Xây Dựng Hệ Khuyến Nghị Cho Diễn Đàn Các Câu Hỏi Về Nghề Nghiệp

Recommender Systems for Q&A Forum

Trần Thị Mỹ Linh[1,2], Dương Thị Hồng Hạnh[1,2], and Nguyễn Trọng Ân[1,2,]
Đỗ Trọng Hợp[1,2,]
18520999@gm.uit.edu.vn, 18520711@gm.uit.edu.vn, 18520434@gm.uit.edu.vn,
hopdt@uit.edu.vn

[1] University of Information Technology, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam

Tóm tắt. Trong bài báo cáo này, chúng em hành xây dựng hệ khuyến nghị chuyên gia cho việc trả lời các câu hỏi của người dùng trên trang careervillage.org sử dụng bộ dữ liệu lấy từ cuộc thi Data Science for Good: CareerVillage.org được diễn ra trên Kaggle vào năm 2019. Chúng em tiến hành xử lý, phân tích bộ dữ liệu sau đó tiến đề ra các giải pháp và bài toán khác trên bộ dữ liệu này đồng thời áp dụng các phương pháp đã học được để thực nghiệm và đánh giá theo 3 phương pháp: Content-based Filtering, Collaborative Filtering và Hybrid Recommendation. Cuối cùng là kết quả thu được cùng với hướng phát triển trong tương lai.

1 Giới thiệu

CareerVillage.org là một tổ chức phi lợi nhuận được xây dựng với mong muốn trở thành Wikipedia trong lĩnh vực tư vấn về nghề nghiệp nhưng với nhiều sự nhiệt tình và thấu hiểu hơn. Sứ mệnh đặt ra của họ là trở thành nguồn thông tin nghề nghiệp đáng tin cậy cho thanh thiếu niên trên toàn cầu.

Và họ đã làm điều đó bằng cách xây dựng một diễn đàn hỏi đáp tương tự như StackOverflow hoặc Quora. Nơi các thanh thiếu niên sẽ nhận được các câu trả lời ngay lập tức, theo yêu cầu cho mọi câu hỏi về mọi nghề nghiệp.

Nền tảng này sẽ cho phép bạn tham gia với những vai trò như Student, Professional.Trong đó, Student là những thanh thiếu niên, sinh viên đến với CareerVillage.org với mong muốn được giải đáp những câu hỏi đặt ra và Professional là những tình nguyện viên đăng kí để trở thành những tư vấn viên, giải đáp những câu hỏi thuộc lĩnh vực chuyên môn của mình.

Tuy nhiên để có thể xây dựng nên một nền tảng hỏi đáp hoàn hảo, họ cần xây dựng một hệ thống đề xuất (Recommender System) các câu hỏi đến với các chuyên gia phù hợp nhất để có thể giúp sinh viên nhận được lời khuyên cần thiết, chất lượng và nhanh chóng nhất. Và để giải quyết vấn đề này, CareerVillage.org đã mang đến Data Science for Good challenge với 5 năm dữ liệu của họ nhằm tìm kiếm các giải pháp tốt nhất cho tổ chức của mình.

Nhận thấy những điều thú vị từ dữ liệu được cung cấp cũng như những mục đích vì cộng đồng mà tổ chức này hướng đến, chúng em đã quyết định sử dụng những dữ liệu

được cung cấp từ CareerVillage.org và áp dụng các kiến thức chuyên môn mà mình đã học được với mong muốn xây dựng nên một hệ thống đề xuất câu hỏi giúp ích được cho tổ chức này.

Trong báo cáo đồ án này, chúng em sẽ tập trung vào giới thiệu các thông tin liên quan đến mục tiêu xây dựng bài toán đề xuất. Trong mục 2, chúng em sẽ trình bày một số công trình nghiên cứu liên quan. Tiếp theo ở mục 3, chúng em trình bày chi tiết về quá trình xây dựng bộ dữ liệu. Trong mục 4, các giải pháp, mô hình được chúng em trình bày và đồng thời, kết quả thử nghiệm sẽ được đánh giá, phân tích ở mục 5. Cuối cùng, mục 6 sẽ là kết luận và hướng phát triển trong tương lai cho các bài toán hỏi đáp nói chung và các bài toán hỏi đáp tự động trên ảnh nói riêng.

2 Bài toán và Bộ dữ liệu

2.1 Mô tả bài toán

Mục tiêu chính của đồ án này là xây dựng hệ khuyến nghị cho trang web các câu hỏi về nghề nghiệp, với 2 đối tượng chính là các chuyên gia và sinh viên. Với từng đối tượng, chúng em đề xuất những bài toán khác nhau.

2.1.1 Bài toán 1: Đối với sinh viên

- Đầu vào: Câu hỏi về nghề nghiệp, học tập mà sinh viên đặt ra.
- Đầu ra: Có 2 loại đầu ra mà chúng em hướng tới:
  - Đề xuất câu hỏi, chủ đề liên quan.
  - Đề xuất những chuyên gia có khả năng cao sẽ trả lời tốt những câu hỏi đó để sinh viên có thể liên lạc, trao đổi trực tiếp.

2.1.2 Bài toán 2: Đối với chuyên gia

- Đầu vào: Thông tin (id) của chuyên gia.
- Đầu ra: Đề xuất những câu hỏi mà chuyên gia này có khả năng có thể trả lời được.

2.2 Bộ dữ liệu

Bộ dữ liệu được CareerVillage.org cung cấp chứa thông tin hoạt động của họ trong vòng 5 năm, gồm 15 file Comma-Separated Values (csv): professionals, matches, comments, tag_users, groups, school_memberships, group_memberships, answers, emails, questions, tags, tag_questions, answer_scores, question_scores, students.

Các file csv này chứa thông tin của 30971 tài khoản sinh viên, 28152 tài khoản chuyên gia cùng 23931 câu hỏi và 51123 câu trả lời tương ứng và một số thông tin liên quan như các tag, nội dung bình luận, nhận xét cũng như điểm đánh giá cho từng câu hỏi và câu trả lời trên diễn đàn . Dưới đây là sơ đồ dữ liệu và quan hệ giữa chúng.

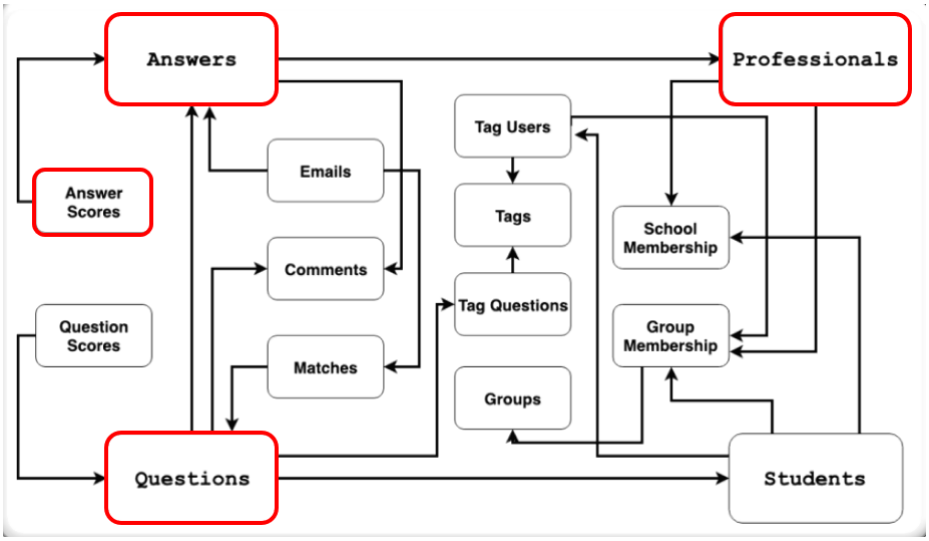

Fig. 1: Sơ đồ dữ liệu và quan hệ giữa chúng

Thực hiện kết hợp các file dữ liệu dựa trên các thuộc tính khóa để trích xuất dữ liệu cho quá trình xây dựng hệ khuyến nghị. Cụ thể chúng em đã các thuộc tính được lấy từ các file chính gồm professionals, answers, questions, answer_scores. Tất cả được thể hiện chi tiết trong bảng bên dưới.

Table 1: Các thuộc tính sử dụng trong quá trình xây dựng hê khuyến nghị.

Thuộc tính	**Vị trí**	**Ý nghĩa**
questions_id	questions	Id của câu hỏi.
questions_title		Tiêu đề của câu hỏi.
questions_body		Nội dung câu hỏi.
answers_id	answers	Id của câu trả lời tương ứng.
answers_author_id (professionals_id)		Id chuyên gia của câu trả lời.
answers_body		Nội dung câu trả lời
scores	answer_scores	Điểm đánh giá của câu trả lời.
professionals_location	professionals	Thông tin của chuyên gia trả lời câu hỏi : Địa chỉ, chuyên ngành, mục nổi bật, ngày tham gia.
professionals_industry		
professionals_headline		
professionals_date_joined		

Xem chi tiết về dữ liệu tại đây.

2.3 Phân tích dữ liệu

2.3.1 Dữ liệu rỗng

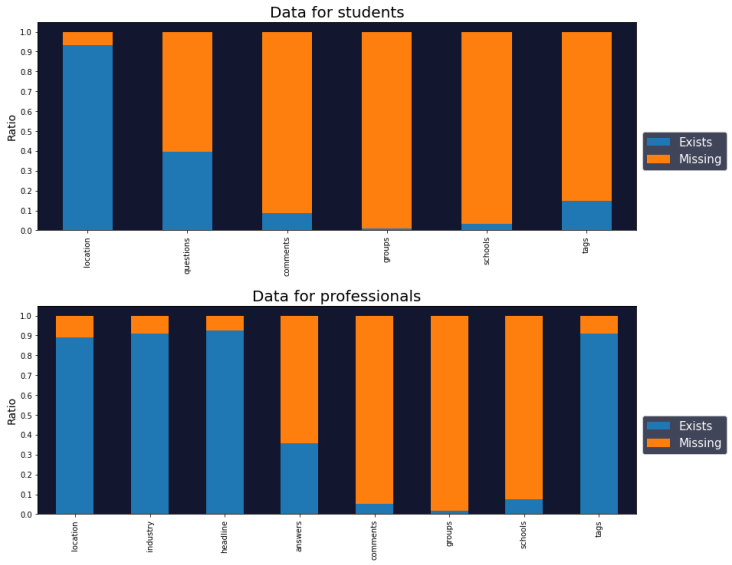

Fig. 2: Biểu diễn tỉ lệ dữ liệu rỗng trên bộ dữ liệu

Như **Hình 2** ta có thể thấy được rằng tỉ lệ rỗng trên dữ liệu cho sinh viên cho thuộc tín: "comments", "groups", "schools" và "tags" là khá lớn trên 90% tương tự với dữ liệu của các chuyên gia là các thuộc tính: "comments", "groups" và "schools" cho thấy sự ảnh hưởng của các thuộc tính này là không lớn có thể xem xét bỏ qua.

2.3.2 Hashtags

Hình dưới mô tả các chủ đề được nhắc đến nhiều trong bộ dữ liệu dựa trên các hashtags được nhắc đến

Fig. 3: Hashtags

2.3.3 Thông tin về câu hỏi và câu trả lời

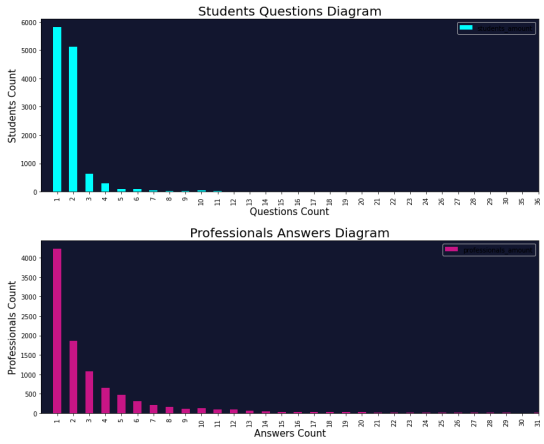

Fig. 4: Số lượng câu hỏi trung bình trên mỗi tài khoản

Phần lớn số câu hỏi của học sinh trên 1 tài khoản dao động trong khoản từ 0 đến 10 mà cao nhất là ở 1 và 2 bởi vì đa số sinh viên tham gia chỉ để hỏi vấn đề của bản thân. Còn về phần câu trả lời của chuyên gia đều hơn từ 1 cho đến 20 mà cao nhất trong đoạn từ 1 đến 10 cho thấy các chuyên gia hoặc tình nguyện viên rất tích cực trong việc trả lời.

2.3.4 Topics

Các chủ đề được hỏi trên CareerVillage rất đa dạng với 7091 chủ đề được xác định thông qua các tag_question. Bên dưới là ảnh thống kê các chủ đề được hỏi nhiều nhất dữ liệu được cung cấp.

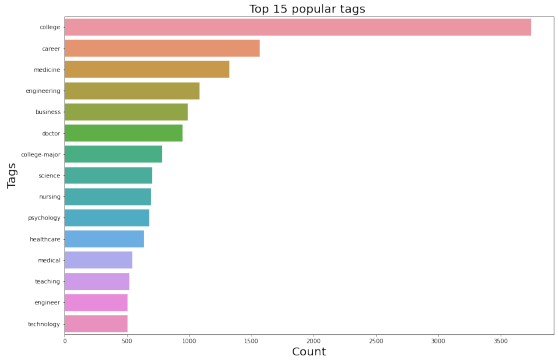

Fig. 5: Top 15 các chủ đề phổ biến nhất được hỏi trên CareerVillage.

3 Phương Pháp

Để giải quyết các bài toán đặt ra, chúng em tiếp cận theo 2 hướng chính: Content-based Filtering và Collaborative Filtering. Sau đó kết hợp chúng tạo thành mô hình Hybrid Recommendation. Mỗi phương pháp chúng em sẽ tiến hành các bước tiền xử lý và xâu dựng hệ khuyến nghị cho phù hợp. Cụ thể như sau:

3.1 Content-based Filtering

Đề xuất được đưa ra dựa vào hồ sơ của người dùng hoặc dựa vào nội dung/thuộc tính của những item tương tự như item mà người dùng đã chọn trong quá khứ.

3.1.1 Tiền xử lý dữ liệu

- Feature Generations: Tạo ra các thuộc tính mới từ những thuộc tính cũ.
  - questions_body_final: thuộc tính được tổng hợp từ lịch sử trả lời câu hỏi của chuyên gia.
- Loại bỏ các thẻ html, url,..
- Xóa bỏ các kí tự đặc biệt (vd: dấu @, dấu #,...)
- Xây dựng pipeline chuyển dữ liệu dạng text về vector:
  - Tách từ: sử dụng RegexTokenizer
  - Loại bỏ stopword
  - Chuyển đổi text thành vector thông qua Word Embedding.

3.1.2 Word embedding

- Word embedding là một kỹ thuật trong Xử lý ngôn ngữ tự nhiên (NLP), bằng cách ánh xạ các từ hoặc cụm từ từ nhóm từ vựng thành các vector số thực. Nó giúp cải thiện độ chính xác của các mô hình ngôn ngữ tự nhiên khác nhau.
- Có nhiều kỹ thuật embedding khác nhau, qua thực nghiệm, nhóm chúng em lựa chọn Word2Vec được tích hợp sẵn trong pyspark.

3.1.3 Cosine similarity Để xây dựng hệ khuyến nghị theo phương pháp Content-based Filtering, chúng em sử dụng 1 độ đo là Cosine similarity [1], tính toán độ tương tự giữa 2 vector, có công thức như sau:

$$similarity = \cos(\theta) = \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1}$$

3.1.4 Phương thức xây dựng hệ khuyến nghị

- Bài toán 1 (Đối với sinh viên):
  - Hướng 1:
    - Tính toán độ tương đồng giữa câu hỏi đầu vào, và các câu hỏi có sẵn trên hệ thống, đề xuất cho sinh viên top 20 những câu hỏi có độ tương đồng cao.

    - Tiếp theo, từ những câu hỏi được đề xuất ở trên, tìm ra những chuyên gia có câu trả lời tốt trong top 20 và đề xuất những chuyên gia đó cho sinh viên.
    - Trọng số chuyên gia là kết quả giữa độ tương đồng câu hỏi nhân với số lượt thích cho câu trả lời(answer_score).
  - Hướng 2:
    - Xây dựng thuộc tính mới là questions_body_final cho mỗi chuyên gia, sau đó tiến hành tính toán độ tương đồng giữa thuộc tính này với câu hỏi đầu vào, lựa chọn những chuyên gia có độ tương đồng cao trong top 20 để đề xuất cho sinh viên.
- Bài toán 2 (Đối với chuyên gia): Xây dựng thuộc tính mới là questions_body_final cho chuyên gia đầu vào, sau đó tiến hành tính toán độ tương đồng giữa thuộc tính này với tập câu hỏi, lựa chọn những câu hỏi có độ tương đồng cao trong top 20 để đề xuất cho chuyên gia.

3.2 Collaborative Filtering

Đưa ra đề xuất dựa trên những items mà người dùng có khả năng ưa thích nhất dựa vào những người dùng có hành vi tương tự.

3.2.1 Tiền xử lý dữ liệu

- Feature Generations: Tạo ra các thuộc tính mới từ những thuộc tính cũ.
  - check: thuộc tính được tạo thêm dựa trên lịch sử trả lời câu hỏi của chuyên gia, nếu chuyên gia có trả lời câu hỏi đó thì check=1, ngược lại check=0
- Chuyển professionals_id, questions_id từ dạng chuỗi riêng biệt thành các số riêng biệt để phù hợp với đầu vào của mô hình.

3.2.2 Mô hình Ở phần này, chúng em tiến hành thực nghiệm trên mô hình ALS (Alternating Least Square).

- ALS là một phương pháp collaborative filtering dựa trên một phép phân rã ma trận (matrix fractorization). Lúc này ma trận ban đầu sẽ được phân tích thành tích của các ma trận items và ma trận users. Thuật toán này yêu cầu ta phải thực hiện tối ưu đồng thời cả 2 ma trận users và ma trận items dựa trên hàm mất mát. Khi cần tối ưu ma trận users ta sẽ cố định ma trận items và dịch chuyển theo phương gradient descent đạo hàm của và ngược lại.
- ALS được triển khải trong Spark với các tham số cần chú ý như sau:
  - numBlocks: số lượng block được sử dụng trong tính toán song song
  - rank: số lượng nhân tố ẩn (latent factor) trong mô hình
  - iterations: số lần lặp
  - lambda: tham số của chuẩn hoá (regularization) trong ALS

3.2.3 Phương thức xây dựng hệ khuyến nghị

- Hướng 1: Sử dụng 3 thuộc tính đầu vào là professionals_id, questions_id và check.
  - Đánh giá và lựa chọn mô hình dựa trên accuracy với labelCol là thuộc tính check và độ đo chúng em đề xuất ở phần 4.1.

- Chọn mô hình tốt nhất để kết hợp trong hybrid.

– Hướng 2: Sử dụng 3 thuộc tính đầu vào là professionals_id, questions_id và score.
  - score: là số điểm mà chuyên gia được vote khi trả lời câu hỏi tương ứng.
  - Đánh giá và lựa chọn mô hình dựa trên RMSE với labelCol là thuộc tính score và độ đo chúng em đề xuất ở phần 4.1.
  - Chọn mô hình tốt nhất để kết hợp trong hybrid.

3.3 Hybrid Recommendation

Là sự kết hợp của hai phương pháp là Content-based Filtering và Collaborative Filtering để cải thiện các hạn chế của các phương pháp đó.

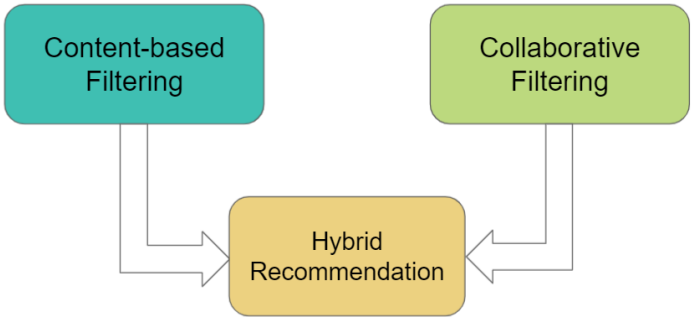

Fig. 6: Phương pháp Hybrid Recommendation

Thực hiện kết hợp dựa trên nhiều tỉ lệ kết hợp để đưa ra mô hình tốt nhất ví dụ 0.5:0.5, 0.75:0.25.

4 Thực nghiệm và Kết quả

4.1 Độ đo đánh giá

– Để dễ dàng trong quá trình đánh giá, so sánh giữa các phương pháp, chúng em đã định nghĩa một độ đo khác làm thước đo chung cho Content-based Filtering, Collaborative Filtering và Hybrid.
– Độ đo trên có thể hiểu là tỉ lệ đề xuất chính xác so với thực tế:
  - Giả sử với câu hỏi A, mô hình đề xuất 10 chuyên gia có khả năng trả lời câu hỏi. Trong thực tế có 8/10 chuyên gia thực sự trả lời thì độ chính xác là 0.8 .
  - Sau đó tính trung bình tất cả tỉ lệ này trên tập test chung, ta sẽ được kết quả cuối cùng.

4.2 Kết quả

– Content-based Filtering:

Table 2: Bảng kết quả thực nghiệm trên Content-based Filtering

	Content-based 1	**Content-based 2**
Question_body	0.4451	0.3322
Question_title	0.4321	0.2439
Question_full	0.3769	0.3326

▷ Nhận xét:

- Với từng loại đầu vào, hiệu suất hệ khuyến nghị xây dựng ở cả hai hướng đều cho kết quả khác nhau.
- Hệ khuyến nghị xây dựng theo hướng 1:
  - Mô hình mang lại kết quả tốt nhất với đầu vào chỉ chứa nội dung câu hỏi với 44.51% và thấp nhất với dữ liệu có cả nội dung và tiêu đề 37.69%.
  - Kết quả thu được với đầu vào là nội dung câu hỏi và tiêu đề câu hỏi mang lại kết quả không chênh lệch nhiều.
- Hệ khuyến nghị xây dựng theo hướng 2:
  - Hệ khuyến nghị mang lại kết quả tốt nhất với đầu vào được tổng hợp cả tiêu đề lẫn nội dung của câu hỏi với 33.26% và thấp nhất với đầu vào là tiêu đề 24.39%. Có thể thấy, khi đầu vào là tiêu đề, question_body_final sẽ tổng hợp thông tin các tiêu đề đã được chuyên gia đã giải đáp. Việc tính toán độ tương đồng giữa một tiêu đề ngắn so với một lượng lớn thông tin đa dạng như vậy sẽ khiến cho việc đề xuất trở nên không tốt.
  - Trường hợp đầu vào chứ nội dung câu hỏi và trường hợp chứa cả nội dung lẫn tiêu đề mang lại kết quả không mấy khác biệt (đều xấp xỉ 33.3%) do phần nội dung của chúng không khác biệt quá lớn.
- Kết quả thu được ở phương pháp 1 tốt hơn so với phương pháp 2. Do so với phương pháp 2 thì phương pháp 1 tiếp cận vấn đề với phạm vi gần hơn, trực tiếp hơn, những chuyên gia được đề xuất là những chuyên gia có khả năng rất cao đã có kinh nghiệm giải quyết vấn đề cụ thể đặt ra. Phương pháp 2 không mang lại kết quả tốt do question_body_final chứa thông tin quá đa dạng.

– Collaborative Filtering:

Table 3: Bảng kết quả thực nghiệm trên Collaborative Filtering

Tham số	Sử dụng thuộc tính **Score**		Sử dụng thuộc tính **Check**	
	RMSE	Our Measure	Accuracy	Our Measure
maxIter=02, rank=10	1.1360	0.7126	0.9756	0.8368
maxIter=15, rank=10	1.0745	0.7444		
maxIter=15, rank=40	1.0483	0.7926		

▷ Nhận xét:

- Lọc cộng tác với thuộc tính Score:
  - Các hệ khuyến nghị xây dựng với các tham số khác nhau mang lại các hiệu suất khác nhau.
  - Trong đó, mô hình mang lại kết quả tốt nhất là mô hình xây dựng với tham số maxIter=15, rank=40 với RMSE=1.0483 và Our measure=79.26%.
- Lọc cộng tác với thuộc tính Check: thu được kết quả khá cao với tỷ lệ đề xuất đúng là 83.68% và độ chính xác hơn 97%.
- Có thể thấy, với cùng tham số mô hình, trên cùng một độ đo đánh giá, mô hình khuyến nghị được xây dựng với thuộc tính Check mang lại kết quả tốt hơn so với mô hình được xây dựng với thuộc tính Score. Tuy nhiên, cần thêm nhiều dữ liệu hơn để có thể so sánh hiệu quả của 2 cách tiếp cận, do dữ liệu hiện tại của Score đang chênh lệch khá nhiều với Check.

– Hybrid Recommendation: Kết hợp 2 mô hình vừa xây dựng tạo thành bảng 4 kết quả như sau:

Table 4: Bảng kết quả thực nghiệm trên Hybrid Recommendation

Tỉ lệ kết hợp	Collaborative_Score		Collaborative_Check	
	Content_1	Content_2	Content_1	Content_2
10 content - 10 collab	0.7460	0.7534	0.7703	0.7787
05 content - 15 collab	0.8068	0.8180	**0.9109**	0.8792

▷ Nhận xét:
- Content-based filtering theo hướng 1 kết hợp với cả 2 mô hình Collaborative đều cho kết quả rốt hơn hướng 2. Trong đó tốt nhất là trên Collaborative với thuộc tính Check và Content_1.
- Khi chuyển đổi tỷ lê kết hợp giữa Content-based và Collaborative từ 10:10 sang 5:15 mang lại kết quả khả quan hơn ở tất cả trường hợp, vì phương pháp Collaborative đạt hiệu suất vượt trội hơn hẳn Content-based.
- Phương pháp kết hợp giữa Collaborative_Check và Content_1 với tỷ lệ 5:15 mang lại kết quả tốt nhất với hiệu suất 91.09%.

5 Kết luận và Hướng phát triển

5.1 Kết luận

– Kết quả tốt nhất thu được với các phương pháp xây dựng hệ khuyến nghị Content-based Filtering, Collaborative Filtering và Hybrid Recommendation lần lượt là 44.51%,83.68%, 91.09%.

– Hạn chế của phương pháp:
- Content-based Filtering: Chỉ dựa trên độ tương quan của câu hỏi mà không liên quan đến thông tin của chuyên gia sẽ làm mất sự đa dạng trong việc tìm các chuyên gia (có những chuyên gia cùng thuộc chủ đề đó nhưng chưa trả lời câu hỏi như vậy) .
- Collaborative Filtering: Các chuyên gia chưa có câu trả lời nào hoặc là chuyên gia mới trên trang chủ CareerVillage.org sẽ không được giới thiệu bởi hệ thống.

$\Longrightarrow$ Việc kết hợp cả hai phương pháp Content-based và Collaborative góp phần giải quyết những hạn chế gặp phải và mang lại kết quả tốt nhất.

5.2 Hướng phát triển

- Thử tìm hiểu và crawl dữ liệu ở các diễn đàn ở Việt Nam và thực nghiệm phương pháp nhóm đã tiến hành.
- Tiếp tục tìm hiểu nhiều hơn để có thể khai thác hết tìm năng từ toàn bộ dữ liệu được cung câp nhằm xây dựng một hệ khuyến nghị hoàn hảo hơn.
- Áp dụng triển khai các phương pháp học sâu.
- Tìm hiểu các hướng giải quyết mới cho bài toán.

6 Nội dung bổ sung

- Ở phần bài toán 1 theo hướng 1, đã thay đổi dựa trên trọng số.
- Với phương pháp Collaborative Filtering thực nghiệm thêm phương pháp bằng cách sử dụng thuộc tính score và thực nghiệm trên nhiều bộ tham số để chọn ra mô hình tốt nhất.
- Kết hợp thêm Collaborative Filtering trên vào Hybrid để so sánh đánh giá.

References

1. Faisal Rahutomo, Teruaki Kitasuka, and Masayoshi Aritsugi. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, volume 4, page 1, 2012.