ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

LẬP HỒ SƠ KHÁCH HÀNG TRỰC TUYẾN VÀ GỢI Ý KHÁCH SẠN

Tp. Hồ Chí Minh, tháng 5 năm 2021

I. Bài toán đặt ra

1. Giới thiệu

KDL chia sẻ các thông tin như Nhận xét, đánh giá, xếp hạng, video, ảnh, bài đăng hoặc lượt thích) lên các nền tảng du lịch như TripAdvisor, Expedia hoặc Booking.com (ảnh mình họa nguồn trích dẫn từ TripAdvisor)

Bài báo này, tác giả đề xuất một công cụ đề xuất dựa trên xếp hạng, đánh giá do nguồn cộng đồng cung cấp và thông tin chính thức từ khách sạn (KS).

Phương pháp tiếp cận hồ sơ gồm:

- Mô hình hóa dựa trên thực thể - khách sạn và khách du lịch được thể hiện bằng các xếp hạng liên quan, giá trị đồng tiền (VfM) và giá trị tình cảm (StV)
- Mô hình hóa dựa trên đặc điểm – khách sạn và khách được đại diện bởi các chủ đề liên quan của KS và xếp hạng tương ứng, VfM, StV

Tác giả áp dụng xếp hạng đa tiêu chí (các thông tin khách sạn, SfM, StV) sau đó đề xuất các bộ lọc dự trên vị trí khách sạn, VfM, StV.

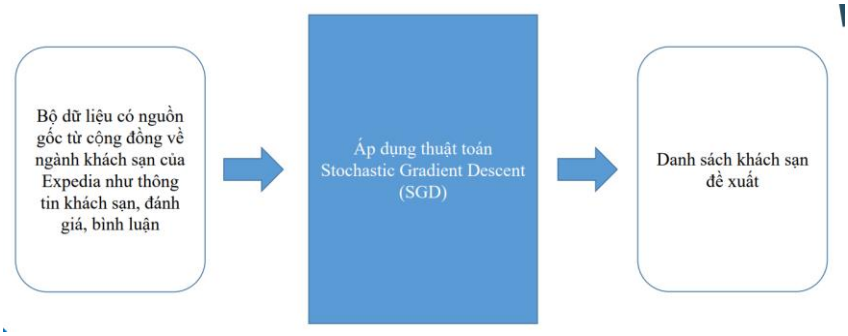
2. Phạm vi

- Tập dữ liệu HotelExpedia, bao gồm 6276 khách sạn và 1090 người dùng, 214342 đánh giá từ 11 địa điểm khác nhau. Mỗi người dùng đã xếp hạng ít nhất 20 khách sạn và mỗi khách sạn có tối thiểu 10 xếp hạng.
- Tuy nhiên, trong bài báo này, tác giả loại bỏ các khách sạn có giá, vì vậy, chỉ còn 3809 khách sạn và 1090 người dùng được xác định và 187892 người đánh giá từ 11 địa điểm khác nhau

3. Đối tượng nghiên cứu

- Hồ sơ khách trực tuyến và thông tin giới thiệu khách sạn như đánh giá, bình luận

4. Mô tả tổng quát bài báo (Đầu vào, tiến trình xử lý, đầu ra)



II. Các nghiên cứu và hướng tiếp cận liên quan

Table 1
Comparison of hotel recommendation research approaches.

Approach	Evaluation	Profiling	Prediction	Post-Filtering
Song et al. (2016)		Rat	ED	
Farokhi et al. (2016)	TripAdvisor	Rat	*k*-means	
Dong and Smyth (2016)	TripAdvisor	Rev	Similarity	
Shrote and Deorankar (2016)		Rev	SA	
Ebadi and Krzyzak (2016)	TripAdvisor	Rat & Rev	SVD & TP	
Sharma et al. (2015)	TripAdvisor	Rat & Rev	NLP	
Hu et al. (2016)	TripAdvisor	Rev	TF-IDF	
Hariri et al. (2011)	TripAdvisor	Rev	CS + LDA	
Current proposal	Expedia	Rat & Rev	SGD	Loc, StV & VfM

1. Hariri et al(2011):

a. Mục tiêu:

- Tạo ra hệ thống khuyến nghị dựa trên mục đích của chuyến du lịch

b. Phương pháp tiếp cận/Kỹ thuật

- Tác giả sử dụng Phân bổ Dirichlet tiềm ẩn (Labelled Latent Dirichlet Allocation) để suy ra mục đích du lịch từ các bài viết đánh giá
- Độ tương tự Cosine (cosine similarity) để ghép nối người dùng với khách sạn tương ứng
- Hệ thống sử dụng dữ liệu của TripAdvisor dể đánh giá kết quả

c. Hạn chế

- Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

2. Sharma et al (2015):

a. Mục tiêu:

- Hệ thống khuyến nghị khách sạn sử dụng quan điểm của người dùng (user preferences)

b. Phương pháp tiếp cận/Kỹ thuật

- Quan điểm của người dùng được trích xuất từ các bài viết đánh giá
- Các bài viết đánh giá được xử lý ngoại tuyến sử dụng kỹ thuật xử lý ngôn ngữ tự nhiên (NPL)
- Hệ thống sử dụng dữ liệu của TripAdvisor để đánh giá kết quả

c. Hạn chế

- Chưa quan tâm đến vấn đề mới như lập hồ sơ dựa vào điểm đánh giá trung bình của chủ đề để đưa râ khuyến nghị mới

- Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

3. **Hu et al (2016):**

   a. **Mục tiêu:**
      - Hệ thống khuyến nghị khách sạn dựa theo mục đích du lịch
   b. **Phương pháp tiếp cận/Kỹ thuật**
      - Mục đích du lịch được suy ra sử dụng các bài viết đánh giá
      - Các bài viết đánh giá được xử lý ngoại tuyến (offline) bằng phân tích tần suất (TF-IDF) từ đó suy ra được mục đích du lịch
      - Hệ thống sử dụng dữ liệu của các khách sạn ở Mỹ được thu thập từ TripAdvisor
   c. **Hạn chế**
      - Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

4. **Ebadi và Krzyzak (2016):**

   a. **Mục tiêu:**
      - Phát triển một hệ thống khuyến nghị khách sạn, kết hợp nhiều điều kiện với nhau (an intelligent hybrid multi-cri-teria)
   b. **Phương pháp tiếp cận/Kỹ thuật**
      - Sử dụng kết hợp cả hình thức đánh giá chấm điểm và các bài viết đánh giá từ TripAdvisor.
      - Ngoài điểm đánh giá, nó áp dụng lập hồ sơ đơn điều kiện để học để dự đoán sở thích của người dùng
      - Ma trận SVD để dự đoán những điểm đánh giá chưa biết
      - Ngoài ra mô hình hóa chủ đề (Topic Modeling) còn được sử dụng để trích xuất nhu cầu của khách hàng từ các bài viết đánh giá
      - Dữ liệu được xử lý ngoại tuyến và đánh giá bằng dữ liệu của TripAdvisor
   c. **Hạn chế**
      - Chưa quan tâm đến vấn đề mới như lập hồ sơ dựa vào điểm đánh giá trung bình của chủ đề để đưa râ khuyến nghị mới
      - Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

5. **Shrote và Deorankar (2016):**

   a. **Mục tiêu:**
      - Đề xuất một phương pháp khuyến nghị sử dụng công nghệ dữ liệu lớn (big data).

b. Phương pháp tiếp cận/Kỹ thuật

- Sử dụng phân tích quan điểm trên các bài viết đánh giá để cá nhân hóa các khuyến nghị
- Hệ thống đánh giá dựa trên dữ liệu không rõ nguồn gốc về các khách sạn ở các thành phố như Dubai, London, Paris...

c. Hạn chế

- Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

6. Dong và Smyth (2016):

a. Mục tiêu:

- Thuật toán trích xuất các thuộc tính của người dùng và khách sạn và xếp hạng các khuyến nghị dựa trên sự tương đồng của người dùng và khách sạn

b. Phương pháp tiếp cận/Kỹ thuật

- Sử dụng phân tích quan điểm của các bài viết đánh giá không kể của người dùng và khách sạn
- Sử dụng thuật toán để đánh giá sự tương đồng giữa người dùng và khách sạn để xếp hạng các khuyến nghị cho người dùng.
- Đánh giá độ chính xác dựa vào dữ liệu của TripAdvisor

c. Hạn chế

- Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

7. Farokhi et al (2016):

a. Mục tiêu:

- Tăng độ chính xác của bộ lọc cộng hưởng (Collaborative Filtering) cho hệ thống khuyến nghị bằng cách kết hợp thêm Fuzzy c-means

b. Phương pháp tiếp cận/Kỹ thuật

- Đầu tiên xác định đánh giá điểm (ratting) đại diện tốt nhất trong nhiều điều kiện
- Sau đó sử dụng đánh giá điểm tổng quát cùng với phân lớp dữ liệu bằng Fuzzy c-mean và k-means để tìm ra lân cận gần nhất
- Cuối cùng dự đoán các điểm đánh giá khách sạn chưa biết bằng hệ số tương quan Pearson (Pearson Correlation coefficient)

c. Hạn chế

- Mô hình chưa được liên tục cập nhật khi có một sự kiện mới xuất hiện do được xử lý ngoại tuyến
- Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

8. Song et al (2016):

a. Mục tiêu:

- Tạo ra hệ thống khuyến nghị dựa vào phân lớp khách sạn bằng đánh giá đa điều kiện (multiple criterial ratings)

b. Phương pháp tiếp cận/Kỹ thuật

- Người dùng phải có một ngưỡng đánh giá nhất định mới nhận được các khuyến nghị cá nhân hóa
- Hệ thống xác định sự tương tự giữa các tiêu chí của người dùng và đánh giá trung bình của khách sạn dựa vào khoảng cách Euclid (Euclidean Distance)

c. Hạn chế

- Chưa sử dụng các thông tin sẵn có như giá trị quan điểm, giá trị đối với chi phí bỏ ra (sentiment value, value of money)

9. Tổng kết

Tóm lại, cách tiếp cận trước đây tập trung vào việc xử lý dữ liệu trực tuyến, phụ thuộc chủ yếu vào đánh giá của thực thể (entity-based rattings) để tạo mô hình khách hàng và khách sạn và có xu hướng bỏ qua những thông tin có sẵn của dữ liệu cộng đồng như vị trí, giá trị quan điểm, giá trị đối với chi phí bỏ ra …. Tác giả đã sử dụng thêm những thông tin này như một bộ lọc để cải thiện hiệu quả của gợi ý

Tác giả đã cải thiện cách tiếp cận của **Shrote cùng với Deorankar (2016)** và **Sharma et al. (2015)** bằng cách tiếp cận với cả thực thể và chủ đề (entity-based and theme-based) cụ thể là cách lập hồ sơ bằng chủ đề không chỉ tránh được các vấn đề khi có đối tượng mới (như hồ sơ của khách sạn mới được dựa vào điểm đánh giá trung bình của chủ đề, giúp cho khách sạn mới có thể được gợi ý) mà còn đảm bảo việc giảm số chiều - một trong những khát khao khi xử lý một lượng dữ liệu lớn.

Tác giả cải thiện kết quả của **Ebadi cùng với Krzyzak (2016)** và **Sharma et al. (2015)** trong việc kết hợp sử dụng chấm điểm và các bài viết đánh giá bằng cách chuẩn hóa đánh giá quan điểm của **Dong cùng với Smyth (2016)** và **Shrote cùng với Deorankar (2016)** và sử dụng các thông tin thu được như một bộ lọc để cải thiện kết quả gợi ý

Tác giả sử dụng cập nhật tăng cường để cải thiện cách xử lý dữ liệu trực tuyến của Song et al. (2016) và Farokhi et al. (2016) dựa vào SGD và giúp cập nhật mô hình liên tục khi có sự kiện mới.

Tác giả đã đóng góp:

- Hệ thống gợi ý trực tuyến cho lĩnh vực du lịch có kết hợp cập nhật tăng cường dựa vào luồng dữ liệu cộng đồng
- Lập hồ sơ dựa vào chủ đề để giảm thiểu vấn đề khi có đối tượng mới
- Bộ lọc gợi ý để tăng độ chính xác của gợi ý.

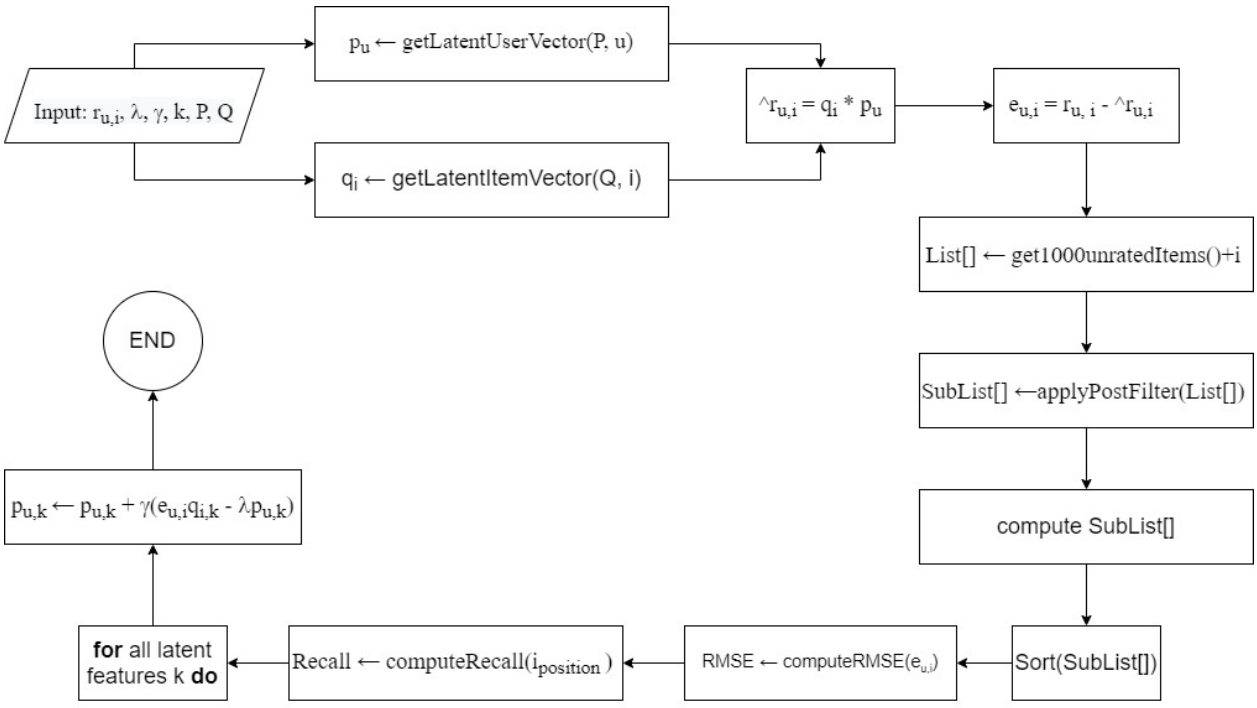
III. Mô hình thuật toán đề xuất cải tiến

Module dự đoán xếp hạng thực hiện lọc đề xuất cộng tác để dự đoán xếp hạng khách sạn chưa được cung cấp bởi khách du lịch. Cụ thể, bài báo này khám phá kỹ thuật SGD để tạo các vectơ đặc trưng tiềm ẩn và cập nhật từng bước mô hình bằng cách sử dụng các phương pháp lập hồ sơ dựa trên khách sạn và dựa trên chủ đề.

Bài báo này tác giả đề xuất 2 thuật toán để xây dựng mô hình dự báo dựa trên cơ sở khách sạn và chủ đề, bên cạnh đó bài báo còn sử dụng bộ lọc sau đề xuất với mục đích cải thiện các đề xuất cuối cùng được cung cấp cho người dùng tận dụng dữ liệu hồ sơ sẵn có.

1. Thuật toán luồng dựa trên khách sạn

a. Mô hình thuật toán



b. Mô tả

Thuật toán mô tả dự đoán xếp hạng bằng cách sử dụng hồ sơ dựa trên khách sạn với các luồng dữ liệu. Gồm các thao tác như sau:

Bước 1: Khởi tạo ma trận xếp hạng với các xếp hạng khách hàng.

Bước 2: Xây dựng ma trận tiềm ẩn cho khách hàng và khách sạn, phân phối ngẫu nhiên một thành phần phạm vi nhỏ từ -0,02 đến 0,02 để đảm bảo các hệ số là khác nhau.

Bước 3: Tạo ra các dự đoán bằng cách sử dụng ma trận tiềm ẩn. Sau đó, tác giả chọn ngẫu nhiên 1000 khách sạn không được xếp hạng cộng với khách sạn mới được xếp hạng, tức là 1001 khách sạn. Thuật toán dự đoán xếp hạng của 1001 khách sạn đã chọn và sắp xếp chúng theo thứ tự giảm dần.

Quá trình tối ưu hóa siêu tham số xác định learning rate tối ưu () và overfitting () bằng cách sử dụng RMSE. Các tham số này được sử dụng để cập nhật ma trận tiềm ẩn của khách hàng và khách sạn. Do đó, đối với mỗi xếp hạng mới ta sẽ thực hiện:

Thêm xếp hạng vào khởi tạo ma trận xếp hạng.

Tính toán sai số dự đoán và các chỉ số độ chính xác.

Cập nhật ma trận tiềm ẩn của khách hàng và khách sạn bằng cách sử dụng các siêu tham số. Cuối cùng, phương pháp tạo ra các dự đoán mới bằng cách sử dụng ma trận tiềm ẩn khách hàng và khách sạn được cập nhật. Phương pháp được lặp lại cho mỗi xếp hạng đầu vào.

Độ phức tạp của thuật toán là Nlog N().

c. Pseudocode

1: Input: $r_{u,i}$, λ, γ, k, P, Q

2: Output: $\hat{r}_{u,i}$

3: $p_u \leftarrow$ getLatentUserVector(P, u)

4: $q_i \leftarrow$ getLatentItemVector(Q, i)

5: $\hat{r}_{u,i} = q_i * p_u$

6: $e_{u,i} = r_{u,i} - \hat{r}_{u,i}$

7: List[] ← get1000unratedItems()+i

8: SubList[] ← applyPostFilter(List[])

9: **for** i ← SubList [] **do**

10: q ← getLatentItemVector(Q, i)

11: $\hat{r}_{u,i} = q_i * p_u$

12: Sort(SubList[])

13: RMSE ← computeRMSE($e_{u,i}$)

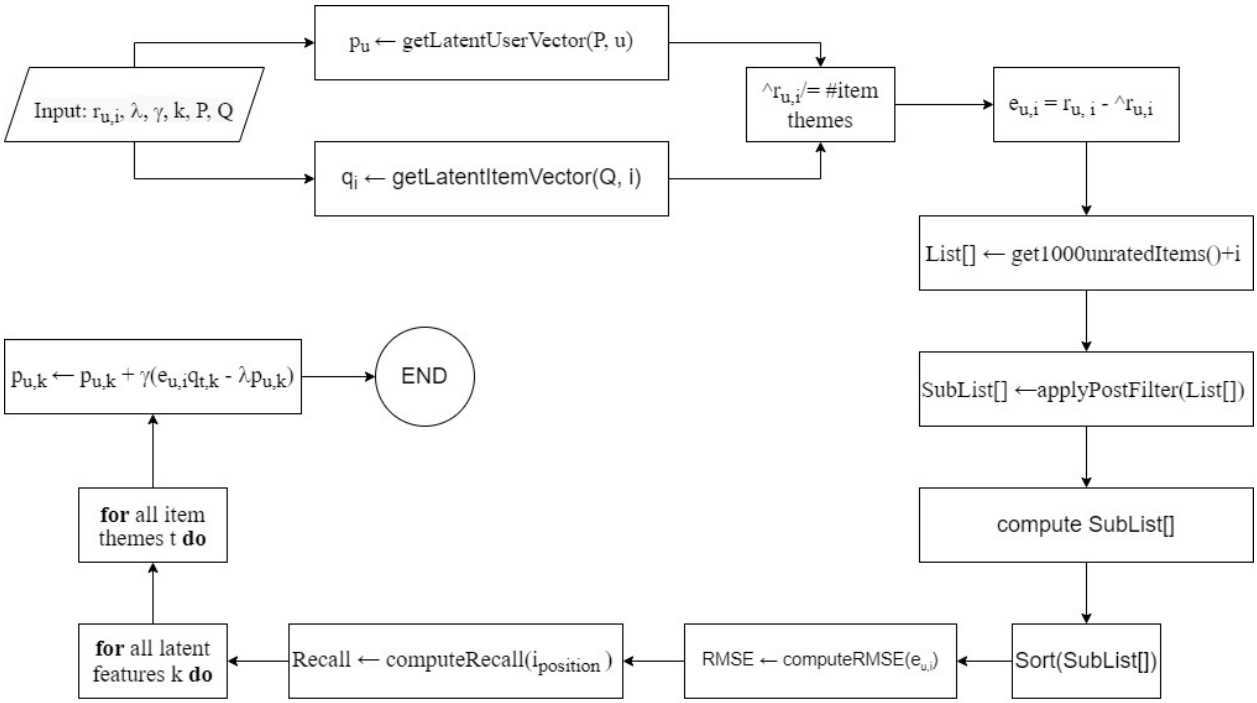
14: Recall ← computeRecall($i_{position}$)

15: **for** all latent features k **do**

16: $p_{u,k} \leftarrow p_{u,k} + \gamma(e_{u,i}q_{i,k} - \lambda p_{u,k})$

2. Thuật toán luồng dựa trên chủ đề

a. Mô hình thuật toán



b. Mô tả

Thuật toán mô tả dự đoán xếp hạng bằng cách sử dụng cấu hình dựa trên chủ đề với các luồng dữ liệu. Hồ sơ của một khách sạn mới dựa trên xếp hạng trung bình của các khách sạn được phân loại cùng chủ đề. Do đó, thuật toán luồng dựa trên chủ đề, gồm các bước:

Bước 1: Tạo ma trận xếp hạng ban đầu với xếp hạng của khách liên quan đến chủ đề của khách sạn

Bước 2: Xây dựng ma trận khách và chủ đề tiềm ẩn phân phối ngẫu nhiên một thành phần từ phạm vi nhỏ −0,02 đến 0,02 để đảm bảo các hệ số khác nhau.

Bước 3: Tạo ra các dự đoán bằng cách sử dụng ma trận tiềm ẩn. Sau đó, tác giả chọn ngẫu nhiên 1000 khách sạn không được xếp hạng cộng với

khách sạn mới được xếp hạng, tức là 1001 khách sạn. Thuật toán dự đoán xếp hạng của các chủ đề liên quan đến 1001 khách sạn đã chọn và sắp xếp chúng theo thứ tự giảm dần.

Quá trình tối ưu hóa siêu tham số tuân theo thuật toán dựa trên khách sạn.

Độ phức tạp của thuật toán là N^2

c. Pseudocode

1: Input: $r_{u,i}$, λ, γ, k, P, Q

2: Output: $\hat{r}_{u,i}$

3: $p_u \leftarrow$ getLatentUserVector(P, u)

4: **for** all item themes t **do**

5: $q_t \leftarrow$ getLatentItemVector(Q, i)

6: $\hat{r}_{u,i} += q_t * p_u$

7: $\hat{r}_{u,i}$ /= #item themes

8: $e_{u,i} = r_{u,i} - \hat{r}_{u,i}$

9: List[] ← getunratedItems1000()+i

10: SubList[] ← applyPostFilter(List[])

11: **for** i ← SubList [] **do**

12: **for** all item themes t **do**

13: $q_t \leftarrow$ getLatentItemVector(Q, t)

14: $\hat{r}_{u,i} += q_t * p_u$

15: $\hat{r}_{u,i}$ /= #item themes

16: Sort(SubList[])

17: RMSE computeRMSE($e_{u,i}$)

18: Recall computeRecall($i_{position}$)

19: **for** all latent features k **do**

20: **for** all item themes t **do**

21: $p_{u,k} \leftarrow p_{u,k} + \gamma(e_{u,i}q_{t,k} - \lambda p_{u,k})$

IV.Bộ dữ liệu thực nghiệm

Bộ dữ liệu HotelExpedia được đề xuất bởi Leal et al. (2017), và nó bao gồm 6276 khách sạn, 1090 người dùng đã xác định và 214.342 đánh giá từ 11 địa điểm khác nhau. Mỗi người dùng đã xếp hạng ít nhất 20 khách sạn và mỗi khách sạn có tối thiểu 10 xếp hạng. Tuy nhiên, đối với các thử nghiệm của tác giả đã loại bỏ các khách sạn không có giá. Do đó, các thử nghiệm đã được thực hiện với 3809 khách sạn, 1090 người dùng được xác định và 187892 người đánh giá từ 11 địa điểm khác nhau.

File	Features
Hotels	hotelId, description, latitude-longitude, **starRating**, guestReviewCount, **price**, amenity, **overall**, recommendedPercent, cleanliness, serviceAndStaff, roomComfort, and hotelCondition
Guests	**nickname**, userLocation, **hotelId**, **overall**, **cleanliness**, **hotelCondition**, **serviceAndStaff**, **roomComfort**, **reviewText**, and **timestamp**

	Post-Filter	RMSE	CV	R@10	CV	TR@10	CV
H-b MRR		*0.190*	*0.2*	*0.170*	*1.8*	*0.126*	*1.6*
	Loc	0.190	0.1	0.309	1.4	0.217	1.3
	Loc & VfM	0.190	0.2	0.454	1.7	0.308	1.4
	Loc & StV	0.190	0.2	0.546	1.4	0.360	1.2
H-b PWRA		*0.168*	*0.0*	*0.175*	*2.1*	*0.144*	*2.0*
	Loc	0.168	0.1	0.316	1.4	0.242	1.2
	Loc & VfM	0.168	0.0	0.476	1.5	0.353	1.3
	Loc & StV	**0.168**	0.1	**0.566**	1.5	**0.410**	1.4
T-b MRR		*0.240*	*0.0*	*0.017*	*4.2*	*0.027*	*2.0*
	Loc	0.240	0.0	0.154	1.3	0.131	1.1
	Loc & VfM	0.240	0.0	0.262	1.0	0.186	0.8
	Loc & StV	0.240	0.0	0.462	0.6	0.286	0.5
T-b PWRA		*0.222*	*0.0*	*0.016*	*5.5*	*0.028*	*1.3*
	Loc	0.222	0.0	0.160	1.8	0.141	1.3
	Loc & VfM	0.222	0.0	0.266	0.8	0.209	0.6
	Loc & StV	**0.222**	0.0	**0.469**	0.4	**0.336**	0.5

Bảng mô tả nội dung của tập dữ liệu, làm nổi bật dữ liệu được sử dụng.

Các thử nghiệm của tác giả, dựa trên dữ liệu đánh giá của người dùng và khách sạn, sử dụng, cụ thể là biệt hiệu của người dùng, nhận dạng khách sạn, mô tả, đánh giá bằng văn bản và như xếp hạng đa tiêu chí, tổng thể, mức độ sạch sẽ, dịch vụ & nhân viên, tình trạng khách sạn và phòng thoải mái. Tập dữ liệu này không chứa xếp hạng rỗng, tức là tất cả người dùng đã xếp hạng các thành phần đa tiêu chí của khách sạn. Trong trường hợp của tập dữ liệu HotelExpedia, MRR là xếp hạng tổng thể.

V. Kết quả thực nghiệm và đánh giá

Tác giả đã tiến hành một số thử nghiệm với bộ dữ liệu của Expedia để đánh giá hiệu quả và tính hữu ích của phương pháp được đề xuất. Các thử nghiệm liên quan đến việc lập hồ sơ dựa trên khách sạn và dựa trên chủ đề bằng cách sử dụng MRR và PWRA cũng như đánh giá dự đoán xếp hạng tương ứng với các bộ lọc sau đề xuất (vị trí, VfM và StV). Cuối cùng, tác giả đánh giá hệ thống bằng cách sử dụng RMSE, Recall@10 và Target Recall@10, cũng như ROC curve. Hơn nữa, tác giả áp dụng phép thử Wilcoxon và McNemar để bỏ các null-hypothesis.

Phương pháp đánh giá xác định thứ tự và phân phối dữ liệu. Dữ liệu được sắp xếp theo thứ tự tạm thời để áp dụng công cụ đề xuất luồng được đề xuất. Mô hình được cập nhật từng bước, sử dụng toàn bộ tập dữ liệu dưới dạng các luồng. Do đó, khi khách xếp hạng khách sạn, phương pháp này không chỉ cập nhật các dự đoán, hồ sơ khách và khách sạn mà còn cập nhật các chỉ số đánh giá.