BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**🙞 🕮 🙜**



**LUẬN VĂN NGÀNH**

**NGÀNH KHOA HỌC MÁY TÍNH**

**Đề tài**

**HỆ THỐNG GỢI Ý**

**CÂU HỎI TƯƠNG TỰ**

**Sinh viên thực hiện : Nguyễn Tấn Pil**

**Mã số : B1812295**

**Khóa : 44**

Cần Thơ, 5/2022

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**🙞 🕮 🙜**



**LUẬN VĂN NGÀNH**

**NGÀNH KHOA HỌC MÁY TÍNH**

**Đề tài**

**HỆ THỐNG GỢI Ý**

**CÂU HỎI TƯƠNG TỰ**

**Giáo viên hướng dẫn: Sinh viên thực hiện:**

**TS.Trần Nguyễn Minh Thư Nguyễn Tấn Pil**

**Mã số: B1812295**

**Khóa : 44**

Cần Thơ, 5/2022

**NHẬN XÉT CỦA GIẢNG VIÊN**

------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Cần Thơ, ngày tháng 12 năm 2022

(GVHD ký và ghi rõ họ tên)

**LỜI CẢM ƠN**

Để hoàn thành đề tài luận văn này, em xin được bày tỏ lòng biết ơn chân thành và sâu sắc đến Thầy Lưu Tiến Đạo – người đã trực tiếp tận tình hướng dẫn, giúp đỡ em trong suốt quá trình thực hiện luận văn, nhờ sự chỉ bảo và hướng dẫn quý giá đó mà bài luận văn 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 suốt khoảng thời gian qua.

Em cũng xin chân thành cảm ơn bạn bè cùng với gia đình luôn 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 thật sự quyết tâm và cố gắng trong suốt quá trình thực hiện đề tài, 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 2022

Người viết

Nguyễn Tấn Pil

**MỤC LỤC**

[DANH MỤC HÌNH 3](#_Toc104224715)

[DANH MỤC BẢNG 4](#_Toc104224716)

[TÓM TẮT 5](#_Toc104224717)

[PHẦN GIỚI THIỆU 6](#_Toc104224718)

[1. Đặt vấn đề 6](#_Toc104224719)

[2. Lịch sử giải quyết vấn đề 7](#_Toc104224720)

[3. Mục tiêu đề tài 9](#_Toc104224721)

[4. Đối tượng và phạm vi nghiên cứu 9](#_Toc104224722)

[5. Phương pháp nghiên cứu 10](#_Toc104224723)

[6. Kết quả đạt được 10](#_Toc104224724)

[7. Bố cục niên luận 10](#_Toc104224725)

[PHẦN NỘI DUNG 11](#_Toc104224726)

[CHƯƠNG 1 - MÔ TẢ BÀI TOÁN 11](#_Toc104224727)

[1. Mô tả chi tiết bài toán 11](#_Toc104224728)

[2. Vấn đề và giải pháp liên quan đến bài toán 12](#_Toc104224729)

[2.1. Công cụ thiết kế website 12](#_Toc104224730)

[2.2. Hệ thống gợi ý 14](#_Toc104224731)

[CHƯƠNG 2 - THIẾT KẾ VÀ CÀI ĐẶT 21](#_Toc104224732)

[1. Thiết kế hệ thống 21](#_Toc104224733)

[1.1 Sơ đồ tổng quát 21](#_Toc104224734)

[1.2 Giới thiệu tập dữ liệu 21](#_Toc104224735)

[1.3 Cấu trúc của website 22](#_Toc104224736)

[2. Cài đặt giải thuật 23](#_Toc104224737)

[2.1 Sơ lược cơ sở dữ liệu 23](#_Toc104224738)

[2.2 Giải thuật chính trong hệ thống 26](#_Toc104224739)

[2.3. Xây dựng hệ thống gợi ý 31](#_Toc104224740)

[CHƯƠNG 3 – GIỚI THIỆU WEBSITE 33](#_Toc104224741)

[1. Giao diện chính hệ thống 33](#_Toc104224742)

[2. Đánh giá 40](#_Toc104224743)

[2.1 Given-N 40](#_Toc104224744)

[2.2 All-but-one 41](#_Toc104224745)

[PHẦN KẾT LUẬN 42](#_Toc104224746)

[1. Kết quả đạt được 42](#_Toc104224747)

[2. Hướng phát triển 42](#_Toc104224748)

[TÀI LIỆU THAM KHẢO 43](#_Toc104224749)

# DANH MỤC HÌNH

[Hình 1: Hôm nay ăn gì? 7](#_Toc103782722)

[Hình 2: Gợi ý từ website Amazon. 8](#_Toc103782723)

[Hình 3: Gợi ý từ Netflix 9](#_Toc103782724)

[Hình 4: Quy trình xây dụng hệ thống gợi ý 15](file:///C:\Users\DELL\OneDrive\Nien_Luan\Nguyen_Tan_Pil_Recipe_Recommendation.docx#_Toc103782725)

[Hình 5: Sơ đồ tổng quát hệ thống gợi ý 21](#_Toc103782726)

[Hình 6: Cấu trúc website 23](#_Toc103782727)

[Hình 7: Giải thuật đăng ký tài khoản 26](#_Toc103782728)

[Hình 8: Giải thuật đánh giá và trả lời đánh giá 27](#_Toc103782729)

[Hình 9: Giải thuật đưa ra gợi ý công thức nấu ăn 29](#_Toc103782730)

[Hình 10: Giải thuật cập nhật thông tin người dùng 30](#_Toc103782731)

[Hình 11: Dữ liệu công thức nấu ăn 31](#_Toc103782732)

[Hình 12: Dữ liệu người dùng 31](#_Toc103782733)

[Hình 13: Tạo ma trân quan hệ 32](#_Toc103782734)

[Hình 14: Tìm các công thức nấu ăn tương tự 32](#_Toc103782735)

[Hình 15: Lọc các công thức có độ tương tự cao nhất 32](#_Toc103782736)

[Hình 16: Giao diện trang chủ hệ thống 33](#_Toc103782737)

[Hình 17: Giao diện tất cả các công thức nấu ăn 33](#_Toc103782738)

[Hình 18: Giao diện các công thức nấu ăn phổ biến nhất 34](#_Toc103782739)

[Hình 19: Giao diện tìm kiếm công thức nấu ăn theo tên 34](#_Toc103782740)

[Hình 20: Giao diện đăng ký tài khoản 35](#_Toc103782741)

[Hình 21: Giao diện đăng nhập 35](#_Toc103782742)

[Hình 22: Giao diện thông tin người dùng 36](#_Toc103782743)

[Hình 23: Giao diện cập nhật thông tin người dùng 37](#_Toc103782744)

[Hình 24: Giao diện chi tiết công thức nấu ăn 38](#_Toc103782745)

[Hình 25: Giao diện đánh giá công thức nấu ăn 38](#_Toc103782746)

[Hình 26: Giao diện các công thức nấu ăn đã được lưu 39](#_Toc103782747)

[Hình 27: Giao diện tạo công thức nấu ăn 39](#_Toc103782748)

[Hình 28: Giao diện thay đổi theme 40](#_Toc103782749)

# DANH MỤC BẢNG

[Bảng 1: Ví dụ các thuộc tính của items 16](#_Toc104222538)

[Bảng 2: Profile người dùng 17](#_Toc104222539)

[Bảng 3: Các items của người dùng 17](#_Toc104222540)

[Bảng 4: Các items sau khi nhân với profile người dùng 18](#_Toc104222541)

[Bảng 5: Tổng điểm theo thành phần dinh dưỡng của người dùng 19](#_Toc104222542)

[Bảng 6: Scale các điểm thành phần về 1 19](#_Toc104222543)

[Bảng 7: Các items mới có thể gợi ý cho người dùng 19](#_Toc104222544)

[Bảng 8: Các items sau khi nhân với điểm thành phần dinh dưỡng người dùng 20](#_Toc104222545)

[Bảng 9: Điểm các items khi dự đoán 20](#_Toc104222546)

[Bảng 10: Cấu trúc dữ liệu công thức nấu ăn 21](#_Toc104222547)

[Bảng 11: Cơ sở dữ liệu công thức nấu ăn 23](#_Toc104222548)

[Bảng 12: Cơ sở dữ liệu người dùng 24](#_Toc104222549)

[Bảng 13: Cơ sở dữ liệu đánh giá 24](#_Toc104222550)

[Bảng 14: Cơ sở dữ liệu trả lời đánh giá 25](#_Toc104222551)

# TÓM TẮT

Ngày nay

# PHẦN GIỚI THIỆU

### 1. Đặt vấn đề

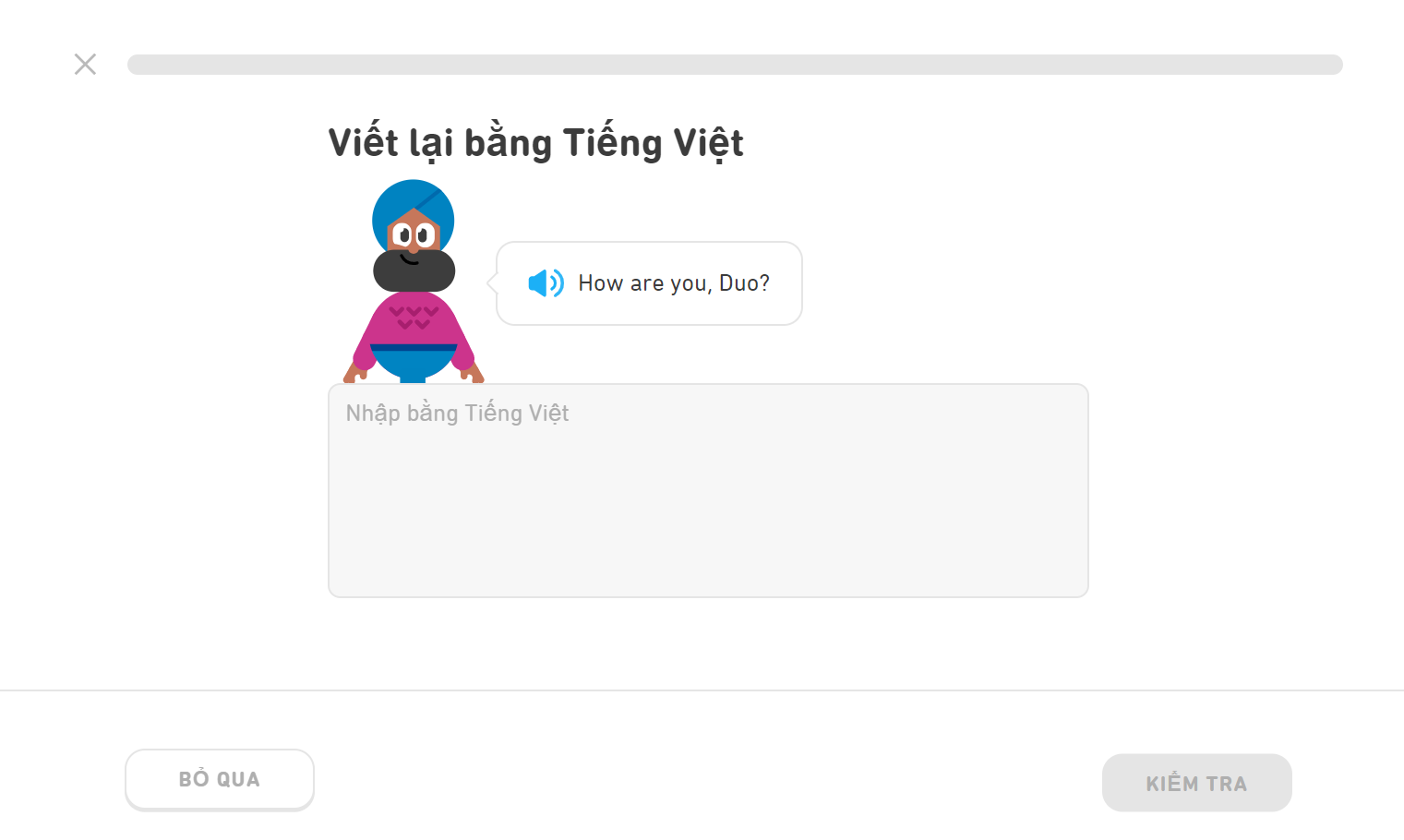
Ngày nay, với sự phát triển mạnh mẽ của Công Nghệ Thông Tin chúng ta có thể học tập, trao đổi và tìm kiếm thông tin một cách nhanh chóng thông qua Internet. Ngày càng có nhiều ứng dụng, website được xây dựng để hỗ trợ cho cuộc sống hằng ngày của chúng ta dễ dàng hơn, các ứng dụng, website có mặt hầu hết trên tất cả các lĩnh vực trong cuộc sống. Có thể kể đến như shoppe, tiki, lazada hỗ trợ việc mua sắm trở nên dễ dàng và thuận lợi hơn. Ngoài ra có website kahoot, quizizz hỗ trợ người dùng có thể tạo các câu hỏi trắc nghiệm và tạo ra các bài thì hoặc trò chơi đó vui một cách dễ dàng và nhanh chóng.

Các hệ thống câu hỏi trắc nghiệm trực tuyến, giảng dạy online đang được áp dụng rộng rãi và phổ biến hơn bao giờ hết. Các hệ thống trắc nghiệm online hỗ trợ rất tốt cho cả giáo viên và các bạn học sinh, sinh viên trong quá trình học tập và kiểm tra năng lực của học sinh, sinh viên, hệ thống trắc nghiệm giúp người ra đề tiết kiệm thời gian ra đề, cũng như thời gian trông coi thi, các câu hỏi thiết kế đa dạng và phong phú và các câu hỏi được hệ thống trộn ngẫu nhiên sẽ hạn chế được tình trạng gian lận trung lúc làm bài, nâng cao chất lượng của bài thi một cách đáng kể. Ngoài ra còn có nhiều ứng dụng, trang web sử dụng hình thức trắc nghiệm để giúp cho việc học tập một cách hiệu quả hơn, các hệ thống có tích hợp hệ thống gợi ý, giúp cho người học có thể học tốt hơn, ghi nhớ tốt hơn.

Nhờ có những hệ thống như trên giúp cho người học dễ dàng học tập và ôn luyện những kiến thức mà mình muốn tiếp thu. Hình thức trắc nghiệm giúp cho người học cảm thấy thích thú trong việc tiếp thu kiến thức, sau khi học những kiến thức bằng nhiều dòng văn bản, người học có thể sử dụng hệ thống trắc nghiệm để ôn luyện lại những kiến thức mà mình đã học, sau khi hoàn thành bài kiểm tra thì có thể biết được kết quả và những lỗi sai. Một số hệ thống còn giúp người học ghi nhớ những lỗi sai và chỉnh sủa giúp họ bằng cách đưa ra các câu hỏi tương tự để học có thể thực hiện lại và ghi nhớ kỹ hơn về câu hỏi đó.

### 2. Lịch sử giải quyết vấn đề

Hệ thống gợi ý đã và đang được sử dụng rộng rãi trong nhiều nền tảng khác nhau như **Duolingo** là trang web cũng như ứng dụng học Tiếng Anh phổ biến trên thế giới, hệ thống có tích hợp hệ thống gợi ý khi chúng ta trả lời sai, hệ thống sẽ đưa ra câu hỏi tương tự với câu hỏi đã trả lời sai.



Hình 2: Gợi ý từ website Duolingo.

### 3. Mục tiêu đề tài

Xây dựng một hệ thống gợi ý các câu hỏi tương tự khi người tham gia trả lời không chính xác và có thể ứng dụng vào các trang web có hệ thống câu hỏi trắc nghiệm.

Tạo ra được một website có giao diện bắt mắt, dễ sử dụng và ứng dụng được hệ thống gợi ý vào trong website.

Tìm hiểu được quy trình để phát triển một hệ thống, ứng dụng. Hiểu được các kiến thức về xây dụng một hệ thống gợi ý, đặc biệt củng cố và ứng dụng được phương pháp gợi ý dựa trên lọc theo nội dung.

Nguyên cứu cách thức hoạt động, mô hình, phương pháp và tầm quan trọng của hệ thống gợi ý.

### 4. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu của đề tài là nghiên cứu về phương pháp gợi ý dựa trên nội dung bằng cách đưa các câu về các vector và sau đó tính độ tương tự của các câu với nhau. Sau khi có độ tương tự thì dựa vào ma trận xây dựng đưa ra danh sách các câu hỏi có độ tương tự cao nhất và thấp nhất.

Phạm vi nghiên cứu của đề tài là xử lý văn bản văn bản tiếng Việt, vector hóa văn bản và tính độ tương tự của các câu hỏi.

### 5. Phương pháp nghiên cứu

Trong đề tài này chúng ta tiếp cận bằng phương pháp gợi ý dựa trên nội dung, sử dụng các phương pháp vector hóa văn bản và các phương pháp tính độ tương tự của văn bản. Về phần thực hiện thì sử dụng ngôn ngữ lập trình Python, các thư viện tách từ, vector hóa văn bản và tính độ tương tự.

### 6. Kết quả đạt được

Xây dựng được trang web tạo và thực hiện kiểm tra các câu hỏi trắc nghiệm.

Xây dựng được mô hình và API gợi ý câu hỏi tương tự bằng phương pháp lọc theo nội dung. Học được cách phân công, lên kế hoạch và giải quyết vấn đề.

### 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, cài đặt giải thuật, biễu diễn cơ sở dữ liệu, trình bày các bước xây dựng hệ thống bằng phương pháp lọc theo nội dung.

**Chương 3** : Kiểm thử hệ thống, đánh giá độ chính xác và demo trang web.

**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. Mô tả chi tiết bài toán

Ngày nay, việc ứng dụng Công Nghệ Thông Tin vào trong việc học tập và làm việc không còn xa lạ với chúng ta nữa, trên Internet có rất nhiều website ứng dụng các hệ thống thi và ôn tập bằng hình thức trắc nghiệm. Các hệ thống thi trắc nghiệm hỗ trợ cho người học có thể không cần phải đến tận nơi để làm bài thi hoặc bài kiểm tra mà có thể trực tiếp thực hiện bài thi tại bất cứ nơi đâu, chỉ cần có các thiết bị thông mình có thể kết nối Internet là được. Trong khoảng thời gian do dịch bệnh covid nhờ có các hệ thống giảng dạy online giúp kết nối giảng viên trong quá trình giảng dạy và thi cử trở nên dễ dạng hơn, cho thấy được hiệu quả của hệ thống giảng dạy online. Không chỉ trong thời gian dịch bệnh, các hệ thống thi và ôn luyện online vẫn là một lựa chọn hàng đầu trong các hình thức thi cử.

### 2. Vấn đề và giải pháp liên quan đến bài toán

Để xây dựng một hệ thống gợi ý thì có nhiều phương pháp như gợi ý dựa theo nội dung, gợi ý dựa trên lọc theo nội dung và kết hợp cả 2 phương pháp trên. Trong đề tài này chúng tôi sẽ sử dụng phương pháp gợi ý dựa trên lọc theo nội dung.

Phương pháp lọc nội dung được triển khai rộng rãi nhất và thành công nhất trong thực tế. Phương pháp này phân tích và đánh giá dựa trên ngượi dùng hoặc mục dữ liệu sau đó tìm ra sự tương đồng thông qua các đánh giá và tạo ra các gợi ý cho người dùng nào đó.

#### 2.1**. Công cụ thiết kế website**

##### 2.1.1. HTML và CSS

**HTML (viết tắt của từ Hypertext Markup Language, hay là “Ngôn ngữ đánh dấu siêu văn bản”)** là một ngôn ngữ đánh dấu tiêu chuẩn cho các tài liệu được thiết kế hiển thị trên trình duyệt web. Nó có thể kết hợp với CSS và JavaScript, HTML là một trongnhững ngôn ngữ quan trọng trong lĩnh vực thiết kế website. HTML có thể nhận từ server hoặc từ local storage và sau đó thực hiện render vào trang web.

Phiên bản chính thức mới nhất của HTML là HTML 4.01 (1999). Sau đó, các nhà phát triển đã thay thế nó bằng XHTML. Hiện nay, phiên bản mới nhất của ngôn ngữ này là HTML5.

Khi làm việc với HTML, chúng ta sẽ sử dụng cấu trúc code đơn giản (tên thẻ và thuộc tính thể) để định nghĩa thẻ và hiển thỉ lên trình duyệt. Ví dụ, chúng ta có thể tạo một đoạn văn và định nghĩa thuộc tính class bằng cách đặt văn bản vào trong cặp tag mở và đóng văn bản <p class=”description”> và </p>.

**CSS** là ngôn ngữ tạo phong cách cho trang web – **Cascading Style Sheet** language. Nó dùng để tạo styles và định kiểu cho những thẻ được viết dưới dạng ngôn ngữ đánh dấu, như là HTML. Nó có thể định kiểu của nhiều trang web cùng lúc để tiết kiệm công sức và tăng khả năng tái sử dụng code cho người viết web. Nó phân biệt cách hiển thị của trang web với nội dung chính của trang bằng cách điều khiển bố cục, màu sắc, và font chữ.

##### 2.1.2. JavasSript

**JavaScript (gọi tắt là JS)** là ngôn ngữ lập trình kích bản phía client cũng như phía server (NodeJS), các mã lệnh được thực thi bởi trình duyệt của người dùng khi cho phép sử dụng JavaScript. Là một công nghệ lõi của **World Wide Web** cùng với HTML và CSS tạo nên một bộ 3 không thể thiếu của nhà phát triển website.

**JavaScript** sử dụng kết hợp với HTML và CSS tạo nên các tính năng tương tác, làm website sinh động hơn, tăng trải nghiệm người dùng. **Javascript** cho phép chúng kiểm soát các hành vi của website tốt hơn so với việc chỉ sử dụng HTML và CSS.

**JavaScript** là ngôn ngữ lập trình đơn giản, dễ học và tìm hiểu cũng như sử dụng nó. Cú pháp của ngôn ngữ có những phương thức đặt tên khá giống với tiếng Anh nên có thể dễ hiểu và sử dụng một cách nhanh chóng.

##### 2.1.3. ReactJS

**ReactJS** là một thư viện JavaScript xây dựng giao diện người dùng có tính hiệu quả và linh hoạt có thể tái sử sử dụng lại thông qua các thành phần gọi là components. ReactJS xây dụng các thành phần với tư tưởng phân chia, từ các thành phần lớn, phức tạp chuyển đổi thành các thành phần nhỏ hơn, dễ quản lý và tái sử dụng hơn. React được tạo ra bởi Jordan Walke, một kỹ sư phần mềm tại Facebook, phát triển và duy trì bởi Facebook và các ứng dụng nổi tiếng như WhatsApp & Instagram sử dụng công nghệ này.

**ReactJS** được dùng để xây dựng các ứng dụng [Single Page Application] (SPA). Một trong những điểm hấp dẫn của ReactJS là nó không chỉ được xây dựng bên phía clients mà còn sử dụng được bên phía server[1].

#### 2.2. Hệ thống gợi ý

**Hệ thống gợi ý** (Recommendation system) là một lớp con của hệ thống lọc thông tin, tìm cách dự đoán “xếp hạng” hoặc “ưa thích” của một người dùng nào đó với một sản phẩm hoặc một đối tượng. Các hệ thống gợi ý thường được sử dụng trong nhiều lĩnh vực khác nhau, đặc biệt là trong các ứng dụng thương mại[4].

Các hệ thống gợi ý được sử dụng ở nhiều lĩnh vực khác nhau, các trường hợp phổ biến như gợi ý các video và bài hát, giới thiệu các sản phẩm cho người dùng mua hàng online trên các trang thương mại điện tử và các trang mạng xã hội. Sau khi người dùng mua một sản phẩm trên trang thương mại điện tử, hệ thống sẽ phân tích các sở thích, các sản phẩm tương tự với sản phẩm mà người dùng đã xem và đưa ra các sản phẩm mà người dùng có thể quan tâm.

Một số hệ thống gợi ý:

* Thương mại điện tử: Azazon, Alibaba, Tiki…
* Giải trí: Youtube, Netflix, Spotify…
* Mạng xã hội: Facebook, Zalo…

Một số minh chứng cụ thể cho thấy hệ thống gợi ý đóng một vai trò quan trọng trong cuộc sống hằng ngày[5]:

Hệ thống gợi ý đóng vai trò rất quan trọng trong các trang web nổi tiếng như Youtube, Amazon, Netflix… Bên cạnh đó, nhiều công ty truyền thông đang phát triển và triển khai các hệ thống gợi ý như một phần dịch vụ cung cấp cho các người dùng. Một trong số các hoạt động tiêu biểu đó là Netflix, nhà chung cấp dịch vụ xem các phim và chương trình trực tuyến theo hình thức đăng ký, đã trao giải thưởng 01 triệu đô la cho nhóm làm tăng độ chính xác lên 10.05%.

Có nhiều hội nghị, hội thảo chuyên về lĩnh vực hệ thống gợi ý, một trong số đó có thể để cập đến ACM Recommeder System(RecSyss), thành lập vào năm 2007 là sự kiện đáng chú ý hằng năm trong lĩnh vự nghiên cứu hệ thống gợi ý. Thêm đó các buổi họp dành riêng cho vấn đề hệ thống gợi ý được tổ chức trong các hội nghị cơ sở dữ liệu, hệ thống thông tin.

Hiện nay trong các trường đại học và sau đại học, các khóa học, học phần hệ thống gợi ý đang trở nên phổ biến, đặt biệt là trong chuyên ngành Khoa học máy tính. Số lượng tài liệu và cộng đồng về hệ thống gợi ý ngày càng tăng.

Các tạp chí khoa học tiến hành xuất bản các số đặt biệt đề cập về nghiên cứu và phát triển trong lính vực hệ thống gợi ý. Các bài báo tiêu biểu như AI Communication (2008), IEEE Intelligent System (2007), Interbational Jourrnal of Electronic Commerece (2006) và ACM Transactions of Inrfomation Systems (2004).

##### 2.2.1. Quy trình xây dựng hệ thống gợi ý

Để xây dựng một hệ thống gợi ý chúng ta cần phải thực hiện các bước như thu thập dữ liệu, chuẩn hóa dữ liệu, huấn luyện mô hình và đưa ra gợi ý, cuối cùng là đánh giá mô hình.

Evaluate

Data

Standardized Data

Predict

Train Model

Hình 4: Quy trình xây dụng hệ thống gợi ý

Thu thập dữ liệu: lấy ví dụ đơn giản chúng ta chỉ quan tâm đến ratings của người dùng thì dữ liệu đã có sẵn trong cơ sở dữ liệu. Tuy nhiên không phải lúc nào chúng ta cũng là tường minh, tùy vào bài toán mà chúng ta cần phải có chiến thuật thu thập dữ liệu. Đây là một bước có vai trò quan trọng, nếu thu thập thông tin không đầy đủ hoặc không chính xác, hệ thống sẽ đưa ra các gợi ý thiếu chính xác.

Chuẩn hóa dữ liệu: để có thể thực hiện quá trình huấn luyện cũng như dự đoán chúng ta phải chuẩn hóa dữ liệu về dạng ma trận dữ liệu, tuy nhiên ma trận dữ liệu còn rất nhiều giá trị rỗng, ta cần đưa chúng về các số 0 thì mới áp dụng được các giải thuật máy học.

Huấn luyện mô hình và dự đoán: dữ liệu trải qua quá trình xử lý và chọn được mô hình phù hợp chúng ta sẽ tiến hành huấn luyện. Sau khi huấn luyện ta có được mô hình, sử dụng mô hình đó thực hiện dữ đoán, lựa chọn ra top N item thỏa mãn điều kiện để gợi ý cho người dùng.

Đánh giá mô hình: giống như các mô hình máy học khác, chúng ta cần phải đánh giá để biết được mức độ chính xác, đô hiệu quả của mô hình. Để giúp người lập trình có thể cài thiện hoặc thay đổi mô hình khác phù hợp hơn.

##### 2.2.2. Phương pháp xây dựng hệ thống gợi dựa trên nội dung

Lọc cộng tác (collaborative filtering) là phương pháp phân tích dữ liệu người dùng để tìm ra điểm tương tự giữa các người dùng có cùng sở thích, thói quen. Phương pháp này hoạt động bằng các xây dựng và phân tích dựa trên ma trận giữa người dùng(user) và mục dữ liệu(items)[7].

Hạn chế của phương pháp lọc cộng tác:

* Dữ liệu thưa.
* Sản phẩm mới.
* Gợi ý thiếu chính xác vì không quan tâm đến dặc điểm của sản phẩm mà người dùng đã chọn.
* Khả năng giải thích cho kết quả gợi ý kém.
* Thiếu minh bạch.

Hệ thống gợi ý dựa trên nội dung với ý tưởng là gợi ý các sản phẩm cho người dùng X tương tự với các **sản phẩm được đánh giá cao** cũng bởi những **người dùng X này**.

Ví dụ như gợi ý phim dựa trên các thuộc tính như thể loại, đạo diễn, diễn viên.

Tìm kiếm các mục dữ liệu tương tự với mục dữ liệu mà **người dùng** đã mua / chọn lựa / xem / đánh giá trước đó **dựa trên thuộc tính** của các mục dữ liệu (ví dụ như màu sắc, giá cả, mô tả mục dữ liệu..) hoặc dựa trên **người dùng.**

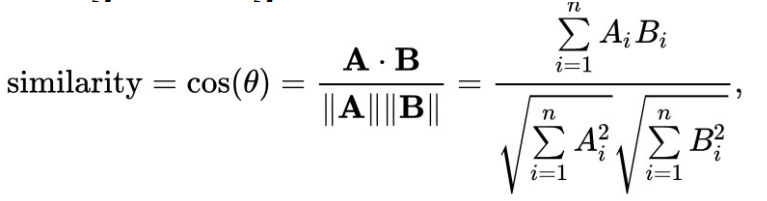
Bảng 1: Ví dụ các thuộc tính của items

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Cabohydrates** |
| item1 | 2.0 | 0.0 | 0.0 | 20.0 | 0.0 |
| item2 | 2.0 | 2.0 | 2.0 | 17.0 | 3.0 |
| itme3 | 4.0 | 0.0 | 0.0 | 30.0 | 0.0 |

Các bước chính:

Biễu diễn mỗi item dưới dạng **một vector thuộc tính.**

Gợi ý các item **tương tự** các item trong quá khứ **của người dùng chính.**

****

Hoăc xây dựng **profile** người dùng theo các thuộc tính **item** và gợi ý **item** có thuôc tính phù hợp với **profile** người.

Dữ liệu xây dựng gợi ý dựa trên item description. Chia thành hai loại:

* Dữ liệu có cấu trúc:
  + Trang phục: loại vải, nhãn hiệu, giá cả…
  + Thức ăn: loại, thương hiệu, dinh dưỡng…
* Dữ liệu phi cấu trúc:
  + Nội dung bài báo.
  + Nội dung email.

Bảng 2: Profile người dùng

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Item 1 | Item 2 | Item 3 | Item 4 |
| User | 5.0 | 2.0 | 4.0 | 1.0 |

Bảng 3: Các items của người dùng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Carbohydrates** |
| Item 1 | 2.0 | 0.0 | 0.0 | 20.0 | 0.0 |
| Item 2 | 2.0 | 2.0 | 2.0 | 17.0 | 3.0 |
| Item 3 | 4.0 | 0.0 | 0.0 | 39.0 | 0.0 |
| Item 4 | 6.0 | 0.0 | 70.0 | 0.0 | 0.0 |

Ta lấy từng item của profile người dùng nhân với các thành phần dinh dưỡng của các item.

Total Fatitem1 = 5.0 \* 2.0 = 10.0

Sugaritem1 = 5.0 \* 0.0 = 0.0

Sodiumitem1 = 5.0 \* 0.0 = 0.0

Saturated Fatitem1 = 5.0 \* 20.0 = 100.0

Carbohydratesitem1 = 5.0 \* 0.0 = 0.0

Bảng 4: Các items sau khi nhân với profile người dùng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Carbohydrates** |
| Item 1 | 10.0 | 0.0 | 0.0 | 100.0 | 0.0 |
| Item 2 | 4.0 | 4.0 | 4.0 | 34.0 | 6.0 |
| Item 3 | 16.0 | 0.0 | 0.0 | 126.0 | 0.0 |
| Item 4 | 6.0 | 0.0 | 70.0 | 0.0 | 0.0 |

Ta tính tổng các items theo thành phần dinh dưỡng.

Total fat = Total Fatitem1 + Total Fatitem2 + Total Fatitem3 + Total Fatitem4

= 10.0 + 4.0 + 16.0 + 6.0

= 36.0

Bảng 5: Tổng điểm theo thành phần dinh dưỡng của người dùng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Carbohydrates** |
| User | 36.0 | 4.0 | 74.0 | 260.0 | 6.0 |

Bảng 6: Scale các điểm thành phần về 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Carbohydrates** |
| User | 0.09 | 0.02 | 0.19 | 0.68 | 0.02 |

Bảng 7: Các items mới có thể gợi ý cho người dùng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Carbohydrates** |
| Item 5 | 0.0 | 5.0 | 3.0 | 2.0 | 6.0 |
| Item 6 | 1.0 | 0.0 | 58.0 | 2.0 | 0.0 |
| Item 7 | 1.0 | 0.0 | 0.0 | 8.0 | 0.0 |
| Item 8 | 3.0 | 24.0 | 19.0 | 17.0 | 11.0 |

Các items mới nhân với các điểm thành phần dinh dưởng đã được scale.

Total Fatitem5 = 0.09 \* 0.0 = 0.0

Sugaritem5 = 0.02 \* 5.0 = 0.1

Sodiumitem5 = 0.19 \* 3.0 = 0.57

Saturated Fatitem5 = 0.68 \* 2.0 = 1.36

Carbohydratesitem5 = 0.02 \* 6.0 = 0.12

Bảng 8: Các items sau khi nhân với điểm thành phần dinh dưỡng người dùng

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Total Fat** | **Sugar** | **Sodium** | **Saturated Fat** | **Carbohydrates** |
| Item 5 | 0.0 | 0.1 | 0.57 | 1.36 | 0.12 |
| Item 6 | 0.09 | 0.0 | 11.02 | 1.36 | 0.0 |
| Item 7 | 0.09 | 0.0 | 0.0 | 5.44 | 0.0 |
| Item 8 | 0.27 | 0.48 | 3.61 | 11.56 | 0.22 |

Bảng 9: Điểm các items khi dự đoán

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Item 5 | Item 6 | Item 7 | Item 8 |
| Pred | 2.15 | 12.47 | 5.33 | 16.14 |

Ta có thể chọn **item 8** để gợi ý cho người dùng vì có số điểm cao nhất là **16.14**.

##### 2.2.3. Phương pháp đánh giá

**Given-N**: Thường được sử dụng để đánh giá trong các lĩnh vực thương mại điện tử. Là một mở rộng của phương pháp k-fold thực hiện trên từng giao dịch thay vì toàn bộ giao dịch. Giao dịch phải có it nhất **N + 1** dữ liệu thì mới có thể thực hiện phương pháp này.

**All-but-one**: Là trường hợp đặc biệt của Given-N, phần <given> là **N-1**, phần test luôn là 1.

## CHƯƠNG 2 - THIẾT KẾ VÀ CÀI ĐẶT

### 1. Thiết kế hệ thống

#### 1.1 Sơ đồ tổng quát

Ban

#### 1.2 Giới thiệu tập dữ liệu

Tập

#### 1.3 Cấu trúc của website

.

### 2. Cài đặt giải thuật

#### 2.1 Sơ lược cơ sở dữ liệu

#### 2.2 Giải thuật chính trong hệ thống

##### 2.2.1 Giải thuật đăng ký tài khoản

#### 2.3. Xây dựng hệ thống gợi ý

## CHƯƠNG 3 – GIỚI THIỆU WEBSITE

### Giao diện chính hệ thống

### 2. Đánh giá

#### 2.1 Given-N

Có tất cả 30 người dùng với 135 công thức nấu ăn được tạo, sử dụng dữ liệu người dùng để thực hiện việc đánh giá. Với mỗi người dùng sử dụng phần **given** là các công thức nấu ăn đầu với số lượng phần **N** của người dùng đã tạo, phần test là các công thức nấu ăn còn lại. Đem phần **given** để dự đoán thành một danh sách các công thức nấu ăn, nếu công thức nấu ăn dự đoán có trong danh sách nấu ăn của người dùng đã tạo thì +1, còn không tồn có thì là 0.

Chia danh sách các công thức nấu ăn thành 2 phần: phần **given** và phần **test**. So sánh các gợi ý với dữ liệu trong tập test để đánh giá.

Trong đề tài này chúng tôi sử dụng **Given-1** và **Given-2** để đánh giá trên **30** người dùng. Sử dụng 1 công thưc (**Given-1**) và 2 công thức (**Given-2**) đầu tiên để dự đoán, nếu trong danh sách dự đoán có công thức nấu ăn của người dùng đã tạo thì là dự đoán đúng.

Với Given-1 ta có 6 dự đoán đúng: 6 / 30 = 0.2 = **20%**.

Với Given-2 ta có 7 dự đoán đúng: 7 / 30 = 0.23 = **23%**.

#### 2.2 All-but-one

Ưu điểm của phương pháp này là cho phép đánh giá các giỏ hàng có kích thước lớn hơn 1 sản phẩm.

Với phương pháp này ta đạt được 12 dự đoán đúng: 12 / 32 = 0.4 = **40%**.

# PHẦN KẾT LUẬN

### 1. Kết quả đạt được

### 2. Hướng phát triển

# TÀI LIỆU THAM KHẢO

[1] **Freetuts**, <https://freetuts.net/>

**[2] Viblo,** Giới thiệu về Hệ thống gợi ý – Recommendation Systems, <https://viblo.asia/p/gioi-thieu-ve-he-thong-goi-y-recommendation-systems-aWj53kEp56m>.

**[3] Lighthouselabs,** How Netflix Utilizes Data Science - Rachel Meltzer, [https://www.lighthouselabs.ca/en/blog/how-netflix-uses-data-to-optimize-their-product#:~:text=The%20Netflix%20Recommendation%20Engine,clusters%20based%20on%20user%20preferences](https://www.lighthouselabs.ca/en/blog/how-netflix-uses-data-to-optimize-their-product%23:~:text=The%20Netflix%20Recommendation%20Engine,clusters%20based%20on%20user%20preferences).

[4] **Wikipedia,** Hệ thống gợi ý, <https://en.wikipedia.org/wiki/Recommender_system>

[5] **Võ Thị Xuân Mai** – Luận văn tốt nghiệp Đại Học Cần Thơ.

[6] **Nguyễn Hùng Dũng, Nguyễn Thái Nghe,** Hệ thống gợi ý sản phẩm trong bán hàng trực tuyến sử dụng kỹ thuật lọc công tác.

[7] **Trần Nguyễn Minh Thư**, Hệ thống gợi ý dựa trên nội dung – slice.