Hệ thống gợi ý bài báo[1]

[*]Phạm Minh Chuẩn, [**]Lê Thanh Hương, [**]Trần Đình Khang

{ Chuanpm@gmail.com; huongthanh@gmail.com; khangtd-fit@mail.hut.edu.vn}

[*]Đại học Sư phạm Kỹ thuật Hưng Yên

[**]Viện CNTT & TT - ĐHBKHN

Tóm tắt. Hệ thống gợi ý là công cụ hỗ trợ ra quyết định, nhằm mục đích cung cấp cho người sử dụng những gợi ý về thông tin, sản phẩm, và dịch vụ phù hợp nhất với yêu cầu và sở thích riêng của từng người tại từng tình huống (ngữ cảnh) hoặc yêu cầu cụ thể. Có nhiều hướng tiếp cận khác nhau để xây dựng hệ thống gợi ý, chẳng hạn như: lọc cộng tác, dựa trên nội dung, dựa trên tri thức và kỹ thuật lai. Trong bài viết này chúng tôi đề xuất mô hình cho bài toán gợi ý bài báo, sử dụng các kỹ thuật khai phá dữ liệu kết hợp với kỹ thuật lọc cộng tác và dựa trên nội dung để xây hệ thống gợi ý bài báo phù hợp nhất cho người dùng.

Từ khóa. Hệ thống gợi ý, lọc cộng tác, dựa trên nội dung, dựa trên tri thức, kỹ thuật lai, khai phá dữ liệu, độ tương đồng.

1. Giới thiệu và phát biểu bài toán

Ngày nay, người sử dụng của các hệ thống thông tin, đặc biệt là các website thương mại, thường khó khăn trong việc tìm kiếm và lựa chọn các thông tin cần thiết và phù hợp để giải quyết một vấn đề nhất định (ví dụ như việc chọn mua một máy ảnh kỹ thuật số phù hợp, hoặc việc lập kế hoạch cho một chuyến đi du lịch); bởi vì người sử dụng có quá nhiều sự lựa chọn, nhưng không có đủ thời gian hoặc "tri thức" để tự đánh giá những lựa chọn này và đưa ra các quyết định tối ưu.

Hệ thống gợi ý (Recommender Systems) ([8], [12]) là các công cụ hỗ trợ ra quyết định, nhằm mục đích cung cấp cho người sử dụng những gợi ý về thông tin, sản phẩm, và dịch vụ phù hợp nhất với yêu cầu và sở thích riêng của từng người tại từng tình huống (ngữ cảnh) yêu cầu cụ thể. Hiện nay, các hệ thống gợi ý đã trở thành một trong những công cụ hữu dụng và phổ biến nhất trong các hệ thống thương mại điện tử (ví dụ như Amazon.com, Barnes&Noble.com, Yahoo! news, TripAdvisor.com,…).

Đối với người dùng là những nhà khoa học hoặc những người nghiên cứu thì việc tìm cho mình một bài báo khoa học để tham khảo phù hợp với lĩnh vực mình nghiên cứu cũng không phải là vấn đề đơn giản. Từ những phân tích trên nhóm đã đề xuất ra một hệ thống gợi ý bài báo đáp ứng được nhu cầu tìm kiếm tài liệu tham khảo bằng tiếng việt cho các nhà khoa học, các nhà nghiên cứu tại Việt Nam.

Bài toán gợi ý bài báo có thể được phát biểu như sau. Hệ thống bao gồm **m** người dùng, và **n** bài báo. Khi người dùng có yêu cầu tìm kiếm các bài báo bằng việc nhập thông tin cần tìm kiếm (các từ khóa) sau đó hệ thống sẽ đưa ra danh sách các bài báo mà theo dự đoán các bài báo này phù hợp nhất với người dùng (có mức độ ưu tiên cao đối với người dùng).

Phát biểu bài toán dạng công thức toán học như sau:

Giả sử

U là tập **m** người dùng trong hệ thống.

I là tập **n** bài báo trong hệ thống.

u là một người dùng trong **U**

i là một bài báo trong I

r_{ui} là đánh giá của người dùng *u* cho bài báo i. r_{ui} nhận giá trị trong tập $X \subset R$. Trong đó, tập X là các số nguyên {1, 2, 3, 4, 5} tương ứng với các mức đánh giá Ghét, Không thích, được, thích, rất thích. .

Bài báo được biểu diễn dưới dạng có cấu trúc như sau:

Item = (PaperID, Title, Date, Magazine, Authors, Keyword1, Keyword2, Domain, References, Content)

- ✓ PaperID: ID của bài báo
- ✓ Title: tiêu đề bài báo, được biểu diễn bằng một vector trọng số của các từ khóa.

[1] Bài báo được hoàn thành với sự tài trợ của Quỹ Phát triển Khoa học và Công nghệ Quốc gia (Nafosted), mã số đề tài 102.01.30.09.

- ✓ Date: ngày đăng của bài báo.
- ✓ Magazine: tạp chí đăng.
- ✓ Authors: Tập thể tác giả, được biểu bởi một vector Id của các tác giả.
- ✓ Keyword1: tập các từ khóa xuất hiện trong Keyword của bài báo, được biểu diễn bằng một vector trọng số của các từ khóa.
- ✓ Keyword2: tập các từ khóa đại diện cho toàn bộ bài báo
- ✓ Domain: lĩnh vực của bài báo.
- ✓ References: tập các bài báo tham khảo, được biểu diễn bởi một vector Id của các bài báo.
- ✓ Content: nội dung của bài báo.

Thông tin về người dùng được biểu diễn như sau:

User = (UserID, Name, sex, Domains, Keywords, Id_author)

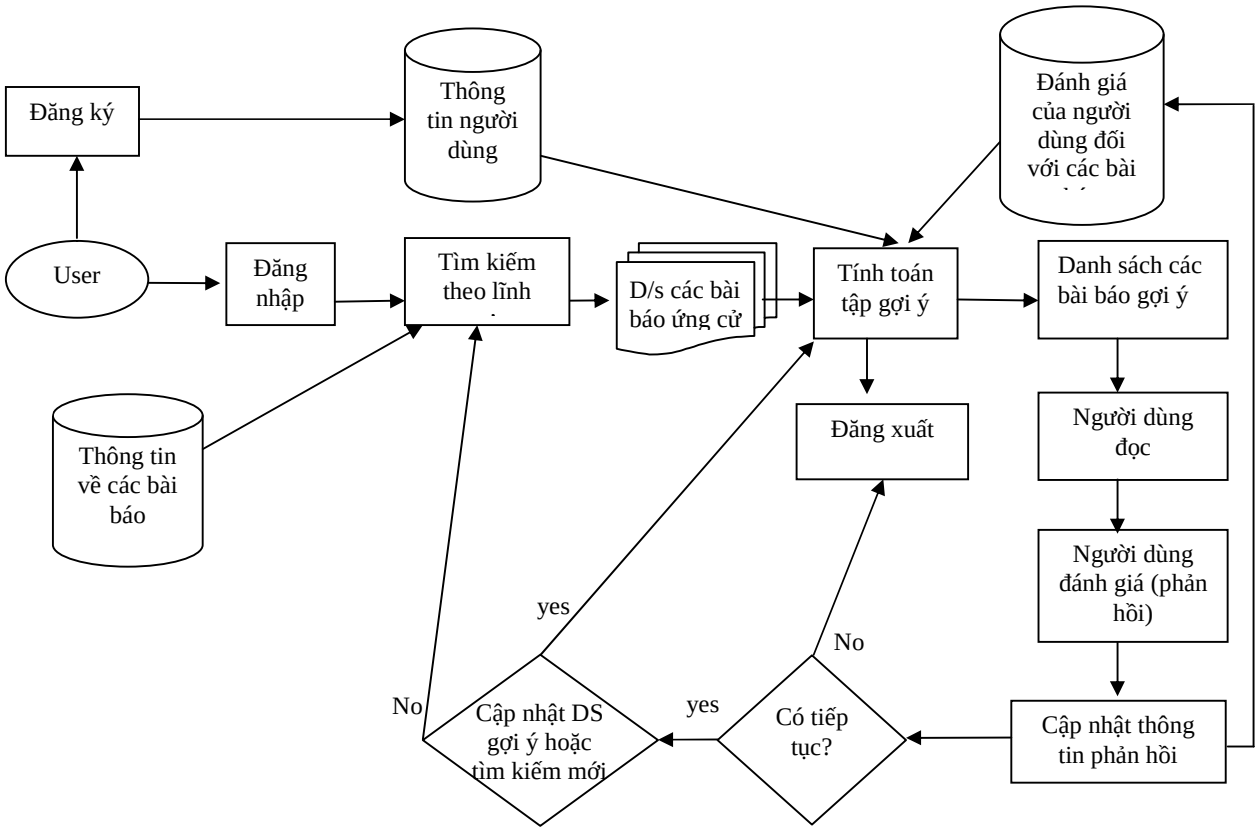
- ✓ Name: tên của người dùng
- ✓ Sex: giới tính
- ✓ Domains: những lĩnh vực nghiên cứu
- ✓ Keywords: những từ khóa yêu thích (tuần suất tìm kiếm)
- ✓ Id_author: tác giả yêu thích (tần suất đọc bài báo của tác giả)

Bảng Đánh giá người dùng đối với các bài báo

- ✓ PaperID
- ✓ UserID
- ✓ Rate

2. Mô hình hệ thống gợi ý bài báo

Mô hình tổng quát của hệ thống gợi ý được trình bày trong hình 1 dưới đây.



Hình 1. Mô hình hệ thống gợi ý bài báo

Khi người dùng sử dụng hệ thống, nếu chưa đăng ký thì hệ thống yêu cầu đăng ký và cung cấp một số thông tin cá nhân cơ bản. Khi đăng nhập vào hệ thống; người dùng nhập các từ khóa cần tìm kiếm; hệ thống dựa trên thông tin về những từ khóa mà người dùng nhập vào tiến hành tìm kiếm trong kho dữ liệu về các bài báo (tìm trong keyword2) đưa ra được tập các bài báo ứng cử, sau đó hệ thống sẽ tính toán tập k các bài báo mà phù hợp nhất với người dùng bằng việc sử dụng các thông tin cá nhân như là những từ khóa và các tác giả yêu thích; và sự tương đồng của các bài bài báo ứng cử với những bài báo mà người dùng đã từng được đánh giá kết hợp với sự đánh giá của những người dùng có cùng sở thích về những bài báo ứng cử; sau đó hệ thống sẽ hiển thị danh sách k bài báo phù hợp nhất cho người dùng, người dùng đọc và phản hồi lại cho hệ thống về các bài báo đó; hệ thống sẽ tiến hành cập nhật thông tin phản hồi; đồng thời nếu người dùng muốn tiếp tục thì sẽ có hai lựa chọn hoặc là cập nhật danh sách gợi ý hoặc là tiến hành tìm kiếm mới và hệ thống tiếp tục theo những lựa chọn tương ứng.

Dựa trên các dữ liệu thu thập là đánh giá của người dùng với các bài báo trong hệ thống (explicit data) kết hợp với thông tin cá nhân của người dùng như là lĩnh vực nhiên cứu, danh sách các bài báo đã đăng, áp dụng các mô hình tính toán, dự đoán đánh giá của một người dùng xác định với một bài báo họ chưa đọc. Cụ thể trong hệ thống này chúng tôi áp dụng kỹ thuật lọc cộng tác (Collaborative filter), dựa trên đánh giá của người dùng về các bài báo;và kỹ thuật dựa trên nội dung (Content based) kết hợp với đánh giá độ tương tự của các bài báo, mức độ phù hợp của một bài báo với người dùng để đưa ra danh sách các bài báo phù hợp cho người dùng trong hệ thống.

Một số tình huống của hệ thống:

Tình huống thứ nhất: Khi có một người sử dụng mới người này chỉ có thông tin cá nhân được thu thập khi đăng ký; chưa có bất kỳ sự đánh giá nào về các bài báo dẫn đến khó khăn cho hệ thống gợi ý để tính toán những bài báo phù hợp nhất đối với người đó.

Giải pháp:

Dựa trên những từ khóa tìm kiếm của người dùng, và những từ khóa yêu thích để tính mức độ phù hợp của bài báo với người dùng và đưa ra những bài báo khởi tạo cho người dùng đó; khi người dùng đọc bài báo và có phản hồi thì hệ thống sẽ cập nhật lại để đưa ra gợi ý mới hơn.

Tình huống thứ hai: Khi có một bài báo mới mà chưa có người dùng nào đánh giá hoặc có ít người dùng đánh giá, lúc đó hệ thống sẽ xử lý như thế nào để gợi ý cho người dùng cụ thể.

Giải pháp:

Sử dụng thông tin về bài báo như là lĩnh vực; thông tin về tác giả bài báo; sự tương tự của các bài báo để đưa ra gợi ý cho những người dùng trong hệ thống. Cụ thể, nếu sử dụng thông tin về tác giả bài báo (tác giả này cũng là một người dùng trong hệ thống) chúng ta có thể tìm ra những người dùng có sở thích tương tự như tác giả bài báo và dự đoán đánh giá của những người sử dụng đó về bài báo; hoặc chúng ta có thể dựa vào sự tương đồng của các bài báo đã được đánh giá bởi một người sử dụng nào đó với bài báo mới từ đó có thể tính toán được xem bài báo này có độ phù hợp nhất với từng người dùng cụ thể,…

Tình huống thứ ba: Với người dùng đã có trong hệ thống, chúng ta có thể áp dụng kỹ thuật lọc cộng tác, kỹ thuật lọc dựa trên nội dung kết hợp các kỹ thuật khai phá dữ liệu để đưa ra gợi ý phù hợp nhất đối với người dùng.

3. Các kỹ thuật sử dụng trong quá trình gợi ý

3.1 Quá trình tiền xử lý dữ liệu

Trước khi đi vào các thuật để tính toán tập gợi ý các bài báo thì hệ thống cần phải tiền xử lý các dữ liệu đầu vào; ở đây chính là tập các bài báo. Với một bài báo lưu trữ dưới dạng một tệp văn bản, và với giả thiết rằng các bài báo đã phân theo từng lĩnh vực nghiên cứu.

Với mỗi bài báo chúng ta sẽ tiến hành phân tách các từ theo phương pháp Maximum Matching. Đây là phương pháp tách từ đơn giản và dễ cài đặt. Phương pháp này sử dụng một từ điển từ vựng để làm cơ sở phân tách các từ.

Sau khi đã phân tách được các từ chúng ta sẽ tiến hành lựa chọn đặc trưng cho từng văn bản. Lựa chọn đặc trưng là việc loại bỏ các từ không mang thông tin khỏi văn bản nhằm nâng cao hiệu quả phân loại và giảm sự

phức tạp tính toán. Có nhiều phương pháp lựa chọn đặc trưng, tuy nhiên trong khuôn khổ bài viết này chúng tôi sử dụng phương pháp Information Gain. Các đặc trưng sau khi đã được trích chọn được lưu trong Keyword2 của bài báo.

Information Gain xác định lượng thông tin của một từ được sử dụng cho việc phân loại dựa trên sự có mặt hay không có mặt của từ đó trong một văn bản.

Bài toán lựa chọn đặc trưng của văn bản được phát biểu như sau: cho $c_1,...,c_k$ là tập hợp gồm k loại (hay k lĩnh vực). $P(c_j)$ là xác suất tìm được một văn bản có trong tập dữ liệu và văn bản đó thuộc loại c_j. P(w) là xác suất xuất hiện của từ w. $P(w, c_j)$ là xác suất tìm được một văn bản thuộc loại c_j và văn bản đó chứa từ w. $P(\overline{w}, c_j)$ là xác suất tìm được một văn bản thuộc loại c_j và văn bản đó không chứa từ w.

Information Gain của một từ w được tính như sau:

$$IG(w,c_i) = \sum_{c \in \{c_i, \bar{c}_i\}} \sum_{w \in \{w_k, \overline{w_k}\}} P(w,c).\log \frac{P(w,c)}{P(w).P(c)} \qquad (1)$$

Theo phương pháp này người ta dựa vào các hàm tính toán $IG(w,c_i)$ của từ w đối với phân loại c_i để quyết định xem có nên lựa chọn từ w làm đặc trưng của tài liệu hay không. Ta cần tính toán IG cho mỗi từ trong tập dữ liệu huấn luyện và những từ có IG nhỏ hơn một ngưỡng cho trước sẽ bị loại bỏ.

Để biểu diễn các bài báo chúng ta sử dụng vector trong không gian K chiều; chúng ta có thể sử dụng mô hình tần suất để biểu diễn sử dụng phương pháp **TF x IDF**:

Phương pháp TF × IDF

Với bài báo d_j được biểu diễn bởi một vector đặc trưng $d_j(w_{1j}, w_{2j},\ldots, w_{kj})$

Trong đó w_{ij} là trọng số của từ khóa t_i trong bài báo d_j

Phương pháp này là tổng hợp của hai phương pháp TF và IDF ([11]), giá trị của ma trận trọng số được tính như sau:

$$w_{ij} = \begin{cases} \left[1+\log\left(\frac{f_{ij}}{\max_z f_{zj}}\right)\right]\log\left(\frac{m}{h_i}\right) & \text{if } h_i \geq 1 \\ 0 \text{ otherwise} \end{cases} \qquad (2)$$

Trong đó, f_{ij} là số lần xuất hiện của từ khóa t_i trong bài báo d_j và $\max_z f_{i,z}$ là số lần xuất hiện lớn nhất của một từ khóa trong tài liệu d_j, m là số lượng bài báo và h_i là số bài báo mà từ khóa t_i xuất hiện.

Phương pháp này kết hợp được ưu điểm của cả hai phương pháp TF và IDF. Trọng số w_{ij} được tính bằng tần suất xuất hiện của từ khóa t_i trong bài báo d_j và *độ hiếm* của từ khóa t_i trong toàn bộ cơ sở dữ liệu.

Việc gán trọng số cho các từ khóa đặc trưng trong Keyword2 sẽ xem xét đến các từ khóa xuất hiện trong tiêu đề bài báo và trong Keyword1 (do người dùng tự nhập), với các từ khóa này thì sẽ được với trọng số có mức độ quan trọng hơn so với các từ khóa đặc trưng khác.

3.2. Một số thuật toán liên quan

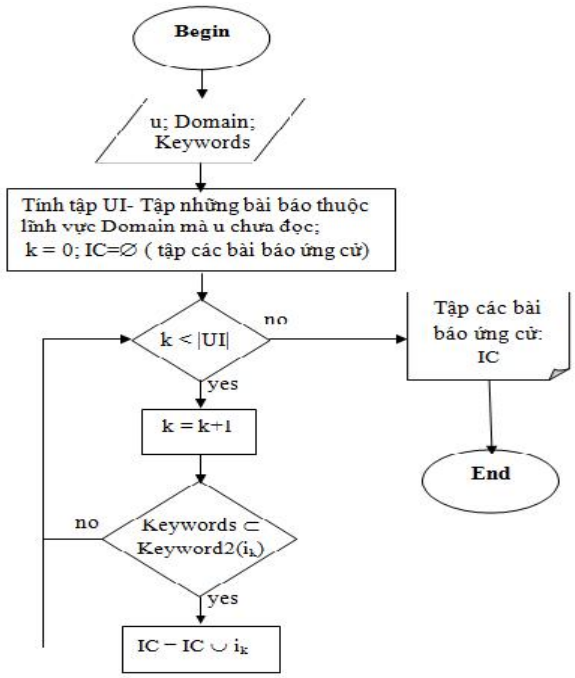
3.2.1 Thuật toán đưa ra danh sách các bài báo ứng cử

Đầu vào: danh sách các từ khóa, các bài báo trong hệ thống với keyword2 tương ứng

Đầu ra: danh sách các bài báo ứng cử

Quá trình xử lý:

- Tìm kiếm các bài báo có chứa các từ khóa cần tìm trong Keyword2
- Nếu không tìm thấy yêu cầu người dùng nhập lại từ khóa



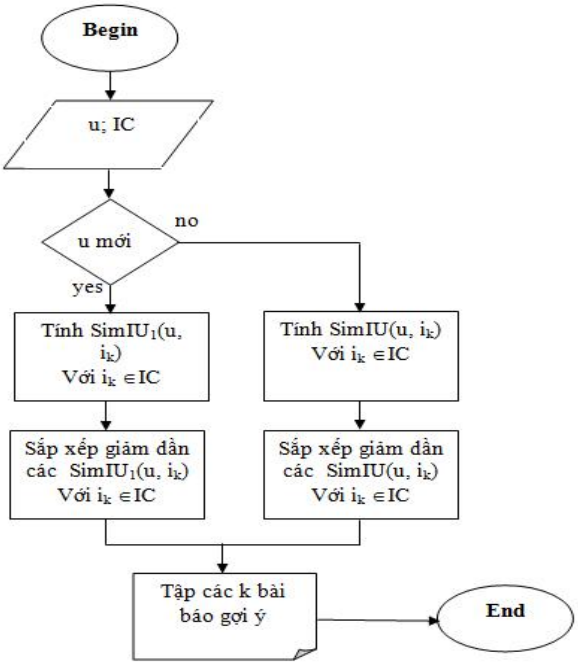
Hình 2. Lưu đồ thuật toán đưa ra danh sách các bài báo ứng cử

3.2.2 Thuật toán đưa ra danh sách các bài báo gợi ý

Đầu vào: Danh sách các bài báo ứng cử

Đầu ra: Danh sách k bài báo phù hợp nhất gợi ý cho người dùng

Quá trình xử lý:



Hình 3. Lưu đồ thuật toán đưa ra danh sách các bài báo gợi ý

Trường hợp thứ nhất: nếu người dùng mới sử dụng hệ thống chưa đọc được nhiều bài báo thì chúng ta sẽ sử dụng công thức (6) để tính toán mức độ phù hợp của các bài báo ứng cử với người dùng dựa trên những từ khóa yêu thích từ đó đưa ra k bài báo phù hợp nhất với người dùng.

Trường hợp thứ 2: khi người dùng đã từng sử dụng hệ thống và có nhiều đánh giá về các bài báo trong hệ thống thì chúng ta sẽ sử dụng công thức (9) để tính độ phù hợp của bài báo ứng cử với người dùng từ đó đưa ra k bài

báo phù hợp nhất với người dùng.; trong công thức này việc xác định các hệ số phù hợp cũng rất quan trọng; chúng tôi sẽ tiến hành xác định giá trị của các hệ số phù hợp qua thực nghiệm.

Vấn đề chọn k bằng bao nhiêu là thích hợp chúng tôi sẽ tiến hành thực nghiệm để xác định.

3.3 Kỹ thuật đánh giá độ tương tự giữa hai bài báo

Việc đánh giá độ tương tự của hai bài báo cũng rất quan trọng; trong bài viết này chúng tôi sẽ đánh giá độ tương tự giữa hai bài báo kết hợp hai cách sau:

Cách 1: dựa trên các từ khóa đặc trưng và trọng số của nó trong hai bài báo

Cách 2: dựa trên sự đánh giá của cùng một tập người dùng với hai bài báo đó

Với cách thứ nhất; chúng tôi dựa trên hai tiêu chí:

- ✓ Số từ khóa giống nhau của hai bài báo (mỗi bài báo sẽ được biểu diễn bởi một tập các từ khóa đặc trưng và trọng số tương ứng của từ khóa)
- ✓ Mức độ quan trọng của các từ khóa với hai bài báo (thể hiện bằng trọng số của từ khóa với bài báo)

Để đánh giá độ tương tự của hai bài báo dựa trên các từ khóa đặc trưng, chúng tôi đề xuất công thức sau:

$$Sim_1(i,j) = \frac{m}{K}\left(1 - \frac{\sum_{k=1}^{m} |w_{ki} - w_{kj}|}{\sum_{k=1}^{m} (w_{ki} + w_{kj})}\right) \quad (3)$$

Trong đó m là tổng số từ khóa giống nhau của hai bài báo; K là tổng số từ khóa của bài báo; w_{ki} là trọng số của từ khóa thứ k trong bài báo i.

Ý tưởng chủ đạo ở đây là, với hai bài báo khác nhau nếu có càng nhiều từ khóa giống nhau thì độ tương đồng càng cao, ngoài ra để đánh giá chính xác hơn chúng ta quan tâm đến trọng số của mỗi từ khóa giống nhau đối với hai bài báo, nếu trọng số của các từ khóa càng gần nhau thì mức độ tương đồng giữa hai bài báo càng cao.

Với cách thứ 2; chúng tôi sử dụng độ tương tự PC (Pearson Correlation) [1].

$$Sim_2(i,j) = \frac{\sum_{u \in U_{ij}} (r_{ui} - \overline{r_i})(r_{uj} - \overline{r_j})}{\sqrt{\sum_{u \in U_{ij}} (r_{ui} - \overline{r_i})^2 \sum_{u \in U_{ij}} (r_{uj} - \overline{r_j})^2}} \quad (4)$$

Trong đó:

U_{ij}: tập tất cả người dùng cùng đánh giá hai bài báo i, j

r_{ui} và r_{uj}: đánh giá của người dùng u cho mỗi bài báo i, j

$\overline{r_i}, \overline{r_j}$: đánh giá trung bình đối với bài báo i, j xét trên tập tất cả những đánh giá của người dùng trong tập U_{ij}.

Công thức tính độ tương tự tổng quát cho hai bài báo được đề xuất như sau:

$Sim(i, j) = w_1 * Sim_1(i, j) + w_2 * Sim_2(i,j)$ (5)

Trong đó w_1, w_2 là các hệ số tương tự được xác định thông qua thực nghiệm, thỏa mãn điều kiện $w_1, w_2 >= 0$; và $w_1 + w_2 = 1$.

Công thức 5 sẽ được sử dụng để tính mức độ phù hợp giữa bài báo với người dùng trình bày ở mục 3.5.

3.4 Kỹ thuật đánh giá mức độ tương tự giữa hai người dùng

Đánh giá mức độ tương tự giữa hai người dùng chúng tôi cũng sử dụng độ tương tự PC, tương tự như công thức 4.

$$SimU(u,v) = \frac{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_{i \in I_{uv}} (r_{ui} - \overline{r_u})^2 \sum_{i \in I_{uv}} (r_{vi} - \overline{r_v})^2}} \quad (6)$$

Trong đó:

I_{uv}: tập tất cả các bài báo được đánh giá bởi hai người dùng u, v

r_{ui} và r_{vi}: đánh giá của người dùng u cho bài báo i

$\overline{r_u}, \overline{r_v}$: đánh giá trung bình của người dùng u, v trong tập các bài báo I_{uv}

3.5 Kỹ thuật đánh giá mức độ phù hợp của một bài báo với một người dùng

Việc đánh giá mức độ phù hợp của một bài báo trong danh sách ứng cử với người dùng sẽ giúp cho quá trình lọc các bài báo gợi ý cho người dùng được thuận tiện hơn. Chúng tôi đề xuất phương pháp đánh giá mức độ phù hợp giữa một bài báo đối với người dùng bằng việc kết hợp ba cách sau:

Cách 1: dựa trên mức độ quan trọng của các từ khóa mà người dùng yêu thích trong bài báo; nếu bài báo nào càng chứa nhiều từ khóa yêu thích và các từ khóa đó có mức độ quan trọng càng cao với bài báo đó thì bài báo đó sẽ có mức độ phù hợp càng cao với người dùng.

Công thức tính độ phù hợp theo cách này được đề xuất như sau:

$$SimIU_1(u,i) = \frac{n}{m} \sum_{k=1}^{m} w_{ki} \quad (7)$$

Trong đó m là tổng số từ khóa người dùng yêu thích, n là số từ khóa mà người dùng yêu thích xuất hiện trong bài báo i; w_{ki} là trọng số của từ khóa k xuất hiện trong bài báo i.

Công thức này có thể được dùng trong trường hợp người dùng mới sử dụng hệ thống chưa đọc và đánh giá nhiều bài báo.

Cách 2: dựa trên ý tưởng của kỹ thuật lọc cộng tác.

Đây là phương pháp quen thuộc, và được sử dụng rộng rãi nhất. Hệ gợi ý dùng phương pháp lọc cộng tác (Collaborative Recommender System) tập hợp thông tin về đánh giá, hoạt động của người dùng; xác định mức độ giống nhau giữa người dùng dựa trên sở thích, từ đó đưa ra các gợi ý từ kết quả so sánh trên.

Chúng ta sẽ sử dụng công thức (6) tính toán mức độ phù hợp giữa những người dùng từ đó đưa ra tập những người dùng (ký hiệu UN) có mức độ phù hợp cao nhất với người dùng **u**.

Sau đó chúng ta sẽ tiến hành xét tập những người dùng UN có đánh giá về bài báo **i** đang xét, xác định đánh giá trung bình đối với bài báo **i** xét trên tập tất cả những đánh giá của người dùng trong tập UN để đưa ra độ phù hợp.

Công thức tính độ phù hợp được đề xuất như sau:

$$SimIU_2(u,i) = \frac{\sum_{un \in UN} r_{un,i}}{5 * |UN|} \quad (8)$$

Trong đó $r_{un,i}$ là đánh giá của người dùng un đối với bài báo **i**; ở đây chúng ta sử dụng thang đánh giá có 5 mức do vậy để chuẩn hóa mức độ phù hợp (nhận giá trị từ 0 đến 1) của bài báo **i** với người dùng u chúng ta sẽ chia cho 5.

Cách 3: dựa trên ý tưởng của phương pháp lọc theo nội dung; với cách này chúng ta sẽ đánh giá mức độ phù hợp của bài báo **i** đối với người dùng **u,** bằng cách đánh giá mức độ tượng tự trung bình giữa những bài báo mà người dùng **u** đã từng đọc trong quá khứ với bài báo **i.**

Giả sử gọi IP là tập những bài báo mà người dùng **u** đã được đánh giá trong quá khứ; khi đó chúng ta sẽ sử dụng công thức (5) để tính độ tương tự giữa bài báo **i** với từng bài báo trong IP và tính trung bình độ tương tự của bài báo **i** với các bài báo đó.

Công thức tính độ phù hợp trong trường hợp này được đề xuất như sau:

$$SimIU_3(u,i) = \frac{\sum_{j \in IP} Sim(i,j)}{|IP|} \quad (9)$$

Công thức tổng quát mà chúng tôi đề xuất để tính toán mức độ phù hợp của bài báo **i** đối với người dùng **u** như sau:

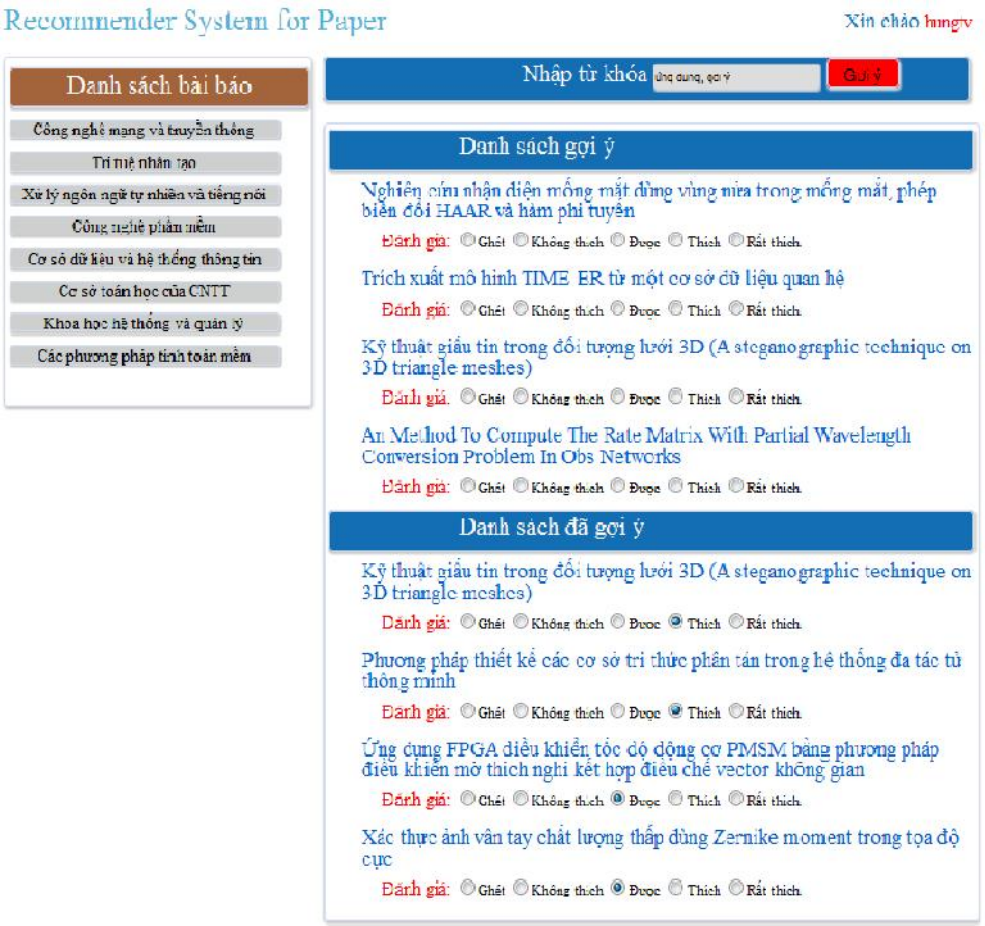
$$SimIU(\mathbf{u},\mathbf{i}) = l_1 * SimIU_1(\mathbf{u},\mathbf{i}) + l_2 * SimIU_2(\mathbf{u},\mathbf{i}) + l_3 SimIU_3(\mathbf{u},\mathbf{i}) \quad (10)$$

Trong đó l_1, l_2, l_3 là các hệ số phù hợp được tính toán trong quá thực nghiệm; l_1, l_2, l_3 >=0 và $l_1 + l_2 + l_3 = 1$.

3.6. Cài đặt ứng dụng và đánh giá kết quả thực nghiệm

Chúng tôi đã tiến hành cài đặt thử nghiệm với tổng số một trăm bài báo thuộc 8 lĩnh vực: Công nghệ mạng và truyền thông, Trí tuệ nhân tạo, Xử lý ngôn ngữ tự nhiên và tiếng nói, Công nghệ phần mềm, Cơ sở dữ liệu và hệ thống thông tin, Cơ sở toán học của tin học, Khoa học hệ thống và quản lý và Các phương pháp tính toán mềm.

Các bài báo trong hệ thống đã được đánh giá bởi tổng số 50 người bao gồm các chuyên gia và các đồng nghiệp trong lĩnh vực Công nghệ Thông tin và toán tin, bước đầu đã đạt được kết quả khả quan, lúc ban đầu với 20 người được gợi ý dựa vào những từ khóa tìm kiếm, dựa vào lĩnh vực nghiên cứu trên tổng số bài báo 45, độ chính xác đạt 10%; sau khi hệ thống đã có dữ liệu đánh giá của người dùng tiếp tục thử nghiệm với số lượng 40 người dùng và 70 bài báo thì độ chính xác đạt tới 19% và khi thử nghiệm với cả 50 người dùng và 100 bài báo thì độ chính xác của quá trình gợi ý đạt 25%.



Hình 4. Giao diện chính của hệ thống gợi ý bài báo

4. Kết luận và hướng phát triển

Trong bài báo chúng tôi đã đề xuất kỹ thuật đánh giá độ tương đồng của hai bài báo dựa trên những từ khóa đặc trưng và dựa trên sự đánh giá của cùng một tập người dùng với hai bài báo đó; trong kỹ thuật này cần phải xác định các hệ số tương tự w_1 và w_2 làm sao cho đạt hiệu quả cao nhất. Ngoài ra bài viết còn đề xuất thêm kỹ thuật đánh giá độ phù hợp giữa một bài báo đối với một người dùng sử dụng hai ý tưởng chủ đạo của phương pháp lọc cộng tác và dựa trên nội dung. Trong kỹ thuật này việc xác định các hệ số phù hợp cũng rất quan trọng. Trong hệ thống của chúng tôi, các hệ số này được xác định qua thực nghiệm.

Hệ thống gợi ý bài báo mà chúng tôi đề xuất bước đầu đã được cài đặt thử nghiệm, tuy nhiên chưa được kiểm chứng bởi một tập đủ lớn dữ liệu trong thực tế, do vậy độ chính xác chưa cao. Chúng tôi sẽ phải làm việc nhiều hơn nữa để hoàn thiện mô hình và đưa ra thử nghiệm với khối lượng lớn các bài báo và người dùng.

5. Tài liệu tham khảo

1. Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B.Kantor,: *Recommender Systems Handbook* , Springer, 13 Dec 2010.

2. Gediminas Adomavicius and Alexander Tuzhilin: *Toward the Next GenerationofRecommender Systems state of the art 2005,* IEEE, 2005.

3. Stefan B. Beckers,: *Recommander system,* Duisburg Publisher ,2006.

4. Matthew R. McLaughlin and Jonathan L. Herlocker , *A Collaborative Filtering Algorithm and Evaluation Metric* , Oregon Univer Publisher, 2004.

5. Daniel Billsus and Michael J. Pazzani , *Learning Collaborative Information Filters*.

6. Robin Burke , Hybrid Recommender Systems.

7. Thomas Tran and Robin Cohen, *Hybrid Recommender Systems for Electronic Commerce*, AAAI, 2000.

8. Resnick, P., & Varian, H. R. (1997), *Recommender systems. Communications of the ACM , 40* (3), 56-58.

9. Two Crows, *Introduction to Data Mining and Knowledge Discovery*, Third Edition, http://www.twocrows.com/booklet.htm.

10. Fabrizio Sebastiani, *Machine Learning in Automated Text Categorization*, ACM Computing Surveys, Vol. 34, No. 1, March 2002.

11. Salton, G., & McGill, M. J. (1986), *Introduction to Modern Information Retrieval,* McGraw-Hill, Inc.

12. Billsus, D., & Pazzani, M. J. (1999), *A personal news agent that talks, learns and explains*. *AGENTS '99: Proceedings of the third annual conference on Autonomous Agents,* (pp. 268-275). ACM.

13. Adomavicius, G., & Tuzhilin, A. (2005), *Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions*. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING , 17*, 734-74.

Ngu

N
đá