Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

HỆ THỐNG GỢI Ý ÁP DỤNG CHO TRANG WEB TỔNG HỢP TIN TỨC TỰ ĐỘNG

Đỗ Thành Nhân[1] và Trần Nguyễn Minh Thư[2]

[1] *THPT Lê Anh Xuân, tỉnh Bến Tre*
[2] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 03/09/2013
Ngày chấp nhận: 21/10/2013

Title:
Recommender system for news aggregation website

Từ khóa:
Hệ thống gợi ý, hệ thống hỗ trợ quyết định

Keywords:
Recommender systems, decision support systems

ABSTRACT

To assist the reader faces the information explosion, we built the recommender system applied for a news website automatically (NewsRES). The NewsRES based on the content-based method and collaborative method. The content-based method is used in comparison the content of information or describing news in order to find out the similar news which the users used to be concerned. The CF method passes the tastes of users to take advice or predictions about unknown tastes for other users. The system is applied to 280 students grade 10, 11 at Le Anh Xuan high school for a week. We gain the results: 30.59% of precision, 94.17% of recall and 45.26% of F-measure.

TÓM TẮT

Việc cập nhật tin tức là nhu cầu không thể thiếu trong thời đại hiện nay. Với trang web tổng hợp tin tức, người đọc sẽ gặp một số trở ngại trong việc tìm đọc những thông tin theo ý thích vì sự gia tăng về số lượng cũng như đa dạng về nội dung của tin tức. Nhằm hỗ trợ người đọc đối mặt với sự bùng nổ thông tin, chúng tôi xây dựng hệ thống gợi ý áp dụng cho một trang web tổng hợp tin tức tự động (NewsRES). NewsRES sử dụng phương pháp lọc theo nội dung (content-based) được thực hiện dựa trên việc so sánh nội dung thông tin hay mô tả tin tức để tìm ra những tin tức tương tự với những gì mà người dùng đã từng quan tâm; phương pháp phối hợp (CF) thông qua các thị hiếu đã được biết đến của một nhóm người dùng để đưa các tư vấn hoặc dự đoán về thị hiếu chưa biết cho một số người dùng khác. Hệ thống này được áp dụng cho 280 học sinh lớp 10, 11 tại trường trung học Lê Anh Xuân, Bến Tre. Kết quả thực nghiệm trên hệ thống NewsRES: Precision 30.59%, Recall 94.17% và F-Measure 45.26%.

1 GIỚI THIỆU

Trong những năm gần đây, hệ thống gợi ý (recommender system) được biết đến như là một sự phát triển quan trọng trong việc giúp người dùng đối mặt với sự bùng nổ thông tin. Hệ thống này được ứng dụng trong nhiều lĩnh vực như thương mại điện tử với Amazon [4], Netflix [12], Ebay [10]; trong lĩnh vực giải trí với MovieLens, Last.fm, Film-Conseil [6]; trong lĩnh vực khác như tin tức trực tuyến netnews [7],…

Kể từ năm 2007, đã có hội nghị chuyên về hệ gợi ý (ACM) là diễn đàn quốc tế hàng đầu cho việc trình bày kết quả nghiên cứu mới, trong lĩnh vực rộng lớn của hệ gợi ý.

Tuy nhiên, các hệ thống gợi ý hiện tại vẫn đòi hỏi phải có nhiều cải tiến hơn nữa để làm cho

phương pháp gợi ý hiệu quả hơn phù hợp với từng lĩnh vực (loại dữ liệu) áp dụng để có thể cung cấp gợi ý phù hợp với từng cá nhân riêng biệt [3], [9].

Trong khuôn khổ nghiên cứu này, chúng tôi muốn hướng tới hệ thống gợi ý áp dụng cho một trang web tổng hợp tin tức tự động. Với trang web tổng hợp tin tức, người đọc sẽ gặp một số trở ngại trong việc tìm đọc những thông tin theo ý thích vì sự gia tăng về số lượng cũng như đa dạng về nội dung của tin tức. Sự ra đời kỹ thuật Really Simple Syndication (RSS)[7] và sự phong phú về số lượng các trang báo điện tử hiện nay là cơ sở để chúng tôi xây dựng một trang tổng hợp tin tức tự động. Trang web này ra đời còn nhằm mục đích tránh bất tiện cho người dùng trong việc phải mở nhiều trang tin tức khác nhau.

2 HỆ THỐNG NEWSRES

Có rất nhiều cách để dự đoán, ước lượng hạng/điểm cho các dữ liệu như sử dụng học máy, lí thuyết xấp xỉ, các thuật toán dựa trên kinh nghiệm… Các hệ thống gợi ý thường được phân thành ba loại dựa trên cách nó dùng để ước lượng các đánh giá về sản phẩm:

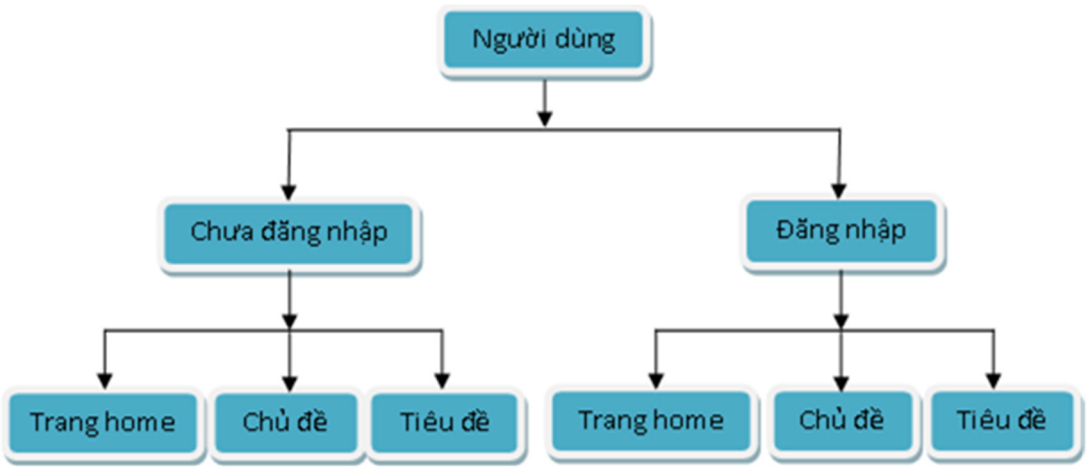
– **Dựa trên nội dung (content-based)**[1]: người sử dụng ðược gợi ý **mục dữ liệu (item)** týõng tự như những mục dữ liệu ðược người sử dụng thích trong quá khứ.

– **Gợi ý phối hợp (collaborative filtering)** [1]: người sử dụng ðược gợi ý mục dữ liệu của những người có cùng **"khẩu vị" và "sở thích"** với mình.

– **Gợi ý hỗn hợp (hybrid)**[1]: kết hợp cả hai tiếp cận ở trên.

2.1 Mô tả hệ thống NewsRES

Hệ thống NEWSRES xây dựng cho hai trường hợp: khi người dùng đăng nhập vào hệ thống hoặc không đăng nhập vào hệ thống như lưu đồ (Hình 1).



Hình 1: Lưu đồ tổng quát

2.1.1 Dữ liệu đầu vào của hệ thống NewsRES

Phân tích dữ liệu của hệ thống sẽ xây dựng để xác định giải thuật sẽ sử dụng. Dữ liệu đầu vào của hệ thống lấy được từ công nghệ RSS ta được:

– Tiêu đề.
– Phân loại/ nhóm tin.
– Tóm tắt.
– Nội dung.
– Ngày tháng.

Lịch sử truy cập của người dùng: Khi người dùng đăng ký thông tin, hệ thống sẽ lưu lại những thông tin người dùng (như nghề nghiệp, sở thích, quan tâm,…). Ngoài ra hệ thống lưu lại lịch sử truy cập tin tức của người dùng như:

– Người dùng đọc thể loại nào bao nhiêu lần trong khoảng thời gian k;

– Người dùng đọc tin "a" rồi đọc tiếp những tin nào.

2.1.2 Đặc trưng của hệ thống NewsRES

Hệ thống gợi ý tin tức là một lĩnh vực giàu tiềm năng bởi số lượng các sản phẩm tư vấn, số lượng người dùng và số lượt sử dụng tương đối nhiều. Tuy nhiên, đi kèm theo đó là các thử thách về các đặc trưng riêng của miền đối tượng tin tức cũng như các đặc trưng chung của người sử dụng gợi ý.

Tin tức là một đối tượng gợi ý đặc biệt, các đặc trưng[10] sau của tin tức giúp đưa ra các giải thuật hữu hiệu hơn trong xây dựng giải thuật cho hệ thống gợi ý tin tức của đề tài:

– Tính thời gian: theo thời gian, tin tức mất đi giá trị. Hệ thống NewsRES gợi ý từ dữ liệu được lấy trong khoảng thời gian 'x' ngày.

– Tính đa quan tâm: tại một thời điểm, người dùng có thể có nhiều mối quan tâm khác nhau. Hệ thống gợi ý phải cung cấp cho người đọc tin tức theo nhiều loại chủ đề chứ không chỉ gợi ý các tin của duy nhất một chủ đề. Ví dụ: họ có thể quan tâm đến cả các thông tin về cả thể thao và chính trị.

2.2 Giải thuật

Hệ thống gợi ý tin tức NewsRES được xây dựng tập trung vào hai giải thuật: giải thuật gợi ý **dựa trên nội dung và** giải thuật **gợi ý phối hợp**.

2.2.1 Áp dụng giải thuật gợi ý dựa trên nội dung (TF-IDF)

Phương pháp lọc theo nội dung được thực hiện dựa trên việc so sánh nội dung thông tin hay mô tả tin tức để tìm ra những tin tức tương tự với những gì mà người dùng đã từng quan tâm để giới thiệu cho người dùng những tin tức này. Lọc dựa trên nội dung thực hiện hiệu quả trên các đối tượng dữ liệu biểu diễn dưới dạng văn bản.

Lọc dựa trên nội dung không gặp phải các vấn đề rất khó giải quyết của lọc cộng tác trên miền đối tượng tin tức: Các tin tức liên tục được sinh ra và cần dễ dàng tiếp cận trong khi quá trình lọc cộng tác không thể tạo ra các sản phẩm chưa từng được đánh giá bởi người dùng khác hoặc những người dùng chưa từng đánh giá một sản phẩm nào. Khó tìm ra được các sản phẩm đã được đánh giá bởi một lượng đủ người dùng vì số lượng quá lớn các tin tức mới và đặt gánh nặng cung cấp thông tin đánh giá lên người dùng.

Vì những lý do trên, hệ thống có áp dụng giải thuật gợi ý dựa trên nội dung để gợi ý một số tin cho người đọc.

Giải thuật dựa trên nội dung

Đầu vào:

– Tập M chứa danh sách các tiêu đề (document).

– Tập Q chứa tiêu đề cơ sở (tiêu đề cần gợi ý).

Đầu ra:

– Tập C: danh sách các tin tức được xếp hạng dựa theo độ ưu tiên gợi ý.

Giải thuật:

B1: Xử lý dữ liệu (Tập M và Q):

– Đưa về chữ thường.
– Loại bỏ từ dừng (stop word).
– Loại bỏ kí tự đặc biệt.
– Loại bỏ chữ số.

B2: Tính D_f & IDF

Ta có công thức:

$$IDF(w)=\log(N/D_f(w))$$

Trong đó:

– N là tổng số lượng tài liệu cần tý vấn cho người sử dụng

– Df(w) là số lượng tài liệu mà một từ nào đó xuất hiện

– w là 1 từ nào đó.

B3: Tính trọng số TF & IDF

Ta có công thức:

$$TF= t_f / f$$

$$W= TF*IDF$$

Trong đó:

– t_f: Số lần xuất hiện của từ t trong tài liệu f.
– f: Tổng số các từ trong tài liệu f.
– W: Trọng số.

B4: Tính Normalizing Vectors

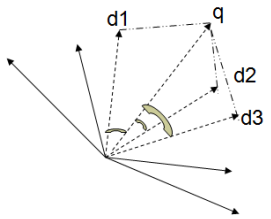
Tìm hiểu mô hình Vector Space Model (VSM):

Vector trong không gian 2 chiều thể hiện là ax+by. Tương tự với không gian n chiều. Mỗi vector là một danh sách các hệ số [a,b] định nghĩa độ lớn của vector trong chiều đó. Mỗi từ trong câu truy vấn là một chiều trong VSM, nếu câu truy vấn có 'n' từ → là một vector n-chiều. Mỗi một tài liệu cũng là một vector nhiều chiều. Như vậy, tiêu đề tin tức cần truy vấn và tiêu đề trong cơ sở dữ liệu là những vector nhiều chiều. Ta cần tính điểm (Score) giữa tiêu đề câu tư vấn và tiêu đề trong cơ sở dữ liệu.

Có hai cách tính Score $\{\vec{q},\vec{d}\}$

– Sử dụng góc giữa $\vec{q},\vec{d}$

– Sử dụng khoảng cách $\vec{q},\vec{d}$



Hình 2 : Hình vector câu truy vấn

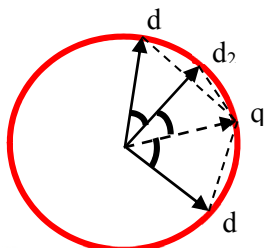
Trong đó:

– $\vec{q}$: tiêu đề tin tức cần tư vấn

– $\vec{d}$: tiêu đề trong cơ sở dữ liệu.

Mỗi tài liệu có độ dài khác nhau thì cách tính theo khoảng cách không còn đúng nữa vì tài liệu nào càng dài thì score càng lớn. Từ đó ta cần **Normalizing Vectors**, làm cho các vector có cùng độ lớn.

Công thức:



$$Q' = \left(\frac{q_1}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \frac{q_2}{\sqrt{\sum_{i=1}^{m} q_i^2}}, \ldots, \frac{q_m}{\sqrt{\sum_{i=1}^{m} q_i^2}}\right)$$

$$d' = \left(\frac{d_1}{\sqrt{\sum_{i=1}^{m} d_i^2}}, \frac{d_2}{\sqrt{\sum_{i=1}^{m} d_i^2}}, \ldots, \frac{d_m}{\sqrt{\sum_{i=1}^{m} d_i^2}}\right)$$

Trong đó: q, d: là trọng số TF*IDF

B5: Tính độ tương đồng của chúng bằng độ đo cosin

$$u(c,s) = \frac{\sum_{i=1}^{K} w_{ic} w_{is}}{\sqrt{\sum_{i=1}^{K} w_{ic}^2} \sqrt{\sum_{i=1}^{K} w_{is}^2}}$$

2.2.2 Áp dụng giải thuật gợi ý phối hợp (CF)

Phương pháp lọc phối hợp được thực hiện thông qua thị hiếu đã được biết đến của một nhóm người dùng để đưa các tư vấn hoặc dự đoán về thị hiếu chưa biết cho một số người dùng khác. Lọc phối hợp sử dụng cơ sở dữ liệu về sở thích của người dùng đối với các item để dự đoán các chủ đề hoặc sản phẩm thêm vào cho một người dùng mới có cùng sở thích.

Hệ thống gợi ý cộng tác khắc phục được nhiều nhược điểm của hệ thống dựa trên nội dung. Một điểm quan trọng là nó có thể xử lý mọi loại dữ liệu và gợi ý một loại sản phẩm, kể cả những sản phẩm mới, khác hoàn toàn so với những gì người dùng từng xem.

Vì những lý do trên, thay vì chỉ cần dùng giải thuật gợi ý đựa trên nội dung, hệ thống đề xuất thêm giải thuật gợi ý phối hợp dựa trên bộ nhớ.

Giải thuật độ tương quan Pearson giữa hai item (CF-ITEM):

Đầu vào:

– Cho tập người dùng u thuộc U là những người cùng đánh giá về hai item i và j

– **Đầu ra:**

– Độ tương quan Pearson giữa item i và item j.

Giải thuật:

B1: Tính trung bình của item thứ I bởi những người dùng khác

B2: Tính Độ tương quan Pearson giữa item i và item j

Công thức:

$$w_{i,j} = \frac{\sum_{u \in U}(r_{u,i} - \overline{r_i})(r_{u,j} - \overline{r_j})}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r_i})^2} \sqrt{\sum_{u \in U}(r_{u,j} - \overline{r_j})^2}}$$

Trong đó:

– $r_{u,i}$: là đánh giá của người dùng u cho item i,

– $\overline{r_i}$: là đánh giá trung bình của item thứ I bởi những người dùng khác.

– Đánh giá người dùng u cho item i: số lần click chuột trên mục tin.

Giải thuật độ tương quan Pearson giữa người dùng (CF-USER):

Đầu vào:

– Cho tập người dùng u thuộc U

Đầu ra:

– Độ tương quan Pearson giữa user U_i và user U_j.

Giải thuật:

B1: Tính trung bình của người dùng U

B2: Tính Độ tương quan Pearson giữa user U_i và user U_j

Công thức:

$$w_{ij} = \frac{\sum_{x \in P_i \cap P_j}(r_{i,x} - \overline{r_i})(r_{j,x} - \overline{r_j})}{\sqrt{\sum_{x \in P_i \cap P_j}(r_{i,x} - \overline{r_i})^2}\sqrt{\sum_{x \in P_i \cap P_j}(r_{j,x} - \overline{r_j})^2}}$$

Trong đó:

– $x \in P_i \cap P_j$: là tập sản phẩm mà người dùng *i* và người dùng *j* cùng đánh giá

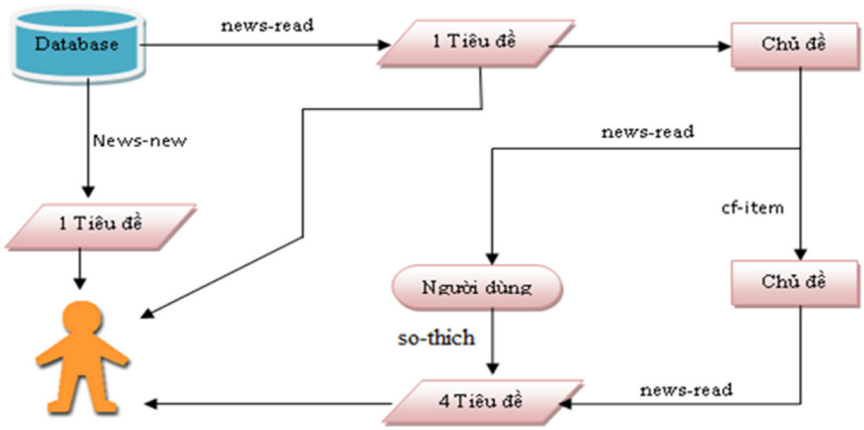
– $r_{i,x}$: là đánh giá của người dùng *i* lên sản phẩm *x*.

– $\overline{r_i}$: là đánh giá trung bình của người dùng *i*.

2.3 Giới thiệu hệ thống NewsRec

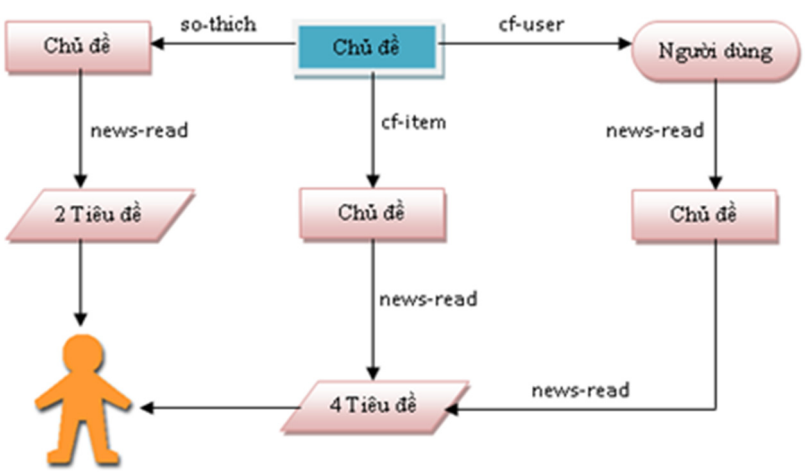
Hệ thống dự đoán thông qua danh sách Top-N tin tức được sắp xếp theo thứ tự giảm dần về độ tương quan. Trong khuôn khổ bài báo này, chúng tôi trình bày 2 trạng thái của người dùng như sau:

Hệ thống gợi ý khi người dùng không đăng nhập: gợi ý một tiêu đề mới nhất trong dữ liệu; một tiêu đề đọc nhiều nhất; lấy chủ đề có tiêu đề đọc nhiều nhất kế tiếp tìm người dùng đọc chủ đề này nhiều nhất dùng giải thuật "so-thich" với người dùng này để gợi ý hai tiêu đề; lấy chủ đề có tiêu đề đọc nhiều nhất dùng giải thuật "cf-item" rồi gợi ý hai tin đọc nhiều nhất như lưu đồ (Hình 3).



Hình 3: Lưu đồ đang ở trang HOME không đăng nhập

Hệ thống gợi ý khi người dùng đăng nhập click vào chủ đề: dùng giải thuật "so-thich" tìm chủ đề gợi ý hai tiêu đề đọc nhiều nhất trong chủ đề này; dùng giải thuật "cf-item" tìm chủ đề tương tự như chủ đề vừa click => gợi ý hai tiêu đề đọc nhiều nhất trong chủ đề,; dùng giải thuật "cf-user" tìm người dùng tương tự với người dùng hiện tại, lấy chủ đề có tiêu đề mà người dùng này đọc nhiều nhất và gợi ý hai tin đọc nhiều nhất trong chủ đề này như lưu đồ (Hình 4).



Hình 4: Lưu đồ click vào chủ đề khi đăng nhập

Chú thích:

– news-new: Những tin mới đăng.

– news-read: Những tin đọc nhiều nhất, trong khoảng thời gian 3 ngày.

– so-thich: truy vấn dựa vào hồ sơ người dùng.

– news-assess: Người dùng đánh giá cao nhất.

– cf-item: Giải thuật độ tương quan Pearson giữa hai item.

cf-user: Giải thuật độ tương quan Pearson giữa người dùng.

2.4 Phương pháp đánh giá

Đề tài trình bày tóm tắt các phương pháp đánh giá hiệu quả hệ thống gợi ý. Đây là một vấn đề rất quan trọng giúp cho người sử dụng có thể lựa chọn đúng mô hình phù hợp với dữ liệu hay ứng dụng trong thực tế.

2.4.1 Nghi thức kiểm tra

Để đánh giá chất lượng của một hệ thống gợi ý chúng ta cần phải đúng cách phân vùng các tập dữ liệu vào một tập huấn luyện và một bộ kiểm tra. Điều rất quan trọng là hiệu suất được tính toán trên dữ liệu mà không có phần trong việc xây dựng các mô hình. Một số chương trình học tập cũng cần một tập hợp xác nhận để tối ưu hóa các thông số mô hình. Bộ dữ liệu thường được chia theo một trong các phương pháp sau:

Holdout: chia tách tập dữ liệu thành hai phần: một tập huấn luyện và một bộ kiểm tra. Những bộ có thể có tỷ lệ khác nhau. Lấy ngẫu nhiên 2/3 tập dữ liệu D để huấn luyện và 1/3 tập dữ liệu còn lại dùng cho bộ kiểm tra, có thể lặp lại quá trình này k lần rồi tính giá trị trung bình.

k-fold: chia tập dữ liệu D thành k phần (fold) bằng nhau, lặp lại k lần, mỗi lần sử dụng k-1 folds để học và 1 fold để kiểm tra, sau đó tính trung bình của k lần kiểm tra. Khi tập dữ liệu D có hơn 300 phần tử, phương pháp thường sử dụng là 10 fold (k=10). Nếu tập D có ít hơn thì leave-1-out (k= số phần tử) được đề nghị sử dụng.

2.4.2 Các chỉ số sử dụng đánh giá

Việc đánh giá hiệu quả của một hệ thống khuyến nghị là cần thiết. Tuy nhiên, việc đánh giá không chỉ phụ thuộc vào dữ liệu mà còn phụ thuộc vào mục tiêu của hệ thống khuyến nghị [2]. Thật vậy, một số hệ thống nhấn mạnh sự đa dạng của các mục trong danh sách khuyến nghị, trong khi những người khác tập trung vào tính mới. Có nhiều chỉ tiêu đánh giá khác nhau nhưng trong khuôn khổ bài báo này, chúng tôi sử dụng precision, recall và F- Measure các chỉ số chủ yếu được sử dụng trong các hệ thống khuyến nghị của thương mại điện tử[10, 11].

$$precision = \frac{tp}{tp + fp}$$

$$recall = \frac{tp}{tp + fn}$$

$$F - measure = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Ở đó:

– true positive (TP): tin tức thú vị gợi ý cho người dùng),

– true negative (TN): tin tức thú vị không gợi ý cho người dùng),

– false negative (FN): tin tức không thú vị không gợi ý cho người dùng),

– false positive (FP) : tin tức không thú vị gợi ý cho người dùng).

3 KẾT QUẢ VÀ THẢO LUẬN

3.1 Xây dựng tập dữ liệu

Kết quả nghiên cứu này được thực nghiệm với tập dữ liệu tin tức (bộ dữ liệu NewsRES). Dữ liệu này được lấy tin tự động từ hai trang web tin tức (vnExpress.net, dantri.com.vn). Ngoày ra, thông tin người dùng được lưu lại từ thông tin đăng ký sử dụng và nhật ký sử dụng của người dùng. Đây là dữ liệu đầu vào của hệ thống đã được mô tả trong mục 2.1.1. Tất cả các dữ liệu này được dùng để xây dựng hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động.

Hệ thống này bước đầu được áp dụng cho học sinh trường THPT Lê Anh Xuân, Bến Tre. Thực nghiệm trên bốn lớp khối 10, 11 (10a, 10c1, 11a, 11c1, 11c2, 11c4, 11c5). NewsRES tính thời điểm thực nghiệm có tổng số 1020 tin và 280 người dùng, có được 229 giao dịch (session), số tin gợi ý cho người dùng (Recommendhistory) 6481 tin, tổng số tin người dùng đọc 1976 tin.

3.2 Phương pháp thử nghiệm

Trong trường hợp phân lớp nhị phân (lớp gợi ý là lớp cần quan tâm, lớp không gợi ý). Trong phân lớp nhị phân, người ta cần tính ma trận phân lớp C có kích thước 2x2 (Bảng 1).

Bảng 1: Bảng ma trận phân lớp C

dự đoán =>	Gợi ý	Không gợi ý
Dùng	True – Positive (tp)	False-Negative (fn)
Không dùng	False – Positive (fp)	True – Negative (tn)

Trong đó:

– True positive (tp): Số tin tức tư vấn chính xác.

– False negative (fn): Số tin tức dùng mà không có tư vấn.

– False positive (fp): Số tin tức tư vấn không chính xác.

– True negative (tn): Số tin tức không dùng không tư vấn.

Các độ đo được tính thông qua precision, recall và F- Measure xác định theo công thức. Giá trị precision, recall càng lớn hiệu quả phương pháp càng cao[2].

$$precision = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ tư\ vấn} = \frac{tp}{tp + fp}$$

$$recall = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ trong\ tập\ người\ dùng} = \frac{tp}{tp + fn}$$

$$F - measure = \frac{2\ x\ precision\ x\ recall}{precision + recall}$$

Ví dụ chúng ta có thể xét giao dịch id là 7e02fa0676, tập dữ liệu có 940 tin trong đó có 15 tin sử dụng là lớp người dùng quan tâm và 913 tin thuộc lớp không dùng. Ta tính được các sai số dự đoán sau:

Bảng 2: Ma trận phân lớp

dự đoán =>	Gợi ý	Không gợi ý
Dùng	12	3
Không dùng	28	897

$$precision = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ tư\ vấn} = \frac{12}{12 + 28} = 30\%$$

$$recall = \frac{Số\ tin\ tức\ tư\ vấn\ chính\ xác}{Tổng\ số\ tin\ tức\ trong\ tập\ người\ dùng} = 80\%$$

$$F - measure = \frac{2\ x\ precision\ x\ recall}{precision + recall} = \frac{2\ x\ 30\%\ x\ 80\%}{30\% + 80\%} = 43.64\%$$

3.3 Kết quả thử nghiệm

Hệ thống NewsRES là mô hình gợi ý tin tức. Độ precision, recall và F-Measure được tính toán dựa trên 1020 tin dùng để tư vấn. Thông qua kiểm nghiệm, chúng tôi được kết quả kiểm nghiệm của mô hình đề xuất trong trường hợp khi người dùng đăng nhập và người dùng không đăng nhập được thể hiện trong hai bảng (Bảng 3, Bảng 4).

Bảng 3: Kết quả thực nghiệm khi người dùng đăng nhập

SessionID	Precision	Recall	F-Measure
084a7a6be3	26.73%	87.10%	40.91%
1663e275a7	36.36%	98.11%	53.06%
4998df2c47	18.75%	100%	31.58%
398ca839ae	25.23%	93.10%	39.70%
24878b0e96	30.53%	100%	46.78%
6afde6425b	50%	100%	66.67%
90b50df2dc	16.67%	100%	28.58%
7e36e77535	23.88%	100%	38.55%
…	…	…	…
36670fc3b6	25%	100%	40%
8dac844637	21.62%	100%	35.55%
8dbc531d17	16.67%	80%	27.59%
5d5c751399	10%	50%	16.67%
27d8e225a4	50%	100%	66.67%
518ee45dc7	50%	100%	66.67%
7e02fa0676	35.71%	83.33%	50%
fbb077a2d0	25%	100%	40%
2486dac3a9	44.44%	100%	61.53%
Tổng	30.59%	94.17%	45.26%

Bảng 4: Kết quả thực nghiệm khi người dùng không đăng nhập

SessionID	Precision	Recall	F-Measure
672991a038	42.86%	100%	60%
24878b0e96	34.48%	100%	51.28%
9f0f1540b2	20%	50%	28.57%
7ed88d577c	33.72%	100%	50.43%
4998df2c47	27.27%	100%	42.85%
1663e275a7	14.29%	100%	25.01%
613b3746d5	40%	100%	57.14%
9b79d29215	20.22%	100%	33.64%
8c1e09d46d	10.34%	75%	18.17%
…	…	…	…
6afde6425b	7.14%	100%	13.33%
24eb7fbfc3	36.36%	100%	53.33%
7e02fa0676	26.92%	77.78%	40%
5a212af2c6	40%	100%	57.14%
4738ac737a	22.22%	100%	36.36%
57102845ad	12.50%	100%	22.22%
ed42411619	22.86%	100%	37.21%
Tổng	25.13%	86%	38%

Kết quả thực nghiệm cho thấy (Precision 30.59%, Recall 94.17% và F-Measure 45.26%.) thì hệ thống NewsRES hiệu quả hơn khi người dùng

không đăng nhập vào hệ thống (Precision 25.13%, Recall 86% và F-Measure 38%).

4 KẾT LUẬN VÀ ĐỀ XUẤT

4.1 Kết luận

Chúng tôi đã trình bày mô hình hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động và hiệu quả kết hợp lọc nội dung và lọc cộng tác để gợi ý tin tức cho người dùng.

Đề tài đã tiến hành thử nghiệm mô hình trên tập dữ liệu (NewsRES) có 940 tin được lấy tự động từ hai trang web vnExpress.net, dantri.com.vn, người sử dụng là học sinh trường THPT Lê Anh Xuân khối 10, 11. Kết quả kiểm nghiệm trên tập dữ liệu NewsRES, ta có kết quả với Precision = 30.59%, Recall = 94.17% , F-Measure = 45.26%.

Theo công trình nghiên cứu, phát triển và ứng dụng CNTT-TT, Lọc cộng tác và lọc theo nội dung dựa trên mô hình đồ thị, năm 2009 của Nguyễn Duy Phương, Từ Minh Phương[5] thì độ đo Precision = 29.2%. Tuy không thể so sánh trực tiếp kết quả thực nghiệm của chúng tôi so với công trình trong bài báo [5], nhưng kết quả này cũng phản ảnh được các hệ thống gợi ý hiện tại chưa đạt được giá trị precision cao như những lĩnh vực nghiên cứu khác.

4.2 Đề xuất

Tiến hành thử nghiệm hệ thống NewsRES với nhiều đối tượng khác nhau (giáo viên, học sinh,…) với khoảng thời gian nhiều hơn.

Tìm kiếm các dữ liệu trong cùng lĩnh vực để so sánh, đối chiếu kết quả nghiên cứu với những giải pháp khác.

Hoàn thiện hệ thống gợi ý người đọc cho trang web tổng hợp tin tức tự động thông qua việc đánh giá kết quả gợi ý và phản hồi của người đọc cũng như trong lúc so sánh với các giải pháp khác để tăng chất lượng của các gợi ý.

Phát triển trên các lĩnh vực khác như tìm kiếm khách sạn, địa điểm du lịch.

TÀI LIỆU THAM KHẢO

1. Gendiminas Adomavicius, Alexander Tuzhilin, Toward the Next Generation of Recommender Systems: A Survey of the State-of-the Art and Possible Extensions.
2. Herlocker Jonathan L., Konstan Jo- seph A., “Evaluating collaborative filtering recommender systems” ACM *Trans. Inf. Syst.*, vol. 22, no. 1, pp. 5–53, 2004.
3. Huang, Z.; Zeng, D. & Chen, A comparative study of recommendation algorithms for e-commerce applications, IEEE Intelligent Systems, 2006.
4. Linden, G.; Smith, B. & York, J. , Amazon.com Recommendations: Item-to-Item Collaborative Filtering, IEEE Internet Computing, IEEE Educational Activities Department, 2003, 7, 76-80.
5. Nguyễn Duy Phương, Từ Minh Phương, 2009, Các công trình nghiên cứu, phát triển và ứng dụng CNTT-TT, Lọc cộng tác và lọc theo nội dung dựa trên mô hình đồ thị.
6. Perny, P. & Zucker, J. D., Preference-based Search and Machine Learning for Collaborative Filtering: the “Film-Conseil” recommender system, Information, Interaction, Intelligence, 2001, 1, 9-48.
7. P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. "GroupLens: An Open Architecture for Collaborative Filtering of Netnews”, Proceedings of the 1994 Computer Supported Cooperative Work Conference, ACM, 1994.
8. RSS, http://en.wikipedia.org/wiki/RSS.
9. Sarwar, B. & al., Analysis of recommendation algorithms for e-commerce EC '00, ACM, 2000, 158-167.
10. Schafer, J. B.; Konstan, J. A. & Riedl, J., E-Commerce Recommendation Applications, Data Min. Knowl. Discov., Kluwer Academic Publishers, 2001, 5, 115-153.
11. Uông Huy Long, 2010, khóa luận tốt nghiệp đại học, giải pháp mở rộng thông tin ngữ cảnh phiên duyệt web người dùng nhằm nâng cao chất lượng tư vấn trong hệ thống tư vấn tin tức.
12. Yehuda Koren, August 2009, The BellKor Solution to the Netflix Grand Prize.