

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 CẢM 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

Mục lục	v
Danh sách hình vẽ	vii
1 Đặt vấn đề và định hướng giải pháp	1
1.1 Đặt vấn đề	1
1.2 Định hướng giải quyết	2
2 Cơ sở lý thuyết	3
2.1 Khảo sát các hệ tư vấn đã có	3
2.1.1 Lọc cộng tác	3
2.1.2 Lọc dựa trên nội dung	4
2.2 Thuật toán Context-matching	4
2.2.1 Tổng quan	5
2.2.2 Input context và output context	5
2.2.3 Các giá trị sử dụng	5
2.2.4 Các bước thực hiện thuật toán	6
2.3 Kỹ thuật Kansei Engineering	6
2.3.1 Tổng quan	6
2.3.2 Mô hình	6
2.3.2.1 Mở rộng miền không gian Kansei	7
2.3.2.2 Mở rộng miền không gian thuộc tính sản phẩm	8
2.3.2.3 Tổng hợp	8
2.4 Computerized Adaptive Testing	8
2.5 Các cơ sở lý thuyết về công nghệ sử dụng	8
2.5.1 Firebase	8
2.5.2 Scrapy	10
2.5.3 Thuật toán $tf - idf$	10
3 Phân tích thiết kế xây dựng hệ thống	11
3.1 Phân tích tổng quát hệ thống	11
3.2 Xây dựng hồ sơ người dùng	13

3.3	Xây dựng mạng từ khoá	14
4	Kết quả thực hiện	15
5	Kết luận và hướng phát triển	16
	Tài liệu tham khảo	17

Danh sách hình vẽ

2.1	Mô hình thuật toán Context Matching	4
2.2	Mô hình tổng quát Kansei Engineering	7
2.3	Quy trình mở rộng miền không gian Kansei	8
2.4	Cơ sở dữ liệu thời gian thực	9
2.5	Lưu trữ cơ sở tri thức chia sẻ giữa các thiết bị	9
2.6	Crawl tài liệu từ thư viện điện tử WorldCat.org	10
3.1	Mô hình kiến trúc ứng dụng	12

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 thống 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 mại 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 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, đề án sẽ trình bày về các kiến thức cơ bản cũng 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 dung

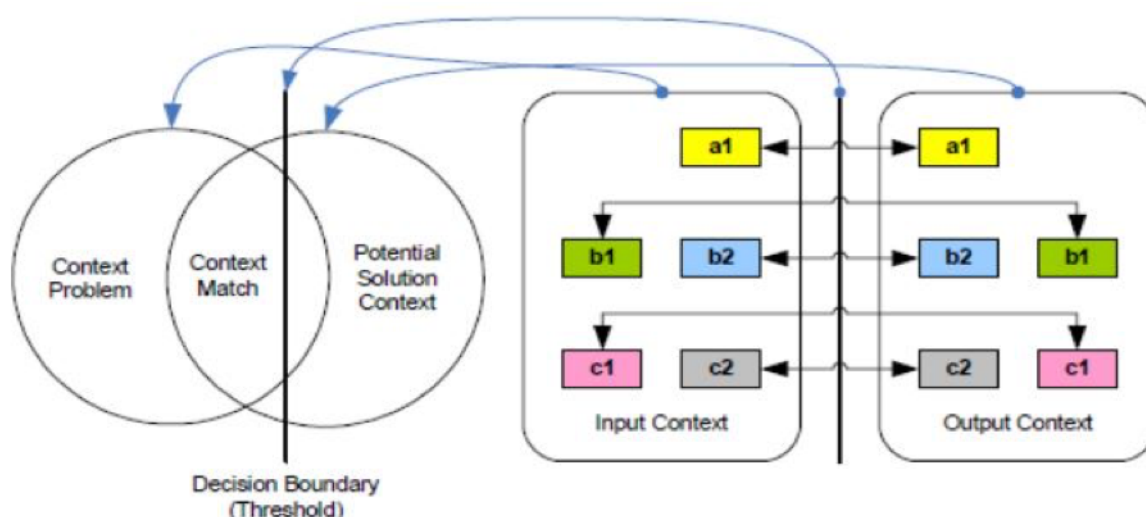
là phương pháp mà người ta quan tâm đến 2 thực thể chính là người dùng và đối tượng được khuyến nghị đến cho người dùng. Quá trình lọc thực hiện bắt đầu từ việc xây dựng một hồ sơ người dùng (user profile) là tập các ưu tiên về sở thích của người dùng về đối tượng, và hồ sơ miêu tả đối tượng. Sau đó 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 cần 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 \rightarrow 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 $(rv) \geq (t)$ 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

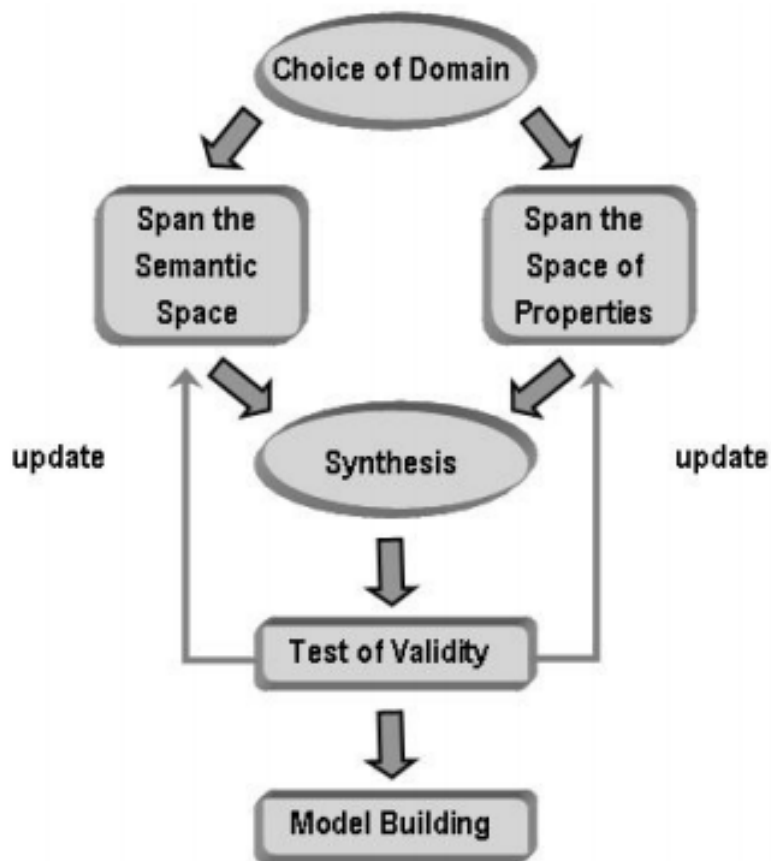
Kansei Engineering là kỹ thuật tích hợp khía cạnh cảm xúc của con người vào trong quá trình xây dựng sản phẩm, nhằm mục đích tạo ra được sản phẩm phù hợp với yêu cầu và mong muốn của người dùng. Nó là việc mang lại sự hài lòng, thoả mãn cho người dùng một cách có khoa học. Để đạt được điều đó, trải nghiệm 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 sau đây:

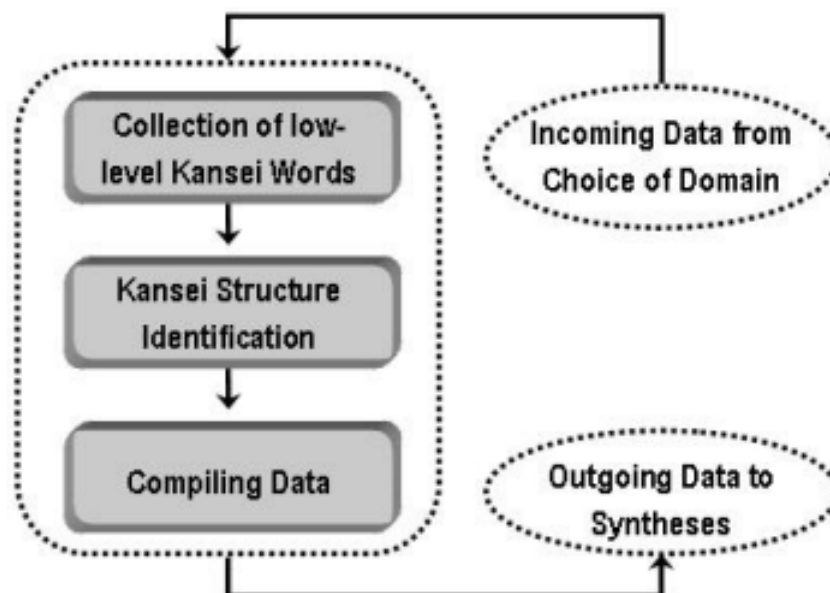


Hình 2.2: Mô hình tổng quát Kansei Engineering

Bắt đầu từ việc chọn ra một miền chủ đề nhất định, 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.

2.3.2.1 Mở rộng miền không gian Kansei

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 tùy 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 phân tích để chọn ra được những từ khoá mang ý nghĩa tổng quát độc lập với nhau.



Hình 2.3: Quy trình mở rộng miền không gian Kansei

2.3.2.2 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.

2.3.2.3 Tổng hợp

ahihi

2.4 Computerized Adaptive Testing

2.5 Các cơ sở lý thuyết về công nghệ sử dụng

2.5.1 Firebase

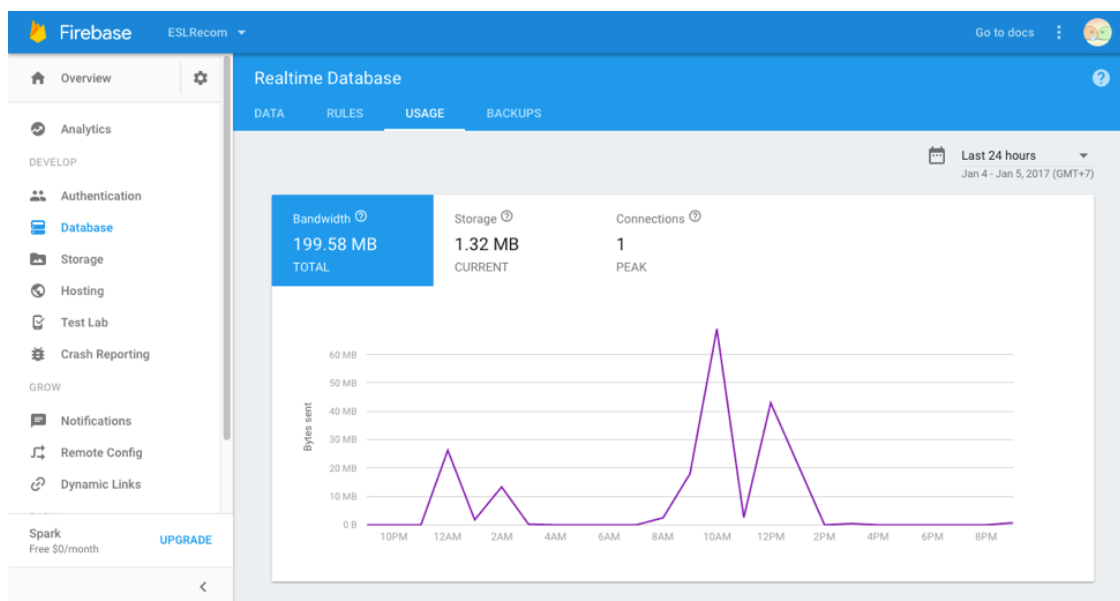
Firebase là một dịch vụ cơ sở dữ liệu thời gian thực hoạt động trên nền tảng đám mây được cung cấp bởi Google nhằm giúp các lập trình phát triển nhanh các ứng dụng bằng cách đơn giản hóa các thao tác với cơ sở dữ liệu.

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

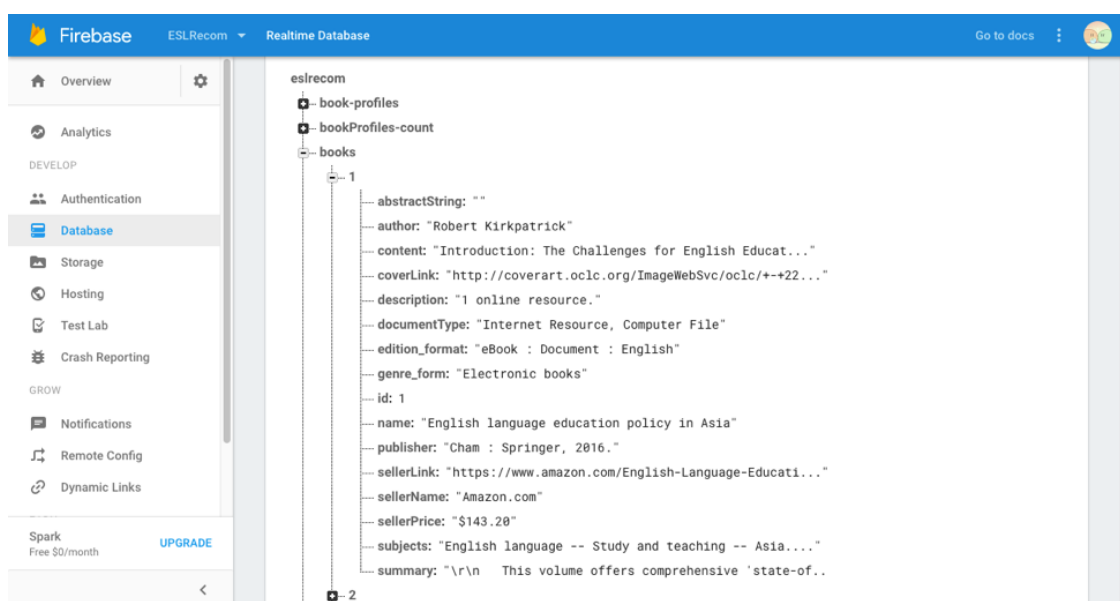
giảm thiểu công việc, và tập trung nâng cao trải nghiệm người dùng.

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.



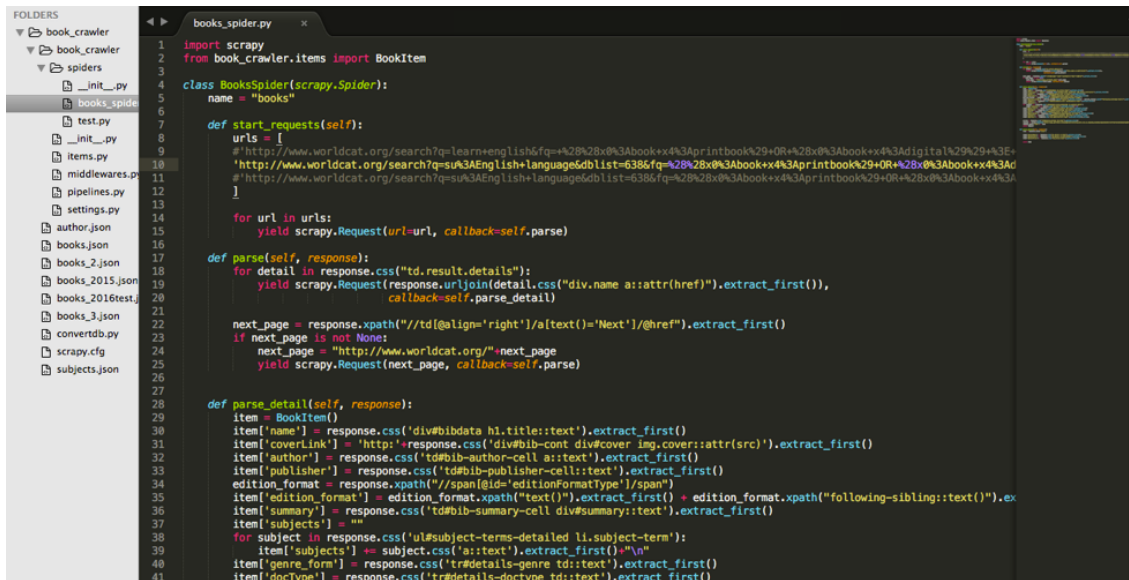
Hình 2.4: Cơ sở dữ liệu thời gian thực



Hình 2.5: Lưu trữ cơ sở tri thức chia sẻ giữa các thiết bị

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 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 2.6: 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:

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

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 để 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 tùy 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

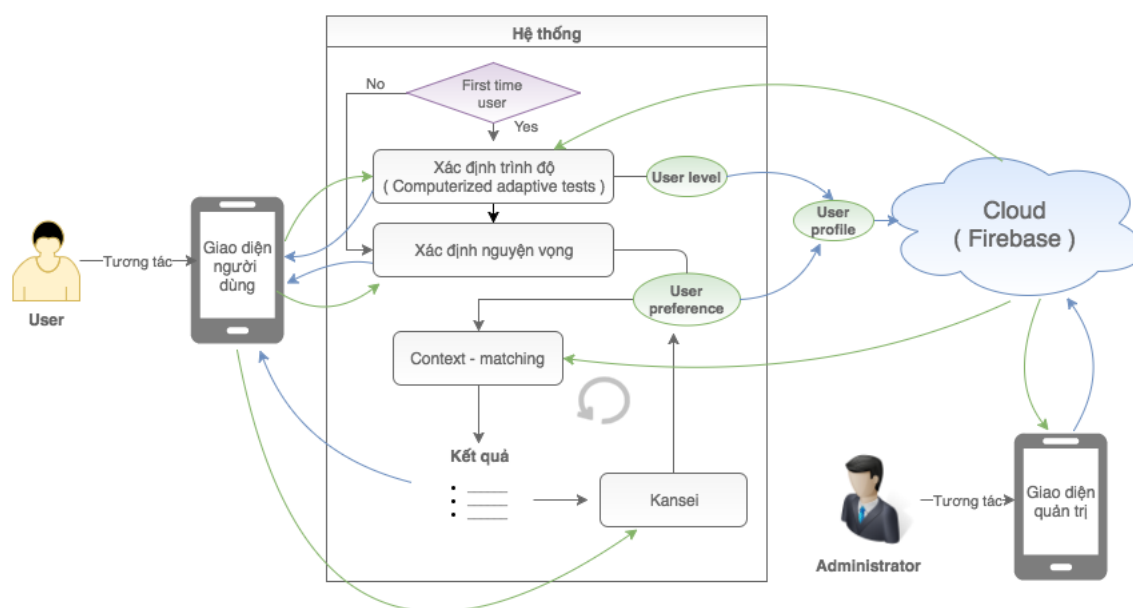
gợi ý cho người dùng về các chủ đề phổ biến. Kết quả bước này sẽ cho ra User preference là các chủ đề người dùng muốn học dưới dạng keyword.

- **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:**
- **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 thể hiện 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 .

Một user profile sẽ gồm có các thuộc tính sau:

- Thời điểm bắt đầu học
- Trình độ nghe hiểu
- Trình độ đọc hiểu
- Trình độ nói
- Trình độ viết văn
- Mục đích học của người dùng

Hệ thống sẽ đưa ra tập câu hỏi với các đáp án cho sẵn để xác định profile người dùng :

- Bạn học tiếng Anh được bao lâu rồi ?
- Trình độ đọc hiểu của bạn thế nào ?
- Bạn viết tiếng Anh tốt đến mức nào ?
- Bạn giao tiếp bằng tiếng Anh có tốt không ?
- Khả năng nghe tiếng Anh của bạn như thế nào ?
- Bạn mong muốn được cải thiện những kĩ năng nào ?

Câu hỏi 1-5 cho ta trình độ của người dùng, mỗi câu hỏi có 5 đáp án tương ứng với các mức: mới học, cơ bản, trung cấp, nâng cao, thành thạo. Dựa vào câu trả lời của người dùng, ta xác định được thuộc tính trong user profile về các kĩ năng đọc, nghe, viết, nói.

Thêm vào đó, để đánh giá chung năng lực của người dùng. Mỗi đáp án được gán với 1 mức điểm đánh giá năng lực tương ứng : a - 0, b - 0.3, c - 0.6, d - 0.85, e - 1. Gọi d là điểm đánh giá của từng câu hỏi, ta có $d = \text{điểm đánh giá đáp án} \times \text{trọng số câu hỏi}$. Từ đây ta tính được điểm đánh giá trình độ của người dùng bằng tổng $d_1+d_2+d_3+d_4+d_5$. Trình độ người dùng sẽ được phân loại vào các mức: mới học (điểm=0), sơ cấp (điểm \leq 0.3), trung cấp(điểm \leq 0.6), cao cấp(điểm \leq 0.85) và thành thạo(điểm \leq 1)

Vd. Anh A học tiếng Anh được trên 5 năm, anh A có thể đọc được tài liệu tiếng Anh chuyên ngành, viết được đoạn văn ngắn bằng tiếng Anh, nghe hiểu hội thoại tự

nhien trong cuộc sống và giao tiếp như người bản xứ. Điểm trình độ của anh A sẽ là : $0.8*0.1 + 1*0.225+0.6*0.225+0.85*0.225+1*0.225 = 0.85625$. => Trình độ của anh A nằm ở mức độ thành thạo.

Câu hỏi cuối cho ta nguyện vọng học của người dùng. Nguyện vọng sẽ ảnh hưởng đến việc chọn cuốn sách dựa trên kĩ năng mà nó cung cấp.

3.3 Xây dựng mạng từ khoá

Chương 4

Kết quả thực hiện

Chương 5

Kết luận và hướng phát triển

Tài liệu tham khảo