Đại học quốc gia TP.HCM
Trường đại học công nghệ thông tin

Môn học: Khai phá dữ liệu trong doanh nghiệp

Lớp: DS317.P11

Báo cáo đề tài

GVHD: ThS. Nguyễn Thị Anh Thư

Nhóm sinh viên thực hiện:

Nguyễn Hữu Nam	MSSV: 22520917
Nguyễn Khánh	MSSV: 22520641
Võ Đình Khánh	MSSV: 22520659
Nguyễn Minh Sơn	MSSV: 22521254
Bùi Hồng Sơn	MSSV: 22521246

Mục lục

1. Tổng quan

Khai phá dữ liệu, đặc biệt là dữ liệu lớn, đã trở thành một lĩnh vực nghiên cứu quan trọng và thu hút sự quan tâm của các nhà khoa học trong những năm gần đây. Các ứng dụng của khai phá dữ liệu rất đa dạng, được triển khai trong nhiều lĩnh vực như kinh doanh, giáo dục, y tế, tài chính, và ngân hàng. Đặc biệt, khai phá dữ liệu trong giáo dục, cụ thể là khai phá dữ liệu lớn, đang là chủ đề thu hút nhiều nghiên cứu nhờ vào tính ứng dụng cao và tiềm năng cải thiện chất lượng giáo dục.

Trong bối cảnh giáo dục trực tuyến hiện nay, người học cần phải tự chủ động và có tinh thần tự giác cao do số lượng môn học đa dạng thuộc nhiều lĩnh vực khác nhau. Họ cần phân bổ thời gian học tập hợp lý cho từng nhóm môn học nhằm bổ sung và nâng cao kiến thức chuyên ngành cần thiết. Tuy nhiên, các nền tảng học tập trực tuyến thường không có ràng buộc cụ thể về thời gian và điểm số, dẫn đến tình trạng nhiều khóa học không được hoàn thành đúng thời hạn, thậm chí bị bỏ dở do người học mất hứng thú.

Vì vậy, công tác cố vấn học tập trên các nền tảng trực tuyến trở nên vô cùng quan trọng để giúp người học cải thiện hiệu suất học tập và gợi ý các khóa học phù hợp với nhu cầu cá nhân. Đây là một bài toán thuộc lĩnh vực khai phá dữ liệu, đặc biệt khi xử lý với số lượng lớn dữ liệu liên quan đến người học và hành vi học tập của họ. Việc nghiên cứu và xây dựng hệ thống khuyến nghị khóa học góp phần quan trọng vào việc cá nhân hóa trải nghiệm học tập, hỗ trợ người dùng lựa chọn các khóa học phù hợp với mục tiêu và nhu cầu học tập.

1.1. Định nghĩa và ngữ cảnh bài toán

Trong bối cảnh các nền tảng học tập trực tuyến, người học thường gặp khó khăn trong việc lựa chọn khóa học phù hợp. Điều này đặt ra nhu cầu xây dựng một hệ thống khuyến nghị giúp cá nhân hóa quá trình học tập của từng người. Bài toán được định nghĩa với đầu vào và đầu ra như sau:

- **Input**: Dữ liệu lớn từ các nền tảng học tập trực tuyến, bao gồm thông tin về người học, thông tin khóa học, và dữ liệu về các hoạt động học tập của người dùng.

- **Output**: Đề xuất top-k khóa học phù hợp nhất với người dùng (trong đó $k \in \mathbb{N}^*$, ví dụ trong nghiên cứu này $k = 10$).

1.2. Ứng dụng

Bài toán khuyến nghị khóa học cho các nền tảng học tập trực tuyến có nhiều ứng dụng thực tiễn, bao gồm:

- **Cá nhân hóa quá trình học tập**: Hệ thống giúp người học lựa chọn các khóa học phù hợp với nhu cầu và trình độ, từ đó cá nhân hóa lộ trình học tập.
- **Tăng tỷ lệ hoàn thành khóa học**: Đề xuất khóa học phù hợp giúp người học duy trì động lực học tập, từ đó tăng tỷ lệ hoàn thành khóa học.
- **Tối ưu hóa lộ trình học tập**: Gợi ý các khóa học tiếp theo dựa trên kỹ năng hiện tại và các khóa học đã hoàn thành.
- **Ứng dụng trong đào tạo doanh nghiệp**: Hỗ trợ doanh nghiệp xây dựng chương trình đào tạo nhân viên hiệu quả, phù hợp với mục tiêu phát triển nghề nghiệp.
- **Nâng cao hiệu quả sử dụng tài nguyên**: Giúp người học tiết kiệm thời gian và tập trung vào các khóa học có giá trị cao.

1.3. Khó khăn và thách thức

Mặc dù có nhiều tiềm năng, bài toán khuyến nghị khóa học vẫn gặp phải các khó khăn và thách thức như:

- **Chất lượng và sự đa dạng của dữ liệu**: Dữ liệu không đồng nhất hoặc không đầy đủ, gây khó khăn trong phân tích hành vi người học.
- **Xử lý dữ liệu lớn**: Khối lượng dữ liệu lớn đòi hỏi khả năng tính toán mạnh mẽ và các thuật toán tối ưu.
- **Lựa chọn đặc trưng quan trọng**: Việc chọn lọc các đặc trưng phù hợp từ bộ dữ liệu lớn đòi hỏi sự cân nhắc về tài nguyên và thời gian.

- **Đánh giá mô hình**: Thiếu dữ liệu rõ ràng về mức độ hài lòng của người học, khiến việc đánh giá hệ thống trở nên khó khăn.

1.4. Các nghiên cứu liên quan

Các phương pháp khuyến nghị chính bao gồm:

- **Matrix Factorization**: Phân rã ma trận để tìm các yếu tố tiềm ẩn.
- **Collaborative Filtering**: Sử dụng thông tin tương đồng giữa người dùng hoặc khóa học.
- **Content-Based Filtering**: Gợi ý dựa trên nội dung và đặc trưng của khóa học.
- **Hybrid Systems**: Kết hợp nhiều phương pháp để tăng hiệu quả.
- **Graph-Based Methods**: Sử dụng đồ thị để biểu diễn mối quan hệ giữa người dùng và khóa học.
- **Neural Collaborative Filtering**: Ứng dụng Deep Learning để học các tương tác phi tuyến.

2. Các công trình nghiên cứu liên quan

Trong lĩnh vực khuyến nghị khóa học, nhiều công trình nghiên cứu đã được thực hiện nhằm cải thiện hiệu quả và độ chính xác của các hệ thống khuyến nghị. Một số phương pháp nổi bật được đề xuất như sau:

2.1. Matrix Factorization

Matrix Factorization (MF) là một kỹ thuật phổ biến trong hệ thống khuyến nghị, được sử dụng để phân rã ma trận user-item nhằm phát hiện các yếu tố tiềm ẩn ảnh hưởng đến hành vi của người dùng. Koren và cộng sự (2009) đã giới thiệu phương pháp này và áp dụng vào hệ thống khuyến nghị, đặc biệt trong bài toán đề xuất phim. Phương pháp này cho phép mô hình hóa sự tương tác giữa người

dùng và các khóa học dựa trên các đặc trưng tiềm ẩn, mang lại kết quả tốt trong nhiều bài toán thực tế.

2.2. Collaborative Filtering

Collaborative Filtering (CF) là một trong những phương pháp lâu đời nhất và hiệu quả nhất trong khuyến nghị. CF được chia thành hai nhánh chính:

- **User-User Collaborative Filtering**: Dựa trên sự tương đồng giữa các người dùng để gợi ý các khóa học mà người dùng có thể quan tâm.
- **Item-Item Collaborative Filtering**: Dựa trên sự tương đồng giữa các khóa học để gợi ý các khóa học mới cho người dùng.

Su và Khoshgoftaar (2009) đã thực hiện một khảo sát toàn diện về CF và các phương pháp cải tiến nhằm khắc phục những hạn chế của nó, bao gồm vấn đề thưa thớt dữ liệu và mở rộng quy mô.

2.3. Content-Based Filtering

Content-Based Filtering (CBF) là phương pháp tập trung vào các đặc trưng của khóa học, dựa trên thông tin về nội dung của các khóa học mà người dùng đã tham gia để đưa ra gợi ý. Burke (2002) đã tổng hợp các phương pháp CBF và chỉ ra rằng việc kết hợp các đặc trưng nội dung của khóa học và hành vi người dùng có thể nâng cao hiệu quả gợi ý.

2.4. Hybrid Recommender Systems

Hybrid Recommender Systems là sự kết hợp giữa Collaborative Filtering và Content-Based Filtering nhằm khắc phục những hạn chế của từng phương pháp riêng lẻ. Burke (2002) đã thực hiện một nghiên cứu toàn diện về các hệ thống khuyến nghị lai, cho thấy rằng việc tích hợp các phương pháp có thể cải thiện hiệu quả và tính tổng quát của hệ thống.

2.5. Graph-Based Recommender Systems

Phương pháp dựa trên đồ thị (Graph-Based Recommender Systems) sử dụng cấu trúc đồ thị để biểu diễn mối quan hệ giữa người dùng và khóa học. Các thuật toán như Random Walk hoặc PageRank được áp dụng để tìm kiếm và khai thác các kết nối trong dữ liệu. Phương pháp này đặc biệt hiệu quả khi xử lý dữ liệu phức tạp với nhiều mối quan hệ đa chiều.

2.6. Neural Collaborative Filtering

Neural Collaborative Filtering (NCF) là một hướng tiếp cận hiện đại, kết hợp Collaborative Filtering với Deep Learning để học các tương tác phi tuyến giữa người dùng và khóa học. He và cộng sự (2017) đã giới thiệu mô hình NCF và cho thấy hiệu suất vượt trội của nó so với các phương pháp truyền thống trong các tập dữ liệu lớn và phức tạp.

2.7. Các hệ thống khuyến nghị khóa học dựa trên dữ liệu lớn

Trong bối cảnh dữ liệu lớn, các nền tảng như MOOCCubeX cung cấp một lượng dữ liệu khổng lồ về hành vi học tập của người dùng. Zhang và cộng sự (2022) đã phát triển một hệ thống khuyến nghị dựa trên dữ liệu từ MOOCCubeX, sử dụng các mô hình đồ thị để cá nhân hóa lộ trình học tập. Kết quả cho thấy hệ thống có khả năng gợi ý các khóa học phù hợp và tăng tỷ lệ hoàn thành khóa học.

2.8. Ứng dụng của các mô hình học sâu

Học sâu (Deep Learning) được áp dụng rộng rãi trong các hệ thống khuyến nghị khóa học hiện đại. Các mạng nơ-ron tích chập (Convolutional Neural Networks - CNNs) và mạng nơ-ron hồi tiếp (Recurrent Neural Networks - RNNs) được sử dụng để phân tích các chuỗi hành vi học tập của người dùng. Ngoài ra, Transformer và mô hình Attention cũng được triển khai để tối ưu hóa việc gợi ý dựa trên các đặc trưng tuần tự và ngữ cảnh.

2.9. Tổng quan các công trình nghiên cứu

Tóm lại, các phương pháp khuyến nghị khóa học đã có nhiều bước tiến đáng kể nhờ vào việc kết hợp các kỹ thuật truyền thống và hiện đại, từ Matrix Factorization đến Neural Collaborative Filtering. Các nghiên cứu không chỉ tập trung vào việc cải thiện độ chính xác mà còn nhấn mạnh đến khả năng mở rộng, xử lý dữ liệu lớn, và cá nhân hóa trải nghiệm học tập cho từng người dùng.

3. Cơ sở lý thuyết

3.1. Phương pháp áp dụng - KGAT

3.2. Phương pháp khác

3.2.1. Content-based Filtering

3.2.2. Matrix Factorization - Bayesian Personalized Ranking

3.2.3. Factorization Machine

Một trong những nhược điểm chính của mô hình BPRMF là không có khả năng mô hình hóa những thông tin bổ trợ giữa người dùng và sản phẩm. Do đó, một phương pháp mở rộng FM đã ra đời. Đây cũng là phương pháp nền móng cho các kĩ thuật DL được ra đời cho bài toán khuyến nghị.

Thuật toán khuyến nghị FM có thể mở rộng ra với các thông tin bổ trợ của người dùng và sản phẩm. Ví dụ như về khuyến nghị cho người dùng một bộ phim, ta có thể xét mức độ ảnh hưởng của các thông tin bổ trợ như: giới tính, tuổi, nghề nghiệp, ... Những thành phần này sẽ được mã hóa thành các vector one-hot hoặc multi-hot vector. Nếu có thêm các dữ liệu dạng số khác, ta có thể thêm vào $\mathbf{x}$ các thành phần tương ứng. Với mỗi thành phần được thêm vào $\mathbf{x}$, ta thêm một cột vector embedding vào $\mathbf{V}$ như hình bên dưới đây. Khi đó, độ quan tâm của người dùng có thể được dựng lên như sau:

$$\hat{y}_{ij} = w_0 + \mathbf{x}\mathbf{w} + \sum_{i=1}^{d} \sum_{j=i+1}^{d} \mathbf{v}_i^T \mathbf{v}_j x_i x_j$$

Trong đó:

- $\mathbf{w_0}$ đóng vai trò như một hệ số bias trong mô hình hồi quy tuyến tính, nó có thể được xem như là một hệ số vô hướng cố định thêm vào kết quả dự đoán cuối cùng để điều chỉnh sự lệch trung bình.
- **xw**: đây là tích vô hướng vector đặc trưng đầu vào (input feature vector) và một vector trọng số $\mathbf{w}$ tương ứng với các đặc trưng của $\mathbf{x}$.
- $\sum_{i=1}^{d} \sum_{j=i+1}^{d} \mathbf{v}_i^T \mathbf{v}_j x_i x_j$: Đây là thành phần tương tác bậc hai giữa các đặc trưng.
  - x_i, x_j lần lượt là phần tử thứ i, thứ j trong feature vector $\mathbf{x}$.
  - $\mathbf{v}_i^T \mathbf{v}_j$ là tích vô hướng giữa các vector embeddings tương ứng với từng đặc trưng đầu vào x_i và x_j.
  - $\sum_{i=1}^{d} \sum_{j=i+1}^{d} \mathbf{v}_i^T \mathbf{v}_j x_i x_j$ là biểu diễn tổng tất cả các cặp tương tác giữa các đặc trưng có trong tập dữ liệu.

Đây chính là ý tưởng chính của FM. Đồng thời, nhờ vào việc $\mathbf{x}$ thường là một vector rất thưa (rất ít thành phần khác 0), việc huấn luyện và dự đoán trở nên rất nhanh ngay cả khi số lượng người dùng và sản phẩm lớn.

3.2.4. Neutral Factorization Machine

4. Phương pháp đề xuất

5. Thực nghiệm

6. Kết luận và hướng phát triển