

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**LUẬN VĂN TỐT NGHIỆP
NGÀNH KHOA HỌC MÁY TÍNH**

Đề tài

**HỆ THỐNG GỢI Ý NHÀ TRỌ, KHÁCH SẠN
Ở CẦN THƠ
(A RECOMMENDATION SYSTEM FOR HOTELS,
BOARDING HOUSES IN CAN THO)**

Sinh viên thực hiện: Nguyễn Hữu Tính

Mã số: B1710355

Khóa: 43

Cần Thơ, 12/2021

**BỘ GIÁO DỤC VÀ ĐÀO TẠO
TRƯỜNG ĐẠI HỌC CẦN THƠ
KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**



**LUẬN VĂN TỐT NGHIỆP
NGÀNH KHOA HỌC MÁY TÍNH**

Đề tài

**HỆ THỐNG GỢI Ý NHÀ TRỌ, KHÁCH SẠN
CẦN THƠ
(A RECOMMENDATION SYSTEM FOR HOTELS,
BOARDING HOUSES IN CAN THO)**

**Giáo viên hướng dẫn:
TS. Trần Nguyễn Dương Chi**

**Sinh viên thực hiện:
Nguyễn Hữu Tính
Mã số: B1710355
Khóa: 43**

Cần Thơ, 12/2017

NHẬN XÉT CỦA GIẢNG VIÊN

Cần Thơ, ngày tháng năm
(GVHD ký và ghi rõ họ tên)

LỜI CẢM ƠN

Để có được bài niên luận này, em xin được bày tỏ lòng biết ơn chân thành và sâu sắc đến Cô Trần Nguyễn Dương Chi – người đã trực tiếp tận tình hướng dẫn, giúp đỡ em. Trong suốt quá trình thực hiện niên luận, nhờ những sự chỉ bảo và hướng dẫn quý giá đó mà bài niên luận này được hoàn thành một cách tốt nhất.

Em cũng xin gửi lời cảm ơn chân thành đến các Thầy Cô Giảng viên Đại học Cần Thơ, đặc biệt là các Thầy Cô ở Khoa CNTT & TT, những người đã truyền đạt những kiến thức quý báu trong thời gian qua.

Em cũng xin chân thành cảm ơn bạn bè cùng với gia đình đã luôn động viên, khích lệ và tạo điều kiện giúp đỡ trong suốt quá trình thực hiện để em có thể hoàn thành bài niên luận một cách tốt nhất.

Tuy có nhiều cố gắng trong quá trình thực hiện niên luận, nhưng không thể tránh khỏi những sai sót. Em rất mong nhận được sự đóng góp ý kiến quý báu của quý Thầy Cô và các bạn để bài niên luận hoàn thiện hơn.

Cần Thơ, ngày tháng 12 năm 2017

Người viết

Nguyễn Hữu Tính

MỤC LỤC

NHẬN XÉT CỦA GIẢNG VIÊN	iii
LỜI CẢM ƠN.....	iv
MỤC LỤC	1
DANH MỤC HÌNH	4
DANH MỤC BẢNG.....	5
ABSTRACT.....	6
TÓM TẮT.....	7
PHẦN GIỚI THIỆU	1
1. Đặt vấn đề:	1
2. Lịch sử giải quyết vấn đề:	2
3. Mục tiêu đề tài: ghi rõ lại.....	3
4. Phương pháp nghiên cứu:	3
5. Đối tượng và phạm vi nghiên cứu:	3
6. Kết quả đạt được:	3
7. Bố cục luận văn:	3
PHẦN NỘI DUNG.....	4
CHƯƠNG 1 MÔ TẢ BÀI TOÁN	4
1.1. Mô tả chi tiết bài toán:	4
1.2. Giới thiệu về hệ thống gợi ý:.....	4
1.2.1. Giới thiệu:	4
1.2.2. Hệ thống gợi ý (recommender systems):	5
1.3. Các kỹ thuật chính trong hệ thống gợi ý:	6
1.3.1. Lọc cộng tác:.....	7
1.3.2. Lọc dựa trên nội dung:.....	8
1.3.3. Hệ thống gợi ý lai (hybrid recommender systems):.....	10
1.3.4. Các kỹ thuật không cá nhân hóa:	11
1.4. Deep learning trong hệ thống gợi ý:.....	12
1.5. Các phương pháp đánh giá:.....	12
1.5.1. Tiêu chí định lượng:	13
1.5.1.1. Đánh giá độ chính xác của các dự đoán:	13
1.5.1.2. Đánh giá việc sử dụng các dự đoán:.....	13
1.5.2. Tiêu chí định tính:	16

1.5.2.1.	Tính mới của các gợi ý:.....	16
1.5.2.2.	Tính đa dạng (Diversity) của các gợi ý:	17
1.5.2.3.	Độ bao phủ (coverage) của các gợi ý:	18
1.5.2.4.	Sự hài lòng của người sử dụng:.....	18
CHƯƠNG 2 THIẾT KẾ VÀ CÀI ĐẶT		20
2.1.	Tập dữ liệu:	20
2.1.1.	Tiền xử lý dữ liệu:.....	22
2.2.	Hệ thống gợi ý nhà trọ, khách sạn sử dụng đề xuất dựa trên một số tiêu chí người dùng cần:.....	31
2.2.1.	Giải pháp:.....	31
2.2.2.	Đề xuất dựa trên tiêu chí người dùng:	31
2.2.3.	Áp dụng vào bài toán thực tế với cơ sở dữ liệu datasets:	32
CHƯƠNG 3 HỆ THỐNG THỬ NGHIỆM.....		37
3.1.	Mục đích của việc xây dựng website:.....	37
3.2.	Sơ đồ hệ thống tổng quát:	38
3.3.	Phân tích hệ thống người dùng website:	39
3.4.	Đặc tả quy trình nghiệp vụ của hệ thống:	40
3.4.1.	Người dùng không có tài khoản:	40
3.4.2.	Người dùng có đăng ký tài khoản trên website:	40
3.4.3.	Người dùng hệ thống (Admin):	40
3.5.	Lập mô hình nghiệp vụ:	41
3.5.1.	Biểu đồ ngữ cảnh hệ thống:	41
3.5.2.	Biểu đồ phân rã chức năng:	42
3.5.3.	Phân rã biểu đồ luồng dữ liệu:	43
3.5.	Thiết kế các bảng dữ liệu:	46
3.5.1.	Bảng User:	46
3.5.2.	Bảng Motel:	47
3.5.3.	Bảng Comment:	47
3.5.4.	Bảng Hotel:	48
3.5.5.	Bảng Reviews:	49
3.5.6.	Bảng Booking:	49
3.5.7.	Bảng Rooms:.....	50
3.6.	Mô tả giao diện của website:	50
3.7.	Giao diện website:	53
3.7.1.	Giao diện trang chủ:.....	53
3.7.3.	Chức năng đăng ký – đăng nhập:.....	57

3.7.4. Chức năng đặt phòng:	58
3.7.5. Chức năng bình luận:	58
PHẦN KẾT LUẬN	60
1. Kết quả đạt được:	60
2. Hạn chế:	60
3. Hướng phát triển:	60
TÀI LIỆU THAM KHẢO	61

DANH MỤC HÌNH

Hình 1.1 Hệ thống gợi ý sản phẩm Amazon.....	5
Hình 1.2 Ma trận biểu diễn dữ liệu RS	6
Hình 1.3 Gợi ý sản phẩm thường được mua	12
Hình 1.4: Sự phổ biến của sản phẩm.....	16
Hình 2.1: Sơ đồ tiền xử lý dữ liệu.....	24
Hình 2.2: Sơ đồ gợi ý nhà trọ - khách sạn.....	33
Hình 3.1: Sơ đồ tổng quát Website gợi ý nhà trọ - khách sạn	38
Hình 3.2: Hệ thống người dùng website	39
Hình 3.3: Biểu đồ ngữ cảnh hệ thống	41
Hình 3.4: Biểu đồ phân rã chức năng.....	42
Hình 3.5: Biểu đồ luồng phân rã cấp 1.0	43
Hình 3.6: Biểu đồ luồng phân rã cấp 2.0	43
Hình 3.7: Biểu đồ luồng phân rã cấp 2.1	44
Hình 3.8: Biểu đồ luồng phân rã cấp 2.2	44
Hình 3.9: Biểu đồ luồng phân rã cấp 2.3	45
Hình 3.10: Giao diện chính của website	51
Hình 3.11: Website gợi ý nhà trọ và khách sạn	52
Hình 3.12: Giao diện đồ họa trang chính	54
Hình 3.13: Giao diện đồ họa khách sạn – nhà trọ	56
Hình 3.14: Đăng ký người dùng.....	57
Hình 3.15: Đăng nhập người dùng	57
Hình 3.16: Chức năng đặt phòng	58
Hình 3.17: Chức năng bình luận	59

DANH MỤC BẢNG

Bảng 1: Tập dữ liệu nhà trọ trước khi tiền xử lý – Datasets_motels.csv	22
Bảng 2: Tập dữ liệu khách sạn trước tiền xử lý – Datasets_hotels.csv	23
Bảng 3: Bảng thay đổi kiểu dữ liệu.....	24
Bảng 4: Bảng quy đổi giá trị ở cột “rating_title”	25
Bảng 5: Dữ liệu bị khuyết (rỗng) của tập dữ liệu Datasets_hotels.csv	25
Bảng 6: Dữ liệu đánh giá của khách hàng.....	28
Bảng 7: Tập dữ liệu nhà trọ sau tiền xử lý dữ liệu.....	29
Bảng 8: Tập dữ liệu khách sạn sau tiền xử lý dữ liệu	30
Bảng 10: Gợi ý nhà trọ.....	36
Bảng 15: <i>Bảng User</i>	47
Bảng 16: <i>Bảng Motel</i>	47
Bảng 17: Bảng Reviews_motel.....	48
Bảng 18: Bảng Hotel.....	49
Bảng 19: Bảng Reviews_hotel.....	49
Bảng 20: Bảng Booking	50
Bảng 21: Bảng Rooms	50

ABSTRACT

Recommender systems can offer relevant information items to users by using data about their past behavior to predict items that users might like. Two successful approaches in the recommender system are collaborative filtering and k-nearest neighbor. In this thesis, another approach is introduced, which is a content-based recommendation system. This research is carried out to collect data of hotels and boarding houses in Cantho city, and building a web-based recommender system for hotel and boarding house in Cantho.

TÓM TẮT

Hệ thống gợi ý có thể đưa ra những mục thông tin phù hợp cho người dùng bằng cách dựa vào dữ liệu về hành vi trong quá khứ của họ để dự đoán những mục thông tin mới trong tương lai mà người dùng có thể thích. Hai nhánh tiếp cận thành công trong hệ thống gợi ý thuộc vào nhóm lọc công tác là mô hình nhân tố tiềm ẩn - xác định mối quan hệ tiềm ẩn trên cả người dùng và mục thông tin; và mô hình láng giềng - phân tích độ tương tự giữa các mục thông tin với nhau hay giữa những người dùng với nhau. Trong luận văn này, giới thiệu một tiếp cận ở nhánh khác là hệ thống gợi ý dựa trên nội dung. Ở đây, bên cạnh việc xây dựng một hệ thống trên nền web để gợi ý nhà trọ - khách sạn tại Cần Thơ, luận văn cũng đã điều chỉnh mô hình đã có để phù hợp hơn với yêu cầu bài toán. Song song đó, luận văn này nhằm tạo được bộ dữ liệu về nhà trọ khách sạn Cần Thơ gồm 408 nhà trọ và 461 khách sạn, kèm với thu thập đánh giá người dùng, từ đó gợi ý nhà trọ hoặc khách sạn phù hợp với nhu cầu người dùng.

PHẦN GIỚI THIỆU

1. Đặt vấn đề:

Việc tìm nơi ở phù hợp khi đến một thành phố mới là điều không dễ dàng trước lượng thông tin khổng lồ hiện nay. Hệ thống gợi ý nhà trọ, khách sạn nhằm giúp sinh viên, người tìm việc, du khách có thể tìm được nơi ở phù hợp tại Cần Thơ.

Một trong những điều đầu tiên cần làm khi lên kế hoạch cho một chuyến đi hay chuẩn bị cho một quá trình học tập xa nhà là tìm một nơi tốt để ở. Việc lựa chọn phòng khách sạn trực tuyến hay nhà trọ có thể là một nhiệm vụ quá sức với hàng ngàn lựa chọn, cho mọi điểm đến. Với tầm quan trọng của những tình huống này, nhiệm vụ giới thiệu khách sạn – nhà trọ cho người dùng được đề xuất.

Mục đích của hệ thống gợi ý nhà trọ - khách sạn là dự đoán và đề xuất nhóm các khách sạn phù hợp với tiêu chí người dùng trong hàng trăm khách sạn – nhà trọ thuộc khu vực tỉnh Cần Thơ.

Với sự phát triển của công nghệ web mới, hệ thống gợi ý (RS) đang nhận được sự quan tâm đáng kể của giới kinh doanh cũng như khách hàng do vai trò của nó đối với thương mại điện tử, chiến lược kinh doanh tinh tế, cải thiện sự hài lòng của khách hàng, v.v. Thành công của hệ thống thương mại điện tử hiện đại và hệ thống đặt chỗ và đặt chỗ trực tuyến phụ thuộc rất nhiều vào

sự hài lòng và tin tưởng của khách hàng. Mariani và cộng sự^[1] đã quan sát thấy rằng du lịch là một trong những nổi tiếng và là một ngành công nghiệp mạnh mẽ trên thế giới có tác động to lớn đến tổng GDP hoặc việc làm. Du lịch gắn liền với khách sạn vì khách du lịch luôn muốn biết về các khách sạn nơi họ sẽ lưu trú trong chuyến tham quan của họ. Trong những năm gần đây, khách sạn trực tuyến đặt phòng đã trở thành một trong những lựa chọn hàng đầu của khách hàng. Một vài hệ thống gợi ý cũng được phát triển trong quá khứ để tạo điều kiện cho khách hàng có những đề xuất khách sạn trước khi tìm phòng hoặc đặt phòng Liu et al.^[2] Tuy nhiên, những hệ thống này là dữ liệu chung và chỉ xử lý dữ liệu đồng nhất; trong khi bản chất của hầu hết dữ liệu trên web là không đồng nhất đó là một nút thắt lớn trong hoạt động của hệ thống gợi ý khách sạn.

Luận văn trình bày một cách tiếp cận thông minh để xử lý dữ liệu không đồng nhất và có kích thước lớn sử dụng máy học để tạo ra các đề xuất thực sự có ích cho khách hàng. Phương pháp lọc cộng tác (CF) là một trong những kỹ thuật phổ biến nhất của Hệ thống gợi ý (Recommender System - RS) để tạo các đề xuất. Luận văn đề xuất một cách tiếp cận gợi ý CF mới trong đó phân tích tình cảm dựa trên quan điểm của khách hàng. Cách tiếp cận này kết hợp phân tích từ vựng, phân tích cú pháp và phân tích ngữ nghĩa để hiểu tình cảm đối với các đặc điểm của khách sạn và hồ sơ của loại khách hàng. Hệ

thống đề xuất đề xuất các khách sạn dựa trên các đặc điểm của khách sạn và khách hàng nhập làm thông tin bổ sung cho đề xuất được cá nhân hóa. Hệ thống được phát triển không chỉ có khả năng xử lý dữ liệu không đồng nhất mà còn đề xuất khách sạn dựa trên loại khách sử dụng các quy tắc mờ.

2. Lịch sử giải quyết vấn đề:

Các kỹ thuật lọc cộng tác CF chủ đạo dựa vào tính tương đồng giữa những người sử dụng. Tương tự về mặt người dùng hoặc mặt hàng được phát hiện bằng cách tính toán các điểm tương đồng về xếp hạng chung của người dùng [4]. Phương pháp CF hoạt động tốt khi có đủ thông tin xếp hạng [6]. Tuy nhiên, hiệu quả của chúng bị ảnh hưởng khi xảy ra sự cố thừa thớt xếp hạng, vì lý do thường có một điểm chung bị hạn chế xếp hạng số giữa những người dùng [7]. Một hạn chế khác là các phương pháp tiếp cận CF không nắm bắt được lý do cho xếp hạng của người dùng và do đó không thể nắm bắt chính xác sở thích của người dùng mục tiêu [8].

Để giải quyết những vấn đề này, một số phương pháp dựa trên nội dung đã được phát triển để đại diện cho người dùng và các mục theo nhiều loại dữ liệu khác nhau, bao gồm các thẻ [9], mô tả các mặt hàng [10], và các yếu tố xã hội [11].

Tóm lại, những kỹ thuật này vẫn còn thiếu sót, đặc biệt khi mức độ thừa thớt xếp hạng là chính, hoặc người dùng không có nhiều xếp hạng lịch sử [6]. Với kịch bản hiện tại của Web, người dùng ngày càng trở nên thoải mái hơn trong việc thể hiện bản thân và chia sẻ quan điểm của họ về những vấn đề liên quan đến các sản phẩm trên nền tảng điện tử sử dụng các bài đánh giá văn bản [12]. Với sự phát triển của công nghệ web mới, hệ thống giới thiệu (RS) đang nhận được sự quan tâm đáng kể của giới kinh doanh cũng như khách hàng do vai trò của nó đối với thương mại điện tử, chiến lược kinh doanh tinh tế, cải thiện sự hài lòng của khách hàng, v.v. Thành công của hệ thống thương mại điện tử hiện đại và hệ thống đặt chỗ và đặt chỗ trực tuyến phụ thuộc rất nhiều vào

sự hài lòng và tin tưởng của khách hàng. Mariani và cộng sự [1] đã quan sát thấy rằng du lịch là một trong những nổi tiếng và là một ngành công nghiệp mạnh mẽ trên thế giới có tác động to lớn đến tổng GDP hoặc việc làm. Du lịch gắn liền với khách sạn vì khách du lịch luôn muốn biết về các khách sạn nơi họ sẽ lưu trú trong chuyến tham quan của họ. Trong những năm gần đây, khách sạn trực tuyến đặt phòng đã trở thành một trong những lựa chọn hàng đầu của khách hàng. Một vài hệ thống gợi ý cũng được phát triển trong quá khứ để tạo điều kiện cho khách hàng có những đề xuất khách sạn trước khi tìm phòng hoặc đặt phòng Liu et al. [2]. Tuy nhiên, những hệ thống này là dữ liệu chung và chỉ xử lý dữ liệu đồng nhất; trong khi bản chất của hầu hết dữ liệu trên web là không đồng nhất đó là một nút thắt lớn trong hoạt động của hệ thống khuyến nghị khách sạn.

3. Mục tiêu đề tài: ghi rõ lại

Mục tiêu của luận văn là tìm hiểu về hệ thống gợi ý, sau đó xây dựng thuật toán dựa trên lý thuyết và đánh giá kết quả trên dữ liệu thực tế.

Thu thập đánh giá của khách hàng về các nhà trọ, khách sạn ở Cần Thơ.

Xây dựng website tư vấn, gợi ý khách sạn, nhà trọ ở Cần Thơ dựa trên nhận xét đã thu thập được sử dụng phương pháp lọc cộng tác.

4. Phương pháp nghiên cứu:

Thu thập dữ liệu đánh giá của khách hàng về các nhà trọ, khách sạn ở Cần Thơ.

Nghiên cứu về cách xây dựng website và các thuật toán mới trong lĩnh vực máy học (Machine Learning) để xây dựng trang web gợi ý khách sạn, nhà trọ ở Cần Thơ dựa trên nhận xét đã thu thập được. Sử dụng các dữ liệu được nhập vào để kiểm tra độ chính xác của thuật toán.

5. Đối tượng và phạm vi nghiên cứu:

❖ Đối tượng nghiên cứu:

- Áp dụng tiền xử lý văn bản trong xử lý ngôn ngữ tự nhiên.
- Cách thức phân tích từ vựng, ngữ nghĩa và cú pháp trong việc phân tích văn bản đánh giá của người dùng.
- Xây dựng hệ thống gợi ý dựa trên nội dung để dự đoán và gợi ý nhà trọ, khách sạn phù hợp với người dùng.

❖ Phạm vi nghiên cứu:

- Thu thập dữ liệu là các văn bản đánh giá, xếp hạng của người dùng về khách sạn, nhà trọ Cần Thơ.

6. Kết quả đạt được:

Qua nghiên cứu xây dựng được mô hình máy học có khả năng dự đoán và gợi ý khách sạn, nhà trọ cho người dùng.

7. Bố cục luận văn:

Phần giới thiệu

Giới thiệu tổng quát về đề tài.

Phần nội dung

Chương 1: Mô tả bài toán.

Chương 2: Thiết kế và cài đặt.

Chương 3: Hệ thống thử nghiệm.

Phần kết luận

Trình bày kết quả đạt được và hướng phát triển hệ thống.

PHẦN NỘI DUNG

CHƯƠNG 1

MÔ TẢ BÀI TOÁN

1.1. Mô tả chi tiết bài toán:

Việc tìm nơi ở phù hợp khi đến một thành phố mới là điều không dễ dàng trước lượng thông tin khổng lồ hiện nay. Hệ thống gợi ý nhà trọ, khách sạn nhằm giúp sinh viên, người tìm việc, du khách có thể tìm được nơi ở phù hợp tại Cần Thơ.

Một trong những điều đầu tiên cần làm khi lên kế hoạch cho một chuyến đi hay chuẩn bị cho một quá trình học tập xa nhà là tìm một nơi tốt để ở. Việc lựa chọn phòng khách sạn trực tuyến hay nhà trọ có thể là một nhiệm vụ quá sức với hàng ngàn lựa chọn, cho mọi điểm đến. Với tầm quan trọng của những tình huống này, nhiệm vụ giới thiệu khách sạn – nhà trọ cho người dùng được đề xuất.

Mục đích của hệ thống gợi ý nhà trọ - khách sạn là dự đoán và đề xuất nhóm các khách sạn phù hợp với tiêu chí người dùng trong hàng trăm khách sạn – nhà trọ thuộc khu vực tỉnh Cần Thơ.

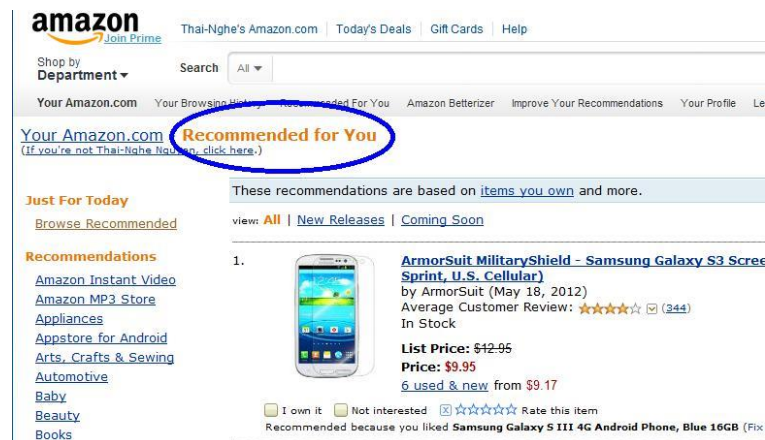
1.2. Giới thiệu về hệ thống gợi ý:

1.2.1. Giới thiệu:

Hệ thống gợi ý (Recommender systems - RS) là một dạng của hệ thống lọc thông tin (Information filtering), thường được sử dụng để dự đoán sở thích của người dùng dựa vào những phản hồi (feedbacks) của họ nhằm gợi ý các sản phẩm (item) mà người dùng có thể thích. RS hiện đang được ứng dụng ở rất nhiều lĩnh vực khác nhau như: trong thương mại điện tử (bán hàng trực tuyến), trong giải trí (âm nhạc, phim ảnh...), trong giáo dục đào tạo (gợi ý nguồn tài nguyên học tập như: sách, báo, ...).

Ví dụ, trong hệ thống bán hàng trực tiếp của Amazon, để tối ưu hóa khả năng mua sắm của khách hàng (user), người ta sẽ quan tâm đến việc những khách hàng nào đã “yêu thích” những sản phẩm (item) nào bằng cách dựa vào dữ liệu quá khứ của khách hàng (có thể là đánh giá số sao của khách hàng đã bình chọn trên sản phẩm, thời gian duyệt (browse) trên sản phẩm hay số click chuột trên sản phẩm, ...), từ

đó hệ thống sẽ dự đoán được khách hàng có thể thích sản phẩm nào và đưa ra gợi ý phù hợp với họ.



Hình 1.1 Hệ thống gợi ý sản phẩm Amazon

Ngoài lĩnh vực thương mại điện tử, RS hiện tại cũng được ứng dụng rất nhiều và thành công trong các lĩnh vực khác nhau như: giải trí (gợi ý bài hát cho người nghe – LastFM – www.last.fm), gợi ý phim ảnh (Netflix – www.netflix.com), gợi ý các video phù hợp (Youtube – www.youtube.com), giáo dục và đào tạo (gợi ý nguồn tài nguyên học tập phong phú như sách, báo, website, ... cho người học). Hệ thống gợi ý không chỉ đơn thuần là một dạng của Hệ thống thông tin mà nó còn là cả một lĩnh vực nghiên cứu đang được rất nhiều nhà khoa học quan tâm. Từ năm 2007 đến nay, hàng năm đều có hội thảo chuyên về hệ thống gợi ý như ACM (ACM RecSys) cũng như các tiểu bang dành riêng cho RS trong các hội nghị lớn khác như ACM KDD, ACM CIKM, ...

1.2.2. Hệ thống gợi ý (recommender systems):

1.2.2.1. Các khái niệm chính:

Trong hệ thống gợi ý, thông thường người ta quan tâm đến ba thông tin chính – người dùng (user), mục tin (item, item ở đây có thể là sản phẩm, bài hát, bộ phim hay bài báo... tùy thuộc vào mỗi hệ thống khác nhau) và phản hồi của người dùng (feedback) trên mục tin đó (các xếp hạng/đánh giá – rating biểu diễn mức độ thích/quan tâm của họ). Các thông tin này sẽ được biểu diễn thông qua ma trận mà trong đó mỗi dòng là một user, mỗi cột là một item và mỗi ô trong ma trận là một giá trị phản hồi biểu diễn ở “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 (nên chú ý rằng mỗi user chỉ xếp hạng cho một hoặc một vài item trong quá khứ, do đó sẽ có rất nhiều ô trống trong ma trận, được gọi là ma trận thưa – sparse matrix).

	Items					
	1	2	...	i	...	m
1	5	3		1	2	
2		2				4
:			5			
u	3	4	?	2	1	
:					4	
n			3	2		

Hình 1.2 Ma trận biểu diễn dữ liệu RS

Nhiệm vụ chính của hệ thống gợi ý là dựa vào các ô đã có giá trị trong ma trận để thông qua mô hình được xây dựng có thể dự đoán ra các ô còn trống, sau đó sắp xếp kết quả dự đoán (có thể là từ cao xuống thấp hay hữu ích đến không hữu ích, ... tùy thuộc vào hệ thống) và chọn ra Top-N items theo thứ tự, từ đó gợi ý chúng cho người dùng.

1.2.2.2. Thông tin phản hồi từ người dùng và hai dạng bài toán chính trong hệ thống gợi ý:

Trong hệ thống gợi ý, giá trị phản hồi (feedback) r_{ui} của mỗi người dùng trên mục tin sẽ được ghi nhận lại để làm cơ sở cho việc dự đoán các giá trị kế tiếp. Tùy thuộc vào từng hệ thống mà giá trị này sẽ có ý nghĩa khác nhau, chẳng hạn như nó có thể để đo độ “phù hợp” hay “mức độ thích” trong các hệ thống thương mại điện tử hay là “năng lực/kết quả thực hiện” của người dùng trong các hệ thống e-learning.

Giá trị r_{ui} có thể được xác định một cách tường minh (explicit feedbacks) như thông qua việc đánh giá/xếp hạng (ví dụ, rating từ ★ đến ★★★★★; hay like (1) và dislike (0),...) mà người dùng u đã bình chọn cho item i ; hoặc r_{ui} có thể được xác định một cách không tường minh (implicit feedbacks) thông qua số lần click chuột, thời gian mà u đã duyệt/xem i ,...

Có 2 dạng bài toán chính trong RS là *dự đoán xếp hạng (rating prediction)* của các hệ thống có phản hồi tường minh như đã trình bày ở trên và *dự đoán mục thông tin (item prediction/recommendation)* là việc xác định xác suất mà người dùng thích mục tin tương ứng.

1.3. Các kỹ thuật chính trong hệ thống gợi ý:

Hiện tại, trong hệ thống gợi ý có rất nhiều giải thuật được đề xuất, tuy nhiên có thể gom chúng vào trong các nhóm chính:

- Nhóm giải thuật lọc theo nội dung (content-based filtering)
- Nhóm giải thuật lọc cộng tác (collaborative filtering)
- Nhóm giải thuật lai ghép (hybrid filtering)
- Nhóm giải thuật không cá nhân hóa (non-personalization)

1.3.1. Lọc cộng tác:

Một cách tiếp cận để thiết kế các hệ thống recommender được sử dụng rộng rãi là lọc cộng tác. Các phương pháp lọc cộng tác dựa trên việc thu thập và phân tích một lượng lớn thông tin về hành vi, hoạt động hoặc sở thích của người dùng và dự đoán những gì người dùng sẽ thích dựa trên sự tương đồng của họ với người dùng khác. Một lợi thế quan trọng của phương pháp lọc cộng tác là nó không dựa vào nội dung phân tích máy và do đó nó có khả năng đề xuất chính xác các mục phức tạp như phim mà không yêu cầu “hiểu biết” về mục đó. Nhiều thuật toán đã được sử dụng để đo lường sự giống nhau của người dùng hoặc sự tương đồng về mặt hàng trong các hệ thống giới thiệu. Ví dụ, cách tiếp cận hàng xóm gần nhất (k-nearest neighbor) và Pearson Correlation được Allen triển khai lần đầu tiên.

Lọc cộng tác dựa trên giả định rằng những người đã đồng ý trong quá khứ sẽ đồng ý trong tương lai và rằng họ sẽ thích các loại mặt hàng tương tự như họ thích trong quá khứ.

Khi xây dựng mô hình từ hành vi của người dùng, sự phân biệt thường được thực hiện giữa các hình thức thu thập dữ liệu rõ ràng và tiềm ẩn.

Thu thập dữ liệu rõ ràng bao gồm:

- Yêu cầu người dùng xếp hạng một mục trên thang trượt.
- Yêu cầu người dùng tìm kiếm.
- Yêu cầu người dùng xếp hạng một bộ sưu tập các mục từ yêu thích đến ít yêu thích nhất.
- Trình bày hai mục cho một người dùng và yêu cầu khách hàng chọn một trong số chúng tốt hơn.
- Yêu cầu người dùng tạo danh sách các mục mà khách hàng thích.

Thu thập dữ liệu ngầm bao gồm:

- Quan sát các mục mà người dùng xem trong cửa hàng trực tuyến.
- Phân tích thời gian xem mục / người dùng.
- Lưu giữ một bản ghi các mục mà người dùng mua trực tuyến.
- Lấy danh sách các mục mà người dùng đã nghe hoặc xem trên máy tính của họ.
- Phân tích mạng xã hội của người dùng và khám phá những lượt thích và không thích tương tự.

Hệ thống gợi ý so sánh dữ liệu đã thu thập với dữ liệu tương tự và khác nhau được thu thập từ những người khác và tính toán danh sách các mục được đề xuất cho người dùng. Một số ví dụ thương mại và phi thương mại được liệt kê trong bài viết về các hệ thống lọc cộng tác.

Một trong những ví dụ nổi tiếng nhất về lọc cộng tác là lọc cộng tác theo từng mục (những người mua x cũng mua y), một thuật toán được phổ biến rộng rãi bởi hệ thống gợi ý của Amazon.com. Các ví dụ khác bao gồm:

- Last.fm đề xuất âm nhạc dựa trên so sánh thói quen nghe của những người dùng tương tự, trong khi Realgeek so sánh xếp hạng sách cho các đề xuất.
- Facebook, MySpace, LinkedIn và các mạng xã hội khác sử dụng tính năng lọc cộng tác để giới thiệu bạn bè, nhóm và các kết nối xã hội khác (bằng cách kiểm tra mạng kết nối giữa người dùng và bạn bè của họ). Twitter sử dụng nhiều tín hiệu và tính toán trong bộ nhớ để giới thiệu cho người dùng của họ rằng họ nên “theo dõi”.
- Các phương pháp lọc cộng tác thường gặp phải ba vấn đề: Cold Start, khả năng mở rộng và sự thưa thớt (sparsity).
- Cold Start: Các hệ thống này thường yêu cầu một lượng lớn dữ liệu hiện có của người dùng để đưa ra các đề xuất chính xác.
- Khả năng mở rộng: Trong nhiều môi trường mà các hệ thống này đưa ra các khuyến nghị, có hàng triệu người dùng và sản phẩm. Do đó, một lượng lớn công suất tính toán thường là cần thiết để tính toán các gợi ý.
- Sparsity: Số lượng các mặt hàng được bán trên các trang web thương mại điện tử lớn là cực kỳ lớn. Những người dùng tích cực nhất sẽ chỉ đánh giá một tập con nhỏ của cơ sở dữ liệu tổng thể. Do đó, ngay cả những mặt hàng phổ biến nhất cũng có rất ít xếp hạng.

Một loại thuật toán lọc cộng tác cụ thể sử dụng hệ số ma trận hóa (matrix factorization), kỹ thuật xấp xỉ ma trận cấp thấp (low-rank matrix approximation).

Các phương pháp lọc cộng tác được phân loại là bộ lọc cộng tác dựa trên bộ nhớ và dựa trên mô hình. Một ví dụ nổi tiếng về các phương pháp dựa trên bộ nhớ là thuật toán dựa trên người dùng và các phương pháp dựa trên mô hình là Kernel-Mapping Recommender.

1.3.2. Lọc dựa trên nội dung:

Một cách tiếp cận phổ biến khác khi thiết kế hệ thống recommender là lọc nội dung. Phương pháp lọc dựa trên nội dung dựa trên mô tả về mặt hàng và hồ sơ về các tùy chọn của người dùng.

Trong hệ thống gợi ý dựa trên nội dung, từ khóa được sử dụng để mô tả các mục và hồ sơ người dùng được xây dựng để chỉ ra loại mục mà người dùng này thích. Nói cách khác, các thuật toán này cố gắng đề xuất các mục tương tự với các mục mà người dùng đã thích trong quá khứ (hoặc đang kiểm tra trong hiện tại). Cụ thể, các mục đề cử khác nhau được so sánh với các mục

được đánh giá trước đây bởi người dùng và các mục phù hợp nhất được đề xuất. Cách tiếp cận này có nguồn gốc từ việc thu thập thông tin và nghiên cứu lọc thông tin.

Để tóm tắt các tính năng của các mục trong hệ thống, một thuật toán trình bày mục được áp dụng. Một thuật toán được sử dụng rộng rãi là biểu diễn tf – idf (còn được gọi là biểu diễn không gian vector).

Để tạo hồ sơ người dùng, hệ thống chủ yếu tập trung vào hai loại thông tin:

- Một mô hình ưu tiên của người dùng.
- Lịch sử tương tác của người dùng với hệ thống gợi ý.

Về cơ bản, các phương thức này sử dụng một hồ sơ mặt hàng (ví dụ, một tập hợp các thuộc tính và tính năng rời rạc) mô tả mục trong hệ thống. Hệ thống tạo hồ sơ dựa trên nội dung của người dùng dựa trên vector trọng số của các đối tượng địa lý. Trọng số biểu thị tầm quan trọng của từng tính năng đối với người dùng và có thể được tính từ các vector nội dung được xếp hạng riêng lẻ bằng nhiều kỹ thuật. Các phương pháp đơn giản sử dụng các giá trị trung bình của vector hạng mục trong khi các phương pháp phức tạp khác sử dụng các kỹ thuật máy học như Bayesian Classifiers, phân tích cụm, cây quyết định và mạng thần kinh nhân tạo (artificial neural networks) để ước tính xác suất người dùng sẽ thích mục đó.

Phản hồi trực tiếp từ người dùng, thường dưới dạng nút thích hoặc không thích, có thể được sử dụng để gán trọng số cao hơn hoặc thấp hơn về tầm quan trọng của các thuộc tính nhất định (sử dụng phân loại Rocchio hoặc các kỹ thuật tương tự khác).

Một vấn đề quan trọng với lọc dựa trên nội dung là liệu hệ thống có thể tìm hiểu các tùy chọn của người dùng từ hành động của người dùng liên quan đến một nguồn nội dung hay không và sử dụng chúng trên các loại nội dung khác. Khi hệ thống bị hạn chế đề xuất nội dung cùng loại với người dùng đang sử dụng, giá trị từ hệ thống đề xuất thấp hơn đáng kể so với các loại nội dung khác từ các dịch vụ khác có thể được đề xuất. Ví dụ: giới thiệu các bài viết tin tức dựa trên việc duyệt tin tức hữu ích nhưng sẽ hữu ích hơn nhiều khi bạn có thể đề xuất âm nhạc, video, sản phẩm, cuộc thảo luận, v.v. từ các dịch vụ khác nhau dựa trên duyệt tin tức.

Pandora Radio là một ví dụ về hệ thống giới thiệu dựa trên nội dung phát nhạc có các đặc điểm tương tự như một bài hát do người dùng cung cấp làm hạt giống ban đầu. Ngoài ra còn có một số lượng lớn các hệ thống gợi ý dựa trên nội dung nhằm cung cấp các đề xuất phim, một vài ví dụ như Rotten Tomatoes, Internet Movie Database, Jinni, Rovi Corporation và Jaman. Các hệ thống gợi ý giới thiệu tài liệu liên quan nhằm mục đích cung cấp các đề xuất tài liệu cho các nhà nghiên cứu. Các chuyên gia y tế công cộng đã nghiên

cứu các hệ thống gợi ý để cá nhân hóa giáo dục sức khỏe và các chiến lược phòng ngừa.

1.3.3. Hệ thống gợi ý lai (hybrid recommender systems):

Nghiên cứu gần đây đã chứng minh rằng một phương pháp lai, kết hợp lọc cộng tác và lọc dựa trên nội dung có thể hiệu quả hơn trong một số trường hợp. Các phương pháp lai có thể được thực hiện theo nhiều cách:

- Bằng cách đưa ra các dự đoán dựa trên nội dung và dựa trên lọc cộng tác riêng biệt và sau đó kết hợp chúng.
- Bằng cách thêm các khả năng dựa trên nội dung vào phương pháp cộng tác (và ngược lại).
- Bằng cách thống nhất các phương pháp tiếp cận thành một mô hình.

Một số nghiên cứu thực nghiệm so sánh hiệu suất của phương pháp lai với các phương pháp cộng tác thuần túy và chứng minh rằng các phương pháp lai có thể cung cấp các khuyến nghị chính xác hơn các phương pháp thuần túy. Những phương pháp này cũng có thể được sử dụng để khắc phục một số vấn đề thường gặp trong hệ thống gợi ý như Cold Start và vấn đề thừa thớt.

Netflix là một ví dụ tốt về việc sử dụng các hệ thống hybrid recommender. Trang web đưa ra các đề xuất bằng cách so sánh thói quen xem và tìm kiếm của những người dùng tương tự (ví dụ: lọc cộng tác) cũng như bằng cách cung cấp những bộ phim có chung đặc điểm với những bộ phim mà người dùng đánh giá cao (lọc dựa trên nội dung).

Một loạt các kỹ thuật đã được đề xuất làm cơ sở cho các hệ thống gợi ý: các kỹ thuật hợp tác (collaborative), dựa trên nội dung (content-based), dựa trên kiến thức (knowledge-based) và nhân khẩu học (demographic techniques). Mỗi kỹ thuật này đều có những thiếu sót, như vấn đề Cold Start cho các hệ thống cộng tác và dựa trên nội dung (phải làm gì với người dùng mới với ít xếp hạng) và tắc nghẽn kỹ thuật tri thức (knowledge engineering bottleneck) trong các phương pháp dựa trên tri thức. Một hệ thống gợi ý lai là một hệ thống trong đó kết hợp nhiều kỹ thuật với nhau để đạt được một số sức mạnh tổng hợp giữa chúng.

- Cộng tác – Collaborative: Hệ thống tạo đề xuất chỉ sử dụng thông tin về hồ sơ xếp hạng cho những người dùng hoặc mục khác nhau. Các hệ thống cộng tác định vị “người dùng/mục” ngang hàng với lịch sử xếp hạng tương tự như người dùng hoặc mục hiện tại và tạo đề xuất sử dụng vùng lân cận này. Các thuật toán dựa trên người dùng và dựa trên hàng gần nhất có thể được kết hợp để giải quyết vấn đề Cold Start và cải thiện kết quả đề xuất.
- Dựa trên nội dung – Content-based: Hệ thống tạo đề xuất từ hai nguồn: các tính năng liên quan đến sản phẩm và xếp hạng mà người dùng đã

cung cấp cho họ. Đề xuất dựa trên nội dung coi đề xuất là sự cố phân loại người dùng cụ thể và tìm hiểu trình phân loại cho lượt thích và không thích của người dùng dựa trên các tính năng của sản phẩm.

- Nhân khẩu học – demographic techniques: Trình giới thiệu nhân khẩu học cung cấp các đề xuất dựa trên hồ sơ nhân khẩu học của người dùng. Sản phẩm được đề xuất có thể được sản xuất cho các mục nhân khẩu học khác nhau, bằng cách kết hợp xếp hạng của người dùng trong các mục đó.
- Dựa trên tri thức – knowledge-based: Trình giới thiệu dựa trên kiến thức gợi ý các sản phẩm dựa trên các suy luận về nhu cầu và sở thích của người dùng. Kiến thức này đôi khi sẽ chứa kiến thức chức năng rõ ràng về cách các tính năng sản phẩm nhất định đáp ứng nhu cầu của người dùng.

Thuật ngữ Hybrid recommender systems được sử dụng ở đây để mô tả bất kỳ hệ thống recommender nào kết hợp nhiều kỹ thuật đề xuất với nhau để tạo dữ liệu đầu ra của nó.

Các kỹ thuật lai cơ bản (hybridization techniques):

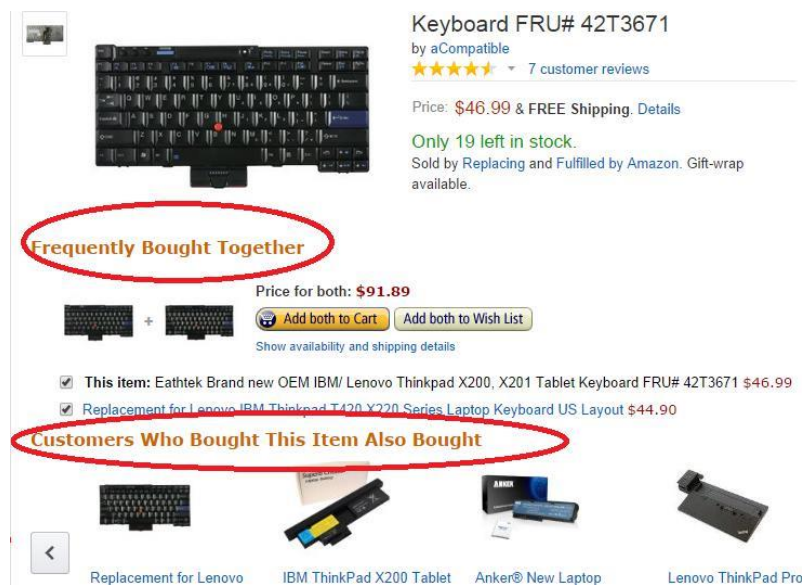
- Có trọng số (Weighted): Điểm số của các thành phần đề xuất khác nhau được kết hợp theo số lượng.
- Chuyển đổi (Switching): Hệ thống chọn giữa các thành phần đề xuất và áp dụng hệ thống được chọn.
- Hỗn hợp (Mixed): Các khuyến nghị từ những người giới thiệu khác nhau được trình bày cùng nhau để đưa ra đề xuất.
- Kết hợp tính năng (Feature Combination): Các tính năng được lấy từ các nguồn tri thức khác nhau được kết hợp với nhau và được đưa ra cho một thuật toán gợi ý duy nhất.
- Tính năng tăng cường (Feature Augmentation): Một kỹ thuật gợi ý được sử dụng để tính toán một tính năng hoặc tập hợp các tính năng, sau đó là một phần của đầu vào cho kỹ thuật tiếp theo.
- Cascade: Các khuyến nghị được ưu tiên nghiêm ngặt, với những ưu tiên thấp hơn phá vỡ các mối quan hệ trong việc tính điểm của những người cao hơn.
- Cấp độ meta (Meta-level): Một kỹ thuật đề xuất được áp dụng và tạo ra một số loại mô hình, sau đó là đầu vào được sử dụng bởi kỹ thuật tiếp theo.

1.3.4. Các kỹ thuật không cá nhân hóa:

Trong nhóm kỹ thuật này, do chúng khá đơn giản, dễ cài đặt nên nên thường được các website/hệ thống tích hợp vào, gồm cả các website thương mại, website tin tức, hay giải trí. Chẳng hạn như trong các hệ thống bán hàng trực

tuyến, người ta thường gợi ý các sản phẩm được xem/mua/bình luận/... nhiều nhất; gợi ý các sản phẩm mới nhất; gợi ý các sản phẩm cùng loại/ cùng nhà sản xuất ...; gợi ý các sản phẩm được mua/chọn cùng nhau. Một ví dụ khá điển hình là thông qua luật kết hợp (như Apriori), Amazon đã áp dụng khá thành công để tìm ra các sản phẩm hay được mua cùng nhau như minh họa trong Hình 4.

Tuy vậy, bất lợi của các phương pháp này là không cá nhân hóa cho từng người dùng, nghĩa là tất cả các user đều được gợi ý giống nhau khi chọn cùng sản phẩm.



Hình 1.3 Gợi ý sản phẩm thường được mua

1.4. Deep learning trong hệ thống gợi ý:

Deep Learning (DL) là một chủ đề nóng trong cộng đồng học máy. Sự phổ biến của việc áp dụng học sâu vào hệ thống khuyến nghị là tương đối chậm, vì chủ đề này chỉ trở nên phổ biến trong năm 2016, với hội thảo Deep Learning for recommender Systems tại ACM RecSys 2016.

Mạng nơ-ron hồi quy (RNN) có một số thuộc tính làm cho chúng trở nên phù hợp để mô hình hóa chuỗi các phiên truy cập của người dùng. Đặc biệt, chúng có khả năng kết hợp đầu vào từ các sự kiện xảy ra trong quá khứ, cho phép dự đoán tốt hơn ý định của người dùng.

1.5. Các phương pháp đánh giá:

Có hai nhóm tiêu chí đánh giá: các tiêu chí định lượng và tiêu chí định tính. Các tiêu chí định lượng được dành riêng cho việc đánh giá số lượng các gợi ý liên quan, chúng tương ứng với độ chính xác^[13]. Với sự phát triển không ngừng, bên cạnh các tiêu chí định lượng thì người ta nghiên cứu thêm các tiêu chí đánh giá mới (tiêu chí định tính) nhằm có những đánh giá chính xác hơn để cải thiện hệ

thống gợi ý. Các tiêu chí định tính được sử dụng để đánh giá chung về chất lượng của hệ thống gợi ý.

1.5.1. Tiêu chí định lượng:

1.5.1.1. Đánh giá độ chính xác của các dự đoán:

Việc đánh giá tính chính xác các dự đoán có thể sử dụng sai số bình phương trung bình (MSE – Mean Square Error), căn của sai số bình phương trung bình (RMSE – Root Mean Square Error), sai số tuyệt đối trung bình (MAE - Mean Absolute Error) (Herlocker J.L *et al*, 2004; Koren.Y, 2009; Trần Nguyễn Minh Thư, 2011).^[13] Tính chính xác của các dự đoán được đo trên n quan sát, trong đó p_i là giá trị dự đoán đánh giá của mục i và r_i là giá trị đánh giá thực tế của mục i .

$$MSE = \frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |p_i - r_i|$$

Các chỉ số này thích hợp cho một cơ sở dữ liệu không phải nhị phân và cho một giá trị dự đoán là số. Nó giúp đo lường mức độ sai số của các dự đoán. Các giá trị đo lường này bằng 0 khi hệ thống đạt được hiệu quả tốt nhất. Giá trị này càng cao thì hiệu quả của hệ thống càng thấp.

Tại cuộc thi nhằm cải thiện độ chính xác của hệ thống gợi ý do Netflix 4 tổ chức, các hệ thống gợi ý đã được đánh giá bởi chỉ số RMSE. Các chỉ số MAE, MSE và RMSE đã được sử dụng để đánh giá hệ thống gợi ý mà kết quả là giá trị dự đoán các đánh giá như hệ thống MovieLens, BookCrossing. Những chỉ số này rất dễ sử dụng để đánh giá, tuy nhiên MAE là biện pháp sử dụng nhiều nhất vì khả năng giải thích trực tiếp của nó.

1.5.1.2. Đánh giá việc sử dụng các dự đoán:

Ngoài việc đánh giá tính chính xác của các dự đoán, một số chỉ số khác như precision, recall và F_score, R_score được dùng để đánh giá việc sử dụng của các dự đoán trong trường hợp cơ sở dữ liệu nhị phân (Herlocker J.L *et al*, 2004; Sarwar, B and G. Karypis, 2000; Breese, J.S. and D. Heckerman, 1998).^[14] Các chỉ số này đánh giá các gợi ý phù hợp cho mỗi người dùng thay vì đánh giá số điểm liên quan đến từng đề nghị. Đề nghị được coi là phù hợp khi người dùng chọn mục dữ liệu từ danh sách những đề nghị đã được gợi ý cho người dùng.

Precision là tỷ lệ giữa số lượng các gợi ý phù hợp và tổng số các gợi ý đã cung

cấp (đã tạo ra). Precision bằng 100% có nghĩa là tất cả các kiến nghị đều phù hợp.

$$Precision = \frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng gợi ý tạo ra}}$$

Recall được định nghĩa bởi tỉ lệ giữa số lượng các gợi ý phù hợp và số lượng các mục dữ liệu mà người dùng đã chọn lựa (xem, nghe, mua, đọc). Recall được sử dụng để đo khả năng hệ thống tìm được những mục dữ liệu phù hợp so với những gì mà người dùng cần.

$$Recall = \frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng sản phẩm mua bởi người dùng}}$$

Precision và Recall được xem là hữu ích trong việc đánh giá một gợi ý. Tuy nhiên, trong một số trường hợp thì precision và recall có giá trị tỉ lệ nghịch với nhau. Ví dụ như số lượng gợi ý mà hệ thống tạo ra là 10, số lượng gợi ý phù hợp là 3, số lượng sản phẩm mua bởi người dùng là 3 thì độ chính xác thấp (30%), tuy nhiên giá trị recall lại cao (100%) nghĩa là độ chính xác thấp nhưng người dùng lại hài lòng bởi vì họ mua có 3 sản phẩm và hệ thống gợi ý đúng cả 3 sản phẩm đó. Trong tình huống đó, chỉ số F-score được sử dụng để đánh giá hiệu quả tổng thể của hệ thống bằng cách kết hợp hài hòa hai chỉ số Recall và Precision.

$$F - score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

R_{score} hay Breese score (Breese, J.S. and D. Heckerman, 1998)^[15] cũng là một trong những chỉ số đánh giá khả năng sử dụng dự đoán nhưng chỉ số này chính xác đến thứ tự của các gợi ý được xây dựng. R_{score} đánh giá vị trí của sản phẩm được chọn bởi người dùng trong danh sách sản phẩm gợi ý được tạo ra bởi hệ thống. Ví dụ, một hệ thống gợi ý cho người dùng 10 sản phẩm sắp xếp theo thứ tự ưu tiên từ cao đến thấp. Nếu người dùng chọn sản phẩm đầu tiên trong danh sách thì hệ thống gợi ý hiệu quả hơn khi người dùng chọn sản phẩm có thứ tự thứ 10. Chỉ số R_{score} được tính dựa vào tỉ lệ giữa thứ tự của mục gợi ý đúng (R_{score}_p) và thứ tự của mục gợi ý đúng tốt nhất (R_{score}_{max}) như công thức sau:

$$Rankscore = \frac{Rankscore_p}{Rankscore_{\max}}$$

$$Rankscore_p = \sum_{i \in h} 2^{\frac{Rank(i)-1}{\alpha}}$$

$$Rankscore_{\max} = \sum_{i=1}^{|T|} 2^{\frac{i-1}{\alpha}}$$

Trong đó:

- h là tập các sản phẩm gợi ý đúng.
 - Rank trả về thứ tự sắp xếp của một sản phẩm trong danh sách gợi ý
 - T là tập tất cả các sản phẩm người dùng quan tâm
- α là chu kỳ nửa phân kỳ (xác suất mà mục dữ liệu trong danh sách gợi ý được chọn là 50%).

Các chỉ số *Precision*, *Recall* và *F_score*, R_{score} thường được sử dụng đối với các hệ thống gợi ý trong lĩnh vực thương mại điện tử. Các chỉ số đánh giá, công thức tương ứng và một số hệ thống gợi ý/ nghiên cứu đã áp dụng các chỉ số tương ứng đó được tổng hợp trong Bảng 2.

STT	Chỉ số	Công thức	Hệ thống đã áp dụng
1.	MAE	$\frac{1}{n} \sum_{i=1}^n (p_i - r_i)$	MovieLens
2.	MSE	$\frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2$	Netflix
3.	RMSE	$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (p_i - r_i)^2}$	BookCrossing
4.	Precision	$\frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng gợi ý tạo ra}}$	EachMovie
5.	Recall	$\frac{\text{Số lượng gợi ý phù hợp}}{\text{Số lượng sản phẩm mua bởi người dùng}}$	Yeong, et al, 2005
6.	F_{score}	$F\text{-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$	MovieLens
7.	R_{score} hay Breese score	$Rankscore = \frac{Rankscore_p}{Rankscore_{\max}}$ $Rankscore_p = \sum_{i \in h} 2^{\frac{Rank(i)-1}{\alpha}}$ $Rankscore_{\max} = \sum_{i=1}^{ T } 2^{\frac{i-1}{\alpha}}$	TaFeng, B&Q (Breese, J.S. and D. Heckerman, 1998, Hsu, C. and H. Chung, 2004)

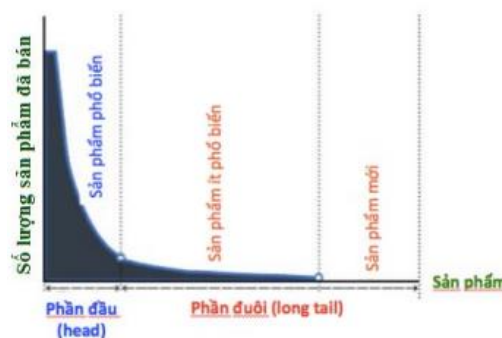
Bảng 2: Các phương pháp đánh giá

1.5.2. Tiêu chí định tính:

Trong những giai đoạn đầu phát triển thì hệ thống gợi ý chỉ sử dụng các độ đo chính xác định lượng như đã đề cập. Tuy nhiên, người dùng ngày càng có yêu cầu cao hơn và nhiều hơn về chất lượng của các gợi ý. Nếu chỉ xét độ chính xác thì không đủ để đánh giá hiệu quả của một hệ thống gợi ý nên cần đưa thêm thuộc tính chất lượng các gợi ý thay vì chỉ sử dụng độ chính xác của các gợi ý. Các chỉ số đánh giá chất lượng có thể là tính đa dạng, tính mới lạ, khả năng cầu may (một gợi ý không mong đợi bởi người dùng nhưng cuối cùng lại phù hợp cho người dùng), phạm vi bao phủ của gợi ý (độ bao phủ của dự báo hay gợi ý). Một số chỉ số định tính sẽ được phân tích chi tiết trong nội dung tiếp theo (Herlocker J.L et al, 2004; Slaney.M, 2006; Takács G., et al, 2007; Yu, C and L. Lakshmanan, 2009).^[16]

1.5.2.1. Tính mới của các gợi ý:

Việc đánh giá tính mới của gợi ý là hiển nhiên để đáp ứng nhu cầu của người sử dụng các sản phẩm mới được tạo ra liên tục. Khái niệm "sản phẩm mới" có thể có nhiều ý nghĩa khi đề cập đến hệ thống gợi ý (Herlocker J.L et al, 2004; Karypis.g, 2001). Tính mới của mục dữ liệu theo quan điểm thời gian (trong trường hợp xuất hiện một sản phẩm mới) hoặc liên quan đến lịch sử của người sử dụng (một sản phẩm mà chưa bao giờ được mua). Điều này xảy ra như một trường hợp đặc biệt mà mục dữ liệu trong hệ thống chưa có thông tin liên quan đến người sử dụng, như thể hiện “sản phẩm mới” trong Hình 3. Vấn đề này cũng được xác định như là một trong những khó khăn của hệ thống gợi ý – vấn đề “thiếu thông tin” (cold start problem).



Hình 1.4: Sự phổ biến của sản phẩm

Một số hệ thống cung cấp rất chính xác gợi ý nhưng không hữu dụng trong thực tế vì không quan tâm đến các tiêu chí định tính trong quá trình xây dựng hệ thống. Ví dụ như hệ thống gợi ý “sữa tươi” cho khách hàng trong một siêu thị ở châu Âu. Đề nghị này là chính xác bởi vì hầu như tất cả các khách hàng đều mua sữa, nhưng nó không phải là hữu ích cho tất cả người dùng đã quen thuộc với sản phẩm này. Các nhà cung cấp nhận thức được điều này trong một

thời gian dài và tổ chức các kệ đựng hàng hoá cho phù hợp. Do đó, việc giới thiệu cho người mua một thực phẩm mới có khả năng để làm hài lòng người mua mà họ không bao giờ nghĩ là cần thiết. Ví dụ thứ hai, liên quan đến sự cần thiết phải có thuộc tính mới trong các gợi ý. Giả sử một người dùng mua quà tặng trước Giáng sinh hai tuần. Trong trường hợp này, việc xây dựng danh sách các gợi ý căn cứ vào mặt hàng phổ biến của các năm trước thì không phù hợp bởi vì người dùng thường tìm cách tặng các sản phẩm mới cho người thân.

Thuộc tính mới được nhấn mạnh như là một chỉ số cần thiết để đánh giá tính hiệu quả của hệ thống gợi ý. G. Shani và ctv cùng với những nghiên cứu của mình đã chỉ ra 3 điểm quan trọng liên quan đến hiệu quả của hệ thống gợi ý. Thứ nhất, tính chính xác và tính mới lạ phải được tính đến để xây dựng được các gợi ý hiệu quả. Thứ hai, yếu tố thời gian là điều cần thiết trong việc đánh giá tính mới của mục dữ liệu. Thứ ba, danh sách gợi ý phù hợp nhất phải kết hợp một tỉ lệ các mục dữ liệu mới và các gợi ý phù hợp khác.

1.5.2.2. Tính đa dạng (Diversity) của các gợi ý:

Sự đa dạng của hệ thống gợi ý đo lường khả năng cung cấp một danh sách các mục dữ liệu được phân phối từ nhiều loại khác nhau. Có thể phân chia sự đa dạng của các gợi ý thành hai loại đa dạng: sự đa dạng cá nhân và đa dạng tổng thể. Loại đa dạng cá nhân quan tâm đến các khái niệm về đa dạng từ quan điểm của người sử dụng. Chỉ số này được tính toán dựa trên trung bình sự khác nhau giữa tất cả các cặp mục dữ liệu đã gợi ý. Ngược lại, sự đa dạng tổng thể là quan tâm đến các mục dữ liệu đã gợi ý hơn là quan tâm đến người dùng. Nếu sự đa dạng tổng thể của hệ thống giới thiệu là lớn, thì sự đa dạng của các gợi ý cá nhân cũng là rất lớn, nhưng điều này không đúng cho chiều ngược lại. Ví dụ, hệ thống cung cấp 3 gợi ý khác nhau cho tất cả người dùng, thì sự đa dạng cá nhân là tương đối cao nhưng sự đa dạng tổng thể là rất thấp (Adomavicius, G. And Y. Kwon, 2008; Adomavicius, G. and Y. Kwon, 2010; Bradley, 2001; Takács G., et al, 2007; Ziegler C., et al, 2005).^[17]

Trong các hệ thống gợi ý truyền thống, sự đa dạng của các gợi ý chưa được quan tâm đến mặc dù chỉ số này rất quan trọng. Trong một số trường hợp, sự đa dạng sẽ trở thành một điều cần thiết. Ví dụ như sự đa dạng của các điểm tham quan cho các kỳ nghỉ lễ trong hệ thống gợi ý các địa điểm du lịch. Với thực tế đó, đã có nhiều nghiên cứu cải thiện hiệu quả của hệ thống gợi ý hướng đến sự đa dạng và các nghiên cứu này cũng đã khẳng định “Nếu chỉ tính đến độ chính xác của các gợi ý để đánh giá chất lượng của một hệ thống là không đủ để đảm bảo sự phù hợp, hiệu quả của những gợi ý cho người dùng” G. Adomavicius.^[18]

1.5.2.3. Độ bao phủ (coverage) của các gợi ý:

Độ bao phủ của hệ thống gợi ý là thước đo số lượng lĩnh vực mà danh sách các sản phẩm gợi ý được tạo ra thuộc về chúng, số lĩnh vực này có bao trùm được hệ thống hay không (Herlocker J.L et al, 2004, Takács G., et al, 2007).^[16] Độ bao phủ của các gợi ý thấp thì thường ít được đánh giá cao bởi người dùng bị giới hạn thông tin về các lĩnh vực của hệ thống và họ cần được tư vấn đa lĩnh vực. Độ bao phủ đã được sử dụng trong đánh giá hệ thống gợi ý bởi một số nhà nghiên cứu như Good et al. 1999, Herlocker et al. 1999, Sarwar et al. 1998.

Hầu hết độ bao phủ được đo bằng số các mặt hàng mà dự đoán có thể được hình thành như là một tỷ lệ phần trăm của tổng số các mặt hàng. Cách dễ nhất để đo loại này là chọn một cách ngẫu nhiên cặp user/item, yêu cầu một dự đoán cho mỗi cặp, và đo tỷ lệ phần trăm mà dự đoán được cung cấp. Giống như chỉ số precision và recall phải được xem xét đồng thời, độ bao phủ (Coverage) thường được kết hợp với chỉ số “accuracy”, vì không thể tăng giá độ bao phủ mà không quan tâm đến việc tạo ra những gợi ý không thuộc hệ thống. Một cách khác để tính độ bao phủ là chỉ xem xét độ bao phủ trên những mặt hàng mà người dùng quan tâm. Độ bao phủ tính theo cách này không được đo trên toàn bộ các sản phẩm mà chỉ quan tâm đến những sản phẩm mà khách hàng đã biết hay đã từng xem qua. Ưu điểm của cách tính này là nó đáp ứng tốt nhu cầu của người dùng.

Độ bao phủ được đo bằng sự phong phú của hồ sơ người dùng để đưa ra gợi ý. Ví dụ trong lọc cộng tác, người dùng phải đánh giá các item trước khi nhận các gợi ý. Việc đo lường này là một loại hình đặc trưng trong nghi thức đánh giá offline.

1.5.2.4. Sự hài lòng của người sử dụng:

Sự hài lòng của người sử dụng là một khía cạnh hơi mơ hồ và phụ thuộc vào từng cá nhân khác nhau và do đó rất khó để đo lường. Theo định nghĩa của (Herlocker J.L et al, 2004) thì sự hài lòng của người dùng được định nghĩa là mức độ mà một người dùng được hỗ trợ trong việc đối phó với các vấn đề quá tải thông tin. Herlocker et al đã phân loại một số phương pháp đánh giá sự hài lòng của người sử dụng.

Phương pháp đánh giá “rõ ràng” (Explicit) và “ngầm hiểu” (Implicit): phương pháp đánh giá một cách rõ ràng nghĩa là hệ thống đo độ hài lòng của người sử dụng bằng cách yêu cầu trực tiếp; phương pháp đánh giá ngầm hiểu thì cần phải đặt ra những giả định và dịch những quan sát được thành những giả định, ví dụ như sự gia tăng doanh số của một cửa hàng chứng tỏ sự hài lòng của khách hàng tăng lên.

➔ Kết quả so với quá trình: việc đánh giá có thể chỉ tập trung vào kết quả, nhưng nó cũng có thể tập trung vào quá trình áp dụng hệ thống gợi ý.

Đánh giá trong khoảng thời gian ngắn (Short term) và khoảng thời gian dài (Long term): đánh giá người sử dụng trong một khoảng thời gian ngắn có thể sẽ thiếu sót thông tin mà nó sẽ trở nên chính xác hơn sau một khoảng thời gian nhất định. Sở thích của người dùng cần phải xem xét đánh giá của người dùng qua một khoảng thời gian dài. Các nghiên cứu điều tra sự hài lòng của người dùng đối với hệ thống gợi ý là rất hiếm và nghiên cứu tập trung trên sự hài lòng của các gợi ý thì càng hiếm hơn. Tiêu chí đánh giá này được sử dụng trong nghiên cứu của Cosley (Cosley D., et al., 2003) và Herlocker (Herlocker J. L., et al., 2000).^[19]

CHƯƠNG 2

THIẾT KẾ VÀ CÀI ĐẶT

2.1. Tập dữ liệu:

Tập dữ liệu là danh sách các nhà trọ, khách sạn thuộc địa phận trên địa bàn tỉnh Cần Thơ gồm 407 nhà trọ và 461 khách sạn được lấy trực tiếp từ các Website trên internet bằng phương pháp Data Crawling được đặt tên lần lượt là `Datasets_motels.csv`, `Datasets_reviews_motels.csv`, `Datasets_hotels.csv`, `Datasets_reviews_hotels.csv`.

`Datasets_motels.csv` là tập dữ liệu chứa danh sách gồm 408 nhà trọ Cần Thơ được thu thập từ các Website ^{[20], [21], [22], [23]} gồm các thông tin sau:

- `id_motels`: id tự tạo cho mỗi một nhà trọ, dùng để liên kết với `datasets_reviews_motels.csv` để biết các nhận xét đánh giá là của nhà trọ nào.
- `tennhatro`: tên các nhà trọ ở Thành phố Cần Thơ
- `diachi`: hiển thị địa chỉ chi tiết của từng nhà trọ
- `toalac`: hiển thị nhà trọ nằm ở quận hay đường nào của thành phố Cần Thơ
- `danhgia`: đánh giá tổng hợp của người dùng về nhà trọ
- `gia`: thể hiện giá của các nhà trọ theo đơn vị việt nam đồng (VND)
- `dientich`: thể hiện diện tích của từng nhà trọ
- `noidung`: nội dung chi tiết của các nhà trọ mà người dùng quan tâm như giá điện, giá nước hàng tháng, nhà trọ có sẵn wifi hay không, gần hay xa các địa điểm như siêu thị hay trường học, ...
- `luotxem`: thể hiện số lượng lượt người dùng ghé vào xem trang nhà trọ, hay nhà trọ đó được quan tâm nhiều hơn so với các nhà trọ khác.
 - ➔ Mỗi hàng thể hiện một nhà trọ có địa chỉ và được tọa lạc ở đâu, có giá cho thuê với diện tích và đánh giá bao nhiêu sao, bao nhiêu lượt xem cùng nội dung chủ trọ đăng trên diễn đàn

`Datasets_reviews_motels.csv` là tập dữ liệu chứa các nhận xét của người dùng về từng nhà trọ. Các nhận xét được thể hiện dưới dạng đánh giá sao (dạng số - integer) và dạng văn bản (dạng text – string):

- `Id_motels`: id của các nhà trọ, dùng để liên kết với `datasets_motels.csv`
- `Rating`: đánh giá sao của người dùng ở từng nhận xét của họ
- `Views`: văn bản đánh giá của người dùng cho biết nhà trọ đó tốt hay xấu, ...
- ➔ Mỗi dòng cho biết đánh giá của mỗi người dùng với nội dung đánh giá như thế nào và thuộc vào khách sạn nào được biết dựa vào `id_motel`

Datasets_hotels.csv là tập dữ liệu chứa danh sách gồm 461 khách sạn trực thuộc thành phố Cần Thơ được thu thập từ các Website [24], [25], [26], [27] gồm các thông tin như sau:

- id_hotels: id của từng khách sạn được tạo ngẫu nhiên
 - name: tên các khách sạn thuộc thành phố Cần Thơ
 - rating: đánh giá tổng hợp của khách sạn
 - number_of_ratings: số lượt nhận xét của người dùng đối với khách sạn
 - rating_title: đánh giá tổng hợp của người dùng được tổng lại xem nhà trọ đó như thế nào như Good, Very good hay Fabulous, ...
 - price: giá của khách sạn được thể hiện dưới đơn vị USD
 - price_free: giá khuyến mãi của khách sạn được thể hiện dưới đơn vị USD
 - room_type: loại phòng của khách sạn
 - address: địa chỉ cụ thể của từng khách sạn
 - location: thể hiện khách sạn thuộc quận hay nằm trên đường nào của thành phố Cần Thơ
 - Distance: khoảng cách của khách sạn so với trung tâm thành phố Cần Thơ
 - Content: Nội dung giới thiệu của khách sạn
 - price_for: số ngày ở lại khách sạn của khách hàng
 - People numbers: thể hiện số người ở hay số giường trong một phòng khách sạn, ...
- ➔ Mỗi dòng cho biết tên khách sạn cùng địa chỉ và nơi tọa lạc của khách sạn đó, tiếp sau đó là giá và giá khuyến mãi cùng với điểm đánh giá và phân loại xem khách sạn đó thuộc lớp đánh giá nào, cuối cùng là cho biết thêm số phòng, loại phòng, khoảng cách so với trung tâm thành phố cùng nội dung giới thiệu về khách sạn đó.

Datasets_reviews_hotels.csv là tập dữ liệu chứa các nhận xét của người dùng về từng khách sạn. Các nhận xét được thể hiện dưới dạng đánh giá sao (dạng số - integer) và dạng văn bản (dạng text – string):

- id_hotels: id của các khách sạn, dùng để liên kết với datasets_hotels.csv
 - user: tên người dùng đã nhận xét bình luận
 - title: tiêu đề của bình luận
 - rating: đánh giá của người dùng ở từng bình luận của họ
 - reviews: văn bản đánh giá của người dùng về khách sạn.
- ➔ Mỗi dòng thể hiện khách sạn được ai đánh giá với tựa đề là gì và đánh giá bao nhiêu điểm kèm theo đó là văn bản đánh giá của người dùng.

2.1.1. Tiền xử lý dữ liệu:

Hai tập dữ liệu chi tiết nhà trọ (Datasets_motels.csv) và khách sạn (Datasets_hotels.csv) khi được lấy về bằng phương pháp Data Crawling sẽ có dạng như sau:

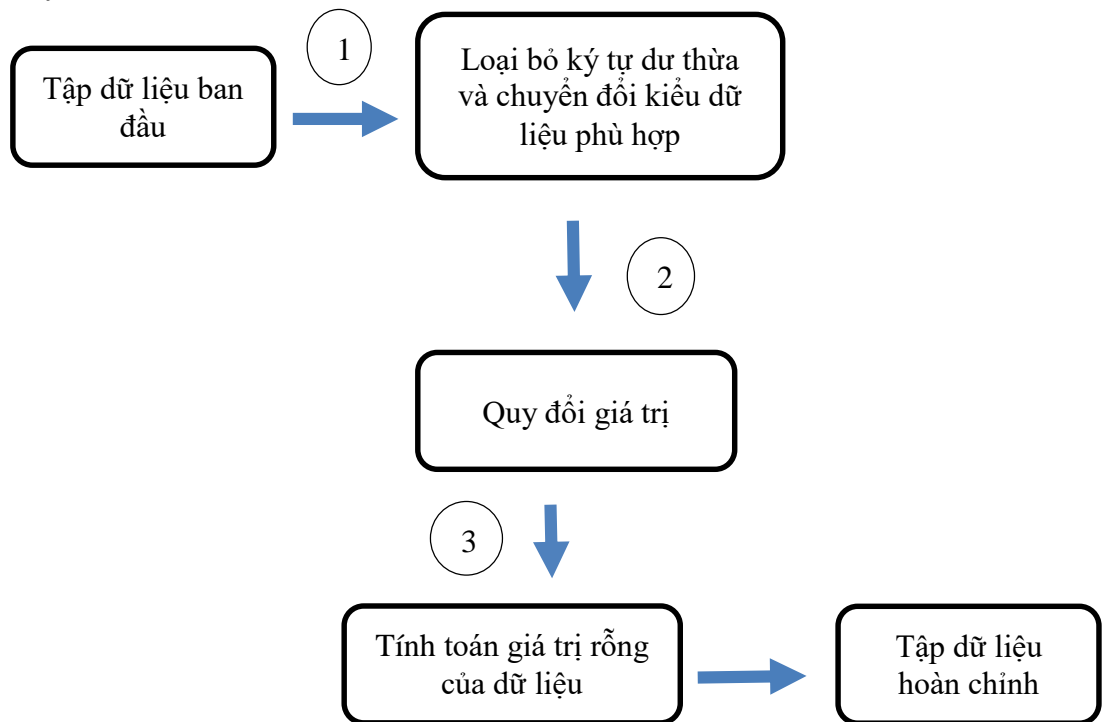
	tenhatro	gia	diachi	quan	dientich	noidun g	Luot xem	danh gia
0	Phòng trọ cho thuê gần cầu đầu sáu	Giá: 1.000.000 vnđ	Địa chỉ: 102/ 60B đường 3 tháng 2 cần thơ	Quận Ninh Kiều, Cần Thơ\n	DT: 26.00m ²	phòng trọ mới, có gác lửng. - hẻm 102 gần cầu...	Lượt xem: 21056	3
1	Nhà trọ mới xây cho thuê 5 phòng	Giá: 2.000.000 vnđ	Địa chỉ: 18/23A Xô Viết Nghệ Tĩnh	Quận Ninh Kiều, Cần Thơ\n	DT: 25.00 m ²	Nhà trọ Trần Quang Diệu (Q. Bình Thủy) thoáng.. .	Lượt xem: 102985	5

Bảng 1: Tập dữ liệu nhà trọ trước khi tiền xử lý – Datasets_motels.csv

	name	location	price	price_for	room_type	beds	rating	rating_title	number_of_ratings	url
0	Lan Vy Hotel	97D Pham Ngoc Thach, Cần Thơ, Việt Nam	US\$ 10	1 night, 1 adult	Deluxe Queen Room with Two Queen Beds	2 large double beds	8.8	Fabulous	95 reviews	https://www.booking.com/hotel/vn/d-home-stay.en...
1	N&D Homestay	88/24 Lê Lai, An Phú, Ninh Kiều, Thành phố Cần Thơ, Việt Nam, 900000	US\$ 15	1 night, 1 adult	Double Room	1 large double bed	7.8	Good	14 reviews	https://www.booking.com/hotel/vn/n-and-d-home-stay...

Bảng 2: Tập dữ liệu khách sạn trước tiền xử lý – Datasets_hotels.csv

Quá trình tiền xử lý dữ liệu được thực hiện gồm 3 bước thông qua sơ đồ dưới đây:



Hình 2.1: Sơ đồ tiền xử lý dữ liệu

Bước 1: Loại bỏ ký tự dư thừa và chuyển đổi kiểu dữ liệu phù hợp:

Đầu tiên ở các cột *gia*, *diachi*, *dientich*, *luotxem*, *quan* của bảng nhà trọ và *price*, *number_of_rating* của bảng khách sạn sẽ loại bỏ các từ không cần thiết với phương thức `replace`

```
df['Tên cột'] = df['Tên cột'].str.replace('chuỗi cũ', 'chuỗi thay thế')
```

Ở đây sẽ thay thế bằng chuỗi rỗng `''`.

Sau khi đã xóa bỏ các ký tự không cần thiết, tiến hành chuyển đổi kiểu dữ liệu của một số cột cần thiết để thuận tiện cho các bước tiếp theo:

Tên cột	Kiểu dữ liệu
Gia	Interger
Luotxem	Interger
Price	Interger
Number_of_rating	Interger

Bảng 3: Bảng thay đổi kiểu dữ liệu

Bước 2: Quy đổi giá trị:

Ở bảng nhà trọ chuyển đổi cột “đanhgia”, “luotxem” và “gia” từ các giá trị cụ thể thành các giá trị thuộc trong một khoảng nhất định. Ví dụ ở cột giá sẽ được quy đổi giá trị như sau:

Giá trị	Quy đổi
1 điểm đến 3 điểm	Kém
3.1 điểm đến 5 điểm	Trung bình
5.1 điểm đến 8 điểm	Khá
8.1 điểm đến 10 điểm	Rất tốt

Bảng 4: Bảng quy đổi giá trị ở cột “rating_title”

Thực hiện tương tự với các cột còn lại cần thực hiện

Bước 3: Tính toán giá trị rỗng của dữ liệu:

Ở bảng khách sạn và nhà trọ, do được lấy từ nhiều nguồn khác nhau nên dữ liệu sẽ bị khuyết (rỗng) ở cột đánh giá người dùng – điển hình là cột rating_title của tập dữ liệu Datasets_hotels.csv:

	name	location	price	price_for	room_type	beds	rating	rating_title	number_of_ratings	url
0	Lan Vy Hotel	97D Pham Ngoc Thach, Cần Thơ, Việt Nam	US\$ 10	1 night, 1 adult	Deluxe Queen Room with Two Queen Beds	2 large double beds	8.8	Fabulous	95 reviews	https://www.booking.com/hotel/vn/d-home-stay.en...
1	VAN A'S BOU TIQUE	14 Nguyễn Thị Minh Khai, Cần Thơ	12	1 night, 1 adult	Double Room	1 large double bed	?	?	5 reviews	https://www.booking.com/hotel/vn/vana-39-s...

Bảng 5: Dữ liệu bị khuyết (rỗng) của tập dữ liệu Datasets_hotels.csv

Do dữ liệu bị khuyết nên sẽ cần xử lý từ bảng reviews của khách sạn (Datasets_reviews_hotels.csv) với đánh giá người dùng, sau đó ánh xạ qua bảng khách sạn với TextBlob và công thức (1):

TextBlob là một thư viện Python để xử lý dữ liệu dạng văn bản. Nó cung cấp một API đơn giản để đi sâu vào các tác vụ xử lý ngôn ngữ tự nhiên (NLP) phổ biến như gắn thẻ từng phần của giọng nói, trích xuất cụm danh từ, phân tích tình cảm, phân loại, dịch thuật và hơn thế nữa. Ở phần này dùng TextBlob để phân tích cảm xúc (Sentiment Analysis) văn bản đánh giá người dùng, sau đó xếp các đánh giá được phân lớp vào các điểm đánh giá nhằm phục vụ cho thao tác tiếp theo.

Sau khi thực hiện xong phân loại đánh giá và có các điểm đánh giá, thực hiện công thức (1) sau để tính toán và ánh xạ chúng qua cột rating_title:

$$WeightedRating(\mathbf{WR}) = \left(\frac{v}{v + m} \cdot \mathbf{R} \right) + \left(\frac{m}{v + m} \cdot \mathbf{C} \right)$$

(1)

Trong đó:

- v là số lượt reviews của khách sạn
- m là số lượt reviews tối thiểu cần thiết để được liệt kê
- R là điểm đánh giá trung bình của khách sạn
- C là đánh giá trung bình trên toàn bộ khách sạn.

Ở đây, lượt reviews – v có sẵn và m là một siêu tham số có thể chọn cho phù hợp vì không có giá trị phù hợp cho m . Có thể coi đây là một bộ lọc tiêu cực sơ bộ, đơn giản là sẽ loại bỏ những phim có số phiếu bầu thấp hơn một ngưỡng nhất định m .

Điểm đánh giá trung bình R và đánh giá trung bình C sẽ được tính ở bảng reviews khách sạn bằng hàm mean () của pandas

$$C = \text{metadata['rating'].mean()}$$

Ví dụ ở khách sạn VANA'S BOUTIQUE là một trong những khách sạn không có đánh giá tổng quát. Sau khi xử lý ở tập dữ liệu datasets_reviews_hotel.csv sẽ có được kết quả như sau:

Id_hotel	Title_rating	rating	Comment_title	user
35	Phòng lớn và một vị trí rất trung tâm	4	Chuyến đi tuyệt vời. Khách sạn phục vụ tốt nhân viên thân thiện và nhiệt tình Cảm ơn bạn Thịnh và bạn Phát bộ phận tiền sảnh đã hỗ trợ và tư vấn các địa điểm ăn uống nhiệt tình. Có dịp sẽ quay lại. View bao đẹp.	KR-ORD
35	Magnifique, Magnifique.	5	Nhân viên khách sạn rất nhiệt tình, siêu dễ thương vừa đến là được mọi người xách đồ cho ngay. Phòng sạch sẽ, đồ ăn ổn hơn các khách sạn khác. Tuy khách sạn đông nhưng các dịch vụ được đáp ứng rất nhanh. Mọi thứ rất ok	mythalbie
35	Địa điểm, địa điểm	5	Chúng tôi được các bạn Huy, Khoa, Thiện hỗ trợ rất nhiệt tình trong quá trình lưu trú tại khách sạn, thức ăn ngon, dịch vụ tốt, giá cả hợp lý, nhân viên vui vẻ dễ thương...5 sao. Xin cảm ơn các bạn. Một trải nghiệm thật thú vị ở vinpearl cần thơ khiến chúng tôi nhớ mãi không quên!	TravelProNewYork
35	Khách sạn này là một lựa chọn tuyệt vời, nó là ...	4	Gia đình tôi đi du lịch tại Cần Thơ đã có những trải nghiệm tuyệt vời tại Vinpearl Cần Thơ. Đặc biệt ấn tượng và cảm ơn các bạn nhân viên hỗ trợ hành lý Khoa, Sang, Huy. Các bạn rất thân thiện, nhiệt tình giúp đỡ chúng tôi vận chuyển và đóng gói hành lý.	Emkaloma

35	khuyến khích	5	Tôi check in ngày 23.7 và out ngày 25.7. Về dịch vụ thì khá hài lòng, chợ nổi náo nhiệt, nhiều khu bán và ăn uống trên sông. Trái cây đa dạng, vườn trái cây rất phong phú. Gia đình tôi rất hài lòng với chuyến du lịch này. Nhân viên Kim yền hỗ trợ tôi về tour du lịch. Cảm ơn khách sạn	LarW3
----	--------------	---	--	-------

Bảng 6: Dữ liệu đánh giá của khách hàng

Khi đó ta có các giá trị sau:

Số lượt reviews của khách sạn	$v = 5$
Số lượt reviews tối thiểu cần thiết để được liệt kê	$m = 1$
Điểm đánh giá trung bình của khách sạn	$R = (4+5+5+4+5)/5 = 4.6$

Đánh giá trung bình trên toàn bộ khách sạn sử dụng toàn bộ dữ liệu đánh giá của Datasets_reviews_hotels.csv: $C = 32.2$

Áp dụng công thức (1) ta được:

$$WeightedRating(\mathbf{WR}) = \left(\frac{v}{v+m} \cdot R \right) + \left(\frac{m}{v+m} \cdot C \right) = 9.2$$

Suy ra rating (VANA'S BOUTIQUE) = **9.2**

⇒ Dựa vào bảng 4 ở bước 2:

Giá trị	Quy đổi
1 điểm đến 3 điểm	Kém
3.1 điểm đến 5 điểm	Trung bình
5.1 điểm đến 8 điểm	Khá
8.1 điểm đến 10 điểm	Rất tốt

⇒ Ta có title_rating = **Tốt**

Lần lượt sử dụng với các dữ liệu rỗng còn lại.

Sau khi hoàn thành bước tiền xử lý dữ liệu, hai bảng nhà trọ và khách sạn có các thuộc tính sau đây:

	tennhatro	gia	diachi	quan	dientich	noidung	Luot xem	danh gia	Gia1	luot xem	Danh gia1
0	Phòng trọ cho thuê gần cầu đầu sáu	1000000	102/60B đườn g 3 tháng 2 cần thơ	Quận Ninh Kiều, Cần Thơ	26	phòng trọ mới, có gác lửng. - hẻm 102 gần cần...	21056	4	1 triệu – 2 triệu	Khá	Rất tốt
1	Nhà trọ mới xây cho thuê 5 phòng	2000000	18/23 A Xô Viết Nghệ Tĩnh	Quận Ninh Kiều, Cần Thơ	25	Nhà trọ Trần Quang Diệu (Q. Bình Thủy) thoáng...	102985	5	1 triệu – 2 triệu	Khá	Rất tốt

Bảng 7: Tập dữ liệu nhà trọ sau tiền xử lý dữ liệu

	name	location	price	price_for	room_type	beds	rating	rating_title	number_of_ratings	url	Price2
0	Lan Vy Hotel	97D Pham Ngoc Thach, Cần Thơ, Việt Nam	10	1 night, 1 adult	Deluxe Queen Room with Two Queen Beds	2 large double beds	8.8	Fabulous	95	https://www.booking.com/hotel/vn/d-home-stay.en...	Giá thấp
1	N&D Homestay	88/24 Lê Lai, An Phú, Ninh Kiều, Thành phố Cần Thơ	15	1 night, 1 adult	Double Room	1 large double bed	7.8	Good	14	https://www.booking.com/hotel/vn/n-and-home-s...	Giá trung bình

Bảng 8: Tập dữ liệu khách sạn sau tiền xử lý dữ liệu

2.2. Hệ thống gợi ý nhà trọ, khách sạn sử dụng đề xuất dựa trên một số tiêu chí người dùng cần:

2.2.1. Giải pháp:

Hệ thống tạo đề xuất từ hai nguồn: các tính năng liên quan đến sản phẩm và xếp hạng mà người dùng đã cung cấp cho họ. Đề xuất dựa trên phân loại người dùng cụ thể và tìm hiểu trình phân loại cho lượt thích và không thích của người dùng dựa trên các tính năng của sản phẩm.

Ý tưởng của thuật toán này là, từ thông tin mô tả của item, biểu diễn item dưới dạng vector thuộc tính. Sau đó dùng các vector này để học mô hình của mỗi user, là ma trận trọng số của user với mỗi item.

2.2.2. Đề xuất dựa trên tiêu chí người dùng:

2.2.2.1. Item profiles:

Trong các hệ thống content-based, cần xây dựng một bộ hồ sơ (profile) cho mỗi item. Profile này được biểu diễn dưới dạng toán học là một "feature vector" n chiều. Trong những trường hợp đơn giản (ví dụ như item là dữ liệu dạng văn bản), feature vector được trực tiếp trích xuất từ item. Từ đó có thể xác định các item có nội dung tương tự bằng cách tính độ tương đồng giữa các feature vector của chúng.

Một số phương pháp thường được sử dụng để xây dựng feature vector là:

- Sử dụng TF-IDF
- Sử dụng biểu diễn nhị phân

2.2.2.2. Xây dựng ma trận TF_IDF:

Giả sử rằng số *users* là N, số *items* là M, *utility matrix* được mô tả bởi ma trận Y. Thành phần ở hàng thứ m, cột thứ n của Y là *mức độ quan tâm* (ở đây là số sao đã *rate*) của *user* thứ n lên sản phẩm thứ m mà hệ thống đã thu thập được. Ma trận Y bị khuyết rất nhiều thành phần tương ứng với các giá trị mà hệ thống cần dự đoán. Thêm nữa, gọi R là ma trận *rated or not* thể hiện việc một *user* đã *rated* một *item* hay chưa. Cụ thể, $r(i,j)$ bằng 1 nếu *item* thứ i đã được *rated* bởi *user* thứ j, bằng 0 trong trường hợp ngược lại.

Mô hình tuyến tính:

Giả sử rằng có thể tìm được một mô hình cho mỗi *user*, minh họa bởi vector cột hệ số $w(i)$ và bias $b(n)$ sao cho *mức độ quan tâm* của một *user* tới một *item* có thể tính được bằng một hàm tuyến tính:

$$y_{mn} = \mathbf{x}_m \mathbf{w}_n + b_n$$

(Chú ý rằng $\mathbf{x}(m)$ là một vector hàng, $\mathbf{w}(n)$ là một vector cột.)

Xét một *user* thứ n bất kỳ, nếu coi training set là tập hợp các thành phần đã được *điền* của $y(n)$, có thể xây dựng hàm mất mát tương tự như Ridge Regression như sau:

$$\mathcal{L}_n = \frac{1}{2} \sum_{m: r_{mn}=1} (\mathbf{x}_m \mathbf{w}_n + b_n - y_{mn})^2 + \frac{\lambda}{2} \|\mathbf{w}_n\|_2^2$$

Trong đó $s(n)$ là số lượng các *items* mà *user* thứ n đã *rated*, là tổng các phần tử thứ n của ma trận *rated or not* R :

$$s_n = \sum_{m=1}^M r_{mn},$$

Vì biểu thức loss function chỉ phụ thuộc vào các *items* đã được *rated*, có thể rút gọn nó bằng cách đặt $\mathbf{y}(n)$ là sub vector của \mathbf{y} được xây dựng bằng cách trích các thành phần khác nhau ở cột thứ n , tức đã được *rated* bởi *user* thứ n trong Utility Matrix \mathbf{Y} . Đồng thời, đặt $\mathbf{X}(n)$ là sub matrix của ma trận *feature* \mathbf{X} , được tạo bằng cách trích các hàng tương ứng với các *items* đã được *rated* bởi *user* thứ n . Khi đó, biểu thức hàm mất mát của mô hình cho *user* thứ n được viết gọn thành:

$$\mathcal{L}_n = \frac{1}{2s_n} \|\hat{\mathbf{X}}_n \mathbf{w}_n + b_n \mathbf{e}_n - \hat{\mathbf{y}}_n\|_2^2 + \frac{\lambda}{2s_n} \|\mathbf{w}_n\|_2^2$$

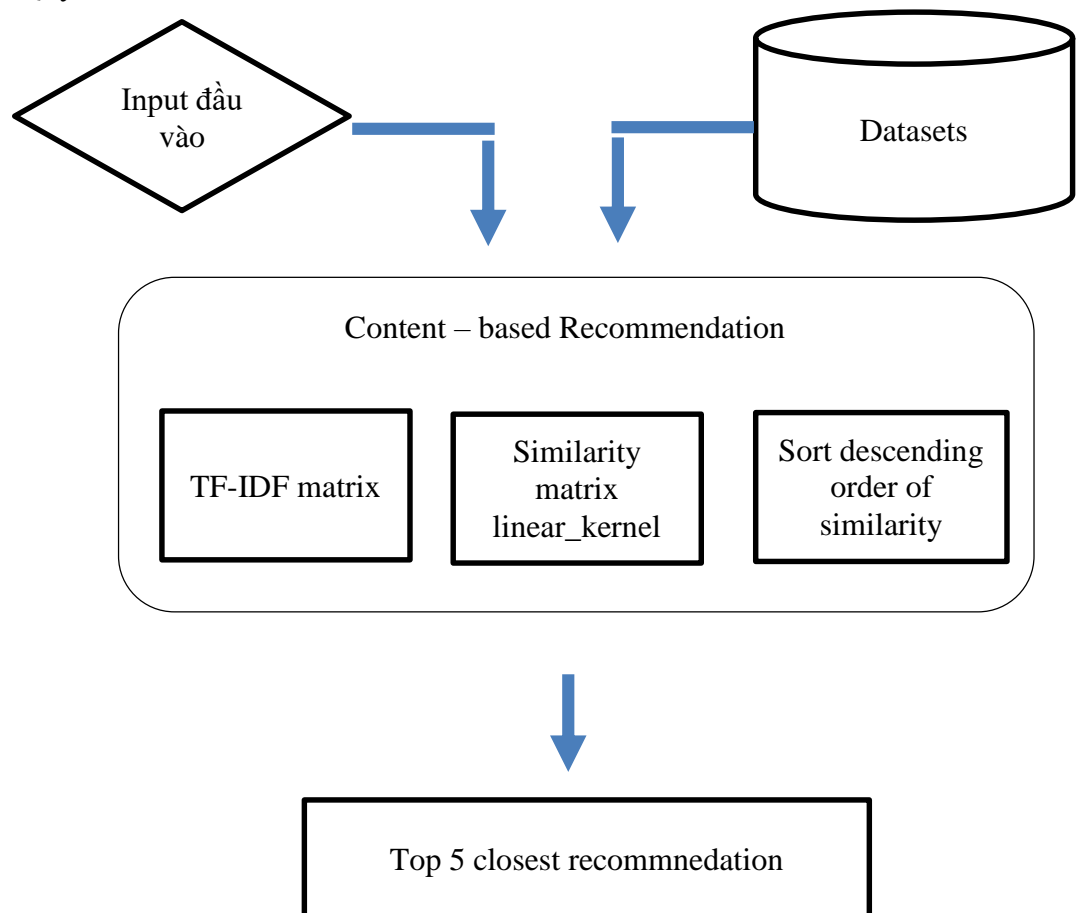
Trong đó, $\mathbf{e}(n)$ là vector cột chứa $s(n)$ phần tử 1.

2.2.3. Áp dụng vào bài toán thực tế với cơ sở dữ liệu datasets:

Phương pháp: Sử dụng phương pháp đề xuất dựa trên nội dung: NLTK, scikit-learning, TFIDF.

Bộ cơ sở dữ liệu thực tế gồm 408 nhà trọ và 461 khách sạn thuộc tỉnh Cần Thơ được chia thành bốn tập dữ liệu như đã nói ở phần *Tập dữ liệu*.

Quy trình đề xuất được hiện hiện như sau:



Hình 2.2: Sơ đồ gợi ý nhà trọ - khách sạn

Đầu tiên, hệ thống xây dựng ma trận TF-IDF cho dữ liệu có trong cơ sở dữ liệu, xây dựng feature vector cho mỗi item dựa trên ma trận dữ liệu cần thiết và feature TF-IDF. TfidfVectorizer về cơ bản sẽ chuyển đổi cột được sử dụng để gợi ý từ dạng text thành number. Tất cả các mô hình khoa học dữ liệu đều chạy trên các giá trị số vì máy tính chỉ có thể hiểu các số 0 và 1. TF-IDF là tần số tài liệu nghịch đảo-Tần số thuật ngữ. Số lượng tính năng mà nó tạo ra bằng tổng số từ riêng biệt được sử dụng trong cột tổng quan và các giá trị tỷ lệ thuận với số lần một từ cụ thể được sử dụng và tỷ lệ nghịch với số tài liệu được sử dụng.

Cách tính ma trận TF-IDF:

- TF: Tần suất kỳ hạn. Đây chỉ đơn giản là tần suất xuất hiện của một từ trong tài liệu.
- IDF: Tần suất Tài liệu Nghịch đảo. Đây là vũ trụ tần số tài liệu trong toàn bộ kho tài liệu.

Giả sử có 6 item được hiển thị trong bảng sau:

Item	Ninh	Kiều	Cái	Bình	Thủy
Item1	21	24	0	0	2
Item2	25	55	20	2	0
Item3	4	113	10	1	14
Item4	6	44	3	0	1
Item5	8	19	1	5	4
Item6	12	49	0	0	5
DF	5000	50000	10000	500000	7000

Tần suất kỳ hạn (TF):

Như đã thấy trong hình trên, đối với item 1, từ “Ninh” có TF là $1 + \log_{10} 21 = 2.322$. Theo cách này, TF được tính cho các thuộc tính khác của từng bài báo. Các giá trị này tạo nên vector thuộc tính cho mỗi bài viết. Thực hiện phép tính này cho tất cả các từ khóa trên tất cả các mục có:

Item	Ninh	Kiều	Cái	Bình	Thủy	Độ dài vector
Item1	2.322219295	2.380211242	0	0	1.325689898	3.800456039
Item2	2.382112402	2.770852012	1.91235468	1.30123569	0	4.004460697
Item3	2.602059991	3.06069784	1.61256202	1	1.235698789	5.380804488
Item4	1.602055991	2.447125634	1.63259874	0	1	3.527276247
Item5	1.903089987	2.658971254	1.25647895	1.45689621	1.625987445	4.257450611
Item6	2.230448921	2.69870125	0	0	1.698754121	4.326697114

Tần suất tài liệu nghịch đảo (IDF):

IDF được tính bằng cách lấy nghịch đảo logarit của tần suất tài liệu trong toàn bộ kho tài liệu. Vì vậy, nếu có tổng cộng 1 triệu tài liệu được trả về bởi truy vấn tìm kiếm và trong số các tài liệu đó, 'Bình' xuất hiện trong 0,5 triệu tài liệu. Do đó, điểm IDF của nó sẽ là: $\log_{10} (10^6 / 500000) = 0,30$.

DF	5000	50000	10000	500000	7000
IDF	2.301029996	1.301029996	2	0.301029996	2.15490196
N	10^6				

Có thể thấy, thuật ngữ phổ biến nhất 'Bình' đã được IDF gán cho trọng số thấp nhất. Độ dài của các vector này được tính bằng **căn bậc hai của tổng các giá trị bình phương** của mỗi thuộc tính trong vector:

Item	Ninh	Kiều	Cái	Bình	Thủy	Độ dài vector
Item1	2.322219295	2.380211242	0	0	1.325689898	3.800456039

Sau khi đã tìm ra trọng số TF -IDF và cả độ dài vector → chuẩn hóa các vector.

Item	Ninh	Kiều	Cái	Bình	Thủy	Tổng độ dài
Item1	2.322219295	2.380211242	0	0	1.325689898	1
Item2	2.382112402	2.770852012	1.91235468	1.30123569	0	1
Item3	2.602059991	3.06069784	1.61256202	1	1.235698789	1
Item4	1.602055991	2.447125634	1.63259874	0	1	1
Item5	1.903089987	2.658971254	1.25647895	1.45689621	1.625987445	1
Item6	2.230448921	2.69870125	0	0	1.698754121	1

Mỗi vector số hạng được chia cho độ dài vector tài liệu để có được vector chuẩn hóa. Vì vậy, đối với từ 'Ninh' trong item 1, vector chuẩn hóa là: $2.322 / 3.800$. Tương tự như vậy, tất cả các thuật ngữ trong tài liệu đều được chuẩn hóa và như bạn có thể thấy (hình trên) mỗi vector tài liệu hiện có độ dài là 1.

Bây giờ, đã có được các vector chuẩn hóa, tính các giá trị cosin để tìm ra sự giống nhau giữa các bài báo. Với mục đích này, chúng tôi sẽ chỉ lấy ba bài báo và ba thuộc tính.

Item	Ninh	Kiều	Bình
Item	0.61103701	0.626296217	0
Item	0.594389962	0.691941368	0.324895184
Item	0.483581962	0.568817887	0.353681311
Item Vector	Cosin Values		
Cos(A1,A2)	0.796554526		
Cos(A2,A3)	0.795934246		
Cos(A1,A3)	0.651734976		

Cos (A1, A2) đơn giản là tổng tích số chấm của các vector số hạng chuẩn hóa từ cả hai bài báo. Cách tính như sau:

$$\text{Cos (A1, A2)} = 0,611 * 0,59 + 0,63 * 0,69 + 0 * 0,32 = 0.7965$$

Như vậy, item 1 và 2 giống nhau nhất và do đó xuất hiện ở hai vị trí đầu của kết quả tìm kiếm. Ngoài ra, Article 1 (M1) thiên về phân tích hơn là Cloud-Analytics có trọng số là 0,611 trong khi đám mây có 0 sau khi chuẩn hóa độ dài.

Để tìm sự giống nhau giữa các nhà trọ - khách sạn và đưa ra gợi ý, sẽ sử dụng cosine_similarity và sắp xếp các điểm tương đồng theo thứ tự giảm

dần. Cosine_Similarity về cơ bản là thước đo mức độ giống nhau giữa 2 vector.

Khi người dùng nhập vào yêu cầu thực thi, hệ thống tạo một chuỗi ánh xạ chỉ mục của ma trận với yêu cầu được nhập để tìm ra các gợi ý phù hợp nhất. Cuối cùng trả về Top 5 gợi ý phù hợp nhất.

	Tên nhà trọ	Giá
27	Phòng trọ sinh viên giá rẻ địa chỉ 286/13 CMT8	700000
26	Cho thuê phòng trọ đầy đủ tiện nghi	1500000
35	Cho thuê phòng trọ đầy đủ tiện nghi tại trung tâm	1500000
13	Cho thuê phòng trọ đầy đủ tiện nghi tại trung tâm	1500000
23	Cho thuê phòng trọ sinh viên giá 700	700000

Bảng 9: Gợi ý nhà trọ

CHƯƠNG 3

HỆ THỐNG THỬ NGHIỆM

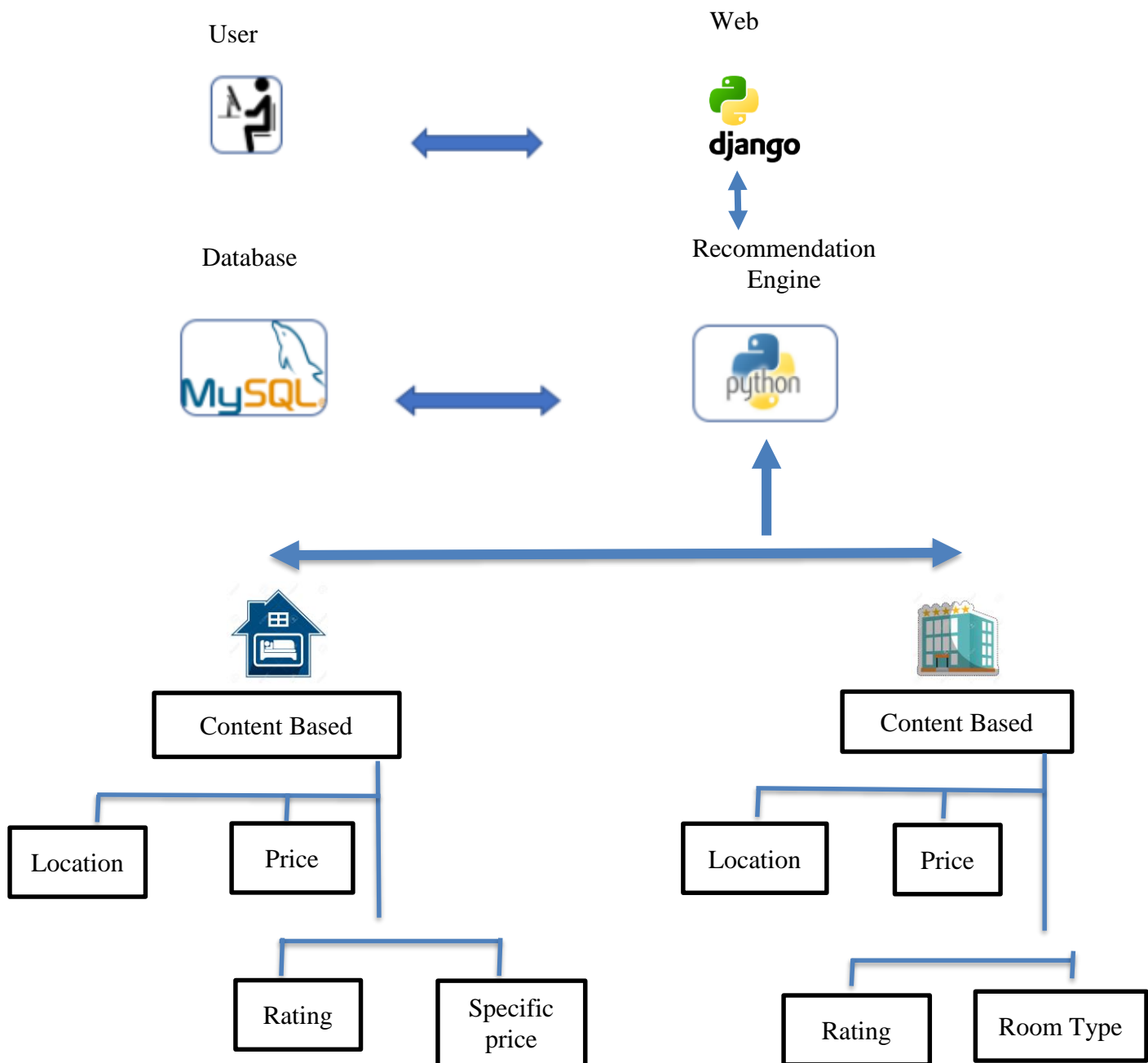
Thiết kế 1 hệ thống thử nghiệm. Hệ thống sử dụng ngôn ngữ lập trình Python, hệ quản trị cơ sở dữ liệu MySQL và framework Django.

3.1. Mục đích của việc xây dựng website:

Xây dựng một Website gợi ý các nhà trọ và khách sạn ở Cần Thơ, cho phép người dùng có thể truy cập vào trang web để tìm kiếm các nhà trọ hoặc khách sạn phù hợp với nhu cầu của bản thân, giúp tiết kiệm thời gian trong việc tìm kiếm nhà trọ, khách sạn – đặc biệt trong thời đại công nghệ 4.0.

3.2. Sơ đồ hệ thống tổng quát:

Sơ đồ tổng quát Website gợi ý nhà trọ - khách sạn Cần Thơ:

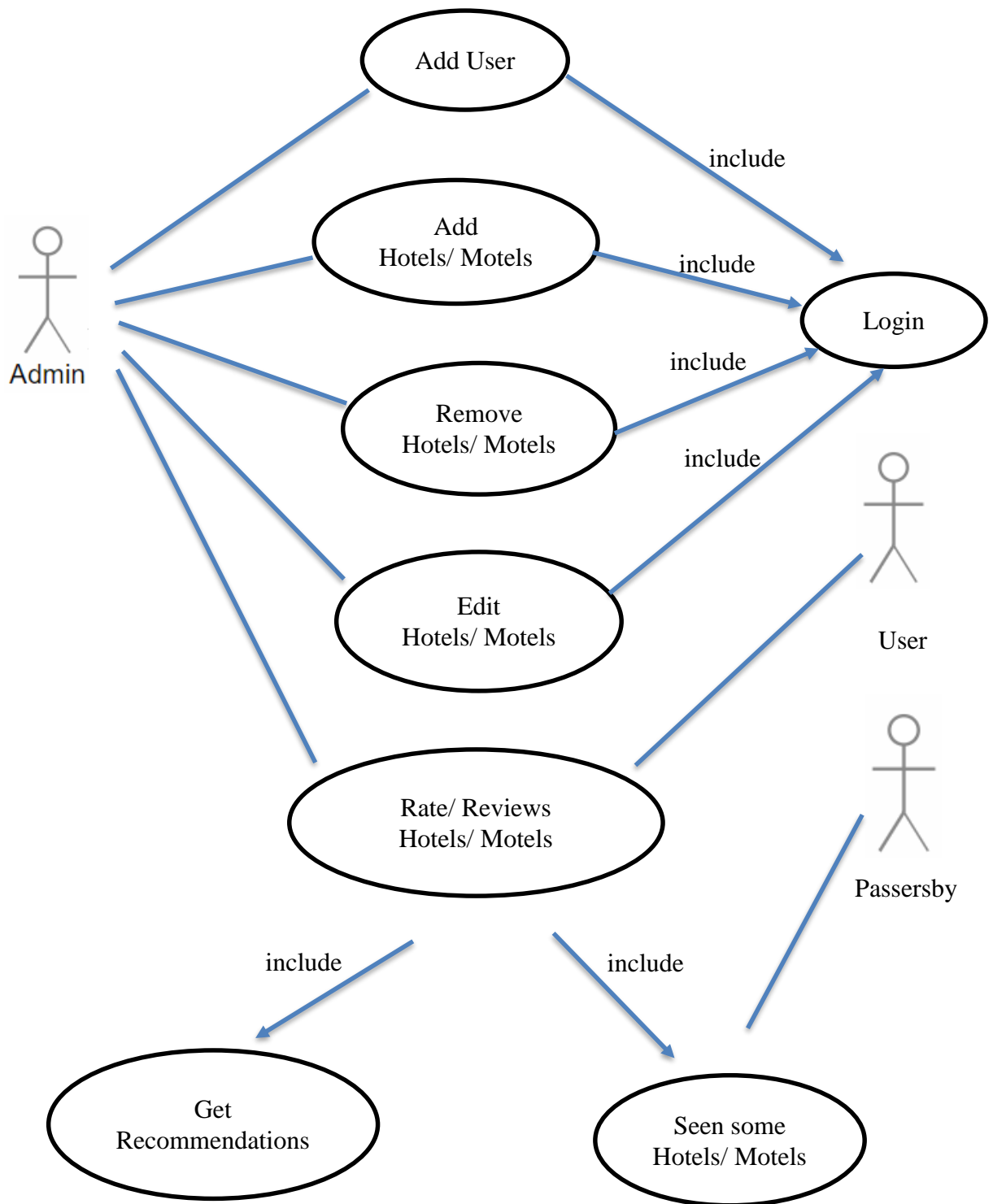


Hình 3.1: Sơ đồ tổng quát Website gợi ý nhà trọ - khách sạn

Những việc Website có thể làm:

- Đề xuất nhà trọ dựa trên nơi tọa lạc và giá của nhà trọ hoặc dựa trên điểm đánh giá và lượt views để đề xuất nhà trọ thích hợp.
- Đề xuất các khách sạn dựa trên nơi tọa lạc, điểm đánh giá và mức giá hoặc dựa trên khoảng cách đến trung tâm Ninh Kiều và loại phòng để đưa ra đề xuất khách sạn hợp lý.

3.3. Phân tích hệ thống người dùng website:



Hình 3.2: Hệ thống người dùng website

Có 3 kiểu người dùng:

- Người dùng không có tài khoản trên Website (khách vãng lai): Khi truy cập vào website có thể xem một số nhà trọ, khách sạn phổ biến ở Cần Thơ ở thời điểm hiện tại.
- Người dùng có tài khoản trên Website: Ngoài việc xem các nhà trọ, khách sạn phổ biến ở thời điểm hiện tại, người dùng có thể tìm kiếm thông qua hệ thống gợi ý của website để chọn ra nhà trọ, khách sạn ưng ý theo yêu cầu của mình, có thể xem chi tiết nhà trọ, khách sạn, đặt phòng cũng như để lại bình luận cho các nhà trọ.
- Người dùng là Admin của hệ thống: Quản lý các tài khoản của người dùng và quản trị nội dung của Website.

3.4. Đặc tả quy trình nghiệp vụ của hệ thống:

3.4.1. Người dùng không có tài khoản:

Đối với người dùng không có tài khoản trên Website sẽ được sử dụng một số chức năng trên website như sau:

Chức năng xem trên website: xem một số nhà trọ và khách sạn được nhiều người chọn hoặc đánh giá cao ở thời điểm hiện tại.

3.4.2. Người dùng có đăng ký tài khoản trên website:

Đối với người dùng có tài khoản trên website, ngoài chức năng của người dùng bình thường (khách vãng lai). Khi đăng nhập vào website, người dùng có thể tìm kiếm nhà trọ, khách sạn theo nhu cầu của mình (giá tiền, đánh giá tốt hay xấu...), xem chi tiết của từng nhà trọ, khách sạn, có thể tiến hành đặt phòng và để lại bình luận (trải nghiệm) của mình sau khi ở nhà trọ hay khách sạn ở thời gian.

Sau đó người dùng có thể đăng xuất ra khỏi tài khoản của mình và sử dụng chức năng của website như người dùng không có tài khoản.

Nếu người dùng không đăng xuất khỏi hệ thống mà trực tiếp tắt trang web thì tài khoản của người dùng sẽ được lưu trên trang web cho lần sử dụng tiếp theo.

3.4.3. Người dùng hệ thống (Admin):

3.4.3.1. Quản trị nội dung Website:

Admin sẽ có quyền xóa bỏ những nhà trọ hoặc khách sạn không hoạt động khỏi hệ thống, thay đổi các quảng cáo (nếu có), xóa những bình luận của người dùng có nội dung không phù hợp.

3.4.3.2. Quản trị người dùng:

Admin sẽ quản lý các quyền của người dùng, cấp quyền và phân quyền cho người dùng.

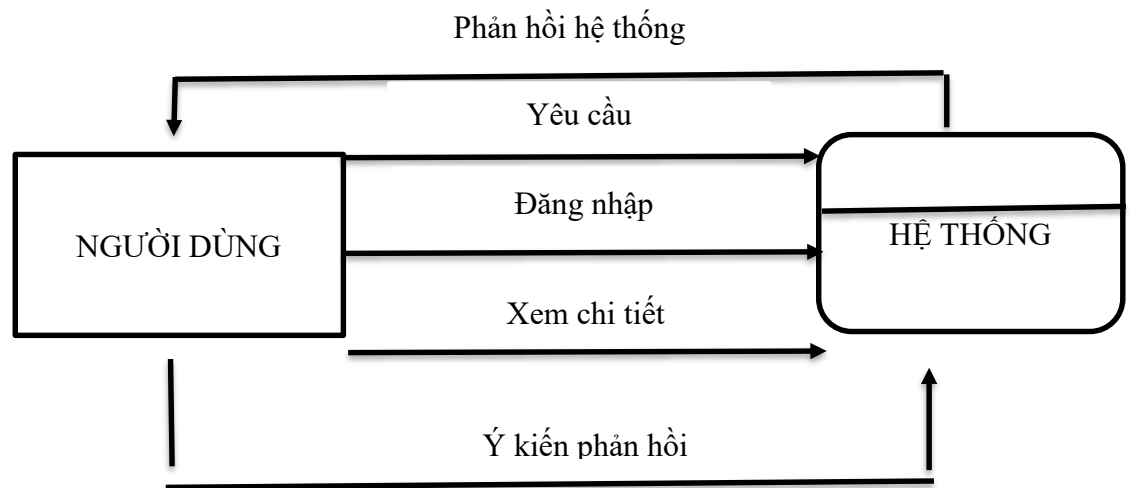
Đăng ký tài khoản: Người dùng có thể đăng ký tài khoản trên website. Người dùng sẽ cung cấp các thông tin mà hệ thống yêu cầu, khi hệ thống

kiểm tra tất cả các thông tin mà người dùng cung cấp điều hợp lệ thì người dùng sẽ được mở một tài khoản mới. Và khách hàng có thể sử dụng các công cụ, chức năng trên website ngay lập tức.

Xóa tài khoản người dùng: Những tài khoản vi phạm nội quy của website sẽ bị Admin xóa tài khoản.

3.5. Lập mô hình nghiệp vụ:

3.5.1. Biểu đồ ngữ cảnh hệ thống:

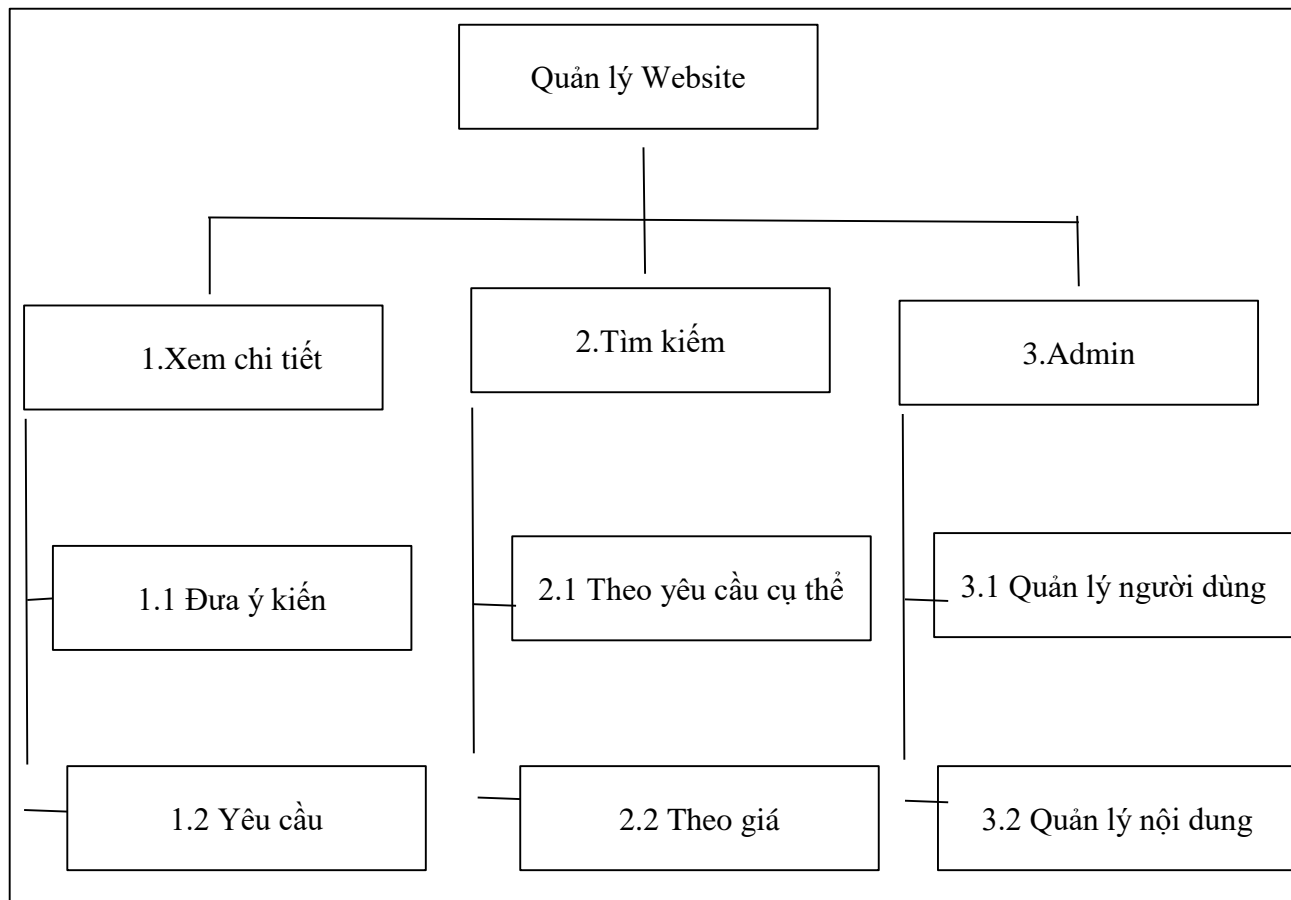


Hình 3.3: Biểu đồ ngữ cảnh hệ thống

Người dùng có thể yêu cầu, đăng nhập, xem chi tiết và để lại ý kiến của mình khi tương tác với hệ thống.

Về hệ thống thì sẽ tương tác với người dùng bằng cách đưa ra phản hồi cho người dùng.

3.5.2. Biểu đồ phân rã chức năng:



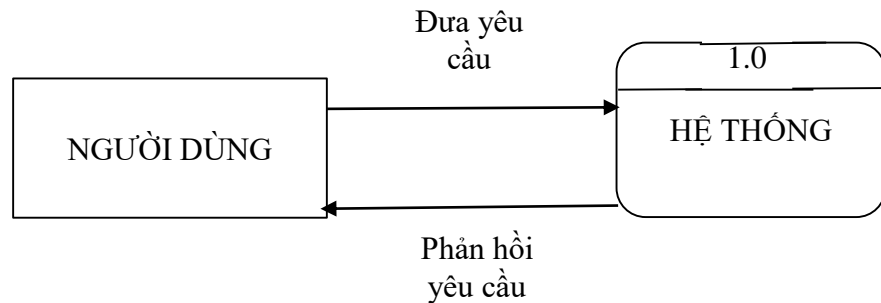
Hình 3.4: Biểu đồ phân rã chức năng

Website có 3 chức năng lớn là xem chi tiết, tìm kiếm và admin:

- Chức năng xem chi tiết có thể đưa ra ý kiến và yêu cầu
- Chức năng tìm kiếm có thể tìm kiếm theo nhiều yêu cầu cụ thể khác nhau
- Admin có quyền quản lý cấp quyền cho người dùng và quản lý nội dung trang web

3.5.3. Phân rã biểu đồ luồng dữ liệu:

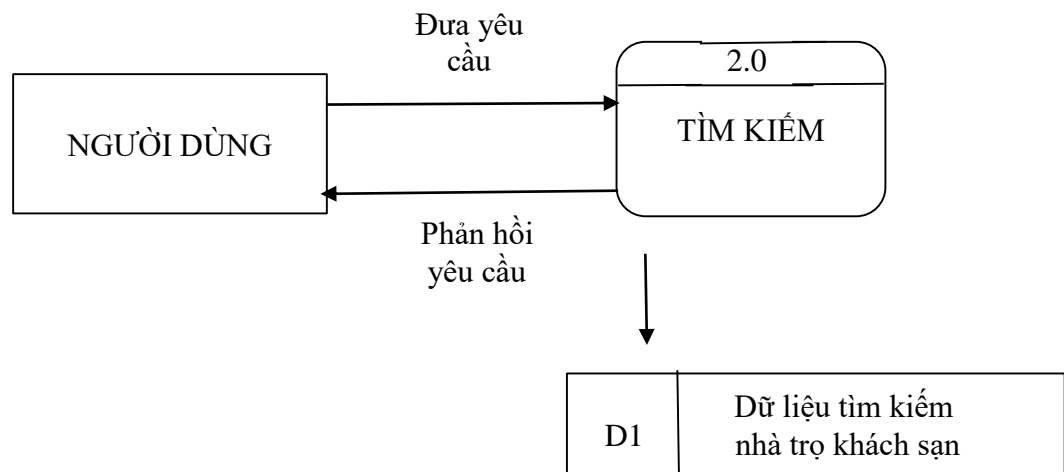
3.5.3.1. Biểu đồ luồng phân rã cấp 1.0:



Hình 3.5: Biểu đồ luồng phân rã cấp 1.0

Khi người dùng gửi yêu cầu, hệ thống website sẽ thực hiện yêu cầu, sau đó trả về kết quả cho người dùng, kết quả trả về được hiển thị trên giao diện website.

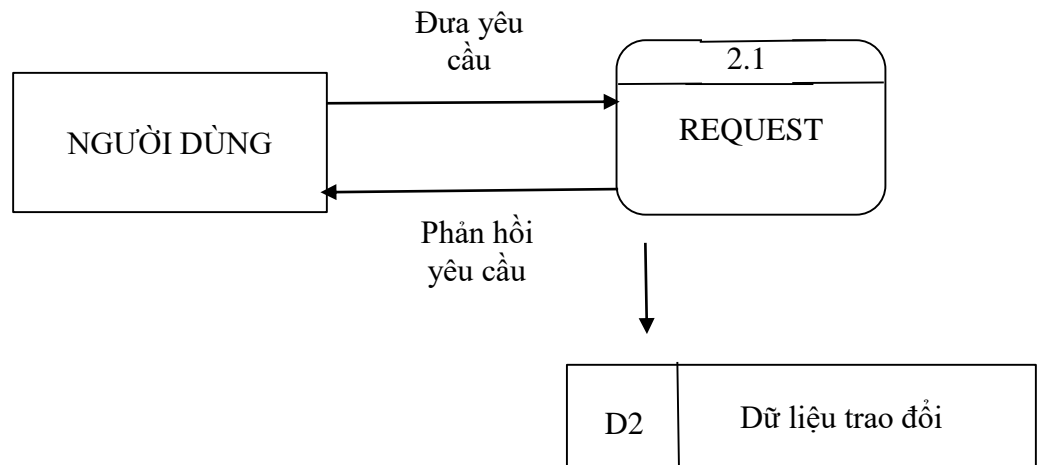
3.5.3.2. Biểu đồ luồng phân rã cấp 2.0:



Hình 3.6: Biểu đồ luồng phân rã cấp 2.0

Khi người dùng nhập yêu cầu, hệ thống sẽ xử lý dữ liệu và trả kết quả gợi ý về cho người dùng.

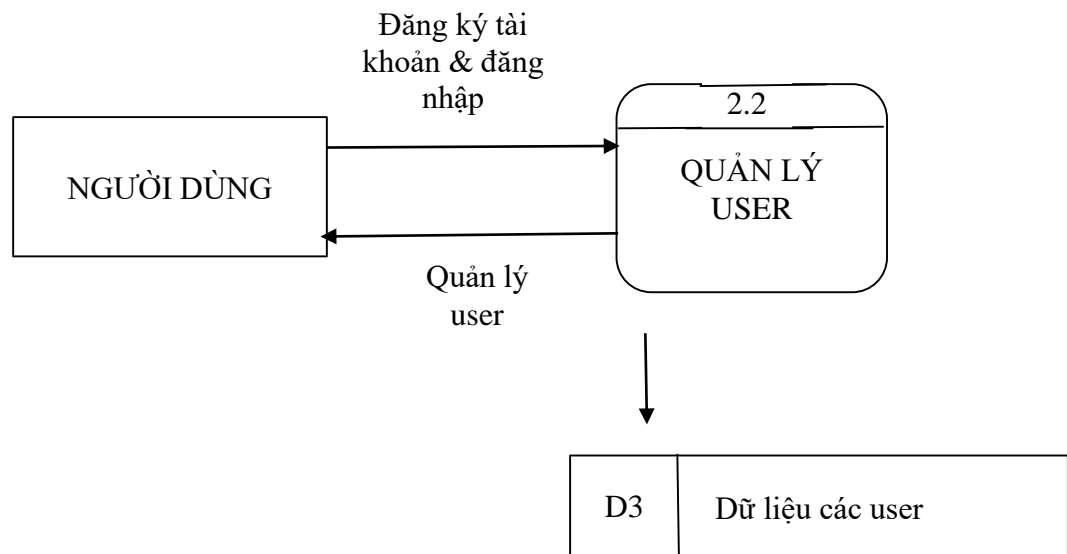
3.5.3.3. Biểu đồ luồng phân rã cấp 2.1:



Hình 3.7: Biểu đồ luồng phân rã cấp 2.1

Khi người dùng nhập yêu cầu, hệ thống sẽ xử lý dữ liệu và trả kết quả gợi ý về cho người dùng.

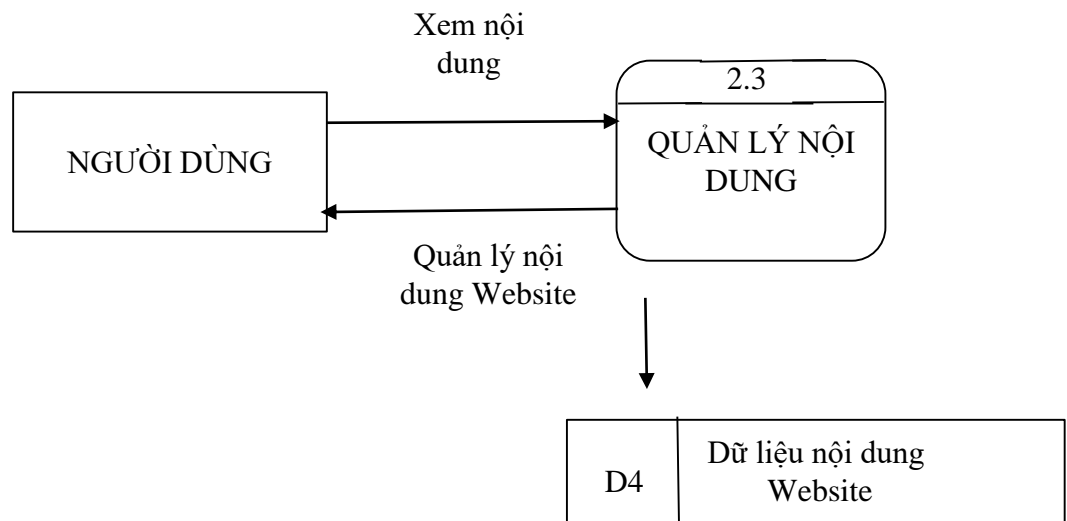
3.5.3.4. Biểu đồ luồng phân rã cấp 2.2:



Hình 3.8: Biểu đồ luồng phân rã cấp 2.2

Tất cả người dùng khi đăng ký tài khoản sẽ được quản lý bởi Admin và được lưu trữ ở bảng User trong cơ sở dữ liệu.

3.5.3.5. Biểu đồ luồng phân rã cấp 2.3:

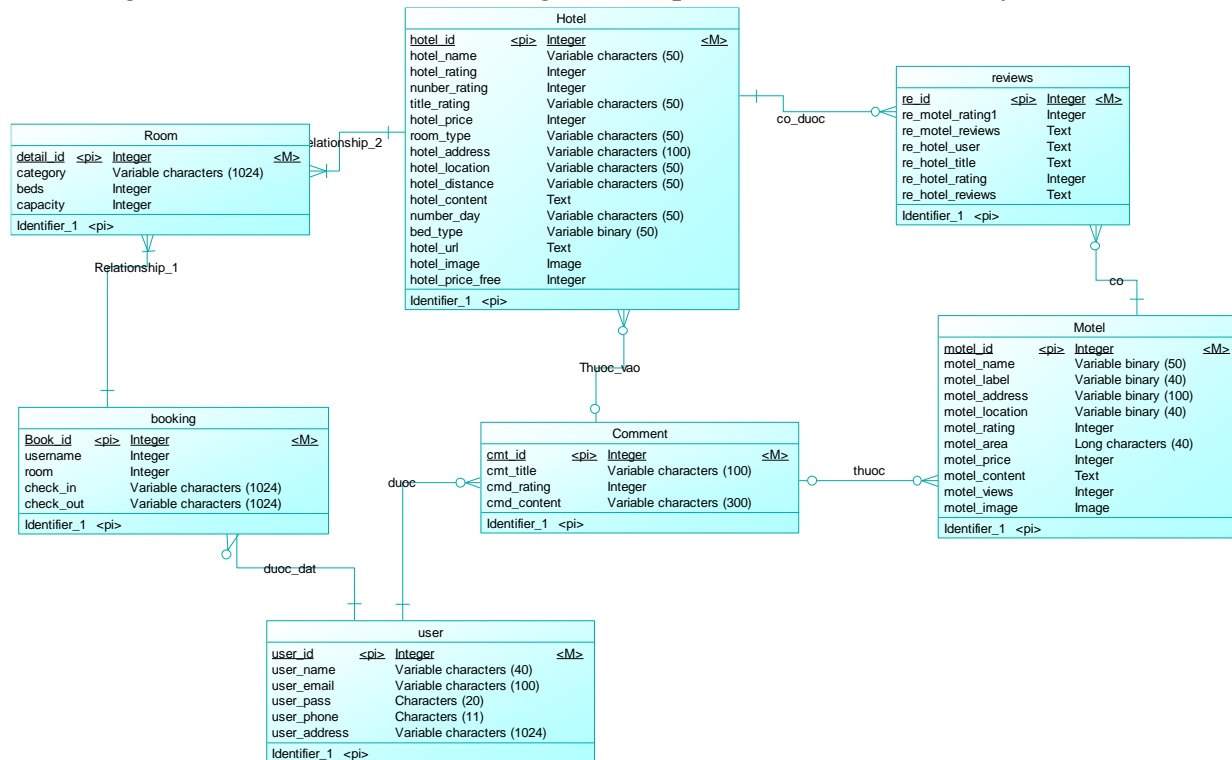


Hình 3.9: Biểu đồ luồng phân rã cấp 2.3

Người dùng có thể xem các trang nội dung và có thể để lại bình luận ở các trang nội dung từng nhà trọ - khách sạn.

3.5. Thiết kế các bảng dữ liệu:

Các bảng dữ liệu được thiết kế sử dụng cho hệ quản trị cơ sở dữ liệu MySQL.



Cơ sở dữ liệu gồm có 7 bảng:

- Bảng User: lưu trữ tài khoản người dùng
- Bảng Motel: lưu trữ thông tin nhà trọ
- Bảng Comment: lưu trữ đánh giá về nhà trọ
- Bảng Hotel: lưu trữ thông tin khách sạn
- Bảng Reviews: lưu trữ đánh giá khách sạn
- Bảng Booking: lưu trữ thông tin đặt phòng
- Bảng Room: lưu trữ thông tin phòng

3.5.1. Bảng User:

Là bảng lưu trữ tài khoản người dùng đăng nhập, thông tin lưu trữ bao gồm:

Tên người dùng (username)

Địa chỉ email (email address)

Mật khẩu người dùng (password)

Field	Type	Collation	Attributes	Null
Username	Varchar(255)	Utf8_general_ci		No
Email address	Varchar(255)	Utf8_general_ci		No
Password	Varchar(20)	Utf8_general_ci		No

Bảng 10: Bảng User

3.5.2. Bảng Motel:

Là bảng lưu trữ các thông tin cần thiết của các nhà trọ, thông tin lưu trữ bao gồm:

Id nhà trọ (MOTEL_ID)

Tên nhà trọ (MOTEL_NAME)

Phân lớp nhà trọ (MOTEL_LABEL)

Địa chỉ nhà trọ (MOTEL_ADDRESS)

Nơi tọa lạc (MOTEL_LOCATION)

Đánh giá sao (MOTEL_RATING)

Diện tích (MOTEL_AREA)

Giá nhà trọ (MOTEL_PRICE)

Nội dung nhà trọ (MOTEL_CONTENT)

Lượt xem của nhà trọ (MOTEL_VIEWS)

Hình ảnh nhà trọ (MOTEL_IMAGE)

Mức giá (MOTEL_PRICE_2)

Phân lớp lượt xem (MOTEL_VIEWS_2)

Field	Type	Collation	Attributes	Null
MOTEL_ID	Int(11)			No
MOTEL_NAME	Varchar(100)	Utf8_general_ci		
MOTEL_LABEL	Varchar(100)	Utf8_general_ci		
MOTEL_ADDRESS	Varchar(100)	Utf8_general_ci		
MOTEL_LOCATION	Varchar(100)	Utf8_general_ci		
MOTEL_RATING	Int(11)			
MOTEL_AREA	Longtext	Utf8_general_ci		
MOTEL_PRICE	Int(11)			
MOTEL_CONTENT	Text	Utf8_general_ci		
MOTEL_VIEWS	Int(11)			
MOTEL_IMAGE	Text	Utf8_general_ci		
MOTEL_PRICE_2	Text	Utf8_general_ci		
MOTEL_VIEWS_2	Text	Utf8_general_ci		

Bảng 11: Bảng Motel

3.5.3. Bảng Comment:

Là bảng lưu trữ những bình luận (reviews) của người dùng về nhà trọ, thông tin lưu trữ gồm:

Id reviews (RE_ID)

Id người dùng (USER_ID)

Id nhà trọ (MOTEL_ID)

Đánh giá sao (RE_MOTEL_RATING1)

Nội dung đánh giá (RE_MOTEL_REVIEWS)

Field	Type	Collation	Attributes	Null
RE_ID	Int(11)			No
USER_ID	Int(11)	Utf8_general_ci		No
MOTEL_ID	Int(11)	Utf8_general_ci		No
RE_MOTEL_RATING1	Int(11)	Utf8_general_ci		
RE_MOTEL_REVIEWS	Text	Utf8_general_ci		

Bảng 12: Bảng Reviews_motel

3.5.4. Bảng Hotel:

Là bảng lưu trữ các thông tin cần thiết của các khách sạn, thông tin gồm:

Id khách sạn (HOTEL_ID)

Tên khách sạn (HOTEL_NAME)

Điểm đánh giá (HOTEL_RATING)

Số lượt đánh giá (NUMBER_RATING)

Phân lớp đánh giá (TITLE_RATING)

Giá (HOTEL_PRICE)

Giá khuyến mãi (HOTEL_PRICE_FREE)

Loại phòng (ROOM_TYPE)

Địa chỉ (HOTEL_ADDRESS)

Nơi tọa lạc (HOTEL_LOCATION)

Khoảng cách so với trung tâm thành phố (HOTEL_DISTANCE)

Nội dung khách sạn (HOTEL_CONTENT)

Số ngày (NUMBER_DAY)

Loại giường ngủ (BED_TYPE)

Url khách sạn (HOTEL_URL)

Hình ảnh khách sạn (HOTEL_IMAGE)

Phân lớp giá khách sạn (HOTEL_PRICE_FREE)

Field	Type	Collation	Attributes	Null
HOTEL_ID	Int(11)			No
HOTEL_NAME	Varchar(50)	Utf8_general_ci		
HOTEL_RATING	Int(11)			
NUMBER_RATING	Int(11)			
TITLE_RATING	Varchar(50)	Utf8_general_ci		
HOTEL_PRICE	Int(11)			
HOTEL_PRICE_FREE	Int(11)			
ROOM_TYPE	Varchar(50)	Utf8_general_ci		

HOTEL_ADDRESS	Varchar(100)	Utf8_general_ci		
HOTEL_LOCATION	Varchar(50)	Utf8_general_ci		
HOTEL_DISTANCE	Varchar(50)	Utf8_general_ci		
HOTEL_CONTENT	Text	Utf8_general_ci		
NUMBER_DAY	Varchar(50)	Utf8_general_ci		
BED_TYPE	Varchar(100)	Utf8_general_ci		
HOTEL_URL	Text	Utf8_general_ci		
HOTEL_IMAGE	Text	Utf8_general_ci		
HOTEL_PRICE_FREE	Text	Utf8_general_ci		

Bảng 13: Bảng Hotel

3.5.5. Bảng Reviews:

Là bảng lưu trữ những bình luận (reviews) của người dùng về nhà trọ, thông tin lưu trữ gồm:

Id reviews (CMT_ID)

Id người dùng (USER_ID)

Id khách sạn (HOTEL_ID)

Tên người dùng (HOTEL_USER)

Tựa đề bài đánh giá (CMT_TITLE)

Số điểm đánh giá (CMD_RATING)

Nội dung đánh giá (CMD_CONTENT)

Field	Type	Collation	Attributes	Null
CMT_ID	Int(11)			No
USER_ID	Int(11)			No
HOTEL_ID	Int(11)			No
HOTEL_USER	Varchar(255)	Utf8_general_ci		No
CMT_TITLE	Varchar(100)	Utf8_general_ci		
CMD_RATING	Int(11)			
CMD_CONTENT	Varchar(300)	Utf8_general_ci		

Bảng 14: Bảng Reviews_hotel

3.5.6. Bảng Booking:

Là bảng lưu trữ những người dùng đặt phòng khách sạn, thông tin lưu trữ bao gồm:

Tên người dùng (username)

Loại phòng (room)

Ngày đến (check_in)

Ngày đi (check_out)

Field	Type	Collation	Attributes	Null
Username	Varchar(20)	Utf8_general_ci		No
Room	Varchar(20)	Utf8_general_ci		No
Check_in	Text	Utf8_general_ci		No
Check_out	Text	Utf8_general_ci		No

Bảng 15: Bảng Booking

3.5.7. Bảng Rooms:

Là bảng lưu trữ các loại phòng, số phòng và số người ở, thông tin lưu trữ bao gồm:

Số thứ tự (number)

Loại phòng (category)

Số phòng (beds)

Số người ở (capacity)

Field	Type	Collation	Attributes	Null
Number	Int(10)			No
Category	Varchar(20)	Utf8_general_ci		No
Beds	Int(10)			No
Capacity	Int(30)			No

Bảng 16: Bảng Rooms

3.6. Mô tả giao diện của website:

Để Website tiện dụng cho người sử dụng, trang web được thể kế theo ý tưởng và bố cục như sau:

Giao diện trang web gồm một giao diện chính và hai giao diện phụ chia thành trang web gợi ý nhà trọ và gợi ý khách sạn.

Phần đầu trang chủ của trang web có chứa logo của trường Đại học Cần Thơ, chức năng đăng nhập và truy xuất đến hai trang phụ tìm kiếm, gợi ý nhà trọ và khách sạn, cùng với đó là các chức năng giúp liên kết đến các phần của trang chủ.

Phần thân của Website (Body): là phần quan trọng nhất của Website. Đầu tiên là một đoạn giới thiệu về Cần Thơ kèm theo hình ảnh một bên, tiếp theo là danh sách một số nhà trọ cũng như khách sạn nổi bật ở Cần Thơ cho người dùng xem tham khảo. Có thể xem nhà trọ hoặc khách sạn riêng biệt, hoặc có thể xem chung tùy vào ý thích của người dùng.

Phần footer: Ghi các thông tin về bản quyền của Website, tên và địa chỉ của chủ sở hữu Website, email liên hệ, ...

Chi tiết cấu trúc giao diện như sau:

LOGO CTU		GIỚI THIỆU	NỔI BẬT	LIÊN HỆ	ĐĂNG NHẬP
TÊN WEBSITE LIÊN KẾT NHÀ TRỌ - KHÁCH SẠN					
HÌNH ẢNH			GIỚI THIỆU SƠ LƯỢC VỀ CẦN THƠ		
MỘT SỐ NHÀ TRỌ NỔI BẬT					
BẢN QUYỀN – THÔNG TIN LIÊN HỆ					

Hình 3.10: Giao diện chính của website

Ở hai trang web gợi ý nhà trọ và khách sạn có cấu trúc giống nhau:

Phần header của trang web gồm logo trường Đại học Cần Thơ, các nút liên kết về trang chính, trang web gợi ý khách sạn, đăng xuất và liên hệ

Phần thân của Website (Body): là phần quan trọng nhất của Website. Đầu tiên là tên của trang web và khung tìm kiếm cho phép người dùng chọn theo một số yêu cầu nhất định. Tiếp theo là phần gợi ý các nhà trọ hoặc khách sạn theo yêu cầu người dùng, cuối cùng là khung google maps định vị khách sạn, nhà trọ.

Chi tiết cấu trúc giao diện như sau:

LOGO CTU		TRANG CHỦ	KHÁCH SẠN	ĐĂNG XUẤT	LIÊN HỆ
TÊN WEBSITE KHUNG TÌM KIẾM					
GỢI Ý NHÀ TRỌ - KHÁCH SẠN					
GOOGLE MAPS ĐỊNH VỊ					
FOOTER					

Hình 3.11: Website gợi ý nhà trọ và khách sạn

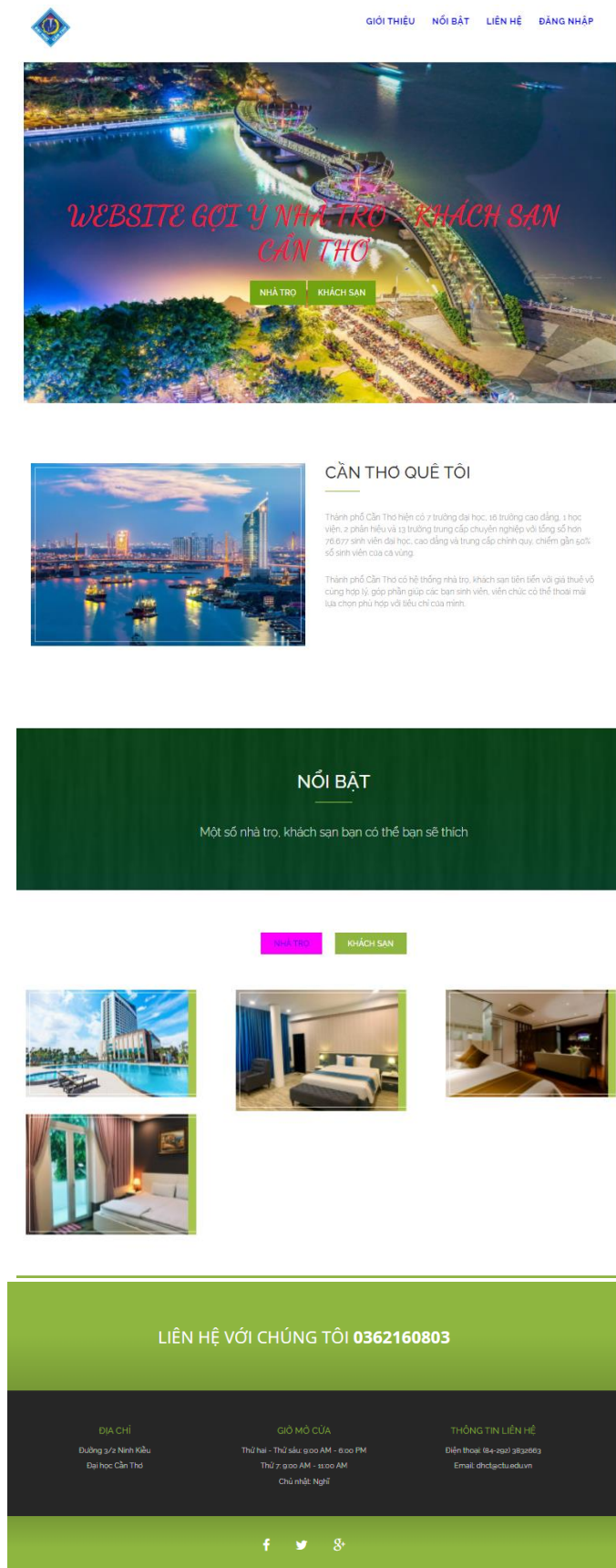
3.7. Giao diện website:

3.7.1. Giao diện trang chủ:

Khi truy cập vào Website, giao diện trang chủ sẽ hiện ra như sau:

Phần đăng nhập và đăng ký cho người dùng: khi truy cập vào trang web, người dùng chưa có tài khoản có thể đăng ký tài khoản trên website. Sau khi đăng ký hoàn tất người dùng có thể đăng nhập vào website, tham gia vào diễn đàn và gửi các yêu cầu và có thể click để đến trang tìm kiếm nhà trọ hoặc khách sạn cần thiết cho yêu cầu của mình.

Khi người dùng không có tài khoản, người dùng có thể xem được một số nhà trọ và khách sạn nổi bật ở trang chủ.



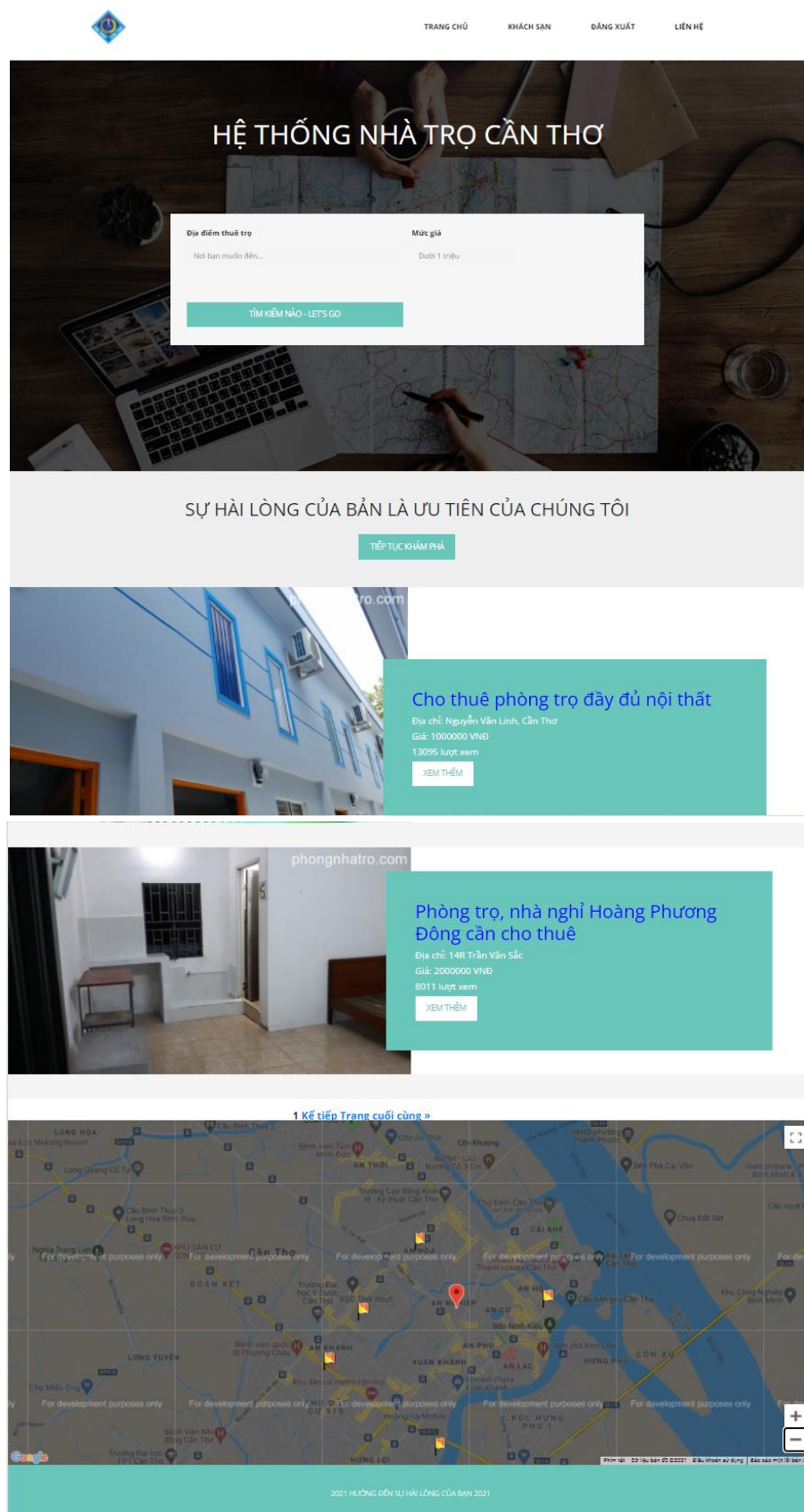
Hình 3.12: Giao diện đồ họa trang chính

3.7.2. Website gợi ý nhà trọ - khách sạn:

Khi vào đến trang gợi ý nhà trọ hoặc khách sạn, giao diện sẽ hiện ra như sau:

Phần header: Người dùng có thể click để về trang chủ và trang khách sạn nếu ở trang nhà trọ và ngược lại, người dùng cũng có thể đăng xuất trực tiếp ở đây nếu không muốn thực hiện chương trình nào nữa.

Phần body: Người dùng có thể chọn một số yêu cầu cụ thể theo gợi ý của hệ thống để được đề xuất top 5 nhà trọ/ khách sạn phù hợp với yêu cầu nhất. Các nhà trọ và khách sạn sẽ được đề xuất cùng với một số thông tin cơ bản như tên, địa chỉ, giá, ... Người dùng có thể sử dụng bản đồ google maps để xem vị trí cụ thể của nhà trọ hay khách sạn đó nằm ở đâu.



Hình 3.13: Giao diện đồ họa khách sạn – nhà trọ

3.7.3. Chức năng đăng ký – đăng nhập:

Khi đăng ký, người dùng cần nhập những thông tin cơ bản cho hệ thống lưu vào như sau:

Tên đăng nhập (username)

Email

Mật khẩu (password)

Đăng ký

Nhập thông tin của bạn

Username:

Email:

Password:

Đăng nhập nhanh

ĐĂNG NHẬP

Xin chào!

Nhập thông tin cá nhân của bạn và bắt đầu hành trình với chúng tôi

Hình 3.14: Đăng ký người dùng

Sau khi đăng ký xong, người dùng có thể đăng nhập vào hệ thống với tên đăng nhập kèm theo password được người dùng tạo lúc đầu

Đăng nhập

Tài khoản của bạn

Username:

Password:

Bạn chưa có tài khoản

ĐĂNG NHẬP

Chào mừng trở lại!

Để giữ kết nối với chúng tôi, vui lòng đăng nhập bằng thông tin cá nhân của bạn

Hình 3.15: Đăng nhập người dùng

➔ Ở cả hai chức năng đăng ký và đăng nhập, người dùng có thể đăng nhập nhanh bằng tài khoản google của mình nếu muốn.

3.7.4. Chức năng đặt phòng:

Khi chọn được khách sạn ưng ý, người dùng có thể tiến hành đặt phòng đó trên hệ thống của website. Khi đặt phòng, người dùng cần nhập vào các thông tin cần thiết gồm:

- Họ tên
- Số điện thoại để liên hệ
- Loại phòng
- Chọn thời gian đặt phòng

ĐẶT PHÒNG

Vui lòng nhập đầy đủ thông tin bên dưới để đặt phòng

Họ tên:
Nguyễn Hữu Tính

Số điện thoại:
0362160803

Loại phòng
Phòng tiêu chuẩn 2 giường (Standard Twin Room)

Thời gian đặt phòng

Từ ngày:
21-11-2021

Đến ngày:
22-11-2021





Proceed

Về chúng tôi

Chúng tôi là một phần của chuỗi khách sạn sang trọng ở Cần Thơ. chúng tôi cung cấp một kỳ nghỉ sang trọng với nhiều giá trị khác nhau các dịch vụ bổ sung và miễn phí sẽ tạo ra bạn ghé thăm chúng tôi hơn và hơn nữa.

Địa chỉ

Cần Thơ, Việt Nam Phone : 02926 254 567
Email : internationalhotelvn@gmail.com

Hình 3.16: Chức năng đặt phòng

3.7.5. Chức năng bình luận:

Người dùng có thể bình luận cho khách sạn hoặc nhà trọ mà mình đã từng sử dụng ở trang nội dung của từng nhà trọ. Người dùng có thể đánh giá nhà trọ / khách sạn bao nhiêu sao (★★★★★) và để lại bình luận ở dạng văn bản cho người dùng khác có thể hiểu biết thêm về nhà trọ - khách sạn.

PHẦN KẾT LUẬN

1. Kết quả đạt được:

- Hoàn thành tập dữ liệu nhà trọ và khách sạn thuộc khu vực tỉnh Cần Thơ
- Xây dựng hệ thống gợi ý dựa trên nội dung
- Xây dựng Website gợi ý nhà trọ, khách sạn Cần Thơ với framework Django với các chức năng:
 - ✓ Gợi ý theo địa điểm, đánh giá của người dùng và gợi ý theo giá của nhà trọ, khách sạn.
 - ✓ Thể hiện được địa điểm của nhà trọ, khách sạn trên bản đồ địa lý.
 - ✓ Cho phép người dùng nhập nhận xét của mình về nhà trọ, khách sạn đó → Thuận lợi cho việc thu thập dữ liệu và cải tiến hệ thống gợi ý trong tương lai.
 - ✓ Thực hiện đặt phòng khách sạn.

2. Hạn chế:

- Gợi ý theo khoảng cách gần một địa điểm nào đó theo khoảng cách (km) chưa thực hiện
- Không kết nối được với các khách sạn để thực hiện gợi ý theo thời gian

3. Hướng phát triển:

Phát triển hệ thống Website với các gợi ý:

- Gợi ý tìm kiếm bằng cách cho người dùng nhập văn bản (dạng text) và gợi ý nhà trọ, khách sạn thích hợp.
- Gợi ý theo khoảng cách gần một địa điểm nào đó theo khoảng cách (km). Ví dụ: Tìm nhà trọ xung quanh đại học Cần Thơ bán kính 500m
- Phát triển kết nối thực tế với các nhà trọ, khách sạn hoàn thành gợi ý nhà trọ, khách sạn theo thời gian
- Phát triển phần xác định giá trị rating và rating title theo cách hai: vector hóa cột user review từ tập dữ liệu datasets_hotels_reviews.csv hoặc datasets_motels_reviews.csv và xây dựng mô hình trên một phần dữ liệu đã có đủ thông tin, phần còn lại đánh giá mô hình. Khi mô hình đã ổn, dùng nó để tính rating và rating_title

TÀI LIỆU THAM KHẢO

- [1] M. M. Mariani, D. Buhalis, C. Longhi, and O. Vitouladitis, "Managing change in tourism destinations: Key issues and current trends," *Journal of Destination Marketing & Management*, vol. 2, no. 4, pp. 269-272, 2014.
- [2] H. Liu, J. He, T. Wang, W. Song, and X. Du, "Combining user preferences and user opinions for accurate recommendation," *Electronic Commerce Research and Applications*, vol. 12, no. 1, pp. 14-23, 2013.
- [4] Francesco, R.; Rokach, L.; Shapira, B. Recommender systems: Introduction and challenges. In *Recommender Systems Handbook*; Springer: Boston, MA, USA, 2015; pp. 1–34.
- [6] Li, C.; Chen, G.; Wang, F. Recommender systems based on user reviews: The state of the art. *User Model. User Adapt. Interact.* 2015, 25, 99–154.
- [7] Yue, M.; Chen, G.; Wei, Q. Finding users' preferences from large-scale online reviews for personalized recommendation. *Electron. Commer. Res.* 2017, 17, 3–29
- [8] He, X.; Chen, T.; Kan, M. Y.; Chen, X. Trirank: Review-aware explainable recommendation by modeling aspects. *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, 19–23 October 2015, Melbourne, Australia.
- [9] Han, H.; Huang, M.; Zhang, Y.; Bhatti, U.A. An Extended-Tag-Induced Matrix Factorization Technique for Recommender Systems. *Information* 2018, 9, 143.
- [10] Alshammari, G.; Jorro - Aragonese, J. L.; Polatidis, N.; Kapetanakis, S.; Pimenidis, E.; Petridis, M. A switching multi-level method for the long tail recommendation problem. *J. Intell. Fuzzy Syst.* 2019, 37, 7189–7198.
- [11] Su, J.-H.; Chang, W.; Tseng, V.S. Effective social content-based collaborative filtering for music recommendation. *Intell. Data Anal.* 2017, 21, S195–S216.
- [12] Zhang, Z.; Zhang, D.; Lai, J. urCF: User Review Enhanced Collaborative Filtering; AMCIS: Bubendorf, Switzerland, 2014.
- [13] Trần Nguyễn Minh Thư, 2011 - Trần Nguyễn Minh Thư, 2011. Abstraction et règles d'association pour l'amélioration des systèmes de recommandation à partir de données de préférences binaires. Phd thesis.

- [14] Herlocker J. L., et al, 2000. Explaining collaborative filtering recommendations. In Proceedings of the 2000 Conference on Computer Supported Cooperative Work, 241–250.
- [15] Breese, J.S. and D. Heckerman, 1998. Empirical analysis of predictive algorithms for collaborative filtering. Morgan Kaufmann, pp. 43–52.
- [16] Herlocker J.L et al, 2004. Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst., vol. 22, no. 1, pp. 5–53.
- [17] Adomavicius, G. and Y. Kwon, 2010. Improving recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering.
- [18] Adomavicius, G. and Y. Kwon, 2010. Improving recommendation diversity using ranking-based techniques. IEEE Transactions on Knowledge and Data Engineering.
- [19] Cosley D., et al, 2003. Is seeing believing?: how recommender system interfaces affect users' opinions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '03). ACM, New York, NY, USA, 585-592. DOI=10.1145/642611.642713 <http://doi.acm.org/10.1145/642611.642713>
- [20] <https://phongtro123.com/>
- [21] <https://sosanhnhha.com/>
- [22] <https://phongnhatro.com/>
- [23] <https://dithuenha.com/>
- [24] <https://www.agoda.com/>
- [25] <https://www.traveloka.com/>
- [26] <https://www.tripadvisor.com.vn/>
- [27] <https://www.booking.com/>