PHÁT TRIỂN HỆ THỐNG KHUYẾN NGHỊ KỸ NĂNG CẦN THIẾT CHO SINH VIÊN ĐẠI HỌC DỰA TRÊN NHU CẦU TUYỂN DỤNG

Nguyễn Thái Anh*

Phan Hồ Viết Trường*

TÓM TẮT

Hệ thống giới thiệu đề xuất các sản phẩm và dịch vụ theo sở thích của người dùng dựa trên lịch sử mua hàng. Chúng hầu hết được áp dụng trong lĩnh vực thương mại điện tử, phim ảnh, tin tức và việc làm. Tuy nhiên, rất nhiều nghiên cứu sử dụng hệ thống thư giới thiệu để giới thiệu các kỹ năng cần thiết cho sinh viên trước khi tốt nghiệp nhằm đáp ứng nhu cầu tuyển dụng của nhà tuyển dụng và cạnh tranh với các ứng viên khác. Bài báo này đề xuất một phương pháp để xây dựng một hệ thống khuyến nghị kỹ năng cần thiết cho sinh viên ngành công nghệ thông tin bằng cách kết hợp học tập không giám sát và trích xuất cụm từ khóa. Cụ thể, nghiên cứu này đã sử dụng kỹ thuật phân cụm để nhóm các công việc mà người sử dụng lao động muốn thuê thành hai nhóm chăm sóc viên trong lĩnh vực tin học như khoa học dữ liệu và kỹ thuật phần mềm. Sau đó, mô hình đo lường mức độ tương đồng giữa các công việc và đối tượng mà sinh viên đã vượt qua để chọn ra 5 công ty phù hợp nhất. Cuối cùng, người mẫu lần lượt rút ra các cụm từ khóa kỹ năng cần thiết từ các tin tuyển dụng của 5 công ty này để giới thiệu cho sinh viên. Dữ liệu kiểm tra được thu thập từ bảng điểm của khoảng 500 sinh viên năm cuối và năm thứ ba theo học ngành công nghệ thông tin tại Đại học Văn Lang, Việt Nam. Kết quả là, phương pháp đề xuất của chúng tôi có hiệu suất tốt hơn các phương pháp tương tự phổ biến như Jaccard, Cosine, TF-IDF và word2vec.

Từ khóa: *đo lường độ tương tự, hệ thống đề xuất, nhúng từ, Fasttext, Yake*

1. Giới thiệu

Hệ thống khuyến nghị là hệ thống đề xuất các sản phẩm và dịch vụ phù hợp với hành vi của người dùng. Hệ thống này mang đến cho người dùng nhiều sự lựa chọn hơn trong việc quyết định mua sản phẩm nào. Hệ thống khuyến nghị được áp dụng rộng rãi trong nhiều lĩnh vực khác nhau như âm nhạc, tin tức, giáo dục, trả lời câu hỏi và thương mại điện tử (Raza, 2021; Nguyen & Phan, 2021; Phuc Do và cộng sự, 2021; Fernandez và cộng sự, 2014). Tuy nhiên, có rất nhiều nghiên cứu sử dụng hệ thống giới thiệu trong tuyển dụng. Sinh viên thường hy vọng rằng họ có thể tìm được một công việc phù hợp sau khi tốt

* Khoa Công nghệ Thông tin, Trường Kỹ Thuật và Công Nghệ Văn Lang, Đại học Văn Lang, Thành phố Hồ Chí Minh

nghiệp. Tuy nhiên, nếu sinh viên đợi đến khi ra trường mới tìm được việc làm thì họ khó có thể đáp ứng được một số yêu cầu từ nhà tuyển dụng. Ví dụ, nhà tuyển dụng cần tuyển lập trình viên đã có kinh nghiệm sử dụng HBase, Redis, máy tính phân tán nhưng sinh viên chỉ học các môn như MS SQL Server, ngôn ngữ C, Java, Python, Lập trình web, Lập trình di động. Do đó, sinh viên khó tìm được việc làm tốt và cạnh tranh với các ứng viên khác.

Tuy nhiên, có rất nhiều nghiên cứu sử dụng thuật toán phân cụm trong các lĩnh vực giáo dục hoặc tìm kiếm việc làm. Việc phân cụm dữ liệu vào lĩnh vực tìm việc sẽ giúp mọi người dễ dàng tìm được công việc phù hợp với chuyên ngành và sở thích của mình. Nghiên cứu này sử dụng thuật toán phân cụm để phân loại thông tin tuyển dụng sao cho phù hợp với hai chuyên ngành Kỹ thuật phần mềm và Khoa học dữ liệu. Từ đó, tác giả có thể dễ dàng xác định được nội dung tuyển dụng của từng chuyên ngành thay vì phải mở từng mẫu thông tin tuyển dụng để đọc, sẽ rất tiết kiệm thời gian. Thông thường, chúng tôi sẽ phân cụm theo số, phân nhóm khách hàng và dữ liệu thống kê. Trong bài báo này, chúng tôi đã sử dụng một thuật toán để có thể phân cụm các tài liệu có điểm tương đồng gần giống nhau, sao cho phù hợp với hầu hết các chuyên ngành.

Gần đây, việc trích xuất từ khóa là một việc rất quan trọng với lượng thông tin bùng nổ. Bài toán trích xuất từ khóa đã giúp giải quyết nhiều vấn đề trong thực tế như tìm kiếm thông tin, tổng hợp tài liệu (Chengyu Sun, 2020; Nakul Sharma, 2021). Nhiều người cần tổng hợp và tóm tắt thông tin để tiện theo dõi thông tin như vậy. Trong bài viết này, chúng tôi đã sử dụng Yake để tự động trích xuất các từ khóa / thuật ngữ là từ / cụm từ tiêu biểu, thể hiện những kỹ năng mà sinh viên CNTT còn thiếu so với nhu cầu tuyển dụng của doanh nghiệp. Qua đó, học viên sẽ nhận ra những kỹ năng còn thiếu sót để sớm trang bị cho phù hợp với nhu cầu của doanh nghiệp.

Nghiên cứu này đề xuất phương pháp xây dựng hệ thống khuyến nghị kỹ năng cần thiết cho sinh viên năm thứ ba và năm cuối theo khả năng học tập của họ. Cụ thể, nghiên cứu thực hiện ba giai đoạn: (1) sử dụng kỹ thuật phân cụm để nhóm các công ty có nội dung tuyển dụng tương tự cho từng chuyên ngành; (2) Sử dụng phương pháp lọc nội dung để đo điểm tương đồng giữa các môn chuyên ngành mà thí sinh đạt điểm cao với yêu cầu của nhà tuyển dụng; (3) sử dụng kỹ thuật trích xuất văn bản để trích xuất từ khóa từ nội dung tuyển dụng của các công ty không phù hợp với kỹ năng của sinh viên. Những đóng góp chính của nghiên cứu này như sau:

- Sử dụng các kỹ thuật cụm để nhóm nhu cầu của người sử dụng lao động theo các lĩnh vực của họ như khoa học dữ liệu và kỹ thuật phần mềm.
- Sử dụng sự giống nhau trong văn bản để chỉ ra các công ty phù hợp với kỹ năng của sinh viên.
- Sử dụng trích xuất cụm từ khóa để giới thiệu những kỹ năng còn thiếu mà sinh viên nên chuẩn bị trước khi tốt nghiệp.

2. Tổng quan nghiên cứu

Hiện tại, các hệ thống khuyến nghị có ba cách tiếp cận phổ biến là lọc cộng tác, lọc nội dung và hệ thống kết hợp (Fayyaz và cộng sự, 2020).

A. Lọc cộng tác

Lọc cộng tác dựa trên công việc thu thập và phân tích một số lượng lớn các hành vi và hoạt động của người dùng để dự đoán mức độ tương tự sở thích của họ với những người khác (Salloum & Rajamanthri, 2021; Phan & Do, 2020; ThoratR và cộng sự, 2015; Mohamed và cộng sự, 2019; Bustos López và cộng sự, 2020). Osman và cộng sự (2021) đề xuất sự kết hợp giữa thông tin theo ngữ cảnh và phân tích cảm tính để cải thiện hệ thống khuyến nghị về sự không rõ ràng của từ ngữ, độ thưa thớt của dữ liệu và tỷ lệ sai sót. Họ đã sử dụng các bài đánh giá dạng văn bản được tích hợp vào ma trận mục người dùng để giảm bớt sự thưa thớt của dữ liệu. Fernandez và cộng sự (2015) đề xuất thuật toán Slope One để dự đoán hồ sơ cá nhân và sử dụng chiến lược thực dụng nhiều lần để giới thiệu phim cho một nhóm. Thuật toán đó nhằm mục đích đề xuất phim cho các nhóm có các thành viên có cùng sở thích và họ không có hồ sơ lịch sử. Cách tiếp cận này không sử dụng máy học nhưng có thể đề xuất chính xác các sản phẩm khác mà người dùng quan tâm. Vì vậy, khi ma trận đánh giá khá lớn, tức là số lượng người dùng và sản phẩm cũng lớn thì thời gian tính toán sẽ tăng lên. Phương pháp này khó có thể giới thiệu sản phẩm theo thời gian thực hoặc gần thời gian thực. Đặc biệt, khi người dùng có những đánh giá trái ngược nhau. Ví dụ, người dùng bình chọn một sản phẩm yêu thích nhưng họ không thích nó, vì vậy không thể giới thiệu chính xác sản phẩm cho những người này.

B. Lọc nội dung

Hệ thống này dựa trên việc rút các mô tả mặt hàng và hồ sơ về các lựa chọn của người dùng trong quá khứ (Nguyen & Phan, 2021). Từ đó, hệ thống so sánh sự giống nhau về nội dung để chọn ra những điểm chung nhất cho một gợi ý. Ví dụ: sử dụng tính năng lọc nội dung trong giáo dục nhằm chấm điểm một kỳ thi lập trình. Nếu hai bài kiểm tra có nhiều điểm giống nhau, chúng tôi kết luận rằng sinh viên đã sao chép bài làm của một bạn khác. (Son & Kim, 2017) đã sử dụng một mạng đa thuộc tính để tính toán mối tương quan giữa các mục. Họ đo lường các mục tương tự thông qua các liên kết. Cách tiếp cận này đảm bảo sự độc lập giữa những người dùng và giải quyết hầu hết các vấn đề khởi động nguội cho người dùng mới hoặc sản phẩm mới. Nhưng cách tiếp cận phải phân tích và bóc tách nội dung sản phẩm. Điều này đã dẫn đến hệ thống giới thiệu sản phẩm quen thuộc. Đồng thời, hệ thống gặp khó khăn khi giới thiệu sản phẩm mới cho người dùng mới.

C. Hệ thống lai

Đây là sự kết hợp của cả lọc dựa trên nội dung và lọc cộng tác. Phương pháp này nhằm giải quyết vấn đề khởi động nguội và đánh giá thưa (Fortes và cộng sự, 2017; Ibrahim & Saidu, 2020). Shruthi và Gripsy (2017) đề xuất một hệ thống kết hợp sử dụng lọc cộng

tác và phân tích nhân khẩu học. Bộ lọc cộng tác được sử dụng để tìm các mục có liên quan cho một mục nhất định. Phân tích nhân khẩu học được sử dụng để hiểu các đặc điểm của người dùng như độ tuổi, giới tính và thành phần chủng tộc của dân số.

Phần còn lại của bài báo này được tổ chức như sau: Trong phần 3, phương pháp luận, chúng tôi đã trình bày chi tiết hơn về phương pháp đề xuất của chúng tôi. Phương pháp này sử dụng ba kỹ thuật như phân cụm văn bản, tương tự văn bản và trích xuất cụm từ khóa. Trong phần 4, trải nghiệm, chúng tôi đưa ra kết quả mà chúng tôi thực hiện để so sánh với các phương pháp hiện đại khác. Chúng tôi liệt kê hai ví dụ đại diện cho hai chuyên ngành công nghệ thông tin và áp dụng so sánh giữa các phương pháp tương tự văn bản. Trong phần 5, kết luận, chúng tôi đã tóm tắt những gì chúng tôi đã đề xuất và triển khai cho một ứng dụng thực tế trong giáo dục. Ngoài ra, chúng tôi đã chỉ ra công việc mà chúng tôi cần làm trong tương lai để cải thiện sự

3. Phương pháp nghiên cứu

Trong phần này, chúng tôi đã trình bày phương pháp đề xuất các kỹ năng cần thiết cho sinh viên để phù hợp với nhu cầu của nhà tuyển dụng, chẳng hạn như phân nhóm dựa trên công việc, đo lường mức độ tương đồng giữa khả năng của sinh viên và nhu cầu của nhà tuyển dụng, đồng thời rút ra các kỹ năng cần thiết cho sinh viên.

Phân nhóm dựa trên công việc

Phân cụm là quá trình nhóm dữ liệu hoặc đối tượng thành các cụm sao cho các đối tượng trong cùng một cụm giống với nhau hơn các đối tượng trong các cụm khác. Trong nghiên cứu này, các tác giả đã sử dụng kỹ thuật phân cụm để phân loại 10.000 công việc thành 2 cụm kỹ thuật phần mềm và khoa học dữ liệu. Nội dung tuyển dụng ở mỗi cụm có điểm giống nhau nhất. Qua đó, chúng tôi dễ dàng có được dữ liệu của từng công ty cho từng chuyên ngành. Đầu tiên, chúng tôi thu thập thông tin khoảng 10.000 việc làm từ các trang web công nghệ thông tin của Việt Nam. Tiếp theo, chúng tôi trích xuất nội dung tuyển dụng và áp dụng thuật toán phân cụm để nhóm chúng thành 2 cụm. Chúng tôi đã sử dụng mô hình LSA để thực hiện phân cụm (Suleman & Korkontzelos, 2020). Mô hình LSA đã sử dụng một loạt các từ để xây dựng một ma trận từ ngữ bằng cách sử dụng công thức sau.

$$M = UEV \tag{1}$$

Trong đó,

M = m × n ma trận

U = ma trận m × n số ít bên trái

E = số thực không âm n × n ma trận đường chéo

Ma trận V = n × m

Thuật toán 1 đã trình bày kỹ thuật phân cụm mà mô hình của chúng tôi đã sử dụng

Algorithm 1: job-based clustering

Data: recruitment text files
Result: two groups in data science and software engineering

```
1  doc_term ← {};
2  topics ← 2;
3  words ← 10;
4  document_list, titles = load_data(D)
5  clean_text = preprocess_data(document_list)
6  dictionary ← corpora.Dictionary(clean_text)
7  for doc ∈clean_text do
8  |   doc_term.append(dictionary.doc2bow(doc))
9  end
10 lsa = LsiModel(doc_term, num_topics = topcis, id2word = dictionary);
11 lsa.print_topics(num_topics = topics, num_word = words)
```

Chúng tôi đã sử dụng thư viện gensim để triển khai mô hình LSA. Trong thuật toán 1, doc_term là một ma trận tài liệu thuật ngữ, các chủ đề chứa 2 cụm {khoa học dữ liệu, kỹ thuật phần mềm}, các từ gồm 10 từ xuất hiện trong mỗi cụm ở các dòng 1,2, 3. Hàm load_data được sử dụng để tải tất cả dữ liệu của các hồ sơ tuyển dụng. Sau đó, chúng tôi mã hóa, loại bỏ các từ dừng và thực hiện dựa trên các tài liệu để tạo ma trận doc_term thuật ngữ tài liệu trên các dòng 4, 5, 6, 7 và 8. Cuối cùng, chúng tôi tạo mô hình LSA và in các từ của mỗi cụm trên các dòng 10 và 11.

Đo lường sự tương đồng giữa khả năng của sinh viên và nhu cầu tuyển dụng của nhà tuyển dụng

Trong phần này, nghiên cứu đã thu thập bảng điểm của gần 1.000 sinh viên ngành Công nghệ thông tin với hai chuyên ngành khoa học dữ liệu và kỹ thuật phần mềm. Đối với mỗi môn học mà sinh viên đạt được điểm trung bình trên 3.0, chúng tôi trích xuất các mô tả về các kỹ năng đầu ra từ đề cương môn học. Đây là kiến thức mà sinh viên thu được và chúng tôi coi đó là khả năng của sinh viên. Về mặt toán học, chúng tôi đặt S là một tập sinh viên với $s_i = \{d_i \mid s_i \in S, d_i \in D \wedge GPA\ (si) \geq 3.0\}$, D là tập nội dung mô tả kỹ năng đầu ra. Chúng tôi xác định vấn đề đo lường mức độ tương đồng của sinh viên với nhu cầu tuyển dụng như sau: Khả năng của sinh viên là đối tượng mà sinh viên đạt được điểm trung bình ≥ 3.0. Các môn học này được liên kết với nhau để tạo thành một mô tả về khả năng của sinh viên. Chúng ta đặt C = {data science, software engineering} là các cụm đã được tập hợp, *c* là nội dung tuyển dụng của một công ty và được định nghĩa như sau: $C_k = \{c_i, ..., c_j \mid c_i \in s_i\}$, trong đó k là giá trị C. Sau khi có nội dung mô tả năng lực của sinh viên và nhu cầu tuyển dụng của doanh nghiệp, chúng tôi sử dụng mô hình Fasttext để chuyển nội dung thành một vector đặc trưng. FastText là một thư viện mã nguồn mở do Facebook tạo ra, hỗ trợ tính năng nhúng từ và đào tạo phân loại văn bản (Zhou và cộng sự, 2020). Cuối cùng, mô

hình đề xuất đã sử dụng thước đo cosine để tìm ra 5 công ty hàng đầu phù hợp nhất với khả năng của sinh viên (Oduntan và cộng sự, 2018).

$$score(s_i, c_i) = cosine(fasttext(s_i), fasttext(c_i)) \quad (2)$$

Trong đó, hàm Fasttext sẽ chuyển đổi nội dung của si và ci thành các vectơ đặc trưng. Hàm cosine được sử dụng để tính độ giống nhau giữa chúng.

Rút trích những kỹ năng cần thiết cho sinh viên

Trích xuất từ khóa là một quy trình trích xuất cụm từ khóa tự động từ văn bản. Trong nghiên cứu này, tác giả đã sử dụng thuật toán YAKE để lấy các cụm từ khóa đại diện cho các kỹ năng của sinh viên và nhu cầu của công ty họ. Về mặt toán học, chúng tôi gọi K_s *là tập hợp các cụm từ khóa về kỹ năng của sinh viên,* K_s *là tập hợp các cụm từ khóa về kỹ năng mà các công ty cần tuyển dụng. Các kỹ năng mô tả khả năng học tập của* sinh viên *được định nghĩa như sau:* $K_{si} = \{k_1, k_2,..., k_j \mid k \in K_s\}$ là những kỹ năng mà sinh viên đạt được, $K_{ci} = \{k_1, k_2,..., k_j \mid k \in K_{ci}\}$ là những kỹ năng mà nhà tuyển dụng cần. Tiếp theo, chúng tôi tính toán các kỹ năng cần thiết cho sinh viên bằng công thức sau:

$$Skills(s_i) = K_{ci} \setminus K_{si} \quad (3)$$

Thuật toán 2 trình bày quá trình đề xuất các kỹ năng cần thiết cho sinh viên theo khả năng học tập của họ.

Algorithm 2: Extracting necessary skills for students

Data: S: learning ability description files, J: recruitment text files
Result: a list of missing skills

```
1  student_missing_skills ← {};
2  job_skill ← {};
3  stopwords = open("stopword.txt").read.splitlines();
4  extractor = yake.extractor(n = 2, stopwords);
5  job_skill = extractor.extract_keyword(J);
6  for s ∈ S do
7  |   student_skill = extractor.extract_keyword(s);
8  |   student_missing_skills.append(student_skill − job_skill);
9  end
10 return student_missing_skills
```

Chúng tôi đã sử dụng thư viện YAKE để thực hiện trích xuất từ khóa. Biến student_missing_skills được sử dụng để lưu trữ các kỹ năng của từng sinh viên, biến job_skill được sử dụng để chứa các kỹ năng yêu cầu từ nhà tuyển dụng ở dòng 1 và 2. Họ sử dụng tệp stopwords.txt để loại bỏ các từ phổ biến và khởi tạo mô hình YAKE trong dòng 3 và 4. Tiếp theo, chúng tôi rút ra các kỹ năng cần thiết từ các công ty ở dòng 5. Cuối cùng, chúng tôi tính toán các kỹ năng cần thiết cho sinh viên ở các dòng 6, 7, 8, 9 và 10.

4. Thực nghiệm

Mô hình của chúng tôi được cài đặt trên máy tính có cấu hình như CPU 11th Gen Intel® Core™ i7-11800h @ 2.30GHz (16 CPUs), ~2.3GHz, RAM: 16 GB. Máy tính cài phần mềm Sublime Text, python 3.9.0. Chúng tôi đã sử dụng các thước đo phổ biến về độ giống văn bản như Jaccard, cosine và TF-IDF (Sohangir & Wang, 2017), (Artama và cộng sự, 2020). Ngoài ra, chúng tôi cũng sử dụng word2vec để đo độ tương tự khi chuyển đổi văn bản sang chiều thấp vectơ (Mikolov và cộng sự, 2013).

A. So sánh sự tương đồng về năng lực học tập của sinh viên với nhu cầu tuyển dụng

Nghiên cứu này chọn ngẫu nhiên 2 doanh nghiệp đại diện cho 2 chuyên ngành khoa học dữ liệu và phần mềm A tin tuyển dụng khoa học dữ liệu:

- Thu thập và giải thích các nguồn dữ liệu bằng các phương pháp thống kê khoa học
- Tạo các nguồn dữ liệu có liên quan đến hoạt động kinh doanh hàng ngày của công ty
- Xác định các quy trình thu thập dữ liệu mới
- Lọc và xóa dữ liệu bất thường Thực hiện phân tích sâu sắc các kết quả
- Khắc phục sự cố nguồn dữ liệu và cơ sở dữ liệu
- Báo cáo, đánh giá và đưa ra các đề xuất phát triển dựa trên dữ liệu kinh doanh đã thu thập
- Xác định và nắm bắt các cơ hội kinh doanh và cơ hội cải tiến quy trình
- Phối hợp chặt chẽ với các nhóm bán hàng, kế toán và quản lý để nhận ra và nắm bắt các cơ hội phát triển Yêu thích và đam mê nghiên cứu và xử lý dữ liệu

Cao đẳng/ Đại học chuyên ngành Toán học, Kinh tế, Khoa học Máy tính, Quản lý Thông tin hoặc Thống kê 2 năm kinh nghiệm làm việc trong lĩnh vực phân tích/ khoa học dữ liệu

- Sử dụng tốt Power BI Hiểu biết về ngành bán lẻ, thương mại điện tử Kỹ năng phân tích, tư duy logic, lập luận chặt chẽ
- Kỹ năng tổ chức công làm việc, làm việc độc lập hoặc chức năng chéo
- Sử dụng tiếng Anh tốt

Kết quả so sánh giữa Fasttext và các phương pháp khác được thể hiện trong bảng 1.

Bảng 1. So sánh sự tương đồng giữa khả năng học tập của sinh viên và nhu cầu tuyển dụng ngành khoa học dữ liệu

MSSV	Fasttext + cosine	Cosine	Tf-idf + cosine	Jaccard	Word2vec + cosine
187IT14044	**0.9799**	0.6231	0.0204	0.1325	0.5028
187IT23746	**0.9798**	0.6194	0.0202	0.1414	0.503
187IT14048	**0.9767**	0.5727	0.0096	0.1422	0.5042
187IT21187	**0.9776**	0.5866	0.0097	0.1432	0.5044
187IT23616	**0.9795**	0.6037	0.0243	0.1447	0.5038
197CT01LHS	**0.9786**	0.5897	0.0148	0.1857	0.5094
197CT31272	**0.9783**	0.6135	0.0146	0.1583	0.5054
197CT22514	**0.9782**	0.5954	0.0142	0.158	0.5048
197CT09890	**0.9778**	0.5855	0.0128	0.1677	0.5047
197CT31311	**0.9761**	0.5627	0.0123	0.1489	0.5056

Trong Bảng 1, Fasttext + cosine cho thấy ba sinh viên có mã số 187IT14044,187IT23746 và 187IT23616 đều có mức độ phù hợp với công việc cao nhất mà nhà tuyển dụng đưa ra lần lượt là 0,9799, 0,9798 và 0,9795. Hai sinh viên có mã số 187IT14048 và 187IT14048 tuy học cùng khóa với hai sinh viên trên nhưng mức độ phù hợp thấp hơn. Điều này được lý giải như sau là do hiện nay, Trường Đại học Văn Lang đang đào tạo theo học chế tín chỉ. Sinh viên đăng ký chuyên ngành yêu thích vào đầu mỗi học kỳ. Vì vậy, những sinh viên cùng khóa học nhưng có số môn học khác nhau, bao gồm cả các môn học không thuộc chuyên ngành Khoa học dữ liệu.

Tiếp theo, chúng tôi đã thử nghiệm với lĩnh vực tuyển dụng kỹ thuật phần mềm:

- Tham gia một phần hoặc toàn bộ quá trình phát triển dự án theo mô hình Agile. Tham gia xây dựng và phát triển phần mềm quản lý trên công nghệ ASP.NET/. Net C #, Net / C ++.
- Thiết kế & viết code ứng dụng Web Tốt nghiệp đại học / cao đẳng chuyên ngành CNTT (phần mềm).
- Có kiến thức về lập trình hướng đối tượng.
- Ưu tiên có ít nhất 1 năm kinh nghiệm phát triển phần mềm thực hành.
- Có kinh nghiệm về công nghệ: ASP.NET/C#/.Net, MVC hoặc C ++. Kỹ năng phát triển web: HTML5 / CSS3 hoặc Openui5.
- Có kiến thức làm việc với các hệ quản trị cơ sở dữ liệu: Database MongoDB / MS SQL Server. Tinh thần trách nhiệm cao trong công việc.
- Tích cực, chủ động, nhiệt tình trong công việc.
- Làm việc theo nhóm, và nhóm. Tinh thần hợp tác.

Kết quả so sánh giữa Fasttext và các phương pháp khác được thể hiện trong bảng 2.

Bảng 2. So sánh mức độ tương đồng giữa năng lực học tập của sinh viên và nhu cầu tuyển dụng ngành kỹ thuật phần mềm

MSSV	Fasttext + cosine	Cosine	Tf-idf +cosine	Jaccard	Word2vec+cosine
187IT14044	**0.9706**	0.3761	0.031	0.0744	0.5048
187IT23746	**0.9702**	0.3727	0.033	0.0793	0.5051
187IT14048	**0.9721**	0.3951	0.0163	0.1007	0.5072
187IT21187	**0.9742**	0.4188	0.0165	0.0951	0.5075
187IT23616	**0.971**	0.4064	0.0413	0.0854	0.5065
197CT01LHS	**0.9707**	0.3786	0.0312	0.1225	0.5156
197CT31272	**0.9697**	0.3712	0.021	0.0983	0.5091
197CT22514	**0.9687**	0.3614	0.0192	0.0982	0.5083
197CT09890	**0.9692**	0.3594	0.0185	0.0966	0.508
197CT31311	**0.9708**	0.3853	0.0208	0.1101	0.5095

Trong bảng 2, tại cột Fasttext + cosine, hai sinh viên có mức độ phù hợp thấp hơn do chưa học đủ với mã số 187IT14048 và 187IT21187 trúng tuyển chuyên ngành hoặc chuyên ngành khác, v.v. Như đã đề cập, do mức độ phù hợp do nhà tuyển dụng đưa ra là 0,9721 và sinh viên được chọn học theo tín chỉ nên có thể thấy mức tương ứng là 0,9742. Tuy nhiên, còn lại hai sinh viên 187IT14048 và 187IT21187 phù hợp với sinh viên ngành kỹ thuật phần mềm.

Qua thử nghiệm trình bày ở bảng 1 và bảng 2, kỹ thuật nhúng từ fasttext + cosine cho kết quả khá cao trên 90% so với các phương pháp còn lại. **Trích xuất các kỹ năng cần thiết bằng mô hình YAKE.**

Chúng tôi đã sử dụng kỹ thuật YAKE để trích xuất từ khóa từ nội dung mô tả khả năng của sinh viên và nhu cầu tuyển dụng của công ty (Campos và cộng sự, 2020). Sau khi trích xuất từ khóa, nghiên cứu đã loại bỏ những từ không có nghĩa trong công nghệ thông tin. Mô hình RAKE được chọn làm cơ sở để so sánh. Chúng tôi chọn ngẫu nhiên 10 sinh viên có điểm tương đồng cao nhất về khoa học dữ liệu và kỹ thuật phần mềm và trích xuất các cụm từ khóa từ những sinh viên này. Bên cạnh đó, chúng tôi cũng kéo các cụm từ khóa từ các tin tuyển dụng.

Trích xuất kỹ năng khoa học dữ liệu

Sau khi trích xuất các cụm từ mô tả khả năng của 10 sinh viên khoa học dữ liệu, chúng tôi đã chọn ra 10 kỹ năng hàng đầu được trình bày trong Bảng 3.

Bảng 3. YAKE trích xuất các kỹ năng từ các nhà tuyển dụng và sinh viên trong ngành khoa học dữ liệu

Các kĩ năng cần thiết từ nhà tuyển dụng	Kĩ năng sinh viên Công nghệ Dữ liệu
dữ liệu lớn	Lập trình Java
sử dụng R	dự đoán hồi quy
học sâu	Học sâu
Nghiên cứu AI	Nghiên cứu AI
máy học	máy học
xử lý ngôn ngữ tự nhiên	hệ thống hỗ trợ quyết định
khai thác dữ liệu	mạng thần kinh
mạng lưới thần kinh nhân tạo	MongoDB
phân tích kinh doanh	Thị giác máy tính
thiết kế phần mềm	Lập trình song song

Trong Bảng 3, chúng ta thấy rằng các kỹ năng mà nhà tuyển dụng yêu cầu nhưng kiến thức mà sinh viên thu được lại thiếu là {Sử dụng R, quy trình ngôn ngữ tự nhiên, khai thác dữ liệu, phân tích kinh doanh và thiết kế phần mềm}. Đây là những kỹ năng mà khoa học dữ liệu.

Sau khi trích xuất các cụm từ mô tả khả năng của 10 sinh viên ngành khoa học dữ liệu, chúng tôi đã chọn ra 10 kỹ năng hàng đầu được trình bày trong Bảng 3: sinh viên cần trang bị trước khi tốt nghiệp.

Trích xuất kỹ năng kỹ thuật phần mềm

Chúng tôi tiếp tục trích xuất các cụm từ mô tả khả năng của 10 sinh viên ngành kỹ thuật phần mềm và có kết quả như sau trong bảng 4.

Bảng 4. YAKE trích xuất các kỹ năng từ các nhà tuyển dụng và sinh viên trong ngành kỹ thuật phần mềm

Các kĩ năng cần thiết từ nhà tuyển dụng	Kĩ năng sinh viên công nghệ phần mềm
cơ sở dữ liệu NoSQL	lập trình hướng đối tượng
kiểm thử phần mềm	lập trình java
kỹ năng thuật toán	cơ sở dữ liệu
suy nghĩ logic	lập trình ứng dụng web
ứng dụng web	lập trình ứng dụng android
cơ sở dữ liệu	kiểm thử phần mềm
thiết kế phần mềm	thiết kế phần mềm
quản lý dự án	kỹ thuật lấy yêu cầu
phát triển API	cấu trúc dữ liệu và thuật toán

Trong Bảng 4, chúng ta thấy rằng các kỹ năng mà nhà tuyển dụng yêu cầu nhưng kiến thức mà sinh viên thu được còn thiếu là {cơ sở dữ liệu nosql, tư duy logic, quản lý dự án, phát triển API}. Đây là những kỹ năng mà sinh viên ngành kỹ thuật phần mềm cần trang bị trước khi ra trường.

5. Kết luận

Trong bài báo này, tác giả đã áp dụng các kỹ thuật phân cụm văn bản, đo độ tương đồng và trích xuất từ khóa để xây dựng hệ thống khuyến nghị kỹ năng cho sinh viên chưa tốt nghiệp. Điều này giúp sinh viên chuẩn bị đầy đủ kiến thức trước khi ra trường và tìm được công việc phù hợp. Nghiên cứu này sử dụng kỹ thuật phân cụm để nhóm việc tuyển dụng vào hai lĩnh vực khoa học dữ liệu và kỹ thuật phần mềm. Sau đó, mô hình thể hiện mô tả các môn học chuyên ngành mà sinh viên đạt được và nhu cầu tuyển dụng của doanh nghiệp thành các vector nhúng từ. Hệ thống đã tiến hành đo lường mức độ tương đồng để gợi ý những doanh nghiệp phù hợp với khả năng học tập của từng sinh viên. Cuối cùng, việc trích xuất từ khóa đã rút ra những kỹ năng cần thiết cho sinh viên. Tuy nhiên, hệ thống có những hạn chế như chỉ thử nghiệm trên tập dữ liệu sinh viên khá nhỏ và không xử lý được dữ liệu lớn. Hệ thống chưa khuyến nghị những môn học mà sinh viên cần học để trang bị thêm kiến thức so với nhu cầu của doanh nghiệp. Trong thời gian tới, tác giả sẽ tiếp tục nghiên cứu các giải pháp xử lý dữ liệu lớn và sử dụng mô hình học máy để cải thiện hiệu suất của hệ thống.

Tài liệu tham khảo

Wartelle, A., Mourad-Chehade, F., Yalaoui, F., Chrusciel, J., Laplanche, D., & Sanchez, S. (2021). Clustering of a health dataset using diagnosis co-occurrences. *Applied Sciences*, *11*(5), 2373.

Nguyen, A. T., Phan, H. V. T. (2021). Developing a Job Recommended System for Students in Information Technology based on Learning Capability Meeting the Needs of Employers. *Proceedings of the 14th National Conference on Fundamental and Applied Information Technology Research (FAIR2021)*.

Chengyu Sun, L. H. (2020). A Review of Unsupervised Keyphrase Extraction Methods Using Within-Collection Resources. *Symmetry*.

Fernández, G., López, W., Olivera, F., Rienzi, B., & Rodríguez-Bocca, P. (2014, September). Let's go to the cinema! A movie recommender system for ephemeral groups of users. *Clei electronic Journal*, *18*(2)

Phan, H. V. T., & Do, P. (2020). BERT+vnKG: Using Deep Learning and Knowledge Graph to Improve Vietnamese Question Answering System. *International Journal of Advanced Computer Science and Applications (IJACSA), 11*(7), 480-487.

Jay Kumar, J. S. (2020). An Online Semantic-enhanced Dirichlet Model for Short Text Stream Clustering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 766-776). Association for Computational Linguistics.

Son, J., & Kim, S. B. (2017). Content-based filtering for recommendation systems using multiattribute networks. *Expert Systems with Applications*, *89*, 404-412.

Juli Rejito, A. S. (2017). Image indexing using color histogram and k-means clustering for optimization CBIR in image database. *Journal of Physics: Conference Series, Volume 893, The Asian Mathematical Conference 2016 (AMC 2016)*. Bali Nusa Dua Convention Center, Bali, Indonesia: IOP Publishing Ltd.

Kalun Ho, J. K.-J. (2020). Learning Embeddings for Image Clustering: An Empirical Study of Triplet Loss Approaches. *arXiv:2007.03123v1 [cs.CV]*.

Artama, M., Sukajaya, I. N., & Indrawan, G. (2020, April). Classification of official letters using TF-IDF method. In *Journal of Physics: Conference Series, 1516*(1), 012001. IOP Publishing.

Bustos López, M., Alor-Hernández, G., Sánchez-Cervantes, J. L., Paredes-Valverde, M. A., & Salas-Zárate, M. D. P. (2020). EduRecomSys: an educational resource recommender system based on collaborative filtering and emotion detection. *Interacting with Computers*, *32*(4), 407-432.

Mohamed, M. H., Khafagy, M. H., & Ibrahim, M. H. (2019, February). Recommender systems challenges and solutions survey. In *2019 international conference on innovative trends in computer engineering (ITCE)* (pp. 149-155). IEEE.

Mohamed Ouhda, K. E. (2018). A Content Based Image Retrieval Method Based on K-Means Clustering Technique. *Journal of Electronic Commerce in Organizations, 16(1)*, 82-96.

Ibrahim, M. S., & Saidu, C. I. (2020). Recommender systems: algorithms, evaluation and limitations. *Journal of Advances in Mathematics and Computer Science*, *35*(2), 121-137.

Nakul Sharma, P. Y. (2021). Keyphrase Extraction And Source Code Similarity Detection - A Survey. *IOP Conf. Series: Materials Science and Engineering*.

Osman, N. A., Mohd Noah, S. A., Darwich, M., & Mohd, M. (2021). Integrating contextual sentiment analysis in collaborative recommender systems. *Plos one*, *16*(3), e0248695.

Oduntan, O. E., Adeyanju, I. A., Falohun, A. S., & Obe, O. O. (2018). A comparative analysis of Euclidean distance and cosine similarity measure for automated essay-type grading. *Journal of Engineering and Applied Sciences*, *13*(11), 4198-4204.

Do, P., Phan, T. H., & Gupta, B. B. (2021). Developing a Vietnamese tourism question answering system using knowledge graph and deep learning. *Transactions on Asian and Low-Resource Language Information Processing*, *20*(5), 1-18.

Thorat, P. B., Goudar, R. M., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, *110*(4), 31-36.

Qusay Bsoul, R. A. (2021). Arabic Text Clustering Methods and Suggested Solutions for Theme-Based Quran Clustering: Analysis of Literature. *Journal of Information Science Theory and Practice*, 15-34.

Suleman, R. M., & Korkontzelos, I. (2021). Extending latent semantic analysis to manage its syntactic blindness. *Expert Systems with Applications*, *165*, 114130.

Raza, S. D. (2021). News recommender system: a review of recent progress, challenges, and opportunities. *Artif Intell Rev*, 749-800.

Fortes, R. S., de Freitas, A. R., & Gonçalves, M. A. (2017). A Multicriteria Evaluation of Hybrid Recommender Systems: On the Usefulness of Input Data Characteristics. In *ICEIS (2)*, 623-633.

Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, *509*, 257-289.

Shruthi, S., & Gripsy, V. J. (2017). An Effective Product Recommendation System for E-Commerce Website Using Hybrid Recommendation Systems. *International Journal of Computer Science & Communication*, *8*(2), 81-88.

Sohangir, S., & Wang, D. (2017). Improved sqrt-cosine similarity measurement. *Journal of Big Data*, *4*(1), 1-13.

Salloum, S., & Rajamanthri, D. (2021). Implementation and evaluation of movie recommender systems using collaborative filtering. *Journal of Advances in Information Technology*.

Sergey V. Kireev, I. L. (2017). Economic Clusters: Concepts and Characteristic Features. *International Journal of Applied Business and Economic Research*, 123-132.

Monica, S., Natalia, F., & Sudirman, S. (2018, June). Clustering tourism object in Bali province using k-means and x-means clustering algorithm. In *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS)* (pp. 1462-1467). IEEE.

Zhou, T., Wang, Y., & Zheng, X. (2020, December). Chinese text classification method using FastText and term frequency-inverse document frequency optimization. In *Journal of Physics: Conference Series, 1693*(1), 012121. IOP Publishing.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Yuan, T., Deng, W., Hu, J., An, Z., & Tang, Y. (2019). Unsupervised adaptive hashing based on feature clustering. *Neurocomputing*, *323*, 373-382.

Fayyaz, Z., Ebrahimian, M., Nawara, D., Ibrahim, A., & Kashef, R. (2020). Recommendation systems: Algorithms, challenges, metrics, and business opportunities. *Applied Sciences*, *10*(21), 7748.