Áp dụng thuật toán mờ vào gợi ý công việc liên quan

1. Giai đoạn xử lý dữ liệu:

Xác định các đặc trưng của công việc và chuẩn bị dữ liệu

Trích xuất các đặc trưng (features) cho mỗi công việc, chẳng hạn như:

- Ngành nghề (field)

- Mức lương (salary)

- Vị trí (location)

- Kinh nghiệm (experience)

- Kỹ năng yêu cầu (skills)

Chuẩn hóa các đặc trưng để đưa chúng về cùng một phạm vi. Việc chuẩn hóa giúp các đặc trưng có mức độ ảnh hưởng ngang nhau trong việc tính toán độ thuộc về và phân cụm.

Các cách chuẩn hóa:

Loại dữ liệu	Phương pháp chuẩn hóa
Dữ liệu liên tục (e.g., lương, kinh nghiệm)	Min-Max Scaling, Z-score Standardization
Dữ liệu phân loại (e.g., ngành nghề, công ty)	One-Hot Encoding, Label Encoding
Dữ liệu nhị phân (e.g., đã ứng tuyển)	Không cần chuẩn hóa hoặc Min-Max Scaling
Dữ liệu không chuẩn (e.g., phân phối lệch)	Robust Scaler

Dữ liệu tỷ lệ (e.g., số giờ làm việc)	Min-Max Scaling, Z-score Standardization

Dữ liệu có giá trị liên tục (Continuous Data)

Những trường như mức lương, kinh nghiệm làm việc, số năm học, độ tuổi, v.v., thường là giá trị liên tục và có thể có phạm vi rất khác nhau.

- Phương pháp chuẩn hóa:

+ Min-Max Scaling (Chuẩn hóa min-max):

Công thức:

$$X' = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Giúp đưa giá trị vào khoảng [0, 1] hoặc [-1, 1] nếu cần. Phương pháp này rất hữu ích khi bạn muốn giữ được dải giá trị ban đầu của dữ liệu nhưng có phạm vi nhất quán.

Ứng dụng: Dùng khi các giá trị có đơn vị hoặc phạm vi khác nhau và cần đưa về một phạm vi chung để dễ so sánh.

+ Z-score Standardization (Chuẩn hóa Z-Score):

Công thức:

$$X' = \frac{X - \mu}{\sigma}$$

Giúp đưa dữ liệu về phân phối chuẩn (mean = 0, standard deviation = 1). Phù hợp với các mô hình yêu cầu dữ liệu có phân phối chuẩn như SVM, Logistic Regression.

Ứng dụng: Dùng khi các giá trị có phạm vi rộng hoặc không đồng nhất và muốn giảm ảnh hưởng của các giá trị cực trị (outliers).

Dữ liệu phân loại (Categorical Data)

Các trường như ngành nghề, vị trí công việc, trình độ học vấn, hay công ty làm việc thường là dữ liệu phân loại (có các giá trị rời rạc).

- Phương pháp chuẩn hóa:

+ One-Hot Encoding (Mã hóa một nóng):

Mỗi giá trị của trường phân loại sẽ được mã hóa thành một cột nhị phân (0 hoặc 1).

Ví dụ: Nếu trường "Ngành nghề" có các giá trị như "Kỹ sư", "Nhân viên văn phòng", "Quản lý", thì ta tạo ra 3 cột với các giá trị 1 hoặc 0 cho mỗi công việc.

Ứng dụng: Dùng khi trường có số lượng nhỏ giá trị phân loại.

+ Label Encoding (Mã hóa nhãn):

Chuyển các giá trị phân loại thành các số nguyên. Ví dụ, "Kỹ sư" = 0, "Nhân viên văn phòng" = 1, "Quản lý" = 2.

- Ứng dụng: Dùng khi các giá trị có thể so sánh thứ tự, nhưng không thích hợp cho các thuật toán yêu cầu có một phạm vi chuẩn như One-Hot.

Dữ liệu nhị phân (Binary Data)

Các trường như “Có kỹ năng A hay không”, “Đã ứng tuyển hay không”, v.v. thường là các giá trị nhị phân (0 hoặc 1).

Phương pháp chuẩn hóa:

Không cần chuẩn hóa:

Các giá trị nhị phân đã có phạm vi từ 0 đến 1, vì vậy không cần phải chuẩn hóa thêm. Tuy nhiên, nếu cần phải chuẩn hóa để phục vụ cho các thuật toán như SVM hoặc neural networks, có thể áp dụng Min-Max Scaling (chuyển về [0,1] nếu cần).

Dữ liệu có phân phối không chuẩn (Non-Normalized Data)

Đối với các dữ liệu có phân phối không đồng đều, cần phải kiểm tra và xử lý các giá trị cực trị (outliers) hoặc phân phối bất thường trước khi chuẩn hóa.

- Phương pháp chuẩn hóa:

+ Robust Scaler:

Công thức:

$$X' = \frac{X - \text{Median}(X)}{\text{IQR}(X)}$$

Giúp chuẩn hóa dữ liệu với tỉ lệ chuẩn bằng cách sử dụng độ lệch chuẩn và trung vị, thay vì trung bình và độ lệch chuẩn như trong chuẩn hóa Z-score. Phương pháp này đặc biệt hữu ích khi dữ liệu có nhiều outliers.

- Ứng dụng: Dùng khi dữ liệu chứa nhiều giá trị ngoại lai hoặc phân phối không chuẩn.

Dữ liệu tỷ lệ (Ratio Data)

Các trường như lương, số giờ làm việc, số năm kinh nghiệm làm việc có thể được xử lý như dữ liệu tỷ lệ, tức là có một điểm gốc và phép toán giữa các giá trị có ý nghĩa.

- Phương pháp chuẩn hóa:

+ Min-Max Scaling hoặc Z-score Standardization đều có thể áp dụng tùy vào yêu cầu của mô hình và mục tiêu của dự án.

- Ứng dụng: Dùng khi dữ liệu có giá trị tỷ lệ và cần đưa về cùng một phạm vi.

Thu thập dấu vết của người dùng

- Truy vết hành vi người dùng, bao gồm các công việc mà người dùng đã xem, lưu, follow, hoặc ứng tuyển.

- Loại trừ các công việc đã ứng tuyển khỏi quá trình gợi ý, vì những công việc này không cần phải gợi ý lại.

- Các dấu vết như "đã xem" có thể được sử dụng để xác định mức độ giám sát (α) cho các công việc trong quá trình phân cụm.

Xác định mức độ giám sát α

Giám sát mức độ (α): Với các công việc mà người dùng đã xem, lưu, hoặc follow, bạn gán cho chúng các mức α khác nhau để quyết định mức độ "giám sát". Ví dụ:

- Công việc đã xem có thể có α = 0.5.

- Công việc đã lưu có thể có α = 0.7.

- Công việc đã follow có thể có α = 0.8.

- Không giám sát các công việc chưa xem/lưu/follow: Những công việc này không có α và không bị giám sát, chỉ sử dụng thuật toán để phân cụm.

Tính toán độ thuộc về (U) và tham số mờ hóa (M)

Áp dụng thuật toán sSMC-FCM để tính toán độ thuộc về (U) của mỗi công việc đối với các cụm "liên quan" và "không liên quan".

Để tính toán tham số Mi và M' cho từng công việc:

- Mi là tham số mờ hóa cho mỗi công việc, tính theo độ đậm đặc của các phần tử (dựa trên khoảng cách với các công việc khác).

- M' được tính để điều chỉnh mức độ giám sát cho từng công việc, từ đó xác định độ thuộc về chính xác hơn.

2. Sử dụng thuật toán SSMC-FCM để phân cụm

Sử dụng thuật toán SSMC-FCM (Self-Supervised Multi-Cluster Fuzzy C-Means) để phân loại các công việc mà bạn đã xem, đã lưu hoặc đã follow vào các cụm liên quan và không liên quan.

- Cụm liên quan: Các công việc có các yếu tố tương đồng cao với công việc X (về ngành nghề, yêu cầu kỹ năng, mức lương, vị trí,...).

- Cụm không liên quan: Các công việc có sự khác biệt lớn hoặc ít tương đồng với công việc X.

3. Chọn ra các công việc trong cụm "liên quan"

Sau khi phân cụm, hệ thống sẽ chỉ gợi ý các công việc từ cụm "liên quan" đến công việc X. Các công việc này có độ thuộc tính cao với công việc X và có thể sẽ phù hợp với bạn.

Các công việc này có thể bao gồm:

- Các công việc tương tự về ngành nghề, kỹ năng, vị trí, hoặc yêu cầu công việc.

- Các công việc có độ tương đồng cao với những công việc bạn đã lưu, đã xem, hoặc đã ứng tuyển trong quá khứ.

Để đo lường sự tương đồng giữa công việc X và các công việc khác, hệ thống có thể sử dụng ma trận độ thuộc từ thuật toán FCM.

Công việc nào có độ thuộc cao vào cụm "liên quan" sẽ được gợi ý với khả năng cao hơn.

Khi bạn xem một công việc mới, hệ thống có thể giám sát hành động này và cho công việc đó một mức độ α nhất định (ví dụ: 0.6 nếu là công việc bạn quan tâm một cách vừa phải).

Sau đó, thuật toán sẽ điều chỉnh độ thuộc của công việc này trong các cụm, và tái phân cụm với các công việc đã được xem để xác định lại công việc nào sẽ được gợi ý.

4. Gợi ý công việc

Dựa vào độ thuộc vào các cụm và mức độ giám sát (α) của các công việc, hệ thống sẽ gợi ý các công việc có độ thuộc cao nhất trong cụm "liên quan" với công việc X.

Các công việc gợi ý này có thể là những công việc bạn chưa từng xem hoặc đã xem, tùy thuộc vào chính sách giám sát và thuật toán của hệ thống.

Giả sử bạn đang xem công việc X với các đặc điểm sau:

Ngành nghề: Kỹ sư phần mềm

Vị trí: TP.HCM

Kinh nghiệm: 3-5 năm

Hệ thống sẽ thực hiện các bước sau:

Xác định công việc đã xem: Bạn đã xem công việc X, hệ thống sẽ lấy các công việc bạn đã lưu, đã follow hoặc đã xem trước đó.

Phân cụm với thuật toán FCM:

Công việc A: Kỹ sư phần mềm, TP.HCM, 3-5 năm kinh nghiệm (độ thuộc cao vào cụm "liên quan").

Công việc B: Kỹ sư phần mềm, Hà Nội, 5-7 năm kinh nghiệm (độ thuộc vào cụm "liên quan").

Công việc C: Nhân viên bán hàng, TP.HCM, 0-2 năm kinh nghiệm (độ thuộc vào cụm "không liên quan").

Gợi ý công việc:

Hệ thống sẽ gợi ý các công việc trong cụm "liên quan" (A, B) với bạn, bởi vì chúng có độ tương đồng cao với công việc X (ngành nghề, vị trí, kinh nghiệm).

Điều chỉnh với giám sát

Nếu bạn chọn "Thích" công việc A và lưu lại, hệ thống sẽ điều chỉnh α của công việc A và tái phân cụm các công việc để đảm bảo rằng những công việc có độ thuộc cao vào bạn sẽ được gợi ý tiếp theo.