GIẢI PHÁP KHUYẾN NGHỊ TIN TỨC TRÊN CỔNG THÔNG TIN ĐIỆN TỬ DỰA THEO PHIÊN

Nguyễn Hoàng Anh, Ngô Xuân Bách*
Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Cổng thông tin hay cổng thông tin điện tử (Web portal) là một hoặc một nhóm trang web mà từ đó người truy cập có thể dễ dàng truy xuất các trang web và các dịch vụ thông tin khác trên mạng máy tính. Cổng thông tin điện tử cần đảm bảo các tính năng về cung cấp nội dung, phân loại nội dung và cá nhân hóa nội dung. Một trong những phương pháp để tăng cường giới thiệu nội dung đến người đọc và cá nhân hóa trên cổng thông tin điện tử là dùng hệ khuyến nghị. Hệ khuyến nghị theo các phương pháp truyền thống thường sử dụng dữ liệu được thu thập trong khoảng thời gian dài của người dùng định danh. Tuy nhiên, trên cổng thông tin điện tử khó thực hiện điều đó vì đa số người dùng ẩn danh, do đó khó áp dụng các phương pháp truyền thống. Một giải pháp cho vấn đề này là giải pháp khuyến nghị dựa trên phiên hoạt động của người dùng, trong đó dữ liệu là một chuỗi các hoạt động tuần tự của người dùng trong một khoảng thời gian xác định gọi là phiên. Trong nghiên cứu này, chúng tôi giải quyết bài toán khuyến nghị tin tức trên cổng thông tin điện tử Học viện Công nghệ Bưu chính Viễn thông dựa trên dữ liệu phiên của người dùng ẩn danh. Chúng tôi nghiên cứu và đánh giá nhóm các phương pháp khuyến nghị trên dữ liệu phiên phổ biến hiện nay bao gồm nhóm cơ bản, nhóm K láng giềng gần nhất, nhóm học sâu. Kết quả cho thấy các phương pháp thuộc nhóm học sâu có kết quả tốt hơn so với các phương pháp học máy thuộc nhóm cơ bản trong khi đó các phương pháp thuộc nhóm cơ bản và nhóm K láng giềng cho kết quả tương đối tốt trong khi thời gian chạy và tài nguyên sử dụng ít hơn hẳn.

Từ khóa: Cổng thông tin điện tử; Hệ khuyến nghị; Khuyến nghị theo phiên.

I. GIỚI THIỆU

Cổng thông tin điện tử là các trang web phong phú về thông tin, thu thập nhiều thông tin hữu ích từ các nguồn khác nhau vào một trang Web "một cửa" duy nhất và cung cấp nó ở dạng nhỏ gọn và dễ cung cấp thông tin cho người dùng cuối [1]. Theo Marjan Mansourvar và các cộng sự [2], cổng thông tin khác trang web thông thường ở những đặc điểm: cổng thông tin lấy người dùng làm trung tâm, có nghĩa là có khả năng cá nhân hóa theo người dùng, cho phép người dùng tự tổ chức thông tin; người dùng và cổng thông tin có thể tương tác qua lại; thông tin trên cổng được cập nhật thường xuyên bởi chủ sở hữu. Vấn đề đáp ứng các yêu cầu đó của cổng thông tin được đặt ra bởi các nhà nghiên cứu và các nhà phát triển hệ thống cổng thông tin. Một trong những giải pháp được đề cập đến là áp dụng hệ khuyến nghị vào trong cổng thông tin điện tử.

Hệ khuyến nghị (Recommender system) là hệ thống cho phép dự đoán trước đánh giá của người dùng đối với một mục tin hoặc một nội dung nào đó [4]. Hệ khuyến nghị là một trong những cách tiếp cận tiên tiến nhất, phổ biến trong thương mại và trong cộng đồng nghiên cứu [4]. Nhiều cổng thông tin điện tử đang sử dụng hệ khuyến nghị để tăng lượng độc giả của họ và cung cấp cho độc giả những nội dung tốt hơn. Hệ khuyến nghị học từ hành vi đọc tin, xếp hạng và nhận xét của người dùng, sau đó quyết định điểm số nhờ sự trợ giúp của hệ thống. Theo cách phân chia truyền thống, hệ khuyến nghị được chia thành các phương pháp: khuyến nghị dựa trên lọc cộng tác [5], khuyến nghị dựa trên nội dung [6], và hệ khuyến nghị lai. Phương pháp lọc cộng tác sử dụng ma trận đánh giá của người dùng đối với sản phẩm để dự đoán đánh giá của người dùng cho những sản phẩm chưa được đánh giá trong khi phương pháp lọc dựa trên nội dung khuyến nghị các sản phẩm có nội tương tự với tin mục người dùng đã thích hoặc tin tức người dùng đã đọc mà không phụ thuộc vào đánh giá của người dùng khác về tin mục đó. Phương pháp khuyến nghị lai [7] là phương pháp kết hợp các phương pháp trên để đưa ra sản phẩm được khuyến nghị cho người dùng. Các phương pháp khuyến nghị kể trên chủ yếu sử dụng dữ liệu người dùng và tin mục được lưu lâu dài trong hệ thống.

Trong nhiều lĩnh vực ứng dụng của hệ khuyến nghị, các mô hình người dùng dài hạn thường không có sẵn cho phần lớn người dùng, ví dụ người dùng lần đầu ghé website hoặc cổng thông tin, người dùng lần đầu mua hàng, hoặc người dùng không đăng nhập vào hệ thống. Do đó, các đề xuất phù hợp phải được xác định dựa trên các loại thông tin khác, thường là các tương tác gần đây nhất của người dùng với cổng thông tin điện tử theo một cách tuần tự trong một khoảng thời gian hạn định gọi là phiên. Các phương pháp khuyến nghị dựa trên phiên là phương án giải quyết hiệu quả bài toán khuyến nghị cho người dùng ẩn danh hoặc người dùng mới khi hệ thống chưa có hoặc không có thông tin lâu dài của người dùng.

Cổng thông tin Học viện Công nghệ Bưu chính Viễn thông (www.ptit.edu.vn) là nơi cung cấp thông tin về các hoạt động đào tạo, công tác giáo vụ, nghiên cứu khoa học, tuyển sinh đến sinh viên, đến cán bộ và đến những đối

Tác giả liên hệ: Ngô Xuân Bách,
Email: bachnx@ptit.edu.vn
Đến tòa soạn: 9/2022, chỉnh sửa: 10/2022, chấp nhận đăng: 12/2022.

tượng quan tâm đến hoạt động nghiên cứu đào tạo của Học viện. Người đọc truy cập vào cổng thông tin Học viện Công nghệ Bưu chính Viễn thông để đọc, theo dõi và tìm kiếm thông tin. Đa số người dùng trên cổng thông tin điện tử đều không có tài khoản mà thuộc về đối tượng người dùng ẩn danh (chỉ được xác định qua thiết bị truy cập). Bài toán đặt ra là khuyến nghị tin bài phù hợp cho người đọc trong trường hợp người dùng mới vào cổng thông tin và không có tài khoản trên hệ thống.

Nghiên cứu này đề xuất giải pháp khuyến nghị tin tức cho người dùng ẩn danh trên cổng thông tin điện tử của Học viện Công nghệ Bưu chính Viễn thông. Các nhóm phương pháp khuyến nghị tập trung vào xác định sở thích ngắn hạn của người dùng thông qua chuỗi tuần tự các hoạt động của người dùng trong một khoảng thời gian ngắn xác định gọi là một phiên. Thông qua phân tích về người dùng ẩn danh đó, các phương pháp khuyến nghị dựa trên phiên hiện tại của người dùng sẽ gợi ý tin tức tiếp theo mà người dùng ẩn danh nên đọc trong phiên. Nhóm các phương pháp khuyến nghị dựa trên phiên là nhóm các phương pháp phổ biến hiện nay bao gồm nhóm các phương pháp cơ bản, nhóm các phương pháp sử dụng K láng giềng gần nhất và nhóm các phương pháp sử dụng học sâu. Kết quả thực nghiệm cho thấy khả năng khuyến nghị tin tức phù hợp cho người dùng ẩn danh trên cổng thông tin điện tử với độ chính xác khá cao.

Phần tiếp theo của bài báo được cấu trúc theo các mục bao gồm: Mục II các nghiên cứu liên quan; mục III trình bầy nội dung về cổng thông tin điện tử và bài toán khuyến nghị trên cổng thông tin điện tử cho dữ liệu dựa trên phiên; mục IV đưa ra phân tích về các nhóm phương pháp khuyến nghị dựa trên dữ liệu phiên phổ biến hiện nay bao gồm nhóm cơ bản, nhóm K láng giềng gần nhất và nhóm dựa trên các phương pháp học sâu. Kết quả thực nghiệm được thực hiện ở mục IV thể hiện kết quả khuyến nghị trên dữ liệu cổng thông tin điện tử đã thu thập được. Nhận xét về kết quả thực nghiệm, đánh giá giữa các phương pháp được đưa ra trong mục VI.

II. CÁC NGHIÊN CỨU LIÊN QUAN

Cổng thông tin điện tử là đối tượng có trong các nghiên cứu hàn lâm. Từ những năm 1970, Yu và các cộng sự [1] đã chỉ ra kiến trúc, đặc điểm của cổng thông tin điện tử, đó chính là cửa ngõ, là nơi trên cùng (on-top) mà người dùng có thể tìm kiếm và truy xuất thông tin từ các nguồn khác nhau ở sau cánh cổng đó. Đến những năm 2000, các nghiên cứu về cổng thông tin điện tử tập trung vào các đặc tính của cổng thông tin điện tử với tương tác của người dùng trên cổng. Mansourvar và các cộng sự [2] chỉ ra các loại cổng thông tin điện tử khác nhau và các đặc tính của cổng thông tin điện tử như: khả năng tìm kiếm, khả năng cung cấp thông tin và khả năng tùy biến hóa theo cá nhân người dùng. Các yêu cầu cần đáp ứng trong việc xây dựng và triển khai cổng thông tin điện tử cũng được các nhóm nghiên cứu đưa ra, trong đó nhấn mạnh các yêu cầu đáp ứng thông tin, tìm kiếm thông tin, nhất là trong những hệ thống cổng thông tin điện tử có lượng thông tin lớn và được phân cấp sâu [8][9]. Một trong những ứng dụng phổ biến cho cổng thông tin điện tử đáp ứng yêu cầu cung cấp thông tin tới người dùng một cách chính xác và phổ biến chính là ứng dụng hệ khuyến nghị [4].

Hệ khuyến nghị là giải pháp phù hợp cho các hệ thống như cổng thông tin khi số lượng thông tin và số lượng người dùng có rất nhiều. Hệ khuyến nghị trong trường hợp này sẽ giúp người dùng lựa chọn các thông tin phù hợp với sở thích khi lượng thông tin bùng nổ [3]. Hệ khuyến nghị theo các phương pháp truyền thống được phân chia thành hệ khuyến nghị dựa trên lọc cộng tác, hệ khuyến nghị dựa trên nội dung và hệ khuyến nghị lai [3].

Các phương pháp vừa nêu đều hoạt động chủ yếu trên mô hình dữ liệu lâu dài của người dùng và mô hình tương tác lâu dài giữa người dùng và tin mục. Tuy nhiên, với các đặc điểm của cổng thông tin điện tử như đã nêu, mô hình dữ liệu người dùng lâu dài thường khó khả thi. Các phương pháp để giải quyết vấn đề này có đề cập đến sử dụng mô hình ngắn hạn của người dùng hay những tương tác trong ngắn hạn của người dùng đối với tin mục trong cổng thông tin điện tử gọi là phiên. Hầu hết các cách tiếp cận cho khuyến nghị dựa trên phiên trong bài báo đều dựa theo cách học trình tự [10]. Các phương pháp tiếp cận ban đầu dựa trên việc xác định các mẫu tuần tự thường xuyên, có thể được sử dụng vào thời điểm cần đề xuất để dự đoán hành động tiếp theo của người dùng. Sau đó, các kỹ thuật khai phá mẫu như vậy cũng được sử dụng cho các vấn đề khuyến nghị mục tiếp theo trong thương mại điện tử hoặc miền âm nhạc.

Trong những nghiên cứu gần đây, có nhiều cách tiếp cận học dữ liệu tuần tự phức tạp hơn được nghiên cứu và triển khai. Các cách tiếp cận để mô hình hóa dữ liệu tuần tự thường dựa trên mô hình chuỗi Markov [11][12] hoặc mạng nơ ron hồi tiếp [13][14][15]. Các nghiên cứu này được ứng dụng trong các lĩnh vực thương mại điện tử và âm nhạc. Một trong các tiếp cận sớm nhất dựa trên quá trình quyết định dựa trên chuỗi Markov được đưa ra bởi Shani và các cộng sự [16]. Nghiên cứu này chỉ ra giá trị của việc sử dụng dữ liệu tuần tự trong lĩnh vực thương mại điện tử nhưng cũng cho thấy rằng mô hình dựa trực tiếp trên chuỗi Markov không thể áp dụng trực tiếp do sự thưa thớt về mặt dữ liệu Do đó, nghiên cứu cũng đề cập tới một vài phương pháp mẹo để vượt qua vấn đề đó. Một thách thức khác khi sử dụng dạng mô hình này là sử dụng bao nhiêu tương tác trước đó để đoán tương tác tiếp theo. Vài tác giả sử dụng hỗn hợp mô hình Markov biến thiên (VMM) và mô hình cây ngữ cảnh để giải quyết vấn đề chiều dài chuỗi biến thiên [16][17].

Những nghiên cứu gần đây nhất cho mô hình hóa dữ liệu tuần tự dựa trên RNN. Zhang và các cộng sự [18] sử dụng RNN để dự đoán click chuột tiếp theo của người dùng trong lĩnh vực quảng cáo. Hidasi và các cộng sự là trong những nhóm đầu tiên sử dụng đơn vị hồi tiếp có cổng (GRU4REC) [13] là một dạng đặc biệt của RNN để khuyến nghị hành động tiếp theo của người dùng. Phương pháp này được gọi là GRU4REC sau đó được mở rộng theo nhiều cách khác nhau trong [19] và [20]. Bên cạnh đó, Jannach và các cộng sự [21] chỉ ra rằng các phương pháp khuyến nghị cho dữ liệu tuần tự dựa trên K láng giềng gần nhất cho kết quả cạnh tranh với nhóm GRU4REC. Trong nghiên cứu đó, nhóm thực hiện thực nghiệm với biến thể mới nhất của GRU4REC và một số biến thể của phương pháp K láng giềng gần nhất.

Trong bài báo này, chúng tôi tiếp cận dựa trên phân tích đặc điểm cốt lõi của cổng thông tin điện tử là người dùng ẩn danh nhưng vẫn đảm bảo được tính riêng, tùy biến cho mỗi người dùng để đề xuất giải pháp khuyến

nghị. Nghiên cứu của chúng tôi tập trung đề xuất giải pháp khuyến nghị tin tức cho cổng thông tin điện tử của Học viện Công nghệ Bưu chính Viễn thông trên tập dữ liệu tương tác của người dùng ẩn danh thu thập được bằng những phương pháp và nhóm phương pháp phổ biến và tiên tiến hiện nay.

III. KHUYẾN NGHỊ TRÊN CỔNG THÔNG TIN ĐIỆN TỬ

A. *Cổng thông tin điện tử*

Cổng thông tin điện tử ngày càng có ảnh hưởng nhiều trong cuộc sống trên khía cạnh truyền thông, chia sẻ thông tin, tìm kiếm thông tin, các hoạt động xã hội.... Cổng thông tin điện tử chính là cửa ngõ (gateway) để người dùng qua đó truy cập các dịch vụ thông tin, các sản phẩm dịch vụ của tổ chứ, của doanh nghiệp. Các thông tin, sản phẩm và dịch vụ trên cổng thôn tin điện tử là rất nhiều và được cập nhật thường xuyên liên tục. Yang và các cộng sự [22] chia dịch vụ cổng thông tin thành ba nhóm dịch vụ khác nhau bao gồm:

- Search: Dịch vụ tìm kiếm là dịch vụ rất quan trọng của cổng thông tin, dịc vụ được thực thi bởi các phương pháp tìm kiếm của cổng thông tin hoặc có thể được hỗ trợ từ bên ngoài (như google).

- Information: cổng thông tin cung cấp nhiều dạng thông tin khác nhau cho người dùng như tin tức, thông báo, hoạt động, tuyển dụng... Người dùng có thể trực tiếp truy cập các loại thông tin này mà không cần có tài khoản và mật khẩu mà chỉ cần thao tác đơn giản là nhấn vào đường liên kết (link) trên cổng thông tin. Để thực hiện được việc này, cổng thông tin cần được chia thành các mục (categories) và các tin mục/sản phẩm cần được người quản trị phân vào các mục đã được phân chia trước đó.

- Personal Service: Dịch vụ này cho phép người dùng tùy biến theo mong muốn trên cổng thông tin. Để thực hiện được dịch vụ này, thông thường người dùng phải đăng ký trên cổng thông tin bằng tài khoản và mật khẩu. Khi đăng nhập tài khoản và mật khẩu riêng, người dùng trên cổng thông tin điện tử có thể nhận được các nội dung tùy biên như chat, email, các dịch vụ được phép sử dụng hoặc có thể tùy biến giao diện theo sở thích cá nhân. Tuy nhiên, việc tùy biến này làm giảm tốc độ truy cập đáng kể do phải xem xét tài khoản và phân quyền trước khi hiển thị và cung cấp dịch vụ.

Mong muốn của hệ thống cổng thông tin điện tử là cung cấp thông tin phù hợp với người dùng, kể cả người dùng không tạo tài khoản một cách nhanh chóng và đúng với mối quan tâm của người dùng nhất, từ đó tăng lưu lượng truy cập và thời gian sử dụng của người dùng.

B. *Bài toán*

Như đã đề cập ở phần trên, việc thấy được thông tin phù hợp với nhu cầu của người dùng trên cổng thông tin điện tử là tương đối khó khăn vì lượng thông tin trên cổng thường rất nhiều và được cập nhật mới liên tục. Hơn nữa, trên cổng thông tin điện tử về tin tức, người dùng thường không có xu hướng tạo tài khoản đăng nhập mà thường có xu hướng chỉ đọc (không tạo tài khoản hay ẩn danh). Do đó việc nắm bắt sở thích lâu dài của người dùng trên cổng thông tin điện tử là khó khả thi.

Một trong những phương pháp phổ biến để tăng cường thông tin hữu ích cho người dùng trên cổng thông tin điện tử chính là áp dụng hệ khuyến nghị [4]. Tuy nhiên, phương pháp được đề cập dựa trên mô hình mối quan tâm tương tự của cá nhân người dùng hoặc theo nhóm người dùng sử dụng dữ liệu người dùng được ghi nhận lâu dài trên hệ thống. Ngoài ra, mối quan tâm của người dùng thay đổi theo thời gian, do đó, nếu sử dụng mô hình mối quan tâm dài hạn của người dùng sẽ không phù hợp với sở thích ngắn hạn của người dùng. Bài toán đặt ra ở đây là cần xác định được sở thích ngắn hạn của người dùng ẩn danh trên cổng thông tin điện tử và đưa ra khuyến nghị về hành động tiếp theo cho người dùng dựa trên chuỗi các hành động đang diễn ra của người dùng đó. Bài toán được đặt ra như dưới đây.

Đầu vào:

- Phiên của người dùng bao gồm:

$l_v = \{v_1, ..., v_j, ...v_n\} (v_j \in V)$

trong đó $v_j \in V$ là tập tin tức trong tập tin tức V.

- Tập tất cả các phiên L

- Ngữ cảnh c trong tập ngữ cảnh C.

Đầu ra:

- Danh sách một hoặc nhiều tin tức được khuyến nghị thỏa mãn điều kiện:

$\hat{l} = \arg\max f(c, l), c \in C, l \in L$

IV. CÁC PHƯƠNG PHÁP KHUYẾN NGHỊ

Wang và các cộng sự [23] đã tổng hợp trong một khảo sát về bài toán khuyến nghị dựa trên phiên (Session-based Recommender Systems) bao gồm mô tả về phiên, bài toán và nhóm các phương pháp khuyến nghị dựa trên dữ liệu phiên. Trong báo cáo [24], Ludewig và các cộng sự cũng đã tiến hành thực nghiệm nhóm các phương pháp khuyến nghị dựa trên dữ liệu phiên cho các tập dữ liệu tiêu chuẩn như 8TRACKS, AOTM, NOWPLAYING, RSC15, TMALL. Từ việc phân tích các kết quả nghiên cứu, nhóm nghiên cứu phân các phương pháp khuyến nghị theo ba nhóm như sau: Nhóm cơ bản bao gồm các phương pháp dựa trên quy tắc xuất hiện đồng thời của các mục trong phiên, các phương pháp trong nhóm này bao gồm Luật kết hợp đơn giản, Chuỗi Markov, Luật tuần tự và Cây ngữ cảnh; nhóm K láng giềng dựa trên sự tương đồng giữa các phiên trong tập dữ liệu, các phương pháp trong nhóm bao gồm K láng giềng dựa trên phiên-SKNN và các biến thể của phương pháp này như V-SKNN, S-SKNN, SF-SKNN, STAN, V-STAN; Nhóm các phương pháp dựa trên các phương pháp học sâu. Trong các phương pháp học sâu, phương pháp GRU4Rec 2016 là phương pháp cơ sở, hầu hết các phương pháp khác trong tập xem xét đều phát triển từ phương pháp này.

A. *Nhóm cơ bản*

1) *Phương pháp luật kết hợp đơn giản (AR)*

Luật kết hợp đơn giản là phiên bản đơn giản của kỹ thuật khai phá luật kết hợp với luật lớn nhất của 2 [24]. Phương pháp được thiết kế để đo độ thường xuyên cùng xẩy ra của hai sự kiện.

Đầu vào:

- Tập phiên trong quá khứ: S_p

- Một phiên s hiện tại của người dùng

Đầu ra:

Top k mục tin i phù hợp với hành động tiếp theo của người dùng tại phiên s.

$s=(s_1,s_2...,s_m)$: Một phiên trong tập các phiên, bao gồm các phần tử (các mục) $s_1,s_2...,s_m$.

S_p: là tập các phiên trong quá khứ

$s_{|s|}$: mục cuối cùng trong phiên s

Cách tính điểm của một mục i với một phiên s:

$$score_{AR}(i,s)=A*B$$

$$A=\frac{1}{\sum_{p\in S_p}\sum_{x=1}^{|p|}1_{EQ}(s_{|s|},p_x)(|p|-1)} \quad (1)$$

$$B=\sum_{p\in S_p}\sum_{x=1}^{|p|}\sum_{y=1}^{|p|}1_{EQ}(s_{|s|},p_x)1_{EQ}(i,p_y)$$

Trong đó $1_{EQ}(a,b)$ có giá trị bằng 1 nếu a bằng b, và có giá trị bằng 0 nếu ngược lại.

2) Phương pháp chuỗi Markov (Markov Chain-MC)

Chuỗi markov, được coi là biến thể của phương pháp AR, tập trung vào sự nối tiếp trong dữ liệu. Luật của chuỗi theo xác suất chuyển giữa hai sự kiện con trong phiên. Phương pháp xác định số lần sự kiện mục q xuất hiện ngay sau mục p, càng nhiều lần thì trọng số của cặp p,q càng cao.

$$score_{MC}(i,s)=A*B$$

$$A=\frac{1}{\sum_{p\in S_p}\sum_{x=1}^{|p|-1}1_{EQ}(s_{|s|},p_x)} \quad (2)$$

$$B=\sum_{p\in S_p}\sum_{y=1}^{|p|-1}1_{EQ}(s_{|s|},p_x).1_{EQ}(i,p_{x+1})$$

Hàm $1_{EQ}(s_{|s|},p_x).1_{EQ}(i,p_{x+1})$ sẽ kiểm tra và chỉ tính điểm khi trong phiên đang xét có tồn tại mục i ngay sau mục $s_{|s|}$. Sau khi tính điểm, phần khuyến nghị giống như phương pháp luật kết hợp đơn giản.

3) Luật tuần tự (Sequence Rules – SR)

Phương pháp luật tuần tự (SR) có thể coi luật tuần tự là biến thể của phương pháp AR và phương pháp MC. Phương pháp này để ý đến thứ tự xuất hiện của các mục trong chuỗi nhưng kém chặt chẽ hơn. Phương pháp này xem xét mục q xuất hiện sau mục p trong một phiên ngay cả khi có các mục giữa p và q.

$$score_{SR}(i,s)=A*B$$

$$A=\frac{1}{\sum_{p\in S_p}\sum_{x=2}^{|p|}1_{EQ}(s_{|s|},p_x).x} \quad (3)$$

$$B=\sum_{p\in S_p}\sum_{x=2}^{p}\sum_{y=1}^{x-1}1_{EQ}(s_{|s|},p_y).1_{EQ}(i,p_x).w_{SR}(x-y)$$

$w_{SR}(x)=1/(x)$ trong đó x là số bước (số mục) từ phiên p tới phiên q.

4) Phương pháp Cây ngữ cảnh (Context tree)

Phương pháp phi tham số dựa trên cấu trúc gọi là cây ngữ cảnh [25]. Phương pháp cây ngữ cảnh được đề ra ban đầu cho việc nén không mất mát thông tin. Đây là phương pháp phi tham số dựa trên mô hình Markov có các biến theo thứ tự.

B. Nhóm K láng giềng gần nhất

Nhóm K láng giềng gần nhất được đặc trưng bởi việc tìm tập hợp k phiên tương đồng nhất so với phiên đang xét [26]. Sau đó thông qua tập k phiên này để khuyến nghị hành động tiếp theo của người dùng tại phiên s đang được xem xét. Cho một sự kiện s tập phiên trong quá khứ S_p, cần tìm mục i phù hợp nhất với hành động tiếp theo của người dùng tại phiên s.

1) Phương pháp SKNN

Thay vì chỉ quan tâm đến hành động cuối cùng trong phiên hiện tại s, phương pháp SKNN so sánh toàn bộ phiên hiện tại (bao gồm tất cả các mục tin) với các phiên trong quá khứ trong tập dữ liệu huấn luyện để xác định mục sẽ khuyến nghị.

Về phương pháp thực hiện, với một phiên s, đầu tiên xác định k phiên hàng xóm gần nhất với phiên gọi là tập N_s bằng cách áp dụng độ đo tương tự của phiên khi phiên được mã hóa thành vector nhị phân trên không gian tất cả các mục tin. Sau đó, tính điểm của mục tin i với phiên s theo công thức:

$$score_{SKNN}(i,s)=\sum_{n\in N_s}sim(s,n).1_n(i) \quad (4)$$

Với $1_n(i)$ =1 nếu phiên n (trong tập k hàng xóm của phiên s) chứa i và bằng 0 nếu ngược lại.

Sau đó, điểm của các mục i được sắp xếp theo thứ tự từ lớn đến nhỏ và chọn ra top k mục để khuyến nghị.

2) Nhóm SKNN mở rộng

Nhóm SKNN mở rộng bao gồm các phương pháp V-SKNN, S-SKNN và SF-SKNN. Theo mô tả, phương pháp SKNN không tính tới thứ tự của các tin mục trong một phiên khi sử dụng chỉ số Jaccard hoặc độ tương tự Cosine để đo khoảng cách. Do đó, thứ tự của các tin mục trong một phiên có thể là một yếu tố có thể xem xét. Theo đó sở thích của người dùng có thể thay đổi trong một phiên, phụ thuộc vào những tin mục họ đã xem. Nhóm SKNN mở rộng gồm các phương pháp như sau:

Phương pháp VSKNN (Vector Multiplication Session-Based kNN): ý tưởng của phương pháp này là nhấn mạnh vào những tin mục gần hơn khi tính toán độ tương tự.

Thay vì mã hóa một phiên theo dạng nhị phân, phương pháp chỉ mã hóa tin mục cuối cùng thành trong chuỗi 1, trọng số của các tin mục khác trong chuỗi được mã hóa thành những số khác tùy thuộc vào vị trí trong chuỗi theo một hàm giảm tuyến tính.

Phương pháp S-SKNN (Sequential Session-based kNN): Biến thể này đặt thêm trọng số vào phần tử xuất hiện phía sau trong phiên. Công thức tính điểm như sau:

$$score_{S-SKNN}(i,s) = \sum_{n \in N_s} sim(s,n).\mathrm{w}_n(s).1_n(i) \qquad (5)$$

với $\mathrm{w}_n(s) = x/|s|$

Phương pháp tính thứ tự của mục trong phiên hàng xóm gần nhất thông qua trọng số $\mathrm{w}_n(s)$. Trong tất cả các phiên hàng xóm của phiên s có chứa mục i, vị trí của mục i càng về cuối phiên thì điểm trọng số của mục i với phiên s càng cao.

Phương pháp SF-SKNN: cũng sử dụng hàm tính điểm của mục i với các phiên hàng xóm gần nhất nhưng theo cách chặt chẽ hơn. Phương pháp chỉ tính điểm của mục tin khi mục này xuất hiện ngay sau mục là mục cuối cùng của phiên đang xét s trong các phiên hàng xóm của phiên s.

$$score_{SF-SKNN}(i,s) = \sum_{n \in N_s} sim(s,n).1_n(s_{|s|},i) \qquad (6)$$

Hàm $1_n(s_{|s|},i)$ chỉ bằng 1 nếu tồn tại chuỗi $(s_{|s|},i)$ trong phiên đang xét.

3) STAN

Phương pháp STAN được giới thiệu tại SIGIR'19 [27]. Phương pháp này dựa trên SKNN nhưng xem xét bổ sung những yếu tố sau cho việc khuyến nghị:

i) vị trí của một mục trong phiên hiện tại.

ii) khoảng cách của các phiên quá khứ với phiên hiện tại.

iii) vị trí của mục có thể được khuyến nghị trong phiên hàng xóm gần nhất.

4) VSTAN

Phương pháp VSTAN là ý tưởng kết hợp của STAN và V-SKNN theo cách tiếp cận đơn lẻ. Phương pháp này kết hợp tất cả ba đặc điểm đã đề cập trên đây của STAN, vốn đã có một số điểm tương đồng với phương pháp V-SKNN. Thêm vào đó, việc tính giá trị các mục bằng số thực có áp dụng giá trị giảm dần theo hàm tuyến tính từ V-SKNN sẽ được áp dụng tại VSTAN

C. Nhóm học sâu

1) Phương pháp GRU4REC

Sử dụng RNNs để mô hình hóa chuỗi người dùng cho hệ khuyến nghị dựa trên phiên [13]. Phương pháp này sử dụng Gated Recurrent Units (GRU) để xử lý vấn đề phạt gradient. Sau đó, kỹ thuật này được cải thiện bằng cách sử dụng các hàm mất mát hiệu quả (CIKM '18).

$$Top1: L_s = -\frac{1}{N_S}\sum_{j=1}^{N_S}\sigma(p_j - p_i) + \sigma(p_j^2) \qquad (7)$$

$$BPR: L_s = -\frac{1}{N_S}\sum_{j=1}^{N_S}\log(\sigma(p_j - p_i)) \qquad (8)$$

2) Phương pháp NARM

Mô hình này mở rộng GRU4REC và cải thiện mô hình hóa phiên của nó với việc giới thiệu bộ mã hóa hỗn hợp với cơ chế chú ý (attention mechanism) [28]. Cơ chế chú ý đặc biệt được sử dụng để xem xét các mục xuất hiện trước đó trong phiên và tương tự với mục được nhấp cuối cùng. Điểm cho mỗi mục ứng viên được tính với một sơ đồ khớp song tuyến dựa trên việc biểu diễn phiên thống nhất.

3) Phương pháp STAMP

Ngược lại với NARM, phương pháp STAMP không dựa trên RNN. Mô hình chú ý ngắn hạn/ưu tiên bộ nhớ được đưa ra có khả năng các mối quan tâm của người dùng nói chung từ bộ nhớ dài hạn và cũng cân nhắc tới những mối quan tâm hiện tại của người dùng từ bộ nhớ ngắn hạn [29]. Mối quan tâm nói chung của người dùng được ghi ở bộ nhớ ngoài được xây dựng từ tất cả các click của người dùng trong quá khứ (bao gồm cả các click gần đây). Cơ chế chú ý được xây dựng trên các click gần đây nhất của người dùng, thể hiện mối quan tâm hiện tại.

4) Phương pháp NEXTITREC

Mô hình này sử dụng mạng CNN thay vì mạng RNN như các mô hình học sâu khác [30]. Mô hình sinh (generative model) được thiết kế để mã hóa tường minh sự phụ thuộc giữa các mục mà cho phép đánh giá trực tiếp phân phối của chuỗi đầu ra.

5) Phương pháp SR-GNN

Phương pháp SR-GNN [31] mô hình hóa các chuỗi phiên thành dữ liệu đồ thị có cấu trúc. Dựa trên đồ thị phiên, SR-GNN có thể ghi lại quá trình chuyển đổi giữa các mục và tạo ra các vector nhúng tương ứng của các mục – là điều khó được phát hiện bởi các phương pháp chuỗi truyền thống như phương pháp dựa trên chuỗi Markov hay các phương pháp dựa trên RNN.

V. THỰC NGHIỆM VÀ KẾT QUẢ

A. Thu thập và xử lý dữ liệu

Nội dung trong phần này sẽ mô tả cách thức thu thập dữ liệu hành động của người dùng trên cổng thông tin điện tử và từ đó xử lý tạo thành dữ liệu phiên hoạt động của người dùng.

1) Mô tả dữ liệu

Dữ liệu được thu thập từ Cổng thông tin điện tử Học viện Công nghệ Bưu chính Viễn thông trong thời gian gần 3 tháng (Từ 26/10/2019 tới 19/2/2020) qua script được cài phía máy khách và được lưu vào dữ liệu mongoDB của máy chủ phân tích.

Dữ liệu gồm nhiều thông tin khác nhau của người dùng truy cập vào cổng thông tin điện tử như các thông tin về thiết bị dùng để truy cập cổng thông tin điện tử (máy tính, máy tính bảng, điện thoại), trang tin cụ thể (post/page) trên cổng thông tin điện tử được người dùng click chuột vào đại diện bởi url của trang tin, thời gian click vào và thời gian thoát, các thao tác của người dùng

trên trang tin bao gồm click chuột, lăn chuột, tải tài liệu, vùng được click trên màn hình. Các thông tin được quan tâm trong nghiên cứu này bao gồm các tin bài trên cổng thông tin điện tử, các chuyên mục các nhấn chuột vào bài viết của thiết bị của người dùng đọc tin trên đó. Do người dùng trên cổng thường ẩn danh nên việc xác định người dùng thông qua ID của thiết bị truy cập vào cổng. Dữ liệu bao gồm các trường chính như trong bảng I.

Bảng I. Các trường dữ liệu

Trường dữ liệu	Kiểu dữ liệu	Mô tả dữ liệu
userID	String	ID của thiết bị người dùng sử dụng
itemID	int	ID của URL người dùng nhấn vào
cateID	int	Chuyên mục tin tức
time	unix time	Thời gian người dùng click

2) Thống kê dữ liệu

Dữ liệu được thống kê bởi các loại dữ liệu và số lượng cụ thể như trong bảng II.

Bảng II. Thống kê dữ liệu Tin bài và Người dùng

Loại dữ liệu	Số lượng	Ghi chú
Actions (url click)	100176	Số click vào url
Devices (users)	42668	Số thiết bị truy cập
Items	941	Số URL được click
Số chuyên mục	20	Số chuyên mục của tin tức

3) Tiền xử lý dữ liệu

Tiền xử lý dữ liệu bao gồm các bước:

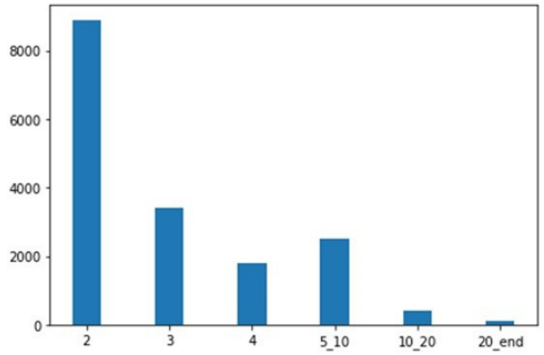
- Lọc dữ liệu cần thiết từ tập dữ liệu thô, trích ra các trường thông tin về userID, urlID, categoryID, timeClick.

- Nhóm các dữ liệu giao dịch theo người dùng, mỗi người dùng là tập các nhấp chuột vào các bài viết.

- Một phiên được xác định là các hoạt động gần nhau trong khoảng thời gian cách nhau không quá 20 phút. Nếu khoảng cách giữa các hoạt động lớn hơn 20 phút thì hoạt động kế tiếp xác định việc bắt đầu một phiên mới. Số nhấp chuột trong phiên nằm trong khoảng từ một tới bảy mươi mốt.

- Các phiên chỉ có một click không mang nhiều ý nghĩa, do vậy, phiên được loại bỏ khỏi tập dữ liệu huấn luyện. Ngoài ra, trong giai đoạn phân chia tập dữ liệu thành train, valid và test, các mục nằm trong tập test mà không nằm trong tập train sẽ bị loại bỏ. Việc loại bỏ cũng diễn ra tương tự với các mục nằm trong tập valid nhưng không nằm trong tập train.

Biểu đồ thể hiện thống kê dữ liệu như hình 1 sau:



Hình 1. Thống kê dữ liệu

Thống kê cho thấy số lượng người dùng click vào chỉ hai phiên chiếm đa số. Các phiên bao gồm nhiều click hơn thì chiếm số lượng ít dần. Có nhóm số ít người dùng duy trì trên cổng từ 10 tới 20 clicks trong một phiên và rất ít người dùng có hơn 20 clicks trong một phiên. Số click chuột trung bình của người dùng là 3.18 clicks trong một phiên.

B. Chia tập dữ liệu và độ đo

Để đảm bảo tính thống kê vào tính khoa học, nhóm nghiên cứu chia tập dữ liệu đã thu thập và xử lý trong phần III thành 5 tập dữ liệu con (PTIT1, PTIT2, PTIT3, PTIT4, PTIT5), mỗi tập được chia theo thời gian xảy ra hoạt động của phiên: Tập huấn luyện 80%, Test 20% theo thứ tự thời gian, trong đó tập Test là tập có thời gian của các phiên gần nhất. Các tập dữ liệu con đều được tiền xử lý giống như phương pháp xử lý trên tập dữ liệu ban đầu. Thống kê tập dữ liệu phiên như trong bảng III:

Bảng III. Chia tập dữ liệu

Tập dữ liệu	Train			Test		
	Events	Sessions	Items	Events	Sesions	Items
PTIT1	9607	2207	498	1838	453	223
PTIT2	9607	2757	514	2267	647	245
PTIT3	9607	2768	437	1955	578	192
PTIT4	9607	2913	402	2198	658	161
PTIT5	9607	2921	316	2359	793	94

Thực nghiệm sử dụng độ đo Hitrate (HR), MRR, Coverage (Cov) và độ phổ biến (POP). Hitrate là độ đo tỉ lệ số khuyến nghị chính xác trên tổng số các khuyến nghị. HR@K tương ứng với số mục nằm trong tập top K mục được khuyến nghị cho một phiên, trên tổng số phiên trong tập test.

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i} \quad (9)$$

MRR tương tự như Hitrate nhưng tính tới thứ tự của mục đúng trong Top k mục được khuyến nghị. Nếu mục đúng nằm tại Top 1 thì giá trị phép tính là 1, nếu không sẽ nhỏ hơn 1.

Độ bao phủ (Coverage-Cov): số lượng các mục tin khác nhau xuất hiện trong top k khuyến nghị. Hệ thống

Bảng IV. Tài nguyên sử dụng

Nhóm	Phương pháp	Training Time (s)	Testing Time (s)	Memory Usage
Nhóm Baseline	AR	0.1788	0.0093	1416872
	MC	0.0632	0.0092	1087728
	SR	0.1060	0.0091	1410152
	CT	3.5326	0.0140	100181040
Nhóm KNN	SKNN	0.0803	0.0142	8498704
	V-SKNN	0.1165	0.0200	9342624
	S-SKNN	0.0834	0.0178	9150968
	SF-SKNN	0.0982	0.0241	10313176
	STAN	0.0927	0.0197	7932168
	VSTAN	0.0967	0.0223	8010872
Nhóm Deep Learning	GRU4REC	303.66	0.0176	3787600
	STAMP	190.29	0.0107	31910832
	NARM	714.84	0.0177	46222768
	Nextitrec	1189.78	0.0178	28985224
	SGNN	346.30	0.0078	113991400

Bảng V. Kết quả lựa chọn siêu tham số

Nhóm	Phương pháp	Hyper-parameter
Cơ bản	AR	Không điều chỉnh
	MC	Không điều chỉnh
	SR	steps: 2, weighting: div
	CT	expert: 'DirichletExpert' history_maxlen: 20
Nhóm KNN	SKNN	k: 1000, sample_size: 5000, similarity: jaccard
	VSKNN	k: 500, sample_size: 1000, weighting: log , weighting_score: quadratic, idf_weighting: 1
	S-SKNN	k: 1000,sample_size: 2500, similarity: cosine
	SF-SKNN	k: 1000, sample_size: 5000, similarity: jaccard
	STAN	k: 500, sample_size: 1000, lambda_spw: 0.00001 , lambda_snh: 80, lambda_inh: 0.905
	VSTAN	k: 500, sample_size: 1000 , similarity: 'cosine', lambda_spw: 3.62, lambda_snh: 100, lambda_inh: 1.81, lambda_ipw: 0.00001, lambda_idf: 1
Nhóm học sâu	GRU4REC	loss: 'top1-max', final_act: 'linear', dropout_p_hidden: 0.5, learning_rate: 0.02, momentum: 0.0, constrained_embedding: True
	STAMP	init_lr: 0.0007, n_epochs: 20, decay_rate: 0.8
	NARM	epochs: 20, lr: 0.007, hidden_units: 100, factors: 50
	Nextitrec	learning_rate: 0.006, iterations: 20, is_negsample: True
	SGNN	hidden_size: 100, out_size: 100, step: 1, nonhybrid: True, batch_size: 100, epoch_n: 10, batch_predict: True, lr: 0.004, l2: 3.00E-05, lr_dc: 0.36666667, lr_dc_step: 3

được đánh giá là khuyến nghị tốt khi khuyến nghị trải rộng trên tập dữ liệu, có nghĩa là *Cov* lớn.

Độ phổ biến (POP): độ đo tần suất các tin tức được hệ thống khuyến nghị trong tập dữ liệu Test. Nếu tần suất cao có nghĩa là các tin tức được khuyến nghị do sự phổ biến của tin tức. Hệ thống được đánh giá là tốt khi khuyến nghị được những tin tức không phổ biến, có nghĩa là *POP* nhỏ.

C. Điều chỉnh siêu tham số

Điều chỉnh siêu tham số thích hợp là điều cần thiết khi so sánh các phương pháp học máy, do đó, các siêu tham số được điều chỉnh cho tất cả các phương pháp học máy với tham số tối ưu là $MRR@10$. Các giá trị của siêu tham số được lặp trong các tập giá trị được liệt kê sau đây: số láng giềng gần nhất k trong tập [50,100,500,1000,1500]; số lượng phiên gần nhất để xem xét trong tập: [500,1000,2500,5000,10000]; độ đo tương tự (similarity) trong tập ['cosine','vec']; hàm mất mát (loss) trong tập: ['bpr-max', 'top1-max'], thao tác cuối cùng(final_act) trong tập: ['elu-0.5','linear'], dropout trong tập: [0.1 tới 0.9], momentum được xét trong tập: [0.1 tới 0.9]; tỉ lệ học (learning_rate) trong tập: [từ 0.01 tới 0.1 và 0.2, 0.3, 0.4, 0.5]; kích cỡ tập con (batch_size) trong tập: [16,32,64,100], số lần huấn luyện toàn bộ tập dữ liệu (epoch_n) trong tập: [10,20,30].

Do độ phức tạp của tính toán, việc điều chỉnh siêu tham số được thực hiện với tối đa lặp 100 lần. Trong mỗi vòng lặp, các yếu tố như learning rate, drop-out, momentum và hàm mất mát được xác định để tìm ra MRR tốt nhất với độ dài là 20. Kết quả các siêu tham số cho mỗi phương pháp n như bảng IV.

D. Độ phức tạp tính toán và bộ nhớ sử dụng

Nhóm nghiên cứu tiến hành thực nghiệm dựa trên phần cứng máy chủ Google colab với các thông số kỹ

Bảng VI. Kết quả thực nghiệm

Nhóm	Phương pháp	MRR@10	HR@10	MRR@5	HR@5	POP@10	Cov@10
Nhóm Baseline	AR	0.5875	0.7676	0.5329	0.6995	0.1881	0.7556
	MC	0.5667	0.7550	0.5604	0.7093	0.1426	0.6991
	SR	0.5657	0.7605	0.5589	0.7105	0.1538	0.7194
	CT	0.5616	0.7583	0.5561	0.7023	0.3109	0.6035
Nhóm KNN	SKNN	0.5269	0.7828	0.5163	0.7031	0.1919	0.7608
	V-SKNN	0.5871	0.8041	0.5788	**0.7424**	0.1575	0.8220
	S-SKNN	0.5582	**0.8063**	0.5486	0.7350	0.2111	0.7703
	SF-SKNN	0.5432	0.7640	0.5371	0.7183	0.1514	0.6877
	STAN	0.5874	0.7888	0.5799	0.7335	0.1804	0.8255
	VSTAN	**0.5880**	0.7904	**0.5806**	0.7364	0.1598	0.8513
Nhóm Deep Learning	GRU4REC	0.5572	0.7579	0.5492	0.6982	**0.0701**	**0.9345**
	STAMP	0.5550	0.7743	0.5464	0.7096	0.1819	0.8428
	NARM	0.5571	0.7926	0.5478	0.7233	0.1802	0.8672
	Nextitrec	0.5526	0.7652	0.5445	0.7057	0.1960	0.8501
	SGNN	0.5648	0.7941	0.5550	0.7210	0.1965	0.8147

thuật bao gồm: Chip: 2vCPU @ 2.20GHz (Intel(R) Xeon(R)), Ram: 13G, SSD: 100G và chạy trên GPU Tesla K80, pci bus id: 0000:00:04.0, compute capability: 3.7. Thời gian chạy/bộ nhớ cho các phương pháp được thống kê trong bảng V.

E. Kết quả thực nghiệm

Kết quả thực nghiệm tại bảng VI với các độ đo: $MRR@K$ và $HR@K$ với $K = 10$, $K = 5$ và $POP@K$, $Cov@K$ với K=10. Giá được in đậm là giá trị tốt nhất trong mỗi độ đo. Kết quả thử nghiệm cho thấy các phương pháp thuộc nhóm học sâu vẫn chiếm ưu thế trong các độ đo khác nhau. Trong đó, phương pháp GRU4REC tỏ ra vượt trội cho việc khuyến nghị các mục tin mới và bao phủ các mục tin. Nhóm KNN cũng cho kết quả khá tốt so với các phương pháp còn lại, đặc biệt với MRR@5 và HR@5 thì phương pháp thuộc nhóm KNN là VSTAN cho kết quả tốt nhất. Điều đó chứng tỏ phương pháp VSTAN thực hiện tốt việc đưa ra số ít các tin tức phù hợp.

Nhóm cơ bản cho kết quả tương đối tốt với độ đo MRR và HR với cả K=10 và K=5. Riêng phương pháp cây ngữ cảnh có độ đo *POP* lớn hơn và *Cov* nhỏ hơn hẳn các phương pháp còn lại, điều này chứng tỏ phương pháp cây ngữ cảnh khuyến nghị các tin tức phổ biến và không khuyến nghị trải rộng các tin tức. Về thời gian chạy và bộ nhớ sử dụng, nhóm Cơ bản và nhóm KNN cho kết quả khả quan (trừ phương pháp cây ngữ cảnh kém hơn), trong khi nhóm các phương pháp học sâu có thời gian thực nghiệm và bộ nhớ sử dụng lớn hơn hẳn, không phù hợp với hệ thống cần huấn luyện lại trong khoảng thời gian ngắn.

VI. KẾT LUẬN

A. Kết luận chính

Việc có thể dự đoán sự quan tâm ngắn hạn của người dùng ẩn danh trên cổng thông tin điện tử đối với một phiên trực tuyến có ý nghĩa lớn trong thực tế cũng như trong nghiên cứu trong những năm gần đây. Trong bài báo này, chúng tôi đã đặt ra và giải quyết bài toán khuyến nghị tin bài cho người dùng ẩn danh trên cổng thông tin điện tử Học viện Công nghệ Bưu chính viễn thông dựa trên chuỗi các hành động của người dùng ẩn danh trong một khoảng thời gian ngắn được gọi là phiên.

Bài báo cũng thực hiện thực nghiệm và so sánh nhóm các phương pháp khuyến nghị dựa trên dữ liệu phiên phổ biến hiện nay bao gồm nhóm cơ bản, nhóm K láng giềng gần nhất và nhóm các phương pháp học sâu. Thực nghiệm cho thấy nhóm phương pháp học sâu có kết quả tốt nhất nhưng thời gian huấn luyện lâu và sử dụng tài nguyên máy tính cao trong khi đó các phương pháp cơ bản và các phương pháp K láng giềng gần nhất cho kết quả tương đối tốt với thời gian thực hiện nhanh (gần như tức thời) và tài nguyên sử dụng thấp. Các phương pháp thuộc nhóm học sâu phù hợp với các hệ thống không huấn luyện dữ liệu thường xuyên còn các phương pháp thuộc nhóm cơ bản và nhóm K láng giềng gần nhất phù hợp với các hệ thống yêu cầu huấn luyện dữ liệu liên tục.

B. Hướng nghiên cứu tiếp theo

Các nghiên cứu và thử nghiệm trong bài báo mới tập trung vào mô hình hóa dữ liệu tuần tự với nhóm các phương pháp Cơ bản, nhóm phương pháp K láng giềng gần nhất và một số phương pháp thuộc nhóm học sâu. Việc mô hình hóa dữ liệu dạng tuần tự có thể được thực hiện bởi các phương pháp nghiên cứu tiến tiến gần đây như sử dụng mô hình Transformer, BERT và các biến thể. Đây có thể là một trong những hướng nghiên cứu tiếp theo của bài báo.

Một hướng nghiên cứu có thể phát triển từ bài báo là kết hợp dữ liệu phiên với các dữ liệu khác để tăng tính chính xác và khả năng của hệ khuyến nghị như dữ liệu về ngữ cảnh, dữ liệu về nội dung, dữ liệu hồ sơ người dùng. Sự kết hợp của dữ liệu phiên với các loại dữ liệu khác có

recommendation systems. Conventional recommender systems often use data collected over a long period of time by identifiable users. However, on the portal it is difficult to do that because most users are anonymous, so it is difficult to apply traditional methods. One solution to this problem is a user session-based recommendation, where the data is a sequence of sequential user activities over a specified period of time called sessions. In this study, we solve the problem of recommending news on the portal of Posts and Telecommunications Institute of Technology based on anonymous user session data. We study and evaluate a group of recommended methods on popular session data, including basic group, K nearest neighbor group, deep learning group. The results show that the methods belonging to the deep learning group have better results than the machine learning methods of the basic group while the methods belonging to the basic group and the K nearest neighbor group give relatively good results while less running time and resource usage.

Keywords: Web portal; Recommendation system; Session-based recommendations**.**

Nguyễn Hoàng Anh, Nhận học vị Thạc sỹ năm 2012, hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Trí tuệ Nhân tạo, Học máy, Hệ khuyến nghị.

Email: anhnh@ptit.edu.vn

Ngô Xuân Bách, Nhận học vị Tiến sỹ năm 2014 tại Nhật Bản. Hiện công tác tại Học viện Công nghệ Bưu chính Viễn thông. Lĩnh vực nghiên cứu: Xử lý ngôn ngữ tự nhiên, Học máy, Hệ khuyến nghị.

Email: bachnx@ptit.edu.vn