

HỆ THỐNG GỢI Ý SẢN PHẨM THEO PHIÊN DỰA TRÊN MÔ HÌNH WORD2VEC PRODUCT RECOMMENDATION SYSTEM BY SESSION BASED ON THE WORD2VEC MODEL

NGUYỄN THÁI NGHE¹, ĐẶNG THỊ KIM QUI^{2,a}

¹Khoa Công nghệ thông tin, Trường Đại học Sư phạm kỹ thuật Vĩnh Long.

²Học viên cao học, Trường Đại học Sư phạm Kỹ thuật Vĩnh Long

^aTác giả liên hệ: dangthikimqui87@gmail.com

Nhận bài(Received):01/03/2024;Phản biện(Reviewed):25/03/2024;Chấp nhận(Accepted):08/04/2024

TÓM TẮT

Hệ thống gợi ý (Recommendation Systems - RS) đang được ứng dụng rộng rãi trong nhiều lĩnh vực (như thương mại điện tử, giải trí, giáo dục,...) nhằm dự đoán sở thích của người dùng nhờ vào thông tin cá nhân hoặc những ý kiến phản hồi (đánh giá) của họ. Đề tài này đề xuất một giải pháp trong xây dựng hệ thống gợi ý sản phẩm theo phiên sử dụng mô hình Word2Vec, nhằm gợi ý các sản phẩm cho người mua sắm trên các trang web bán hàng online. Kết quả đạt được của đề tài là Recall@10=27,3% và MRR@10=15,4%. Mặc dù kết quả thực nghiệm của hệ thống chưa cao lắm, nhưng đề tài này có thể làm tiền đề đóng góp vào việc phát triển và cải thiện các hệ thống gợi ý thông minh trong tương lai.

Từ khóa: Hệ thống gợi ý theo phiên, hệ thống gợi ý sản phẩm, Word2vec

ABSTRACT

Recommendation Systems (RS) are being widely applied in many fields (such as e-commerce, entertainment, education,...) to predict user preferences based on personal information, employees or their feedback (evaluation). This research proposes a solution in building a session-based product recommendation system using word2vec model, to suggest products to shoppers on online sales websites. The results achieved by the project are Recall@10=27.3% and MRR@10=15.4%. Although the experimental results of the system are not very high, this topic can serve as a premise to contribute to the development and improvement of intelligent recommendation systems in the future.

Keywords: Session recommendation system, product recommendation system, Word2vec.

1. GIỚI THIỆU

Trong thời đại số hoá, hệ thống gợi ý ngày càng trở nên quan trọng và cần thiết để cung cấp trải nghiệm tốt hơn cho người dùng. Bài toán gợi ý hiện nay cũng đang được các nhà nghiên cứu công nghệ thông tin quan tâm để cải thiện ngày càng tốt hơn những hệ thống đang sử dụng. Hiện nay, ở nước ta cũng có rất nhiều đề tài nghiên

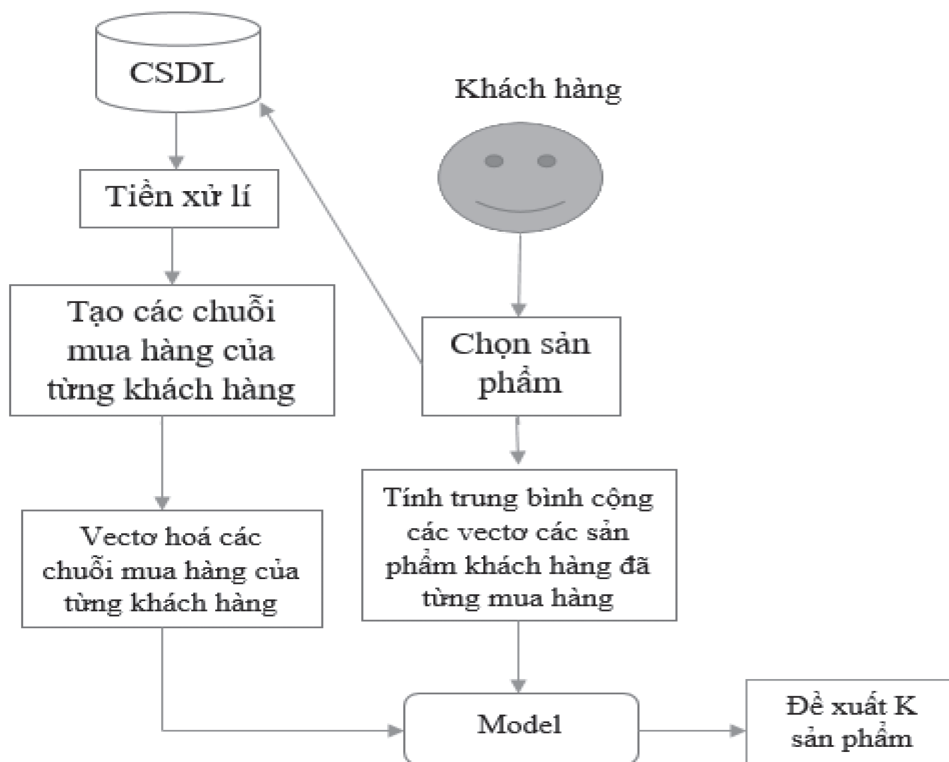
cứu về lĩnh vực gợi ý này bằng các thuật toán lọc cộng tác, lọc nội dung, phân rã ma trận... Trong bài viết này tác giả mong muốn tìm hiểu và giới thiệu một giải pháp mới trong hệ thống gợi ý. Đó là sử dụng mô hình Word2vec để xây dựng hệ thống gợi ý theo phiên. Hy vọng rằng đề tài này là tiền đề quan trọng cho những công trình nghiên cứu tiếp theo về hệ thống gợi ý.

Một số nghiên cứu liên quan:

- Đề tài “Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác” của nhóm tác giả: Nguyễn Hùng Dũng, Nguyễn Thái Nghe thuộc đơn vị Trường Đại học Cần Thơ, công bố năm 2014. Trong đề tài này, tác giả giới thiệu kỹ thuật lọc cộng tác dựa trên mô hình láng giềng (mô hình lân cận) để gợi ý sản phẩm trong hệ thống bán hàng trực tuyến. Kết quả giá trị trung bình RMSE=90%

- Đề tài: “Một giải pháp trong xây dựng Hệ thống gợi ý bài hát” của nhóm tác giả: Nguyễn Tấn Phong, Nguyễn Thái Nghe thuộc trường Đại học Cần Thơ công bố năm 2014. Trong đề tài này, tác giả giới thiệu một giải pháp xây dựng Hệ thống gợi ý bài hát dựa vào phản hồi tiềm ẩn (implicit feedback) từ người dùng để gợi ý những bài hát mà họ có thể thích nghe. Kết quả giá trị trung bình AUC=74,2%

2. XÂY DỰNG HỆ THỐNG GỢI Ý



Hình 1. Sơ đồ tổng quan hệ thống gợi ý

- Hệ thống gợi ý được xây dựng theo trình tự: Cơ sở dữ liệu là lịch sử duyệt web của khách hàng được đưa vào hệ thống để train, tạo ra các chuỗi mua hàng của từng khách hàng. Tiếp theo dùng kỹ thuật Word2vec để tạo ra model. Khách hàng chọn sản phẩm mình yêu thích trên hệ thống bán hàng. Dựa vào lịch sử duyệt website của khách hàng đó mà hệ thống sẽ dự đoán các sản phẩm tiếp theo mà khách hàng muốn mua. Từ đó mà có thể gợi ý một

cách chính xác nhất cho khách hàng.

- Đầu tiên là tạo ra các chuỗi mua hàng do khách hàng thực hiện cho cả 2 tập dữ liệu train và test. Nghĩa là trong tập dữ liệu của chúng ta có 4.372 khách hàng thì hệ thống có 4.372 chuỗi

- Xây dựng bộ Word2vec Embeddings cho các sản phẩm

- Khi khách hàng đã chọn 1 sản phẩm, hệ thống sẽ lấy trung bình cộng của tất cả

các vectơ của các sản phẩm khách hàng đã mua cho đến nay và sử dụng vectơ kết quả này để tìm các sản phẩm tương tự. Để làm được điều đó, hệ thống sử dụng hàm để lấy danh sách ID sản phẩm và đưa ra vectơ 100 chiều, là giá trị trung bình của vectơ của sản phẩm trong danh sách đầu vào. Tiếp theo là hệ thống sẽ so sánh và tính độ tương đồng của các vectơ để trả về 10 sản phẩm có vectơ có độ tương đồng cao nhất để gợi ý cho khách hàng.

3. PHƯƠNG PHÁP ĐÁNH GIÁ

Đề tài sử dụng hai đại lượng Recall@K và MRR@K để đánh giá mô hình

Recall@K (Recall at K) đo lường tỷ lệ các sản phẩm có liên quan được xác định chính xác trong đề xuất K sản phẩm trên tổng số sản phẩm có liên quan trong tập dữ liệu. Nói một cách đơn giản hơn, nó cho biết bạn có thể tìm thấy chính xác bao nhiêu sản phẩm có liên quan.

$$\text{Recall@K} = \frac{\text{Number of relevant items in K}}{\text{Total number of relevant items}}$$

Number of relevant items in K: Số sản phẩm có liên quan trong top K sản phẩm mà hệ thống đã gợi ý.

Total number of relevant items: Tổng số sản phẩm có liên quan có trong tập dữ liệu.

MRR@K (Mean Reciprocal Rank at K). Xếp hạng đối ứng trung bình là lựa chọn tốt hơn cho các ứng dụng trong đó thứ tự của các vấn đề đề xuất (ví dụ: các đề xuất được xếp hạng thấp hơn là chỉ hiển thị khi người dùng cuộn xuống cuối trang). Trong trường hợp này, có MRR cao hơn là rất quan trọng, cho thấy các đề xuất tốt nhất nằm gần đầu.

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$

4. TIẾN HÀNH THỰC HIỆN

4.1 Công nghệ sử dụng

Đề tài được thực nghiệm trên nền Google colab phiên bản free, kết hợp thuật toán Gensim để vectơ hoá mã khách hàng.

4.2 Tập dữ liệu

Ứng dụng mô hình Word2vec xây dựng hệ thống gợi ý dựa trên tập dữ liệu OnlineRetail[12]. Bộ dữ liệu này được thu thập từ ngày 01/12/2010 đến ngày 09/12/2011 và chứa lịch sử mua hàng của 4.372 khách hàng và 3.684 sản phẩm.

4.3. Tiền xử lý dữ liệu

Bảng dữ liệu gồm có 8 cột: InvoiceNo: Số hóa đơn, StockCode: Mã hàng hóa, Descriptip.

Bảng dữ liệu gồm có 8 cột: InvoiceNo: Số hoá đơn, StockCode: Mã hàng hoá, Description: Mô tả hàng hoá, Quantity: Số lượng, InvoiceDate: Ngày lập hoá đơn, UnitPrice: Đơn giá, CustomerID: Mã khách hàng, Country Quốc gia.

Dữ liệu được load lên hệ thống sẽ được kiểm tra các dòng còn thiếu dữ liệu và các dòng đó sẽ được xóa bỏ. Chuyển đổi kiểu dữ liệu cột StockCode sang dạng chuỗi.

Sau khi tiền xử lý, tập dữ liệu được chia làm 2 phần, 95% là dữ liệu train và 5% làm dữ liệu test.

4.4 Cấu hình các tham số

Tiến hành train data với thư viện Gensim, Gensim là một thư viện xử lý ngôn ngữ tự nhiên nguồn mở dùng để vectơ hoá dữ liệu. Hệ thống đã tiến hành thay đổi 4 tham số: min_count, epochs, workers và sg. Mỗi bộ giá trị đề tài thực hiện 5 lần chạy và lấy kết quả trung bình. Bộ tham số Min_count=1; epochs=100; workers=10; sg=1; là cho kết quả cao nhất và thời gian thực hiện là phù hợp nhất. Trong đó:

- **min_count** là số lần tối thiểu mà một từ trong từ vựng phải có để word2vec tạo ra phân nhúng cho nó. Vì trong tập dữ liệu có một số sản phẩm hiếm nên giá trị này được đặt là 1 để tất cả các ID sản phẩm trong tập huấn luyện đều có phân nhúng.

- **epochs** trong đề tài này chúng tôi có thay đổi số kỉ nguyên (epochs) ở các giá trị khác nhau từ 10, 20, 100, 200, 1000. Tuy nhiên chúng tôi nhận thấy số epochs = 200 là cho kết quả cao và có thời gian thực hiện phù hợp nhất.

- **workers** kiểm soát số lượng luồng độc lập thực hiện đào tạo đồng thời. Với cấu hình phần cứng mà google colab hỗ trợ thì tham số workers = 10 là hợp lý nhất.

- **sg** là viết tắt của từ **skip-gram**, vì chúng ta dự đoán mặt hàng tiếp theo mà khách hàng quan tâm là phụ thuộc vào mặt

hàng mà khách hàng vừa chọn nên tham số sg=1 là phù hợp nhất.

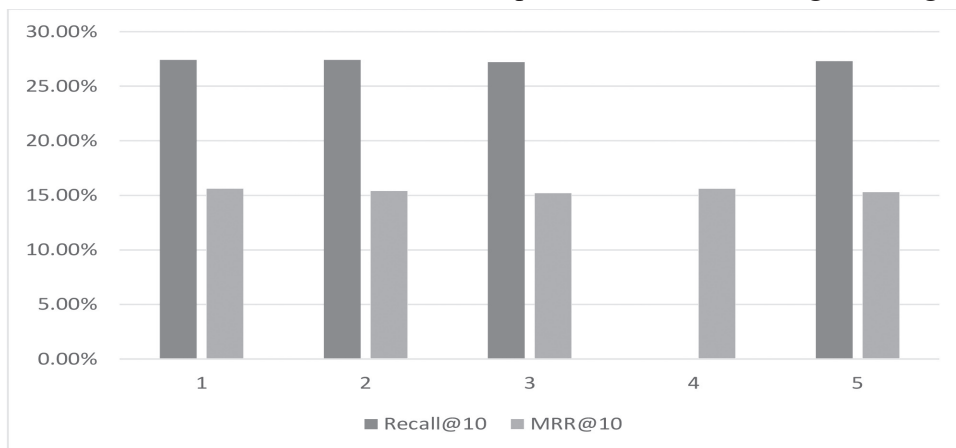
4.5 Các kết quả thực nghiệm

Kết quả thực nghiệm với bộ tham số phần 4.4

Bảng kết quả thực nghiệm

Lần chạy	Recall@10	MRR@10
1	27,4%	15,6%
2	27,4%	15,4%
3	27,2%	15,2%
4	27,5%	15,6%
5	27,3%	15,3%
Trung bình	27,33%	15,42%

Qua bảng số liệu kết quả cho ta thấy tỉ lệ đề xuất chính xác khá cao. Trung bình là trên 27% sản phẩm được đề xuất là có liên quan đến nhu cầu của người dùng.



Hình 2. Biểu đồ hiển thị kết quả qua 5 lần chạy thực nghiệm

Qua biểu đồ 1 cho ta thấy kết quả Recall@10 và MRR@10 của hệ thống tương đối ổn định.

5. KẾT LUẬN

Chúng tôi đã nghiên cứu và xây dựng một hệ thống gợi ý sản phẩm dựa trên phân tích ngữ nghĩa sử dụng Word2Vec. Mục tiêu của đề tài là tạo ra một hệ thống thông minh có khả năng gợi ý các sản phẩm phù

hợp cho người dùng dựa trên sở thích và hành vi trước đó của họ.

Kết quả gợi ý của hệ thống đạt trên 27% cho thấy hệ thống gợi ý sản phẩm dựa trên Word2Vec có hiệu suất tốt. Hệ thống có khả năng gợi ý các sản phẩm phù hợp với người dùng dựa trên sở thích và hành vi trước đó của họ. Đồng thời, việc sử dụng phân tích ngữ nghĩa giúp cải thiện

chất lượng gợi ý, giúp người dùng tìm kiếm và khám phá các sản phẩm một cách hiệu quả hơn.

Khó khăn mà đề tài này chưa giải quyết đó chính là việc gợi ý cho khách hàng các sản phẩm mới. Bởi vì các sản phẩm mới thì chưa có trong bất cứ phiên làm việc nào nên chắc chắn các sản phẩm mới sẽ không có trong các mục đề xuất của hệ thống.

Tổng kết lại, đề tài này đã đạt được mục tiêu xây dựng một hệ thống gợi ý sản phẩm dựa trên phân tích ngữ nghĩa sử

dụng Word2Vec. Kết quả nghiên cứu cho thấy hệ thống có hiệu suất tốt và có tiềm năng ứng dụng rộng rãi trong việc cải thiện trải nghiệm người dùng và tăng cường khả năng tìm kiếm và khám phá sản phẩm.

Mặc dù kết quả của đề tài khá cao. Tuy nhiên, còn nhiều khía cạnh có thể được nghiên cứu và cải tiến trong tương lai. Ví dụ, có thể nghiên cứu các phương pháp khác nhau để cải thiện mô hình ngữ nghĩa, hoặc mở rộng hệ thống để áp dụng cho các lĩnh vực khác nhau.

TÀI LIỆU THAM KHẢO

- [1] Nguyễn Hùng Dũng, Nguyễn Thái Nghe. 2014. Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác. Tạp chí Khoa học Trường Đại học Cần Thơ, số 31a (2014), trang 36-51. ISSN: 1859-2333.
- [2] Nguyễn Tấn Phong, Nguyễn Thái Nghe. 2014. Một giải pháp trong xây dựng Hệ thống gợi ý bài hát. Trang 149-154, kỷ yếu hội thảo quốc gia lần thứ XVII: Một số vấn đề chọn lọc của CNTT&TT (@2014). Nhà xuất bản Khoa học và kỹ thuật. ISBN: 978-604-67-0426-3
- [3] Huỳnh Lý Thanh Nhân, Nguyễn Thái Nghe. 2013. Hệ thống dự đoán kết quả học tập và gợi ý lựa chọn môn học. Kỷ yếu hội thảo quốc gia lần thứ XVI: Một số vấn đề chọn lọc của CNTT&TT (@2013), trang 110-118. Nhà xuất bản Khoa học kỹ thuật. ISBN: 987- 604-67-0251-1.
- [4] Nguyễn Thái Nghe. “Chương 2. Hệ thống gợi ý: Kỹ thuật và ứng dụng - PDF FreeDownload.” Accessed: Jan.01, 2024. [Online]. Available: <https://docplayer.net/71285438-Chuong-2-he-thong-goi-y-ky-thuat-va-ung-dung.html>
- [5] Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. “Session-based Recommendations with Recurrent Neural Networks.” Proceedings of the 4th International Conference on Learning Representations (ICLR), 2016.
- [6] Ruining He, Wang-Cheng Kang, and Julian McAuley. “Sequence-Aware Recommendation through Personalized Heterogeneous Embedding.” Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2017.
- [7] Ludovic Denoyer and Thanh-Nam Hoang. “Context-Aware Session-Based Recommendations with Recurrent Neural Networks.” Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM), 2016.
- [8] Dawen Liang, Jun Zhu, and Jianqiang Huang. “Neural Attentive Session-based Recommendation.” Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS), 2017.