MÔ HÌNH KẾT HỢP NGƯỠNG SỞ THÍCH VÀ LUẬT SỐ ĐÔNG CHO DỰ ĐOÁN XẾP HẠNG TRONG HỆ THỐNG GỢI Ý

Lê Huỳnh Quốc Bảo[1], Quách Nguyễn Đạt[2], Nguyễn Thái Nghe[3]
[1] Trung tâm Công nghệ Phần mềm, Trường Đại học Cần Thơ
[2] Trung tâm Chất lượng Nông lâm thủy sản vùng 5
[3] Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ
lhqbao@ctu.edu.vn, nguyendat.nafi5@gmail.com, ntnghe@cit.ctu.edu.vn

TÓM TẮT *- Hệ thống gợi ý (Recommender Systems – RS) hiện đang được sử dụng trong nhiều lĩnh vực để dự đoán "sở thích" (thói quen/ nhu cầu/ năng lực/…) của người dùng từ đó gợi ý cho họ những mục thông tin (item) phù hợp nhất. Thương mại điện tử ở Việt Nam hiện đang phát triển mạnh, do vậy RS sẽ mở ra nhiều tiềm năng trong cả nghiên cứu lẫn ứng dụng.*

Bài viết này đề xuất một tiếp cận mới trong dự đoán xếp hạng của hệ thống gợi ý, đó là việc sử dụng luật bình chọn số đông kết hợp với ngưỡng sở thích nhằm xác định giá trị xếp hạng của người dùng trên các mục thông tin. Phương pháp này khá đơn giản nhưng lại cho kết quả tốt. Kết quả thử nghiệm trên các tập dữ liệu chuẩn cho thấy phương pháp được đề xuất có thời gian thực hiện nhanh hơn đáng kể so với các phương pháp truyền thống dựa trên lọc cộng tác trong khi độ chính xác cũng được cải thiện trong phần lớn các trường hợp thử nghiệm. Chính vì thế, đây có thể là một hướng tiếp cận hữu ích trong lĩnh vực dự đoán xếp hạng của RS.

Từ khóa *- hệ thống gợi ý, lọc cộng tác, bình chọn số đông, ngưỡng sở thích.*

1. Giới thiệu

Hiện nay thương mại điện tử (e-commerce) giúp người dùng có thể tiếp cận với sản phẩm một cách dễ dàng và nhanh chóng hơn so với các phương thức mua bán truyền thống. Chính vì tính tiện ích của nó dẫn đến sự bùng nổ của các các website bán hàng, giới thiệu sản phẩm trực tuyến. Điều này thuận lợi cho khách hàng do thông tin đa dạng, phong phú, tuy nhiên cũng gây khó khăn trong việc lựa chọn sản phẩm phù hợp do có quá nhiều thông tin. Thế nên, sự trợ giúp và tư vấn cho khách hàng là rất quan trọng và cần thiết để họ có thể lựa chọn được sản phẩm mà phù hợp với sở thích của mình.

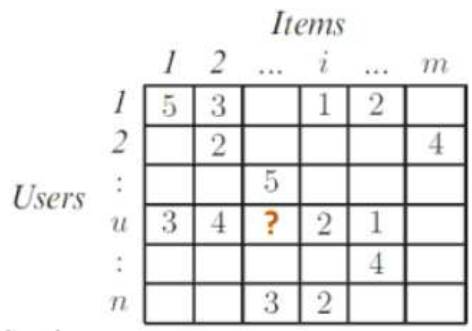
Hệ thống gợi ý (RS) đóng vai trò như một người trung gian đưa ra các gợi ý sản phẩm phù hợp với sở thích của người dùng. Bằng cách thu thập thông tin về sở thích (thông qua các phản hồi của người dùng trên sản phẩm), hệ thống sẽ gợi ý các sản phẩm phù hợp nhất.

Đã có nhiều công trình nghiên cứu về các hệ thống gợi ý sử dụng các kỹ thuật khác nhau như: Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác [3]; Xây dựng hệ thống gợi ý phim dựa trên mô hình nhân tố láng giềng [15]; Hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động [8];...

Trong nội dung bài viết này, chúng tôi đề xuất một hướng tiếp cận mới sử dụng luật bình chọn số đông kết hợp ngưỡng sở thích cho hệ thống gợi ý dự đoán xếp hạng. Chúng tôi không đi sâu vào việc xây dựng một hệ thống gợi ý như thế nào, mà chỉ tập trung giới thiệu một phương pháp mới, đơn giản nhưng hiệu quả, cho dự đoán xếp hạng trong hệ thống gợi ý.

2. Hệ thống gợi ý và vấn đề dự đoán xếp hạng (rating prediction)

Trong RS, ba thông tin cơ bản về user, item và rating được biểu diễn thông qua một ma trận như trong Hình 1. Ở đó, mỗi dòng là một user, mỗi cột là một item, và mỗi ô là một giá trị xếp hạng (rating) biểu diễn "mức độ thích" của user trên item tương ứng. Các ô có giá trị là những item mà các user đã xếp hạng trong quá khứ. Những ô trống là những item chưa được xếp hạng (điều đáng lưu ý là mỗi user chỉ xếp hạng cho một vài item trong quá khứ, do vậy có rất nhiều ô trống trong ma trận này – còn gọi là ma trận cực thưa – sparse matrix).



Hình 1. Ma trận biểu diễn xếp hạng của người dùng-mục tin

Như vậy, nhiệm vụ chính của RS là dựa vào các ô đã có giá trị trong ma trận này (dữ liệu thu được từ quá khứ), để dự đoán các ô còn trống (của user hiện hành), sau đó sắp xếp kết quả dự đoán (ví dụ, từ cao xuống thấp) và chọn ra Top-N items theo thứ tự, từ đó gợi ý chúng cho người dùng.

Một cách hình thức:

- Gọi U là một tập hợp n người dùng (user), $|U| = n$, và u là một người dùng cụ thể nào đó ($u \in U$)

- Gọi I là một tập hợp m mục thông tin (item), $|I| = m$, và i là một mục thông tin cụ thể nào đó($i \in I$).

- Gọi R là một tập hợp các giá trị dùng để ước lượng 'sở thích' (preference) của người dùng, và $r_{ui} \in R$ ($R \subset \mathcal{R}$) là xếp hạng của người dùng u trên mục thông tin i.

Lưu ý rằng giá trị r_{ui} có thể được xác định một cách tường minh (explicit feedback) như thông qua việc đánh giá/xếp hạng (ví dụ, rating từ 1 đến 5; hay like (1) và dislike (0),…) mà u đã bình chọn cho i – trong trường hợp này gọi là dự đoán xếp hạng (rating prediction); hoặc r_{ui} có thể được xác định một cách không tường minh (implicit feedback) như số lần click chuột, thời gian mà u đã duyệt/xem i,…

- Gọi Dtrain $\subseteq$ U × I × R là tập dữ liệu huấn luyện

- Gọi $D^{test} \subseteq U \times I \times R$ là tập dữ liệu kiểm thử.

- Gọi r: $U \times I \rightarrow R$ $\quad\quad$ $(u, i) \rightarrow r_{ui}$

Mục tiêu của RS là tìm một hàm $\hat{r}$: $U \times I \rightarrow \mathcal{R}$

Sao cho $\xi(r, \hat{r})$ thỏa mãn một điều kiện nào đó. Ví dụ, nếu ξ là một hàm ước lượng lỗi như Root Mean Squared Error (RMSE) thì nó cần phải được tối tiểu.

$$RMSE = \sqrt{\frac{1}{|D^{test}|}\sum_{u,i,r \in D^{test}}(r_{ui} - \hat{r}_{(u,i)})^2} \quad (1)$$

Có 2 dạng dự đoán phổ biến trong RS là dự đoán xếp hạng như đã thấy ở trên và dự đoán mục thông tin (item prediction) – xác định xác suất mà người dùng thích mục tin tương ứng. Tuy nhiên, trong khuôn khổ bài viết này, chúng tôi chỉ quan tâm đến lĩnh vực dự đoán xếp hạng.

Hiện tại, trong RS có rất nhiều giải thuật được đề xuất, tuy nhiên ta có thể gom chúng vào trong 3 nhóm chính:

- Nhóm giải thuật lọc cộng tác (Collaborative Filtering): trong nhóm này, các giải thuật chủ yếu dựa trên các kỹ thuật:

+ Phương pháp láng giềng (Neighborhood-based, còn gọi là Memory-based), trong đó hoặc là dựa trên dữ liệu quá khứ của người dùng "tương tự - similarity" (user-based approach), hoặc là dựa trên dữ liệu quá khứ của những item "tương tự" (item-based approach). Trong đó sử dụng mô hình nhân tố láng giềng là một điển hình.

+ Dựa trên mô hình (Model-based): Nhóm này liên quan đến việc xây dựng các mô hình dự đoán dựa trên dữ liệu thu thập được trong quá khứ. Như mô hình Bayesian, các mô hình nhân tố tiềm ẩn (latent factor models).

- Nhóm giải thuật lọc trên nội dung (Content-based Filtering): Gợi ý các item dựa vào hồ sơ (profiles) của người dùng hoặc dựa vào nội dung (attributes) của những item tương tự như item mà người dùng đã chọn trong quá khứ.

- Kết hợp cả 2 cách trên.

Trong phần tiếp theo, chúng tôi sẽ trình bày chi tiết thuật toán User_KNN – được ứng dụng rất nhiều trong RS và hướng tiếp cận bình chọn số đông kết hợp ngưỡng sở thích cho hệ thống gợi ý dự đoán xếp hạng.

3. Thuật toán về mô hình láng giềng trong lọc cộng tác (k-Nearest Neighbors)

Phương pháp lọc cộng tác có đặc trưng cơ bản là nó thường sử dụng toàn bộ dữ liệu đã có để dự đoán đánh giá của một người dùng nào đó về sản phẩm mới. Nhờ lợi thế là nó có khả năng đưa trực tiếp dữ liệu mới vào bảng dữ liệu, do đó nó đạt được khá nhiều thành công khi được áp dụng vào các ứng dụng thực tế. Cũng do đó mà các kỹ thuật này thường đưa ra các dự đoán chính xác hơn trong các hệ trực tuyến – nơi mà ở đó luôn có dữ liệu mới được cập nhật.

Thông thường, có hai cách tiếp cận của lọc cộng tác theo mô hình K láng giềng: hệ dựa trên người dùng (User_KNN) – tức dự đoán dựa trên sự tương tự giữa các người dùng và hệ dựa trên sản phẩm (Item_KNN) – dự đoán dựa trên sự tương tự giữa các sản phẩm. Hệ dựa trên người dùng (User_KNN) xác định sự tương tự giữa hai người dùng thông qua việc so sánh các đánh giá của họ trên cùng sản phẩm, sau đó dự đoán đánh giá sản phẩm i bởi người dùng u, hay chính là đánh giá trung bình của những người dùng tương tự với người dùng u. Độ tương tự giữa người dùng u và người dùng u' có thể được tính theo Pearson (L. Herlocker et al., 1999) vì phân tích thực nghiệm cho thấy rằng đối với hệ dựa trên người dùng thì tính độ tương tự theo Pearson sẽ tốt hơn so với một vài cách khác như độ tương tự theo cấp

bậc của Spearman (Spearman's rank correlation) hay độ tương tự theo bình phương trung bình (mean squared difference). Công thức tính độ tương tự theo Pearson và Cosine như sau:

$$sim_{pearson}(u,u') = \frac{\sum_{i\in I_{uu'}}(r_{ui}-\bar{r}_u)(r_{u'i}-\bar{r}_{u'})}{\sqrt{\sum_{i\in I_{uu'}}(r_{ui}-\bar{r}_u)^2}\sqrt{\sum_{i\in I_{uu'}}(r_{u'i}-\bar{r}_{u'})^2}} \quad (2)$$

$$sim_{cosine}(u,u') = \frac{\sum_{i\in I_{uu'}} r_{ui}.r_{u'i}}{\sqrt{\sum_{i\in I_{uu'}} r_{ui}^2}\sqrt{\sum_{i\in I_{uu'}} r_{u'i}^2}} \quad (3)$$

Trong đó:

$I_{uu'}$ là một tập các item được đánh giá bởi *u* và *u'*

$\bar{r}_u$ là giá trị đánh giá trung bình trên tất cả các item của người dùng *u*.

$\bar{r}_{u'}$ là giá trị đánh giá trung bình trên tất cả các item của người dùng *u'*.

Đưa ra được những dự đoán hoặc lời gợi ý là một bước quan trọng trong hệ tư vấn lọc cộng tác. Sau khi tính toán độ tương tự giữa các người dùng hay giữa các sản phẩm, chúng ta có thể dự đoán đánh giá của người dùng *u* trên sản phẩm *i* theo công thức (P. Resnick et al., 1994) như sau:

$$\hat{r}_{ui} = \bar{r}_u + \frac{\sum_{u'\in K_u} sim(u,u').(r_{u'i}-\bar{r}_{u'})}{\sum_{u'\in K_u}|sim(u,u')|} \quad (4)$$

Trong đó:

$\hat{r}_{ui}$ chính là dự đoán cho người dùng *u* trên sản phẩm *i*

Sim(u,u'): độ tương tự giữa người dùng *u* và *u'*.

K_u là số người dùng có độ lận cận gần người dùng *u*.

Chúng tôi biểu diễn giải thuật lọc cộng tác dựa trên người dùng lân cận gần nhất (User_KNN) sử dụng độ tương tự Pearson bằng ngôn ngữ giả để dự đoán độ thích cho người dùng *u* trên sản phẩm *i* như sau:

1: **procedure** USERKNN-CF($\bar{r}_u$,*r*,D^{train})
2: **for** *u*=1 to *N* **do**
3: Tính Sim_uu' , sử dụng công thức (2)
4: **end for**
5: **Sort** Sim_uu' // sắp xếp giảm dần độ tương tự
6: **for** *k*=1 to *K* **do**
7: $K_u \leftarrow k$ // Các người dùng k gần nhất của u
8: **end for**
9: **for** *i* = 1 to *M* **do**
10: Tính $\hat{r}_{ui}$, sử dụng công thức (4)
11: **end for**
12: **end procedure**

Trong đó:

$\bar{r}_u$: đánh giá trung bình của người dùng *u* trên tất cả các item

r: đánh giá của người dùng trên tập huấn luyện

K: người dùng k gần nhất

N: số người dùng

M: số item

D^{train}: tập dữ liệu huấn luyện

4. Phương pháp sử dụng luật bình chọn số đông kết hợp ngưỡng sở thích

Trong phần này chúng tôi sẽ đề xuất phương pháp Luật bình chọn số đông kết hợp ngưỡng sở thích (MASSVOTING) dựa trên ý tưởng xem xét sự tương đồng về sở thích của các user đối với các item, được mô tả cụ thể như sau:

- Giả sử, cho ma trận dữ liệu user-item-rating như Hình 2.

- Gọi U là tập tất cả các user có trong cơ sở dữ liệu.

- Xét tại user u

+ Gọi I_u là tập các item cần dự đoán xếp hạng của user u và $I_{u'}$ là tập các item của user u' tương ứng với các item cần dự đoán xếp hạng của user u có giá trị xếp hạng $\geq T$ (với T là ngưỡng sở thích nhận giá trị thuộc tập $R = \{1, 2, 3, 4, 5\}$).

+ Gọi $I_{uu'}$ là tập các item được đánh giá bởi cả user u và u' (với $u' \in U$) có sở thích $\geq T$.

+ Xét item i thuộc $I_{u'}$, với mỗi item i' thuộc $I_{uu'}$ tăng giá trị (số lần được đánh giá xếp hạng $\geq T$) cho item thứ i trong tập I_u lên 1.

Sau khi duyệt qua tất cả các user, mỗi item trong tập I_u sẽ chứa số lần được đánh giá xếp hạng $\geq T$.

Sắp xếp I_u theo thứ tự giảm dần và chọn ra top N items để gợi ý cho user u.

Item/ User	1	2	3	4	5
1	1	4	5	2	3
2	5	1	2.995	5	2
3	4	1	2	5	2
4	2.501	3	4	3.495	4
5	2	1	3.013	5	2
6	3.991	4	1	2	4
7	5	3	2	1	4
8	3	4	4	3.464	5
9	2	4	4	2	4.499798
10	3	4	1	1.507	2

Hình 2. Ma trận dữ liệu user-item-rating

Giải thuật luật bình chọn số đông kết hợp ngưỡng sở thích để gợi ý item i cho user u được biểu diễn bằng ngôn ngữ giả như sau:

```
1:  function MASSVOTING(D^train , u, i, threshold)
2:      r̂_ui = 0                                  // Khởi tạo giá trị cần dự đoán
3:      for u'=1 to N do                          // Duyệt qua tất cả user ngoại trừ user u
4:          if (r_u'i ≥ threshold) then           // Nếu giá trị xếp hạng của user u' cho item i lớn hơn hoặc bằng ngưỡng
5:              r̂_ui ← count(I_uu')               // Đếm số item trong I_uu' và cộng dồn vào r̂_ui
6:          end if
7:      end for
8:      return r̂_ui                               // Trả kết quả dự đoán của user u cho item i
9:  end function
```

Trong đó:

D^{train}: tập dữ liệu huấn luyện – đầu vào

u: user đang xét – đầu vào

i: item đang xét – đầu vào

threshold: là ngưỡng sở thích nhận giá trị thuộc tập $R = \{1, 2, 3, 4, 5\}$ – đầu vào

N: số người dùng

$r_{u'i}$: giá trị xếp hạng của user u' cho item i

$I_{uu'}$: là tập các item được xếp hạng bởi cả user u và $u' \geq threshold$

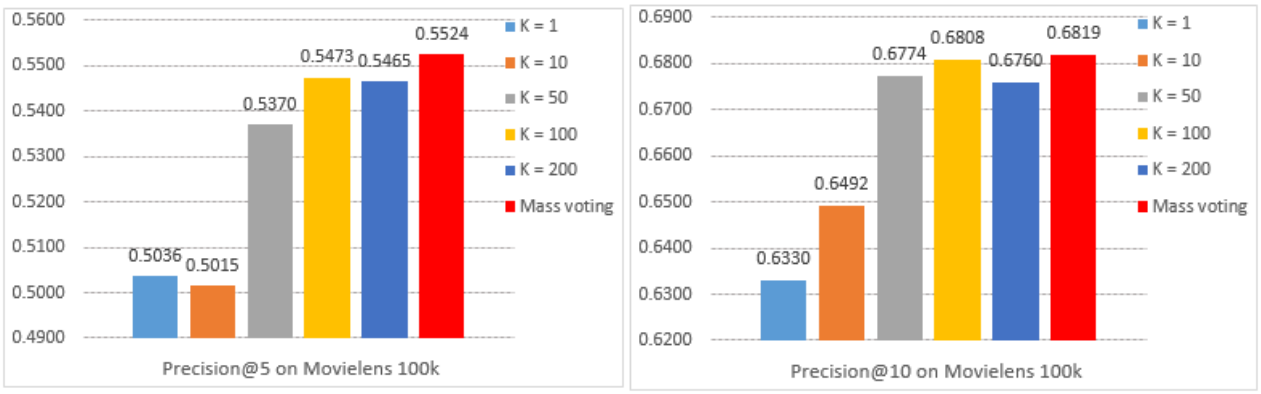
$\hat{r}_{ui}$: giá trị cần dự đoán cho item i của user u – đầu ra

5. Kết quả thực nghiệm

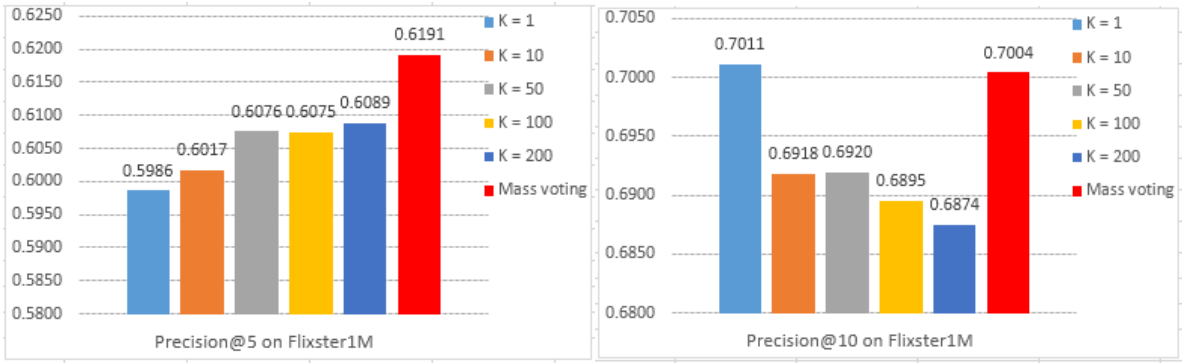
Dữ liệu: Sử dụng 4 tập dữ liệu là MovieLens[1] 100K (943 users, 1.682 items và 100.000 rating), MovieLens 1M (6.040 users, 3.952 items, 1.000.209 ratings), Flixster[2] 1K (2.000 users, 4.000 items, 1.089 ratings) và Flixster 10K (10.000 users, 13.000 items, 10.118 ratings).

Độ đo: Để đánh giá hiệu quả của giải thuật được đề xuất, chúng tôi cài đặt giải thuật luật bình chọn số đông kết hợp ngưỡng sở thích đã trình bày ở phần 4 và kỹ thuật User-KNN để so sánh đối chiếu; sử dụng nghi thức kiểm tra k-fold cross validation với k = 5, dùng độ đo precision@5 và precision@10 (tỉ lệ item gợi ý đúng được trả về trong 5 và 10 items).

Kết quả thực nghiệm trên các tập dữ liệu được trình bày trong các hình dưới đây



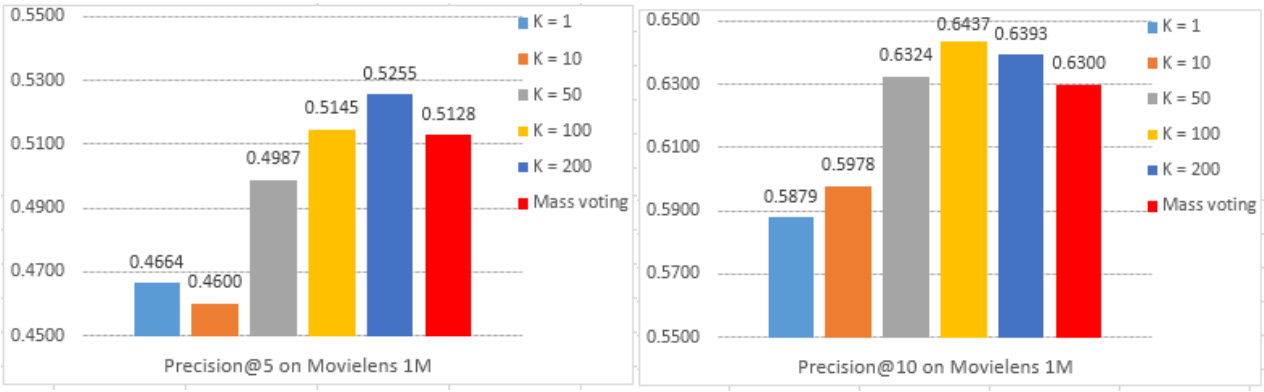
Hình 3. So sánh độ chính xác của các phương pháp trên tập dữ liệu Movielens 100k



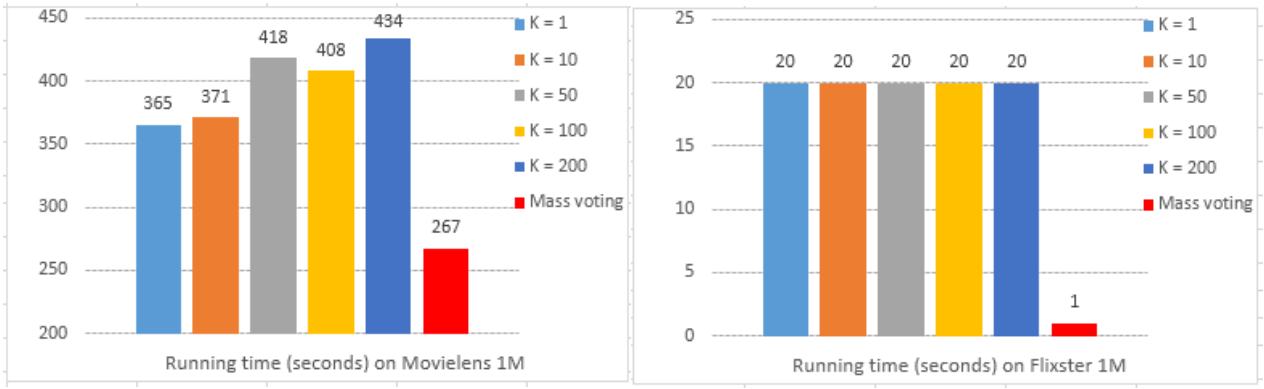
Hình 4. So sánh độ chính xác của các phương pháp trên tập dữ liệu Flixster 1k

[1] © 2014 GroupLens, http://grouplens.org/datasets/movielens/

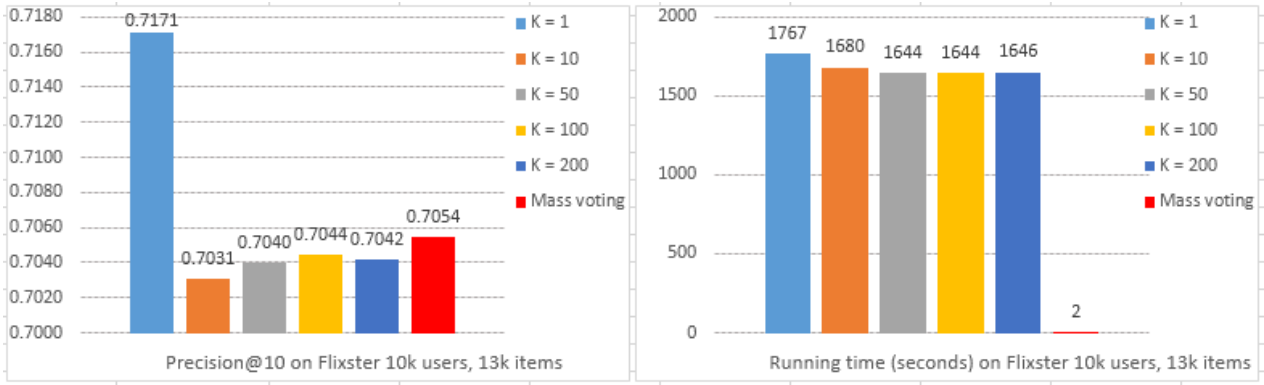
[2] http://www.recsyswiki.com/wiki/Flixster



Hình 5. So sánh độ chính xác của các phương pháp trên tập dữ liệu Movielens 1M



Hình 6. So sánh **thời gian thực thi** (tính bằng giây) của các phương pháp trên 2 tập dữ liệu lớn: Movielens 1M và Flixster 1k



Hình 7. So sánh độ chính xác và **thời gian thực thi** của các phương pháp trên tập dữ liệu Flixster 10k (10k users, 13k items)

Nếu như kỹ thuật User-KNN cần nhiều thời gian thực nghiệm để tìm siêu tham số K thì luật bình chọn số đông kết hợp ngưỡng sở thích khá đơn giản, không cần siêu tham số nhưng cho kết quả gợi ý khá tốt. Đối với các tập dữ liệu MovieLens, luật bình chọn số đông kết hợp ngưỡng sở thích cho kết quả gợi ý tương đương với kết quả gợi ý khi dùng kỹ thuật User-KNN trong khi thời gian thực thi giải thuật có phần tốt hơn, đặc biệt nếu chọn ngưỡng sở thích càng lớn thời gian thực thi giải thuật càng giảm. Điều này thể hiện rõ khi thực nghiệm trên các tập dữ liệu Flixster, khi ma trận xếp hạng cực thưa, thời gian thực thi giải thuật luật bình chọn số đông kết hợp ngưỡng sở thích càng thể hiện ưu điểm vượt trội so với kỹ thuật User-KNN mặc dù kết quả gợi ý thấp hơn kỹ thuật User-KNN nhưng không đáng kể.

6. Kết luận

Qua bài viết này, chúng tôi đã giới thiệu về hệ thống gợi ý, vấn đề dự đoán xếp hạng, các nhóm giải thuật chính trong hệ thống gợi ý và đề xuất một phương pháp mới, đơn giản nhưng khá hiệu quả: Phương pháp luật bình chọn số đông kết hợp ngưỡng sở thích để gợi ý sản phẩm cho người dùng trong hệ thống gợi ý dự đoán xếp hạng. Kết quả thử nghiệm trên các tập dữ liệu chuẩn cho thấy phương pháp được đề xuất có thời gian thực hiện nhanh hơn đáng kể so với các phương pháp truyền thống dựa trên lọc cộng tác trong khi độ chính xác cũng được cải thiện.

Ưu điểm của phương pháp đề xuất là công thức tính toán đơn giản, dễ hiểu, giải thuật dễ cài đặt, kết quả gợi ý tốt, thời gian thực thi nhanh (so với User KNN) và phù hợp với sở thích của người dùng. Tuy nhiên nó có thể không hiệu quả trên tập dữ liệu nhỏ.

TÀI LIỆU THAM KHẢO

[1] R. M. Bell, Y. Koren, „Scalable collaborative filtering with jointly derived neighborhood interpolation weights", In Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), (pp. 43-52). Washington, USA. IEEE CS, 2007.

[2] L. Bottou, "Stochastic learning", In O Bousquet & U. von Luxburg (Eds.), Advanced Lectures on Machine Learning, Lecture Notes in Artificial Intelligence, LNAI 3176, (pp. 146-168). Berlin, Germany: Springer Verlag, 2004.

[3] N. H. Dũng, N. T. Nghe, "Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc cộng tác", Tập chí Khoa học Trường Đại học Cần Thơ, pp. 36-51, 2014.

[4] Z. Gantner, L. Drumond, C. Freudenthaler, S. Rendle, L. Schmidt-Thieme, "Learning attribute-to-feature mappings for cold-start recommendations", In Proceedings of the 10th IEEE International Conference on Data Mining (ICDM 2010). IEEE Computer Society, 2010.

[5] K. R. Koedinger, R. S. J. d. Baker, K. Cunningham, A. Skogsholm, B. Leber, J. Stamper, "A data repository for the EDM community: The PSLC DataShop", In C. Romero, S. Ventura, M. Pechenizkiy, & R. S. J. d. Baker (Eds.), Handbook of educational data mining. Boca Raton, FL: CRC Press, 2010.

[6] Y. Koren, R. Bell, C. Volinsky, "Matrix factorization techniques for recommender systems", IEEE Computer Society Press, 42(8), 30-37, 2009.

[7] N. Manouselis, H. Drachsler, R. Vuorikari, H. Hummel, R. Koper, "Recommender systems in technology enhanced learning", In P. B. Kantor, F. Ricci, L. Rokach, & B. Shapira (Eds.), 1st Recommender Systems Handbook, (pp. 1-29). Berlin, Germany: Springer, 2010.

[8] Đ. T. Nhân, T. N. M. Thư, "Hệ thống gợi ý áp dụng cho trang web tổng hợp tin tức tự động", Tạp chí Khoa học Trường Đại học Cần Thơ, 2013.

[9] I. Pilaszy, D. Tikk, "Recommending new movies: Even a few ratings are more valuable than metadata", In Proceedings of the Third ACM Conference on Recommender Systems (RecSys 2009), (pp. 93-100). New York, NY: ACM, 2009.

[10] F. Ricci, L. Rokach, B. Shapira, P.B. Kantor, "Recommender Systems Handbook", Springer.

[11] X. Su , T.M. Khoshgoftaar, "A survey of collaborative filtering techniques", Advances in Artificial Intelligence, 2009, 4:1-4:19, 2011, 2009.

[12] G. Takacs, I. Pilaszy, B. Nemeth, D. Tikk, "Scalable collaborative filtering approaches for large recommender systems (special topic on mining and learning with graphs and relations)", Journal of Machine Learning Research, 10, 623-656, 2009.

[13] N. Thai-Nghe, L. Drumond, T. Horvath, A. Krohn-Grimberghe, A. Nanopoulos, L. Schmidt-Thieme, "Factorization techniques for predicting student performance", Chapter 6 in Educational Recommender Systems and Technologies: Practices and Challenges (ERSAT 2011), in O.C. Santos & J.G. Boticario (eds.), IGI Global, 2011.

[14] N. Thai-Nghe, Z. Gantner, L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data", In Proceeding of the IEEE International Joint Conference on Neural Networks (IJCNN 2010), 1-8, IEEE Xplore, 2010.

[15] T. V. Viêm, T. Y. Yến, N. T. Nghe, "Xây dựng hệ thống gợi ý phim dựa trên mô hình nhân tố láng giềng", Tạp chí Khoa học Trường Đại học Cần Thơ, 2013.

A COMPOUND MODEL OF USER PREFERENCE THRESHOLD AND VOTING RULE FOR RATING PREDICTION IN RECOMENDER SYSTEMS

Lê Huỳnh Quốc Bảo, Quách Nguyễn Đạt, Nguyễn Thái Nghe

Abstract *- Recommender Systems (RS) are widely used in many fields such as e-commerce, entertainment, education, etc. The purpose of RS is to predict user preferences/behaviors/etc. Based on the prediction results, the system can recommend appropriate items to the users.*

This study proposes a new approach for rating prediction in RS. This approach is a compound of voting rule and user preference threshold to predict the rating of the user. This approach is quite simple and easy to implement but it is effective. Experimental results on standard data sets in RS show that the proposed approach performs much faster than the well-known collaborative filtering approach while its accuracy is also improved in most of the cases. Thus, this could be a promising approach for rating prediction in recommender systems.

Keywords *– Recommender systems, collaborative filtering, voting rule, preference threshold.*