lượng của lớp

Số lớp	**ACL-600**		**ACL-200**		**RefSeer**	
	R@10	**MRR**	**R@10**	**MRR**	**R@10**	**MRR**
1	0.739	0.358	0.662	0.383	0.525	0.285
2	**0.756**	**0.379**	**0.718**	**0.403**	**0.582**	**0.307**
3	0.743	0.340	0.702	0.372	0.531	0.276

Kết quả thực hiện điều chỉnh siêu tham số của mô hình được trình bày trong Bảng 3.12 chứng minh rằng khi số lớp = 2 thì mô hình RHN-DualLCR thu được kết quả tốt nhất. Khi thử điều chỉnh số lượng lớp (1 hoặc 3) thì kết quả thực nghiệm cho thấy rằng mô hình nâng cao hoạt động tốt nhất khi giá trị số lớp được đặt thành 2.

3.6. Kết luận chương 3

Nội dung của chương 3 tập trung vào các mô hình kết hợp lọc nội dung và lọc cộng tác. Tiêu biểu trong số này là mô hình DualLCR do Medić và Šnajder [12] [13] xây dựng. Luận án đã cải tiến mô hình DualLCR bằng cách áp dụng các kết quả nghiên cứu gần đây của mạng nơ-ron hồi quy của học sâu cũng như các thành tựu nghiên cứu trong xử lý ngôn ngữ tự nhiên, đặc biệt là ngôn ngữ trong các bài báo khoa học. Luận án đã tích hợp ScispaCy [91] để phân tách các câu trong bối cảnh trích dẫn và tóm tắt của bài báo, sử dụng mô hình SciBERT [18] để thực hiện phép nhúng cho văn bản khoa học của các bài báo ở đầu vào và thay thế BiLSTM [14] bằng RHN [20], một mô hình đã mở rộng kiến trúc BiLSTM để cho phép chuyển tiếp giữa các bước có độ sâu lớn hơn một. Luận án đã đánh giá hiệu suất của mô hình được đề xuất bằng ba bộ dữ liệu ACL-200, ACL-600 và RefSeer và đạt được kết quả cải thiện đáng kể khi so sánh với mô hình của Medić và Šnajder trong 2 bài báo gốc [12] [13] khi sử dụng cùng hai tiêu chuẩn đánh giá Recall@10 và MRR. Mô hình RHN-DualLCR cũng đã chứng minh được hiệu quả hơn so với 3 mô hình khuyến nghị trích dẫn tiên tiến hiện nay, đó là các mô hình: HAtten [16], NCN [10] [11] và BERT-GCN [15]. Hơn nữa, nội dung nghiên cứu của chương 3 này cũng bao gồm phần tiến hành thực nghiệm điều chỉnh để xem xét các siêu tham số khác nhau có thể ảnh hưởng như thế nào đến hiệu suất của mô hình RHN [20] và luận án đã đề xuất các cách sử dụng các kết quả thực nghiệm này để nâng cao hiệu quả của mô hình trong tương lai.

Các kết quả nghiên cứu của chương 3 này đã được công bố tại công trình [CT2] và [CT4] trong phần "Danh mục các công trình đã công bố của tác giả"

Áp dụng các mô hình khuyến nghị RHN-DualLCR kết hợp giữa lọc nội dung và lọc cộng tác cho các bộ dữ liệu chứa không chỉ nội dung của bài báo mà còn thông tin về đánh giá từ cộng đồng học thuật là lựa chọn hiệu quả. Lý do là vì lọc nội dung giúp

phân tích các đặc điểm của bài báo dựa trên nội dung thực tế, trong khi lọc cộng tác tận dụng thông tin đánh giá từ cộng đồng để xếp hạng xem bài báo nào tốt hơn để trích dẫn. Khi kết hợp cả hai phương pháp, mô hình có thể tận dụng tốt cả hai nguồn dữ liệu: nội dung của bài báo để tìm các bài có nội dung tương tự và thông tin đánh giá từ cộng đồng học thuật để cải thiện khả năng khuyến nghị những bài báo được đánh giá cao. Điều này đặc biệt quan trọng trong bối cảnh học thuật, nơi đánh giá từ cộng đồng giúp xác thực giá trị và độ tin cậy của các bài báo, làm cho mô hình khuyến nghị trở nên chính xác và hiệu quả hơn. Tuy nhiên trong các bài báo khoa học, ngoài nội dung bài báo (đã nghiên cứu ở chương 2) và thông tin học thuật (nghiên cứu ở chương 3), thì thông tin về liên kết trích dẫn giữa các bài báo cũng phải được quan tâm đúng mức trong các hệ thống khuyến nghị. Đó cũng là nội dung nghiên cứu của chương 4.

CHƯƠNG 4. MÔ HÌNH KHUYẾN NGHỊ TRÍCH DẪN MỚI SỬ DỤNG SCIBERT VÀ GRAPHSAGE

Trong các mô hình khuyến nghị trích dẫn dựa trên đồ thị thì mạng đồ thị được sử dụng để mô hình hóa thông tin của bài báo. Mạng đồ thị của các bài báo chứa nhiều đối tượng được thiết lập mối quan hệ với nhau, chẳng hạn như tác giả, nội dung bài báo, địa điểm và thời gian công bố, từ khóa và tiêu đề. Do đó, các mô hình dựa trên đồ thị có thể khai thác các mối quan hệ và ngữ nghĩa có ý nghĩa giữa các đối tượng này để đưa ra các đề xuất trích dẫn mạnh mẽ hơn và khắc phục các vấn đề liên quan đến mô hình lọc nội dung hay lọc cộng tác truyền thống. Trong phạm vi nghiên cứu của chương 4 này, luận án xây dựng một mô hình khuyến nghị trích dẫn mới bằng cách kết hợp hai phương pháp: (1) lọc nội dung: sử dụng SciBERT [18] để biểu diễn dữ liệu văn bản hay ngữ cảnh; và (2) lọc dựa vào đồ thị: sử dụng GraphSAGE [19] để biểu diễn các liên kết trích dẫn giữa các bài báo.

Các đề xuất và kết quả nghiên cứu của chương 4 này được công bố tại công trình [CT3] và [CT5] trong phần "Danh mục các công trình đã công bố của tác giả"

4.1. Thảo luận các vấn đề còn tồn tại của mô hình khuyến nghị trích dẫn hiện nay

Phần lớn các phương pháp tiếp cận hiện tại đối với bài toán khuyến nghị trích dẫn nhận biết ngữ cảnh chỉ tập trung vào nội dung của cả ngữ cảnh trích dẫn và các bài báo khoa học [13] [16] [93] [94]. Cách tiếp cận này nhằm mục đích kết nối khoảng cách ngữ nghĩa giữa các yếu tố này mà không xem xét thông tin vượt ra ngoài nội dung ngữ nghĩa của các bài báo khoa học. Thực tế trong các công bố khoa học có những yếu tố bổ sung như tác giả, thông tin hội nghị/tạp chí và năm xuất bản, có mức độ quan trọng khác nhau trong việc hỗ trợ các nhà nghiên cứu hiểu được sự tương đồng về ngữ nghĩa giữa các bài báo khoa học và bối cảnh trích dẫn. Ví dụ: một tác giả được liên kết với một bài báo khoa học có thể là đồng tác giả với các bài báo liên quan khác. Tương tự như vậy, một hội nghị hay tạp chí công bố một bài báo khoa học cụ thể cũng có thể xuất bản các bài báo khác có chủ đề tương tự. Vì vậy, việc bổ sung thông tin về tác giả và nơi xuất bản (tên hội nghị hoặc tạp chí) được kỳ vọng sẽ làm tăng hiệu quả của hệ thống khuyến nghị

trích dẫn. Ngoài ra, thông tin về năm xuất bản của bài báo cũng cần được quan tâm đúng mức đối với các mô hình khuyến nghị trích dẫn. Cụ thể, khi tìm tài liệu trích dẫn thì các nhà nghiên cứu luôn có xu hướng trích dẫn những bài báo mới nhất và cập nhật nhất. Dựa vào những giả định này, có thể phát hiện ra rằng hiệu suất khuyến nghị trích dẫn theo ngữ cảnh không chỉ bị ảnh hưởng hoàn toàn bởi sự giống nhau về mặt ngữ nghĩa giữa ngữ cảnh trích dẫn và các bài báo khoa học mà nó cũng phụ thuộc vào các yếu tố khác, chẳng hạn như tác giả, địa điểm và năm xuất bản của các bài báo này.

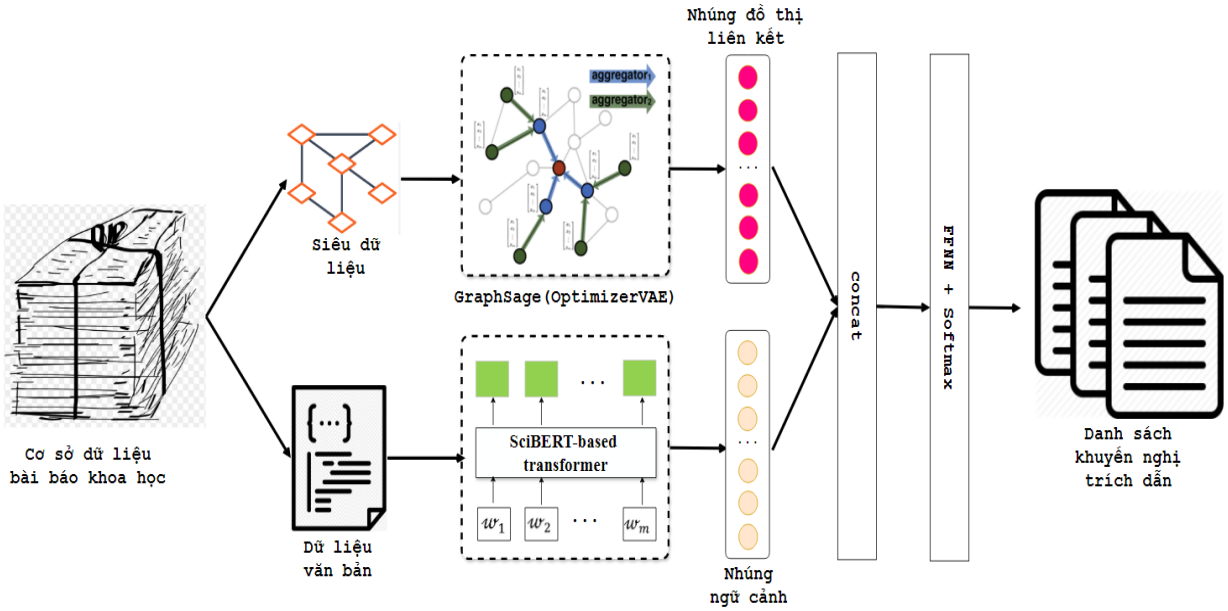
Một công bố gần đây của nhóm Jeong [15] đã đề xuất mô hình BERT-GCN, trong đó kết hợp BERT để mã hóa ngữ cảnh trích dẫn và thông tin của bài báo, còn mạng đồ thị tích chập GCN để mã hóa các thông tin siêu dữ liệu (tác giả, địa điểm và năm công bố) của bài báo. Các công bố của các nhóm nghiên cứu khác [11] [12] cũng đã bước đầu đưa thêm các thông tin siêu dữ liệu vào mô hình của họ. Tuy nhiên các mô hình này đều có thể cải thiện được nếu áp dụng các thành tựu nghiên cứu về mạng tích chập đồ thị trong thời gian gần đây.

Được thúc đẩy bởi những kết quả ấn tượng trong các nghiên cứu gần đây về xử lý ngôn ngữ tự nhiên cũng như các mô hình học sâu dựa trên đồ thị, hơn nữa các bài báo khoa học luôn chứa cả dữ liệu nội dung văn bản (ngữ cảnh trích dẫn, tóm tắt) và siêu dữ liệu (liên kết trích dẫn giữa các bài báo, tác giả, thông tin địa điểm và năm xuất bản) hoàn toàn có thể được biểu diễn dưới dạng đồ thị. Do đó trong chương 4 này luận án đề xuất một mô hình học sâu mới cho bài toán khuyến nghị trích dẫn bằng cách kết hợp các kết quả nghiên cứu tốt nhất hiện tại về phép nhúng theo ngữ cảnh và phép nhúng đồ thị tích chập. Nội dung chi tiết của mô hình mới này sẽ được trình bày trong phần 4.2.

4.2. Xây dựng mô hình khuyến nghị trích dẫn mới với SciBERT và GraphSAGE

Nội dung phần này mô tả chi tiết xây dựng một mô hình khuyến nghị trích dẫn nhận biết ngữ cảnh hoàn toàn mới bằng cách kết hợp hai thành tựu nghiên cứu cập nhật nhất hiện nay cho các kỹ thuật học biểu diễn dữ liệu dựa trên văn bản và ngữ cảnh SciBERT [18] và đồ thị biểu thị liên kết trích dẫn GraphSAGE [19]. Như đã đề cập ở chương 3, SciBERT là một biến thể của mô hình xử lý ngôn ngữ tự nhiên BERT [68] nhưng đã được điều chỉnh đặc biệt cho các nhiệm vụ trong lĩnh vực phân tích văn bản

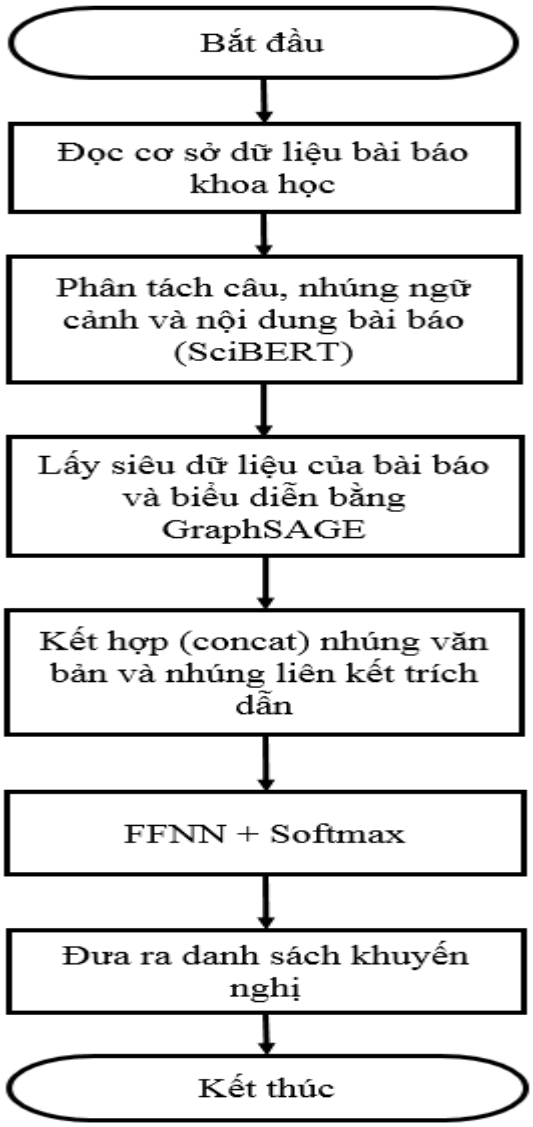
khoa học và y sinh. Việc sử dụng SciBERT được huấn luyện trước để biểu diễn câu theo ngữ cảnh sẽ mang lại hiệu quả cao. Dữ liệu khoa học chẳng hạn như các bài báo, thì ngoài dữ liệu văn bản vẫn còn chứa nhiều siêu dữ liệu khác nhau như liên kết trích dẫn giữa các bài báo, tác giả, thông tin địa điểm và năm xuất bản. Do đó luận án ứng dụng mô hình GraphSAGE để mô tả mối liên kết trích dẫn giữa các bài báo và rút ra biểu diễn đã được huấn luyện từ chúng.



Hình 4.1. Sơ đồ kiến trúc của mô hình SciBERT-GraphSAGE

Hình 4.1 minh họa kiến trúc của mô hình khuyến nghị trích dẫn SciBERT-GraphSAGE, trong đó bao gồm một bộ mã hóa ngữ cảnh để lấy các phần nhúng văn bản thông qua SciBERT và một bộ mã hóa liên kết trích dẫn để trích xuất các thành phần nhúng đồ thị bằng GraphSAGE. Thông tin đầu vào của mô hình SciBERT-GraphSAGE lấy từ cơ sở dữ liệu các bài báo khoa học bao gồm siêu dữ liệu (địa điểm (tạp chí, sách hay hội nghị), năm công bố, tên tác giả và liên kết trích dẫn của bài báo) và dữ liệu văn bản (ngữ cảnh trích dẫn, tóm tắt và tiêu đề của bài báo). SciBERT được sử dụng để tạo ra vectơ nhúng có kích thước cố định cho các dữ liệu văn bản, trong khi GraphSAGE được sử dụng để mô hình hóa các liên kết trích dẫn. Mỗi nút trong GraphSAGE được liên kết với một vectơ đặc điểm là siêu dữ liệu của bài báo. Đầu ra của GraphSAGE là một vectơ nhúng cho mỗi nút (biểu diễn một bài báo), nắm bắt cấu trúc của đồ thị trích dẫn - tức là cách bài báo được kết nối với các bài báo khác. Sau khi có được các vectơ

nhúng dữ liệu văn bản (từ SciBERT) và nhúng đồ thị trích dẫn (từ GraphSAGE), chúng sẽ được nối lại thành một vectơ duy nhất bằng cách sử dụng hàm concat của TensorFlow. Thao tác nối lại chỉ đơn giản là thêm hai vectơ (nhúng ngữ cảnh và nhúng đồ thị trích dẫn) vào một vectơ dài hơn. Nếu nhúng ngữ cảnh có chiều (d_1) và nhúng đồ thị trích dẫn có chiều (d_2), vectơ nhúng nối lại kết quả sẽ có chiều ($d_1 + d_2$). Những phần nhúng kết hợp này sau đó được chuyển đến lớp phân loại (bao gồm mạng nơ-ron truyền thẳng và hồi quy softmax) để dự đoán khả năng trích dẫn giữa bài báo hiện tại và bài báo ứng viên.



Hình 4.2. Sơ đồ khối xử lý mô hình SciBERT-GraphSAGE

4.2.1. Tiếp cận học mô hình biểu diễn các bài báo khoa học bằng SciBERT

SciBERT chuyển đổi các cặp ngữ cảnh trích dẫn và văn bản tham chiếu thành mã có thể đọc được bằng máy tính. Ngữ cảnh trích dẫn xuất hiện dưới dạng một đoạn văn bản bao gồm (*n*) từ trước và sau vị trí trích dẫn "[1]", hoặc "*Ebesu and Fang, 2017*", v.v… Tài liệu tham khảo đại diện cho một bài báo được đề cập trong ngữ cảnh trích dẫn. Do đó, mô hình nắm bắt được bản chất ngữ nghĩa chính xác của các cặp văn bản đầu vào thông qua các thuộc tính cấu trúc của nó. Ban đầu, cặp đầu vào có cấu trúc là "*[CLS] text_a [SEP] text_b [SEP]*". Cấu trúc này sau đó được chuyển đổi thành mã định danh có thể đọc được trên máy tính, bao gồm nhúng của mã thông báo, vị trí và phân đoạn bằng cách sử dụng "*từ vựng*" của SciBERT. Các thành phần nhúng này được tổng hợp để tạo từ nhúng cho văn bản trong lớp mã hóa SciBERT. Sau đó, ma trận nhúng từ cho từng kích thước theo tập (*batch*) sẽ trải qua quá trình xử lý trong lớp chú ý nhiều đầu (*multi-headed attention layer*) để tính toán mối liên hệ giữa mỗi từ và tất cả các từ khác trong câu. Cuối cùng, kết quả thu được từ cơ chế chú ý của nhiều đầu được đưa vào các lớp xử lý thêm và chuẩn hóa (*Add and Norm*) để đạt được sự chuẩn hóa. Quá trình này được minh họa như công thức (4.1).

$$LN(x) = \propto \times \frac{x - \mu}{\delta} + \beta \qquad (4.1)$$

trong đó *x* là ma trận nhúng từ; μ, σ là giá trị trung bình và độ lệch chuẩn của các phần tử trong *x*, tức là $\mu = \frac{1}{d}\sum_{k=1}^{d} x_k$ và $\sigma^2 = \frac{1}{d}\sum_{k=1}^{d}(x_k - \mu)^2$, tỉ lệ $\propto$ và vectơ điều chỉnh β là các tham số của mô hình.

Văn bản đầu vào được đưa vào mạng nơ-ron truyền thẳng FNN để xử lý trên từng vị trí trong cùng một trạng thái. Mạng nơ-ron này bao gồm hai phép biến đổi tuyến tính, như được hiển thị trong công thức (4.2).

$$FNN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \qquad (4.2)$$

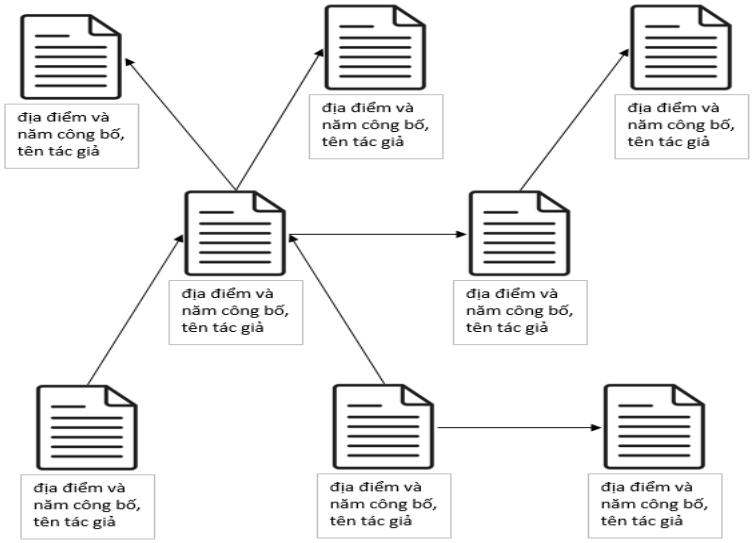
trong đó (W_1), (b_1), (W_2) và (b_2) là các tham số mô hình của mạng nơron truyền thẳng.

Kết quả xử lý đầu ra của mạng nơ-ron truyền thẳng đầy đủ kết hợp cùng với đầu ra từ mô hình GraphSAGE sẽ được chuẩn hóa thêm một lần nữa. Sau đó, việc tính toán

xử lý kết nối dư được thực hiện để lấy được đầu ra của toàn bộ mô hình SciBERT-GraphSAGE.

4.2.2. Bộ mã hóa đồ thị liên kết trích dẫn sử dụng GraphSAGE

Nhóm nghiên cứu của Hamilton [19] đã giới thiệu GraphSAGE (**Graph SA**mpling and **AG**gregation) như một phiên bản mở rộng và cải tiến của mạng tích chập đồ thị GCN [49]. Khái niệm cốt lõi đằng sau mô hình GraphSAGE là rút ra các biểu diễn cấu trúc cục bộ có thứ tự cao và quy nạp của các nút từ các đồ thị biểu diễn trích dẫn đã cho. Khác với mạng tích chập đồ thị GCN trước đó, việc tổng hợp tính năng cho nút mục tiêu dựa vào tập hợp con được lấy mẫu của các thuộc tính nút lân cận của nó, thay vì dựa vào tập hợp lân cận đã hoàn chỉnh như mạng tích chập đồ thị GCN [49]. Với thuộc tính này, việc áp dụng các mô hình GraphSAGE vào khuyến nghị trích dẫn đã trở nên nổi bật. Các nghiên cứu gần đây đã chứng minh rằng GraphSAGE được coi là một nền tảng học máy dựa trên đồ thị mạnh mẽ được thiết kế để học cách biểu diễn từ các đồ thị tỷ lệ lớn một cách hiệu quả. Bằng cách coi các bài báo học thuật là các nút trong đồ thị, siêu dữ liệu (tác giả, thông tin về địa điểm và năm xuất bản) là thuộc tính của nút và mối liên kết trích dẫn giữa các bài báo là các cạnh của đồ thị, mạng nơ-ron trích dẫn tạo thành cấu trúc đồ thị tự nhiên, trong đó các bài báo được kết nối với nhau thông qua mối quan hệ trích dẫn của chúng. Chi tiết về quá trình này có thể được tham khảo trong Hình 4.3. Tận dụng khả năng kết nối vốn có này, GraphSAGE cung cấp một cách tiếp cận đầy hứa hẹn để nâng cao khuyến nghị trích dẫn. Nó vượt trội trong việc học cách biểu diễn nút bằng cách lấy mẫu và tổng hợp thông tin từ các nút lân cận trong đồ thị. Bằng cách này, nó không chỉ có thể nắm bắt được ý nghĩa ngữ nghĩa của từng bài báo mà còn mã hóa ngữ cảnh được cung cấp bởi các mối quan hệ trích dẫn của chúng. Do đó, các biểu diễn nút đạt được phản ánh vị trí của các bài báo trong bối cảnh học thuật rộng hơn, nhằm nắm bắt những điểm tương đồng, ảnh hưởng và mối liên hệ theo chủ đề giữa các bài báo.



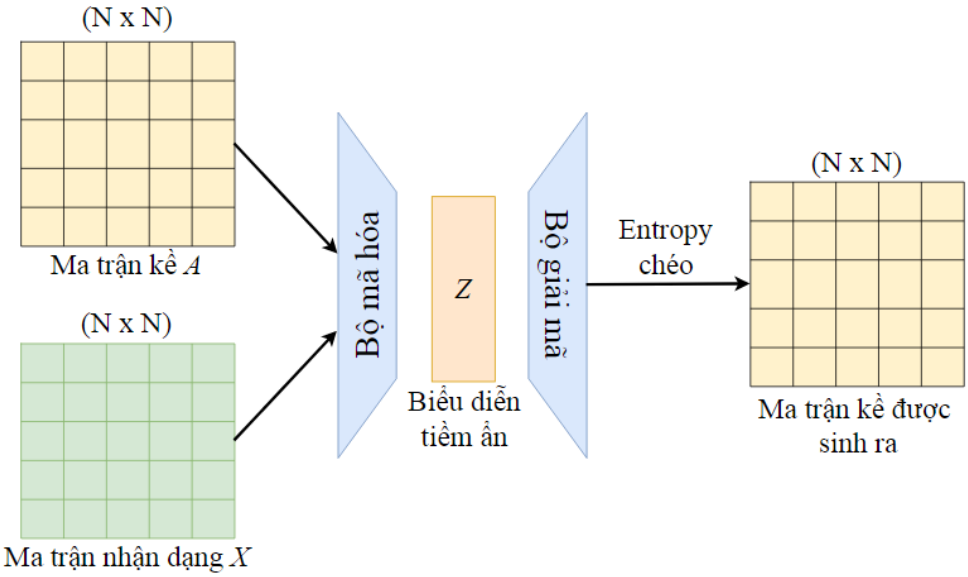
Hình 4.3. Tạo các nút và cạnh cho GraphSAGE từ siêu dữ liệu của các bài báo

Vai trò của mô hình GraphSAGE là trừu tượng hóa thông tin từ mạng liên kết trích dẫn thông qua một mạng nơ-ron chập. GraphSAGE được xây dựng như là một mô hình suy luận của bộ mã hóa tự động đồ thị biến đổi (*Variational Graph Auto-Encoders, VGAE*) [95] để có thể nắm bắt được biểu diễn quá trình huấn luyện tiềm ẩn của việc nhúng dữ liệu đồ thị từ GraphSAGE. Công thức của lớp GraphSAGE dành cho VGAE được thể hiện trong công thức (4.3).

$$GraphSAGE(X, A) = \tilde{A}ReLU(\tilde{A}XW_0)\ W_1 \qquad (4.3)$$

Mô hình GraphSAGE này sử dụng hai ma trận làm đầu vào: ma trận nhận dạng (X) và ma trận kề (A). Cả hai ma trận này đều là ma trận vuông (N x N), với (N) là số lượng bài báo khoa học ở đầu vào. Sau khi học được huấn luyện từ mô hình GraphSAGE đầu tiên, tham số của lớp là (W_0) được sử dụng làm ma trận trọng số cho lớp thứ hai. Mỗi lớp sử dụng phương pháp lan truyền từng lớp một. ($\tilde{A}$) là ma trận kề đã được chuẩn hóa dựa trên ma trận đường chéo (D) như được thể hiện trong công thức (4.4).

$$\tilde{A} = D^{-1/2}AD^{-1/2} \qquad (4.4)$$



Hình 4.4. Cấu trúc của mô hình VGAE

Mô hình VGAE là một loại mô hình tổng quát học cách mã hóa và giải mã dữ liệu dạng đồ thị trong một không gian tiềm ẩn. Mô hình VGAE huấn luyện biểu diễn tiềm ẩn bằng cách tối thiểu hóa hàm mất mát giữa lớp mã hóa và giải mã như trong công thức (4.5).

$$L = E_{q(Z|X,A)}[\log p(A|Z)] - KL[q(Z|X,A)||p(Z)] \qquad (4.5)$$

Lớp bộ mã hóa của VGAE huấn luyện biểu diễn Z bằng cách giảm phân kỳ Kullback–Leibler [96] giữa phân phối chuẩn (*normal distribution*) từ mô hình GraphSAVE và phân phối chuẩn Gaussian, như được tính toán trong công thức (4.6).

$$q(Z|X,A) = \prod_{i=1}^{N} q(z_i|X,A), q(z_i|X,A) = \aleph(z_i|\mu_i, diag(\delta_i^2)) \qquad (4.6)$$

Bước tiếp theo, lớp giải mã học ma trận kề dựa trên ma trận biểu diễn (Z) của lớp mã hóa. Các biến tiềm ẩn (*latent variable*) (z_i) và (z_j) là giá trị tích bên trong của bài báo (i) và (j). Một ma trận kề được tạo ra dựa trên biến tiềm ẩn thông qua tích bên trong giữa các vectơ biểu diễn của bài báo, như trong công thức (4.7). Lớp mã hóa xác định ma trận biểu diễn (Z) bằng cách tối thiểu hóa sự khác biệt giữa ma trận kề (A) được sinh ra bởi lớp mã hóa và ma trận kề thực tế.

$$p(A|Z) = \prod_{i=1}^{N} \prod_{j=1}^{N} p(A_{ij}|z_i, z_j), p(A_{ij} = 1|z_i, z_j) = \sigma(z_i|z_i) \qquad (4.7)$$

4.3. Tiến hành thực nghiệm với mô hình SciBERT-GraphSAGE

Nội dung của phần này sẽ mô tả về việc thiết lập các thực nghiệm để kiểm tra hiệu quả của mô hình khuyến nghị trích dẫn SciBERT-GraphSAGE cũng như các kết quả thu được từ thực nghiệm cũng như các đánh giá cho các kết quả này sẽ được trình bày chi tiết.

4.3.1. Cài đặt mô hình SciBERT-GraphSAGE

Luận án đã xây dựng mô hình SciBERT-GraphSAGE cho hệ thống khuyến nghị trích dẫn bằng cách kết hợp bộ mã hóa ngữ cảnh trích dẫn SciBERT và mã hóa liên kết trích dẫn GraphSAGE. Để có thể mã hóa ngữ cảnh trích dẫn, luận án đã khởi tạo SciBERT bằng mô hình được huấn luyện trước được cung cấp bởi nhóm của Beltagy[25] [18]. Tương tự, tác giả của luận án đã chỉnh sửa mã nguồn của mô hình GraphSAGE[26] [19] để có thể mã hóa đồ thị liên kết trích dẫn. Tất cả các mô hình được xây dựng với Python phiên bản 3.8.5 và TensorFlow phiên bản 2.7.0. Luận án đã trích xuất các vectơ ngữ cảnh nhúng và vectơ đồ thị trích dẫn bằng cách sử dụng các lớp SciBERT và GraphSAGE, được xây dựng thông qua các quy trình huấn luyện riêng biệt. Trong SciBERT, số lượng lớp chú ý (*number of attention heads*) là 12, ngăn xếp (*stack*) bộ mã hóa là 12 và trình tối ưu hóa Adam optimizer [97] được sử dụng. Tốc độ học (η) là 2×10^{-5}, epsilon (ϵ) là 1×10^{-6}, với tham số beta 1 (β_1) được đặt là: 0.9, beta 2 (β_2) được đặt là 0.999 và tốc độ giảm trọng số là 0.01. Mô hình cũng thiết lập độ dài chuỗi tối đa là 128, bộ đệm (*padding*) là 0 nếu độ dài ngắn hơn 128 và kích thước ẩn là 768. Đối với GraphSAGE, số vòng lặp huấn luyện (*epochs*) là 200, kích thước ẩn đầu tiên tương ứng với với số lượng bài báo trong bộ dữ liệu và thứ nguyên ẩn thứ hai là 768, kích thước tập (*batch*) giống với tổng kích thước tài liệu (giảm độ dốc toàn lô), trình tối ưu hóa là OptimizerVAE [95] và tốc độ học là 0.01. Mô hình SciBERT-GraphSAGE được chạy trên môi trường Linux 3.10.0-x86-64, bộ xử lý NVIDIA GPU H100. Đối với cấu hình

[25] https://github.com/allenai/scibert
[26] https://github.com/williamleif/GraphSAGE

tham số của mô hình, sau khi chạy thử nghiệm với nhiều giá trị, luận án đã tìm thấy giá trị cài đặt tốt nhất cho từng bộ dữ liệu thử nghiệm như trong Bảng 4.1 sau:

Bảng 4.1. Kết quả của điều chỉnh siêu tham số trong mô hình theo tập dữ liệu

Mô hình	Tham số	FullTextPeerRead	ACL-200	RefSeer
SciBERT	Kích thước lô	16	32	64
	Tần suất tập dữ liệu	5	10	20
	Số lần lặp huấn luyện	30	60	120
GraphSAGE	Số chiều ẩn	[7,071; 768]	[19,606; 768]	[33,559; 768]
	Số lần lặp huấn luyện	200	500	1,000
	Số nút lân cận	5	10	20
	Kích thước lô	32	64	128

4.3.2. Mô tả về bộ dữ liệu thực nghiệm

Trong các mô hình khuyến nghị kết hợp giữa lọc nội dung và lọc đồ thị, việc áp dụng cho các bộ dữ liệu không chỉ chứa nội dung của bài báo mà còn bao gồm cả các liên kết trích dẫn là một lựa chọn lý tưởng. Mô hình lọc nội dung phân tích các đặc điểm của bài báo dựa trên nội dung thực tế, trong khi mô hình lọc đồ thị tận dụng các liên kết trích dẫn giữa các bài báo để xây dựng một mạng lưới liên kết khoa học. Điều này cho phép mô hình hiểu rõ hơn về mối quan hệ giữa các bài báo, không chỉ dựa trên sự tương đồng về nội dung mà còn dựa trên cách chúng được các bài báo khác trích dẫn. Việc kết hợp cả hai phương pháp giúp tối ưu hóa khả năng khuyến nghị, đặc biệt trong các tập dữ liệu học thuật, nơi các liên kết trích dẫn đóng vai trò quan trọng trong việc xác định tầm quan trọng và độ ảnh hưởng của một bài báo. Nhờ vậy, mô hình kết hợp lọc nội dung và lọc đồ thị có thể đưa ra các gợi ý chính xác và giá trị hơn, phù hợp với nhu cầu nghiên cứu và tham khảo. Với các lý do trên, chương 4 này đã đánh giá tính hiệu quả của mô hình SciBERT-GraphSAGE trên ba bộ dữ liệu tiêu chuẩn thường được sử dụng cho mô hình khuyến nghị trích dẫn hiện nay, bao gồm: ACL-200 [12], RefSeer [12] và FullTextPeerRead [15]. Nghiên cứu này đã sử dụng các bộ dữ liệu này để đánh giá hiệu suất vượt trội của mô hình SciBERT-GraphSAGE với 5 mô hình khuyến nghị trích dẫn

tiên tiến hiện nay là: CACR [93], BERT-GCN [15], HAtten [16], DualLCR [12], DualLCR theo thiết kế [13]. Số liệu thống kê của ba bộ dữ liệu này được hiển thị ở trong Bảng 4.2.

Bảng 4.2. Thống kê 3 bộ dữ liệu (số lượng ngữ cảnh trích dẫn và bài báo)

Tên bộ dữ liệu	**Số lượng ngữ cảnh**			**Số bài báo**	**Năm công bố**
	Huấn luyện	**Xác thực**	**Kiểm tra**		
ACL-200	30,390	9,381	9,585	19,711	2009 - 2015
FullTextPeerRead	9,363	1,043	6,841	4,898	2007 - 2017
RefSeer	3,521,582	124,551	126,021	624,957	- 2014

ACL-200 là bộ dữ liệu về ngữ cảnh trích dẫn được trích xuất từ ACL Anthology Network (AAN) [98]. Đây là một tập hợp các bài báo từ lĩnh vực ngôn ngữ học tính toán. Bộ dữ liệu này chứa 49,356 bối cảnh trích dẫn từ 19,771 bài báo được chọn ngẫu nhiên trong bộ dữ liệu AAN. Các bài báo này được xuất bản trong khoảng thời gian từ năm 2009 đến năm 2015. Mỗi ngữ cảnh trích dẫn gồm một câu trước, một câu sau và một câu chứa dấu trích dẫn, sao cho kích thước ngữ cảnh trích dẫn là ±200 ký tự. Bộ dữ liệu được chia thành ba tập hợp con: huấn luyện (30,390 bối cảnh), xác thực (9,381 bối cảnh) và kiểm tra (9,585 bối cảnh). Bộ huấn luyện và xác nhận được sử dụng để huấn luyện và điều chỉnh các mô hình khuyến nghị trích dẫn cục bộ, trong khi bộ kiểm tra được sử dụng để đánh giá.

FullTextPeerRead là bộ dữ liệu này được giới thiệu gần đây trong công bố của nhóm Jeong [15]. Bộ dữ liệu này chứa các câu ngữ cảnh cho các tài liệu tham khảo được trích dẫn và siêu dữ liệu bài báo, làm cho nó trở thành một bộ dữ liệu được tổ chức tốt cho bài toán khuyến nghị trích dẫn theo nhận biết ngữ cảnh. Bộ dữ liệu này dựa trên PeerRead [99], là tập hợp các bài báo và đánh giá khoa học từ lĩnh vực xử lý ngôn ngữ tự nhiên và học máy. Bộ dữ liệu PeerRead chứa 14,792 bài báo và 10,930 bài đánh giá từ nhiều nơi công bố khác nhau như ACL, NAACL, EMNLP, ICML, NIPS,.v.v… Bộ dữ liệu FullTextPeerRead mở rộng bộ dữ liệu PeerRead bằng cách thêm toàn bộ văn bản của bài báo, trích xuất bối cảnh trích dẫn từ các bài báo, và cung cấp siêu dữ liệu bổ sung như tác giả, địa điểm, năm xuất bản,.v.v… Bộ dữ liệu FullTextPeerRead cũng lọc ra các

bài báo có ít hơn 5 trích dẫn hoặc không bằng tiếng Anh. Bộ dữ liệu này chứa 4,898 bài báo và 17,247 bối cảnh trích dẫn. Mỗi ngữ cảnh trích dẫn bao gồm một câu trước, một câu sau và câu chứa vị trí trích dẫn. Bộ dữ liệu cũng cung cấp tiêu đề và bản tóm tắt của từng bài báo trích dẫn và từng bài báo được trích dẫn. Bộ dữ liệu này được chia thành ba tập con: huấn luyện (9,363 bối cảnh), xác thực (1,043 bối cảnh) và kiểm tra (6,841 bối cảnh). Bộ huấn luyện và bộ xác thực được sử dụng để huấn luyện và điều chỉnh các mô hình khuyến nghị trích dẫn nhận biết ngữ cảnh, trong khi bộ kiểm tra được sử dụng để đánh giá.

RefSeer là bộ dữ liệu về ngữ cảnh trích dẫn được trích xuất từ cơ sở dữ liệu CiteSeerX, là một tập hợp các bài báo từ nhiều lĩnh vực khoa học máy tính khác nhau. Bộ dữ liệu này chứa 3,772,154 ngữ cảnh trích dẫn từ 624,957 bài báo được xuất bản trên CiteSeerX cho đến năm 2014. Mỗi ngữ cảnh trích dẫn bao gồm một câu trước, một câu sau và câu chứa dấu vị trí trích dẫn. Bộ dữ liệu cũng cung cấp tiêu đề và tóm tắt của từng bài báo trích dẫn và bài báo được trích dẫn, cũng như siêu dữ liệu của các bài báo, chẳng hạn như tác giả, địa điểm, năm xuất bản,.v.v…Bộ dữ liệu được chia thành ba tập con: huấn luyện (3,521,582 ngữ cảnh), xác thực (124,551 ngữ cảnh) và kiểm tra (126,021 ngữ cảnh). Bộ huấn luyện và xác nhận được sử dụng để huấn luyện và điều chỉnh các mô hình khuyến nghị trích dẫn cục bộ, trong khi bộ kiểm tra được sử dụng để đánh giá.

4.3.3. Phương pháp đánh giá mô hình

Để đánh giá hiệu quả của mô hình SciBERT-GraphSAGE, luận án sử dụng 3 chỉ số đánh giá thông dụng cho các hệ thống khuyến nghị trích dẫn là độ chính xác trung bình MAP, xếp hạng đối ứng trung bình MRR và Recall@K. Hai chỉ số xếp hạng đối ứng trung bình MRR và Recall@K đã được giới thiệu ở chương 2 và chương 3 nên phần này sẽ giới thiệu về chỉ số độ chính xác trung bình MAP.

MAP là một tiêu chí đánh giá hiệu suất tổng quan của mô hình trên toàn bộ danh sách khuyến nghị. Nó đo lường mức độ chính xác trung bình của các khuyến nghị trên tất cả các truy vấn. Điều này giúp đánh giá xem mô hình có thể đưa ra nhiều kết quả đúng trong danh sách khuyến nghị không, và ở vị trí nào trong danh sách đó. MAP là một thước đo hữu ích để đánh giá chất lượng toàn diện của mô hình khi người dùng quan

tâm đến toàn bộ danh sách được khuyến nghị, thay vì chỉ vị trí đầu tiên. Bởi vì mô hình SciBERT-GraphSAGE dựa trên sự phù hợp về ngữ nghĩa và các liên kết trích dẫn để sắp xếp thứ tự ưu tiên của các kết quả khuyến nghị đưa ra cho người sử dụng, hơn nữa còn phải so sánh hiệu suất của mô hình này với các mô hình tiên tiến hiện nay, do đó việc đánh giá theo tiêu chí MAP là cần thiết. Dưới đây là bảng phân tích các khái niệm chính liên quan đến MAP:

· Độ chính xác (*Precision*) là tỷ lệ kết quả phù hợp trong số các kết quả được truy xuất. Ví dụ: nếu công cụ tìm kiếm trả về 10 kết quả và 4 trong số đó có liên quan thì độ chính xác là 0.4.

· Độ chính xác trung bình (*Average Precision*, *AP*): Giá trị trung bình của các giá trị độ chính xác ở từng kết quả liên quan trong danh sách xếp hạng. Ví dụ: nếu danh sách xếp hạng là [R, N, R, R, N, N, R, N], trong đó R biểu thị kết quả có liên quan và N biểu thị kết quả không liên quan, AP được tính là AP = (1/1 + 2/3 + 3/4 + 4/7) / 4 = 0.71.

Chỉ số MAP được tính bằng giá trị trung bình của các giá trị AP trên một tập hợp truy vấn như công thức (4.8):

$$MAP = \frac{\sum_{i=1}^{n} AP(i)}{n} \quad (4.8)$$

Ví dụ: nếu có 3 truy vấn và giá trị AP của chúng là 0.71, 0.82 và 0.65 thì MAP được tính là MAP = (0.71 + 0.82 + 0.65) / 3 = 0.73.

4.4. Đánh giá kết quả thực nghiệm và thảo luận

Trong phần này, luận án trình bày phân tích so sánh toàn diện về hiệu suất của mô hình SciBERT-GraphSAGE với một số mô hình khuyến nghị trích dẫn cục bộ tiên tiến có kết quả tốt nhất gần đây. Mục tiêu của phân tích này là đánh giá chặt chẽ điểm mạnh và khả năng của mô hình SciBERT-GraphSAGE với các mô hình tiên tiến hiện nay trên các tiêu chuẩn và bộ dữ liệu thường dùng. Bằng cách tiến hành so sánh chuyên sâu, luận án mong muốn cung cấp những hiểu biết sâu sắc về những cải thiện hiệu suất đã đạt được bằng phương pháp tiếp cận mới của mình và nêu bật những đóng góp tiềm năng của phương pháp đó cho lĩnh vực này. Để chứng minh những ưu điểm của mô hình mới, luận án đã đưa mô hình SciBERT-GraphSAGE vào một quy trình đo điểm chuẩn nghiêm

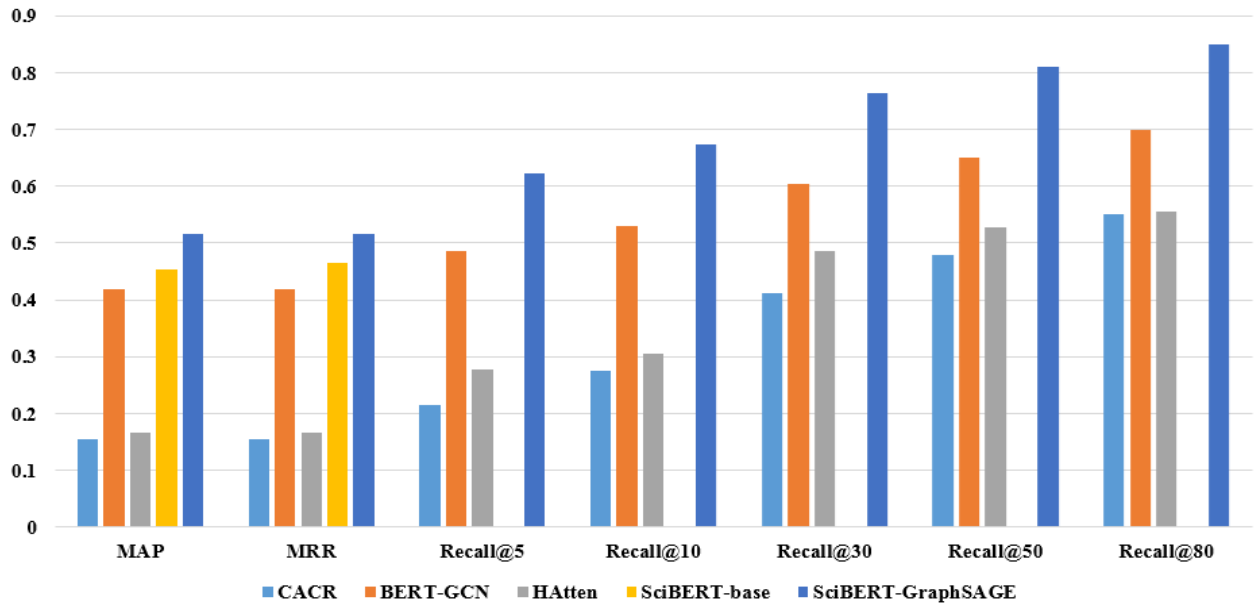
ngặt để so sánh với 5 mô hình tiên tiến nhất được công nhận rộng rãi trong cộng đồng nghiên cứu về khuyến nghị trích dẫn cục bộ. Việc đánh giá được thực hiện trên ba bộ dữ liệu được sử dụng phổ biến và thường được sử dụng: ACL-200 [12], RefSeer [12] và FullTextPeerRead [15]. Những bộ dữ liệu này bao gồm nhiều vấn đề phức tạp, phản ánh các tình huống và thách thức mà trong thế giới thực hay gặp phải với các hệ thống khuyến nghị trích dẫn cục bộ. Bằng cách sử dụng các bộ dữ liệu đa dạng như vậy, luận án mong muốn nắm bắt được sự hiểu biết toàn diện về hiệu suất của mô hình SciBERT-GraphSAGE trong nhiều bối cảnh khác nhau. Để đảm bảo sự công bằng và tính khoa học khi so sánh, kết quả thực nghiệm của các mô hình được mang ra so sánh ở 2 Bảng 4.3 và Bảng 4.4 đều lấy từ các công bố của tác giả các mô hình đó.

Với bộ dữ liệu FullTextPeerRead [15], mô hình SciBERT-GraphSAGE được so sánh với các mô hình tiên tiến như sau: (1) mô hình CACR [93] có cả bộ mã hóa bài báo và bộ mã hóa ngữ cảnh trích dẫn dựa trên mô hình LSTM; (2) mô hình BERT-GCN [15] kết hợp BERT cho bộ mã hóa văn bản và GCN cho bộ mã hóa siêu dữ liệu; (3) mô hình HAtten [16] bao gồm giai đoạn tìm nạp trước và giai đoạn xếp hạng lại; (4) mô hình SciBERT-base [100] được huấn luyện để dự đoán bài báo mà sẽ được trích dẫn từ một ngữ cảnh. Kết quả so sánh được thể hiện ở Bảng 4.3.

Bảng 4.3. Kết quả so sánh hiệu suất của mô hình SciBERT-GraphSAGE với 4 mô hình tiên tiến nhất trên bộ dữ liệu FullTextPeerRead

Mô hình	MAP	MRR	Recall@5	Recall@10	Recall@30	Recall@50	Recall@80
CACR [93]	0.1551	0.1549	0.2154	0.2761	0.4128	0.4794	0.5516
BERT-GCN [15]	0.4181	0.4179	0.4864	0.5291	0.6093	0.6495	0.6994
HAtten [16]	0.1672*	0.1670	0.2780*	0.3060	0.4850*	0.5270*	0.5560*
SciBERT-base [100]	0.454	0.466	-	-	-	-	-
SciBERT-GraphSAGE	**0.5162**	**0.5163**	**0.6217**	**0.6744**	**0.7636**	**0.8099**	**0.8504**

Trong kết quả công bố của họ, nhóm của Gu [16] không trình bày kết quả đánh giá thử nghiệm với các tiêu chí MAP và Recall@K, K = 5, 30, 50, 80 cho mô hình HAtten của họ. Để so sánh được, luận án đã tiến hành thực nghiệm lại mô hình HAtten[27] với bộ dữ liệu FullTextPeerRead và đánh dấu (*) trong các thành tích này. Các tác giả của mô hình SciBERT-base [100] cũng không công bố thực nghiệm của họ với tiêu chí Recall@K. Kết quả từ Bảng 4.3 cho thấy, trong số các mô hình được công bố gần đây thì kết quả thử nghiệm của 2 mô hình BERT-GCN [15] và SciBERT-base [100] với bộ dữ liệu FullTextPeerRead cho thành tích khả quan nhất. Tuy nhiên, bằng cách kết hợp SciBERT và GraphSAGE là những mô hình mới hơn và cải tiến so với BERT [68] và GCN [49], mô hình SciBERT-GraphSAGE thậm chí còn vượt trội hơn so với 2 mô hình này từ 22% đến 28% ở tất cả các chỉ số so sánh. Những kết quả này có thể được thấy rõ hơn trong Hình 4.5.



Hình 4.5. Kết quả so sánh hiệu năng của mô hình SciBERT-GraphSAGE với 4 mô hình tiên tiến nhất trên tập dữ liệu FullTextPeerRead

Với hai bộ dữ liệu ACL-200 [12] và RefSeer [12], mô hình SciBERT-GraphSAGE tiếp tục được so sánh với các thành tựu nghiên cứu được công bố gần đây cho bài toán khuyến nghị trích dẫn cục bộ: (1) mô hình HAtten [16] bao gồm giai đoạn

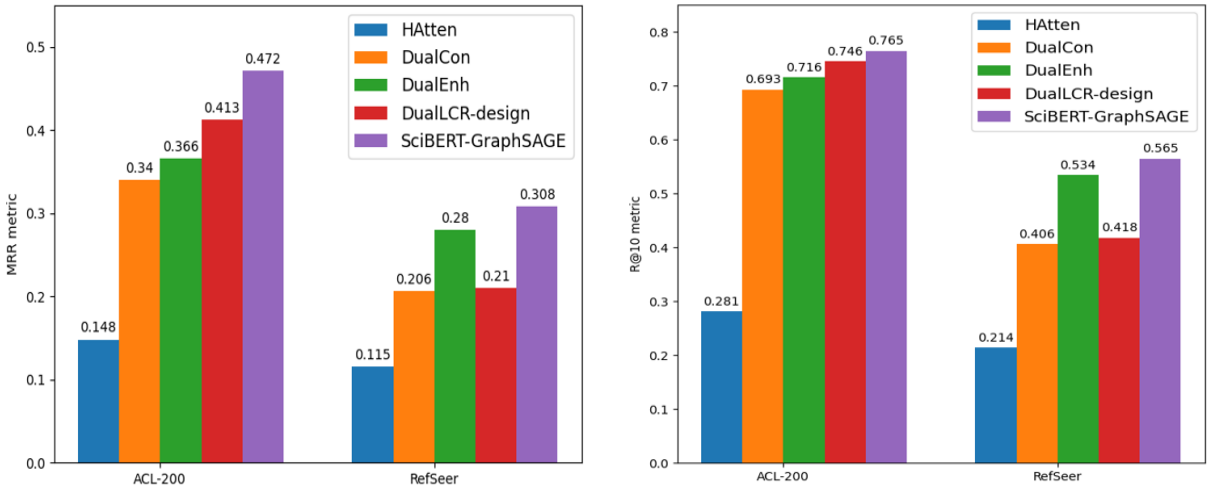
[27] https://github.com/nianlonggu/Local-Citation-Recommendation

tìm nạp trước và giai đoạn sắp xếp lại; (2) mô hình FLCR [101] có khả năng mở rộng cho nhiều chủ đề, nhằm cải thiện hiệu suất bằng cách tận dụng thông tin từ nhiều nguồn chủ đề khác nhau. (3) mô hình DualEnh và DualCon [12] sử dụng cả nội dung ngữ nghĩa và dữ liệu thông tin học thuật để đánh giá chất lượng của từng bài báo ứng viên trích dẫn tiềm năng; (4) mô hình DualLCR-design [13] đã tối ưu hóa các tham số để mang lại hiệu suất tốt nhất. Kết quả so sánh được thể hiện ở Bảng 4.4.

Bảng 4.4. Kết quả so sánh hiệu suất của mô hình SciBERT-GraphSAGE với 4 mô hình tiên tiến nhất trên 2 bộ dữ liệu ACL-200 và RefSeer

Bộ dữ liệu	**ACL-200**		**RefSeer**	
Mô hình	**MRR**	**Recall@10**	**MRR**	**Recall@10**
HAtten [16]	0.148	0.281	0.115	0.214
FLCR [101]	0.331	0.604	0.196	0.376
DualCon [12]	0.340	0.693	0.206	0.406
DualEnh [12]	0.366	0.716	0.280	0.534
DualLCR-design [13]	0.413	0.746	0.210	0.418
SciBERT-GraphSAGE	**0.472**	**0.765**	**0.308**	**0.565**

Kết quả thực nghiệm từ Bảng 4.4 cũng cho thấy, trong các công bố gần đây, Medić và Šnajder [12] [13] đã đề xuất các mô hình có kết quả tốt nhất. Với bộ dữ liệu ACL-200, mô hình DualEnh [12] của họ thu được kết quả vượt trội so với các mô hình khác, trong khi với bộ dữ liệu RefSeer, mô hình DualLCR-design [13] cũng thu được kết quả tốt hơn. Tuy nhiên, mô hình SciBERT-GraphSAGE thậm chí còn mang lại thành tích tốt hơn. Với bộ dữ liệu ACL-200, sử dụng số liệu MRR và Recall@10, kết quả của luận án lần lượt tốt hơn 14% và 3%, trong khi với bộ dữ liệu RefSeer, mô hình SciBERT-GraphSAGE cũng cho kết quả tốt hơn lần lượt là 10% và 6% cho các chỉ số tương ứng. Những kết quả này có thể được thấy rõ hơn trong Hình 4.5.



Hình 4.6. Kết quả so sánh hiệu năng của mô hình SciBERT-GraphSAGE với 4 mô hình tiên tiến nhất trên 2 tập dữ liệu ACL-200 và RefSeer

4.5. So sánh với mô hình RHN-DualLCR

Để tăng tính kết nối của các chương cũng như so sánh giữa các mô hình được đề xuất với nhau, luận án thực hiện đánh giá hiệu suất của mô hình SciBERT với mô hình RHN-DualLCR đã được đề cập ở chương 3.

Bảng 4.5. Kết quả so sánh hiệu suất của 2 mô hình SciBERT-GraphSAGE và RHN-DualLCR

Bộ dữ liệu	**ACL-200**		**RefSeer**	
Mô hình	**MRR**	**Recall@10**	**MRR**	**Recall@10**
RHN-DualLCR	0.403	0.748	0.307	0.582
SciBERT-GraphSAGE	0.472	0.765	0.308	0.565

Dựa trên kết quả thực nghiệm được hiển thị trong Bảng 4.5, có thể đánh giá hai mô hình SciBERT-GraphSAGE và RHN-DualLCR trên hai bộ dữ liệu ACL-200 và RefSeer với các tiêu chí Recall@10 và MRR như sau:

Trên bộ dữ liệu ACL-200, SciBERT-GraphSAGE có hiệu suất cao hơn RHN-DualLCR về cả hai tiêu chí MRR và Recall@10. Với tiêu chí MRR, SciBERT-GraphSAGE đạt 0.472, trong khi RHN-DualLCR đạt 0.403. Điều này cho thấy SciBERT-GraphSAGE có khả năng dự đoán vị trí đúng của các trích dẫn tốt hơn trên bộ dữ liệu ACL-200. Với tiêu chí Recall@10, SciBERT-GraphSAGE đạt 0.765, cao hơn so

với 0.748 của RHN-DualLCR, thể hiện rằng mô hình SciBERT-GraphSAGE có khả năng truy xuất trích dẫn chính xác trong top-10 tốt hơn trên bộ dữ liệu ACL-200.

Trên bộ dữ liệu RefSeer, RHN-DualLCR có hiệu suất tốt hơn SciBERT-GraphSAGE ở tiêu chí Recall@10. Recall@10 của RHN-DualLCR đạt 0.582, cao hơn so với 0.565 của SciBERT-GraphSAGE. Điều này chỉ ra rằng mô hình RHN-DualLCR có khả năng thu hồi trích dẫn chính xác trong top-10 tốt hơn SciBERT-GraphSAGE trên bộ dữ liệu RefSeer, mặc dù chênh lệch không lớn. Tuy nhiên, khi xét về tiêu chí MRR, SciBERT-GraphSAGE có hiệu suất cao hơn một chút so với RHN-DualLCR. MRR của SciBERT-GraphSAGE đạt 0.308, trong khi RHN-DualLCR đạt 0.307, cho thấy khả năng dự đoán vị trí đúng của các trích dẫn của SciBERT-GraphSAGE tốt hơn chút ít.

SciBERT-GraphSAGE với khả năng lọc theo đồ thị hoạt động tốt hơn trên cả hai bộ dữ liệu với tiêu chí MRR, điều này có thể giải thích bởi khả năng của mô hình trong việc tận dụng mối quan hệ giữa các trích dẫn thông qua mô hình GraphSAGE. RHN-DualLCR, với cách tiếp cận lọc cộng tác, có hiệu suất tốt hơn về Recall@10 trên bộ dữ liệu RefSeer lớn hơn. Điều này có thể phản ánh rằng với bộ dữ liệu lớn hơn, phương pháp lọc cộng tác có thể tìm kiếm tốt hơn các trích dẫn trong top-10. Việc bộ dữ liệu ACL-200 nhỏ hơn có thể giúp mô hình SciBERT-GraphSAGE dễ dàng khai thác và tận dụng các mối quan hệ đồ thị, trong khi bộ dữ liệu lớn hơn như RefSeer giúp phương pháp lọc cộng tác của RHN-DualLCR phát huy tác dụng trong việc truy xuất chính xác.

4.6. Kết luận chương 4

Chương 4 đã tập trung vào nghiên cứu mô hình kết hợp giữa lọc nội dung và lọc dựa vào đồ thị. Luận án đã đề xuất một mô hình mới cho bài toán khuyến nghị trích dẫn, đó là mô hình lai ghép SciBERT-GraphSAGE, là sự kết hợp của SciBERT để biểu diễn nội dung của bài báo và đồ thị GraphSAGE để biểu diễn các liên kết trích dẫn. Mô hình SciBERT [18] là phiên bản chuyên biệt của BERT [68] đã được huấn luyện đặc biệt cho các nhiệm vụ trong lĩnh vực nghiên cứu khoa học và kỹ thuật. GraphSAGE [19] là một nền tảng học máy mạnh mẽ dựa trên đồ thị được thiết kế để tìm hiểu các biểu diễn từ đồ thị có số lượng nút lớn một cách hiệu quả. Bằng cách coi các bài báo học thuật là các nút, siêu dữ liệu là các thuộc tính của nút và các trích dẫn là các cạnh trong đồ thị, mạng

trích dẫn tạo thành một cấu trúc đồ thị tự nhiên, trong đó các bài báo được kết nối với nhau thông qua các mối quan hệ trích dẫn của chúng. Kết quả thực nghiệm khi so sánh trên 3 bộ dữ liệu tiêu chuẩn thường dùng (ACL-200 [12], RefSeer [12] và FullTextPeerRead [15]) với 6 mô hình tiên tiến nhất (CACR [93], BERT-GCN [15], HAtten [16], DualEnh và DualCon [12], DualLCR-design [13] và SciBERT-base [100]) đều cho kết quả vượt trội ở 3 chỉ số (MAP, MRR và Recall@K), chứng tỏ phương pháp tiếp cận của luận án là đúng đắn và đã đạt được những kết quả thực sự nổi bật.

Các kết quả nghiên cứu của chương 4 này đã được công bố tại công trình [CT3] và [CT5] trong phần "Danh mục các công trình đã công bố của tác giả".

KẾT LUẬN VÀ KIẾN NGHỊ

Với số lượng và tốc độ phát triển như vũ bão của các ấn bản khoa học ngày nay, việc tìm kiếm tài liệu thích hợp để trích dẫn ngày càng trở nên khó khăn hơn đối với hàng triệu người làm khoa học công nghệ. Sử dụng các công cụ tìm kiếm hiện tại để tìm kiếm dựa trên từ khóa chung thì đều phải mất công sức để sàng lọc thêm từ các kết quả nhận được và không làm giảm đáng kể khối lượng công việc của khảo sát tài liệu. Vì vậy, xây dựng hệ thống khuyến nghị trích dẫn với mục đích cung cấp cho các nhà khoa học những gợi ý mang tính xây dựng, đồng thời nâng cao hiệu quả và chất lượng công việc tìm kiếm bài viết đã thu hút sự quan tâm của cộng đồng các nhà nghiên cứu trong hai thập kỷ gần đây.

Cho đến nay, đã có rất nhiều các cách tiếp cận khác nhau được đưa ra để giải quyết cho bài toán này. Tuy nhiên, các giải pháp cho bài toán khuyến nghị trích dẫn vẫn còn có những hạn chế nhất định. Hạn chế thứ nhất liên quan đến việc mô hình khuyến nghị chưa được cung cấp đủ thông tin ngữ cảnh của các bài báo khoa học. Hạn chế thứ hai liên quan đến việc các mô hình vẫn chưa sử dụng những thành tựu mới nhất của các nghiên cứu về học sâu. Hạn chế thứ ba liên quan đến việc các mô hình khuyến nghị trích dẫn hiện tại đang tập trung vào ngữ cảnh trích dẫn và nội dung của bài báo ứng viên mà chưa quan tâm đúng mức đến siêu dữ liệu của bài báo, như là tên tác giả, năm và địa điểm công bố của bài báo. Mục tiêu nghiên cứu của luận án hướng tới góp phần giải quyết ba hạn chế đã đề cập ở trên.

1. Kết luận các đóng góp chính của luận án

Các kết quả nghiên cứu đã trả lời được các câu hỏi và vấn đề nghiên cứu đặt ra, góp phần giải quyết được những hạn chế trong cả 3 hướng tiếp cận của các mô hình khuyến nghị trích dẫn hiện nay. Cụ thể, luận án đã có những đóng góp chính như sau:

1) Với hướng tiếp cận lọc nội dung, đề xuất các giải pháp để nâng cao hiệu suất cho mô hình mạng nơ-ron trích dẫn NCN. Bao gồm:

- Thêm thông tin phụ trợ bổ sung vào mô hình NCN [10] [11]. Tiêu đề của mỗi bài báo khoa học là thông tin quan trọng khi xem xét trích dẫn, tuy nhiên mô hình

NCN đã bỏ qua yếu tố này. Do đó, để cải thiện hiệu suất trích dẫn, luận án đã thêm cả tiêu đề của bài báo như là một thông tin đầu vào cho mô hình.

- Áp dụng mô hình xử lý ngôn ngữ tự nhiên BERT [68] để tiền xử lý dữ liệu trích dẫn trước khi đưa vào mô hình, cũng như sử dụng cơ chế chú ý trong phần mở rộng mô hình NCN.
- Tiến hành các thử nghiệm sâu rộng trong tập dữ liệu arXiv tiêu chuẩn [10] [11] để cho thấy hiệu suất của mô hình sau khi cải tiến so với mô hình NCN tiên tiến gần đây. Hơn nữa, luận án cũng cung cấp các khảo sát kỹ lưỡng về ảnh hưởng của các siêu tham số vào mô hình NCN mới để chứng minh tính hiệu quả của nó cả về độ ổn định và khả năng mở rộng cho việc triển khai trong thực tế.

Các đề xuất và kết quả nghiên cứu này được công bố tại công trình [CT1] trong phần "Danh mục các công trình đã công bố của tác giả".

2) Với hướng tiếp cận kết hợp lọc nội dung và lọc cộng tác, đề xuất một mô hình mới tên là RHN-DualLCR, trong đó bao gồm các giải pháp để nâng cao hiệu suất cho mô hình kép DualLCR cho bài toán khuyến nghị trích dẫn đã được công bố bởi Medić và Šnajder [12] [13]. Bao gồm:

- Sử dụng ScispaCy [91] để tách các câu trong các bài báo khoa học trước khi tiền xử lý. Thay thế cơ chế nhúng văn bản theo phương pháp AI2 của Bhagavatula [87] bằng SciBERT [18].
- Để đạt được sự trình bày tuần tự tốt hơn về thông tin văn bản trích dẫn đã được nhúng thông qua SciBERT ở bước trước, tiếp tục thay thế bộ nhớ hai chiều (BiLSTM) trong mô hình hiện tại bằng mạng hồi quy mới (Recurrent Highway Network - RHN) [20].
- Đánh giá hiệu suất của mô hình RHN-DualLCR so với 2 mô hình trước cải thiện là DualLCR [12] và DualLCR-Design [13]. Thực hiện đánh giá trên 3 bộ dữ liêu tiêu chuẩn là ACL-200, ACL-600 và RefSeer với các chỉ số hiệu suất MRR và Recall@10 để chứng minh sự cải tiến rõ ràng của mô hình RHN-DualLCR. Luận án cũng đã trình bày các khảo sát tinh chỉnh các siêu tham số của mô hình mạng

hồi quy (RHN) [20] để chứng minh tính hiệu quả của nó về cả tính ổn định và khả năng mở rộng cho việc triển khai trong thực tế.

Các đề xuất và kết quả nghiên cứu này được công bố tại công trình [CT2] và [CT4] trong phần "Danh mục các công trình đã công bố của tác giả".

3) Với hướng tiếp cận lọc nội dung kết hợp với lọc dựa vào đồ thị, đề xuất mô hình khuyến nghị trích dẫn mới SciBERT-GraphSAGE bằng cách kết hợp 2 thành tựu mới nhất trong xử lý ngôn ngữ tự nhiên SciBERT [18] và tạo dữ liệu nhúng của liên kết trích dẫn bằng đồ thị GraphSAGE [19]. Bao gồm:

- Sử dụng mô hình xử lý ngôn ngữ tự nhiên cho các bài báo khoa học SciBERT [18] để biểu diễn các thông tin dạng văn bản như tiêu đề và tóm tắt của bài báo trích dẫn và được trích dẫn, ngữ cảnh trích dẫn bài báo.
- Sử dụng mô hình đồ thị GraphSAGE [19] để biểu diễn mối liên hệ trích dẫn trong các bài báo. Mỗi bài báo được xem như là 1 nút, các siêu dữ liệu như là tên tác giả, địa điểm xuất bản và năm xuất bản được xem như thuộc tính của nút, và mỗi quan hệ trích dẫn là một cạnh của đồ thị. Dữ liệu được mã hóa từ cả SciBERT và GraphSAGE đều được móc nối (*concat*) với nhau và chuyển đến mạng nơ-ron tiếp liệu. Sau đó, lớp đầu ra hồi quy softmax được tạo ra và entropy chéo được sử dụng làm hàm mất mát cho quá trình huấn luyện.
- Cuối cùng, để chứng minh hiệu quả của mô hình SciBERT-GraphSAGE, luận án đã xây dựng lại 3 bộ dữ liệu thường dùng ACL-200[18], RefSeer[18] và FullTextPeerRead [14] bằng cách thêm thuộc tính cho các nút (trong trường hợp này là mỗi nút là một bài báo khoa học). Luận án cũng phân tích kết quả đánh giá với năm mô hình khuyến nghị trích dẫn hiện đại nhất và cung cấp những hiểu biết sâu sắc quan trọng cho sự phát triển trong tương lai của các phương pháp khuyến nghị trích dẫn cục bộ.

Các đề xuất và kết quả nghiên cứu này được công bố tại công trình [CT3] và [CT5] trong phần "Danh mục các công trình đã công bố của tác giả".

2. Hạn chế và khả năng áp dụng của các mô hình trong thực tế

Mặc dù các mô hình khuyến nghị trích dẫn được đề xuất trong luận án đã mang lại những cải tiến đáng kể về mặt hiệu suất, đặc biệt trong việc nâng cao hiệu suất của các mô hình khuyến nghị trích dẫn hiện nay, nhưng chúng vẫn tồn tại một số hạn chế cần được xem xét. Những hạn chế này không chỉ ảnh hưởng đến hiệu quả hoạt động của mô hình trong các điều kiện cụ thể mà còn đặt ra những thách thức khi triển khai vào môi trường thực tế với dữ liệu phức tạp và đa dạng. Trong phần này, luận án phân tích những khía cạnh cần được cải thiện và những điểm yếu tiềm tàng của các mô hình, nhằm đưa ra những định hướng phát triển trong các nghiên cứu tương lai.

Với mô hình Enhanced-NCN ở chương 2, do tiếp cận theo hướng lọc nội dung hạn chế là chưa sử dụng hết các thông tin khác của bài báo như là siêu dữ liệu hoặc liên kết trích dẫn. Ngoài ra, nếu tiêu đề không chứa nhiều thông tin hoặc không liên quan chặt chẽ đến nội dung trích dẫn, mô hình Enhanced-NCN có thể không khai thác được lợi thế từ nguồn dữ liệu này. Về mặt hiệu suất, mặc dù mặc dù có cải thiện hơn mô hình NCN ban đầu, nhưng vẫn chưa thể so sánh được với các mô hình lọc kết hợp ở các chương 3 và chương 4.

Với mô hình RHN-DualLCR ở chương 3, mặc dù đã kết hợp giữa SciBERT và RHN để cải thiện hiệu suất so với mô hình DualLCR ban đầu, thì vẫn có những hạn chế nhất định. Ngoài dữ liệu văn bản của bài báo thì đầu vào của mô hình RHN-DualLCR yêu cầu siêu dữ liệu của bài báo, nên mô hình này sẽ gặp khó khăn với các bộ dữ liệu mà siêu dữ liệu của các bài báo không đầy đủ. Ngoài ra, luận án mới thử nghiệm mô hình RHN-DualLCR trên các bộ dữ liệu tiêu chuẩn chứ chưa thực nghiệm trên các bộ dữ liệu lớn hơn trong thực tế. Thêm nữa, so sánh với mô hình SciBERT-GraphSAGE thì nhìn chung hiệu suất của RHN-DualLCR không tốt bằng. Kết quả so sánh này đã được trình bày chi tiết ở mục 4.5 của chương 4.

Với mô hình SciBERT-GraphSAGE ở chương 4 được kết hợp các thành tựu tiên tiến trong xử lý ngôn ngữ tự nhiên SciBERT và tạo dữ liệu nhúng từ đồ thị liên kết trích dẫn SciBERT. Mặc dù đây là mô hình cho thấy hiệu suất tốt hơn các mô hình tiên tiến hiện nay cũng như 2 mô hình đã được đề xuất ở chương 2 và chương 3, mô hình

SciBERT-GraphSAGE cũng có những hạn chế nhất định. Hạn chế đầu tiên là mô hình này yêu cầu bộ dữ liệu đầu vào phải có đầy đủ thông tin liên kết trích dẫn của các bài báo. Hạn chế thứ hai là chi phí tính toán cao do mô hình có tính toán đồ thị liên kết trích dẫn GraphSAGE. Hạn chế thứ ba là cũng như mô hình RHN-DualLCR, trong luận án cũng mới chỉ thực nghiệm mô hình SciBERT-GraphSAGE trên 3 bộ dữ liệu tiêu chuẩn thường dùng cho các mô hình khuyến nghị tiên tiến hiện nay chứ chưa thực nghiệm trên các bộ dữ liệu lớn hơn trong thực tế.

3. Kiến nghị các hướng phát triển trong tương lai

Khuyến nghị trích dẫn tự động sẽ đóng một vai trò quan trọng trong việc trợ giúp các nhà khoa học tìm được các tài liệu khoa học phù hợp và thích đáng để trích dẫn trong các nghiên cứu của họ. Tuy nhiên, từ hướng nghiên cứu của luận án này, vẫn còn một số vấn đề cần được tiếp tục nghiên cứu trong tương lai để cải thiện hiệu quả của hệ thống khuyến nghị như sau:

- Trong phạm vi nghiên cứu của luận án mới chỉ thực nghiệm các mô hình cải tiến và xuất của mình với các bộ dữ liệu arXiv CS (chương 2), ACL-ARC, RefSeer (chương 3) và FullTextPeerRead (chương 4). Tuy nhiên, các mô hình này vẫn cần phải thực nghiệm hiệu quả trên các bộ bộ dữ liệu lớn hơn như DBLP, PubMed, OpenCorpus[28] hay S2ORC[29]. Việc thử nghiệm các mô hình với các bộ dữ liệu khác nhau sẽ là cơ sở để điều chỉnh các mô hình cho phù hợp hơn với các dữ liệu bài báo trong thực tế.
- Đồ thị liên kết trích dẫn GraphSAGE mà luận án đã sử dụng trong mô hình SciBERT-GraphSAGE ở chương 4 thì hiện nay vẫn được tiếp tục nghiên cứu và phát triển với các biến thể mới hơn như là E-GraphSAGE [102] hoặc các đồ thị liên kết trích dẫn khác [47] [50]. Có cơ sở để tin tưởng rằng có tiềm năng áp dụng các biến thể này vào bài toán khuyến nghị trích dẫn để có thể có những kết quả tốt hơn.

[28] https://api.semanticscholar.org/corpus
[29] https://github.com/allenai/s2orc

- Hơn nữa, luận án giữ quan điểm rằng có một con đường đầy hứa hẹn để nghiên cứu sâu hơn nằm ở việc sử dụng những tiến bộ gần đây trong việc nhúng mạng không đồng nhất cho các hệ thống khuyến nghị [54] [55] [103] hoặc mạng tích chập đồ thị [47] [51], sẽ được áp dụng để giải quyết các thách thức của vấn đề khuyến nghị trích dẫn.

DANH MỤC CÔNG TRÌNH CÔNG BỐ LIÊN QUAN ĐẾN LUẬN ÁN

TẠP CHÍ KHOA HỌC QUỐC TẾ

[CT1]	**Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enhanced context-aware citation recommendation with auxiliary textual information based on an auto-encoding mechanism," *Applied Intelligence*, 53(14), 2023, pp. 17381–17390 ISSN/eISSN: 0924-669X/1573-7497 https://doi.org/10.1007/s10489-022-04423-1 Scopus indexed (Q2), SCIE IF:5.086 (Q2)
[CT2]	**Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enhancing local citation recommendation with recurrent highway networks and SciBERT-based embedding", *Expert Systems with Applications* (*ESWA*), Volume 243, 2024, 122911, ISSN 0957-4174 https://doi.org/10.1016/j.eswa.2023.122911 Scopus indexed (Q1), SCIE IF:8.5 (Q1)
[CT3]	**Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Ngoc Thanh Nguyen and Bay Vo, "A hybrid citation recommendation model with SciBERT and GraphSAGE", *IEEE Transactions on Systems, Man and Cybernetics: Systems*, Volume 55, Issue 2, pp. 852-863, Feb. 2025. ISSN/eISSN: 2168-2216/2168-2232 https://doi.org/10.1109/TSMC.2024.3490774 Scopus indexed (Q1), SCIE IF:8.7 (Q1)

HỘI NGHỊ KHOA HỌC

[CT4]	**Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "Enrich textual information for Hierarchical-Attention Text Encoder in Local Citation Recommendation", *Kỷ yếu Hội thảo quốc gia lần thứ XXV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông* – Hà Nội, 8-9/12/2022 (National Symposium of Selected ICT Problems – VNICT(@) 2022) ISBN:978-604-67-2508-4.
[CT5]	**Thi N. Dinh**, Phu Pham, Giang L. Nguyen, Bay Vo, "A context-aware citation recommendation model with SciBERT and GraphSAGE", *Kỷ yếu Hội thảo quốc gia lần thứ XXVI: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông* – Bắc Ninh, 5-6/10/2023 (National Symposium of Selected ICT Problems – VNICT(@) 2023) ISBN: 978-604-67-2746-0.

DANH MỤC TÀI LIỆU THAM KHẢO

[1] O. Küçüktunç, E. Saule, K. Kaya and Ü. V. Çatalyürek, "Diversifying Citation Recommendations," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 4, pp. 1-21, 2015.

[2] L. Bornmann and R. Mutz, "Growth rates of modern science: a bibliometric analysis based on the number of publications and cited references," *The Journal of the Association for Information Science and Technology (JASIST)*, vol. 66, no. 11, pp. 2215-2222, 2015.

[3] S. Fortunato, C. T. Bergstrom, K. Borner and J. A. Evans, "Science of science," *Science*, vol. 359, p. eaao0185, 2018.

[4] M. Ware and M. Mabe, "The STM Report: An overview of scientific and scholarly journal publishing," 2015 STM: International Association of Scientific, Technical and Medical Publishers, 2015.

[5] S. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. Lam, A. Rashid, J. Konstan and J. Riedl, "On the recommending of citations for research papers," in *The 2002 ACM conference on computer supported cooperative work*, 2002.

[6] Z. Medić and J. Šnajder, "A Survey of Citation Recommendation Tasks and Methods," *Journal of Computing and Information Technology*, vol. 28, no. 3, pp. 183-205, 2020.

[7] M. Färber and A. Jatowt, "Citation recommendation: approaches and datasets," *International Journal on Digital Libraries 21*, p. 375–405, 2020.

[8] Z. Ali, I. Ullah, A. Khan, A. U. Jan and K. Muhammad, "An overview and evaluation of citation recommendation models," *Scientometrics*, vol. 126, no. 5, pp. 4083-4119, 2021.

[9] S. Ma, C. Zhang and X. Liu, "A review of citation recommendation: from textual content to enriched context," *Scientometrics,* vol. 122, p. 1445–1472, 2020.

[10] T. Ebesu and Y. Fang, "Neural Citation Network for Context-Aware Citation Recommendation," in *The 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'2017)*, 2017.

[11] M. Färber, T. Klein and J. Sigloch, "Neural Citation Recommendation: A Reproducibility Study," in *The 10th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2020)*, 2020.

[12] Z. Medić and J. Šnajder, "Improved Local Citation Recommendation Based on Context Enhanced with Global Information," in *Proceedings of the First Workshop on Scholarly Document Processing (EMNLP)*, 2020.

[13] Z. Medić and J. Šnajder, "An empirical study of the design choices for local citation recommendation systems," *Expert Systems with Applications,* vol. 200, p. 116852, 2022.

[14] S. Hochreiter and J. Schmidhuber, "Long Short-term Memory," *Neural Computation,* vol. 9, no. 8, p. 1735–1780, 1997.

[15] C. Jeong, S. Jang, H. Shin, E. Park and S. Choi, "A context-aware citation recommendation model with BERT and graph convolutional networks," *Scientometrics,* vol. 124, no. 3, pp. 1907-1922, 2020.

[16] N. Gu, Y. Gao and R. H. Hahnloser, "Local Citation Recommendation with Hierarchical-Attention Text Encoder and SciBERT-Based Reranking," in *44th European Conference on Information Retrieval (44th ECIR)*, 2022.

[17] T. Dai, L. Zhu, Y. Wang and K. M. Carley, "Attentive Stacked Denoising Autoencoder with Bi-LSTM for Personalized Contextaware Citation Recommendation," in *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[18] I. Beltagy, K. Lo and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language (EMNLP-IJCNLP)*, 2019.

[19] W. L. Hamilton, R. Ying and J. Leskovec, "Inductive representation learning on large graphs," in *The 31st International Conference on Neural Information Processing Systems(NIPS'17)*, 2017.

[20] J. G. Zilly, R. K. Srivastava, J. Koutník and J. Schmidhuber, "Recurrent Highway Networks," *arXiv:1607.03474v5,* 2017.

[21] H. Oldenburg, "Epistle Dedicatory," *Philosophical Transactions of the Royal Society,* 1665.

[22] K. E. White, "Science and Engineering Publication Output Trends: 2014 Shows Rise of Developing Country Output while Developed Countries Dominate Highly Cited Publications," *National Science Foundation; Social, Behavioral and Economic Sciences,* 2017.

[23] P. O. Larsen and M. V. Ins, "The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index," *Scientometrics,* vol. 84, no. 3, pp. 575-603, 2010.

[24] B. Kurilla, "Can too much science be a bad thing? Growth in scientific publishing as a barrier to science communication (2019)," http://geekpsychologist.com/can-too-much-science-be-a-bad-thing-growth-in-scientific-publishing-as-a-barrier-to-science-communication/. Accessed 19 June 2019.

[25] R. Van Noorden, "Scientists May Be Reaching a Peak in Reading Habits," *Nature News,* 2014.

[26] W. Ammar, "Construction of the Literature Graph in Semantic Scholar," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[27] J. Zhang and L. Zhu, "Citation recommendation using semantic representation of cited papers' relations and content," *Expert Systems With Applications,* vol. 187, 2022.

[28] Y. Koren and R. Bell, "Advances in Collaborative Filtering," in *Recommender Systems Handbook*, 2022, p. 91–142.

[29] H. Liu, X. Kong, X. Bai, W. Wang, T. M. Bekele and F. Xia, "Context-Based Collaborative Filtering for Citation Recommendation," *IEEE Access*, vol. 3, pp. 1695-1703, 2015.

[30] T. Bansal, D. Belanger and A. McCallum, "Ask the GRU: Multi-task Learning for Deep Text Recommendations," in *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems*, 2016.

[31] K. Sugiyama and M.-Y. Kan, "Exploiting potential citation papers in scholarly paper recommendation," in *Proceedings of the 13th ACM/IEEE-CS joint conference on digital libraries (JCDL)*, 2013.

[32] L. Galke, F. Mai, I. Vagliano and A. Scherp, "Multi-modal adversarial autoencoders for recommendations of citations and subject labels," in *The 26th conference on user modeling, adaptation and personalization*, 2018.

[33] Z. Ali, G. Qi, P. Kefalas, W. A. Abro and B. Ali, "A graph-based taxonomy of citation recommendation models," *Artificial Intelligence Review,* vol. 53, pp. 5217-5260, 2020.

[34] S. Khusro, Z. Ali and I. Ullah, "Recommender Systems: Issues, Challenges, and Research Opportunities," in *Information Science and Applications (ICISA) 2016*, Springer, 2016, p. 1179–1189.

[35] B. Alhijawi and Y. Kilani, "A collaborative filtering recommender system using genetic algorithm," *Information Processing & Management,* vol. 57, no. 6, 2020.

[36] Q. V. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *31st International Conference on Machine Learning, ICML 2014*, 2014.

[37] L. F. Ribeiro, P. H. Saverese and D. R. Figueiredo, "struc2vec: Learning Node Representations from Structural Identity," in *KDD '17: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

[38] J. Han, Y. Song, W. X. Zhao, S. Shi and H. Zhang, "Hyperdoc2vec: Distributed Representations of Hypertext Documents.," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)(Volume 1: Long Papers)*, 2018.

[39] X. Kong, M. Mao, W. Wang, J. Liu , B. Xu, "VOPRec: Vector representation learning of papers with text information and structural identity for recommendation," *Transactions on Emerging Topics in Computing*, pp. 1-12, 2019.

[40] Y. Zhang and Q. Ma, "Citation recommendations considering content and structural context embedding," in *International Conference on Big Data and Smart Computing*, 2020.

[41] W. Huang, S. Kataria, C. Caragea, P. Mitra, C. L. Giles and L. Rokach, "Recommending Citations Translating Papers into References," in *The 21st ACM International Conference on Information & Knowledge Management*, 2012.

[42] W. Huang, Z. Wu, C. Liang, P. Mitra and L. C. Giles, "A Neural Probabilistic Model for Context based Citation Recommendation," in *The 29th AAAI Conference on Artificial Intelligence*, 2015.

[43] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford, "Okapi at TREC-3," in *Proceedings of the Third Text REtrieval Conference (TREC 1994)*, 1994.

[44] M. Färber and A. Sampath, "HybridCite: A Hybrid Model for Context-AwareCitation Recommendation," in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*, 2020.

[45] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, pp. 5998-6008, 2017.

[46] P. Pham, L. Nguyen, B. Vo and U. Yun, "Bot2Vec: A general approach of intra-community oriented representation learning for bot detection in different types of social networks," *Information Systems, 101771*, 2021.

[47] S. Wu, F. Sun, W. Zhang, X. Xie and B. Cui, "Graph Neural Networks in Recommender Systems: A Survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1-37, 2022.

[48] M. Péter and K. Attila, "Influential Performance of Nodes Identified by Relative Entropy in Dynamic Networks," *Vietnam Journal of Computer Science*, vol. 8, no. 1, pp. 93-112, 2021.

[49] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[50] X. He, K. Deng, X. Wang, Y. Li, Z. YongDong and M. Wang, "LightGCN: Simplifying and Powering Graph Convolution Network for Recommendation,"

in *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[51] Y. Feng, H. You, Z. Zhang, R. Ji and Y. Gao, "Hypergraph Neural Networks," in *Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 3558-3565.*, 2019.

[52] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò and B. Yoshua, "Graph Attention Networks," in *International Conference on Learning Representations (ICLR)*, 2018.

[53] Y. Li, D. Tarlow, M. Brockschmidt and R. Zemel, "Gated Graph Sequence Neural Networks," in *International Conference on Learning Representations (ICLR)*, 2017.

[54] P. Pham, L. T. Nguyen, N. T. Nguyen, R. Kozma and B. Vo, "A hierarchical fused fuzzy deep neural network with heterogeneous network embedding for recommendation," *Information Sciences,* vol. 620, pp. 105-124, 2023.

[55] P. Pham, L. T. Nguyen, N. T. Nguyen, W. Pedrycz, U. Yun, J. C. W. Lin and B. Vo, "An approach to semantic-aware heterogeneous network embedding for recommender systems," *IEEE Transactions on Cybernetics,* vol. 53, no. 9, pp. 6027 - 6040, 2023.

[56] Z. Wang, G. Lin, H. Tan, Q. Chen and X. Liu, "CKAN: Collaborative Knowledge-aware Attentive Network for Recommender Systems," in *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[57] C. Chen, W. Ma, M. Zhang, Z. Wang, X. He, C. Wang, Y. Lu and S. Ma, "Graph Heterogeneous Multi-Relational Recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021.

[58] X. Chen, H.-j. Zhao, S. Zhao, J. Chen and Y.-p. Zhang, "Citation recommendation based on citation tendency," *Scientometrics,* vol. 121, no. 2, pp. 937-956, 2019.

[59] L. Yang, Z. Zhang, X. Cai and L. Guo, "Citation Recommendation as Edge Prediction in Heterogeneous Bibliographic Network: A Network Representation Approach," *IEEE Access,* vol. 7, pp. 23232 - 23239, 2019.

[60] X. Cai, Y. Zheng, L. Yang, T. Dai and L. Guo, "Bibliographic network representation based personalized citation recommendation," *IEEE Access,* vol. 7, pp. 457-467, 2019.

[61] Z. Ali, G. Qi, K. Muhammad, B. Ali and W. A. Abro, "Paper recommendation based on heterogeneous network embedding," *Knowledge-Based Systems,* vol. 210, 2020.

[62] W. Wang, "Venue topic model‐enhanced joint graph modelling for citation recommendation in scholarly big data," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 20.1*, pp. 1-15, 2020.

[63] J. Chen, Y. Liu, S. Zhao and Y. Zhang, "Citation recommendation based on weighted heterogeneous information network containing semantic linking," in *2019 IEEE international conference on multimedia and expo*, 2020.

[64] L. Guo, X. Cai, H. Qin and F. Heo, "A content-sensitive citation representation approach for citation recommendation," *Journal of Ambient Intelligence and Humanized Computing,* vol. 13, no. 12, 2022.

[65] B. Perozzi, R. Al-Rfou and S. Skiena, "DeepWalk: online learning of social representations," in *KDD '14: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2014.

[66] C. Pornprasit, X. Liu, P. Kiattipadungkul, N. Kertkeidkachorn, K.-S. Kim, T. Noraset, S.-U. Hassan and S. Tuarob, "Enhancing citation recommendation using citation network embedding," *Scientometrics*, vol. 127, no. 1, pp. 233-264, 2022.

[67] S. Ganguly and V. Pudi, "Paper2vec: Combining Graph and Text Information for Scientific Paper Representation," in *European Conference on Information Retrieval (ECIR): Advances in Information Retrieval*, 2020.

[68] D. Jacob, C. Ming-Wei, L. Kenton and T. Kristina, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2019)*, 2019.

[69] J. Wang, L. Zhu, T. Dai and Y. Wang, "Deep memory network with Bi-LSTM for personalized context-aware citation recommendation," *Neurocomputing,* vol. 410, no. 0925-2312, pp. 103-113, 2020.

[70] X. Cai, N. Wang, L. Yang and X. Mei, "Global-local neighborhood based network representation for citation recommendation," *Applied Intelligence,* vol. 52, no. 9, p. 10098–10115, 2022.

[71] M. Färber, V. Zinecker, I. B. Cartus, S. Celis and M. Duma, "C-Rex: A Comprehensive System for Recommending In-Text Citations with Explanations," in *Companion Proceedings of the Web Conference*, 2021.

[72] Y. Zhang and Q. Ma, "Recommending Multiple Positive Citations for Manuscript via Content-Dependent Modeling and Multi-Positive Triplet," in *WI-IAT'21: IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2021.

[73] Y. Zhang and Q. Ma, "MP-BERT4REC: Recommending Multiple Positive Citations for Academic Manuscripts via Content-Dependent BERT and Multi-

Positive Triplet," *IEICE Transactions on Information and Systems*, vol. E105.D, 2022.

[74] Z. Ali, G. Qi, K. Muhammad, P. Kefalas and S. Khusro, "Global citation recommendation employing generative adversarial network," *Expert Systems with Applications*, vol. 180, 2021.

[75] L. Wang, Y. Rao, Q. Bian and S. Wang, "Content-Based Hybrid Deep Neural Network Citation Recommendation Method," in *International Conference of Pioneering Computer Scientists, Engineers and Educators (ICPCSEE 2020)*, 2020.

[76] P. Yadav, N. Remala and N. Pervin, "RecCite: A Hybrid Approach to Recommend Potential Papers," in *2019 IEEE International Conference on Big Data (Big Data)*, 2019.

[77] Y. Xie, S. Wang, W. Pan, H. Tang and Y. Sun, "Embedding Based Personalized New Paper Recommendation," in *CCF Conference on Computer Supported Cooperative Work and Social Computing*, 2021.

[78] J. C. Chang, A. X. Zhang, J. Bragg, A. Head, K. Lo, D. Downey and D. S. Weld, "CiteSee: Augmenting Citations in Scientific Papers with Persistent and Personalized Historical Context," in *CHI '23: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023.

[79] W. Li, C. Chang, C. He, Z. Wu, J. Gou and B. Peng, "Academic Paper Recommendation Method Combining Heterogeneous Network and Temporal Attributes," in *CCF Conference on Computer Supported Cooperative Work and Social Computing*, 2021.

[80] Q. He, J. Pei, D. Kifer, P. Mitra and L. Giles, "Context-aware Citation Recommendation," in *The 19th International Conference on World Wide Web 2010*, 2010.

[81] Y. Liang and L. K. Lee, "A Systematic Review of Citation Recommendation Over the Past Two Decades," *International Journal on Semantic Web and Information Systems* , vol. 19, no. 1, pp. 1-22, 2023.

[82] R. Collobert and J. Weston, "A Unified Architecture for Natural Language Processing: Deep Neural Networks with Multitask Learning," in *The 25th International Conference on Machine Learning*, 2008.

[83] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk and Y. Bengio, "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[84] D. Bahdanau, K. Cho and Y. Bengio, "Neural Machine Translation by Jointly Learning to Align and Translate," in *Proceedings of the 3rd International Conference on Learning Representations 2015 (ICLR 2015)*, 2015.

[85] Y. Zhang and Q. Ma, "Dual Attention Model for Citation Recommendation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.

[86] L. Yang, Z. Zang, X. Cai and T. Dai, "Attention-Based Personalized Encoder-Decoder Model for Local Citation Recommendation," *Computational Intelligence and Neuroscience*, vol. 2019, p. 1232581:1–1232581:7, 2019.

[87] C. Bhagavatula, S. Feldman, R. Power and W. Ammar, "Content-based citation recommendation," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 2018.

[88] S. Gerschgorin, "Über die Abgrenzung der Eigenwerte einer Matrix," *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na,* vol. 6, pp. 749-754, 1931.

[89] T. Mikolov and G. Zweig, "Context dependent recurrent neural network language model," in *2012 IEEE Spoken Language Technology Workshop (SLT)*, 2012.

[90] M. Hutter, "The human knowledge compression contest," http://prize.hutter1.net/, 2012.

[91] M. Neumann, D. King, I. Beltagy and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing," *arXiv:1902.07669.*, 2019.

[92] A. Cohan, S. Feldman, I. Beltagy, D. Downey and D. S. Weld, "SPECTER: Document-level Representation Learning using Citation-informed Transformers," in *ACL*, 2020.

[93] L. Yang, Y. Zheng, X. Cai, H. Dai, D. Mu, L. Guo and T. Dai, "A LSTM based model for personalized context-aware citation recommendation," *IEEE Access*, vol. 6, p. 59618–59627, 2018.

[94] A. Brack, A. Hoppe and R. Ewerth, "Citation Recommendation for Research Papers via Knowledge Graphs," in *Proceedings of Linking Theory and Practice of Digital Libraries: 25th International Conference on Theory and Practice of Digital Libraries*, 2021.

[95] T. N. Kipf and M. Welling, "Variational graph auto-encoders," in *NIPS Workshop on Bayesian Deep Learning*, 2016.

[96] J. R. Hershey and P. A. Olsen, "Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 2007.

[97] D. Kingma and J. B. Adam, "Adam: A Method for Stochastic Optimization," *arXiv e-prints*, 2014.

[98] D. Radev, P. Muthukrishnan, V. Qazvinian and A. Abu-Jbara, "The ACL anthology network corpus," *Language Resources and Evaluation,* vol. 47, no. 4, pp. 919-944, 2013.

[99] D. Kang, W. Ammar, B. Dalvi, M. v. Zuylen, S. Kohlmeier, E. Hovy and R. Schwartz, "A dataset of peer reviews(peerread): Collection, insights and NLP applications," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 2018.

[100] M. Ohagi and A. Aizawa, "Pre-trained Transformer-Based Citation Context-Aware Citation Network Embeddings," in *2022 ACM/IEEE Joint Conference on Digital Libraries (JCDL),* 2022.

[101] M. J. Yin, B. Wang and C. Ling, "A fast local citation recommendation algorithm scalable to multi-topics," *Expert Systems With Applications,* vol. 238, 2024.

[102] W. W. Lo, S. Layeghy, M. Sarhan, M. Gallagher , M. Portmann, "A Graph Neural Network based Intrusion Detection System for IoT," *IEEE/IFIP Network Operations and Management Symposium,* 2022.

[103] Z. Huang, X. Xu, H. Zhu and M. Zhou, "An Efficient Group Recommendation Model With Multiattention-Based Neural Networks," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 31, no. 11, pp. 4461-4474, 2020.