ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG

———————*———————

ĐỒ ÁN
TỐT NGHIỆP ĐẠI HỌC

NGÀNH CÔNG NGHỆ THÔNG TIN

XÂY DỰNG HỆ THỐNG TƯ VẤN TÀI LIỆU HỌC TIẾNG ANH E-LEARNING

Sinh viên thực hiện: **Hoàng Đức Việt**
Lớp AS - K57

Giáo viên hướng dẫn: TS.
Phạm Văn Hải

HÀ NỘI 06-2017

PHIẾU GIAO NHIỆM VỤ ĐỒ ÁN TỐT NGHIỆP

1. Thông tin về sinh viên

Họ và tên sinh viên: Hoàng Đức Việt

Điện thoại liên lạc: 0902239958 Email: cooro1994@gmail.com

Lớp: AS Hệ đào tạo: Chính quy

Đồ án tốt nghiệp được thực hiện tại: Đại học Bách Khoa Hà Nội

Thời gian làm ĐATN: Từ ngày / / đến / /

2. Mục đích nội dung của ĐATN

- Nghiên cứu tìm hiểu về hệ thống tư vấn và các kỹ thuật cơ bản.
- Nghiên cứu các đặc tính của Kansei ảnh hưởng đến tâm lý và hành vi lựa chọn của con người.
- Đề xuất xây dựng một hệ thống tư vấn tài liệu học tiếng Anh E-learning dành cho người dùng điện thoại Android.

3. Các nhiệm vụ cụ thể của ĐATN

- Tìm hiểu về hệ thống tư vấn, hệ thống gợi ý và các kỹ thuật cơ bản.
- Tìm hiểu về giải thuật Context-matching và ứng dụng trong việc tư vấn dựa trên nguyện vọng và năng lực của người dùng.
- Tìm hiểu về Kansei Engineering và ứng dụng của nó trong việc cá nhân hoá kết quả tư vấn.
- Phân tích, thiết kế và cài đặt hệ thống tư vấn tài liệu học tiếng Anh E-learning dành cho người dùng điện thoại Android.

4. Lời cam đoan của sinh viên:

Tôi - Hoàng Đức Việt - cam kết ĐATN là công trình nghiên cứu của bản thân tôi dưới sự hướng dẫn của TS. Phạm Văn Hải.

Các kết quả nêu trong ĐATN là trung thực, không phải là sao chép toàn văn của bất kỳ công trình nào khác.

Hà Nội, ngày tháng năm

Tác giả ĐATN

Hoàng Đức Việt

5. Xác nhận của giáo viên hướng dẫn về mức độ hoàn thành của ĐATN và cho phép bảo vệ

Hà Nội, ngày tháng năm

Giáo viên hướng dẫn

TS. Phạm Văn Hải

LỜI CÁM ƠN

TÓM TẮT NỘI DUNG ĐỒ ÁN TỐT NGHIỆP

Sự phát triển mạnh mẽ của khoa học công nghệ đi kèm với sự phát triển của Internet và điện thoại thông minh. Với lợi thế nhỏ gọn, tiện dụng và thông minh, smartphone đã trở thành một phần không thể tách biệt trong cuộc sống, hỗ trợ rất nhiều cho con người trong các hoạt động hằng ngày. Một trong những công dụng hữu ích phải kể đến đó là việc học tập trên điện thoại. Tuy nhiên, với lượng thông tin ngày càng nhiều và đa dạng như hiện nay, việc tìm kiếm cho được khoá học hay tài liệu để học phù hợp với mục đích và trình độ của bản thân mình là một vấn đề nan giải chưa có hướng giải quyết.

Đồ án sẽ đề xuất hệ thống tư vấn tài liệu học tiếng Anh cho người dùng trên điện thoại di động giúp cho người dùng có thể lựa chọn được tài liệu học phù tiếng Anh phù hợp với bản thân mình. Hệ thống được xây dựng dựa vào việc thu thập thông tin về trình độ và nguyện vọng của người dùng, đồng thời dựa vào đánh giá của người dùng về những kết quả tư vấn. Qua đó xây dựng được hồ sơ người dùng và tư vấn ra những kết quả phù hợp với họ.

Nội dung đồ án sẽ được trình bày dưới các phần sau:

- **Chương 1:** Đặt vấn đề và định hướng giải pháp
- **Chương 2:** Cơ sở lý thuyết
- **Chương 3:** Phân tích thiết kế xây dựng ứng dụng
- **Chương 4:** Kết quả thực hiện
- **Chương 5:** Kết luận và hướng phát triển.

Mục lục

Danh sách hình vẽ

Danh sách bảng

Chương 1

Đặt vấn đề và định hướng giải pháp

1.1 Đặt vấn đề

Sự phát triển mạnh mẽ của khoa học công nghệ đi kèm với sự phát triển của Internet và điện thoại thông minh. Với lợi thế nhỏ gọn, tiện dụng và thông minh, smartphone đã trở thành một phần không thể tách biệt trong cuộc sống, hỗ trợ rất nhiều cho con người trong các hoạt động hằng ngày. Người dùng smartphone ngoài chức năng liên lạc, họ còn sử dụng nhiều loại dịch vụ khác nhau như giải trí, định vị, mua sắm, thanh toán trực tuyến,... Một trong những công dụng hữu ích phải kể đến đó là việc học tập trên điện thoại.

Nhiều chuyên gia nhận định, cùng với sự phát triển của Internet, Giáo dục trực tuyến (E-Learning) đang dần trở nên phổ biến nhờ tính tiện dụng, tương tác cao và nhu cầu rất lớn từ cộng đồng học tập. Với chiếc smartphone trên tay, chỉ bằng một vào thao tác tìm kiếm đơn giản, người dùng đã có thể truy cập đến vô vàn các khoá học, sách tham khảo, bài giảng online khác nhau. Tuy nhiên, với lượng thông tin ngày càng nhiều và đa dạng như hiện nay, không phải khoá học, tài liệu nào cũng phù hợp với mục đích và trình độ học của người dùng. Việc theo học một khoá học không phù hợp sẽ dẫn đến việc người học mất dần hứng thú, động lực học, gây ra tốn kém thời gian mà hiệu quả thu được là không cao.

Để giải quyết bài toán tìm kiếm và lựa chọn thông tin cần thiết phù hợp với nhu cầu người dùng, các hệ thống thông tin thường tích hợp một hệ lọc để đưa ra chỉ những thông tin mà người dùng có thể quan tâm. Hệ thống này được gọi là hệ thông tư vấn, hay hệ gợi ý (Recommender System). Hệ thống tư vấn dựa trên các thông tin thu thập được từ người dùng, phân tích xử lý và đối chiếu với cơ sở tri thức, từ đó đưa ra được những thông tin hữu dụng giúp cho người dùng đạt được mục đích của mình. Hiện nay trên thế giới đã có rất nhiều hệ thống tư vấn được tích hợp ứng dụng trong nhiều lĩnh vực khác nhau như thương mai điện tử, phim ảnh, âm nhạc, sách,... Tuy nhiên rất ít trong số đó dùng cho mục đích tư vấn tài liệu học, và hầu hết các hệ thống giáo dục trực tuyến không được tích hợp chức năng tư vấn.

Để giải quyết vấn đề trên, đồ án này đề xuất một hệ thống tư vấn tài liệu học tiếng Anh cho người dùng trên điện thoại di động giúp cho người dùng có thể lựa chọn được tài liệu học phù tiếng Anh phù hợp với bản thân mình. Hệ thống được xây dựng dựa vào việc thu thập thông tin về trình độ và nguyện vọng của người dùng, đồng thời dựa vào đánh giá của người dùng về những kết quả tư vấn. Qua đó xây dựng được hồ sơ người dùng và tư vấn ra những kết quả phù hợp với họ.

1.2 Định hướng giải quyết

Hệ thống tư vấn sẽ được xây dựng theo mô hình client-server. Trong đó client sẽ là ứng dụng điện thoại di động được viết trên nền tảng Android đóng vai trò một giao diện tương tác, thu thập và xử lý thông tin người dùng và trả về kết quả tư vấn. Server lưu trữ trên nền tảng Firebase của Google sẽ đóng vai trò là hệ cơ sở tri thức, lưu giữ thông tin hồ sơ người dùng, tài liệu tiếng Anh và mạng lưới từ khoá (keyword network) sử dụng thực hiện tư vấn.

Trên client, người dùng sẽ được yêu cầu cung cấp về thông tin trình độ cũng như mong muốn học của họ, cụ thể là :

<trình độ>

- Thời điểm bắt đầu học
- Trình độ đọc hiểu
- Vốn từ vựng
- Vốn ngữ pháp

<nguyện vọng>

- Mục đích học của người dùng

Từ tập câu hỏi trên và xây dựng thành profile người dùng. Dựa trên profile này, hệ thống tiến hành context - match giữa hồ sơ người dùng và thuộc tính của từng tài liệu và chấm điểm độ phù hợp, sau đó hiện kết quả cho người dùng về khoá học, tài liệu tương ứng với trình độ và nguyện vọng của họ. Người dùng sau đó sẽ tiến hành đánh giá xem các kết quả tư vấn trả về có phù hợp với họ hay không. Hồ sơ người dùng sẽ có sự thay đổi dựa trên các đánh giá đó. Cứ tiếp tục như vậy, các kết quả tư vấn tiếp theo sẽ càng ngày chính xác đúng với nhu cầu cá nhân của người dùng hơn.

Chương 2

Cơ sở lý thuyết

Trong nội dung của chương này, đồ án sẽ trình bày về các kiến thức cơ bản cũng như như các thuật toán được sử dụng trong hệ thống tư vấn.

2.1 Khảo sát các hệ tư vấn đã có

Hệ thống tư vấn đã trở thành một đề tài khá phổ biến trong khoảng thời gian gần đây. Trong lĩnh vực xây dựng các hệ thống tư vấn trong quá khứ, người ta đã làm việc và nghiên cứu khá nhiều và ứng dụng rộng rãi trên các lĩnh vực khác nhau. Hầu hết công việc chủ yếu tập trung phát triển những phương pháp gợi ý những những đối tượng ưa thích đến cho người dùng. Ví dụ như những trang web gợi ý những bộ phim, gợi ý những quyển sách mà người dùng có thể yêu thích. Hệ thống tư vấn thông thường sẽ tiếp cận và giải quyết vấn đề theo 1 trong 2 hướng: Lọc dựa trên nội dung (*content-based filtering*) hoặc lọc cộng tác (*collaborative filtering*). Tuy nhiên trên thực tế, việc áp dụng cả 2 hướng để giải quyết vấn đề cũng thường được cân nhắc trong việc giải quyết bài toán trong thực tế. (*hybrid recommender system*)

2.1.1 Lọc cộng tác

là phương pháp được xây dựng dựa trên lý thuyết : "Những người có cùng hứng thú/mong muốn/sở thích về một vấn đề gì đó trong quá khứ thì có thể họ sẽ cũng có cùng hứng thú/mong muốn/sở thích trong tương lai". Ví dụ: hai người dùng A và B có chung sở thích ăn uống, họ đã mua các đồ ăn giống nhau. Nếu B còn thích thêm cả Cocacola nữa thì rất có thể A cũng thích, nên ta có thể gợi ý cho A mua thêm Cocacola.

Phương pháp này thực hiện việc thu thập và đánh giá một lượng lớn thông tin về hành vi, sở thích của người dùng để tiên đoán sở thích của họ dựa trên sự giống nhau về thông tin giữa các người dùng. Ưu điểm của phương pháp này là nó không phải phụ thuộc vào việc nhận định, đánh giá để “hiểu được” nội dung của đối tượng tư vấn mà vẫn có thể đưa ra được kết quả thoả mãn mong muốn của người dùng. Tuy nhiên mặt

hạn chế của phương pháp trên là nó cần một lượng lớn dữ liệu người dùng đa dạng để có thể hoạt động chính xác được. Và việc tính toán hành vi của từng người dùng cũng tiêu hao một lượng lớn tài nguyên máy tính.

2.1.2 Lọc dựa trên nội dụng

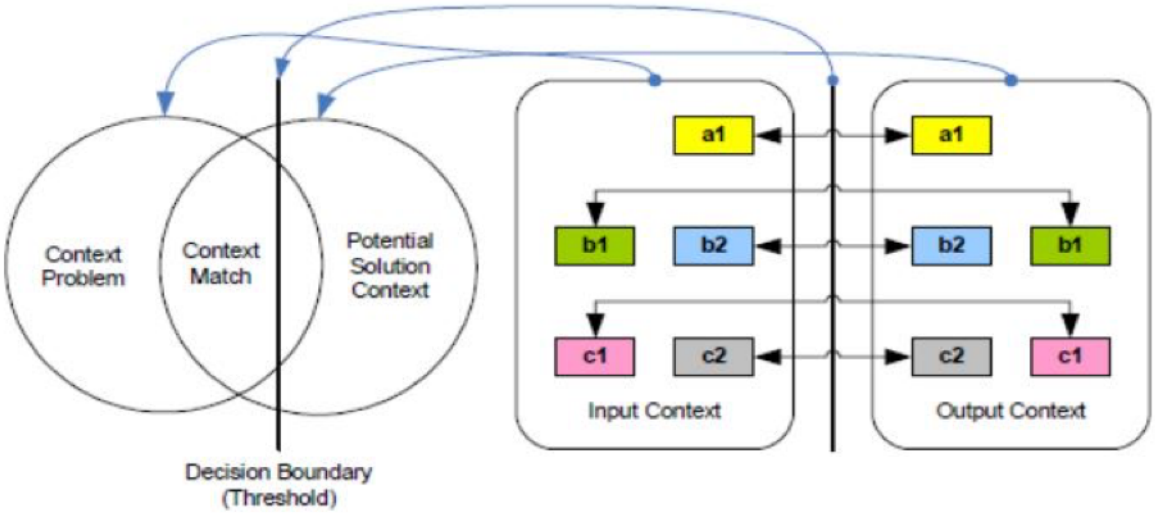
là phương pháp mà người ta quân tâm đến 2 thực thể chính là người dùng và đối tượng được khuyến nghị đến cho người dùng. Quá trình lọc thực hiện bắt đầu từ việc xây dựng một hồ sơ người dùng (user profile) là tập các ưu tiên về sở thích của người dùng về đối tượng, và hồ sơ miêu tả đối tượng. Sau đó tiến hành lọc kết quả dựa vào phương pháp context-matching. Đây là phương pháp bắt nguồn từ lĩnh vực nghiên cứu triết xuất thông tin và lọc thông tin.

Hệ thống tư vấn trong tài liệu này áp dụng phương pháp lọc dựa trên nội dung, và cân nhắc thêm lọc cộng tác khi hệ thống đã phát triển sau này.

2.2 Thuật toán Context-matching

Thuật toán context-matching sử dụng để giải quyết bài toán cần Context-matching với Partial-matching (khi mà khả năng đạt được "perfect match" là thấp). Xây dựng các input context, output context và hàm đánh giá giữa các properties của chúng. Mục tiêu cuối cùng là một giá trị Boolean thể hiện sự phù hợp, hay không phù hợp của thực thể được tính đến.

Mô hình thuật toán được biểu diễn ở hình sau :



Hình 2.1: Mô hình thuật toán Context Matching

2.2.1 Tổng quan

Cấu trúc cơ bản của thuật toán là $ON < event > IF < condition > THEN < action >$. Trong đó $< event >$ có thể là context data cần tính (event khởi động), hoặc là kết quả trả về của một lần so sánh ở phía trên, $< condition >$ là cách so sánh giữa một input context property với một output context property sẽ được báo cáo dưới đây, $< action >$ là kết quả của việc đánh giá trên $< condition >$ và nó có thể là việc tiếp tục so sánh input và output tiếp theo hoặc là giá trị Boolean quyết định độ phù hợp của context (action cuối cùng).

2.2.2 Input context và output context

Input context là nguyên mẫu để so sánh, các kết quả phù hợp là những kết quả phải có độ match với input context lớn nhất định. Output context là giải pháp tiềm năng cho vấn đề cần giải quyết trong bài toán, là một số lượng những thực thể đem so sánh với input context, được trích xuất từ cấu trúc dữ liệu. Từ input và output context, cần xây dựng một bộ các context properties là những thuộc tính bên trong quyết định context. Việc so sánh sẽ là so sánh giữa các context property.

2.2.3 Các giá trị sử dụng

Giả sử áp dụng thuật toán cho bộ context properties $[a1, b1, b2, c1, c2]$. Ta có :

- w: giá trị trong khoảng [0.10...1.00] phản ánh mức độ ưu tiên của thuộc tính
- e: giá trị đánh giá xem thuộc tính ở input và output có match nhau không, e được biểu diễn trong khoảng [0...1]
- $av = (e * w)$: giá trị thực tế đánh giá độ match của thuộc tính, nằm trong khoảng [0.00...1.00]
- $sav = av(a_1) + av(b_1) + av(b_2) + av(c_1) + av(c_2)$: tổng giá trị av
- mpv: giá trị sav cao nhất có thể đạt được. Nó thể hiện trường hợp mà tất cả các context property đều match nhau.
- $rv = \frac{sav}{mpv}$: giá trị trả về sau so sánh. Nó đánh giá tổng quan xem kết quả context match đến mức độ nào, nằm trong khoảng [0.00...1.00].
- t: giá trị cho ngưỡng đạt [0.10...1.00]. Context match phù hợp khi so sánh ta có giá trị $rv > t$.

2.2.4 Các bước thực hiện thuật toán

- Bước 1: Đánh giá context match cho từng context property -> rút ra giá trị e .
- Bước 2: Thiết lập giá trị w đã được định sẵn cho từng context property
- Bước 3: Tính giá trị av của thuộc tính $av = (e * w)$
- Bước 4: Tính tổng các giá trị av của cả quá trình đánh giá $sav = av(a_1) + av(b_1) + av(b_2) + av(c_1) + av(c_2)$
- Bước 5: Tính giá trị sav cao nhất có thể đạt $mpv = w(a_1) + w(b_1) + w(b_2) + w(c_1) + w(c_2)$
- Bước 6: Tính giá trị kết quả trả về $rv = \frac{sav}{mpv}$
- Bước 7: Sử dụng giá trị ngưỡng đã có sẵn để xác định xem output context có match với input context hay không . IF $(rv) >= (t)$ THEN $context - match = true$ [1] or IF $(rv) < (t)$ THEN $context - match = false$ [0]

2.3 Kỹ thuật Kansei Engineering

2.3.1 Tổng quan

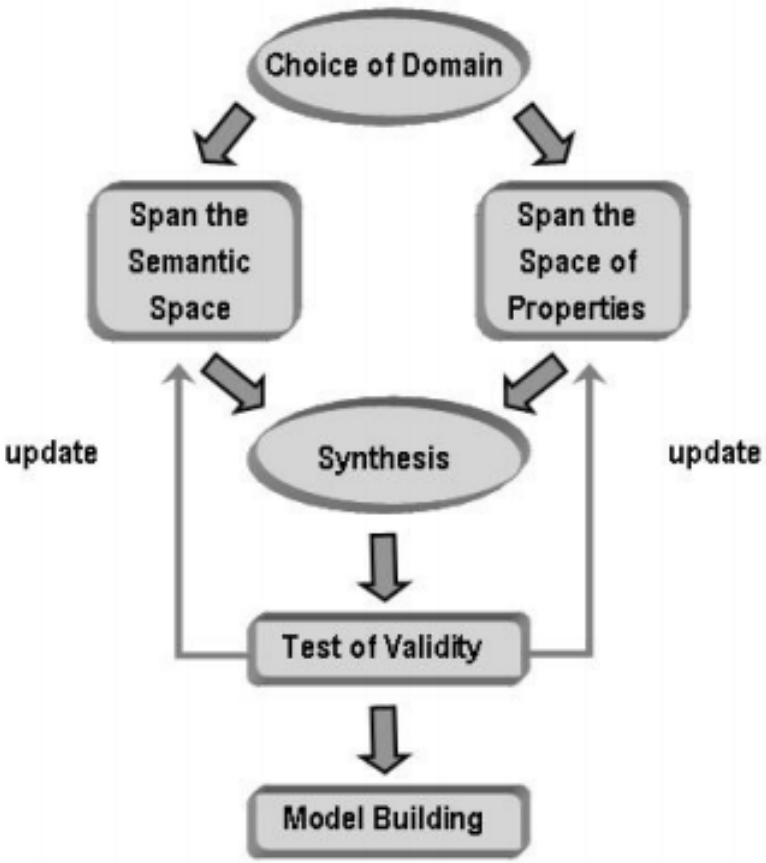
Kansei Engineering là kỹ thuật tích hợp khía cạnh cảm xúc của con người vào trong quá trình xây dựng sản phẩm, nhằm mục đích tạo ra được sản phẩm phù hợp với yêu cầu và mong muốn của người dùng. Nó là việc mang lại sự hài lòng, thoả mãn cho người dùng một cách có khoa học. Để đạt được điều đó, trải nghiệm của người dùng dựa trên các sản phẩm tương tự được thu thập và phân tích, từ đó thiết lập mô hình dự đoán mối quan hệ giữa cảm xúc của con người và các đặc tính vật lý của sản phẩm.

3 điểm chính được chú trọng trong Kansei Engineering đó là :

- Làm thế nào để hiểu được cảm xúc nội tâm của người dùng
- Làm thế nào để phản ánh hiểu biết đó vào trong việc phát triển sản phẩm
- Làm thế nào để xây dựng một hệ thống có tổ chức theo định hướng Kansei Engineering

2.3.2 Mô hình

Mặc dù nhiều mô hình Kansei Engineering khác nhau phục vụ cho các bài toán cụ thể khác nhau. Nhưng về cơ bản, tất cả đều tuân theo mô hình tổng quát [4] sau đây:



Hình 2.2: Mô hình tổng quát Kansei Engineering

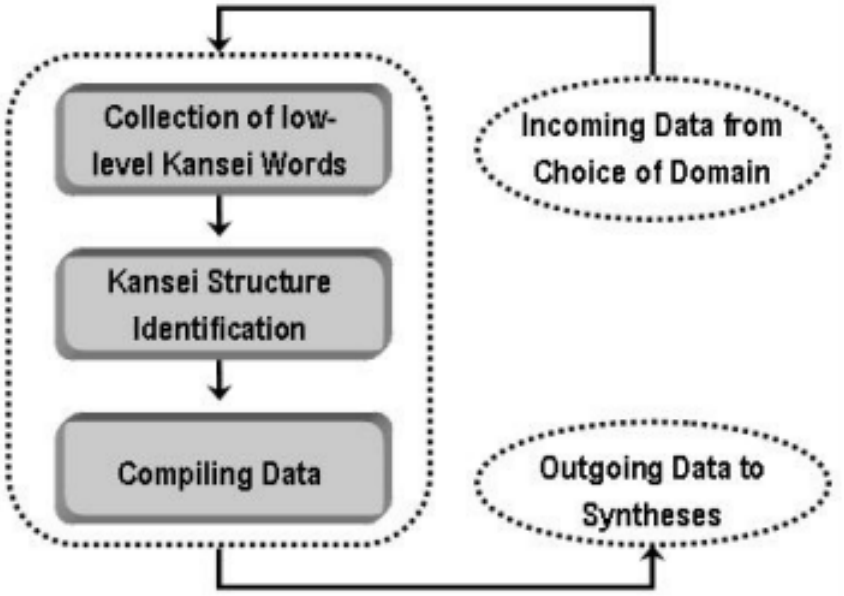
Chọn miền chủ đề

Chủ đề mang ý nghĩa là ý tưởng tổng quát đằng sau sản phẩm. việc lựa chọn miền chủ đề trong đó bao gồm lựa chọn loại sản phẩm sẽ xét đến, đối tượng người dùng sử dụng và các đặc tả cụ thể khác. Tiếp theo đó, miền không gian Kansei và miền không gian thuộc tính sản phẩm sẽ được xác định. Chúng sẽ được phân tích trong bước tổng hợp để tìm ra mối quan hệ giữa thuộc tính của sản phẩm và cảm xúc của người dùng. Từ đó xác định được thuộc tính sản phẩm sẽ ảnh hưởng đến người dùng như thế nào.

Mở rộng miền không gian Kansei

Osgood et al[2] đề xuất rằng mọi vật đều có thể được miêu tả bằng miền không gian vector cảm xúc. Từ miền chủ đề đã xác định, các từ khoá Kansei Word sẽ được thu thập. Kansei Word là các danh từ, tính từ cảm xúc mà người dùng có thể sử dụng để miêu tả về sản phẩm. Số lượng từ khoá thu thập được sẽ đa dạng tuỳ theo từng loại chủ đề khác nhau, giao động từ 100 đến 1000 từ khoá khác nhau. Tuy nhiên, do yếu tố chủ quan của con người, một số từ khoá có thể mang ý nghĩa tương đồng hoặc gần giống nhau. Do vậy tập từ khoá này sau đó sẽ được nhóm lại với nhau bằng phương

pháp thủ công hoặc toán học. cuối cùng, chúng phân tích để chọn ra được những từ khoá đại diện mang ý nghĩa tổng quát độc lập với nhau.



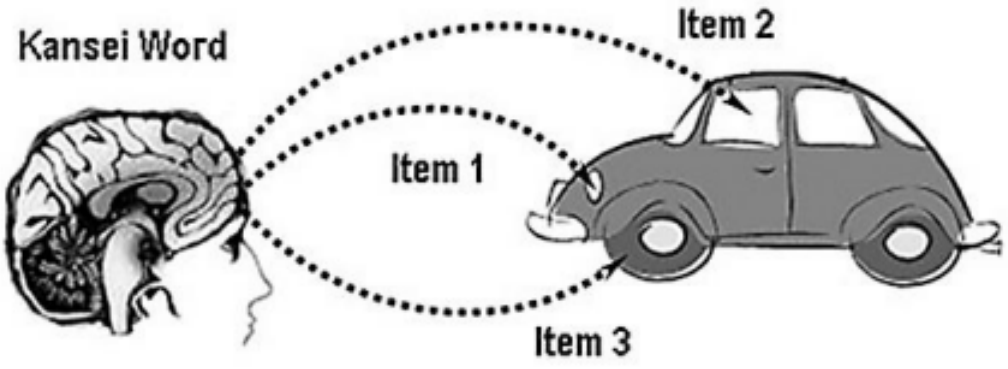
Hình 2.3: Quy trình mở rộng miền không gian Kansei

Mở rộng miền không gian thuộc tính sản phẩm

Tương tự như bước Mở rộng miền không gian Kansei, ở bước này, các thuộc tính của sản phẩm như hình thức, màu sắc, thể loại, nội dung..v..v... được thu thập từ các sản phẩm khác. Các sản phẩm được chọn để khai thác thuộc tính có thể là các sản phẩm đang lưu hành trên thị trường, đề xuất của khách hàng, hoặc thậm chí là ý tưởng thiết kế mới. Tương tự, do có thể có sự tương đồng giữa các thuộc tính gây ảnh hưởng đến độ chính xác nên các thuộc tính được chọn ra là các thuộc tính tiêu biểu nhất của sản phẩm.

Tổng hợp

Trong bước tổng hợp, 2 miền không gian từ khoá Kansei và thuộc tính sản phẩm sẽ được móc nối vào với nhau. Mỗi từ khoá Kansei Word sẽ tương ứng với một hoặc nhiều thuộc tính của sản phẩm, ảnh hưởng trực tiếp đến các thuộc tính đó. Như trong nghiên cứu về thiết kế lon bia của Ishihara et al. (1998)[1] cho thấy rằng, cảm giác "đắng" của người uống chịu ảnh hưởng bởi màu sắc và hình dạng logo lon bia, với màu đen và logo vuông cho người uống cảm giác "rất đắng", trong khi màu trắng và logo hình bầu dục tạo cảm giác ngược lại.



Hình 2.4: Pha tổng hợp

Có nhiều phương pháp tổng hợp, trong đó phổ biến là:

- Category Identification
- Regression Analysis/Quantification Theory Type I
- Rough Sets Theory
- Genetic Algorithm
- Fuzzy Sets Theory

Kiểm tra độ chính xác

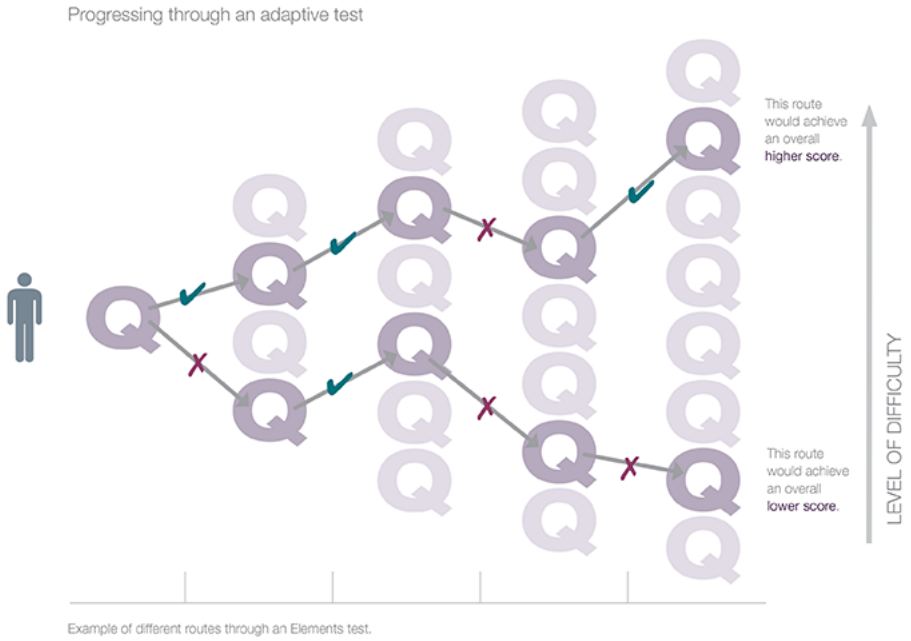
Trước khi mô hình đã xây dựng có thể đem vào sử dụng. Nó cần được phải được kiểm tra độ chính xác, đánh giá xem nó có đủ độ tin cậy và phù hợp với thực tế hay không. Trong trường hợp mô hình không đạt yêu cầu, cần có sự thay đổi trong các bước trên rồi tiếp tục quy trình đánh giá lại cho đến khi ra được kết quả đạt yêu cầu cuối cùng.

2.4 Computerized Adaptive Testing

2.4.1 Tổng quan

Computerized Adaptive Testing - Bài thi tương tác tuỳ biến qua máy tính là hình thức làm bài thi trên máy tính. Trong đó, nội dung và độ khó của câu hỏi sẽ được tuỳ chỉnh sao cho phù hợp với năng lực của thí sinh dự thi. Sau mỗi câu hỏi, trình độ của thí sinh được cập nhập, quyết định độ khó của câu hỏi tiếp theo. Nếu thí sinh trả lời tốt các câu hỏi trước đó, hệ thống sẽ đưa ra các câu hỏi có độ khó cao hơn. Ngược lại,

nếu thí sinh trả lời sai nhiều, độ khó của câu hỏi tiếp theo sẽ được giảm xuống. Và khi đến một ngưỡng nhất định, đã xác định chắc chắn được trình độ của thí sinh thì bài kiểm tra sẽ được hoàn tất.



Hình 2.5: Computerized Adaptive Test

2.4.2 Mô hình

Mô hình CAT cơ bản gồm 4 bước sau đây:

- Chọn ra câu hỏi từ tập câu hỏi dựa vào trình độ thí sinh
- Câu hỏi được chọn sẽ được thí sinh trả lời, đúng hoặc sai
- Trình độ của thí sinh sẽ được cập nhập, dựa vào kết quả của tất cả những câu hỏi đã trả lời
- Lặp lại quá trình trên cho đến khi một ngưỡng quy định trước đạt được.

2.4.3 Ưu điểm

Với phương thức khác hoàn toàn với bài kiểm tra trên giấy thông thường, với CAT mỗi thí sinh được nhận các bài thi hoàn toàn khác nhau. Hơn nữa, số lượng câu hỏi thí sinh cần phải trả lời sẽ được giảm thiểu. Thí sinh sẽ không phải trả lời những câu hỏi quá khó so cũng như quá dễ so với trình độ của họ, do đó thời gian làm bài trung bình sẽ ngắn hơn so với bài kiểm tra thông thường song vẫn cho ra kết quả chính xác tương tự.

2.5 Các cơ sở lý thuyết về công nghệ sử dụng

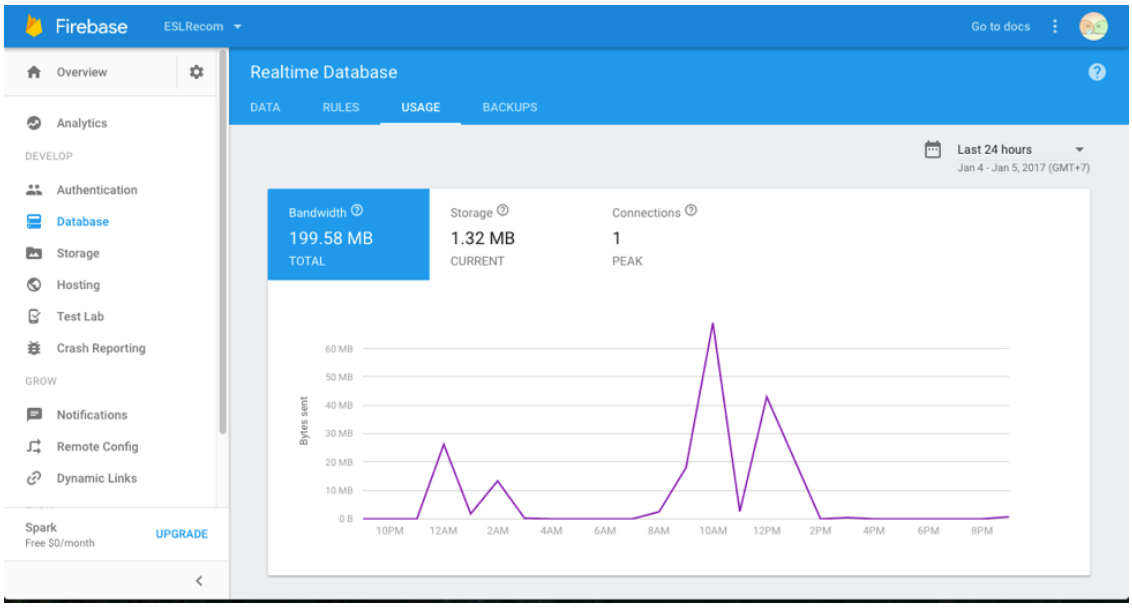
2.5.1 Firebase

Firebase là một dịch vụ cơ sở dữ liệu thời gian thực hoạt động trên nền tảng đám mây được cung cấp bởi Google nhằm giúp các lập trình phát triển nhanh các ứng dụng bằng cách đơn giản hóa các thao tác với cơ sở dữ liệu.

Firebase hỗ trợ tối đa đối với những ứng dụng Backend, nó bao gồm các tiện ích lưu trữ dữ liệu, xác thực người dùng, static, hosting,... Vì thế giúp cho lập trình viên giảm thiểu công việc, và tập trung nâng cao trải nghiệm người dùng.

Realtime Database – Cơ sở dữ liệu thời gian thực

Firebase lưu trữ dữ liệu database dưới dạng JSON và thực hiện đồng bộ database tới tất cả các client theo thời gian thực. Cụ thể hơn là bạn có thể xây dựng được client đa nền tảng (cross-platform client) và tất cả các client này sẽ cùng sử dụng chung 1 database đến từ Firebase và có thể tự động cập nhật mỗi khi dữ liệu trong database được thêm mới hoặc sửa đổi.



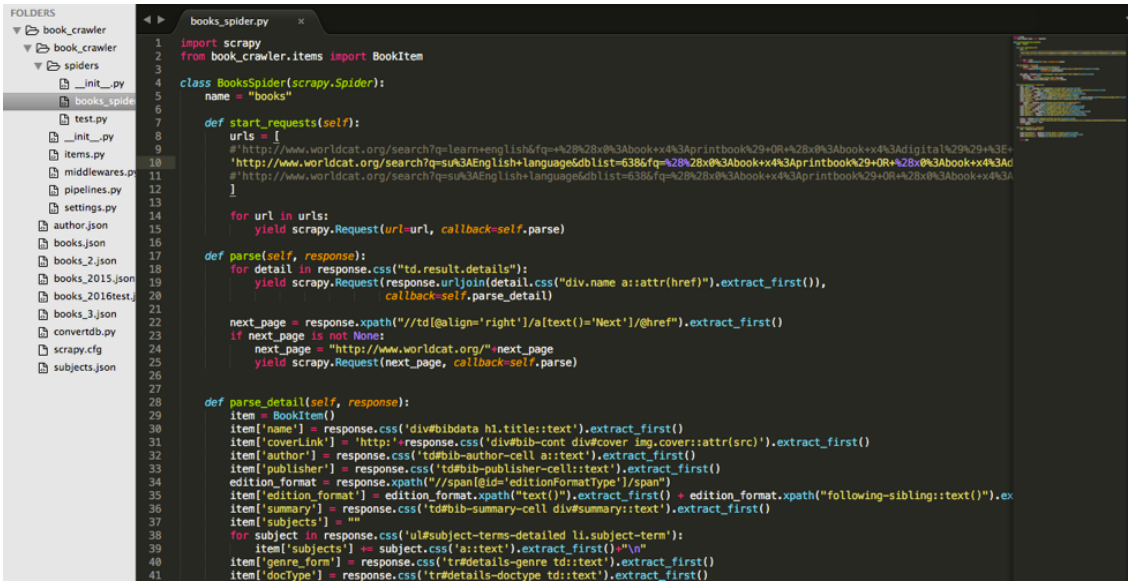
Hình 2.6: Cơ sở dữ liệu thời gian thực

2.5.2 Scrapy

Scrapy là một framework được viết bằng Python, nó cấp sẵn 1 cấu trúc tương đối hoàn chỉnh để thực hiện việc crawl và extract data từ website một cách nhanh chóng và dễ dàng. Bạn muốn lấy dữ liệu từ các website nhưng dữ liệu đó quá lớn để copy rồi paste vào database của bạn, scrapy hỗ trợ bạn làm điều đó. Việc lấy dữ liệu website

Hình 2.7: Lưu trữ cơ sở tri thức chia sẻ giữa các thiết bị

hoàn toàn tự động nhanh chóng và việc sử dụng scrapy cũng rất đơn giản giúp bạn tiếp kiệm được nhiều thời gian và công sức.



Hình 2.8: Crawl tài liệu từ thư viện điện tử WorldCat.org

2.5.3 Thuật toán $tf - idf$

Viết tắt của thuật ngữ tiếng Anh term frequency – inverse document frequency,$tf - idf$ là thuật toán dùng để tìm trọng số của một từ trong văn bản thu được qua thống kê thể hiện mức độ quan trọng của từ này trong một văn bản, mà bản thân văn bản đang xét nằm trong một tập hợp các văn bản.

Nguyên lý cơ bản của $tf - idf$ là "độ quan trọng của một từ sẽ tăng lên cùng với số lần xuất hiện của nó trong văn bản và sẽ giảm xuống nếu như từ đó xuất hiện trong nhiều văn bản khác". Do đó trọng số của một từ t trong tài liệu f sẽ được tính bằng $tf * idf$, với tf là độ phổ biến của từ t trong tài liệu f và idf là nghịch đảo độ phổ biến của từ t trong các tài liệu còn lại của tập tài liệu. Cụ thể theo công thức sau đây:

$$tf(t,d) * idf(t,D) = \frac{f_d(t)}{\max_{w \in d} f_d(w)} * \log \frac{|D|}{|d \in D : t \in d|} \tag{2.1}$$

Lấy ví dụ một văn bản chứa 100 từ, trong đó từ "cat" xuất hiện 3 lần. Giá trị tf của "cat" sẽ là $\frac{3}{100} = 0.03$. Tiếp theo, giả dụ có 10 triệu văn bản và từ "cat" xuất hiện ở 1000 văn bản trong đó. Giá trị idf sẽ là $\log \frac{10,000,000}{1,000} = 4$. Vậy trọng số $tf - idf$ của từ khoá "cat" sẽ là $0.03 * 4 = 0.12$.

Chương 3

Phân tích thiết kế xây dựng hệ thống

3.1 Phân tích tổng quát hệ thống

Các yêu cầu của hệ thống:

- Cung cấp các phương pháp để xác định được trình độ, nguyện vọng và các yếu tố ưu tiên khác trong việc học tiếng Anh của người dùng thông qua giao diện đơn giản, dễ tương tác. Qua đó, đưa ra được các kết quả tư vấn là tài liệu, sách, video, bài giảng Tiếng Anh..v..v... tương ứng.
- Cung cấp giao diện quản lý cho quản trị viên dễ dàng thực hiện các thao tác quản lý thông tin người dùng, quản lý hệ cơ sở tri thức gồm tập câu hỏi kiểm tra và tài liệu tiếng Anh.

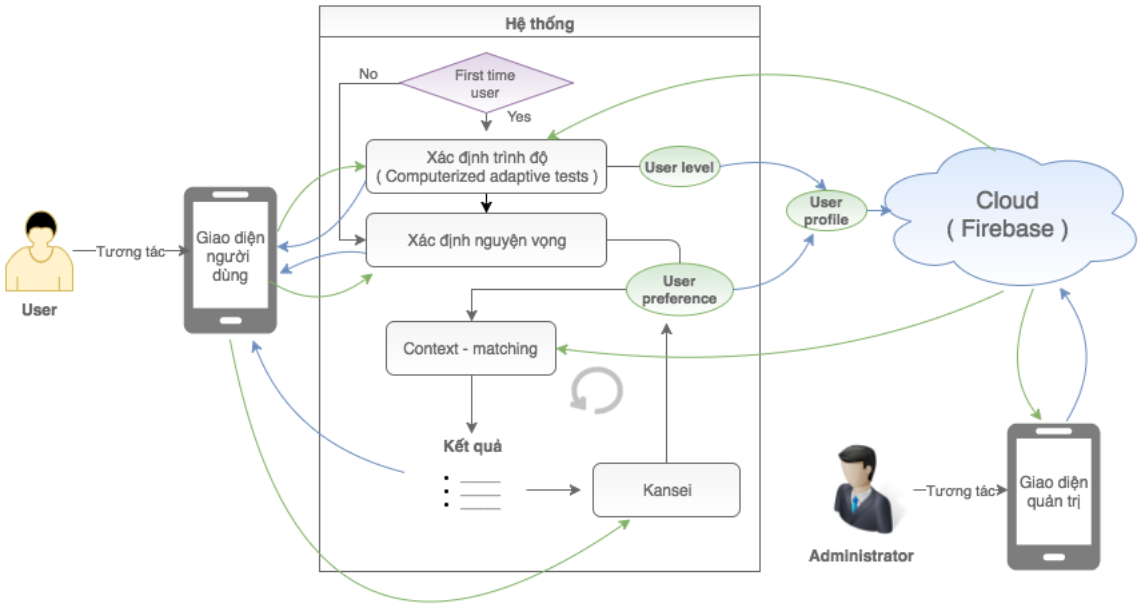
Qua việc khảo sát các hệ thống tư vấn đã và đang được triển khai, đồng thời để giải quyết các yêu cầu đặt ra ở trên, hệ thống đề xuất xây dựng trong đề tài này sẽ có cấu trúc gồm các thành phần sau:

- **Module xác định trình độ:** có nhiệm vụ xác định trình độ của người sử dụng, thông qua việc thực hiện bài kiểm tra General English Test . Sử dụng kĩ thuật Computerized Adaptive Testing, các câu hỏi đưa ra cho người dùng sẽ được tuỳ biến sao cho độ khó phù hợp với năng lực của người dùng. Nhờ vậy, số lượng câu hỏi mà người dùng cần trả lời để xác định được trình độ của họ là ít hơn bài kiểm tra truyền thống, song vẫn cho ra kết quả chính xác như tương tự. Kết quả của bước này sẽ cho ra User level bao gồm trình độ đọc hiểu, vốn từ vựng và vốn ngữ pháp.
- **Module xác định nguyện vọng:** có nhiệm vụ nhận input về nguyện vọng từ người dùng, cụ thể là chủ đề mà người dùng muốn học. Có thể đưa ra các

gợi ý cho người dùng về các chủ đề phổ biến. Kết quả bước này sẽ cho ra User preference là các chủ đề người dùng muốn học dưới dạng keyword.

- **Context-matching:** thực hiện nhận thông tin User level và User preference tổng hợp thành User profile. Sau đó sử dụng thuật toán Context-matching tiến hành matching với profile tài liệu và trả về những kết quả có độ khớp cao nhất.
- **Kansei:** kết quả sau khi Context Matching sẽ được trả về cho người dùng đánh giá trên thang cảm xúc từ "rất thích" cho đến "rất ghét". Dựa vào đánh giá, những thuộc tính trong profile tài liệu ứng với "thích" sẽ được cập nhập vào User preference và lấy nó làm cơ sở context matching các kết quả tiếp theo.
- **Hệ cơ sở tri thức:** sử dụng cơ sở dữ liệu online của Firebase làm cơ sở tri thức cho hệ thống. Nhiệm vụ của nó là trao đổi thông tin với client, cập nhập thông tin mới đảm bảo tính đồng bộ cho toàn hệ thống.
  Dữ liệu được lưu trữ bao gồm:
  - Dữ liệu câu hỏi và đáp án General English Test
  - Thông tin người dùng : id, loại người dùng, trình độ, nguyện vọng.
  - Dữ liệu tài liệu học Tiếng Anh: tên, loại tài liệu, tác giả, miêu tả, nội dung ..v..v... và profile của tài liệu dưới dạng một tập keyword.

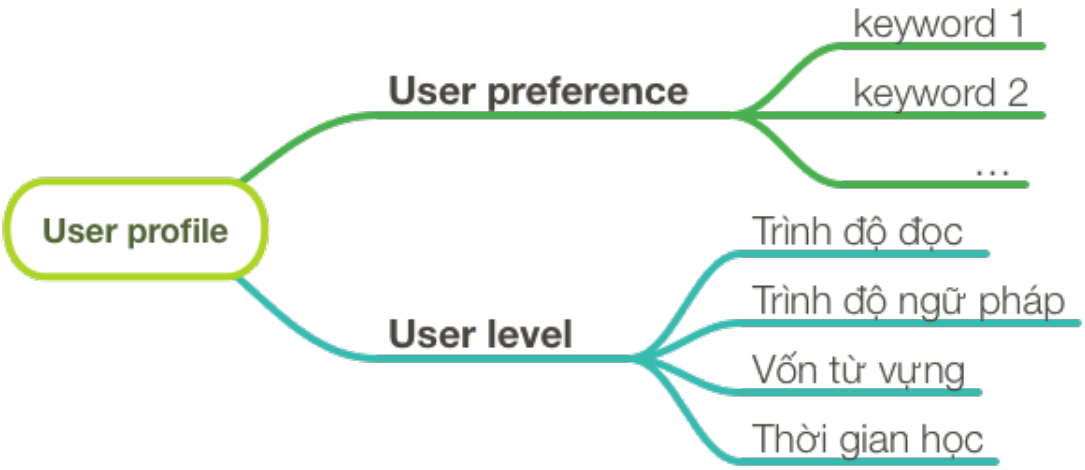
Sau đây là mô hình kiến trúc của ứng dụng:



Hình 3.1: Mô hình kiến trúc ứng dụng

3.2 Xây dựng hồ sơ người dùng

User profile bao gồm trình độ và nguyện vọng của người dùng. Những thông tin này sẽ được dùng trong quá trình matching với các tập dữ liệu bằng thuật toán Context-matching .



Hình 3.2: User profile

3.2.1 Xác định trình độ người dùng

Để xác định trình độ Tiếng Anh của người dùng, tất cả các khía cạnh sau đây cần được khai thác:

- Thời điểm bắt đầu học
- Trình độ đọc hiểu
- Trình độ ngữ pháp
- Vốn từ vựng

Bài kiểm tra tương tác CAT sẽ được sử dụng để đánh giá trình độ.Phương thức chọn câu hỏi và đánh giá bài thi CAT trong đề tài được xây dựng dựa trên mô hình Thử tỉ lệ xác suất nối tiếp [3] (Sequential probability ratio test).

Nguyên lý căn bản nằm bên trong mô hình này là xác suất hợp lý rời rạc. Giả sử từ quan sát thực tế cho thấy thí sinh có trình độ tiếng Anh xuất sắc đạt trung bình 85/100 điểm trong bài thi, trong khi thí sinh có trình độ thấp chỉ đạt 35/100 điểm. Dưới góc nhìn hệ thống có thể coi nó như tập luật $if....else$ sau đây:

1. Thí sinh có trình độ xuất sắc, đã nắm vững kiến thức và hiểu rõ câu hỏi cũng như phương pháp giải quyết, do vậy khả năng mà họ trả lời đúng câu hỏi sẽ là 85 %.

   Prob(Correct|Master) = .85 (P_m)

   Prob(Incorrect|Master) = .15

2. Ngược lại, thí sinh có trình độ thấp, trả lời phần lớn dựa trên may rủi, giác quan thứ 6 của bản thân, do vậy khả năng mà họ trả lời đúng câu hỏi sẽ là 35 %.

   Prob(Correct|Nonmaster) = .35 (P_n)

   Prob(Incorrect|Nonmaster) = .75

Trong bài kiểm tra CAT, với câu hỏi bất kì phù hợp trình độ được chọn ra trong tập câu hỏi đưa cho thí sinh. Quan sát trả lời của thí sinh, xác suất khả năng trình độ sẽ được tính bằng:

$$PR = \frac{P_m^t(1-P_m)^f}{P_m^t(1-P_m)^f} \tag{3.1}$$

trong đó P_m = khả năng thí sinh trình độ xuất sắc trả lời đúng câu hỏi
P_n = khả năng thí sinh trình độ thấp trả lời đúng câu hỏi
t = tổng số câu hỏi thí sinh trả lời đúng
f = tổng số câu hỏi thí sinh trả lời sai
Giá trị PR sau đó sẽ được đem so sánh với tập luật:

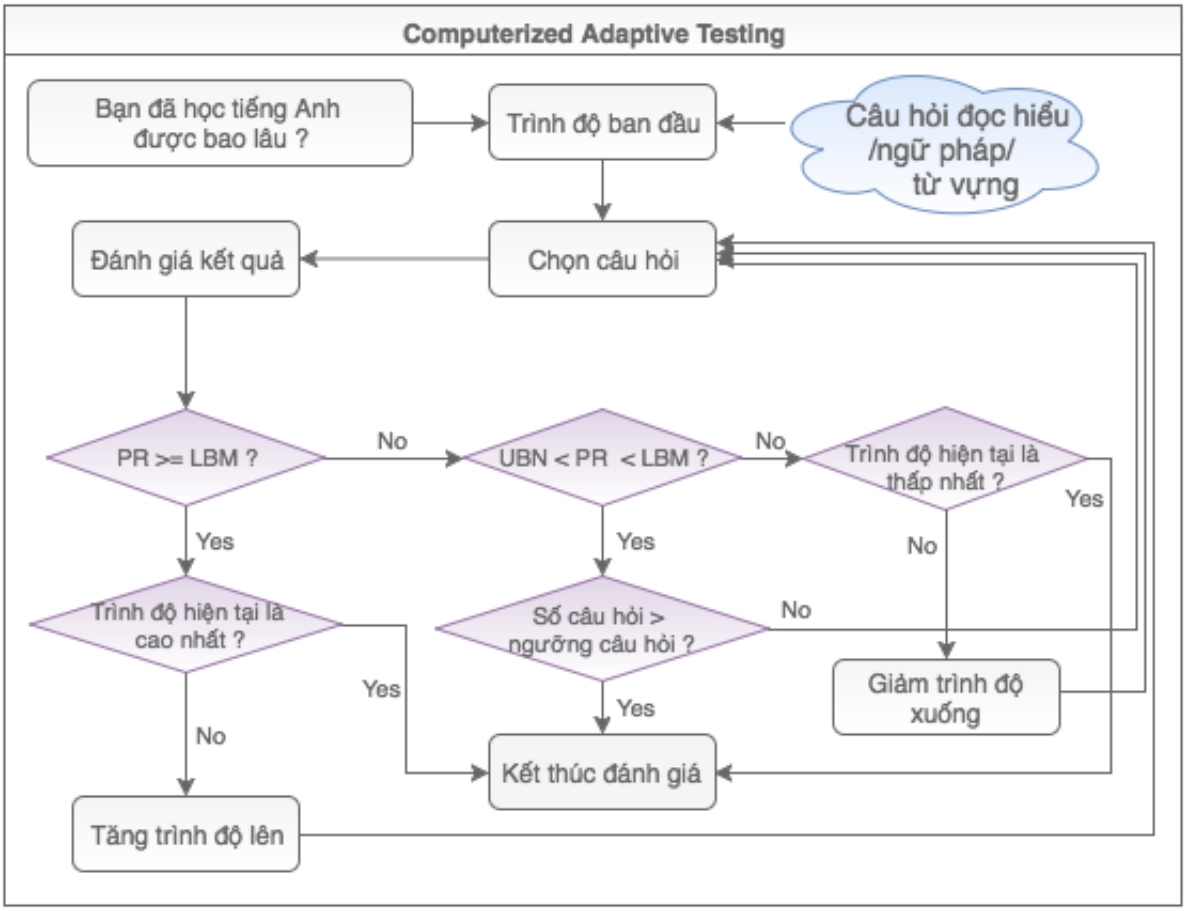
- Nếu $PR > UBN$ (Upper Bound Nonmastery: giá trị ngưỡng trên của độ không thuần thục) -> trình độ người dùng thấp hơn câu hỏi hiện tại, chọn câu hỏi tiếp theo ở trình độ thấp hơn để tiếp tục đánh giá .
- Nếu $PR < LBM$ (Lower Bound Mastery: giá trị ngưỡng dưới của độ thuần thục) -> trình độ người dùng nằm trên câu hỏi hiện tại, chọn câu hỏi tiếp theo ở trình độ cao hơn để tiếp tục đánh giá.
- Nếu $UBN < PR < LBM$, kết quả hiện tại chưa đủ để đánh giá trình độ người dùng, chọn câu hỏi tiếp theo ở cùng trình độ để tiếp tục đánh giá.

Trình độ của thí sinh được xác định khi PR lớn hơn giá trị UBN hoặc nhỏ hơn giá trị LBM. Tuy nhiên, xét trong trường hợp thực tế có khả năng xảy ra việc số lượng câu hỏi thí sinh trả lời sai bằng với số lượng câu hỏi thí sinh trả lời đúng. Điều này dẫn đến tình trạng thí sinh trả lời một số lượng lớn câu hỏi nhưng vẫn không xác định được trình độ. Điều kiện dừng sẽ được đưa vào để giải quyết trường hợp đó.

Hệ thống sẽ dừng bài kiểm tra đánh giá nếu như:

- Nếu $PR > UBN$ và *"độ khó hiện tại là thấp nhất"*
- Nếu $PR < LBM$ và *"độ khó hiện tại là cao nhất"*
- Nếu $UBN < PR < LBM$ và số lượng câu hỏi hiện tại vượt *ngưỡng câu hỏi*

Sau dây là mô hình đánh giá trình độ người dùng sử dụng trong đề tài:



Hình 3.3: Mô hình bài kiểm tra tương tác

Trước tiên, trình độ ban đầu của người dùng sẽ được xác định qua câu hỏi "Bạn đã học tiếng Anh được bao lâu rồi". Trình độ của một người nào đó thường tỉ lệ thuận với thời gian họ bỏ ra để học. do vậy ta có thể phần nào phán đoán được thông qua thông tin thời gian học. Đây là bước tiền đề trước khi đi vào thực hiện bài thi.

Tập câu hỏi đánh giá trình độ ban đầu, gồm hơn 70 câu hỏi lấy từ trang http://www.englishjet.com/. Hình thức của câu hỏi ở dạng trắc nghiệm với 4 đáp án. Chúng được phân loại vào từng tập câu hỏi khác nhau theo các *chủ đề* {*đọc hiểu, từ vựng, ngữ pháp*} và *trình độ* {*nhập môn, cơ bản, trung bình, khá, cao cấp*}. Dựa trên trình độ hiện tại của người dùng, hệ thống sẽ chọn bất kì một câu hỏi trong tập câu hỏi cùng trình độ ra để kiểm tra. Sau khi kết thúc một chủ đề, trình độ hiện tại của

người dùng sẽ được sử dụng làm trình độ ban đầu trong đánh giá chủ đề tiếp theo.

You can exchange the gift .....

A. so long that

B. while

C. as long as

D. meanwhile

E. whether

..... it is returned within seven days.

Việc đánh giá kết quả được thực hiện theo tập luật đã đề cập ở phía trên. Dựa vào quan sát thử nghiệm trong thực tế, hệ thống đề xuất trong đề tài sử dụng các giá trị $UBN = 0.02$, $LBM = 7$ và *ngưỡng câu hỏi* $= 5$.

Ví dụ, một thí sinh học tiếng Anh được 4 năm làm bài kiểm tra trình độ. Hệ thống sẽ dự đoán trình độ của anh ta nằm ở mức trung bình và đưa ra câu hỏi ở mức trình độ đó. Bắt đầu với chủ đề *từ vựng*, thí sinh này trả lời đúng liên tiếp 3 câu hỏi. PR của anh ta lúc đó sẽ là $\frac{0.85^3(1-0.85)^0}{0.35^3(1-0.35)^0} \approx 14.3236 > LBM = 7$. Điều kiện tăng trình độ thoả mãn, trình độ thí sinh được nâng lên mức khá. Tiếp tục quá trình đánh giá, lần này với câu hỏi trình độ khá, thí sinh lần lượt đạt kết quả Đúng-Sai-Đúng-Đúng-Sai, tương ứng với $PR = \frac{0.85^3(1-0.85)^2}{0.35^3(1-0.35)^2} \approx 3.305$. Do $UBN < PR < LBM$ và số lượng câu hỏi đã trả lời đạt ngưỡng, việc đánh giá trình độ từ vựng của thí sinh kết thúc. Kết quả là thí sinh đạt trình độ ***khá*** về *từ vựng*. Lấy trình độ *khá* làm trình độ ban đầu, hệ thống tiếp tục đánh giá chủ đề tiếp theo. Kết quả thu được là ***trung bình*** về *ngữ pháp* và ***cao cấp*** về *đọc hiểu*. Lấy trung bình 3 kết quả thu được, hệ thống kết luận trình độ tiếng Anh của thí sinh là ***khá***.

3.2.2 Xác định nguyện vọng người dùng

Nguyện vọng người dùng được xác định một cách đơn giản bằng việc người dùng trực tiếp nhập nội dung mình muốn học. Hệ thống sẽ cung cấp cho người dùng một giao diện nhập dữ liệu đơn giản. Người dùng sử dụng bàn phím sẽ nhập nguyện vọng của mình vào dưới dạng các keyword, phân tách nhau bởi dấu cách.

Để hỗ trợ xác định nguyện vọng, hệ thống sẽ đưa ra các keyword gợi ý dựa trên nội dung người dùng nhập vào. Các keyword gợi ý này được tổng hợp bằng việc phân

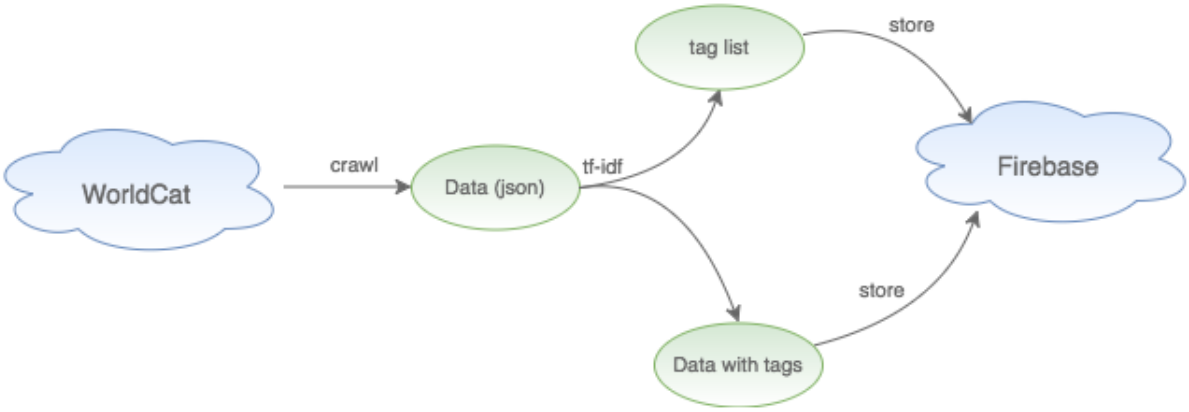
tích keyword các tài liệu Tiếng Anh sử dụng thuật toán $tf - idf$. Nội dung của công đoạn phân tích keyword sẽ được trình bày cụ thể ở phần sau.

Sau hai bước trên, một user profile như ví dụ sau được xây dựng.

```
'userProfile': {
        'id': 1,
        'vocabProfi': 'upper-intermediate',
        'grammarProfi': 'intermediate',
        'readingProfi': 'advanced',
        'overallProfi': 'upper-intermediate',
        'preference': ['ietls', 'grammar', 'advanced']
    }
```

3.3 Thu thập và xử lý tài liệu học tiếng Anh

Đây là bước tiền xử lý trước khi đi vào xây dựng hệ thống.



Hình 3.4: Mô hình thu thập và xử lý tài liệu học tiếng Anh

3.3.1 Thu thập dữ liệu

Tài liệu học tiếng Anh hệ thống sử dụng được lấy từ www.worldcat.org . WorldCat được biết đến như là một CSDL liên hợp toàn cầu. Là một thư viện điện tử chứa đựng dữ liệu từ 72,000 thư viện ở hơn 170 quốc gia và vùng lãnh thổ, WorldCat chứa một lượng dữ liệu khổng lồ gồm hơn 330 triệu bản ghi, với số lượng ngôn ngữ cực kỳ đa dạng, gồm gần 500 ngôn ngữ trên toàn thế giới, trong đó tiếng Anh chiếm khoảng 38%, tiếng Đức khoảng 13%, tiếng Pháp khoảng 9%, ngoài ra là các ngôn ngữ khác như tiếng Tây Ban Nha, tiếng Trung, tiếng Nhật, tiếng Hàn, và bao gồm cả tiếng Việt (cho dù chỉ là một tỷ lệ nhỏ). Nhiều chuyên gia đánh giá rằng WorldCat có thể bao

gồm tới trên 70% lượng tài liệu có trên toàn cầu, và là bộ CSDL thư mục toàn diện nhất thế giới từ trước tới giờ.

Sử dụng framework ***Scrapy***, function sau đây được viết để trích xuất dữ liệu từ WorldCat.

```
import scrapy
from book_crawler.items import BookItem

class BooksSpider(scrapy.Spider):
  name = "books"
  page = 0

  def start_requests(self):
        url = 'http://www.worldcat.org/search?q=kw%3Aenglish&fq
           =yr%3A2014..2017+%3E+%3E+-mt%3Afic+%3E+ln%3Aeng&qt=
           advanced&dblist=638'
        yield scrapy.Request(url=url, callback=self.parse)

  def parse(self, response):
        for detail in response.css("td.result.details"):
              yield scrapy.Request(response.urljoin(detail.
                 css("div.name a::attr(href)").extract_first()
                 ), callback=self.parse_detail)
        next_page = response.xpath("//td[@align='right']/a[text
           ()='Next']/@href").extract_first()
        BooksSpider.page += 1
        if next_page is not None and BooksSpider.page <= 500:
              next_page = "http://www.worldcat.org/"+
                 next_page
              yield scrapy.Request(next_page, callback=self.
                 parse)

  def parse_detail(self, response):
        item = BookItem()
        item['name'] = response.css('div#bibdata h1.title::text
           ').extract_first()
        item['coverLink'] = 'http:'+response.css('div#bib-cont
           div#cover img.cover::attr(src)').extract_first()
        item['author'] = response.css('td#bib-author-cell a::
           text').extract_first()
        item['publisher'] = response.css('td#bib-publisher-cell
           ::text').extract_first()
```

```
        edition_format = response.xpath("//span[@id='
            editionFormatType']/span")
        item['edition_format'] = edition_format.xpath("text()")
            .extract_first() + edition_format.xpath("following-
            sibling::text()").extract_first()
        item['summary'] = response.css('td#bib-summary-cell div
            #summary::text').extract_first()
        item['subjects'] = ""
        for subject in response.css('ul#subject-terms-detailed
            li.subject-term'):
                item['subjects'] += subject.css('a::text').
                    extract_first()+"\n"
        item['genre_form'] = response.css('tr#details-genre td
            ::text').extract_first()
        item['docType'] = response.css('tr#details-doctype td::
            text').extract_first()
        item['note'] = response.css('tr#details-notes td::text'
            ).extract_first()
        item['description'] = response.css('tr#details-
            description td::text').extract_first()
        item['content'] = response.css('tr#details-contents td
            ::text').extract_first()
        item['abstract'] = response.css('div.abstracttxt::text'
            ).extract_first()
        item['onlineName'] = response.css('div#links-all856 p a
            ::text').extract_first()
        item['onlineLink'] = response.css('div#links-all856 p a
            ::attr(title)').extract_first()
        oclcno = response.css('tr#details-oclcno td::text').
            extract_first()
        if oclcno is not None:
                request = scrapy.Request('http://www.worldcat.
                    org/wcpa/servlet/org.oclc.lac.ui.buying.
                    AjaxBuyingLinksServlet?serviceCommand=
                    getBuyingLinks&oclcno='+oclcno, callback=self
                    .parse_seller)
                request.meta['item'] = item
                yield request

    def parse_seller(self, response):
        item = response.meta['item']
        item['sellerName'] = response.css('td.seller a::text').
```

```
            extract_first()
        item['sellerLink'] = response.css('td.seller a::attr(
            href)').extract_first()
        item['sellerPrice'] = response.css('td.price::text').
            extract_first()
        yield item
```

Để phục vụ cho quy mô thử nghiệm thuật toán, 5000 tài liệu tiếng Anh phát hành trong thời điểm từ năm 2014 đến 2017 được thu thập. Chúng bao gồm sách, e-book, bản ghi âm, file audio, video..v..v... có nội dung nằm trong các chủ đề đa dạng khác nhau:

- Đọc hiểu
- Nghe hiểu
- Hội thoại, giao tiếp
- Viết văn
- IETLS
- TOEIC
- Chuyên ngành pháp luật
- Chuyên ngành y tế
- Chuyên ngành toán học

.....

Tài liệu trích xuất được lưu trữ dưới định dạng json:


```
{
        "abstract": "Approximately 500 words and their definitions...",
        "author": "Steven J Matthiesen",
        "content": "Success on the TOEFL --",
        "coverLink": "http://coverart.oclc.org/ImageWebSvc/oclc/...",
        "description": "1 online resource (vii, 344 pages)",
        "docType": "Internet Resource, Computer File",
        "edition_format": "eBook : Document : English : 6th edition",
        "genre_form": "Electronic books",
        "name": "Barron's essential words for the TOEFL : test...",
        "note": "Previous edition: 2011.",
        "onlineLink": "https://www.overdrive.com/search?q=D2F0C4CD-...",
        "onlineName": "OverDrive",
```

```
        "publisher": "Hauppauge : Barron's, 2014.",
        "sellerLink": "https://www.amazon.com/Essential-Words-TOEFL...",
        "sellerName": "Amazon.com",
        "sellerPrice": "$9.59",
        "subjects": "Test of English as a Foreign Language -- Study..."
}
```

3.3.2 Xử lý dữ liệu

Dữ liệu thu thập được xử lý bằng thuật toán $tf - idf$ để xác định các từ khoá tiêu biểu đại diện cho nội dung của tài liệu. Văn bản đầu vào của thuật toán $tf - idf$ bao gồm nội dung của các trường "*name*", "*abstract*" và "*subject*".

```
import math
import json
import jsonpickle
from textblob import TextBlob as tb

def tf(word, blob):
    return blob.words.count(word) / len(blob.words)

def n_containing(word, bloblist):
    return sum(1 for blob in bloblist if word in blob.words)

def idf(word, bloblist):
    return math.log(len(bloblist) / (1 + n_containing(word, bloblist)))

def tfidf(word, blob, bloblist):
    return tf(word, blob) * idf(word, bloblist)
```

Kết quả thu được là tập từ khoá ứng với mỗi tài liệu như sau:

Tài liệu	**Tags**
Dictionary of medical terms	medicine, terms, dictionary, zymotic, surgery, specialisations, pathology, anatomical, euphemistic, diagnosis
Pronouncing and defining dictionary of music	musicians, music, bio-bibliography, pronouncing, defining, dictionary
Teaching reading vocabulary	reading, comprehension, lecture, subjects, vocabulary, english, curriculum
....	....

Bảng 3.1: Ví dụ kết quả phân tích dữ liệu

Để thu được những tag tiêu biểu nhất, hệ thống áp dụng điều kiện ràng buộc {*mỗi từ khoá phải xuất hiện ít nhất trong 5 tài liệu*} để loại đi các từ khoá thiểu số.

Sau bước xử lý dữ liệu, ta xây dựng được profile của tài liều tiếng Anh như ví dụ sau:

```
'documentProfile': {
        'id': 12,
        'author': 'John S Kwan',
        'name' : 'English spelling',
        'abtract': '...',
        'content': '...',
        ....
        'tags': ['spelling', 'punctuation', 'orthography', '
           rules', 'self-study']
    }
```

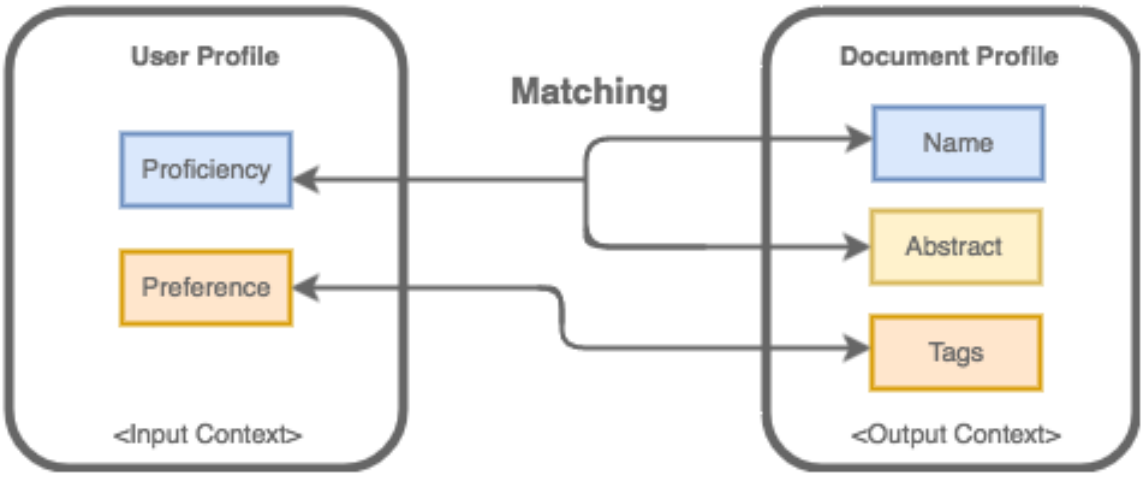
Ngoài ra, một bảng chứa các từ khoá sắp xếp theo thứ tần xuất xuất hiện giảm dần cũng được xây dựng để hỗ trợ người dùng trong bước xác định nguyện vọng như đã đề cập ở bên trên. Chúng sẽ được lưu trữ trong CSDL của Firebase và trả về client mỗi khi có request gửi lên.

```
'tagList': [
        {'name': 'teaching', 'score': 429},
        {'name': 'writing', 'score': 389},
        {'name': 'vocabulary', 'score': 383},
        {'name': 'juvenile', 'score': 351},
        {'name': 'problems', 'score': 313},
        {'name': 'exercises', 'score': 304},
 ...
]
```

3.4 So sánh và đưa ra tư vấn tài liệu học

Sau khi có được *User Profile* người dùng, kết hợp với *Document Profile* đã xử lý trước lấy từ cơ sở tri thức Firebase. Hệ thống sẽ thực hiện so sánh sử dụng thuật toán Context Matching như đã trình bày ở (2.2.1) để đưa ra kết quả tư vấn.

Cụ thể, với bài toán tư vấn tài liệu Tiếng Anh dựa vào trình độ và nguyện vọng của người dùng, *Input Context* và *Output Context* được sử dụng là *User Profile* và *Document Profile*. Các thuộc tính được quan tâm đến tương ứng gồm {*"trình độ"*, *"nguyện vọng"*} và {*"tên tài liệu"*, *"mô tả"*, *"tập từ khoá"*}.



Hình 3.5: So sánh độ tương đồng giữa người dùng và tài liệu

3.4.1 Tính giá trị match e

Do đặc thù bài toán với số lượng thuộc tính của *User Profile* và *Document Profile* không tương đồng, cộng với việc số liên kết được xét đến trong quá trình Context-Matching là ít, việc áp dụng hoàn toàn cách tính của thuật toán được đề xuất sẽ cho ra kết quả có độ chính xác và độ đa dạng tương đối thấp.

Cụ thể là với cách tính giá trị e truyền thống, e chỉ có thể mang một trong hai giá trị 1 - phù hợp, 0 - không phù hợp. Xét một ví dụ có input context là *User profile* {*"Intermediate"*, *"grammar, ielts, video"*} và các output context cần so sánh gồm:

1. ***DP1*** {*"Study English - Intermediate Level. [Series 1], Episode 25"*, *"video interviews with native speakers on topics with relevance to IELTS"*, *"grammar, ielts, video"*}
2. ***DP2*** {*"Funny phonics & silly spelling."*, *"For English learners of intermediate proficiency"*, *"grammar, ielts, video, funny, phonetic, spelling, literature, conversation"*}
3. ***DP3*** {*"6 IELTS Grammar Tests."*, *"For English learners of intermediate proficiency"*, *"grammar, ielts, test"*}

Dễ thấy, nếu áp dụng đúng nguyên mẫu thuật toán.

Ta có $e_{DP1}(1) = e_{DP2}(1) = e_{DP3}(1) = 1$ (do cả 3 tài liệu đều thuộc trình độ Intermediate) và $e_{DP1}(2) = e_{DP2}(2) = 1$, $e_{DP3}(2) = 0$ ($DP1$ và $DP2$ chứa đủ từ khoá nguyện vọng của người dùng, trong khi $DP3$ thiếu mất *"video"*)

Sau khi kết thúc tính toán, ta sẽ có độ khớp của 3 tài liệu trên với người dùng là : $rv_{DP1} = rv_{DP2} > rv_{DP3}$. Kết quả này có 2 nhược điểm:

- Nội dung tài liệu $DP1$ chắc chắn sẽ phù hợp hơn so với $DP2$ do tập từ khoá của $DP1$ trùng khớp hoàn toàn với nguyện vọng người dùng, trong khi của $DP2$ chỉ khớp một phần. Tuy nhiên, giá trị khớp rv của 2 tài liệu này lại là như nhau.
- Tài liệu $DP3$ tương đối phù hợp với nguyện vọng của người dùng, thậm chí có thể còn phù hợp hơn $DP2$ thì lại có rv thấp hơn. Tuy thiếu mất từ khoá *"video"*, nhưng trong tình huống này, nguyện vọng chính của người dùng học là học ngữ pháp IELTS. Dù hình thức học không phải là video đi chăng nữa thì nó vẫn là kết quả phù hợp với nguyện vọng chấp nhận được.

Vì vậy, để giải quyết các nhược điểm trên, đề tài đề xuất một cách tính giá trị e mới phù hợp với điều kiện bài toán hơn như sau:

Khi so sánh proficiency - name/abstract: Giữ nguyên cách tính như cũ. Tìm xem từ khoá trình độ người dùng có trong tên hoặc miêu tả của tài liệu hay không -> $e(1) = 1|0$. Thông thường, giá trị trình độ tổng thể sẽ được dùng để so sánh. Trong trường hợp người dùng muốn tìm kiếm về tài liệu liên quan đến từ vựng, ngữ pháp và đọc hiểu thì sẽ sử dụng giá trị trình độ tương ứng để so sánh.

Khi so sánh preference - tags: Giá trị e [0.00...1.00] được tính bằng trung bình cộng của (Số nguyện vọng khớp/Tổng số nguyện vọng) và (Giá trị phổ biến của các từ khoá khớp/Tổng giá trị phổ biến của tập từ khoá trong tài liệu):

$$e = \frac{1}{2}\left(\frac{|P \cap T|}{|P|} + \frac{\sum\limits_{P \cap T} ts}{\sum\limits_{T} ts}\right) \tag{3.2}$$

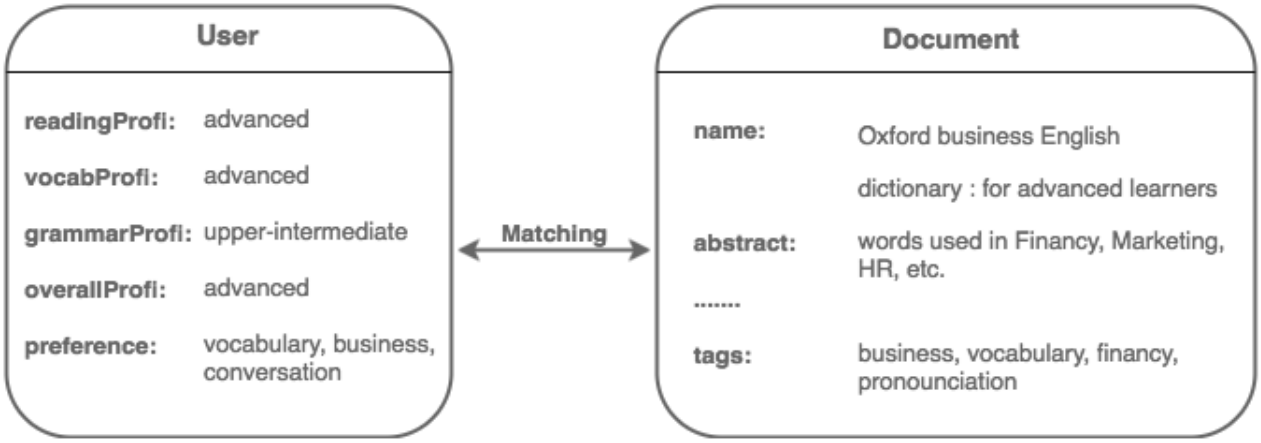
trong đó P = tập nguyện vọng người dùng

T = tập từ khoá của tài liệu

ts = giá trị phổ biến của từ khoá, là số lần xuất hiện của từ khoá đó trên tất cả tài liệu (3.3.2)

3.4.2 Ví dụ với case study

Một người dùng thực hiện bài kiểm tra và đạt kết quả trình độ cao cấp (*advanced*). Anh ta có nhu cầu học từ vựng tiếng Anh văn phòng. Hệ thống sẽ thử matching User Profile này với cuốn sách "*Oxford business English dictionary : for learners of English*".



Hình 3.6: Input context & Output context

Quá trình Context-Matching sẽ diễn ra như sau:

Bước 1: Đánh giá match và xác định e

No	Input	Output	e
1	advanced	Oxford business English dictionary : for advanced learners words used in Financy, Marketing, HR, etc.	1
2	vocabulary, business, conversation	business, vocabulary, financy, pronounciation	0.6788

Bảng 3.2: Xác định e

Ta có $e_1 = 1$,

Giá trị phổ biến của "vocabulary", "business", "conversation", "financy", "pronounciation" lần lượt là 383, 288, 158, 52, 248.

$$=> e_2 = \frac{1}{2}\left(\frac{2}{3} + \frac{383+288}{383+288+52+248}\right) = 0.6788$$

Bước 2: Xác định w

Giá trị w được chuyên gia quy định sẵn trong hệ thống, trong ví dụ này, chúng có gía trị $w_1 = 0.37$ & $w_2 = 0.78$

Bước 3: Tính giá trị av

$av_1 = e_1 * w_1 = 0.37$
$av_2 = e_2 * w_2 = 0.5294$

Bước 4: Tổng sav

$$sav = av_1 + av_2 = 0.8994$$

Bước 5: Tính giá trị match lớn nhất mpv

$$mpv = w_1 + w_2 = 1.15$$

Bước 6: Tính độ phù hợp rv

$$rv = \frac{sav}{mpv} = \frac{0.8994}{1.15} = 0.782$$

Bước 7: So sánh với giá trị ngưỡng t

Giá trị ngưỡng t được chuyên gia quy định sẵn trong hệ thống, trong ví dụ này, nó có giá trị $t = 0.65$.

Do $rv = 0.782 > t =>$ Kết quả là tài liệu được xét có phù hợp với nguyện vọng và trình độ người dùng.

3.5 Áp dụng Kansei Engineering để cải thiện kết quả tư vấn

Chương 4

Kết quả thực hiện

Chương 5

Kết luận và hướng phát triển

Tài liệu tham khảo

[1] Ishihara K. Ishihara, S. and M. Nagamachi. Hierarchical kansei analysis of beer can using neural network. In *Proceedings of Human Factors in Organizational Design and Management - VI*, pages 421–425, 1998. 8

[2] G.J. Suci Osgood, C.E. and P.H. Tannenbaum. The measurement of meaning. 7

[3] Theodore W Frick R. Edwin Welch. Computerized adaptive testing in instructional settings. *Educational Technology Research and Development*, pages 47–62, 1993. 16

[4] Simon Schütte. Engineering emotional values in product design-kansei engineering in development. *Linköpings Universitet*, pages 55–64, 2005. 6