@@

A green and orange logo

Description automatically generated

**TẬP ĐOÀN VIỄN THÔNG QUÂN ĐỘI**

**TÀI LIỆU MÔ TẢ**

**Nghiên cứu đề xuất phương pháp khuyến nghị gói phim**

Hà Nội, 2023



# Mục lục

[Mục lục 2](#_Toc156835694)

[Danh mục hình ảnh 4](#_Toc156835695)

[Danh mục bảng biểu 6](#_Toc156835696)

[Phần 1. Đánh giá và tìm hiểu bài toán 7](#_Toc156835697)

[1.1. Tổng quan bài toán 7](#_Toc156835698)

[1.2. Đề xuất luồng giải pháp xây dựng mô hình 14](#_Toc156835699)

[1.3. Các vấn đề cần giải quyết 16](#_Toc156835700)

[Phần 2. Phân tích và xử lý dữ liệu 22](#_Toc156835701)

[2.1. Xử lý dữ liệu 22](#_Toc156835702)

[2.1.1. Tổng hợp các đặc trưng nhóm mô hình 22](#_Toc156835703)

[2.1.2. Tổng hợp các đặc trưng xử lý nhóm dữ liệu categorical 23](#_Toc156835704)

[2.1.3. Kết hợp dữ liệu và xử lý dữ liệu nhiễu 24](#_Toc156835705)

[2.1.4. Thực hiện chuẩn hóa scale 25](#_Toc156835706)

[2.2. Phân tích dữ liệu 26](#_Toc156835707)

[Phần 3. Các mô hình nghiên cứu đề xuất 29](#_Toc156835708)

[3.1. Mô hình đề xuất cơ sở mạng tích chập đồ thị 29](#_Toc156835709)

[3.1.1. Cơ sở lí thuyết 29](#_Toc156835710)

[3.1.2. Kết quả thực nghiệm và phân tích 33](#_Toc156835711)

[3.2. Mô hình đề xuất cơ sở mạng mở rộng Scale-Graph 39](#_Toc156835712)

[3.2.1. Cơ sở lí thuyết 39](#_Toc156835713)

[3.2.2. Kết quả thực nghiệm và phân tích 44](#_Toc156835714)

[3.3. Mô hình đề xuất cơ sở VAE-CF 46](#_Toc156835715)

[3.3.1. Cơ sở lí thuyết 46](#_Toc156835716)

[3.3.2. Kết quả thực nghiệm và phân tích 47](#_Toc156835717)

[3.4. Mô hình khuyến nghị mở rộng quy mô áp dụng kiến trúc Transformer theo phiên tương tác của khách hàng 51](#_Toc156835718)

[3.4.1. Xây dựng thuật toán và kiến trúc mô hình 51](#_Toc156835719)

[3.4.2. Thực nghiệm và phân tích 53](#_Toc156835720)

[3.5. Tổng hợp và phân tích 57](#_Toc156835721)

# Danh mục hình ảnh

[Hình 1.1. Hình mô tả ý tưởng lọc cộng tác truyền thống trong các hệ khuyến nghị phim và chương trình trên các nền tảng số 8](#_Toc148803408)

[Hình 1.2. Quy trình lọc cộng tác phổ biến trong các hệ khuyến nghị phim và chương trình trên các nền tảng số 8](#_Toc148803409)

[Hình 1.3. Một số vấn đề thiên khiến nổi bật trong các hệ khuyến nghị 9](#_Toc148803410)

[Hình 1.4. Một ví dụ về kỹ thuật ensemble các mô hình học máy trong các hệ khuyến nghị 10](#_Toc148803411)

[Hình 2.1. Đặc trưng về thông tin phim 22](#_Toc148803412)

[Hình 2.2. Đặc trưng về đánh giá của người dùng cho phim 23](#_Toc148803413)

[Hình 2.3. Một số đặc trưng về thông tin hồ sơ của người dùng 23](#_Toc148803414)

[Hình 2.4. Tổng hợp các đặc trưng mô tả thể loại của bộ phim 24](#_Toc148803415)

[Hình 2.5. Kết hợp dữ liệu tên và thể loại tạo trường thông tin mô tả phim 24](#_Toc148803416)

[Hình 2.6. Xử lý dữ liệu NaN trong ma trận tiện ích của người dùng và phim 25](#_Toc148803417)

[Hình 2.7. Kết quả trực quan hóa dữ liệu theo phân bố giới tính của người dùng 26](#_Toc148803418)

[Hình 2.8. Kết quả trực quan hoá dữ liệu theo phân bố độ tuổi và nghề nghiệp của người dùng 27](#_Toc148803419)

[Hình 2.9. Trực quan hoá dữ liệu theo số lượt được tương tác của mỗi loại phim 27](#_Toc148803420)

[Hình 2.10. Trực quan hoá dữ liệu theo điểm số được yêu thích của mỗi loại phim 28](#_Toc148803421)

[Hình 2.11. Trực quan hoá dữ liệu mô tả phân bố số lần tương tác của các người dùng với các bộ phim trong hệ thống 28](#_Toc148803422)

[Hình 3.1. Kiến trúc đồ thị người dùng-sản phẩm 29](#_Toc148803423)

[Hình 3.2. Minh họa huấn luyện LightGCN (trái) và UltraGCN (phải). LightGCN cần thực hiện lặp lại quá trình truyền thông báo 𝐿- lớp để có được phần nhúng cuối cùng cho quá trình đào tạo, trong khi UltraGCN có thể “bỏ qua” việc truyền thông báo đó để khiến các phần nhúng được đào tạo trực tiếp, phần lớn cải thiện hiệu quả huấn luyện và giúp tăng tốc độ triển khai thực tế. 31](#_Toc148803424)

[Hình 3.3. Minh họa các thành phần chính trong đồ thị người dùng-người dùng của mô hình UltraGCN 32](#_Toc148803425)

[Hình 3.4. Minh họa các thành phần chính trong đồ thị sản phẩm-sản phẩm của mô hình UltraGCN 32](#_Toc148803426)

[Hình 3.5. Biểu đồ minh họa sự thay đổi của hàm mất mát trong quá trình huấn luyện mô hình LightGCN - Round 1 37](#_Toc148803427)

[Hình 3.6. Minh họa kiến trúc mô hình IGCCF 39](#_Toc148803428)

[Hình 3.7. Biểu đồ minh họa hiệu suất của mô hình IGCCF trong quá trình các round trên bộ dữ liệu mẫu từ TV360 41](#_Toc148803429)

[Hình 3.8. Minh họa kiến trúc mô hình dự đoán ensemble 43](#_Toc148803430)

[Hình 3.9. Biểu đồ minh họa hiệu suất của mô hình IGCCF trong quá trình các round trên bộ dữ liệu mẫu từ TV360 49](#_Toc148803431)

# Danh mục bảng biểu

[Bảng 1.1. Tóm tắt các vấn đề cần giải quyết 17](#_Toc148803432)

[Bảng 3.1. Tổng hợp các Round thực hiện fine tune model UltraGCN 33](#_Toc148803433)

[Bảng 3.2. Tổng hợp các Round thực hiện fine tune model VAECF 37](#_Toc148803434)

[Bảng 3.3. Tổng hợp các Round thực hiện fine tune model EBPR 40](#_Toc148803435)

[Bảng 3.4. Tổng hợp các Round thực hiện fine tune model UltraGCN + VAECF 45](#_Toc148803436)

# Đánh giá và tìm hiểu bài toán

## Tổng quan bài toán

**Tổng hợp và mô tả bối cảnh bài toán xây dựng và khuyến nghị gói phim cho hệ thống TV360:**

Trong các dịch vụ quảng cáo điện tử, gói sản phẩm là một chiến lược tiếp thị quan trọng để hỗ trợ các chiến dịch quảng cáo, thu hút khách hàng và tăng doanh thu cho doanh nghiệp. Nó thường nhóm một tập hợp các sản phẩm liên quan mà người dùng tiêu thụ tổng thể trong các trường hợp, ví dụ: tổng giá giới hạn hoặc các mục đích cụ thể. Để minh họa, các gói phim hành động hướng đến người dùng có sở thích theo thể loại phim đó được mô tả trong Hình 1. Trong trường hợp này, họ có thể mong đợi một gói như vậy với một loạt các bộ phim liên quan phù hợp, thay vì những khuyến nghỉ phim lẻ rời rạc, không có mối liên hệ nào. Đặc biệt, trong các bối cảnh sự kiện như Giáng Sinh, Tết, v.v, các gói phim gồm các phim/ chương trình liên quan sẽ mang đến cho người dùng trải nghiệm trọn vẹn hơn. Đồng thời, nó thúc đẩy các nhóm sản phẩm ít được tương tác trong hệ thống, giúp giảm vấn đề “đuôi dài” trong kinh doanh.

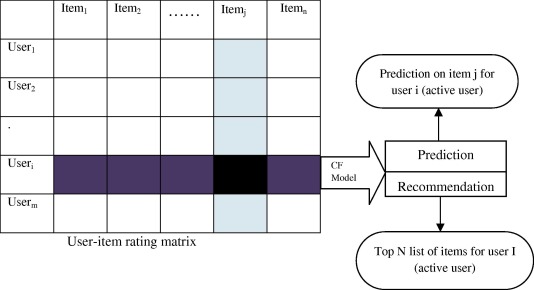
Như đã trình bày ở trên, các gói tổ chức và giới thiệu các sản phẩm có liên quan cao đến khách hàng và giúp họ tránh được những lựa chọn tẻ nhạt. Nó mang lại lợi ích cho cả khách hàng và người bán vì ít nhất những lý do sau. (1) Các gói có thể giúp nâng cao trải nghiệm người dùng theo nhiều cách khác nhau, ví dụ: các mặt hàng thay thế có thể tập hợp các sản phẩm liên quan để so sánh tốt hơn; trong khi các mặt hàng bổ sung có thể mở rộng tầm nhìn của người dùng để giúp thoát khỏi những lựa chọn đơn điệu. (2) Các gói có thể tăng doanh thu bán hàng cho TV360, ví dụ: bằng cách cho người dùng tiếp cận các mặt hàng mới mà họ có thể chưa từng xem xét riêng lẻ. (3) Chi phí mua/bán theo gói có thể thấp hơn so với việc mua/bán từng sản phẩm riêng lẻ cho khách hàng/người bán. Ví dụ: khuyến mãi và miễn phí giao hàng được cung cấp nếu sản phẩm trong một đơn hàng vượt quá một số tiền nhất định trong hệ thống thương mại điện tử [3]. Theo nghĩa này, gói là một giải pháp đôi bên cùng có lợi cho cả khách hàng và người bán. Mặc dù gần đây đã có một số phương pháp được đề xuất để đề xuất gói nhưng chúng vẫn có một số hạn chế cố hữu. 1) Hầu hết các phương pháp tiếp cận đều được xây dựng dựa trên các tập dữ liệu nhiễu, trong đó các gói được xác định bằng phương pháp phỏng đoán chưa được xác minh. Ví dụ, hầu hết các phương pháp trước đây chỉ đơn giản coi các sản phẩm được đồng mua là các gói tổng hợp, mặc dù thực tế là rất nhiều sản phẩm như vậy được tiếp cận chung mà không có mục đích cơ bản chung. Những người khác trực tiếp coi danh sách do người dùng tạo dưới dạng gói trong các miền cụ thể như phim lẻ, phim bộ và chương trình tạp kĩ. Ngoài việc bị giới hạn trong các lĩnh vực cụ thể đó, những chiến lược khuyến nghị đó chưa cố gắng đạt được sự hiểu biết về lý do căn bản đằng sau việc đóng gói sản phẩm. Một số công việc khác tận dụng các gói do nhà bán lẻ xác định. Tuy nhiên, việc có được những gói như vậy là một quá trình lâu dài, tốn nhiều công sức và tốn kém; và do đó, kích thước của các tập dữ liệu bó như vậy bị hạn chế. 2) Các nghiên cứu hiện tại chỉ giới hạn ở các nhiệm vụ cụ thể, ví dụ: chúng thường đi sâu trực tiếp vào nhiệm vụ đề xuất gói, với giả định phi thực tế rằng hệ thống có thể quan sát được các gói lịch sử của người dùng. Chúng tôi lập luận rằng để hỗ trợ đề xuất gói, cần phải thực hiện các bước trung gian như phát hiện các gói từ phiên của người dùng [35, 56]; hơn nữa, một tập hợp các tác vụ phụ trợ, ví dụ: tự động đặt tên gói, cần được xem xét để đề xuất gói thành công trong các ứng dụng thực tế. Trong nghiên cứu này, chúng tôi lùi lại một bước và xem xét quy trình đề xuất gói từ góc độ trải nghiệm người dùng toàn diện. Đầu tiên, để hiểu rõ hơn cơ sở lý luận đằng sau việc đóng gói sản phẩm, chúng tôi thiết kế một cách tinh tế một nhiệm vụ tìm nguồn cung ứng từ cộng đồng với mục tiêu tận dụng trí thông minh cũ của đám đông để giúp gắn nhãn các gói tiềm năng và ý định phản hồi ẩn trong phiên của người dùng ở ba miền (Điện tử, Quần áo và Thực phẩm). Thiết kế của chúng tôi lấy cảm hứng từ các nghiên cứu gần đây [35, 56] cho thấy người dùng có xu hướng khám phá các sản phẩm có mối tương quan cao (thay thế hoặc bổ sung) với mục đích chung trong một phiên [35, 56]. Kết quả là chúng tôi có được ba bộ dữ liệu chất lượng cao với thông tin meta phong phú, đặc biệt là ý định gói. Sau đó, bằng cách sử dụng dữ liệu đó, chúng tôi đề xuất một số nhiệm vụ quan trọng có liên quan với nhau để hỗ trợ đề xuất gói, như được minh họa trong Hình 2. Trên thực tế, người dùng có thể tương tác (ví dụ: nhấp hoặc mua) với nhiều mục trong một phiên [35, 56], với hoặc không có ý định chung. MỘT

Nếu người dùng đã ghé thăm một số nền tảng xem phim trực tuyến, hệ thống hướng đến mục tiêu có thể giúp người dùng thấy một hạng mục như "Đoán Xem Bạn Sẽ Thích" hiển thị danh sách các bộ phim thú vị mà có thể phù hợp với sở thích và mong muốn xem phim của họ. Đây chính là mục tiêu chính của hệ khuyến nghị nói chung và bài toán gợi ý phim nói riêng. Hệ thống khuyến nghị phim là một loại hệ thống lọc hướng đến việc cung cấp thông tin cá nhân hóa cho người dùng, từ đó cải thiện trải nghiệm của người dùng và thúc đẩy lợi nhuận kinh doanh.

Lọc cộng tác (Collaborative Filtering - CF) là phương pháp chính mà các hệ thống khuyến nghị sử dụng để "đoán" sở thích cuả người dùng. Phương pháp này giả định rằng người dùng sẽ quan tâm đến những sản phẩm mà những người dùng tương tự khác đã thể hiện sự quan tâm. Trong ngữ cảnh này, những người dùng tương tự là những người dùng khác đã tương tác lịch sử với chủ yếu những sản phẩm giống nhau mà người dùng chủ thể đã tương tác (Hình 1.1). Lọc cộng tác dự đoán độc lập với miền dữ liệu dành cho bối cảnh nội dung không thể được mô tả dễ dàng và đầy đủ bằng siêu dữ liệu. Kỹ thuật lọc cộng tác đã và đang được áp dụng một số hướng giải pháp trong hệ thống khuyến nghị phim và chương trình của nền tảng TV360. Các kỹ thuật lọc cộng tác hoạt động bằng cách xây dựng cơ sở dữ liệu (ma trận mục người dùng) về các tùy chọn đối với các mục của người dùng. Sau đó, nó kết hợp những người dùng có mối quan tâm và sở thích phù hợp bằng cách tính toán những điểm tương đồng giữa hồ sơ của họ để đưa ra đề xuất phim hoặc nội dung mà khách hàng có khả năng quan tâm kế tiếp. Những người dùng như vậy xây dựng một nhóm được gọi là vùng lân cận. Người dùng nhận được đề xuất về những mặt hàng mà trước đây anh ta chưa xếp hạng nhưng đã được người dùng ở khu vực lân cận đánh giá tích cực. Các khuyến nghị do CF đưa ra có thể là dự đoán hoặc khuyến nghị. Dự đoán là một giá trị số, , biểu thị điểm dự đoán của mục j cho người dùng i, trong khi Khuyến nghị là danh sách N mục hàng đầu mà người dùng sẽ thích nhất như trong Hình 1.2.



Hình mô tả ý tưởng lọc cộng tác truyền thống trong các hệ khuyến nghị phim và chương trình trên các nền tảng số



Quy trình lọc cộng tác phổ biến trong các hệ khuyến nghị phim và chương trình trên các nền tảng số

## Khảo sát các giải pháp liên quan

**Đầu vào**: Dữ liệu đầu vào được trích xuất và thống kê từ hệ quản trị cơ sở dữ liệu của hệ thống, bao gồm các dữ liệu sau:

* Định danh hồ sơ của người dùng
* Hành vi tương tác với phim của người dùng:
  + Thời gian tương tác
  + Thời điểm tương tác
  + Định danh phim được người dùng sử dụng
  + Mức độ và tần suất tương tác
* Thông tin của phim bao gồm:
  + Thông tin về phim hoặc tập phim
  + Thông tin về mô tả phim
  + Thông tin về thể loại phim
  + Thông tin về thời điểm phát hành phim

**Đầu ra**: Nghiên cứu đề xuất phương pháp nâng cao mô hình gợi ý phim khai thác các kỹ thuật kết hợp hiệu suất các mô hình để dự đoán phim tổng quan áp dụng cho nền tảng TV360. Mục tiêu kì vọng của các mô hình nghiên cứu đề xuất là phát triển nâng cao biểu diễn thông tin người dùng và phim dựa trên các tín hiệu cộng tác bậc cao, sau đó kết hợp các mô hình với hiệu suất khác nhau, nhằm đảm bảo tính tổng quan và chắc chắn của kết quả cuối cùng cho người dùng. Qua đó, với mỗi một người dùng (tài khoản) thuộc hệ thống sẽ được gợi ý các bộ phim phù hợp với nhu cầu và sở thích mong muốn của người dùng đó.

**Các mô hình nghiên cứu được khảo sát thử nghiệm và lựa chọn đề xuất**:

* Xây dựng mô hình mạng đồ thị tích chập cho khuyến nghị phim bộ
* Xây dựng mô hình VAE-CF Variational Autoencoders for Collaborative Filtering
* Xây dựng mô hình Explainable Bayesian Personalized Ranking
* Xây dựng kịch bản cho kỹ thuật ensemble các nhóm mô hình dựa trên bỏ phiếu ngưỡng cứng
* Xây dựng kịch bản cho kỹ thuật ensemble các nhóm mô hình dựa trên bỏ phiếu ngưỡng mềm

**Danh sách phương pháp xử lý các đặc trưng bao gồm**:

* xử lý dữ liệu nhiễu,
* xử lý các đặc trưng null,
* rời rạc hóa,
* chuẩn hóa,
* kết hợp đặc trưng, lựa chọn đặc trưng.

**Các phương pháp đánh giá chính**:

* Recall@K
* NDCG@K
* *(với K=30 và K=50).*

## Khảo sát các giải pháp liên quan

## Đề xuất luồng giải pháp xây dựng mô hình

Nhóm nghiên cứu tiến hành đánh giá tìm hiểu bài toán và phân tích các tập đặc trưng cần thiết cho mô hình phù hợp với dữ liệu đầu vào. Tập dữ liệu đầu vào có nhiều đặc trưng khác nhau bao gồm thông tin định danh người dùng, thông tin đặc trưng phim, hành vi tương tác, thời gian tương tác, yếu tố về độ thưa của dữ liệu. Các hành vi được xử lý, trực quan hóa và kiểm tra chất lượng theo từng phương pháp cụ thể trước khi đưa vào mô hình. Qua đó, triển khai các mô hình và phân tích kết quả thực nghiệm để lựa chọn giải pháp phù hợp với bối cảnh hiện tại và phát triển cho các giai đoạn tiếp theo của dự án.

**Phân tích và trực quan hóa dữ liệu:** Bước này cần tiến hành thống kê và khai thác các khía cạnh như số lượng người dùng, số lượng sản phẩm (phim), phân bố người dùng theo tần suất, các phân bố tương tác giữa người dùng và phim, v.v. Phân tích và trực quan hóa dữ liệu để có góc nhìn sâu rộng về hiện trạng dữ liệu, kết hợp với bước Tiền xử lý dữ liệu để đưa đầu vào phù hợp cho các mô hình.

**Tiền xử lý dữ liệu:** Làm sạch và xử lý trước dữ liệu để loại bỏ các giá trị bị thiếu, xử lý các giá trị ngoại lai và chuyển đổi dữ liệu thành định dạng có thể sử dụng được. Bước này là cần thiết để đảm bảo chất lượng của mô hình dự đoán. Cụ thể, do các đặc trưng thực tế có tính chất phức tạp, đa dạng, nên cần phân nhóm các dạng đặc trưng để tiền xử lý bao gồm các bước làm sạch và chuẩn hóa, xử lý giá trị NULL, rời rạc hóa, tạo ra các đặc trưng mới từ dữ liệu gốc để phù hợp với đầu vào mô hình.

**Xác định biến mục tiêu:** Xác định biến mục tiêu cho mô hình. Trong trường hợp này, để dự đoán danh sách các bộ phim có khả năng tương tác với người dùng cao nhất, ta có thể quy về bài toán dự đoán liên kết trong đồ thị với mỗi node trong đồ thị biểu hiện cho người dùng hoặc sản phẩm và học biểu diễn mỗi node dựa trên các kỹ thuật tính toán đồ thị phổ.

**Lựa chọn mô hình:** Chọn một mô hình khai thác các kỹ thuật dựa trên đồ thị phổ phù hợp để dự đoán khả năng tương tác giữa người dùng và phim. Qua quá trình khảo sát và nghiên cứu (phần 1.2), tập các mô hình giải pháp đa dạng được đề xuất dựa trên các kỹ thuật học biểu diễn và nâng cấp khác nhau, bao gồm:

* Xây dựng mô hình mạng đồ thị tích chập cho khuyến nghị phim bộ
* Xây dựng mô hình VAE-CF Variational Autoencoders for Collaborative Filtering
* Xây dựng mô hình Explainable Bayesian Personalized Ranking
* Xây dựng kịch bản cho kỹ thuật ensemble các nhóm mô hình dựa trên bỏ phiếu ngưỡng cứng
* Xây dựng kịch bản cho kỹ thuật ensemble các nhóm mô hình dựa trên bỏ phiếu ngưỡng mềm

**Huấn luyện mô hình:** Điều chỉnh mô hình đã chọn trên dữ liệu huấn luyện bằng cách sử dụng các tính năng và biến mục tiêu. Sau khi xây dựng các model based của các mô hình, thực hiện fine-tunning để điều chỉnh mô hình đã được huấn luyện ban đầu trở nên phù hợp hơn với tập dữ liệu đầu vào. Fine-tunning giúp cải thiện hiệu suất dự đoán và tăng độ tin cậy của mô hình. Thực hiện phân tích kết quả dự đoán để hiểu rõ hơn về hiệu suất và hành vi của mô hình. Xem xét các dự đoán sai và xác định nguyên nhân gây ra sự không chính xác, từ đó cải thiện mô hình.

**Đánh giá mô hình:** Đánh giá hiệu suất của mô hình bằng cách sử dụng các số liệu đánh giá thích hợp như Recall và NDCG với top-K các sản phẩm được khuyến nghị.

Điều quan trọng cần lưu ý là việc xây dựng một mô hình dự đoán chính xác đòi hỏi phải có kiến thức về miền, kiến thức chuyên môn về phân tích dữ liệu và xem xét cẩn thận các yếu tố khác nhau có thể ảnh hưởng đến các hoạt động của mô hình.

## Các vấn đề cần giải quyết

Trong giai đoạn đầu tiên, các công việc cần giải quyết được tóm tắt trong Bảng 1.1 dựa theo các vấn đề chính về dữ liệu, mục tiêu bài toán và mô hình đề xuất như sau:

* Xử lý dữ liệu
* Phân tích các đặc trưng dữ liệu
* Trực quan hóa dữ liệu
* Xác định các mô hình/ giải pháp phù hợp với bài toán
* Triển khai cài đặt các mô hình/ giải pháp đề xuất:
* Mô hình mạng đồ thị tích chập cho khuyến nghị phim bộ
* Mô hình VAE-CF Variational Autoencoders for Collaborative Filtering
* Mô hình Explainable Bayesian Personalized Ranking
* Các kỹ thuật ensemble các nhóm mô hình dựa trên bỏ phiếu ngưỡng cứng
* Các kỹ thuật ensemble các nhóm mô hình dựa trên bỏ phiếu ngưỡng mềm
* Phân tích kết quả thực nghiệm của các mô hình/ giải pháp đề xuất

Tóm tắt các vấn đề cần giải quyết

| **STT** | **Công việc** |
| --- | --- |
| 1 | Khảo sát các nghiên cứu liên quan |
| 2 | Đề xuất giải pháp xây dựng model |
| 3 | Tổng hợp features nhóm model |
| 4 | Tổng hợp features xử lý nhóm dữ liệu categorical |
| 5 | Thực hiện thống kê và feature engineering nhóm dữ liệu không hợp lệ |
| 6 | Thực hiện feature engineering chuẩn hóa dữ liệu |
| 7 | Thực hiện feature engineering xử lý nhóm dữ liệu categorical |
| 8 | Trực quan hóa và phân tích đặc trưng tương tác của các phim trong hệ thống |
| 9 | Trực quan hóa và phân tích đặc trưng hành vi sử dụng của người dùng trong hệ thống |
| 10 | Tạo các feature mới từ các đặc trưng hành vi người dùng |
| 11 | Tổng hợp và phân tích các phim bộ bias trong hệ thống trong các tập dữ liệu |
| 12 | Tổng hợp và xử lý các nhóm phim bộ ít tương tác trong các tập dữ liệu |
| 13 | Tổng hợp và xử lý các nhóm người dùng ít tương tác trong các tập dữ liệu |
| 14 | Tổng hợp và xử lý các nhóm người dùng ngoại lai trong các tập dữ liệu |
| 15 | Lựa chọn feature |
| 16 | Xây dựng mô hình cơ sở mạng đồ thị tích chập cho khuyến nghị phim bộ |
| 17 | Thực nghiệm mô hình cơ sở mạng đồ thị tích chập cho khuyến nghị phim bộ Round 1 |
| 18 | Thực nghiệm mô hình cơ sở mạng đồ thị tích chập cho khuyến nghị phim bộ Round 2 |
| 19 | Đánh giá kết quả thực nghiệm mô hình mạng đồ thị tích chập |
| 20 | Trực quan hóa kết quả thực nghiệm mô hình mạng đồ thị tích chập |
| 20 | Phân tích kết quả thực nghiệm mô hình mạng đồ thị tích chập |
| 21 | Xây dựng mô hình cơ sở VAE-CF cho khuyến nghị thiên khiến phổ biến |
| 22 | Thực nghiệm mô hình cơ sở VAE-CF cho khuyến nghị thiên khiến phổ biến Round 1 |
| 23 | Thực nghiệm mô hình cơ sở VAE-CF cho khuyến nghị thiên khiến phổ biến Round 2 |
| 24 | Đánh giá kết quả thực nghiệm mô hình cơ sở VAE-CF |
| 25 | Trực quan hóa kết quả thực nghiệm mô hình VAE-CF |
| 26 | Phân tích kết quả thực nghiệm mô hình cơ sở VAE-CF |
| 27 | Xây dựng mô hình cơ sở Explainable Bayesian Personalized Ranking EBPR cho khuyến nghị thiên khiến lựa chọn |
| 28 | Thực nghiệm mô hình cơ sở Explainable Bayesian Personalized Ranking cho khuyến nghị thiên khiến lựa chọn Round 1 |
| 29 | Thực nghiệm mô hình cơ sở Explainable Bayesian Personalized Ranking cho khuyến nghị thiên khiến lựa chọn Round 2 |
| 30 | Đánh giá kết quả thực nghiệm mô hình cơ sở EBPR |
| 31 | Trực quan hóa kết quả thực nghiệm mô hình cơ sở EBPR |
| 32 | Phân tích kết quả thực nghiệm mô hình cơ sở EBPR |
| 33 | Xây dựng các phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF |
| 34 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng cứng Round 1 |
| 35 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng cứng Round 2 |
| 36 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng cứng Round 3 |
| 37 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng mềm Round 1 |
| 38 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng mềm Round 2 |
| 39 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng mềm Round 3 |
| 40 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở VAE-CF dựa trên kỹ thuật voting ngưỡng mềm Round 4 |
| 41 | Tổng hợp và đánh giá kết quả phương pháp đề xuất |
| 42 | Phân tích kết quả phương pháp đề xuất |
| 43 | Xây dựng các phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR |
| 44 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng cứng Round 1 |
| 45 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng cứng Round 2 |
| 46 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng cứng Round 3 |
| 47 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 1 |
| 48 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 2 |
| 49 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 3 |
| 50 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 4 |
| 51 | Tổng hợp và đánh giá kết quả phương pháp đề xuất |
| 52 | Phân tích kết quả phương pháp đề xuất |
| 53 | Xây dựng các phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, mô hình cơ sở VAE-CF và mô hình cơ sở EBPR |
| 54 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, , mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng cứng Round 1 |
| 55 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, , mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng cứng Round 2 |
| 56 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, , mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng cứng Round 3 |
| 57 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 1 |
| 58 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 2 |
| 59 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 3 |
| 60 | Thực nghiệm phương pháp ensemble mô hình cơ sở mạng đồ thị tích chập, mô hình cơ sở VAE-CF và mô hình cơ sở EBPR dựa trên kỹ thuật voting ngưỡng mềm Round 4 |
| 61 | Tổng hợp và đánh giá kết quả phương pháp đề xuất 3 |
| 62 | Phân tích kết quả phương pháp đề xuất 3 |
| 63 | Tổng hợp và đánh giá kết quả của các mô hình cơ sở |
| 64 | Tổng hợp và đánh giá kết quả của các phương pháp ensemble đề xuất |
| 65 | Trực quan hóa thống kê danh sách phim bộ dự đoán bởi phương pháp ensemble 1 |
| 66 | Trực quan hóa thống kê danh sách phim bộ dự đoán bởi phương pháp ensemble 2 |
| 67 | Trực quan hóa thống kê danh sách phim bộ dự đoán bởi phương pháp ensemble 3 |
| 68 | Phân tích kết quả so sánh giữa các mô hình cơ sở và các mô hình ensemble đề xuất |

## 

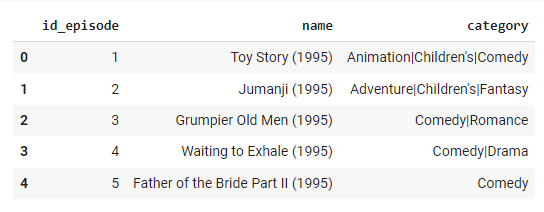
# Phân tích và xử lý dữ liệu

## Xử lý dữ liệu

Với nguồn dữ liệu đầu vào lớn, cần đánh giá tìm hiểu bài toán và phân tích các tập đặc trưng cần thiết cho mô hình. Tập dữ liệu đầu vào có nhiều đặc trưng khác nhau bao gồm hành vi sử dụng dịch vụ phim/chương trình của người dùng; các đặc trưng thông tin của phim (tên, thể loại, ngày sản xuất, ...); thông tin thời điểm và thời lượng tương tác của người dùng với phim. Các hành vi được xử lí, trực quan hóa và kiểm tra chất lượng theo từng phương pháp cụ thể trước khi đưa vào mô hình để lựa chọn những đặc trưng phù hợp nhất tương ứng với mỗi mô hình đề xuất.

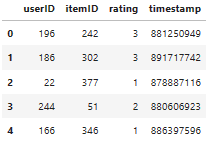
### Tổng hợp các đặc trưng nhóm mô hình

Các đặc trưng nhóm dữ liệu mô hình là tập hợp các thuộc tính hoặc thông tin mô tả về dữ liệu có tính chất nhóm hoặc tương tác trong mô hình học máy và học sâu. Việc xử lý các đặc trưng này cùng một lúc sẽ giúp cải thiện hiệu suất dự đoán của mô hình. Đặc trưng nhóm mô hình có thể kể đến: Các đặc trưng nhóm dữ liệu thông tin phim, hồ sơ người dùng và các đặc trưng nhóm đánh giá của người dùng qua phim.



Đặc trưng về thông tin phim

Đối với bộ dữ liệu mẫu để phục vụ nghiên cứu này, chúng ta quy ước *userID* và *itemID* tương ứng với mã định danh cho hồ sơ người dùng và phim (sản phẩm). Tương tác giữa người dùng và các bộ phim được xác định bằng đánh giá của người dùng với phim đó.



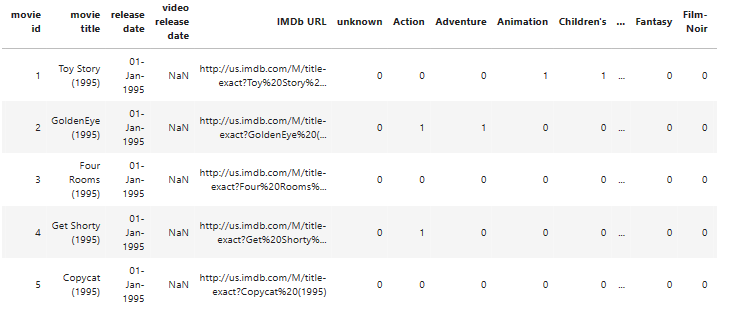
Đặc trưng về đánh giá của người dùng cho phim



Một số đặc trưng về thông tin hồ sơ của người dùng

### Tổng hợp các đặc trưng xử lý nhóm dữ liệu categorical

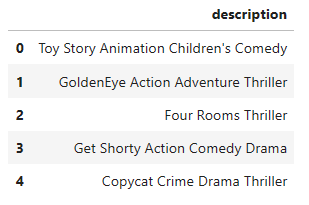
Trong học máy, dữ liệu categorical (biến phân loại) là loại dữ liệu mà các giá trị của nó thuộc vào một tập hợp hữu hạn các nhóm hoặc danh mục. Dữ liệu categorical không thể thực hiện các phép toán số học như các dữ liệu dạng số (numerical). Thay vào đó, chúng được mô tả bằng các nhãn hoặc tên đại diện cho từng danh mục hoặc nhóm riêng biệt. Đặc trưng (feature) của dữ liệu categorical là các biến đại diện cho các nhóm dữ liệu này. Việc xử lý nhóm dữ liệu categorical nhằm tạo biểu diễn đầu vào cho các mô hình thích hợp.



Tổng hợp các đặc trưng mô tả thể loại của bộ phim

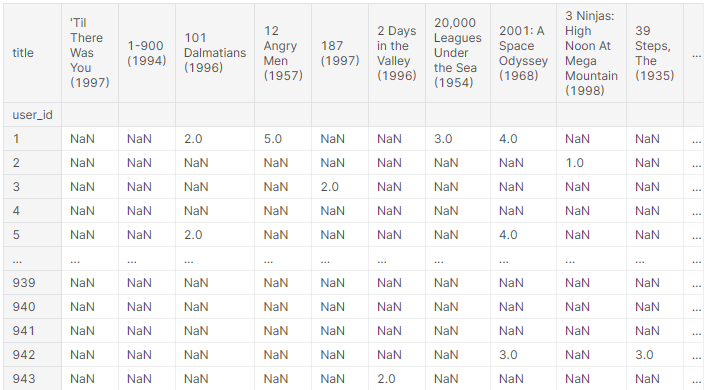
### Kết hợp dữ liệu và xử lý dữ liệu nhiễu

Ensemble data là một khái niệm được sử dụng khi ta kết hợp (ensemble) các tập dữ liệu từ nhiều nguồn khác nhau để tạo ra một tập dữ liệu lớn hơn hoặc cải thiện tính đại diện của dữ liệu. Ý tưởng chính là kết hợp dữ liệu từ nhiều nguồn để có được cái nhìn tổng quát hơn và đa dạng hơn về tập dữ liệu mục tiêu. Kết quả của việc kết hợp dữ liệu người dùng và phim như sau:



Kết hợp dữ liệu tên và thể loại tạo trường thông tin mô tả phim

Làm sạch và chuẩn hóa dữ liệu tiêu dùng là quá trình loại bỏ các giá trị lỗi, không chính xác hoặc thiếu sót trong dữ liệu và chuyển đổi nó thành định dạng chuẩn để phân tích và sử dụng. Cụ thể trong bài toán sẽ tìm các dòng dữ liệu NaN và xử lý các giá trị NaN đó.



Xử lý dữ liệu NaN trong ma trận tiện ích của người dùng và phim

Với dữ liệu văn bản mô tả phim, chúng ta cần đưa qua các mô hình ngôn ngữ để trích chọn các đặc trưng quan trọng thể hiện được ngữ nghĩa của văn bản mô tả ứng với mỗi phim. Trong quá trình thực nghiệm nghiên cứu này, nhóm nghiên cứu đề xuất sử dụng mô hình "Sentence-BERT: Nhúng câu bằng cách sử dụng mạng Siamese BERT." Những biểu diễn nhúng qua mô hình của mỗi thông tin mô tả phim có thể được so sánh độ tương đồng về mặt ngữ nghĩa, ví dụ như bằng cách sử dụng độ tương đồng cosine.

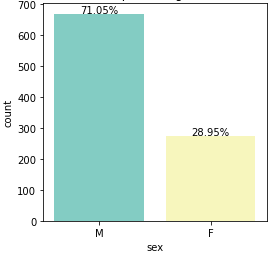
### Thực hiện chuẩn hóa scale

Mục tiêu của việc chuẩn hóa scale là chuyển đổi dữ liệu ban đầu thành dạng có trung bình bằng 0 và độ lệch chuẩn bằng 1, đảm bảo rằng các đặc trưng có phạm vi giá trị tương đương và không ảnh hưởng mạnh đến mô hình.

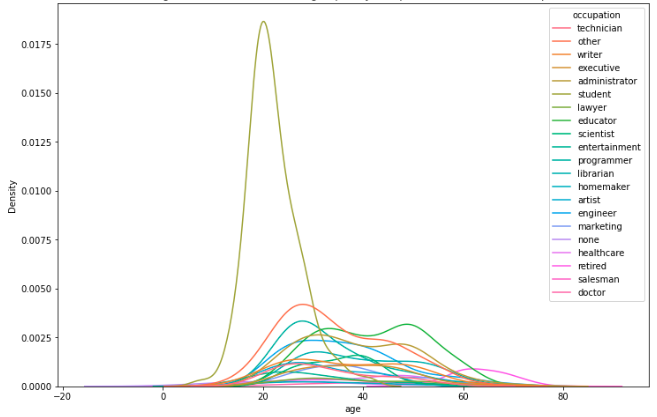
Chúng tôi sử dụng StandardScaler cho việc chuẩn hóa dữ liệu. Nó trừ đi trung bình của từng đặc trưng và chia kết quả cho độ lệch chuẩn của từng đặc trưng. Kết quả cuối cùng là dữ liệu có trung bình bằng 0 và độ lệch chuẩn bằng 1.

## Phân tích dữ liệu

Trong bài toán gợi ý phim sử dụng các mô hình gợi ý sản phẩm có tính cá nhân hóa với thông tin người dùng, nhóm nghiên cứu tiến hành một số bước trực quan hoá dữ liệu theo các thông tin hồ sơ và hành vi sử dụng sản phẩm của người dùng trong hệ thống.

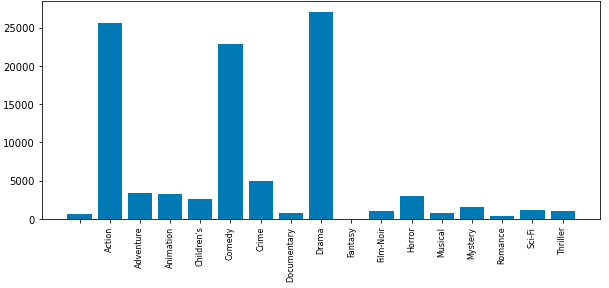


Kết quả trực quan hóa dữ liệu theo phân bố giới tính của người dùng

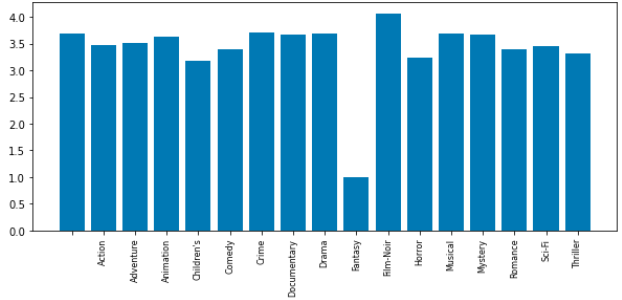


Kết quả trực quan hoá dữ liệu theo phân bố độ tuổi và nghề nghiệp của người dùng

Nhóm nghiên cứu tiến hành phân tích và xác định thể loại phim nào nhận được nhiều tương tác nhất (hình 2.9) và thể loại phim nào nhận được nhiều đánh giá tích cực nhất từ người dùng (hình 2.10).

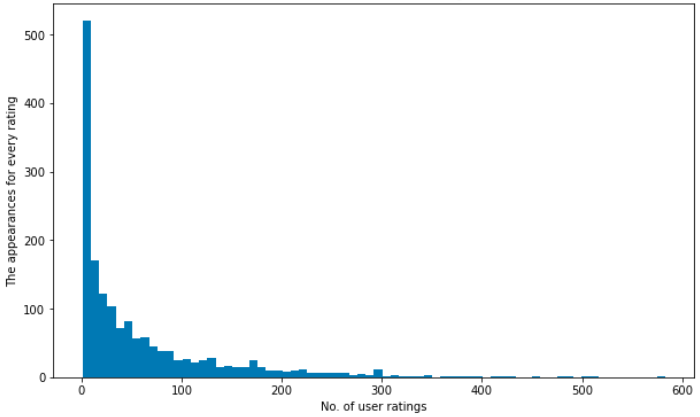


Trực quan hoá dữ liệu theo số lượt được tương tác của mỗi loại phim



Trực quan hoá dữ liệu theo điểm số được yêu thích của mỗi loại phim

Hình 2.11 mô tả phân bố số lần tương tác của các người dùng với các bộ phim trong hệ thống. Phần lớn các người dùng trong hệ thống chỉ tương tác dưới 10 bộ phim, và tồn tại một lượng không nhỏ người dùng đã tương tác hơn 100 bộ phim.



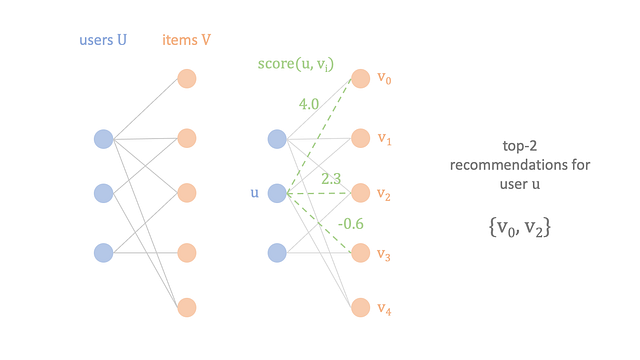
Trực quan hoá dữ liệu mô tả phân bố số lần tương tác của các người dùng với các bộ phim trong hệ thống

# Các mô hình nghiên cứu đề xuất

## Mô hình đề xuất cơ sở mạng tích chập đồ thị

### Cơ sở lí thuyết

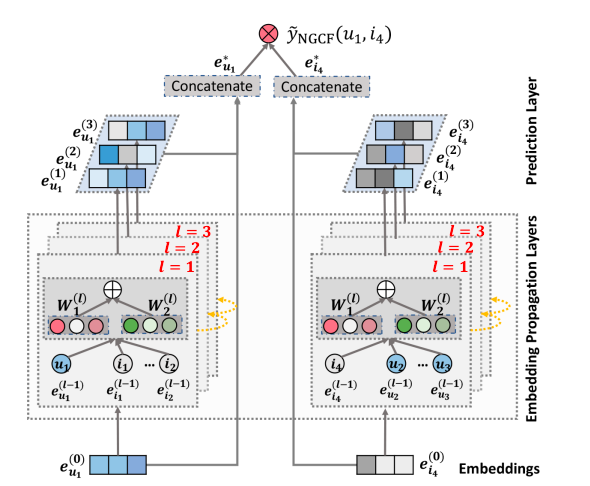
Một cách tốt nhất để hình dung tương tác trong một hệ thống gợi ý là sử dụng các kỹ thuật xây dựng đồ thị với cấu trúc người dùng và các sản phẩm (phim) là các node, và các cạnh giữa chúng biểu thị tương tác người dung và sản phẩm (hoặc các node là người dùng/sản phẩm, và cạnh thể hiện độ tương đồng giữa chúng). Những tương tác này có thể là người dùng đánh giá tích cực, hoặc click vào xem thông tin của sản phẩm (phim). Các mô hình đề xuất sẽ tận dụng cấu trúc đồ thị lưỡng phân để khai thác kết nối bậc cao và ý tưởng lọc cộng tác. Cấu trúc đồ thị này là cơ sở cho phép chúng ta nhận định cách xây dựng các mô hình mạnh mẽ, cụ thể như UltraGCN, GDE và IGCCF được thảo luận và đánh giá trong phần 3 này.



Kiến trúc đồ thị người dùng-sản phẩm

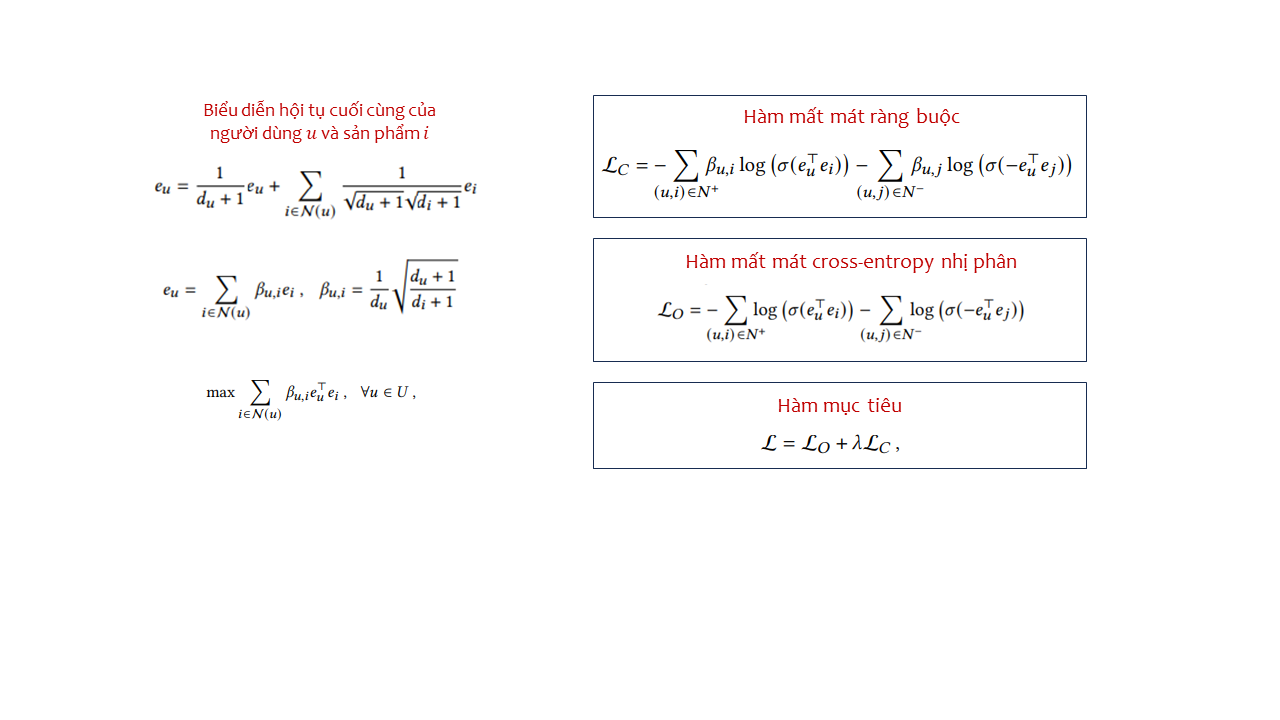
Việc xây dựng đồ thị người dùng – sản phẩm của các mô hình đề xuất đều dựa trên các node biểu thị cho người dùng hoặc sản phẩm, còn các cạnh thể hiện mối quan hệ tương tác giữa chúng (hình 3.1). Với dữ liệu đầu vào là một tập hợp người dùng và sản phẩm (hay tập hợp các bộ phim ), các mô hình hướng đến tạo ra giá trị thực scalar cho mỗi người dùng và phim . Đối với mỗi người dùng , chúng ta muốn gợi ý hạng mục có điểm số cao nhất mà họ chưa tương tác với (các mô hình đề xuất được đánh giá với và ).

Lấy cảm hứng từ những nghiên cứu tiên phong về quá trình huấn luyện các mô hình dựa trên GCN, chúng tôi nhận thấy rằng việc truyền thông điệp (tức là tổng hợp vùng lân cận) trên một đồ thị lớn thường tốn thời gian đối với CF. Đặc biệt, việc xếp chồng nhiều lớp truyền thông điệp có thể dẫn đến sự hội tụ chậm của các mô hình dựa trên GCN trong các tác vụ CF. Mặc dù các mô hình như LightGCN đã được đơn giản hóa cho việc huấn luyện, nhưng hoạt động truyền thông điệp vẫn chiếm ưu thế trong quá trình huấn luyện. Ví dụ: trong các thử nghiệm nghiên cứu, LightGCN ba lớp mất hơn 700 epochs để hội tụ đến kết quả tốt nhất trên tập dữ liệu Amazon Books, điều này vẫn khó có thể chấp nhận được trong môi trường công nghiệp. Làm thế nào để nâng cao hiệu quả của các mô hình GCN mà vẫn giữ được hiệu quả theo khuyến nghị vẫn là một bài toán mở. Để giải quyết thách thức này, trong nhiệm vụ này, chúng tôi nghiên cứu đề xuất về sự cần thiết của các lớp chuyển thông điệp rõ ràng trong CF và cuối cùng đề xuất một dạng GCN cực kỳ đơn giản (được đặt tên là UltraGCN) mà không cần chuyển mesage để có đề xuất hiệu quả. LightGCN được xác định ba hạn chế quan trọng: 1) Trọng số được gán trên các cạnh trong quá trình truyền thông điệp là phản trực giác, có thể không phù hợp với CF. 2) Quá trình nhân giống kết hợp đệ quy các loại cặp quan hệ khác nhau (bao gồm cặp người dùng-sản phẩm, cặp sản phẩm-sản phẩm và cặp người dùng-người dùng) vào mô hình, nhưng không nắm bắt được tầm quan trọng khác nhau của chúng. Điều này cũng có thể gây ra các mối quan hệ ồn ào và thiếu thông tin, gây nhầm lẫn cho việc đào tạo mô hình. 3) Vấn đề làm mịn quá mức đã hạn chế việc sử dụng quá nhiều lớp thông báo truyền trong LightGCN. Do đó, thay vì thực hiện truyền thông báo rõ ràng, UltraGCN tìm cách ước tính trực tiếp giới hạn của tích chập đồ thị lớp vô hạn thông qua việc mất ràng buộc, dẫn đến mô hình GCN cực kỳ đơn giản. Thiết kế dựa trên tổn thất của UltraGCN rất linh hoạt, cho phép chúng tôi điều chỉnh thủ công tầm quan trọng tương đối của các loại mối quan hệ khác nhau và cũng tránh được vấn đề làm mịn quá mức bằng cách lấy mẫu âm tính. Điều này cuối cùng mang lại một mô hình UltraGCN đơn giản nhưng hiệu quả, dễ thực hiện và đào tạo hiệu quả. Hơn nữa, chúng tôi cho thấy Ultra GCN đạt được những cải tiến đáng kể so với các mẫu CF hiện đại. UltraGCN thường đạt được hiệu suất tăng tốc hơn 10 lần khi huấn luyện với LightGCN.



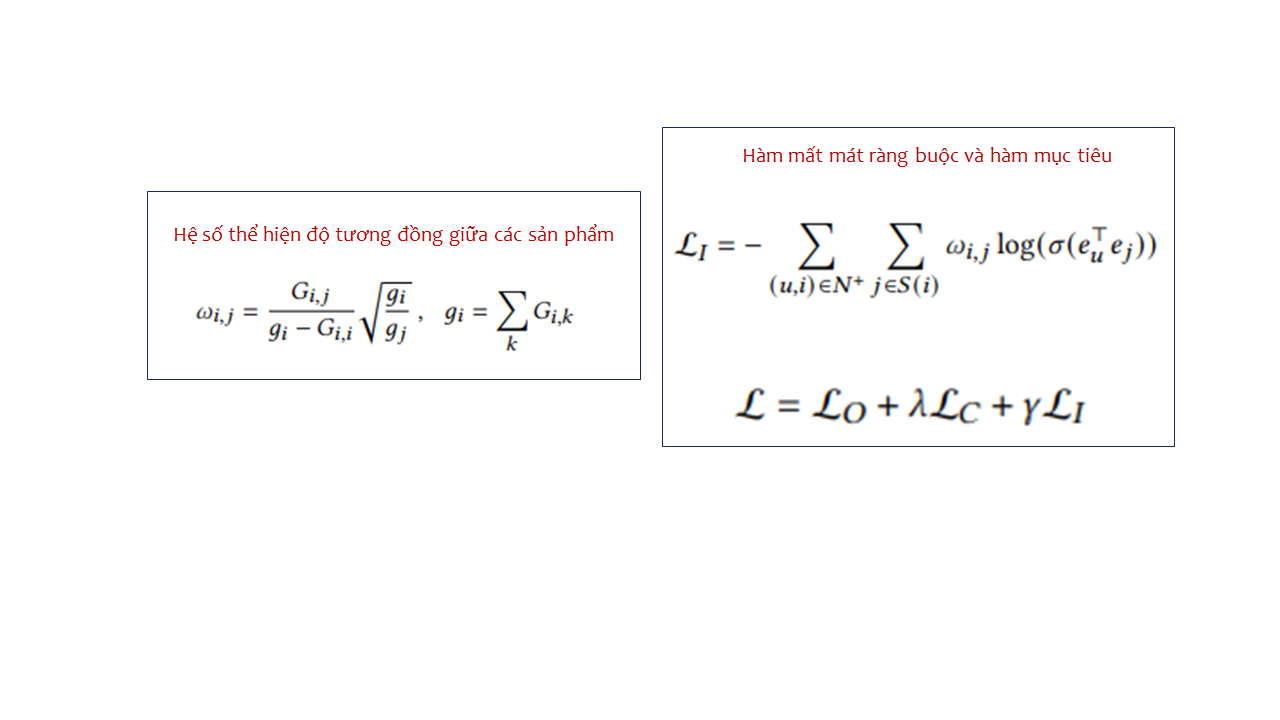
Minh họa huấn luyện LightGCN (trái) và UltraGCN (phải). LightGCN cần thực hiện lặp lại quá trình truyền thông báo 𝐿- lớp để có được phần nhúng cuối cùng cho quá trình đào tạo, trong khi UltraGCN có thể “bỏ qua” việc truyền thông báo đó để khiến các phần nhúng được đào tạo trực tiếp, phần lớn cải thiện hiệu quả huấn luyện và giúp tăng tốc độ triển khai thực tế.

UltraGCN bao gồm hai thành phần chính: học biểu diễn từ đồ thị người dùng-người dùng, và học biểu diễn trên đồ thị sản phẩm-sản phẩm.



Minh họa các thành phần chính trong đồ thị người dùng-người dùng của mô hình UltraGCN

UltraGCN không dựa vào việc truyền thông điệp rõ ràng để có thể tìm hiểu riêng các mối quan hệ khác theo cách linh hoạt hơn. Điều này cũng cho phép mô hình được điều chỉnh tầm quan trọng tương đối của các mối quan hệ khác nhau theo cách thủ công. Đặc biệt, ta cần nhấn mạnh rằng UltraGCN có thể mở rộng linh hoạt để mô hình hóa nhiều đồ thị quan hệ khác nhau, chẳng hạn như đồ thị người dùng-người dùng, đồ thị sản phẩm và thậm chí cả đồ thị tri thức. UltraGCN chủ yếu chứng minh việc sử dụng nó trên đồ thị xảy ra đồng thời giữa sản phẩm- sản phẩm, đồ thị này đã được chứng minh là hữu ích cho đề xuất. Trước tiên, chúng tôi xây dựng đồ thị sự xuất hiện của sản phẩm - sản phẩm bằng cách liên kết các mục có sự xuất hiện đồng thời, tạo ra ma trận liền kề có trọng số sau 𝐺 ∈ R|𝐼 |× |𝐼 | trong đó mỗi mục biểu thị sự xuất hiện đồng thời của hai sản phẩm.



Minh họa các thành phần chính trong đồ thị sản phẩm-sản phẩm của mô hình UltraGCN

### Kết quả thực nghiệm và phân tích

Nhóm nghiên cứu tiến hành xây dựng và thực nghiệm mô hình Đơn giản hóa cực độ mạng tích chập đồ thị UltraGCN trên bộ dữ liệu mẫu từ TV360 với khoảng 1.000.000 tương tác giữa khách hàng và các bộ phim. Mô hình được cài đặt thực nghiệm sử dụng các gói hỗ trợ sau:

* python 3.7.9
* pytorch 1.4.0
* numpy 1.19.2
* scipy 1.1.0
* tensorboard 2.4.0

Kết quả thực nghiệm được tổng hợp với 5 kịch bản khác nhau và thu kết quả cuối cùng đánh giá trên Recall@K và NDCG@K (đánh giá chính trên độ đo Recall@30). Các kết quả thực nghiệm qua các Round được mô tả và tổng hợp chi tiết dưới đây.

**Round** **1**:

Mô hình khởi tạo mặc định với áp dụng phân phối Gaussian với giá trị trung bình là 0 và độ lệch chuẩn 10−4 để khởi tạo các phần nhúng với số chiều là 128. Mô hình áp dụng chính quy hóa 𝐿2 với trọng số 10−4 và đặt tốc độ học thành 10−3, kích thước batch là 1024, tỷ lệ lấy mẫu âm 𝑅 là 200 và kích thước của bộ lân cận 𝐾 là 10.

**Round** **2**:

Trong round 2, mô hình được sửa kích thước nhúng thành 64, tương đương với các thực nghiệm dựa trên GCN gần đây và các mô hình GDE, IGCCF để giữ nguyên mức số lượng tham số để so sánh công bằng.

Kết quả thực nghiệm được tổng hợp ở bảng bên dưới:

Tổng hợp các Round thực hiện fine tune model UltraGCN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall@30** | **Recall@50** | **NDCG@30** | **NDCG@50** |
| **Round 1** | 0.2763 | 0.1885 | 0.3493 | 0.5410 |
| **Round 2** | 0.2665 | 0.1844 | 0.3335 | 0.5305 |

Qua kết quả tổng hợp ở bảng 3.1, ta có thể thấy hiệu suất của mô hình thể hiện như sau:

* Với bộ siêu tham số khởi tạo cơ sở cho mô hình NGCF ở Round 1 mang lại kết quả tốt nhất xét theo toàn diện các độ đo. Tuy nhiên, với mỗi độ đo, các kết quả của Round 1 chỉ thể hiện cao hơn kết quả xếp thứ hai khoảng
* Với các siêu tham số đã được cố định, việc tăng thêm giá trị batch size hoặc tăng trọng số phân rã có thể giúp mô hình đạt được hiệu suất tốt hơn.
* Bên cạnh đó, việc tăng thêm một lớp tính toán cho đồ thị hoặc tăng thêm số chiều ẩn có thể khiến hiệu năng mô hình suy giảm.

Vì vậy, cuối cùng Pinterest đã triển khai PinSage làm hệ thống đề xuất của họ và mọi người thực sự vui mừng về sự thay đổi cũng như sự thay đổi trong mô hình. Nhưng nó thực sự là gì và nó có liên quan như thế nào với GraphSAGE?

Nếu chúng ta nhìn vào tiêu đề PinSage, nó sẽ như sau

Mạng thần kinh tích chập đồ thị cho các hệ thống đề xuất quy mô web

Họ đang đề cập đến điều gì trong tiêu đề của họ? Hãy để tôi chia tiêu đề thành ba từ khóa chính

Mạng nơ-ron tích chập đồ thị

Quy mô web

Hệ thống gợi ý

Mạng nơ-ron tích chập đồ thị

Viết tắt của GCN, là một mô hình trong đó cả Đồ thị và Mạng thần kinh đều xuất hiện, điều này đang trở nên thú vị phải không? Về cơ bản, GCN rất hữu ích cho việc biểu diễn tính năng của các nút trong biểu đồ bằng Mạng thần kinh bằng cách xem xét các nút lân cận trên k-hop. Tích chập được đề cập ở đây không phải là tích chập thông thường mà chúng ta thấy trong Deep Learning với hình ảnh, nhưng khái niệm này vẫn giống nhau.

Hệ thống đề xuất quy mô web

Điều này có nghĩa là họ đang phát triển GCN cho quy mô web, nghĩa là dành cho một biểu đồ lớn và cũng là một hệ thống đề xuất, có nghĩa là họ đang cố gắng tạo ra một số loại biểu diễn các nút mà họ đào tạo trên một và sau đó họ sử dụng biểu diễn đã được đào tạo để đề xuất nhiệm vụ cho các nút không nhìn thấy được.

PinSage

Hãy để tôi giới thiệu một chút về PinSage là gì. PinSage là một Hệ thống đề xuất được xây dựng dựa trên GraphSAGE và hệ thống này phát triển các phần nhúng nút chất lượng cao bằng cách sử dụng cấu trúc biểu đồ và thông tin tính năng của nút. Và mô hình này hiện đang được triển khai trên Pinterest.

Hãy để tôi chọn một đoạn văn nhỏ từ tờ giấy PinSage, nó có nội dung như sau

GraphSAGE là một biến thể quy nạp của GCN mà chúng tôi sửa đổi để tránh hoạt động trên toàn bộ đồ thị Laplacian. Về cơ bản, chúng tôi cải tiến GraphSAGE bằng cách loại bỏ giới hạn lưu trữ toàn bộ biểu đồ trong bộ nhớ GPU, sử dụng các bước đi ngẫu nhiên có độ trễ thấp để lấy mẫu các vùng lân cận biểu đồ trong kiến trúc nhà sản xuất-người tiêu dùng. — PinSage

So finally Pinterest has deployed PinSage as their recommender system and everyone is really excited about the change and the shift in paradigm. But what is it really and how is it related to GraphSAGE?

If we look at PinSage title it reads as follows

Graph Convolutional Neural Networks for Web-Scale Recommender Systems

What are they mentioning in their title? Let me break down the title down into three main keywords

Graph Convolutional Neural Networks

Web-Scale

Recommender System

Graph Convolutional Neural Networks

Short for GCN, is a model in which both Graphs and Neural Networks comes into picture, already this is getting exciting right? GCN are basically useful for feature representation of nodes in a graph using Neural Networks by taking into consideration of neighbours on k-hops. The convolution mentioned here is not a normal convolution that we see in Deep Learning with images, but the concept is same.

Web-Scale Recommender System

This is saying that they are developing a GCN for web-scale which means for a large graph and also that it is a recommender system, which means that they are trying to generate some kind of representation of nodes in which they train on one and then they use the trained representation for recommending tasks for unseen nodes.

PinSage

Let me give a small intro on what PinSage is. PinSage is a Recommender System which is built on top of GraphSAGE, and this develops high quality node embeddings using the graph structure and the node feature information. And this model is currently deployed in Pinterest.

Let me pick a small paragraph from PinSage paper, it reads as follows

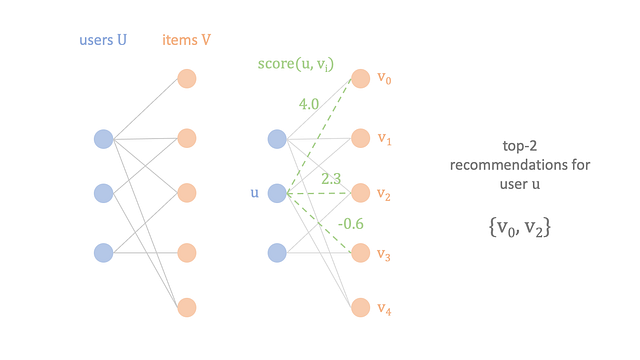
GraphSAGE is an inductive variant of GCNs that we modify to avoid operating on the entire graph Laplacian. We fundamentally improve upon GraphSAGE by removing the limitation that the whole graph be stored in GPU memory, using low-latency random walks to sample graph neighbourhoods in a producer-consumer architecture. — PinSage

## Mô hình đề xuất cơ sở mạng mở rộng Scale-Graph

### Cơ sở lí thuyết

Những tiến bộ gần đây trong mạng mạng nơ ron học sâu dành cho dữ liệu có cấu trúc biểu đồ đã mang lại hiệu suất vượt trội trên các điểm chuẩn của hệ thống đề xuất. Tuy nhiên, việc làm cho các phương pháp này trở nên thiết thực và có thể mở rộng cho các nhiệm vụ đề xuất ở quy mô web với hàng tỷ mục và hàng trăm triệu người dùng vẫn là một thách thức. Ở đây chúng tôi mô tả một công cụ đề xuất sâu quy mô lớn mà chúng tôi đã phát triển và triển khai tại Pinterest. Chúng tôi phát triển thuật toán PinSage Mạng chuyển đổi đồ thị (GCN) hiệu quả về dữ liệu, kết hợp các bước đi ngẫu nhiên hiệu quả và tích chập đồ thị để tạo ra các phần nhúng của các nút (tức là các mục) kết hợp cả cấu trúc biểu đồ cũng như thông tin tính năng của nút. So với các phương pháp GCN trước đây, chúng tôi phát triển một phương pháp mới dựa trên các bước đi ngẫu nhiên hiệu quả cao để cấu trúc các cấu trúc và thiết kế một chiến lược đào tạo mới dựa trên các ví dụ đào tạo ngày càng khó hơn để cải thiện độ mạnh mẽ và độ hội tụ của mô hình. Chúng tôi triển khai PinSage tại Pinterest và đào tạo nó trên 7,5 tỷ ví dụ trên biểu đồ có 3 tỷ nút biểu thị các ghim và bảng cũng như 18 tỷ cạnh. Theo số liệu ngoại tuyến, nghiên cứu người dùng và thử nghiệm A/B, PinSage tạo ra các đề xuất chất lượng cao hơn so với các lựa chọn thay thế dựa trên biểu đồ và học sâu tương đương. Theo hiểu biết của chúng tôi, đây là ứng dụng nhúng đồ thị sâu lớn nhất cho đến nay và mở đường cho một thế hệ hệ thống đề xuất quy mô web mới dựa trên kiến trúc tích chập đồ thị

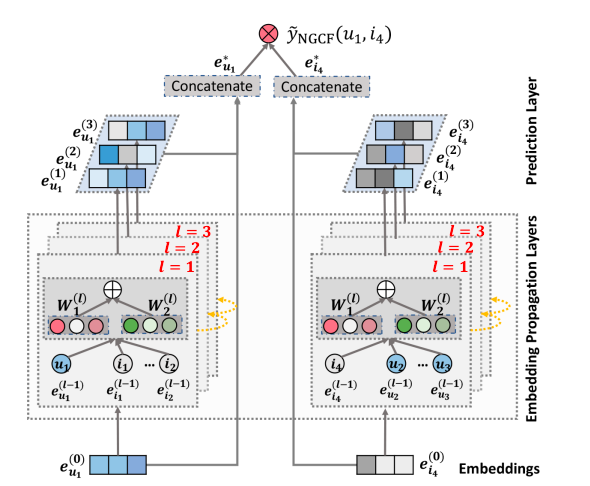
Một cách tốt nhất để hình dung tương tác trong một hệ thống gợi ý là sử dụng các kỹ thuật xây dựng đồ thị với cấu trúc người dùng và các sản phẩm (phim) là các node, và các cạnh giữa chúng biểu thị tương tác người dung và sản phẩm (hoặc các node là người dùng/sản phẩm, và cạnh thể hiện độ tương đồng giữa chúng). Những tương tác này có thể là người dùng đánh giá tích cực, hoặc click vào xem thông tin của sản phẩm (phim). Các mô hình đề xuất sẽ tận dụng cấu trúc đồ thị lưỡng phân để khai thác kết nối bậc cao và ý tưởng lọc cộng tác. Cấu trúc đồ thị này là cơ sở cho phép chúng ta nhận định cách xây dựng các mô hình mạnh mẽ, cụ thể như UltraGCN, GDE và IGCCF được thảo luận và đánh giá trong phần 3 này.



Kiến trúc đồ thị người dùng-sản phẩm

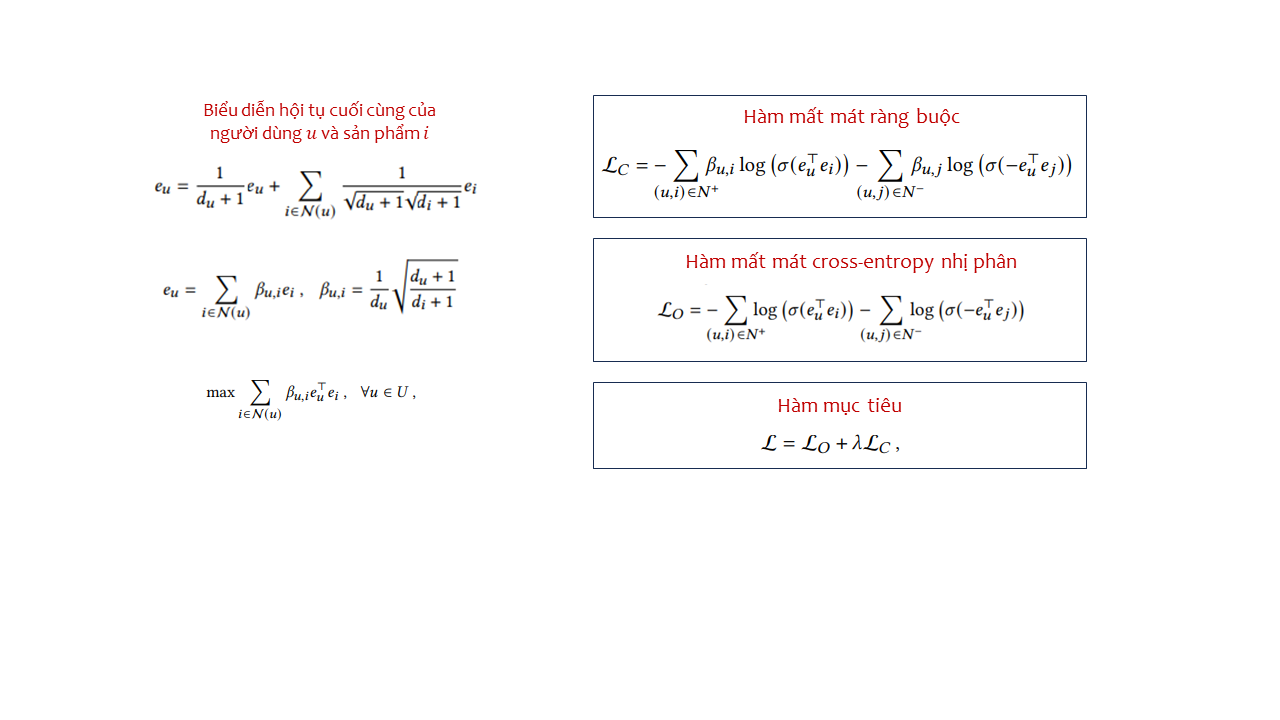
Việc xây dựng đồ thị người dùng – sản phẩm của các mô hình đề xuất đều dựa trên các node biểu thị cho người dùng hoặc sản phẩm, còn các cạnh thể hiện mối quan hệ tương tác giữa chúng (hình 3.1). Với dữ liệu đầu vào là một tập hợp người dùng và sản phẩm (hay tập hợp các bộ phim ), các mô hình hướng đến tạo ra giá trị thực scalar cho mỗi người dùng và phim . Đối với mỗi người dùng , chúng ta muốn gợi ý hạng mục có điểm số cao nhất mà họ chưa tương tác với (các mô hình đề xuất được đánh giá với và ).

Lấy cảm hứng từ những nghiên cứu tiên phong về quá trình huấn luyện các mô hình dựa trên GCN, chúng tôi nhận thấy rằng việc truyền thông điệp (tức là tổng hợp vùng lân cận) trên một đồ thị lớn thường tốn thời gian đối với CF. Đặc biệt, việc xếp chồng nhiều lớp truyền thông điệp có thể dẫn đến sự hội tụ chậm của các mô hình dựa trên GCN trong các tác vụ CF. Mặc dù các mô hình như LightGCN đã được đơn giản hóa cho việc huấn luyện, nhưng hoạt động truyền thông điệp vẫn chiếm ưu thế trong quá trình huấn luyện. Ví dụ: trong các thử nghiệm nghiên cứu, LightGCN ba lớp mất hơn 700 epochs để hội tụ đến kết quả tốt nhất trên tập dữ liệu Amazon Books, điều này vẫn khó có thể chấp nhận được trong môi trường công nghiệp. Làm thế nào để nâng cao hiệu quả của các mô hình GCN mà vẫn giữ được hiệu quả theo khuyến nghị vẫn là một bài toán mở. Để giải quyết thách thức này, trong nhiệm vụ này, chúng tôi nghiên cứu đề xuất về sự cần thiết của các lớp chuyển thông điệp rõ ràng trong CF và cuối cùng đề xuất một dạng GCN cực kỳ đơn giản (được đặt tên là UltraGCN) mà không cần chuyển mesage để có đề xuất hiệu quả. LightGCN được xác định ba hạn chế quan trọng: 1) Trọng số được gán trên các cạnh trong quá trình truyền thông điệp là phản trực giác, có thể không phù hợp với CF. 2) Quá trình nhân giống kết hợp đệ quy các loại cặp quan hệ khác nhau (bao gồm cặp người dùng-sản phẩm, cặp sản phẩm-sản phẩm và cặp người dùng-người dùng) vào mô hình, nhưng không nắm bắt được tầm quan trọng khác nhau của chúng. Điều này cũng có thể gây ra các mối quan hệ ồn ào và thiếu thông tin, gây nhầm lẫn cho việc đào tạo mô hình. 3) Vấn đề làm mịn quá mức đã hạn chế việc sử dụng quá nhiều lớp thông báo truyền trong LightGCN. Do đó, thay vì thực hiện truyền thông báo rõ ràng, UltraGCN tìm cách ước tính trực tiếp giới hạn của tích chập đồ thị lớp vô hạn thông qua việc mất ràng buộc, dẫn đến mô hình GCN cực kỳ đơn giản. Thiết kế dựa trên tổn thất của UltraGCN rất linh hoạt, cho phép chúng tôi điều chỉnh thủ công tầm quan trọng tương đối của các loại mối quan hệ khác nhau và cũng tránh được vấn đề làm mịn quá mức bằng cách lấy mẫu âm tính. Điều này cuối cùng mang lại một mô hình UltraGCN đơn giản nhưng hiệu quả, dễ thực hiện và đào tạo hiệu quả. Hơn nữa, chúng tôi cho thấy Ultra GCN đạt được những cải tiến đáng kể so với các mẫu CF hiện đại. UltraGCN thường đạt được hiệu suất tăng tốc hơn 10 lần khi huấn luyện với LightGCN.



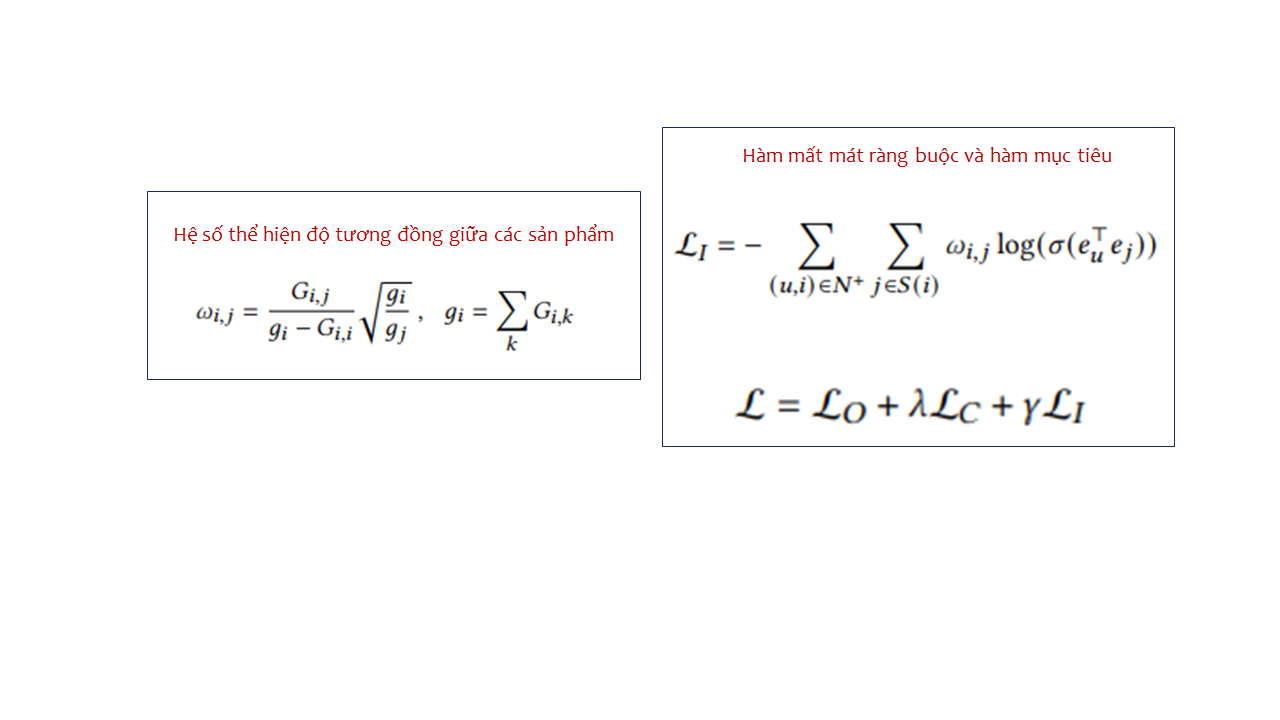
Minh họa huấn luyện LightGCN (trái) và UltraGCN (phải). LightGCN cần thực hiện lặp lại quá trình truyền thông báo 𝐿- lớp để có được phần nhúng cuối cùng cho quá trình đào tạo, trong khi UltraGCN có thể “bỏ qua” việc truyền thông báo đó để khiến các phần nhúng được đào tạo trực tiếp, phần lớn cải thiện hiệu quả huấn luyện và giúp tăng tốc độ triển khai thực tế.

UltraGCN bao gồm hai thành phần chính: học biểu diễn từ đồ thị người dùng-người dùng, và học biểu diễn trên đồ thị sản phẩm-sản phẩm.



Minh họa các thành phần chính trong đồ thị người dùng-người dùng của mô hình UltraGCN

UltraGCN không dựa vào việc truyền thông điệp rõ ràng để có thể tìm hiểu riêng các mối quan hệ khác theo cách linh hoạt hơn. Điều này cũng cho phép mô hình được điều chỉnh tầm quan trọng tương đối của các mối quan hệ khác nhau theo cách thủ công. Đặc biệt, ta cần nhấn mạnh rằng UltraGCN có thể mở rộng linh hoạt để mô hình hóa nhiều đồ thị quan hệ khác nhau, chẳng hạn như đồ thị người dùng-người dùng, đồ thị sản phẩm và thậm chí cả đồ thị tri thức. UltraGCN chủ yếu chứng minh việc sử dụng nó trên đồ thị xảy ra đồng thời giữa sản phẩm- sản phẩm, đồ thị này đã được chứng minh là hữu ích cho đề xuất. Trước tiên, chúng tôi xây dựng đồ thị sự xuất hiện của sản phẩm - sản phẩm bằng cách liên kết các mục có sự xuất hiện đồng thời, tạo ra ma trận liền kề có trọng số sau 𝐺 ∈ R|𝐼 |× |𝐼 | trong đó mỗi mục biểu thị sự xuất hiện đồng thời của hai sản phẩm.



Minh họa các thành phần chính trong đồ thị sản phẩm-sản phẩm của mô hình UltraGCN

### Kết quả thực nghiệm và phân tích

Nhóm nghiên cứu tiến hành xây dựng và thực nghiệm mô hình Đơn giản hóa cực độ mạng tích chập đồ thị UltraGCN trên bộ dữ liệu mẫu từ TV360 với khoảng 1.000.000 tương tác giữa khách hàng và các bộ phim. Mô hình được cài đặt thực nghiệm sử dụng các gói hỗ trợ sau:

* python 3.7.9
* pytorch 1.4.0
* numpy 1.19.2
* scipy 1.1.0
* tensorboard 2.4.0

Kết quả thực nghiệm được tổng hợp với 5 kịch bản khác nhau và thu kết quả cuối cùng đánh giá trên Recall@K và NDCG@K (đánh giá chính trên độ đo Recall@30). Các kết quả thực nghiệm qua các Round được mô tả và tổng hợp chi tiết dưới đây.

**Round** **1**:

Mô hình khởi tạo mặc định với áp dụng phân phối Gaussian với giá trị trung bình là 0 và độ lệch chuẩn 10−4 để khởi tạo các phần nhúng với số chiều là 128. Mô hình áp dụng chính quy hóa 𝐿2 với trọng số 10−4 và đặt tốc độ học thành 10−3, kích thước batch là 1024, tỷ lệ lấy mẫu âm 𝑅 là 200 và kích thước của bộ lân cận 𝐾 là 10.

**Round** **2**:

Trong round 2, mô hình được sửa kích thước nhúng thành 64, tương đương với các thực nghiệm dựa trên GCN gần đây và các mô hình GDE, IGCCF để giữ nguyên mức số lượng tham số để so sánh công bằng.

Kết quả thực nghiệm được tổng hợp ở bảng bên dưới:

Tổng hợp các Round thực hiện fine tune model UltraGCN

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall@30** | **Recall@50** | **NDCG@30** | **NDCG@50** |
| **Round 1** | 0.2763 | 0.1885 | 0.3493 | 0.5410 |
| **Round 2** | 0.2665 | 0.1844 | 0.3335 | 0.5305 |

Qua kết quả tổng hợp ở bảng 3.1, ta có thể thấy hiệu suất của mô hình thể hiện như sau:

* Với bộ siêu tham số khởi tạo cơ sở cho mô hình NGCF ở Round 1 mang lại kết quả tốt nhất xét theo toàn diện các độ đo. Tuy nhiên, với mỗi độ đo, các kết quả của Round 1 chỉ thể hiện cao hơn kết quả xếp thứ hai khoảng
* Với các siêu tham số đã được cố định, việc tăng thêm giá trị batch size hoặc tăng trọng số phân rã có thể giúp mô hình đạt được hiệu suất tốt hơn.
* Bên cạnh đó, việc tăng thêm một lớp tính toán cho đồ thị hoặc tăng thêm số chiều ẩn có thể khiến hiệu năng mô hình suy giảm.

## Mô hình đề xuất cơ sở VAE-CF

### Cơ sở lí thuyết

Lọc cộng tác là một trong những phương pháp được áp dụng rộng rãi nhất trong các hệ thống gợi ý. Lọc cộng tác dự đoán những mục nào người dùng sẽ thích bằng cách khám phá và khai thác các mẫu tương tự giữa người dùng và các sản phẩm. Đối với ứng dụng vào hệ thống khuyến nghị phim Viettel, chúng tôi mở rộng bộ mã hóa tự động biến thiên (VAES) [để lọc cộng tác cho phản hồi ngầm. VAEs khái quát hóa các mô hình hệ số tiềm ẩn tuyến tính và cho phép chúng ta khám phá các mô hình biến tiềm ẩn xác suất phi tuyến tính, được hỗ trợ bởi mạng thần kinh, trên các bộ dữ liệu đề xuất quy mô lớn. Chúng tôi đề xuất một mô hình tạo thần kinh với khả năng có điều kiện đa cực. Mặc dù được sử dụng rộng rãi trong mô hình hóa ngôn ngữ và kinh tế, khả năng đa thức dường như ít được nghiên cứu trong tài liệu lọc cộng tác, đặc biệt là trong bối cảnh các mô hình nhân tố tiềm ẩn. Các hệ thống gợi ý thường được đánh giá bằng cách sử dụng các thước đo dựa trên xếp hạng, chẳng hạn như độ chính xác trung bình trung bình và mức tăng tích lũy chiết khấu chuẩn hóa. Việc giảm thiểu tổn thất xếp hạng Top-N rất khó để tối ưu hóa trực tiếp. Ở đây, chúng tôi chỉ ra rằng khả năng đa thức rất phù hợp để mô hình hóa dữ liệu phản hồi tiềm ẩn và là đại diện gần hơn cho việc mất xếp hạng so với các hàm khả năng phổ biến hơn như Gaussian và logistic. Mặc dù đề xuất thường được coi là một vấn đề về dữ liệu lớn (do số lượng lớn người dùng và các mục thường có trong hệ thống đề xuất), chúng tôi lập luận rằng, ngược lại, nó đại diện cho một vấn đề “dữ liệu nhỏ” đầy thách thức: hầu hết người dùng chỉ tương tác với một tỷ lệ nhỏ các mặt hàng và mục tiêu của chúng tôi là cùng nhau đưa ra suy luận sáng suốt về sở thích của từng người dùng. Để tận dụng các tín hiệu thưa thớt từ người dùng và tránh trang bị quá mức, chúng tôi xây dựng một mô hình biến tiềm ẩn xác suất chia sẻ sức mạnh thống kê giữa người dùng và vật phẩm. Về mặt thực nghiệm, chúng tôi cho thấy rằng việc sử dụng phương pháp Bayesian có nguyên tắc sẽ mạnh mẽ hơn bất kể sự khan hiếm dữ liệu. Mặc dù VAEs đã được nghiên cứu rộng rãi để tạo mô hình và tạo hình ảnh, nhưng đáng ngạc nhiên là có rất ít công việc áp dụng VAEs vào hệ thống gợi ý. Chúng tôi nhận thấy rằng có hai điều chỉnh cần thiết để có được kết quả tiên tiến nhất với VAEs trong nhiệm vụ này:

* Nghiên cứu sử dụng khả năng đa thức để phân phối dữ liệu. Chúng tôi cho thấy rằng sự lựa chọn đơn giản này hiện thực hóa các mô hình hoạt động tốt hơn các khả năng logistic và Gaussian được sử dụng phổ biến hơn.
* Mô hình diễn giải lại và điều chỉnh mục tiêu VAE tiêu chuẩn mà chúng được quy định quá mức.
* Rút ra các kết nối giữa thuật toán học tập từ việc chính quy hóa được đề xuất của chúng tôi và nguyên tắc tắc nghẽn thông tin cũng như phân biệt entropy tối đa. Kết quả là một công thức tạo ra các giải pháp thiết thực cho vấn đề quan trọng này.

### Kết quả thực nghiệm và phân tích

Nhóm nghiên cứu tiến hành xây dựng và thực nghiệm mô hình cơ sở VAE-CF trên bộ dữ liệu mẫu từ TV360 với khoảng 1.000.000 tương tác giữa khách hàng và các bộ phim. Mô hình được cài đặt thực nghiệm sử dụng các gói hỗ trợ sau:

* Python 3.8
* matplotlib == 3.3.3

Kết quả thực nghiệm được tổng hợp với 2 kịch bản khác nhau và thu kết quả cuối cùng đánh giá trên Recall@K và NDCG@K (đánh giá chính trên độ đo Recall@30). Các kết quả thực nghiệm qua các Round được mô tả và tổng hợp chi tiết dưới đây.

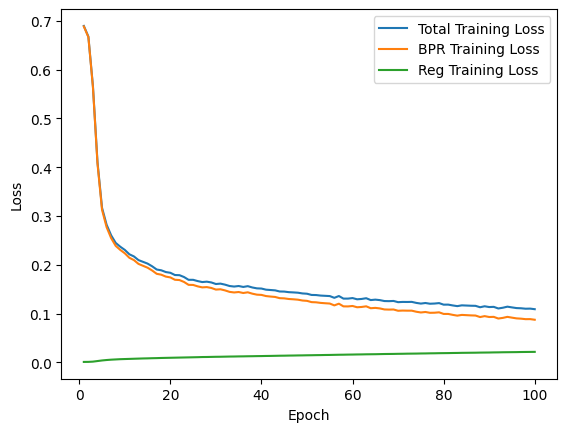
**Round** **1**:

* Khởi tạo ra các đặc tính phổ cần thiết cho tập dữ liệu với số lượng đặc điểm được làm mịn bằng cách điều chỉnh 'smooth\_ratio = 0.2'.

**Round** **2**:

Tinh chỉnh tốc độ học của mô hình đề xuất GDE.

**Nhận xét đánh giá**:



Biểu đồ minh họa sự thay đổi của hàm mất mát trong quá trình huấn luyện mô hình

Kết quả thực nghiệm được tổng hợp ở bảng bên dưới:

Tổng hợp các Round thực hiện fine tune model VAECF

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall@30** | **Recall@50** | **NDCG@30** | **NDCG@50** |
| **Round 1** | 0.2763 | 0.1885 | 0.3493 | 0.5410 |
| **Round 2** | 0.2665 | 0.1844 | 0.3335 | 0.5305 |

Qua kết quả tổng hợp ở bảng 3.2 và các hình minh họa quá trình huấn luyện mô hình, ta có thể thấy hiệu suất của mô hình GDE thể hiện như sau:

* Với bộ siêu tham số cho mô hình LightGCN ở Round 2 mang lại kết quả tốt nhất xét theo hầu hết các độ đo. Tuy nhiên, với độ đo Precision@20, kết quả của Round 2 chỉ thể hiện thấp hơn kết quả của Round 3 là .
* Kết quả của Round 2 và Round 3 cho thấy việc thay đổi batch size với mô hình LightGCN trong quá trình thực nghiệm không gây ra ảnh hưởng đáng kể về hiệu suất mô hình.
* Đối với LightGCN, kết quả thực nghiệm cho thấy việc tăng thêm số chiều ẩn có thể khiến hiệu năng mô hình cải thiện được khoảng .
* Bên cạnh đó, việc tăng thêm một lớp tính toán cho đồ thị cũng khiến hiệu suất mô hình suy giảm tương tự NGCF, có thể do nó gây ra vấn đề về việc dự đoán bị giảm tính cá nhân hóa khi dữ liệu mẫu ít.
* Đặc biệt, khi tăng trọng số phân ra, hiệu suất mô hình LightGCN bị giảm đáng kể, chỉ còn khoảng so với kết quả tốt nhất của mô hình.

Trên cùng tập dữ liệu và các độ đo đánh giá, hiệu suất của mô hình LightGCN đã thấy rằng nó vượt trội hơn mô hình NGCF với các kịch bản thực nghiệm tương tự như chúng ta đã thực hiện ở trên. Điều này đến từ một khía cạnh sâu rộng hơn của nền tảng học sâu, đây là một trong những kết quả mâu thuẫn đối với học máy trên đồ thị. Chúng ta có thể giả thuyết rằng việc biến đổi đặc trưng và kích hoạt phi tuyến thực hiện bởi mô hình NGCF làm tăng độ phức tạp trong quá trình huấn luyện mà không cải thiện đáng kể khả năng học của mô hình. Hầu hết các tham số học được trong cả hai mô hình đều nằm trong các nhúng biểu diễn ban đầu của người dùng và sản phẩm, không phải trong các ma trận trọng số. Điều này có nghĩa là LightGCN gần như có khả năng biểu diễn như NGCF nếu chúng ta chỉ tính các tham số mà không cần xem xét ma trận trọng số. LightGCN tập trung vào sức mạnh của việc sử dụng trực tiếp cấu trúc đồ thị để đạt được hiệu suất cao hơn.

Khuyến nghị theo thời gian thực được cá nhân hóa là một tính năng quan trọng được áp dụng rộng rãi đối với các nền tảng trực tuyến như NetFlix, VieOn hay TV360. Trong khi những tiến bộ gần đây trong mô hình học sâu đã mang lại kết quả đầy hứa hẹn trong các đề xuất theo phiên, các hệ thống được thiết lập như GRU4Rec dựa trên mạng nơ ron hồi quy, và SASRec dựa trên mạng học sâu tự chú ý thường gặp khó khăn trong việc duy trì độ chính xác và khả năng mở rộng khi xử lý với khối lượng sản phẩm lớn. Để giải quyết những hạn chế này, chúng tôi đề xuất mô hình khuyến nghị mở rộng quy mô áp dụng kiến trúc Transformer theo phiên tương tác của khách hàng (*Recommender with Scalable Transformer-based User’s Session – STUS*) bằng cách sử dụng các hàm mất mát và lấy mẫu âm được tối ưu hóa tích hợp các chức năng lấy mẫu âm tính top-K và hàm mất mát theo danh sách phim để nâng cao độ chính xác của khuyến nghị.

SAS-RS = SASRec và GRU-RS = GRU4Rec, STUS = TRON

**Định nghĩa bài toán:**

Hệ thống khuyến nghị theo phiên dự đoán sản phẩm tiếp theo mà khách hàng sẽ tương tác dựa trên các hoạt động trước đó của họ trong phiên. Phiên là một chuỗi các tương tác giữa người dùng và sản phẩm theo thời gian thực, được biểu thị dưới dạng trong đó là độ dài phiên. Các sản phẩm mà khách hàng đã tương tác trong một phiên được coi là mẫu dương, ký hiệu là Ngược lại, các sản phẩm mà người dùng chưa tương tác được gọi là mẫu âm, được biểu thị dưới dạng và là tổng số các mục có sẵn. Việc đào tạo một mô hình để thực hiện dự đoán sản phẩm tiếp theo trên thường không khả thi do kích thước lớn của trong kịch bản thực tế. Do đó, một cách tiếp cận phổ biến là huấn luyện mô hình để phân biệt giữa mẫu dương và mẫu âm, điều này có thể đạt được thông qua lấy mẫu âm. Một thách thức lớn trong việc lấy mẫu âm là tính hiệu quả. Việc lấy mẫu trực tiếp từ có thể tốn kém về mặt tính toán vì nó yêu cầu loại trừ các sản phẩm có trong . Vấn đề này trở nên nghiêm trọng khi tăng số lượng mẫu âm, dẫn đến thời gian đào tạo kéo dài, ảnh hưởng đến hiệu suất triển khai thực tế.

## Mô hình khuyến nghị mở rộng quy mô áp dụng kiến trúc Transformer theo phiên tương tác của khách hàng

### Xây dựng thuật toán và kiến trúc mô hình

Kiến trúc mô hình đề xuất trong chương này được dựa theo một kỹ thuật khuyến nghị áp dụng kiến trúc Transformer theo phiên có thể mở rộng bằng cách sử dụng lấy mẫu âm được tối ưu hóa. Được thúc đẩy bởi các hạn chế về khả năng mở rộng và hiệu suất của các mô hình phổ biến như SASRec và GRU4Rec, mô hình khuyến nghị mở rộng quy mô áp dụng kiến trúc Transformer theo phiên tương tác của khách hàng (nhóm nghiên cứu Trung tâm Phân tích Dữ liệu có thể đặt tên là *Recommender with Scalable Transformer-based User’s Session – STUS*) bằng cách sử dụng các hàm mất mát và lấy mẫu âm được tối ưu hóa tích hợp các chức năng lấy mẫu âm tính top-K và hàm mất mát theo danh sách phim để nâng cao độ chính xác của khuyến nghị.

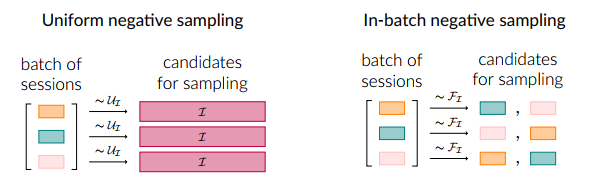
TRON cải tiến dựa trên SASRec, nâng cao khả năng mở rộng và hiệu suất cho các bộ dữ liệu quy mô lớn trên hệ thống TV360. Những cải tiến chính bao gồm:

1. Lấy mẫu âm tính: nhiều mẫu đồng nhất và mẫu trong lô

2. Chỉ lan truyền ngược trên các mẫu âm top-k

3. Hàm mất mát: áp dụng softmax được lấy mẫu

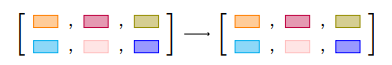
Mô hình sử dụng kết hợp lấy mẫu âm từ phân phối thống nhất và phân phối mật độ theo kinh nghiệm trên bộ sản phẩm . Hãy xem xét một lô bao gồm phiên người dùng. Tại mỗi bước thời gian trong mỗi phiên người dùng , chúng tôi lấy mẫu và Các mẫu này sau đó được ghép nối để tạo thành một vectơ ngẫu nhiên có số chiều với . Các mẫu âm cho toàn bộ lô được biểu diễn dưới dạng .



Mô tả quá trình lấy mẫu âm

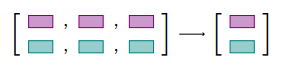
Quá trình lấy mẫu này có thể được thực hiện theo nhiều cách khác nhau, dẫn đến hình dạng âm cực khác nhau có tác động đáng kể đến việc truyền dữ liệu giữa CPU và GPU:

* Lấy mẫu âm theo từng phần tử:



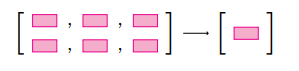
Mô tả quá trình lấy mẫu âm theo từng phần tử

* Lấy mẫu âm theo phiên



Mô tả quá trình lấy mẫu âm theo phiên

* Lấy mẫu âm theo lô



Mô tả quá trình lấy mẫu âm theo lô

Cơ chế chỉ lan truyền ngược trên top-K các mẫu âm: chúng tôi lấy mẫu một tập hợp các mẫu âm N(s,t) và đạt được điểm r(s,t,j) cho mỗi sản phẩm j của phiên s tại bước thời gian t trong N(s,t). Áp dụng hàm top-k cho các sản phẩm được tính điểm, chúng ta chọn các giá trị âm top-k KN(s,t) dựa theo điểm r(s,t,j). Những sản phẩm top-k này sau đó được sử dụng để cập nhật trong bước lan truyền ngược, trong khi những sản phẩm còn lại sẽ bị loại bỏ.

Hàm mất mát: Áp dụng Softmax được lấy mẫu.

### Thực nghiệm và phân tích

Nhóm nghiên cứu tiến hành xây dựng và thực nghiệm mô hình cơ sở EBPR cho khuyến nghị phim bộ trên bộ dữ liệu mẫu từ TV360 với khoảng 1.000.000 tương tác giữa khách hàng và các bộ phim. Mô hình được cài đặt thực nghiệm sử dụng các gói hỗ trợ sau:

* python 3.8
* matplotlib == 3.3.3

Kết quả thực nghiệm được tổng hợp với 5 kịch bản khác nhau và thu kết quả cuối cùng đánh giá trên Recall@K và NDCG@K (đánh giá chính trên độ đo Recall@30). Các kết quả thực nghiệm qua các Round được mô tả và tổng hợp chi tiết dưới đây.

Trong quá trình thực nghiệm, nhóm nghiên cứu có sử dụng hàm mất mát phổ biến như BCE (Entropy chéo nhị phân), BPR, hay SSM (sampled softmax).

**Round** **1**:

* Khởi tạo các tham số mô hình với số chiều nhúng là 64, kích thước batch là 1024, và đánh giá hiệu suất dựa theo Recall@30.
* Cài đặt 1 tầng tích chập cần thực hiện trong mô-đun học biểu diễn nhúng sản phẩm.

.

**Round** **2**:

Tinh chỉnh thêm tầng tích chập đồ thị.

Tổng hợp các Round thực hiện fine tune model EBPR

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Recall@30** | **Recall@50** | **NDCG@30** | **NDCG@50** |
| **Round 1** | 0.462 | 0.554 | 0.647 | 0.713 |
| **Round 2** | 0.367 | 0.446 | 0.527 | 0.582 |

Biểu đồ minh họa hiệu suất của mô hình EBPR trong quá trình các round trên bộ dữ liệu mẫu từ TV360

**Nhận xét đánh giá**:

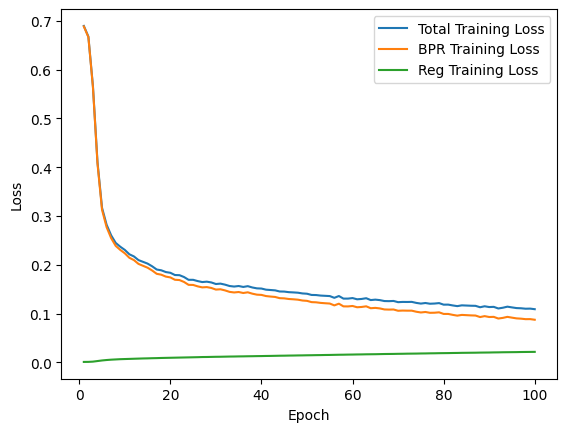
Qua kết quả tổng hợp ở bảng 3.3 và các hình minh họa quá trình huấn luyện mô hình, ta có thể thấy hiệu suất của mô hình EBPR thể hiện như sau:

* Việc xác định số tầng tích chập của đồ thị có ảnh hưởng rõ ràng và đáng kể nhất đến hiệu suất mô hình. Với cùng một bộ tham số, khi tăng lên hai tầng tích chập có thể khiến việc học biểu diễn sản phẩm bị kém đi trong điều kiện dữ liệu thưa, làm hiệu suất mô hình giảm hơn .
* Đánh giá trên các bộ dữ liệu thương mại điện tử quy mô lớn có liên quan cho thấy TRON cải thiện chất lượng đề xuất của các phương pháp hiện tại trong khi vẫn duy trì tốc độ đào tạo tương tự như SASRec. Thử nghiệm A/B trực tiếp mang lại tỷ lệ nhấp tăng 18,14% so với SASRec, nêu bật tiềm năng của TRON trong môi trường thực tế. Để nghiên cứu sâu hơn, chúng tôi cung cấp quyền truy cập vào mã nguồn của chúng tôi và tập dữ liệu ẩn danh.

## Tổng hợp và phân tích

Tổng hợp và đánh giá kết quả của các mô hình cơ sở:

**Nhận xét đánh giá**:



Biểu đồ minh họa sự thay đổi của hàm mất mát trong quá trình huấn luyện mô hình

Tổng hợp và đánh giá kết quả của các phương pháp ensemble đề xuất:

Biểu đồ minh họa hiệu suất của mô hình ensemble trong quá trình các round trên bộ dữ liệu mẫu từ TV360

Qua kết quả tổng hợp ở bảng 3.7, ta có thể thấy hiệu suất của mô hình thể hiện như sau:

* Việc xác định số layer của đồ thị lưỡng phân có ảnh hưởng đáng kể đến hiệu suất dự đoán phim cho người dùng. Với cùng một bộ tham số, khi giảm số lượng layer từ 2 xuống 1, hiệu suất gợi ý của mô hình trên các độ đo giảm đi khoảng từ
* Số lượng layer trong đồ thị đồng xuất hiện trong khoảng có ảnh hưởng không đáng kể đến hiệu suất gợi ý phim của mô hình.
* Bên cạnh đó, số lượng top-K người dùng tương tác chung nhiều nhất trong đồ thị đồng xuất hiện cũng là tham số quan trọng trong việc học biểu diễn người dùng qua các bước tính toán của đồ thị

Qua các kết quả tổng hợp và phân tích thực nghiệm các mô hình đề xuất ở phần 3, chúng ta có thể thấy hiệu suất của mô hình thể hiện như sau:

* Các kết quả đánh giá mô hình gợi ý phim trên các độ đo không chênh lệch nhau quá nhiều, dao động khoảng dưới
* Mô hình UltraGCN vẫn đạt hiệu quả cao nhất so với 3 mô hình đề xuất còn lại trên hầu hết các độ đo