ĐẠI HỌC QUỐC GIA HÀ NỘI

TRƯỜNG ĐẠI HỌC CÔNG NGHỆ

BÁO CÁO BÀI TẬP LỚN CUỐI KÌ

Môn học: Trí tuệ nhân tạo

ĐỀ TÀI: TÌM HIỂU VỀ MODEL UNSUPERVISED LEARNING VÀ HỆ THỐNG RECOMMENDATION SYSTEM TRONG HỌC MÁY

Giảng viên: TS. Trần Hồng Việt

Lớp: INT3401 20

Nhóm: 13

Sinh viên thực hiện:

Lê Thanh Tùng	21021387
Tô Minh Tuấn	21021386
Đinh Thái Tuấn	21021385
Ngô Thanh Tùng	21021388

Hà Nội- 2024

MỤC LỤC

LỜI MỞ ĐẦU

Trong thập kỷ qua, học máy (Machine Learning) đã trở thành một trong những lĩnh vực phát triển nhanh nhất và có ảnh hưởng sâu rộng đến nhiều ngành công nghiệp. Từ nhận dạng hình ảnh, xử lý ngôn ngữ tự nhiên đến dự đoán hành vi người dùng, các thuật toán học máy đang thay đổi cách chúng ta tương tác với công nghệ và thế giới xung quanh. Trong số các phương pháp học máy, học không giám sát (Unsupervised Learning) nổi lên như một công cụ mạnh mẽ, giúp khai thác và hiểu rõ hơn về dữ liệu mà không cần sự can thiệp của con người.

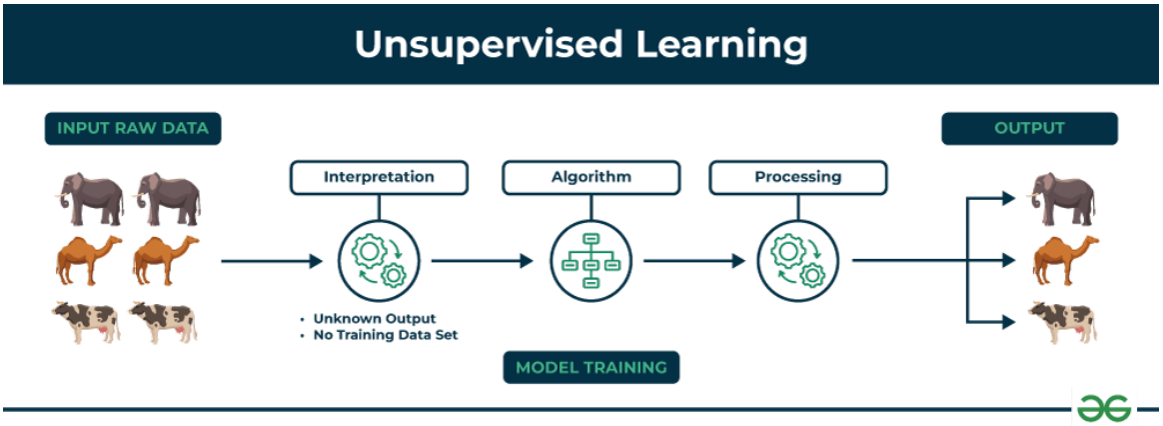
Một trong những ứng dụng quan trọng của học không giám sát là trong hệ thống gợi ý (Recommendation System). Các hệ thống này sử dụng các kỹ thuật học không giám sát để phân tích và khám phá các mẫu ẩn trong dữ liệu người dùng, chẳng hạn như hành vi mua sắm, sở thích xem phim hoặc thói quen nghe nhạc. Thông qua các phương pháp như phân cụm (clustering) và học đặc trưng (feature learning), các hệ thống gợi ý có thể đưa ra các đề xuất cá nhân hóa mà không cần dữ liệu nhãn trước. Điều này không chỉ cải thiện trải nghiệm người dùng mà còn tối ưu hóa doanh thu và tăng sự hài lòng của khách hàng.

Bài báo cáo dưới đây sẽ giúp làm rõ hơn về hai khái niệm Unsupervised learning và Recommendation system.

I. UNSUPERVISED LEARNING

1. Khái niệm unsupervised learning

- Unsupervised learning là một phân nhánh của học máy, xử lý dữ liệu không có nhãn, khác với học có giám sát mà dữ liệu có nhãn cụ thể.
- Trong unsupervised learning, các thuật toán được thiết kế để tự mình tìm hiểu mẫu và mối quan hệ từ dữ liệu mà không cần hướng dẫn hoặc chỉ dẫn rõ ràng từ con người.
- Các thuật toán unsupervised learning không nhận bất kỳ thông tin đầu ra cụ thể nào từ người dùng mà chỉ phân tích cấu trúc dữ liệu để tìm ra các nhóm hoặc mẫu ẩn.
- Phương pháp này thường được sử dụng trong phân tích khám phá dữ liệu, giúp hiểu rõ cấu trúc và đặc điểm cơ bản của dữ liệu.



2. Ứng dụng của unsupervised learning

- Phân khúc khách hàng: Học tập không giám sát có thể được sử dụng để phân khúc khách hàng thành các nhóm dựa trên nhân khẩu học, hành vi hoặc sở thích của họ. Điều này có thể giúp doanh nghiệp hiểu rõ hơn về khách hàng của mình và nhắm mục tiêu đến họ bằng các chiến dịch tiếp thị phù hợp hơn.
- Phát hiện gian lận: Học tập không giám sát có thể được sử dụng để phát hiện gian lận trong dữ liệu tài chính bằng cách xác định các giao dịch đi chệch khỏi mô hình dự kiến. Điều này có thể giúp ngăn chặn gian lận bằng cách gắn cờ các giao dịch này để điều tra thêm.
- Hệ thống đề xuất: Học tập không giám sát có thể được sử dụng để đề xuất các mục cho người dùng dựa trên hành vi hoặc sở thích trong quá khứ của họ. Ví dụ: hệ thống đề xuất có thể sử dụng phương pháp học không giám sát để xác định những người dùng có cùng sở thích về phim và sau đó đề xuất những bộ phim mà những người dùng đó yêu thích.

- Xử lý ngôn ngữ tự nhiên (NLP): Học không giám sát được sử dụng trong nhiều nhiệm vụ NLP khác nhau, bao gồm mô hình hóa chủ đề, phân cụm tài liệu và gắn thẻ từng phần của lời nói.
- Phân tích hình ảnh: Học không giám sát được sử dụng trong nhiều nhiệm vụ phân tích hình ảnh, bao gồm phân đoạn hình ảnh, phát hiện đối tượng và nhận dạng mẫu hình ảnh.

3. Nguyên lý hoạt động của unsupervised learning

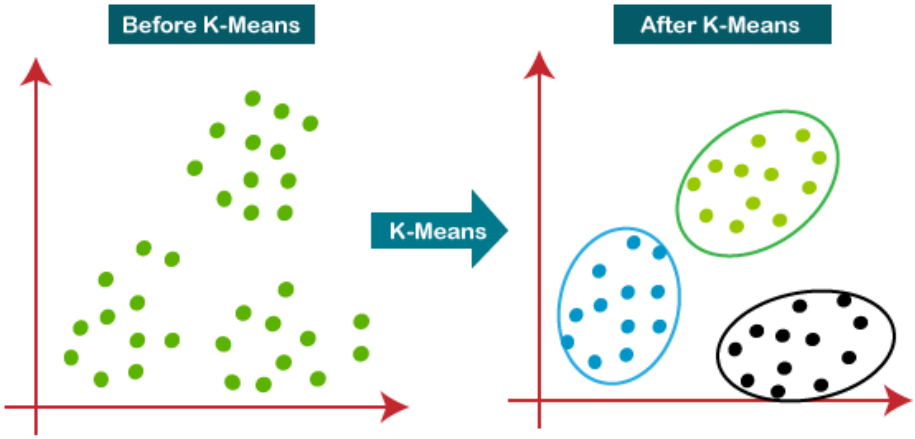
- Khám phá cấu trúc ẩn: Tự động phát hiện mối quan hệ và mẫu trong dữ liệu không được gắn nhãn.
- Phân cụm: Nhóm dữ liệu thành các cụm dựa trên sự tương đồng, giúp phân loại dữ liệu một cách có ý nghĩa.
- Giảm chiều: Áp dụng các kỹ thuật như phân tích thành phần chính (PCA) để giảm số lượng biến, giữ lại thông tin quan trọng nhất.
- Không sử dụng nhãn dữ liệu: Làm việc trực tiếp với dữ liệu thô, không cần dữ liệu đã được phân loại hoặc gắn nhãn trước.
- Phát hiện mẫu: Tìm các mẫu hoặc xu hướng thường xuyên xuất hiện trong dữ liệu để hiểu rõ hơn về cấu trúc dữ liệu.
- Tự học (Self-organizing maps): Sử dụng mạng tự tổ chức để phân loại và đại diện dữ liệu theo cách mà các mẫu tương tự nhau được nhóm gần nhau.
- Khám phá độ tương quan: Xác định và phân tích mối liên hệ giữa các biến để hiểu biết sâu hơn về dữ liệu.

4. Các thuật toán Unsupervised Learning

Có rất nhiều thuật toán được sử dụng trong học không giám sát với các mục đích khác nhau, như: phân cụm, giảm chiều, phát hiện ngoại lệ, học biểu diễn, …Tuy nhiên, trong báo cáo dưới đây sẽ chỉ đề cập đến thuật toán phân cụm(clustering).

4.1. Thuật toán K-means Clustering

❖ Thuật toán K-means là một trong những thuật toán phổ biến nhất trong lĩnh vực học không giám sát (unsupervised learning) được sử dụng rộng rãi để thực hiện phân cụm dữ liệu (clustering). Mục tiêu của thuật toán này là phân chia n quan sát thành k nhóm (hay cụm) sao cho mỗi quan sát thuộc về cụm với trung tâm gần nhất (trung tâm cụm thường được gọi là centroid).



❖ Công thức của thuật toán K-means

- Khởi tạo: Chọn ngẫu nhiên K điểm dữ liệu ban đầu làm các centroid.
- Gán nhãn cụm: Mỗi điểm dữ liệu được gán vào cụm có centroid gần nhất. Khoảng cách thường được tính bằng khoảng cách Euclidean, được tính bằng công thức:

$$\text{Khoảng cách} = \sqrt{\sum_{i=1}^{d}(x_i - c_i)^2}$$

trong đó x_i là giá trị của điểm dữ liệu trên chiều i, c_i là giá trị của centroid trên chiều i, và d là số chiều của không gian dữ liệu.

- Cập nhật centroid: Sau khi tất cả các điểm đã được gán vào các cụm, centroid mới của mỗi cụm được tính toán lại bằng cách lấy trung bình cộng các điểm trong cụm đó:

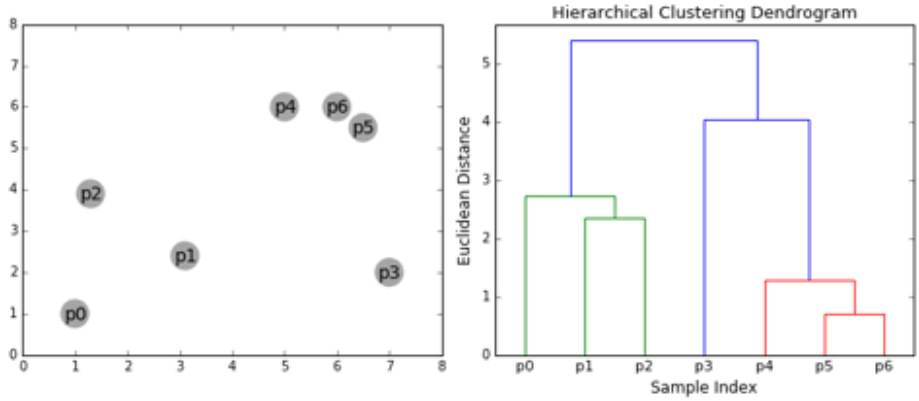
$$c_j = \frac{1}{|S_j|}\sum_{x_i \in S_j} x_i$$

trong đó Sj là tập hợp các điểm dữ liệu trong cụm j và $|Sj|$ là số lượng điểm trong cụm đó.

- Lặp lại: Quá trình gán nhãn cụm và cập nhật centroid được lặp lại cho đến khi đạt được điều kiện dừng (ví dụ như các centroid không thay đổi nữa giữa các lần lặp, hoặc đã đạt số lần lặp tối đa, hoặc sự thay đổi trong tổng khoảng cách giữa các điểm dữ liệu và centroid của chúng dưới một ngưỡng nhất định).

4.2. Thuật toán phân cụm theo cấp bậc

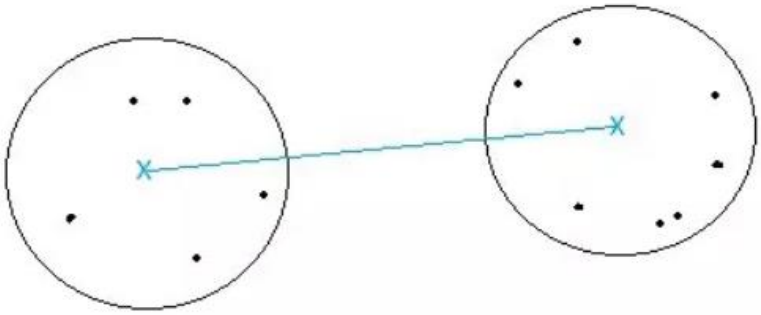
❖ Trong lĩnh vực học không giám sát (unsupervised learning), thuật toán phân cụm theo cấp bậc (hierarchical clustering) được sử dụng rộng rãi để phân chia dữ liệu thành các cụm dựa trên độ tương tự của chúng. Có hai hướng tiếp cận chính: Agglomerative (tổng hợp) và Divisive (chia tách). Trong bối cảnh học không giám sát, mục tiêu là tổ chức dữ liệu thành các cấu trúc có ý nghĩa mà không có bất kỳ nhãn nào được cung cấp trước.



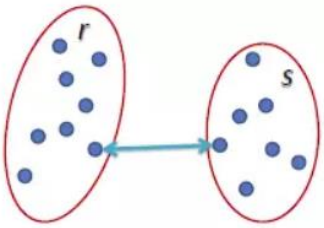
❖ Thuật toán Agglomerative Clustering

- Bước 1: Khởi tạo
    - Bắt đầu bằng cách coi mỗi điểm dữ liệu là một cụm riêng lẻ. Vì vậy, nếu có N điểm dữ liệu, ta có N cụm.
- Bước 2: Tính toán khoảng cách
    - Tính toán khoảng cách giữa các cụm bằng cách sử dụng một số tiêu chí như khoảng cách Euclidean, khoảng cách Manhattan, hoặc bất kỳ phương pháp nào khác phù hợp với dữ liệu và yêu cầu phân tích.
- Bước 3: Gộp cụm
    - Tìm cặp cụm có khoảng cách thấp nhất và gộp chúng lại với nhau, tạo thành một cụm mới.
- Bước 4: Cập nhật khoảng cách
    - Khi hai cụm *Ci* và *Cj* được gộp thành một cụm mới *Ck*, ta cần cập nhật khoảng cách giữa *Ck* và tất cả các cụm còn lại. Có nhiều phương pháp để cập nhật khoảng cách này, bao gồm:

    ▪ Centroid-linkage: Sát nhập hai cụm có khoảng cách giữa hai tâm của hai cụm này là nhỏ nhất.

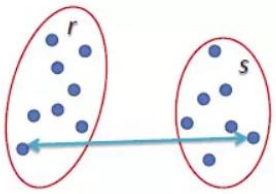


- Single Linkage: Khoảng cách giữa hai cụm là khoảng cách ngắn nhất giữa bất kỳ thành viên nào của cụm này đến bất kỳ thành viên nào của cụm kia.



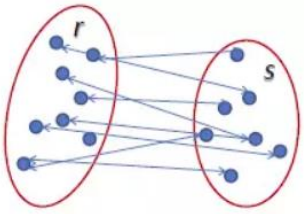
$$L(r,s) = \min(D(x_{ri}, x_{sj}))$$

- Complete Linkage: Khoảng cách giữa hai cụm là khoảng cách lớn nhất giữa bất kỳ thành viên nào của cụm này đến bất kỳ thành viên nào của cụm kia.



$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

- Average Linkage: Khoảng cách giữa hai cụm là trung bình cộng của khoảng cách giữa mọi cặp thành viên của hai cụm.



$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

- Bước 5: Lặp lại
  - Lặp lại quá trình từ Bước 3 cho đến khi tất cả dữ liệu được gộp vào một cụm duy nhất hoặc đạt đến số lượng cụm mong muốn.
- Bước 6: Tạo Dendrogram
  - Dendrogram là một biểu đồ biểu diễn quá trình gộp các cụm và giúp xác định số lượng cụm phù hợp bằng cách cắt dendrogram ở mức độ phù hợp.

❖ Thuật toán Divisive Clustering: đây là thuật toán thứ 2 được sử dụng trong phân cụm theo cấp bậc, với nguyên lý hoạt động hoàn toàn ngược lại với Agglomerative Clustering.

5. Ưu và nhược điểm của Unsupervised learning

- Ưu điểm
  - Không yêu cầu dữ liệu được gắn nhãn: Không giống như học có giám sát, học không giám sát không yêu cầu dữ liệu được gắn nhãn, việc thu thập này có thể tốn kém và mất thời gian.
  - Có thể khám phá các mẫu ẩn: Thuật toán học không giám sát có thể xác định các mẫu và mối quan hệ trong dữ liệu mà con người có thể không rõ ràng.
  - Có thể được sử dụng cho nhiều nhiệm vụ khác nhau: Học không giám sát có thể được sử dụng cho nhiều nhiệm vụ khác nhau, chẳng hạn như phân cụm, giảm kích thước và phát hiện bất thường.
  - Có thể được sử dụng để khám phá dữ liệu mới: Học tập không giám sát có thể được sử dụng để khám phá dữ liệu mới và thu được những hiểu biết sâu sắc mà các phương pháp khác có thể không thực hiện được.

- Nhược điểm
  - Khó đánh giá: Có thể khó đánh giá hiệu suất của các thuật toán học không giám sát vì không có nhãn hoặc danh mục được xác định trước để so sánh kết quả.

- Có thể khó diễn giải: Có thể khó hiểu được quá trình ra quyết định của các mô hình học tập không giám sát.
- Có thể nhạy cảm với chất lượng của dữ liệu: Thuật toán học không giám sát có thể nhạy cảm với chất lượng của dữ liệu đầu vào. Dữ liệu nhiễu hoặc không đầy đủ có thể dẫn đến kết quả sai lệch hoặc không chính xác.
- Có thể tốn kém về mặt tính toán: Một số thuật toán học không giám sát, đặc biệt là các thuật toán xử lý dữ liệu nhiều chiều hoặc bộ dữ liệu lớn, có thể tốn kém về mặt tính toán

II. RECOMMENDATION SYSTEM

1. Khái niệm

- Hệ thống gợi ý (hoặc hệ thống đề xuất) là một lớp con của hệ thống lọc thông tin cung cấp đề xuất cho các mục phù hợp nhất với một người dùng cụ thể.
- Hệ thống đề xuất có thể hoạt động dựa trên một hoặc nhiều loại dữ liệu đầu vào, từ âm nhạc đến tin tức, sách và truy vấn tìm kiếm. Chúng cũng có thể được sử dụng cho các mục tiêu cụ thể như tìm kiếm nhà hàng hoặc hẹn hò trực tuyến. Ngoài ra, chúng cũng có thể được áp dụng trong các lĩnh vực chuyên biệt như khám phá bài báo nghiên cứu hoặc đề xuất dịch vụ tài chính.

2. Nguyên lí hoạt động

- Thu thập dữ liệu: Hệ thống thu thập thông tin từ người dùng, bao gồm lịch sử hoạt động, đánh giá, phản hồi, và thông tin hồ sơ cá nhân.
- Tiền xử lý dữ liệu: Dữ liệu thu thập được có thể cần được tiền xử lý để loại bỏ nhiễu, chuẩn hóa và mã hóa để sử dụng trong các mô hình học máy.
- Xây dựng mô hình: Hệ thống sử dụng các kỹ thuật máy học và khai phá dữ liệu để xây dựng mô hình dự đoán sở thích của người dùng. Các phương pháp phổ biến bao gồm collaborative filtering, content-based filtering, và hybrid approaches.
- Tạo ra đề xuất: Khi mô hình đã được huấn luyện, hệ thống sử dụng nó để tạo ra các đề xuất cá nhân hóa dựa trên sở thích của người dùng. Điều này có thể là sản phẩm, nội dung, thông tin, hoặc bất kỳ thứ gì có thể quan tâm đến người dùng.
- Đánh giá và cải thiện: Hệ thống liên tục đánh giá hiệu suất của các đề xuất bằng cách so sánh giữa các đề xuất được tạo ra và phản hồi thực tế từ người dùng. Dựa trên phản hồi này, hệ thống có thể điều chỉnh mô hình để cải thiện chất lượng của các đề xuất trong tương lai.

- Triển khai và vận hành: Sau khi được xây dựng và kiểm thử, hệ thống được triển khai và duy trì để cung cấp đề xuất cho người dùng thực tế.
-

3. Các thuật toán

Có nhiều thuật toán khác nhau được sử dụng trong hệ thống đề xuất để tạo ra các đề xuất cá nhân hóa dành cho người dùng. Dưới đây là một số thuật toán phổ biến:

❖ Collaborative Filtering (Lọc cộng tác): Là một phương pháp trong hệ thống đề xuất dựa trên việc sử dụng thông tin về sự tương tác giữa người dùng và mục để đưa ra các đề xuất cho người dùng mới. Phương pháp này chia thành hai dạng chính: User-Based Collaborative Filtering và Item-Based Collaborative Filtering.

- User-Based Collaborative Filtering: Trong User-Based Collaborative Filtering, hệ thống đề xuất cho một người dùng các mục mà những người dùng tương tự đã thích.
  - Quá trình hoạt động của thuật toán bao gồm các bước sau:
    - Xác định sự tương đồng giữa người dùng: Thuật toán tính toán độ tương đồng giữa người dùng dựa trên hành vi của họ. Các độ đo tương đồng thường được sử dụng như cosine similarity hoặc correlation coefficient.

      Độ tương tự Cosine:

$$\text{sim}(u,v) = \frac{\sum_{i \in I_u \cap I_v} r_{u,i} \cdot r_{v,i}}{\sqrt{\sum_{i \in I_u} r_{u,i}^2} \cdot \sqrt{\sum_{i \in I_v} r_{v,i}^2}}$$

      Trong đó:

      u, v là hai người dùng,

      I_u, I_v là tập hợp các mục mà hai người dùng đã đánh giá,

      $r_{u,i}$, $r_{v,i}$ là đánh giá của người dùng u và v đối với mục i.

    - Tìm hàng xóm gần nhất: Sau khi có ma trận tương đồng giữa các người dùng, hệ thống chọn ra một số người dùng gần nhất (hàng xóm) với người dùng mục tiêu dựa trên mức độ tương đồng.
    - Đưa ra đề xuất: Cuối cùng, hệ thống sử dụng hành vi và sở thích của hàng xóm gần nhất để đưa ra các đề xuất cho người dùng mục tiêu. Các mục mà hàng xóm đã thích và mà người dùng chưa thấy có thể được đề xuất cho người dùng mục tiêu.

$$\hat{r}_{u,i} = \frac{\sum_{v \in N(u)} \text{sim}(u,v) \cdot r_{v,i}}{\sum_{v \in N(u)} |\text{sim}(u,v)|}$$

Trong đó:

$\hat{r}_{u,i}$ là xếp hạng dự đoán của người dùng u đối với mục i,

N(u) là tập hợp các người dùng tương tự nhất với người dùng u.

	u_0	u_1	u_2	u_3	u_4	u_5	u_6
i_0	5	5	2	0	1	?	?
i_1	3	?	?	0	?	?	?
i_2	?	4	1	?	?	1	2
i_3	2	2	3	4	4	?	4
i_4	2	0	4	?	?	?	5

Ví dụ về utility matrix dựa trên số sao một user rate cho một item. Có thể thấy hành vi của u0 giống với u1 nhất, từ đó suy ra u0 sẽ thích i2 vì u1 cũng thích item này.

- Ưu điểm:
  - Dễ triển khai và hiệu quả.
  - Không cần thông tin về nội dung của mục.
  - Hiệu quả khi có ít dữ liệu.
- Nhược điểm:
  - Dễ bị ảnh hưởng bởi sự hiện diện của các người dùng nhiễu hoặc có sở thích đặc biệt.
  - Không hiệu quả khi có nhiều người dùng và mục.
- User-Based Collaborative Filtering thường được sử dụng trong các hệ thống có số lượng người dùng ít và sở thích của họ dễ dàng nhận biết được.

- Item-Based Collaborative Filtering: Trong Item-Based Collaborative Filtering, hệ thống đề xuất các mục dựa trên sự giống nhau giữa chúng và

các mục mà người dùng đã tương tác với trước đó. Quá trình hoạt động bao gồm các bước giống với User- based Collaborative Filtering.

- Ưu điểm
  - Thích hợp khi số lượng mục lớn hơn số lượng người dùng.
  - Tính toán độ phức tạp thấp hơn so với User-Based Collaborative Filtering.
- Nhược điểm:
  - Khó khăn trong việc xác định sự tương đồng giữa các mục.
  - Không hiệu quả khi có ít dữ liệu.
- Item-Based Collaborative Filtering thường được sử dụng trong các hệ thống có số lượng mục lớn và khi thông tin về người dùng ít và không đủ để tạo ra các đề xuất chất lượng cao.

❖ Content-Based Filtering (Lọc dựa trên nội dung): là một phương pháp trong hệ thống đề xuất, sử dụng thông tin về các đặc điểm của các mục để đưa ra các đề xuất cá nhân hóa cho người dùng. Ý tưởng chính là nếu một người dùng đã thích một mục nào đó, họ có khả năng cao sẽ thích các mục có đặc điểm tương tự.

- Ưu điểm:
  - Không phụ thuộc vào thông tin về người dùng khác.
  - Có thể đề xuất các mục mới hoặc không được phổ biến nhiều.
- Nhược điểm:
  - Đòi hỏi một mô hình đủ phức tạp để biểu diễn các mục.
  - Khó khăn trong việc xây dựng các vector biểu diễn mục cho các mục phức tạp.
- Content-Based Filtering thường được sử dụng trong các hệ thống mà thông tin về mục phong phú và sẵn có, và trong các trường hợp mà thông tin về người dùng ít và không đủ để tạo ra các đề xuất chất lượng cao.

❖ Hybrid recommendations approaches: là một chiến lược trong hệ thống đề xuất, kết hợp các phương pháp khác nhau như Collaborative Filtering, Content-Based Filtering, và các kỹ thuật khác để cải thiện chất lượng của các đề xuất. Bằng cách này, các hệ thống có thể tận dụng ưu điểm của từng phương pháp và giảm thiểu nhược điểm của chúng.

- Một số cách thức phổ biến để kết hợp các phương pháp:
  - Weighted Hybrid:
    - Hoạt động:

        - Mỗi phương pháp được gán một trọng số tương ứng.
        - Khi đưa ra đề xuất cho một người dùng, các đề xuất từ mỗi phương pháp được tính toán và sau đó được trọng số lại để tạo ra đề xuất cuối cùng.
    - Switching Hybrid:
        - Hoạt động:
            - Mỗi phương pháp được áp dụng độc lập để tạo ra các đề xuất.
            - Hệ thống sử dụng một cơ chế chuyển đổi để quyết định phương pháp nào nên được sử dụng để tạo ra đề xuất cuối cùng cho mỗi người dùng hoặc mục tiêu.
    - Feature Combination Hybrid:
        - Hoạt động:
            - Các đặc điểm của các phương pháp được kết hợp lại để tạo ra các đặc điểm mới.
            - Một mô hình học máy được sử dụng để học mối quan hệ giữa các đặc điểm mới này và sở thích của người dùng.
            - Dự đoán sở thích của người dùng dựa trên các đặc điểm mới này để tạo ra các đề xuất.
    - Cascade Hybrid:
        - Hoạt động:
            - Các phương pháp được sử dụng theo một chuỗi tuần tự.
            - Các đề xuất từ phương pháp đầu tiên được sử dụng làm đầu vào cho phương pháp thứ hai và tiếp tục cho đến khi đưa ra đề xuất cuối cùng.
- Ưu điểm:
    - Cải thiện chất lượng của các đề xuất bằng cách tận dụng ưu điểm của các phương pháp khác nhau.
    - Linh hoạt và có thể được điều chỉnh để phù hợp với nhu cầu cụ thể của hệ thống.
- Nhược điểm:
    - Đòi hỏi kiến thức chuyên môn cao để thiết kế và triển khai.
    - Tăng độ phức tạp của hệ thống.

4. Ưu, nhược điểm

- Ưu điểm:
  - Tăng trải nghiệm người dùng: Hệ thống đề xuất giúp cải thiện trải nghiệm của người dùng bằng cách đề xuất các mục phù hợp với sở thích và nhu cầu của họ, giúp họ khám phá và tiêu thụ nội dung mới một cách dễ dàng.
  - Tăng doanh số bán hàng: Trong các nền kinh tế dựa trên nền tảng như thương mại điện tử hoặc dịch vụ trực tuyến, các hệ thống đề xuất có thể giúp tăng doanh số bằng cách đề xuất các sản phẩm hoặc dịch vụ phù hợp với sở thích của người dùng.
  - Tăng sự tương tác: Bằng cách đề xuất các mục có thể quan tâm, hệ thống đề xuất có thể tăng sự tương tác của người dùng với nền tảng, dẫn đến sự tương tác tăng cường, sự trung thành và sự cam kết cao hơn.
  - Tiết kiệm thời gian và công sức: Thay vì phải tìm kiếm và lựa chọn các mục từ một danh sách lớn, người dùng có thể dễ dàng tìm thấy các mục phù hợp với sở thích của họ một cách nhanh chóng thông qua các đề xuất.
- Nhược điểm:
  - Hiệu suất không đồng đều: Hệ thống đề xuất có thể không luôn đưa ra các đề xuất chính xác, và có thể gặp phải vấn đề khi đưa ra các đề xuất cho người dùng có sở thích đặc biệt hoặc không phổ biến.
  - Hiệu quả dựa vào dữ liệu: Hiệu suất của hệ thống đề xuất phụ thuộc vào tính sẵn có và chất lượng của dữ liệu, đặc biệt là trong các hệ thống dựa trên lọc cộng tác.
  - Nguy cơ làm hạn chế mang tính định kiến: Nếu hệ thống không được thiết kế hoặc điều chỉnh đúng cách, có nguy cơ đề xuất một cách hạn chế, chỉ đưa ra các mục tương tự với những gì người dùng đã tiêu thụ trước đó, làm hạn chế sự đa dạng và khám phá.
  - Vấn đề quyền riêng tư và đạo đức: Thu thập và sử dụng dữ liệu cá nhân để tạo ra các đề xuất cá nhân hóa có thể gây ra các vấn đề về quyền riêng tư và đạo đức, đặc biệt là trong việc sử dụng dữ liệu cá nhân mà không được sự cho phép của người dùng.

5. Ứng dụng

Hệ thống đề xuất (recommendation system) đã có ứng dụng rộng rãi trong nhiều lĩnh vực khác nhau trong thực tế. Dưới đây là một số ví dụ tiêu biểu:

- Thương mại điện tử:

- Amazon: Sử dụng hệ thống đề xuất để đề xuất sản phẩm phù hợp với sở thích và hành vi mua hàng của khách hàng, giúp tăng doanh số bán hàng và cải thiện trải nghiệm người dùng.
- Netflix: Đề xuất các bộ phim và chương trình truyền hình dựa trên lịch sử xem của người dùng, đồng thời cung cấp đề xuất cá nhân hóa cho mỗi tài khoản.

- Dịch vụ streaming âm nhạc và video:
  - Spotify, Apple Music: Đề xuất các bài hát và danh sách phát dựa trên thị hiếu âm nhạc và lịch sử nghe nhạc của người dùng.
  - YouTube: Đề xuất video dựa trên lịch sử xem, thích và hành vi tương tác của người dùng.
- Mạng xã hội và truyền thông:
  - Facebook, Instagram, Twitter: Đề xuất nội dung và người dùng mới dựa trên sở thích, quan hệ mạng xã hội và hoạt động trước đó của người dùng.
  - Reddit: Đề xuất bài viết và nhóm dựa trên sở thích và hoạt động của người dùng trong cộng đồng.
- Du lịch và nhà hàng:
  - Booking.com, Airbnb: Đề xuất các chỗ ở phù hợp dựa trên lịch sử đặt phòng và sở thích của du khách.
  - TripAdvisor, Yelp: Đề xuất nhà hàng, điểm tham quan và hoạt động dựa trên đánh giá, bình luận và lịch sử tìm kiếm của người dùng.
- Tin tức và nội dung truyền thông:
  - Google News, Flipboard: Đề xuất tin tức và bài viết dựa trên sở thích đọc và lịch sử tìm kiếm của người dùng.
  - Medium: Đề xuất bài viết và tác giả dựa trên sở thích đọc và tương tác của người dùng.

Tài liệu tham khảo:

https://www.guru99.com/vi/supervised-vs-unsupervised-learning.html
https://pythonistaplanet.com/applications-of-unsupervised-learning/
https://hands-on.cloud/ml-unsupervised-learning-guide/
https://www.geeksforgeeks.org/k-means-clustering-introduction/
https://www.w3schools.com/python/python_ml_hierarchial_clustering.asp
https://www.analyticsvidhya.com/blog/2021/07/recommendation-system-understanding-the-basic-concepts/

https://www.nvidia.com/en-us/glossary/recommendation-system/
https://viblo.asia/p/tong-quan-ve-recommender-system-recommender-system-co-ban-phan-1-924lJGBb5PM
https://sloanreview.mit.edu/article/the-hidden-side-effects-of-recommendation-systems/
https://www.miquido.com/blog/perks-of-recommendation-systems-in-business/