DOI: https://doi.org/10.59294/HIUJS.32.2024.707

Xây dựng hệ khuyến nghị hỗn hợp áp dụng cho trang web thông tin rào cản kỹ thuật đối với thương mại

Nguyễn Minh Đế*, Lê Văn Hạnh và Tô Hoài Việt
Trường Đại học Quốc tế Hồng Bàng

TÓM TẮT

Bài toán đáp ứng được nhu cầu khách hàng về sản phẩm, dịch vụ là một trong nền tảng quan trọng hàng đầu của bên cung cấp. Bên cung cấp dùng nhiều phương pháp để cố gắng đưa ra các đề xuất sản phẩm, dịch vụ phù hợp với từng khách hàng và phía khách hàng có tương tác với sản phẩm, dịch vụ có quan tâm. Bên cung cấp luôn lưu thông tin người dùng cũng như lưu vết lại tất cả lịch sử giao dịch để lần sau phục vụ yêu cầu phù hợp hơn. Trong bài viết này giới thiệu một hệ khuyến nghị hỗn hợp thông tin Technical Barriers to Trade (TBT) dựa vào phản hồi ẩn và áp dụng cho trang web một điểm truy cập TBT cấp tỉnh/thành. Hệ khuyến nghị xây dựng vận dụng kết hợp kỹ thuật lọc theo nội dung và kỹ thuật lọc cộng tác tương ứng với hai phương pháp Mô hình không gian vector (kết hợp TF-IDF) và Phân tích ma trận Matrix Factorization. Bài báo đã cài đặt hệ khuyến nghị hỗn hợp trên vào một trang web ứng dụng và xây dựng một cơ sở dữ liệu thông tin TBT thu thập từ nhiều nguồn. Hệ thống trên được thử nghiệm với cơ sở dữ liệu cho thấy giải pháp này hoàn toàn thích hợp để tích hợp vào trang web của các điểm truy cập TBT.

Từ khóa: *hệ khuyến nghị hỗn hợp/lai, lọc theo nội dung, lọc cộng tác, mô hình không gian vector, TF-IDF, phân tích thừa số ma trận*

1. TỔNG QUAN

1.1. Đặt vấn đề

Hiện nay với sự phát triển của khoa học công nghệ và ứng dụng công nghệ thông tin giúp cho các nhà cung cấp hàng hóa, dịch vụ có thể hoạt động trực tuyến và cung cấp sản phẩm, dịch vụ cho người dùng khắp nơi trên thế giới. Do đó, bài toán nắm bắt thị hiếu, sở thích của người dùng là việc căn bản mà bên phía cung cấp phải cần giải quyết thật tốt. Theo thống của Amazon vào năm 2020, Amazon đã bán hơn 12 triệu sản phẩm, có 9 triệu khách hàng thành viên ở Hoa Kỳ [1]. Theo Tổ chức Thương mại Thế giới (World Trade Organization WTO) [2], từ 1995 đến 2020 có 51,431 bản thông báo tài liệu hàng rào cản kỹ thuật trong thương mại (Technical Barriers to Trade TBT) khác nhau do các quốc gia công bố chính thức. TBT là các tiêu chuẩn, quy chuẩn kỹ thuật mà một nước áp dụng đối với hàng hóa nhập khẩu và/hoặc quy trình nhằm đánh giá sự phù hợp của hàng hóa nhập khẩu đối với các tiêu chuẩn, quy chuẩn kỹ thuật đó. Từ 11/01/2007, Việt Nam đã chính thức trở thành thành viên của Tổ chức Thương mại Thế giới và bắt đầu thực hiện các cam kết gia nhập WTO, trong đó có cam kết thực thi toàn bộ các nghĩa vụ của Hiệp định TBT. Tính chung cả năm 2023, tổng xuất khẩu hàng hóa Việt Nam ước tính đạt 355,5 tỷ USD và số lượng và thể loại hàng hóa đủ ở các lĩnh vực [3]. Trong hợp tác thương mại toàn cầu, tiêu chuẩn (TC) và quy chuẩn kỹ thuật (QCKT) đã trở thành thước đo và chuẩn mực để so sánh, đánh giá chất lượng của sản phẩm hàng hóa và dịch vụ. Do đó, việc xây dựng hệ thống thông tin TBT cho các mạng lưới TBT Việt Nam (điểm truy cập cấp tỉnh) rất quan trọng và có ích cho cá nhân, doanh nghiệp có nhu cầu.

1.2. Bài toán

Người dùng khi có nhu cầu sản phẩm, dịch vụ thường không có đủ thời gian để xem xét, lựa chọn do sự phát triển mạnh lĩnh vực kinh doanh, giải trí trực tuyến với số lượng và chủng loại sản

Tác giả liên hệ: ThS. Nguyễn Minh Đế
Email: denm1@hiu.vn

phẩm, dịch vụ rất lớn. Như vậy, xét về cung cầu sản phẩm, dịch vụ có hai vấn đề phát sinh: Phía người dùng không biết nên xem chọn hàng hóa, dịch vụ nào phù hợp với sở thích, nhu cầu của bản thân; Phía bên cung cấp cần biết rõ và đúng sở thích, thị hiếu của phía người dùng để có kế hoạch hành động phù hợp và gia tăng lợi ích. Do đó, việc xây dựng một hệ khuyến nghị và tích hợp nó vào hệ thống công nghệ thông tin để hỗ trợ hoạt động của tổ chức là một phần quan trọng trong chiến lược hoạt động.

Về tổng quát, hệ khuyến nghị nói chung cần phải xây dựng được ma trận biểu diễn mối tương quan Người dùng-Sản phẩm M_{mxn}. Ma trận này biểu diễn mức độ quan tâm của người dùng với mỗi sản phẩm. Tập giá trị x_{ij} với $i \in m$ và $j \in n$ là phần tử của M_{mxn} mang ý nghĩa các giá trị trọng số (mức độ quan tâm người dùng i đối với sản phẩm j). Tập giá trị x_{ij} trong M_{mxn} thường thiếu nhiều các giá trị x_{ij}. Hệ khuyến nghị đưa ra giá trị xếp hạng dự đoán $\hat{r}_{ij}$ của người dùng u_i cho sản phẩm i_j chưa có tương tác (xếp hạng). Để giải bài toán trên cần xác định hàm $r(u_i, i_j)$ để ước lượng giá trị xếp hạng của người dùng u_i cho sản phẩm i_j sao cho sai số giữa giá trị dự đoán $\hat{r}_{ij}$ với các giá trị xếp hạng r_{ij} đã biết trong ma trận tương tác là nhỏ nhất.

1.3. Phương pháp tiếp cận giải quyết

Hệ khuyến nghị phải mô phỏng được quá trình ra quyết định của người dùng theo các cách tiếp cận như: Lọc nội dung (Content-Based Filtering); Lọc cộng tác (Collaborative Filtering); Hỗn hợp (Hybrid). Trong nghiên cứu [4] có giới thiệu các phương pháp tiếp cận xây dựng hệ thống khuyến nghị, có 3 cách tiếp cận chính:

I. Lọc theo nội dung (Content-Based CB):

Hệ khuyến nghị ghi nhận thông tin từng người dùng cụ thể mà có quan tâm đến từng thuộc tính của sản phẩm xác định, sau đó khuyến nghị sản phẩm tương tự nội dung. Công trình [5] đã liệt kê các kỹ thuật để thực hiện lọc theo nội dung, trong đó kỹ thuật CB chủ yếu dựa vào phản hồi đặc trưng của người dùng về sản phẩm. Phản hồi của người dùng được phân theo hai cách: Phản hồi ngầm phản ánh gián tiếp sở thích của người dùng; Phản hồi rõ ràng trực tiếp chỉ ra lựa chọn của người dùng. Như vậy, bài toán hoặc là tập trung vào phân lớp (dự đoán xem người dùng thích hay không thích một mặt hàng) hoặc là hồi quy (dự đoán mức độ đánh giá mà người dùng đưa ra cho một sản phẩm, dịch vụ). Thông thường, các phương pháp dựa trên tiếp cận nội dung sẽ thực hiện theo hai hướng: Dựa trên bộ nhớ, thực hiện tính toán độ tương tự giữa nội dung sản phẩm, dịch vụ với hồ sơ người dùng xác định mà dùng các độ đo tương tự (Cosine, Euclidean, …); Dựa trên mô hình học từ dữ liệu mà có dùng các kỹ thuật học máy giám sát để phân các đối tượng khuyến nghị thành những đối tượng người dùng có quan tâm (giá trị 1) hay không quan tâm (giá trị 0).

ii. Lọc cộng tác (Collaborative Filtering CF):

Tiếp cận CF được xem là tiếp cận thành công nhất để xây dựng các hệ thống khuyến nghị và ứng dụng rộng rãi trong lĩnh vực thương mại điện tử [4]. Lọc cộng tác thực hiện tư vấn (gợi ý) các sản phẩm, dịch vụ cho một người dùng nào đó dựa trên mối quan tâm, sở thích của những người dùng tương tự đối với các sản phẩm, dịch vụ đó. Lọc cộng tác được xem là một trong cách tiếp cận chính trong xây dựng các hệ thống tư vấn và kỹ thuật này được chia thành hai dạng chính:

- Memory–based: Lọc cộng tác dựa trên việc ghi nhớ toàn bộ dữ liệu. Kỹ thuật này vận dụng các thuật toán tính toán tương tự, lân cận.
- Model–based: Lọc cộng tác dựa trên các mô hình phân lớp, dự đoán. Kỹ thuật này vận dụng các thuật toán gom cụm, phân lớp giám sát, thừa số hóa ma trận (Matrix Factorization).

iii. Hỗn hợp/ lai (Hybrid):

Hai cách tiếp cận xây dựng trên đều có các điểm mạnh, cũng như các điểm yếu. Để tận dụng những điểm mạnh và hạn chế điểm yếu của những tiếp cận khác nhau, nhiều nghiên cứu đã tập trung phát triển các hệ khuyến nghị dựa trên việc kết hợp các tiếp cận khác nhau, được gọi là tiếp cận hỗn hợp/ lai Hybrid Approach và đã cho các kết quả tốt [6].

1.4. Phương pháp sử dụng

Trong nghiên cứu này đã xây dựng hệ khuyến nghị hỗn hợp có vận dụng hai phương pháp là mô hình không gian vector và phân tích ma trận tương ứng với hai cách tiếp cận khuyến nghị lọc theo nội dung và lọc cộng tác:

- Phương pháp gợi ý dựa trên nội dung với TF-IDF và Vector Space Model (VSM):

Phương pháp gợi ý dựa trên nội dung đã được nghiên cứu từ lâu và phương pháp TF-IDF với VSM

đã có kết quả rất tốt [7]. Term Frequency - Inverse Document Frequency TF-IDF là một trong các kỹ thuật cơ bản trong xử lý ngôn ngữ giúp đánh giá mức độ quan trọng của một từ trong văn bản. TF-IDF còn là một phương thức thống kê được biết đến rộng rãi nhất để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều đoạn văn bản khác nhau. VSM là một mô hình đại số tuyến tính biểu diễn dạng văn bản thành dạng một vector, trong đó các phần tử có thể biểu diễn mức độ quan trọng của một từ (TF-IDF) hoặc dạng có mặt hoặc vắng mặt của một túi từ (Bag of Words) trong đoạn văn bản. Không gian vector VSM còn được chuyển đổi từ dạng biểu diễn TF-IDF của văn bản.

- Phân tích ma trận Matrix Factorization MF:
  MF là kỹ thuật phân rã ma trận và là kỹ thuật của lọc cộng tác được sử dụng phổ biến nhất do tính hiệu quả vào nhiều lĩnh khác nhau trong đó nổi trội là lĩnh vực thương mại điện tử [8]. Ý tưởng chính của Matrix Factorization là đặt người dùng và sản phẩm vào trong cùng một không gian thuộc tính ẩn. MF chia một ma trận lớn X thành hai ma trận có kích thước nhỏ hơn là W và H, sao cho có thể xây dựng lại X từ hai ma trận nhỏ hơn này càng chính xác càng tốt, nghĩa là $X \sim W \times H^T$. Trong đó với $x \in X$ là một vector hồ sơ sản phẩm.

1.5. Đóng góp của nghiên cứu

Trong khuôn khổ nghiên cứu có hai đóng góp:

a. Đề xuất một hệ khuyến nghị sử dụng phương pháp hỗn hợp để kết hợp đặc trưng người dùng và thông tin (bao gồm tài liệu bản tin TBT có đính kèm nhúng nhận diện (Identity ID Embedding)). Phương pháp hỗn hợp đề xuất này đã tận dụng có kết hợp sức mạnh thống kê phân tích để biểu diễn nội dung các thông tin TBT dạng vector và sức mạnh phân tích MF.

b. Áp dụng hệ thống đã xây dựng để phát triển một ứng dụng web cho cổng thông tin điện tử TBT (điểm truy cập TBT) có phân hệ tự động đề xuất thông tin TBT phù hợp và chính xác cho từng người dùng xác định. Các thực nghiệm ở đây đều được thực hiện với tập dữ liệu được xây dựng thành một cơ sở dữ liệu CSDL được thu thập từ các nguồn chính thức và qua quá trình thực nghiệm mô phỏng.

Phần còn lại của bài báo như sau: Phần 2, phần kế tiếp mô tả bài toán và phân tích dữ liệu thông tin TBT; Phần 3, trình bày phương pháp sử dụng và mô tả hệ thống thông tin TBT; Phần 4, cài đặt môi trường, thực hiện thử nghiệm và đánh giá; Phần 5, thảo luận và kết luận.

2. MÔ TẢ BÀI TOÁN, PHÂN TÍCH DỮ LIỆU VÀ ĐẶC TRƯNG CỦA THÔNG TIN TBT

2.1. Bài toán

Hiệp định TBT là một trong số 29 văn bản pháp lý nằm trong Hiệp định WTO, trong đó tài liệu (bản thông báo) TBT là thành phần quan trọng nhất của Hiệp định TBT [2]. Mỗi quốc gia thuộc WTO đều phải có nghĩa vụ cung cấp minh bạch và công khai các TBT này. Hiện nay, các cá nhân và doanh nghiệp ở Việt Nam còn lúng túng trong việc áp dụng tiêu chuẩn và quy chuẩn kỹ thuật, do không biết phải áp dụng tiêu chuẩn gì cho phù hợp và sản phẩm khi đưa ra thị trường có đạt tiêu chuẩn hay không. Do đó, một điểm truy cập TBT cung cấp thông tin cho cá nhân, tổ chức xác định một các phù hợp và chính xác là một nội dung cần phải có và là mối quan tâm hàng đầu. Một điểm truy cập TBT theo hướng dẫn TBT Việt Nam phải đáp ứng được các nội dung: Tin tức; Giới thiệu về chức năng nhiệm vụ của điểm TBT của Bộ hoặc địa phương; Hoạt động thông báo; Hoạt động hỏi đáp. Đa số các tỉnh thành Việt Nam đều có trang web TBT của địa phương, phần "Tin tức sự kiện" các trang web đa số đơn thuần là liệt kê tin tức và sự kiện TBT tiếng Việt theo thời gian gần đây nhất và còn có hiệu lực hay không. Theo định nghĩa [4], bài toán khuyến nghị thông tin TBT được khai báo với đầu vào và đầu ra:

I. Đầu vào

- Đầu vào 1: Một tập hợp tất cả người dùng trong hệ thống U; Có m người dùng; Mỗi người dùng $u_i \in U$ có các đặc điểm $u_i = \{u_{i1}, u_{i2}, \ldots u_{ik}\}$, với k là số lượng đặc điểm của người dùng tạo thành vector đặc trưng người dùng có k chiều.
- Đầu vào 2: Cho *I* là tập tất cả thông tin TBT; Có n thông tin; Mỗi thông tin $i_j \in I$ có các đặc điểm đặc trưng V với $v_j = \{v_{j1}, v_{j2}, \ldots v_{jl}\}$, với l là số lượng đặc điểm của thông tin tạo thành vector thông tin có l chiều.
- Đầu vào 3: Dữ liệu tương tác/phản hồi được xếp hạng $r_{ij} \in \mathbb{R}$ là giá trị xếp hạng của người dùng u_i đối với thông tin i_j.
- Với đầu vào như trên, xây dựng mô hình biểu diễn mối tương quan qua ma trận M_{mxn}:

Bảng 1. Ma trận biểu diễn

	Thông tin				
Người dùng		1	2	...	N
	1	x_{11}			x_{1n}
	2		x_{22}	$x_{..}$	
	...		$x_{..2}$	$x_{..}$	$x_{..n}$
	m	x_{m1}	x_{m2}	$x_{..}$	x_{mn}

ii.*Đầu ra:* Danh sách thông tin TBT $i_j \in I$ có độ phù hợp người dùng u_i thuộc U nhất.

2.2. Phân tích thông tin TBT

Căn cứ vào khung phân loại văn bản trên Cổng thông tin TBT Việt Nam, các văn bản TBT các điểm truy cập TBT được phân loại theo thành 4 loại chính: Các thông báo của Việt Nam; Các thông báo của các nước thành viên WTO; Tranh chấp thương mại; Văn bản pháp luật. Trong nghiên cứu này, đối với thông tin TBT thì chỉ xét đến hai loại sau:

Bản thông báo TBT trong và ngoài nước

Tài liệu TBT được sử dụng trong bài báo này là các bản thông báo tóm tắt TBT, được tải xuống trực tiếp từ [2]. Bản thông báo tóm tắt sử dụng ở đây tất cả đều là tiếng Việt (thông báo Việt Nam) và tiếng Anh (thông báo nước ngoài) và có các thông tin theo khuôn mẫu sau, theo đúng thứ tự từ trên xuống:

(1) Thông tin chung: Mã số; Ngày thông báo; Ngôn ngữ gốc.

(2) Thông tin chi tiết gồm có 11 thành phần

a) Thành viên: Tên quốc gia thông báo theo quy cách chuẩn chung.

b) Cơ quan: thông tin chi tiết đến cơ quan thuộc quốc gia.

c) Điều: theo danh sách trong Hiệp định TBT gồm có 15 Điều.

d) Sản phẩm: theo mã số thuộc khung HS hoặc ICS, với Hệ thống mã hóa và mô tả hàng hóa hài hòa (Harmonized System HS) và Tiêu chuẩn Quốc tế (International Classification for Standards ICS) của Tổ chức Tiêu chuẩn hóa Quốc tế (International Organization of Standardization ISO).

e) Tiêu đề: gồm có 3 phần tiêu đề, số trang và ngôn ngữ.

f) Mô tả: đoạn văn bản ngắn có độ dài trung bình 100 từ mô tả vắn tắt nội dung.

g) Mục tiêu và lý do: một câu văn trình bày mục tiêu hoặc lý do ra thông báo này.

h) Tài liệu văn bản có liên quan: Đoạn văn liệt kê các tài liệu văn bản có liên quan.

I) Ngày: gồm 2 loại (Ngày đề xuất thông báo; Ngày có hiệu lực).

j) Thời hạn: khoảng thời gian cho phép ý kiến, thường là 60 ngày.

k) Tài liệu: thông tinh chính thức cơ quan ban hành thuộc quốc gia và địa chỉ để lấy Tài liệu thông tin TBT bản chi tiết và đầy đủ.

Trong bài báo này chỉ cần sử dụng 8 đặc điểm của Bản tóm tắt thông báo TBT làm vector đặc trưng có 8 chiều, các đặc điểm được sử dụng: 1) Mã số; 1) Ngôn ngữ gốc; 2. a); 2. d); 2. e); 2. f); 2. i); 2. k).

Tin tức TBT

Tin tức TBT là các tin tức thuộc loại tranh chấp thương mại hoặc văn bản pháp luật được hệ thống thu thập từ các điểm truy cập TBT trong nước và một số trang web nước ngoài nhờ công nghệ RSS (Feeds - Really Simple Syndication). Một tin tức thu được từ các nguồn sẽ có giống nhau về khung nội dung: Tiêu đề; Phân loại/nhóm tin; Tóm tắt; Nội dung; Ngày tháng; Đường dẫn; Tác giả/cơ quan.

2.3. Đặc trưng của bài toán thông tin TBT

Người dùng trong bài toán khuyến nghị thông tin TBT chính là các tác nhân truy cập ứng dụng trang web của điểm truy cập TBT. Người dùng khi truy cập vào trang web sẽ được chia làm hai loại: Đăng nhập (có thông tin đăng ký); Không đăng nhập (không đăng ký). Đặc trưng từng loại người dùng như sau:

- Đối với người dùng có đăng ký thì thông tin (hồ sơ profile) của người dùng ảnh hưởng tới việc chọn và đọc một thông tin TBT được đặc trưng bởi các thông tin: Địa chỉ (Quốc gia, Tỉnh/thành, ...); Lĩnh vực kinh doanh (sản phẩm nông nghiệp, công nghiệp nằm trong khung HS hoặc ICS); Quan tâm (đọc trong k giây); ... Tất cả thông tin này sẽ được hệ thống lưu lại trong Cơ sở dữ liệu riêng.

- Đối với người dùng không đăng ký thì sử dụng một phương pháp riêng để xác định địa chỉ MAC (Media Access Control) của thiết bị máy tính mà người dùng sử dụng truy cập trang web. Địa chỉ MAC thường được chỉ định bởi nhà sản xuất bộ điều khiển giao diện mạng (Network Interface Card NIC) và được lưu trữ trong phần cứng này và thường được mã hóa số nhận dạng của nhà sản xuất NIC đã đăng ký. MAC cũng có thể được biết đến như một địa chỉ phần cứng Ethernet (EHA), địa chỉ phần cứng hoặc địa chỉ vật lý và giúp xác định các thiết bị được kết nối với một mạng nhất định. Như vậy, đối với nhiều lượt truy cập khác nhau từ một địa chỉ MAC thì đều được xác định là cùng một người dùng, nên hồ sơ cùng là một nguồn.

Người dùng thực hiện các thao tác trên trang web, hệ khuyến nghị lưu vết các lịch sử giao dịch và các trạng thái sử dụng của người dùng như: Chọn và đọc bản thông tin TBT (Bản tóm tắt tài liệu TBT, Tin trong nước, tin ngoài nước, …); Yêu cầu đặt mua chính thức bản đầy đủ tài liệu TBT; Yêu cầu tư vấn về sản phẩm hàng hóa có mã số HS hoặc ICS. Ngoài ra, hệ thống còn lưu lại lịch sử truy cập tin tức như: Đọc tiêu đề thông tin lại bao nhiêu lần trong khoảng thời gian k; Đọc một thông tin TBT và đọc tiếp các thông tin TBT khác.

3. PHƯƠNG PHÁP VẬN DỤNG VÀ MÔ TẢ HỆ THỐNG

Hệ thống khuyến nghị thông tin TBT với bài toán đã xác định ở phần trên được xây dựng dựa vào hai loại giải thuật: Giải thuật gợi ý dựa trên nội dung; Giải thuật gợi ý cộng tác. Giải thuật khuyến nghị dựa trên nội dung vận dụng mô hình không gian vector VSM và giải thuật khuyến nghị cộng tác vận dụng kỹ thuật phân tích thừa số ma trận MF, hai phương pháp sẽ được tích hợp vào hệ thống để phát sinh ra các khuyến nghị phù hợp với người dùng.

3.1. Mô hình không gian vector (VSM)

Phương pháp Lọc theo nội dung được thực hiện dựa trên việc so sánh nội dung mô tả thông tin TBT để tìm ra thông tin tương tự với những gì mà người dùng đã từng quan tâm để giới thiệu tương ứng. Mô hình VSM để biểu diễn tài liệu văn bản ngôn ngữ tự nhiên ở dạng các vector nhiều chiều, dựa vào các từ ngữ trong văn bản. Mô hình VSM được sử dụng trong bài báo để lựa chọn và quyết định tài liệu TBT nào là thích hợp nhất với một người dùng xác định trước.

Trong bài báo này còn sử dụng phương pháp thống kê để xác định độ quan trọng của một từ trong đoạn văn bản trong một tập nhiều đoạn văn bản khác nhau là TF-IDF. TF-IDF chuyển đổi dạng biểu diễn văn bản thành dạng không gian vector nên thường được sử dụng như một trọng số trong việc khai phá dữ liệu văn bản.

TF (Term frequency): Tần suất xuất hiện của 1 từ trong 1 văn bản.

TF(t, d) = (Số lần xuất hiện từ t) / (Tổng số từ)

Giải thuật: Tính TF

Input: term, doc

Output: giá trị tf

1. Khởi tạo tham số, thủ tục và thành phần cần thiết
2. r ← 0; l = len(doc)
3. Duyệt từng từ word trong doc
   Nếu word = term ❼ r++
4. Return r / l

Hình 1. Giải thuật tính TF

IDF (Inverse Document Frequency): Dùng để đánh giá mức độ quan trọng của 1 từ trong văn bản. Khi tính tf mức độ quan trọng của các từ là như nhau. IDF(t, D) = log_e(Số văn bản trong tập D/Số văn bản chứa từ t trong tập D.

Giải thuật: Tính IDF

Input: term, docs

Output: giá trị idf

1. Khởi tạo tham số, thủ tục và thành phần cần thiết
2. r ← 0; l = len(docs)
3. Duyệt từng văn bản doc trong docs
   - 3.1. Duyệt từng từ word trong doc
     - Nếu word = term thì {r++; break;}
4. Return math.log(l/r, math.e)

Hình 2. Giải thuật tính IDF

Thực hiện tính toán tổng hợp cho TF-IDF dùng công thức: TF-IDF(t, d, D) = TF(t, d) * IDF(t, D). Với tf là số lần xuất hiện của từ t trong tài liệu f; f là tổng số các từ trong tài liệu f; W là trọng số tính được. Những từ có tf-idf là những từ xuất hiện nhiều trong văn bản này và xuất hiện ít trong văn bản khác. Việc tính giá trị này giúp lọc ra những từ phổ biến và giữ lại những từ có giá trị cao trong văn bản (keyword).

Thực hiện tính toán Normalizing Vectors: Mỗi vector là một danh sách các hệ số [a, b] để định nghĩa độ lớn của vector trong chiều đó. Trong VSM [9], có mỗi từ trong câu truy vấn là một chiều, với câu truy vấn có 'n' từ sẽ có được một vector n-chiều. Mỗi thông tin TBT là một vector có nhiều chiều. Tiêu đề thông tin TBT $\vec{q}$ và tiêu đề cần so sánh trong cơ sở dữ liệu thu thập được $\vec{d}$ cần được tính toán thành điểm số xác định. Để tính điểm số so sánh xác định này có hai cách: Góc giữa $\vec{q}$ và $\vec{d}$; Khoảng cách giữa $\vec{q}$ và $\vec{d}$. Mỗi thông tin TBT có độ dài khác nhau nên việc tính toán trong nghiên cứu này chọn góc trong không gian vector tính điểm so sánh giữa $\vec{q}$ và $\vec{d}$. Công thức như sau:

$$q' = \left(\frac{q_1}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \frac{q_2}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \dots, \frac{q_m}{\sqrt{\sum_{i=1}^{m} q_i^2}}\right), \text{với q là trọng số TF*IDF} \quad (3.1.1)$$

$$d' = \left(\frac{d_1}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \frac{qd_2}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \dots, \frac{d_m}{\sqrt{\sum_{i=1}^{m} q_i^2}}\right), \text{với q là trọng số TF*IDF} \quad (3.1.2)$$

Chuẩn hóa độ dài của tài liệu: Với cái tài liệu dài (hơn) nên có tần suất các từ xuất hiện cao hơn (higher term frequencies). Một từ giống nhau sẽ có khả năng xuất hiện thường xuyên nên có nhiều từ hơn nên tăng khả năng xuất hiện của các từ trùng. Sự "chuẩn hóa cosine" giảm sự ảnh hưởng của tài liệu dài (so với tài liệu ngắn). Một vector được chuẩn hóa (về độ dài) với cách chia từng phần của nó cho độ dài của nó. Công thức tương đồng cosin:

$$\cos(\vec{q}, \vec{d}) = \frac{\vec{q},\ \vec{d}}{\|\vec{q}\|\|\vec{d}\|} = \frac{\sum_{i=1}^{|v|} q_i d_i}{\sqrt{\sum_{i=1}^{|v|} q_i^2}\sqrt{\sum_{i=1}^{|v|} d_i^2}}, \quad (3.1.3)$$

với q_i là trọng số tf-idf của từ i trong thông tin, d_i là trọng số tf-idf của từ i trong tài liệu $cos(\vec{q}, \vec{d})$

Đối với vector được chuẩn hóa về độ dài thì sự tương đồng cosine là tích vô hướng của hai vector (scalar product) với công thức:

$$\cos(\vec{q}, \vec{d}) = \vec{q} \,.\, \vec{d} = \sum_{i=1}^{|v|} q_i d_i \quad (3.1.4)$$

Từ các thành phần tính toán trên, bài báo xây dựng giải thuật để phát sinh khuyến nghị theo nội dung thông tin TBT qua các bước sau:

Giải thuật: Phát sinh khuyến nghị theo nội dung CB

Input: Tập T chứa danh sách các tiêu đề thông tin TBT; Tập O chứa danh sách các tiêu đề thông tin TBT khuyến nghị.

Output: Tập O chứa danh sách các tiêu đề thông tin TBT khuyến nghị đã xếp hạng theo độ ưu tiên

1. Xử lý dữ liệu cho tập T
2. Xử lý dữ liệu cho tập O
3. Tính: IDF(w)=log(N/D_f (w))

Với N là tổng số lượng tài liệu cần tý vấn cho người sử dụng; D_f *(w) là số lượng tài liệu mà một từ nào đó xuất hiện; w là 1 từ cần tính*

4. Tính: W= t_f/f*IDF

Với t_f *là số lần xuất hiện của từ t trong tài liệu f; f là tổng số các từ trong tài liệu f.*

5. Tính Normalizing Vectors
6. Tính độ tương đồng của các Normalizing Vectors bằng độ đo cosin
7. Xuất kết quả theo khung tập O

Hình 3. Giải thuật phát sinh khuyến nghị theo nội dung CB

3.2. Phân tích ma trận với MF

Kỹ thuật phân tích MF [10] sử dụng lại định nghĩa mỗi thông tin được mô tả bằng một vector x mà được gọi là hồ sơ thông tin (item profile). Phân tích ma trận là tìm một vector hệ số w tương ứng với mỗi người dùng sao cho với đánh giá (ratings) đã biết mà người dùng đó cho thông tin mà xấp xỉ với với công thức: y ≈ xw. Trong công thức có X là ma trận của toàn bộ hồ sơ thông tin, mỗi hàng tương ứng với 1 thông tin, W là ma trận của toàn bộ người dùng, mỗi cột tương ứng với 1 người dùng. Kỹ thuật này cố gắng xấp xỉ Y ∈ R_{mxn} bằng tích của hai ma trận X ∈ R_{mxk} và W ∈ R_{kxn} với k được chọn sao cho nhỏ hơn m và n.

Mô hình bài toán với các đại lượng $w_i^T x_j$ được biểu diễn với m người dùng và n thông tin:

Bảng 2. Ma trận người dùng và thông tin

Thông số	Thông tin 1 (x_1)	Thông tin 2 (x_2)	...	Thông tin n (x_n)
Người dùng 1 (w_1)	x_1 w_1	x_1 w_2	...	x_1 w_n
Người dùng 2 (w_2)	x_2 w_1	x_2 w_2	...	x_2 w_n
...	...	...	...	...
Người dùng m (w_m)	x_m w_1	x_m w_2	...	x_m w_n

Ma trận trên là đặc trưng của người dùng và thông tin trên với W ∈ R_{kxn} và X ∈ R_{kxm}, ma trận Y ∈ R_{mxn} có Y ≈ W^T X. Như vậy, kỹ thuật phân tích này sử dụng *k* nhân tố tiềm ẩn (latent factors) mô tả người dùng *u* và *k* nhân tố tiềm ẩn mô tả cho thông tin *i*. Trong đó có *x* là một vector của hồ sơ thông tin. MF đã khai thác sự tương đồng giữa các người dùng trong các tùy chọn và tương tác của người dùng đó để cung cấp các đề xuất tương tự. Giả sử một người dùng u không chọn đọc thông tin i, có thể giải quyết bằng cách tìm một người dùng u khác mà có cùng sở thích tương tự với người dùng u. Hệ thống lấy xếp hạng được đánh giá cho thông tin i và xem như là xếp hạng cho người dùng u.

Thực tế, giá trị m và n rất lớn nên kỹ thuật này lặp đi lặp lại việc tối ưu ma trận khi cố định ma trận còn lại và có thể áp dụng việc lấy tích của hai vector X x W. Mỗi vector có độ dài k là một số nhỏ hơn nhiều so với m và n, do đó việc tính toán này không cần đòi hỏi bộ nhớ lưu trữ tính toán quá lớn. Điều này làm

cho kỹ thuật MF phù hợp với các tập dữ liệu người dùng và thông tin rất lớn, cụ thể khi lưu trữ ma trận X và W chỉ với bộ nhớ nhỏ do chỉ cần chứa được k(m + n) phần tử thay vì lưu m^2 hoặc n^2.

Xử lý thiên lệch và mất mát trong quá tình phân tích ma trận X

Trong quá trình hệ thống điền giá trị vào ma trận X, giá trị x_{ij} điền vào tương ứng với người dùng i đã đánh giá (đọc) thông tin j, giá trị x_{ij} này thực tế sẽ có sự thiên lệch của người dùng và/hoặc thông tin. Ví dụ, người dùng có thể thấy thông tin TBT trên điểm truy cập thì cứ đọc để tăng hiểu biết (chọn đọc tùy tiện do có dư thời gian, thể hiện sự dễ tính), hoặc người dùng "khó tính, không vừa ý và không có thời gian nhiều để đọc kỹ". Vấn đề thiên lệch này có thể giải quyết bằng cách thêm các biến thiên lệch "bias", biến này sẽ phụ thuộc mỗi người dùng/ thông tin và có thể được tối ưu cùng với X và W. Với một người dùng i và thông tin j với vector đặc trưng tương ứng lần lượt là w_i và x_j, độ yêu thích của người dùng tới thông tin đó có thể được mô tả bởi công thức có thêm hệ số:

$$\hat{y} \approx w_i^T x_j + b_j + d_i + a \quad (3.2.1)$$

Các hệ số như sau:

- b_j là hệ số tự do ứng với người dùng i thể hiện việc người này có "khó tính, sở thích đặc biệt" hay không;
- d_i là hệ số tự do, ứng với thông tin j thể hiện thông tin nhìn chung có được quan tâm hay không;
- a là hệ số tự do, thể hiện thiên hướng chung của bộ dữ liệu.

Việc xây dựng hàm mất mát cũng được dựa trên các thành phần giá trị x_{ij} đã được điền của ma trận Y, việc xây dựng hàm mất mát như sau:

$$L(X,W,b,d) = \frac{1}{2s}\sum_{j=1}^{n}\sum_{i:r_{ij}}(x_i w_j + b_i + d_j + a - y_{ij})^2 + \frac{1}{2}(\|X\|_F^2 + \|W\|_F^2 + \|b\|_2^2 + \|d\|_2^2) \quad (3.2.2)$$

Tiến hành cố định *X, b* và tối ưu *W, d* và ngược lại, cố định *W, d* và tối ưu *X, b*, tương ứng theo các công thức:

$$w_j = w_j - \eta\frac{1}{s}X_j(X_j^T w_j + b_j + d_j - y_{ij}) + \gamma w_j) \quad (3.2.3)$$

$$w_j = w_j - \eta\frac{1}{s}X_j(X_j^T w_j + b_j + d_j - y_{ij}) + \gamma w_j) \quad (3.2.4)$$

$$x_i = x_i - \eta\frac{1}{s}W_i(W_i^T x_i + b_j + d_j - y_{ij}) + \gamma x_i \quad (3.2.5)$$

$$d_i = d_i - \eta\frac{1}{s}1^T(W_i^T x_i + b_j + d_j - y_{ij}) \quad (3.2.6)$$

Từ các công thức (3.2.1), (3.2.2), (3.2.3), (3.2.4), (3.2.5), xây dựng giải thuật cài đặt và tính toán để thu được các ma trận *X, b, W, d* để từ đó dự đoán các đánh giá người dùng chưa biết.

Giải thuật: Phát sinh khuyến nghị theo cộng tác CF

Input: Tập T chứa danh sách các tiêu đề thông tin TBT; Tập O chứa danh sách các tiêu đề thông tin TBT khuyến nghị.

Output: Tập O chứa danh sách các tiêu đề thông tin TBT khuyến nghị đã xếp hạng theo độ ưu tiên

1. Xử lý dữ liệu cho tập T bao gồm chuẩn bị dữ liệu
2. Xử lý dữ liệu cho tập O
3. Định nghĩa mô hình với MF
  3.1. Chuẩn bị bộ 3 giá trị: Người dùng; Thông tin TBT; Số yếu tố ẩn/kích thước nhúng *(latent factors/ embedding size)*
  3.2. Khởi tạo Mô hình với các tham số
  3.3. Duyệt Mô hình theo hướng tới
  3.4. Bước huấn luyện Mô hình
  3.5. Tính giá trị tối ưu
  3.6. Xác thực Mô hình
  3.7. Xác định cuối vòng huấn luyện
  3.8. Xác định cuối vòng xác thực Mô hình
4. Huấn luyện Mô hình với các tham số
5. Tính RMSE cho Mô hình
6. Xuất kết quả theo khung tập O

Hình 4. Giải thuật phát sinh khuyến nghị theo cộng tác CF

3.3. Mô tả hệ thống

Hệ thống được xây dựng và tích hợp vào trang web cung cấp thông tin TBT, giúp người dùng có thể tương tác với thông tin mong muốn tương tác. Khi người dùng truy cập vào trang web, phân hệ tự động sẽ khuyến nghị danh mục từng loại thông tin TBT mà người dùng hiện thời có thể quan tâm. Với danh sách này, hệ thống còn cung cấp thông tin chi tiết về các thông tin TBT như: Tiêu đề; Phân loại/nhóm tin; Tóm tắt; Nội dung; Ngày tháng; Đường dẫn; Tác giả … Hệ thống trong quá trình hoạt động luôn thu thập phản hồi từ người dùng và lưu vết quá trình truy cập hệ thống. Đối với từng thông tin TBT, hệ thống sẽ được ghi nhận phản hồi người dùng dưới dạng phản hồi tiềm ẩn (implicit feedback) và tự động ghi nhận lại giá trị phản hồi được tính bằng thời gian đọc là k giây. Giá trị này được hệ thống lưu vào CSDL và dùng để huấn luyện mô hình và thực hiện chức năng khuyến nghị.

Hệ thống khuyến nghị xây dựng luôn cho ra kết quả dự đoán thông qua một danh sách kiểu Top-N tin tức được sắp xếp giảm dần theo độ tương quan theo từng nhóm danh mục một. Như trình bày ở trên, hệ thống xuất ra kết quả danh sách tin tức TBT có phân mục giống nhau cho cả hai loại người dùng đăng nhập và không đăng nhập. Khi người dùng tương tác với một thông tin TBT trong khoảng thời gian k giây, hệ thống bắt đầu xử lý để có thể phát sinh danh mục chứa danh sách thông tin TBT. Danh sách này được sắp xếp theo thứ tự ưu tiên giảm dần theo giá trị dự đoán nhằm giúp cho người dùng có thể xác định thông tin TBT mà phù hợp nhất. Ngoài ra, hệ thống còn khuyến nghị thêm các loại thông tin TBT theo các phương pháp đơn giản khác, với mỗi loại là 2 bản tin TBT: Cùng thể loại; Cùng cơ quan/chính phủ ban hành; Được đọc nhiều nhất; Đã hết hiệu lực hoặc bắt đầu có hiệu lực.

Phương pháp phát sinh khuyến nghị theo danh sách kiểu Top-N theo giải thuật sau:

Giải thuật: Phát sinh khuyến nghị theo danh sách kiểu Top-N

Input: Tập hợp đầu vào hệ thống

Output: Danh sách Top-N có phân mục

1. Nếu người dùng không đăng nhập:

1.1. Gợi ý 2 tiêu đề thông tin mới nhất

1.2. Gợi ý 1 tiêu đề đọc nhiều nhất, và tìm người dùng đọc chủ đề này nhiều nhất để gợi ý ra 2 tiêu đề thông tin tương ứng

1.3. Lấy chủ đề có tiêu đề đọc nhiều nhất rồi dùng giải thuật CF để gợi ý hai thông tin đọc nhiều nhất

2. Nếu người dùng đăng nhập:

2.1. Dùng giải thuật CB tìm chủ đề gợi ý hai tiêu đề đọc nhiều nhất trong chủ đề này

2.2. Dùng giải thuật CF tìm chủ đề tương tự như chủ đề vừa chọn/đọc để gợi ý hai tiêu đề thông tin đọc nhiều nhất

2.3. Dùng giải thuật CF tìm người dùng tương tự với người dùng hiện tại, lấy chủ đề có tiêu đề mà người dùng này đọc nhiều nhất để gợi ý hai thông tin đọc nhiều nhất trong chủ đề này

Hình 5. Giải thuật phát sinh khuyến nghị theo danh sách kiểu Top-N

4. THỰC NGHIỆM VÀ KẾT QUẢ

4.1. Xây dựng tập dữ liệu và cấu hình cài đặt

Các kết quả nghiên cứu này được thực nghiệm với tập dữ liệu thông tin TBT được lưu trong CSDL. Dữ liệu được thu thập với phương pháp thủ công và tự động từ [2] và từ nguồn các trang web chính thức cung cấp thông tin TBT cấp quốc gia và cơ quan thuộc quốc gia nhờ kỹ thuật RSS. Ngoài ra, thông tin người dùng được lưu trữ lại từ thông tin đăng ký sử dụng và nhật ký sử dụng của người dùng. Tất cả các dữ liệu trên được sử dùng để xây dựng CSDL cho hệ thống khuyến nghị áp dụng cho trang web

thông tin TBT. Hệ thống xây dựng đã chạy thực nghiệm với tập dữ liệu: Người dùng có 245 gồm hai loại (Đăng ký; Không đăng ký) với 8,586 giao tác; Tập dữ liệu TBT có 32,134 Bản thông báo tóm tắt TBT của 198 quốc gia, với 1,425 tin tức TBT. Hệ thống khuyến nghị được phát triển trên môi trường .NET sử dụng ngôn ngữ ASP.NET, C#, Python và Hệ quản trị cơ sở dữ liệu SQL Server.

4.2. Dữ liệu thực nghiệm

Hoạt động hệ thống khuyến nghị theo cách thức là hiển thị ra giao diện người dùng danh sách thông tin (Top-N) theo thứ tự giảm dần về độ tương quan. Hệ thống này triển khai thực nghiệm với 2 trạng thái của một người dùng khi sử dụng ứng dụng Web thuộc hệ thống (Người dùng đăng nhập; Người dùng không đăng nhập). Đối với tiêu đề của thông tin TBT thì trong bài báo chỉ xét theo các tiêu chí: Mới nhất; Tương tác (Đọc; Xem; Chọn; Tải xuống; Chia sẻ; …); … Tập dữ liệu thử nghiệm được chia làm 4 bộ khác nhau:

Bảng 3. Bảng dữ liệu ứng với 4 bộ dữ liệu

File huấn luyện	Số lượng bản ghi	File kiểm	Số lượng bản ghi
Train_1.mdf	124534	Test_1.mdf	17572
Train_2.mdf	124534	Test_2.mdf	17572
Train_3.mdf	124534	Test_3.mdf	17572
Train_4.mdf	124534	Test_4.mdf	17572

4.3. Tiêu chuẩn đánh giá

Trong nghiên cứu này sử dụng tiêu chuẩn đánh giá Root Mean Square Error (RMSE) để đo lường độ chính xác các dự đoán. Công thức RMSE:

$$RMSE = \sqrt{\frac{1}{n}\sum_{u,i}(r_{ui} - \mathrm{r}_{ui})^2}\,, \qquad (4.3.1)$$

với

– r_{ui} là thời gian đọc (k số giây thực đọc) thông tin TBT i

– r_{ui} là thời gian đọc (k số giây dự đoán do hệ thống xác định) thông tin TBT i

– n là tổng số dự đoán đánh giá

4.4. Kết quả thực nghiệm và đánh giá

Sau khi tiến hành thực nghiệm thì nhóm nghiên cứu thu được kết quả như sau:

Bảng 4. Kết quả RMSE ứng với 4 bộ dữ liệu

Bộ dữ liệu 1	Bộ dữ liệu 2	Bộ dữ liệu 3	Bộ dữ liệu 4	$RMSE_{tb}$
1.295	1.295	1.245	1.199	1.287

Với kết quả bảng trên sử dụng đơn vị đo là giây, các giá trị RMSE cho 4 bộ dữ liệu đều lân cận 1.3 giây. Theo thống kê trong bộ dữ liệu thu thập, giá trị trung bình một người dùng có đọc một thông tin TBT (trung bình có khoảng 150-170 từ tiếng Việt) lân cận 16 giây. Như vậy, với giá trị $RMSE_{tb}$ bằng 1,287 trên, các giá trị dự đoán đã chênh lệch lân cận 10% với giá trị thực đọc của người dùng.

5. KẾT LUẬN

Bài báo giới thiệu một giải pháp xây dựng hệ thống khuyến nghị thông tin TBT dựa vào phản hồi tiềm ẩn từ người dùng để đề xuất thông tin TBT mà có thể họ quan tâm tìm hiểu. Hệ thống được xây dựng hoàn thiện có tích hợp giải thuật khuyến nghị vào hệ thống và luôn thu thập phản hồi người dùng (nếu có), cuối cùng đánh giá hiệu quả hệ thống dựa trên phản hồi này. Thực nghiệm cho thấy giải pháp đề trên xuất hoàn toàn có thể tích hợp vào các điểm truy cập TBT cấp tỉnh thành. Các nghiên cứu trong tương lai có thể triển khai các hướng sau:

- Tiến hành thử nghiệm trên hệ thống tiếng Anh cho các đối tượng khác nhau (cá nhân, tổ chức nước ngoài) và kéo dài khoảng thời gian thực nghiệm cũng như đánh giá.
- Thu thập thêm dữ liệu TBT cũng như so sánh với các giải pháp khác nhau để tăng cường chất

lượng khuyến nghị.

- Kết hợp lọc nội dung với lọc cộng tác và áp dụng thêm phương pháp học sâu (Deep Learning).

LỜI CẢM ƠN

Nghiên cứu này được Trường Đại học Quốc tế Hồng Bàng cấp kinh phí thực hiện dưới mã số đề tài GVTC 17.39.

TÀI LIỆU THAM KHẢO

[1] E. Dayton, "Amazon Statistics You Should Know: Opportunities to Make the Most of America's Top Online Marketplace". https://www.bigcommerce.com/blog/amazon-statistics/(ngày truy cập cuối 10-8-2024).

[2] Technical barriers to trade" wto.org. https://www.wto.org/english/tratop_e/tbt_e/tbt_e.htm; https://eping.wto.org/en/Search/Index (ngày truy cập cuối 10-8-2024).

[3] Tổng cục thống kê, "Số liệu xuất nhập khẩu các tháng năm 2023" gso.gov.vn. https://www.gso.gov.vn/du-lieu-va-so-lieu-thong-ke/2023/03/so-lieu-xuat-nhap-khau-cac-thang-nam-2023/ (ngày truy cập cuối 10-8-2024).

[4] F.O Isinkaye, Y.O. Folajimi and B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation", Egyptian Informatics Journal, 16(3), p.261-273, 2015.

[5] J. Umair, S. Kamran and L. Suhuai, "A Review of Content-Based and Context-Based Recommendation Systems", International Journal: Emerging Technologies in Learning, 2021.

[6] A. Chaudhari, H. Alhussian and Roshani Raut, "A Hybrid Recommendation System: A Review", IEEE Access, DOI: 10.1109/ACCESS.2024.3480693, 2024.

[7] M. Beegum and A. S, R. Vijayan, "Synonym Insensitive Searching: A Novel Synonym Weighted-Vector Space Model for Document Retrieval", International Conference on Computational Systems and Communication (ICCSC), 2nd in 2023.

[8] I. Abdul, "Nonnegative Matrix Factorization: A Review," Recent Research Reviews Journal, 2 (2), 324-342, 2023.

[9] L. Gao, Y. Liu, Q. Chen,…and Yan Wang, "A user-knowledge vector space reconstruction model for the expert knowledge recommendation system", Information Sciences, Volume 632, June 2023, Pages 358-377.

[10] N. Liu and J. Zhao, "Recommendation System Based on Deep Sentiment Analysis and Matrix Factorization", in IEEE Access, vol. 11, pp. 16994-17001, 2023, doi: 10.1109/ACCESS.2023.3246060.

Building a hybrid recommendation system applied to website of access point for information technical barriers to trade (TBT)

Nguyen Minh De, Le Van Hanh and To Hoai Viet

ABSTRACT

The problem of meeting customer needs for products, services is one of the most important foundations of the supplier. The supplier uses many methods to try to make product and service recommendations suitable for each customer, and the customer interacts with interesting products and services. The supplier often stores user information as well as all transaction histories to serve more appropriate requests for next time. This paper presents a hybrid recommendation system for Technical Barriers to Trade (TBT) information based on implicit feedback and applies it to the website of TBT access point. This hybrid recommendation system is built using a combination of content-based filtering techniques and collaborative filtering techniques corresponding to two methods: Vector Space Model (TF-IDF); Matrix Factorization. The paper has built a database collected from many sources and installed the above hybrid

recommendation system for the website of TBT access point and combined it with TBT information collected from many sources. The system was tested with a data set stored in the above database, showing that the solution is perfectly suitable for integrating into TBT access points.

Keywords: *hybrid recommendation system, content-based, collaborative filtering, vector space model, TF-IDF, matrix factorization*

Received: 28/08/2024
Revised: 30/10/2024
Accepted for publication: 19/11/2024