Ứng Dụng Đặc Trưng Đa Phương Thức Trong Hệ Khuyến Nghị Sách Tiếng Việt Trong Thương Mại Điện Tử

Lê Xuân Bình[1,2] **Thái Minh Lâm**[1,2] **Mã Kim Phát**[1,2]
[1]Khoa Khoa học và Kỹ thuật Thông tin, Trường Đại học Công nghệ Thông tin,
Thành phố Hồ Chí Minh, Việt Nam
[2]Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam
{22520131, 22520745, 22521071}@gm.uit.edu.vn

Tóm tắt nội dung

Sự phát triển nhanh chóng của thương mại điện tử đã định hình lại hành vi mua sách, trong đó sở thích của người dùng ngày càng bị ảnh hưởng bởi các yếu tố linh hoạt khác nhau. Trong công trình này, chúng tôi trình bày một nghiên cứu thực nghiệm về đề xuất sách đa phương thức cho thị trường sách online Việt Nam. Đầu tiên, chúng tôi xây dựng một bộ dữ liệu sách Việt Nam bằng cách thu thập dữ liệu từ một nền tảng thương mại điện tử lớn trong nước và tích hợp dữ liệu tương tác người dùng-sản phẩm với các tính năng đa phương thức phong phú. Thứ hai, chúng tôi tiến hành các thí nghiệm sâu rộng để đánh giá và so sánh các mô hình đề xuất đa phương thức hiện đại nhất với các mô hình đề xuất truyền thống. Kết quả thí nghiệm cho thấy các mô hình đa phương thức vượt trội hơn đáng kể so với các phương pháp truyền thống. Cuối cùng, chúng tôi triển khai mô hình hoạt động tốt nhất trong kiến trúc triển khai hướng đến Dữ liệu lớn để đề xuất theo thời gian thực. Dữ liệu và mã nguồn được lưu ở đây.

1 Giới thiệu

Sự bùng nổ của thương mại điện tử đã thay đổi căn bản hành vi mua sách: người dùng ngày nay ra quyết định dựa trên nhiều yếu tố như hình ảnh bìa, tóm tắt nội dung và đánh giá trực tuyến (Zhang et al., 2019). Mặc dù các hệ thống khuyến nghị truyền thống hoạt động hiệu quả trên dữ liệu tương tác, chúng thường gặp hạn chế với dữ liệu thưa và chưa khai thác triệt để thông tin ngữ nghĩa từ dữ liệu phi cấu trúc (Su and Khoshgoftaar, 2009). Các nghiên cứu hiện có về đa phương thức thường chỉ xem hình ảnh hay văn bản là đặc trưng bổ sung đơn lẻ, thiếu các đánh giá sâu sắc về hiệu quả thực sự của việc kết hợp các phương thức này (Deldjoo et al., 2020). Đặc biệt, sự thiếu hụt một bộ dữ liệu chuẩn hóa chứa đầy đủ thông tin văn bản và hình ảnh cho sách tiếng Việt đang là rào cản lớn cho nghiên cứu trong nước.

Nghiên cứu này tập trung vào việc giải quyết các thách thức trên thông qua ba đóng góp chính:

- Xây dựng bộ dữ liệu: Chúng tôi giới thiệu bộ dữ liệu sách tiếng Việt tích hợp đầy đủ đặc trưng đa phương thức.
- Đánh giá thực nghiệm: Chúng tôi triển khai và so sánh các mô hình SOTA để làm rõ tác động của việc tích hợp đặc trưng đa phương thức so với các phương pháp đơn phương thức.
- Triển khai hệ thống: Chúng tôi xây dựng một hệ thống khuyến nghị hoàn chỉnh, tối ưu hóa cho khả năng mở rộng và tích hợp công nghệ dữ liệu lớn, phù hợp với yêu cầu thực tế của thương mại điện tử.

2 Các công trình liên quan

Về hệ khuyến nghị sách: Các bộ dữ liệu chuẩn trong lĩnh vực khuyến nghị sách đã có sự phát triển từ bộ dữ liệu thưa Book-Crossing (Ziegler et al., 2005) đến các tài nguyên quy mô lớn như Amazon Books (McAuley et al., 2015) và Goodreads (Wan and McAuley, 2018). Mặc dù các bộ dữ liệu này cung cấp lịch sử tương tác người dùng phong phú, chúng chủ yếu tập trung vào các siêu dữ liệu có cấu trúc. Các nghiên cứu về khuyến nghị trên sách cũng được triển khai trong các công bố (Devika et al., 2021; Mathew et al., 2016; Rajpurkar et al., 2015; Kurmashov et al., 2015) với cách tiếp cận Lọc cộng tác, Dựa trên nội dung và các phương pháp lai giữa các hướng tiếp cận này. Xu hướng chuyển dịch gần đây hướng tới tích hợp đa phương thức được thể hiện rõ trong nghiên cứu của (Spillo et al., 2025), qua việc bổ sung các tín hiệu văn bản và hình ảnh cho bộ dữ liệu DBbook. Tuy nhiên, các tài nguyên này hầu như chỉ dành cho tiếng Anh. Cho đến nay, vẫn chưa có một bộ dữ liệu đa phương thức chuẩn hóa, quy mô lớn nào được xây dựng riêng cho thị trường sách Việt Nam.

Các hướng tiếp cận khuyến nghị cốt lõi: Các nghiên cứu hiện hành phân loại chiến lược khuyến nghị thành ba trụ cột chính: (1) Lọc cộng tác (Collaborative Filtering - CF) và các biến thể nơ-ron (NCF), vốn mô hình hóa các tương tác tiềm ẩn giữa người dùng và mục tiêu (Su and Khoshgoftaar, 2009; He et al., 2017); (2) Lọc dựa trên nội dung (Content-Based Filtering - CBF), tận dụng sự tương đồng của siêu dữ liệu (Lops et al., 2010); và (3) Khuyến nghị dựa trên tri thức (Knowledge-aware Recommendation), sử dụng Đồ thị tri thức nhằm giảm thiểu sự thưa thớt dữ liệu (Guo et al., 2020). Dù có hiệu quả mạnh mẽ, các mô hình này thường coi dữ liệu phi cấu trúc chỉ là thông tin phụ trợ.

Kết hợp đa phương thức và Hiệu quả theo miền: Các hệ thống khuyến nghị đa phương thức (MRS) thường sử dụng các đặc trưng văn bản và hình ảnh thô để bổ trợ cho ma trận tương tác (He and McAuley, 2015; Deldjoo et al., 2020). Tuy nhiên, hiện có một lỗ hổng quan trọng trong việc đánh giá tầm quan trọng của từng phương thức đối với các miền cụ thể (Liu et al., 2024). Nhiều mô hình hiện nay bỏ qua thực tế rằng khả năng dự đoán của một phương thức (ví dụ: phần tóm tắt sách so với hình ảnh bìa) thay đổi đáng kể tùy thuộc vào bối cảnh. Nghiên cứu của chúng tôi giải quyết vấn đề này bằng cách phân tích hiệu quả cụ thể của các đặc trưng hình ảnh và văn bản tiếng Việt trong một khung khuyến nghị thống nhất.

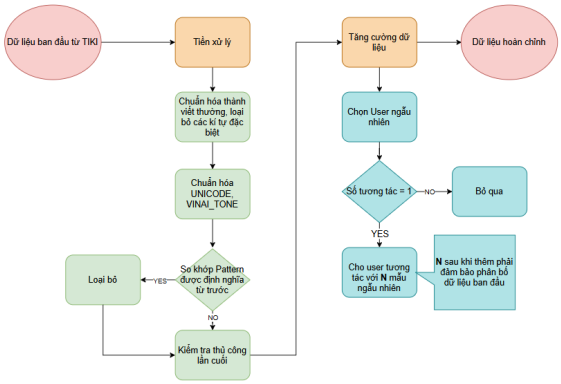
3 Bộ dữ liệu

3.1 Thu thập dữ liệu

Chúng tôi đã xây dựng một bộ dữ liệu mới bằng cách thu thập dữ liệu từ Tiki.vn[1], một nền tảng thương mại điện tử nổi bật về sách tại Việt Nam. Quá trình trích xuất dữ liệu được thực hiện thông qua phương pháp kết hợp: sử dụng BeautifulSoup[2] để truy xuất nhanh các siêu dữ liệu và đánh giá của người dùng, và Selenium[3] để thu thập các phần mô tả văn bản động. Dữ liệu thô bao gồm tiêu đề sách, hình ảnh bìa, mô tả văn bản, giá cả và lịch sử đánh giá của người dùng.

3.2 Tiền xử lý và Tăng cường dữ liệu

Hình 1 tổng quan hóa quy trình tiền xử lý và tăng cường dữ liệu của chúng tôi. Dữ liệu thô trải qua quá trình làm sạch kỹ lưỡng, bao gồm chuẩn hóa Unicode, sửa lỗi chính tả cho văn bản tiếng Việt

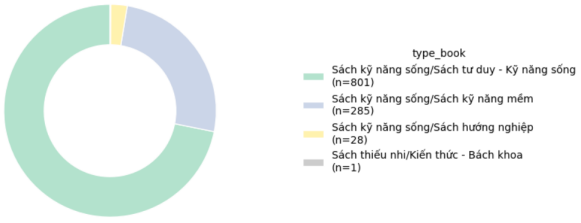


Hình 1: Chiến lượt xử lý và tăng cường dữ liệu

và loại bỏ các nhiễu không liên quan như hashtag, thẻ HTML và biểu tượng cảm xúc (emoji). Để giải quyết vấn đề dữ liệu thưa thớt và duy trì sự đa dạng của dữ liệu trong quá trình lọc k-core trước khi đưa vào mô hình huấn luyện, chúng tôi đã thực hiện tăng cường dữ liệu: đối với các mục có ít hơn 5 lượt đánh giá, chúng tôi tạo ra các đánh giá tổng hợp và gán chúng cho các mã định danh khách hàng (customer ID) giả định với điều kiện số lượt tương tác ban đầu bằng của các khách hàng này là 1 và số lượng mẫu gán không gây ra thay đổi lớn trên phân phối dữ liệu ban đầu. Chiến lược này giúp ngăn chặn việc mất đi các item ít được đánh giá và duy trì mật độ cấu trúc của ma trận tương tác.

3.3 Phân tích dữ liệu và Thống kê

Bộ dữ liệu sau khi làm sạch bao gồm 1.115 cuốn sách với bốn thể loại chính (được thống kê chi tiết tại Hình 2) và 26.050 lượt đánh giá. Các thuộc tính chi tiết được cung cấp trong phần Phụ lục A.

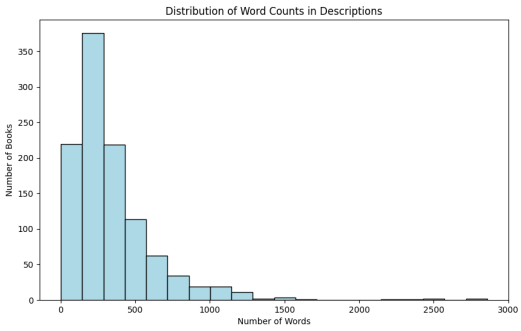


Hình 2: Thống kê thể loại sách

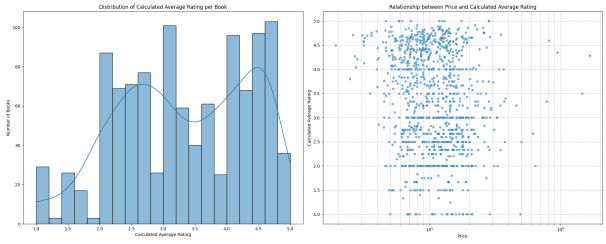
Chúng tôi đã thực hiện phân tích để hiểu rõ các phân phối cơ bản của dữ liệu (Hình 4).

Phân phối của các xếp hạng trung bình cho thấy một sự lệch dương đáng kể, với sự tập trung cao của các mức điểm từ 4.0 đến 5.0. Điều này cho thấy tâm thế ủng hộ chung của khách hàng, đòi hỏi chiến lược khuyến nghị phải có khả năng phân biệt hiệu quả giữa các mục đều có xếp hạng cao.

[1] TIKI
[2] beautifulsoup4
[3] Selenium



Hình 3: Phân phối độ dài tóm tắt sách.



Hình 4: Phân phối đánh giá trung bình (Trái) và Mối quan hệ giữa giá cả và đánh giá (Phải).

Ngoài ra, phân tích mối quan hệ giữa giá sách (theo thang log) và xếp hạng trung bình cho thấy không có sự tương quan rõ rệt. Kết quả này gợi ý rằng giá cả không phải là yếu tố quyết định chính đến sự hài lòng của khách hàng trong bộ dữ liệu này, từ đó củng cố việc chúng tôi tập trung vào nội dung, ảnh bìa và sở thích người dùng thay vì các yếu tố về giá cả.

Đối với nội dung văn bản thể hiện ở phần tóm tắt mô tả sách, chúng tôi trực quan biểu đồ histogram thể hiện phân phối độ dài nội dung (thể hiện qua Hình 3). Phân bố độ dài mô tả sách cho thấy sự đa dạng lớn về số lượng từ. Trung bình các mô tả có độ dài khoảng từ 100-400 từ, với đỉnh phân bố rơi vào khoảng 150-300 từ. Một số mô tả ngoại lệ vượt quá 1000 từ.

4 Thiếp lập thực nghiệm

Chúng tôi áp dụng tiêu chí lọc như sau để đảm bảo mô hình hoạt động tốt:

- Loại bỏ bất kỳ bản ghi sách nào thiếu hình ảnh bìa hoặc mô tả văn bản.
- Lọc Core-k: Áp dụng bộ lọc $k = 3$ để đảm bảo mỗi người dùng và mục tiêu còn lại đều có ít nhất ba tương tác liên quan.

Cách tiếp cận phân chia dữ liệu huấn luyện - đánh giá có sự khác biệt tùy thuộc vào phương pháp thực hiện. Cụ thể, đối với hướng tiếp cận lọc cộng tác (Collaborative Filtering) và khuyến nghị dựa trên nội dung (Content-based Filtering), chúng tôi áp dụng chiến lược Leave-one-out. Theo đó, toàn bộ lịch sử tương tác giữa người dùng và sản phẩm được sắp xếp theo thời gian; sản phẩm cuối cùng trong danh sách tương tác của mỗi người dùng sẽ được đưa vào tập đánh giá (test set), trong khi tất cả các tương tác trước đó được sử dụng làm tập huấn luyện (training set).

Đối với hướng tiếp cận sử dụng các đặc trưng đa phương thức, để tối ưu hóa quá trình huấn luyện và tinh chỉnh siêu tham số cho các mô hình học sâu phức tạp, chúng tôi phân chia bộ dữ liệu thành ba phần riêng biệt: huấn luyện (training set), phát triển (dev set) và kiểm thử (test set). Cụ thể, chúng tôi mở rộng chiến lược Leave-one-out bằng cách trích xuất hai sản phẩm cuối cùng của mỗi người dùng: sản phẩm cuối cùng được đưa vào tập kiểm thử, sản phẩm ngay trước đó được sử dụng làm tập phát triển để phục vụ việc lựa chọn mô hình tối ưu, và toàn bộ các sản phẩm còn lại cấu thành nên tập huấn luyện. Phương pháp này đảm bảo rằng mô hình không chỉ được đánh giá trên dữ liệu chưa biết mà còn được hiệu chỉnh một cách khách quan trước khi đưa ra kết quả cuối cùng. Chúng tôi sử dụng các mô hình tiền huấn luyện để trích xuất đặc trưng cho từng phương thức trong Bảng 1.

Bảng 1: Encoder cho văn bản & hình ảnh

Text Encoder	ViSoBERT(Nguyen et al., 2023)
Image Encoder	ViT(Dosovitskiy et al., 2021), ResNet-152(He et al., 2015)

Sau đó, chúng tôi áp dụng MMRec (Zhou, 2023) như một khung thực nghiệm nhằm huấn luyện và đánh giá các mô hình khuyến nghị đa phương thức. Quy trình tổng quát gồm: đóng gói dữ liệu, thiết lập cấu hình, huấn luyện mô hình và đánh giá hiệu suất.

5 Thực nghiệm và Đánh giá kết quả

5.1 Độ đo đánh giá

Để đánh giá hiệu năng của các mô hình khuyến nghị, chúng tôi sử dụng bốn độ đo tiêu chuẩn tại các ngưỡng $k \in \{5, 10\}$. Gọi R_u là tập hợp k mục được khuyến nghị hàng đầu cho người dùng u, và T_u là tập hợp các mục dữ liệu thực tế (các mục mà người dùng đã thực sự tương tác).

Precision@k và Recall@k Precision đo lường tỷ lệ các mục được khuyến nghị là có liên quan, trong khi Recall đo lường tỷ lệ các mục có liên quan trong dữ liệu thực tế đã được hệ thống khuyến nghị thành công:

$$Precision@k = \frac{|R_u \cap T_u|}{k} \tag{1}$$

$$Recall@k = \frac{|R_u \cap T_u|}{|T_u|} \tag{2}$$

Mean Average Precision (mAP@k) mAP là độ đo xem xét đến thứ hạng của các mục có liên quan trong danh sách. Trước tiên, chúng tôi tính toán Độ chính xác trung bình (Average Precision - AP) cho một người dùng cụ thể, trong đó $rel(i)$ là một biến chỉ báo nhị phân có giá trị bằng 1 nếu mục tại vị trí thứ i có liên quan và bằng 0 nếu ngược lại:

$$AP@k = \frac{1}{\min(|T_u|, k)} \sum_{i=1}^{k} (Precision@i \times rel(i)) \tag{3}$$

Sau đó, mAP được tính bằng giá trị trung bình của $AP@k$ trên tất cả người dùng U trong tập kiểm thử.

Normalized Discounted Cumulative Gain (NDCG@k) NDCG đánh giá chất lượng xếp hạng bằng cách áp dụng hình phạt đối với các mục liên quan nằm ở vị trí thấp trong danh sách thông qua cơ chế suy giảm logarit. Độ đo DCG (Discounted Cumulative Gain) được định nghĩa như sau:

$$DCG@k = \sum_{i=1}^{k} \frac{2^{rel(i)} - 1}{\log_2(i+1)} \tag{4}$$

Giá trị NDCG thu được bằng cách chuẩn hóa DCG theo IDCG (Ideal DCG) — đây là giá trị DCG tối đa đạt được khi danh sách được xếp hạng một cách hoàn hảo:

$$NDCG@k = \frac{DCG@k}{IDCG@k} \tag{5}$$

5.2 Kết quả thực nghiệm

Bảng 2 và 3 tóm tắt hiệu năng của các mô hình khuyến nghị được đánh giá dựa trên tất cả các độ đo đã nêu.

Tất cả các kiến trúc đa phương thức (VBPR (He and McAuley, 2015), LATTICE(Chiappa et al., 2023), FREEDOM (Zhou and Shen, 2023), MMGCN (Wei et al., 2019), SLMRec (Tao et al., 2022)) đều đạt kết quả vượt trội so với baseline BPR (Rendle et al., 2012) (mô hình chỉ dựa trên tương tác thuần túy). Điều này khẳng định rằng việc tích hợp thông tin hình ảnh và văn bản giúp giảm bớt vấn đề dữ liệu thưa thớt (sparsity) và cải thiện độ chính xác của hệ thống gợi ý. Trong số các mô hình được thử nghiệm, MMGCN đạt hiệu năng cao nhất trên hầu hết các chỉ số. Cụ thể, kiến trúc **MMGCN** với thông tin Text (Type) đạt R@5 = 0.0311 và NDCG@10 = 0.0243, cao gấp khoảng gần 3 lần so với mô hình BPR. Kết quả này cho thấy khả năng của mạng đồ thị (GNN) trong việc lan truyền thông tin đa phương thức qua cấu trúc người dùng.

Dựa trên kết quả thực nghiệm, đặc biệt là chỉ số **R@10**, cho thấy việc lựa chọn **Encoder** đóng vai trò quyết định đến hiệu suất tổng thể của mô hình. Sự khác biệt rõ rệt trong các thông số đo lường đã khẳng định tầm quan trọng then chốt của thành phần này, cụ thể như sau:

- Đối với **Textual Encoder**, ViSoBERT chứng minh được sự ổn định và hiệu suất vượt trội khi xử lý dữ liệu tiếng Việt. Thực nghiệm cho thấy thông tin Text (Type) (thể loại sản phẩm) thường đem lại kết quả khả quan hơn so với Text (Desc) (mô tả sản phẩm). Điều này có thể lý giải bởi tính cô đọng của dữ liệu thể loại, giúp mô hình tập trung khai thác các đặc trưng phân loại trọng tâm mà không bị nhiễu bởi các chi tiết dư thừa. Đặc biệt, khi tích hợp đa phương thức, sự kết hợp giữa thông tin thể loại và đặc trưng hình ảnh tạo ra hiệu ứng cộng hưởng tích cực, giúp tối ưu hóa hiệu suất tổng thể. Minh chứng rõ nhất là ở mô hình SLMRec, khi đặc trưng thể loại luôn giúp mô hình đạt kết quả vượt trội hơn so với thông tin mô tả, bất kể được kết hợp với loại Visual Encoder nào.

- Đối với **Visual Encoder**, hiệu quả của các đặc trưng hình ảnh phụ thuộc chặt chẽ vào kiến trúc mô hình và cơ chế hòa trộn thông tin (fusion). Trong các kiến trúc khai thác tương tác đa phương thức dựa trên cấu trúc như LATTICE và FREEDOM, ViT chiếm ưu thế nhờ khả năng trích xuất đặc trưng toàn cục và ngữ nghĩa chuyên sâu, từ đó cải thiện đáng kể độ chính xác của kết quả truy vấn. Ngược lại, với các dòng mô hình dựa trên phân tách nhân tố (factorization) hoặc đồ thị (graph-based) như VBPR và MMGCN, ResNet-152 lại tỏ ra vượt trội hơn. Điều này minh chứng rằng đối

với các cấu trúc này, các đặc trưng thu được từ lớp tích chập (CNN) cung cấp độ ổn định và tính biểu diễn cao hơn so với cơ chế Attention của ViT.

Để đánh giá một cách khách quan hiệu quả của các mô hình khuyến nghị đa phương thức hiện đại, chúng tôi tiến hành so sánh với các nhóm phương pháp nền (baseline) truyền thống, bao gồm lọc cộng tác dựa trên người dùng (User-Based Collaborative Filtering) và khuyến nghị dựa trên nội dung (Description-Based và Typebook-Based). Nhìn chung, các phương pháp truyền thống cho hiệu suất khá thấp: User-based (Cosine) chỉ đạt Recall@5 = 0.0057 và NDCG@10 = 0.0065; trong khi Typebook-based (Cosine) nhỉnh hơn nhẹ với Recall@5 = 0.0141 và NDCG@10 = 0.0088. Trái lại, mô hình đa phương thức MMGCN (Text-Type) đạt Recall@5 = 0.0311 và NDCG@10 = 0.0243, cho thấy lợi thế rõ rệt khi khai thác đồng thời tín hiệu tương tác và đặc trưng nội dung. So với baseline tốt nhất trong nhóm truyền thống, mô hình đề xuất cải thiện khoảng ~ 2.2 lần về Recall@5 và ~ 2.8 lần về NDCG@10; đồng thời nếu so với User-based, mức cải thiện lần lượt là ~ 5.5 lần (Recall@5) và ~ 3.7 lần (NDCG@10).

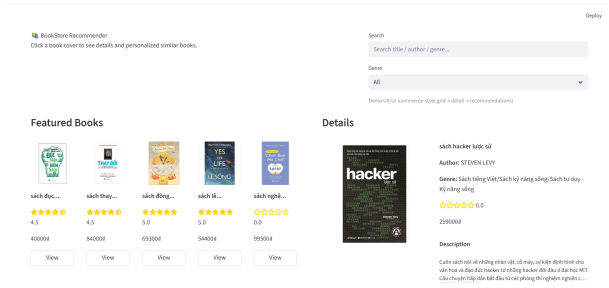
Cuối cùng, việc kết hợp đồng thời cả văn bản và hình ảnh (Text + Image) không phải lúc nào cũng cho kết quả cao nhất ở mọi chỉ số so với việc dùng đơn lẻ một phương thức mạnh (như Description Type).

6 Triển khai hệ thống

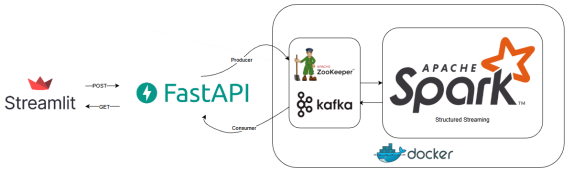
Sau khi thực nghiệm, chúng tôi chọn ra mô hình có kết quả tốt nhất để xây dựng hệ thống khuyến nghị.

Hệ thống được triển khai giả định trên nền tảng Dữ Liệu Lớn được minh hoạ trong Hình 5 và 6. Kiến trúc sử dụng Streamlit cho giao diện dùng, giao tiếp thông qua FastAPI[4] để xử lý các yêu cầu không bộ. Luồng dữ liệu được điều phối thông qua Apache Kafka[5], đảm bảo khả năng chịu lỗi và băng thông cao. Công nghệ xử lý cốt lõi là Apache Spark[6] (Structured Streaming), tích hợp mô hình đề đa phương thức được đào tạo từ trước để thực hiện suy luận theo thời gian thực. Đặc biệt, để dễ dàng triển khai các môi trường và nền tảng khác nhau chúng tôi đóng gói cả hệ thống bằng Docker[7] để đảm bảo triển khai liền mạch và tính nhất quán.

[4] FastAPI
[5] kafka
[6] Spark
[7] Docker



Hình 5: Giao diện hệ thống ứng dụng khuyến nghị sách



Hình 6: Kiến trúc pipeline xử lý dữ liệu và tích hợp mô hình khuyến nghị thời gian thực.

Giao diện của hệ thống của chúng tôi được minh họa trong Hình 5.

7 Hạn chế

Số lượng mẫu sách vẫn còn khá hạn chế so với ứng dụng thực tiễn, điều này dẫn đến ma trận user–item bị thưa, khiến các phương pháp học dựa trên tương tác gặp khó khăn trong việc học được biểu diễn ổn định và tổng quát. Trong bối cảnh dữ liệu thưa, mô hình dễ bị lệch về một nhóm nhỏ người dùng hoặc các sản phẩm phổ biến, làm suy giảm khả năng gợi ý chính xác cho các người dùng ít tương tác và các sách thuộc nhóm long–tail.

Bên cạnh đó, chất lượng và mức độ đầy đủ của dữ liệu nội dung vẫn chưa đồng đều. Mô tả sách (description) có thể bị ngắn, thiếu thông tin hoặc mang tính quảng cáo. Điều này làm giảm hiệu quả của các tín hiệu dựa trên nội dung mô tả và có thể gây nhiễu khi kết hợp đa phương thức.

Ngoài ra, dữ liệu đa phương thức hiện tại chủ yếu tập trung vào ảnh bìa và văn bản mô tả, trong khi nhiều tín hiệu quan trọng khác (ví dụ: nhà xuất bản, năm xuất bản, ngôn ngữ, chủ đề chi tiết, trích đoạn nội dung, hoặc thông tin tác giả có cấu trúc) chưa được khai thác. Việc thiếu các thuộc tính giàu ngữ nghĩa này có thể giới hạn khả năng mô hình phân biệt các cuốn sách có bìa/mô tả tương tự nhưng nội dung khác nhau.

Bảng 2: Kết quả thực nghiệm của các phương pháp tiếp cận truyền thống.

	Pre@5	**Rec@5**	**mAP@5**	**NDCG@5**	**Pre@10**	**Rec@10**	**mAP@10**	**NDCG@10**
User-based (Cosine)	0.0011	0.0057	0.0027	0.0034	0.0016	0.0156	0.0039	0.0065
User-based (Pearson)	0.0006	0.0028	0.0008	0.0013	0.0004	0.0042	0.0011	0.0018
Content-based (Cosine)	0.0011	0.0057	0.0026	0.0034	0.0017	0.0170	0.0041	0.0070
Content-based (Pearson)	0.0011	0.0057	0.0026	0.0034	0.0017	0.0170	0.0041	0.0070
Typebook-based (Cosine)	**0.0028**	0.0141	**0.0052**	**0.0074**	**0.0018**	**0.0183**	**0.0058**	**0.0088**
Typebook-based (Pearson)	0.0019	**0.0099**	0.0022	0.0039	0.0014	0.0141	0.0027	0.0053

Bảng 3: Kết quả thực nghiệm của các mô hình đa phương thức dựa trên các Encoder khác nhau.

Model	**Modality**	**Encoder**	**P@5**	**R@5**	**mAP@5**	**N@5**	**P@10**	**R@10**	**mAP@10**	**N@10**
BPR	-	-	0.0023	0.0113	0.0042	0.006	0.0021	0.0212	0.0054	0.0091
VBPR	Text (Desc)	ViSoBERT	0.0011	0.0057	0.0025	0.0033	0.0013	0.0127	0.0034	0.0055
	Text (Type)	ViSoBERT	0.0011	0.005	0.0028	0.0035	0.0014	0.0141	0.0039	0.0062
	Image	ViT (CLS)	0.0006	0.0028	0.001	0.0014	0.0006	0.0057	0.0014	0.0024
	Image	ResNet-152	0.0025	0.0127	0.0078	0.009	0.002	0.0198	0.0085	0.0111
	Text (Type) + Image	ViSoBERT + ResNet-152	0.0025	0.0127	0.0066	0.0081	0.0016	0.0156	0.0071	0.0091
LATTICE	Text (Desc)	ViSoBERT	0.0028	0.0141	0.0073	0.009	0.0033	0.0325	0.0097	0.0148
	Text (Type)	VisoBERT	0.0011	0.0057	0.005	0.0051	0.0024	0.024	0.0075	0.0112
	Image	ViT (CLS)	0.004	0.0198	0.0081	0.011	0.0031	0.0311	0.0097	0.0147
	Image	ResNet-152	0.002	0.0099	0.0023	0.0041	0.0018	0.0184	0.0035	0.0069
	Text (Type) + Image	ViSoBERT + ViT (CLS)	0.0042	0.0212	0.0095	0.0124	0.0033	0.0325	0.0111	0.0161
FREEDOM	Text (Desc)	ViSoBERT	0.0037	0.0184	**0.013**	0.0143	0.0027	0.0269	0.014	0.0169
	Text (Type)	VisoBERT	0.002	0.0099	0.0033	0.0049	0.0021	0.0212	0.0047	0.0084
	Image	ViT (CLS)	0.0023	0.0113	0.0045	0.0061	0.0025	0.0255	0.0064	0.0107
	Image	ResNet-152	0.0006	0.0028	0.0012	0.0016	0.0014	0.0141	0.0027	0.0053
	Text (Desc) + Image	ViSoBERT + ViT (CLS)	0.004	0.0198	0.0117	0.0137	0.0028	0.0283	0.0128	0.0165
MMGCN	Text (Desc)	ViSoBERT	0.0023	0.0113	0.0044	0.0061	0.0051	0.0509	0.0101	0.0193
	Text (Type)	VisoBERT	0.0023	0.0113	0.0041	0.0059	0.0048	0.0481	0.0089	0.0177
	Image	ViT (CLS)	**0.0062**	**0.0311**	0.0108	0.0157	0.0047	0.0467	0.0126	0.0204
	Image	ResNet-152	0.0054	0.0269	0.0101	0.0143	0.0052	0.0523	0.013	0.022
	Text (Desc) + Image	ViSoBERT + ResNet-152	**0.0062**	**0.0311**	0.0112	**0.0161**	**0.0057**	**0.0566**	**0.0145**	**0.0243**
SLMRec	Text (Desc) + Image	ViSoBERT + ViT (CLS)	0.0017	0.0085	0.0027	0.0041	0.003	0.0297	0.0051	0.0106
	Text (Type) + Image	ViSoBERT + ViT (CLS)	0.0048	0.024	0.0065	0.0107	0.0041	0.041	0.0086	0.016
	Text (Desc) + Image	ViSoBERT + ResNet-152	0.0014	0.0071	0.0024	0.0035	0.0033	0.0325	0.0054	0.0114
	Text (Type) + Image	ViSoBERT + ResNet-152	0.0042	0.0212	0.0075	0.0109	0.0035	0.0354	0.0095	0.0156

8 Kết luận

Mô hình khuyến nghị đa phương thức giải quyết được vấn đề dữ liệu bị thưa, là khuyết điểm của các mô hình khuyến nghị truyền thống như lọc cộng tác và khuyến nghị dựa trên nội dung. Mặc dù chúng tôi đã sử dụng embedding tiên tiến - VisoBERT cho phương pháp truyền thống là Content-based và Collaborative Filtering, tuy nhiên các backbone của các mô hình đa phương thức tận dụng tốt hơn dẫn đến việc cho ra kết quả vượt trội. Bên cạnh đó, các phương pháp truyền thống chỉ tính toán tương đồng cục bộ giữa hai thực thể. Ngược lại, kiến trúc dựa trên đồ thị như MMGCN, LATTICE cho phép thông tin từ một người dùng hoặc sản phẩm lan truyền qua nhiều lớp, giúp tìm ra những mối liên hệ ẩn mà các phương pháp truyền thống không thể phát hiện.

Chúng tôi cũng thành công trong việc triển khai mô hình khuyến nghị tiên tiến trên bối cảnh dữ liệu lớn, tận dụng các công nghệ hiện đại nhằm có thể là luồng triển khai tham khảo khi thực hiện trên bối cảnh dữ liệu lớn thực tế.

Cuối cùng, điều cần nhấn mạnh là hiệu năng của các mô hình khuyến nghị đa phương thức phần lớn dựa vào các Encoder mà chúng ta lựa chọn. Cho nên, tùy vào mục đích khuyến nghị chúng ta có thể sử dụng các Encoder phù hợp như huấn luyện sẵn hoặc fine-tune theo nhu cầu của bài toán.

9 Lời cảm ơn

Nghiên cứu này được thực hiện cho Đồ án môn học Hệ khuyến nghị - DS300 của Trường Đại học Công nghệ Thông tin - ĐHQG TPHCM dưới sự giảng dạy và hướng dẫn của thầy ThS. Huỳnh Văn Tín.

10 Tài liệu tham khảo

References

Alberto Silvio Chiappa, Alessandro Marin Vargas, Ann Zixiang Huang, and Alexander Mathis. 2023. Latent exploration for reinforcement learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yashar Deldjoo, Markus Schedl, Paolo Cremonesi, and Gabriella Pasi. 2020. Recommender systems leveraging multimedia content. 53(5).

PV Devika, K Jyothisree, PV Rahul, S Arjun, and Jayasree Narayanan. 2021. Book recommendation system. In *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering*, 34(8):3549–3568.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *Preprint*, arXiv:1512.03385.

Ruining He and Julian McAuley. 2015. Vbpr: Visual bayesian personalized ranking from implicit feedback. *Preprint*, arXiv:1510.01784.

Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *Proceedings of the 26th international conference on world wide web*, pages 173–182.

Nursultan Kurmashov, Konstantin Latuta, and Abay Nussipbekov. 2015. Online book recommendation system. In *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, pages 1–4. IEEE.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. 2024. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 57(2):1–17.

Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro. 2010. Content-based recommender systems: State of the art and trends. pages 73–105. Springer.

Praveena Mathew, Bincy Kuriakose, and Vinayak Hegde. 2016. Book recommendation system through content based and collaborative filtering method. In *2016 International conference on data mining and advanced computing (SAPIENCE)*, pages 47–52. IEEE.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 43–52.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *Preprint*, arXiv:2310.11166.

Sushama Rajpurkar, Darshana Bhatt, Pooja Malhotra, MSS Rajpurkar, and MDR Bhatt. 2015. Book recommendation system. *International Journal for Innovative Research in Science & Technology*, 1(11):314–316.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2012. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*.

Giuseppe Spillo, Elio Musacchio, Cataldo Musto, Marco de Gemmis, Pasquale Lops, and Giovanni Semeraro. 2025. See the movie, hear the song, read the book: Extending movielens-1m, last. fm-2k, and dbbook with multimodal data. In *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, pages 847–856.

Xiaoyuan Su and Taghi M. Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Adv. Artif. Intell.*, 2009:421425:1–421425:19.

Zhulin Tao, Xiaohao Liu, Yewei Xia, Xiang Wang, Lifang Yang, Xianglin Huang, and Tat-Seng Chua. 2022. Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, 25:5107–5116.

Mengting Wan and Julian McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM conference on recommender systems*, pages 86–94.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, Richang Hong, and Tat-Seng Chua. 2019. Mmgcn: Multi-modal graph convolution network for personalized recommendation of micro-video. In *Proceedings of the 27th ACM international conference on multimedia*, pages 1437–1445.

Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. 2019. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys*, 52(1):1–38.

Xin Zhou. 2023. Mmrec: Simplifying multimodal recommendation. pages 1–2.

Xin Zhou and Zhiqi Shen. 2023. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM international conference on multimedia*, pages 935–943.

Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th international conference on World Wide Web*, pages 22–32.

A Phụ lục

Phụ lục mô tả các thuộc tính trong bộ dữ liệu của chúng tôi, bao gồm bảng dữ liệu về sách và bảng dữ liệu về tương tác đánh giá của người dùng với sách.

Bảng 4: Bảng thông tin sách

Thuộc tính	Mô tả
product_id	ID của sách
product_name	Tên của sách
authors	Tên tác giả
price	Giá bán của sách
seller_id	ID của người bán sách
seller_type	Phân loại người bán sách (OFFICIAL_STORE, TRUSTED_STORE)
rating_average	Điểm đánh giá trung bình của sách
review_count	Số lượng đánh giá của sách
order_count	Số lượng sách đã bán
url	Liên kết sản phẩm
image	Liên kết ảnh bìa sản phẩm
description	Mô tả nội dung sách
type_book	Thể loại sách
product_index	Cấp phát động tăng dần của ID sách

Bảng 5: Bảng tương tác người dùng với sách

Thuộc tính	Mô tả
customer_id	ID của người dùng
product_id	ID của sách được tương tác
rating	Điểm đánh giá của người dùng dành cho sách
content	Nội dung bình luận của người dùng
customer_index	Cấp phát động tăng dần của ID người dùng
product_index	Cấp phát động tăng dần của ID sách