HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Bùi Văn Minh

NGHIÊN CỨU, XÂY DỰNG HỆ THỐNG KHUYẾN NGHỊ PHIM TỰ ĐỘNG

Chuyên ngành : Khoa học máy tính

Mã số : 60.48.01.01

TÓM TẮT LUẬN VĂN THẠC SĨ

HÀ NỘI – 2017

Luận văn được hoàn thành tại:

HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG

Người hướng dẫn khoa học: **TS. Nguyễn Văn Thủy**

Phản biện 1: ……

Phản biện 2: ……

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ .... ngày .... tháng .... năm.....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông

MỤC LỤC

MỞ ĐẦU

Cùng với sự phát triển bùng nổ của internet, các kênh tìm hiểu thông tin, giải trí, thương mại điện tử… cũng phát triển nhanh chóng và mạnh mẽ. Giờ đây, gần nhưcó thể tìm kiếm được mọi thứ trên internet, từ tài liệu, sách, truyện, phim, video đến mặt hàng, sản phẩm…

Mỗi người, khi có một nhu cầu mua bán hoặc giải trí, sẽ có 2 cách để thực hiện. Thứ nhất, người đó có thể đến một địa điểm bán hàng hoặc vui chơi, nơi đó có những nhân viên có thể khuyến nghị về vấn đề của khách hàng hoặc khách hàng có thể thỏa thích xem qua những sản phẩm trên kệ hàng. Thứ 2, người có nhu cầu sử dụng internet để tìm kiếm.

Điểm yếu so với cách thứ nhất của cách thứ 2 – sử dụng internet chính là ở các trang web truyền thống thiếu đi một nhân viên tư vấn cho khách hàng truy cập vào trang web của mình. Có nhiều giải pháp được đưa ra như lập một kênh trò chuyện trực tuyến giữa nhân viên bán hàng và người dùng, gọi điện thoại tư vấn. Như vậy, với một trang web có lượng truy cập lớn, số nhân viên trực cũng phải cần rất nhiều. Điều này đòi hỏi chi phí cao. Vì vậy, hệ thống khuyến nghị tự động ra đời, giải quyết được các vấn đề đó.

Việc lựa chọn đề tài trên với các mục đích sau:

- Nghiên cứu tổng quan về hệ thống khuyến nghị và các phương pháp học máy.

- Nghiên cứu sâu về các thuật toán PMF (Probabilistic Matrix Factorization), BPMF (Bayesian Probabilistic Matrix Factorization), ALS (Alternating Least Squares).

- Ứng dụng và xây dựng hệ thống khuyến nghị phim tự động với tập dữ liệu MovieLens

Luận văn sẽ được trình bày trong 3 chương chính với nội dung như sau:

Chương 1. Tổng quan về hệ thống khuyến nghị

Chương 1 sẽ giới thiệu về hệ thống khuyến nghị và khái quát về phương pháp khuyến nghị. Sau đó sẽ giới thiệu về một số nghiên cứu có liên quan đến đề tài.

Chương 2. Phương pháp khuyến nghị

Chương 2 sẽ giới thiệu về tập dữ liệu thử nghiệm MovieLens và trình bày chi tiết về các thuật toán PMF (Probabilistic Matrix Factorization), BPMF (Bayesian Probabilistic Matrix Factorization), ALS (Alternating Least Squares).

Chương 3. Thử nghiệm, đánh giá và xây dựng hệ thống khuyến nghị phim tự động

Chương 3 sẽ trình bày về phương pháp thử nghiệm thuật toán trong hệ thống khuyến nghị, từ đó, áp dụng vào thử nghiệm, đánh giá thuật toán. Sau đó là quá trình xây dựng hệ thống khuyến nghị phim tự động.

CHƯƠNG 1: TỔNG QUAN VỀ HỆ THỐNG KHUYẾN NGHỊ

1.1. Giới thiệu về hệ thống khuyến nghị

Kể từ khi thương mại hình thành và phát triển, những câu hỏi lớn được đặt ra: "làm sao để bán được nhiều hàng hóa?", "làm sao để khách hàng hài lòng với sản phẩm?"… Vấn đề quyết định cho câu trả lời của những câu hỏi đó chính là quảng cáo và tư vấn sản phẩm đến người tiêu dùng, hay nói cách khác là khuyến nghị đến người tiêu dùng.

Hiện nay, mọi hệ thống có hiển thị quảng cáo trên internet đều sử dụng hệ thống khuyến nghị để đưa ra những quảng cáo, đề xuất tốt nhất có thể cho người dùng. Để làm được điều đó, hệ thống khuyến nghị sử dụng các thuật toán để phân tích, dự đoán dựa trên dữ liệu hành vi người dùng được lưu lại. Nhờ đó, quảng cáo mang tính cá nhân hóa tới người dùng. Hệ thống sẽ biết chính xác từng người sử dụng có nhu cầu gì, muốn mua gì, xem gì để đưa ra khuyến nghị.

1.2.Khái quát về phương pháp khuyến nghị

1.2.1. Khái quát bài toán

Trước khi luận văn trình bày khái quát về phương pháp khuyến nghị, cần làm rõ 2 thuật ngữ được sử dụng: người dùng (user) và sản phẩm (item). Thứ nhất, khái niệm người dùng ở đây là người sử dụng hệ thống để thực hiện các thao tác mua bán, giao dịch, xem, đánh giá… Thứ hai, khái niệm sản phẩm là mặt hàng được giao bán, video, bộ phim, bản nhạc, bài báo… Trong hầu hết các hệ thống khuyến nghị, dữ liệu được sẵn sàng dưới dạng đánh giá của người dùng về sản phẩm.

Bài toán được đặt ra như sau: Với các dữ liệu tập người dùng U, tập các sản phẩm I và tập dữ liệu D = {u, i, $r_{u,i}$}, trong đó, u ∈ U,

i ∈ I, $r_{u,i}$ là đánh giá của người dùng u cho sản phẩm i. Cần dự đoán đánh giá sản phẩm của một người dùng thứ n nào đó u_n cho sản phẩm i_m.

Để đánh giá độ chính xác của việc dự đoán, luận văn sử dụng độ đo Root Mean Squared Error (RMSE).

1.2.2. Phương pháp sử dụng dữ liệu khuyến nghị

Để giải quyết bài toán khuyến nghị, có nhiều phương pháp có thể giải quyết được việc đó. Mỗi phương pháp sử dụng dữ liệu người dùng - sản phẩm theo những cách khác nhau. Nhìn chung, có thể phân loại cách sử dụng dữ liệu người dùng - sản phẩm thành 2 nhóm: Sử dụng dữ liệu rõ ràng và sử dụng dữ liệu ẩn.

a.Dữ liệu rõ ràng

Những dữ liệu rõ ràng được sử dụng trong các thuật toán khuyến nghị như: Hồ sơ của sản phẩm và người sử dụng, dữ liệu đánh giá của toàn bộ hệ thống người dùng và sản phẩm.

b. Dữ liệu ẩn

Dựa trên dữ liệu đánh giá người dùng và những dữ liệu rõ ràng, có thể phân tích ra những dữ liệu ẩn, có ích cho việc dự đoán. Cụ thể, trong hệ khuyến nghị phim tự động, những dữ liệu ẩn: Movie biases, xu hướng người dùng, sở thích của người dùng

1.3. Một số ứng dụng thực tế của hệ thống khuyến nghị

1.3.1. Dịch vụ Netflix

Netflix là một kênh phim truyện có hơn 75 triệu người theo dõi ở hơn 190 quốc gia và mỗi người trong số họ đều có trải nghiệm khác nhau mỗi khi họ đăng nhập. Công ty có khoảng 1.000 người, trụ sở đặt tại Thung lũng Silicon, họ là những người chịu trách nhiệm về kiến trúc sản phẩm và thuật toán cá nhân hóa được đặt lại mỗi 24

giờ một lần để đảm bảo người dùng khám phá nội dung chính xác những gì họ muốn xem trong số 13.000 phim trong bất kỳ thời điểm nào.

Để làm được như vậy, Netflix đã sử dụng những thuật toán vừa có tính cá nhân hóa cao, vừa có độ đa dạng cho người dùng. Ngoài ra, với việc Netflix đầu tư rất nhiều tiền vào sản xuất những chương trình mới, không loại trừ khả năng có những tác động nhất định lên thuật toán để hướng người xem đến những chương trình mới đó.

1.3.2. Dịch vụ YouTube

Tương tự như Netflix, mục tiêu của hệ khuyến nghị trên YouTube là vừa có tính cá nhân hóa cao, vừa có độ đa dạng cho người dùng. Tuy nhiên, YouTube có những thách thức rất lớn trong vấn đề khuyến nghị cho hơn 1 tỉ người dùng và:

- Lượng video tải lên YouTube là vô cùng lớn
- Phần lớn video này có siêu dữ liệu kém như tiêu đề và mô tả không đầy đủ hoặc không liên quan.
- Số liệu sẵn có cho hệ khuyến nghị trên YouTube để đo lường sự quan tâm của người dùng là rất mơ hồ so với những số liệu sẵn có cho các hệ thống khuyến nghị khác như Amazon.

Trước khi tạo các video ứng viên đề xuất, hệ thống sẽ xác định một tập hợp các video có liên quan mà người dùng có thể xem sau khi xem một video hạt giống nhất định. Hệ thống sau đó kết hợp các quy tắc liên kết của video liên quan với hoạt động của người dùng trên trang web. Một khi điều này được thực hiện, nó lưu vết đường dẫn của các video liên quan tới hạt giống này thiết lập để tạo ra các video khuyến nghị ứng viên. Hãy suy nghĩ về hạt giống đặt ra như là trung tâm của một trang web và các ứng cử viên đề nghị tiềm năng như các điểm trên web mở rộng ra ngoài từ trung tâm. Điểm càng gần trung tâm của web thì càng có nhiều liên quan đến hạt giống và xa hơn thì ít liên quan hơn. Khi một loạt các khuyến nghị ứng viên

đã được tạo ra, chúng được xếp hạng theo các tín hiệu khác nhau, có thể được tổ chức thành ba nhóm: Chất lượng video, đặc trưng của người dùng, đa dạng hóa. Hệ thống khuyến nghị của YouTube đã làm tốt để cải thiện trải nghiệm của người dùng. Các video được đề xuất chiếm khoảng 60% số nhấp chuột trên trang chủ.

1.3.2. Website thương mại Amazon

Doanh thu của Amazon đã tăng lên rất lớn (khoảng 35% nhờ vào khuyến nghị) dựa trên việc tích hợp thành công các đề xuất qua trải nghiệm mua hàng - từ lúc khám phá sản phẩm cho đến khi rời khỏi website. Việc cho phép đề xuất được cá nhân hóa trong thương mại điện tử có lẽ là lý do số một cho các công cụ đề xuất để hạn chế những vấn đề như là vấn đề đuôi dài (long tail) - các sản phẩm hiếm hoi, mờ nhạt không phổ biến và không dẫn đến doanh thu.

Các thuật toán đề xuất của nhà bán lẻ khổng lồ dựa trên các yếu tố như: lịch sử mua hàng của người dùng, các mặt hàng trong giỏ hàng, mặt hàng họ đánh giá và thích, và những gì khách hàng khác đã xem và mua. Tuy nhiên, đối với một nhà bán lẻ với nhiều mặt hàng như Amazon, thách thức đó là những khuyến nghị nào sẽ xuất hiện và theo thứ tự nào – một vấn đề được biết đến như là "học cách xếp hạng" trong khoa học dữ liệu. Một vấn đề thứ cấp là sự đa dạng – làm thế nào để hiển thị những sự lựa chọn đa dạng của các sản phẩm trong đề xuất cho khác hàng. Amazon có thể đạt được mức độ quan tâm cao của khách hàng với thuật toán dựa trên quy trình là lọc cộng tác item – to – item.

CHƯƠNG 2: PHƯƠNG PHÁP KHUYẾN NGHỊ

Chương 2 sẽ giới thiệu cụ thể về phương pháp cơ bản của khuyến nghị là khuyến nghị dựa trên nội dung, lọc cộng tác và một số thuật toán khác.

2.1. Khuyến nghị dựa trên nội dung

Ý tưởng chính của phương pháp khuyến nghị này là: Khuyến nghị cho người dùng u những sản phẩm tương tự với những sản phẩm đã được người dùng u đánh giá cao từ trước đó.

2.1.1. Phương pháp dự đoán

a. Hồ sơ sản phẩm

Cần xây dựng hồ sơ cho mỗi sản phẩm i. Hồ sơ sản phẩm được xây dựng dưới dạng vector, là tập các đặc trưng của sản phẩm đó.

b. Hồ sơ người dùng

Có thể xây dựng được hồ sơ người dùng là vector đánh giá trung bình của người dùng u đó cho từng thành phần trong vector đặc trưng của sản phẩm.

c. Dự đoán

Để dự đoán độ "yêu thích" của người dùng u và sản phẩm i:

$$\boldsymbol{l(u,i) = \cos(u,i) = \frac{u.i}{\|u\|\|i\|}} \qquad 2.1)$$

2.1.2. Ưu điểm

- Không bị ảnh hưởng bởi vấn đề khởi đầu lạnh (cold start) hay vấn đề thưa thớt dữ liệu.
- Có thể khuyến nghị cho những người dùng có sở thích riêng.
- Có thể khuyến nghị những sản phẩm mới hoặc sản phẩm không phổ biến.
- Là phương pháp trực quan, dễ dàng giải thích được.

2.1.3. Nhược điểm

- Khó khăn trong việc tìm kiếm, xây dựng vector đặc trưng phù hợp của sản phẩm.
- Gặp vấn đề khi khuyến nghị cho người dùng mới: xây dựng hồ sơ người dùng.

- Không bao giờ khuyến nghị những sản phẩm nằm ngoài hồ sơ của người dùng.

- Không thể khai thác những thông tin đáng giá cho việc dự đoán từ người dùng khác.

2.2.Lọc cộng tác

Nói chung, lọc cộng tác là quá trình lọc thông tin hoặc mẫu sử dụng các kỹ thuật liên quan đến sự hợp tác giữa nhiều nguồn dữ liệu [14]. Có nhiều cách tiếp cận đểgiải quyết bài toán lọc cộng tác: Cách tiếp cận dựa trên bộ nhớ (memory-based); Cách tiếp cận dựa trên mô hình (model-based); Kết hợp nhiều cách tiếp cận: Kết hợp các thuật toán của cách tiếp cận dựa trên bộ nhớ và thuật toán của cách tiếp cận dựa trên mô hình với nhau để đưa ra kết quả tốt hơn.

2.2.1. Cách tiếp cận dựa trên bộ nhớ

Phương pháp lọc cộng tác với cách tiếp cận dựa trên bộ nhớ có đặc trưng cơ bản là thường sử dụng toàn bộ dữ liệu đã có để dự đoán đánh giá của một người dùng nào đó về sản phẩm mới. Cách tiếp cận dựa trên bộ nhớ thường được chia làm 2 loại: dựa trên người dùng và dựa trên sản phẩm.

a.Dựa trên người dùng

Phương pháp được tóm tắt với 2 bước như sau:

- Bước 1: Tìm kiếm những người dùng có đánh giá tương tự với người dùng cần được dự đoán.

- Bước 2: Sử dụng đánh giá từ những người dùng được tìm thấy ở bước 1 để tính toán dự đoán cho người cần được dự đoán.

b. Dựa trên sản phẩm.

Phương pháp được tóm tắt thành 2 bước như sau:

- Bước 1: Xây dựng một ma trận để xác định mối quan hệ giữa các cặp sản phẩm với nhau.

- Bước 2: Kiểm tra thị hiếu của người dùng cần dự đoán bằng cách kiểm tra ma trận và kết hợp dữ liệu của người dùng đó.

2.2.2. Cách tiếp cận dựa trên mô hình

Trong cách tiếp cận này, các mô hình được phát triển bằng cách sử dụng các khai phá dữ liệu khác nhau, các thuật toán học máy

để dự đoán đánh giá của người dùng về các mặt hàng chưa được đánh giá.

2.2.3. Một số hạn chế của lọc cộng tác

Lọc cộng tác gặp phải một số khó khăn như sau:

a. Vấn đề thưa thớt dữ liệu

b. Khả năng mở rộng.

c. Từ đồng nghĩa

d. Cừu xám (gray sheep)

2.3. PMF (Probabilistic Matrix Factorization)

Probabilistic Matrix Factorization (PMF) [11] là một mô hình tuyến tính xác suất với nhiễu quan sát Gaussian.

Giả sửcó N người dùng và M phim. Đặt R_{ij}là giá trị đánh giá của người sử dụng i cho phim j, $U_i$và V_jđại diện cho các vector đặc trưng người dùng và vector đặc trưng phim D chiều. Sự phân bố có điều kiện đối với các đánh giá quan sát được $R \in \mathbb{R}^{N\times M}$ và sự phân bố trước đối với $U \in \mathbb{R}^{D\times N}$ và $V \in \mathbb{R}^{D\times N}$ được cho bởi:

$$p(R|U,V,\propto) = \prod_{i=1}^{N}\prod_{j=1}^{M}\left[\aleph\left(R_{ij}|U_i^T V_j, \alpha^{-1}\right)\right]^{I_{ij}} \tag{2.2}$$

$$p(U|\propto_U) = \prod_{i=1}^{N}\aleph(U_i|0,\propto_U^{-1} I) \tag{2.3}$$

$$p(V|\propto_V) = \prod_{j=1}^{M}\aleph\left(V_j|0,\propto_V^{-1} I\right) \tag{2.4}$$

Trong đó $\aleph(x|\mu,\alpha^{-1})$ biểu thị sự phân bố Gaussian với trung bình μ và độ chính xác α, và I_{ij}là biến chỉ số, bằng 1 nếu người dùng i đánh giá phim j và bằng 0 nếu không có đánh giá.

2.4. BPMF

2.4.1. Mô hình

Sự phân bố tiên nghiệm với vector người dùng và phim được giả định là Gaussian:

$$P(U|\mu_U, \Lambda_U) = \prod_{i=1}^{N} \aleph(U_i|\mu_U, \Lambda_U^{-1}) \tag{2.5}$$

$$P(V|\mu_V, \Lambda_V) = \prod_{i=1}^{M} \aleph(V_i|\mu_V, \Lambda_V^{-1}) \tag{2.6}$$

Tiếp tục đặt các Gaussian-Wishart tiên nghiệm (Gaussian-Wishart priors) vào các siêu tham số (hyperparameter) người dùng và phim $\Theta_U = \{\mu_U, \Lambda_U\}$ và $\Theta_V = \{\mu_V, \Lambda_V\}$:

$$p(\Theta_U, \Theta_0) = p(\mu_U|\Lambda_U)p(\Lambda_U) = \aleph(\mu_U|\mu_0, (\beta_0\Lambda_U)^{-1})\mathcal{W}(\Lambda_U|W_0, v_0) \tag{2.7}$$

$$p(\Theta_V, \Theta_0) = p(\mu_V|\Lambda_V)p(\Lambda_V) = \aleph(\mu_V|\mu_0, (\beta_0\Lambda_V)^{-1})\mathcal{W}(\Lambda_V|W_0, v_0) \tag{2.8}$$

Ở đây $\mathcal{W}$ là sự phân bố của Wishart với v_0 độ tự do và một ma trậnW_0có kích thước D×D :

$$\mathcal{W}(\Lambda_U|W_0, v_0) = \frac{1}{C}|\Lambda|^{(v_0-D-1)/2} exp\left(-\frac{1}{2}Tr(W_0^{-1}\Lambda)\right) \tag{2.9}$$

Trong đó, C là hằng số chuẩn hóa, $\Theta_0 = \{\mu_0, v_0, W_0\}$. Trong các thử nghiệm, thiết lập v_0= D và W_0cho ma trận nhận diện cho cả siêu tham số người dùng và phim, μ_0= 0 theo đối xứng.

2.4.2. Dự đoán

Sự phân bố dự đoán của giá trị đánh giá R_{ij}^*cho người dùng i và phim j thu được bằng tách (marginalizing) các tham số mô hình và siêu tham số. Vì đánh giá chính xác sự phân bố dự đoán này là khó phân tích do sự phức tạp của hậu nghiệm, cần phải dùng đến suy luận gần đúng.Các phương pháp dựa trên MCMC, sử dụng phương trình xấp xỉ Monte Carlo cho sự phân bố dự đoán của phương trình trên được đưa ra bởi:

$$p\left(R_{ij}^*|R, \Theta_0\right) \approx \frac{1}{K}\sum_{k=1}^{K} p\left(R_{ij}^*|U_i^{(k)}, V_j^{(k)}\right) \tag{2.10}$$

Các mẫu $\{U_i^{(k)}, V_j^{(k)}\}$ được tạo ra bằng cách chạy một chuỗi Markov mà phân bố cố định (stationary distribution) là sự phân bố hậu nghiệm về các tham số mô hình và siêu tham số$\{U, V, \Theta_U, \Theta_V\}$.

2.4.3. Kết luận

Do việc sử dụng các tiên nghiệm liên hợp cho các tham số và siêu tham số trong mô hình BPMF, sự phân bố có điều kiện xuất phát từ sự phân bố hậu nghiệm là dễ dàng để lấy mẫu.Sự phân bố có điều kiện đối với các vector đặc trưng của phim và các siêu tham số của phim có dạng chính xác giống nhau.

2.5. ALS (Alternating Least Squares)

2.5.1. Phân loại ma trận cơ bản cho vấn đề khuyến nghị

Mô hình phân loại ma trận phổ biến nhất cho hệ khuyến nghị là đánh giá của người dùng u cho sản phẩm i:

$$\hat{r}_{ui} = x_u^T y_i \tag{2.11}$$

Với $x_u^T = (x_u^1, x_u^2, \dots, x_u^N)$ là một vector liên kết với người dùng, và vector liên kết với item $y_i^T = (y_i^1, y_i^2, \dots, y_i^N)$. Kích thước của vector là hạng của mô hình và các thành phần được gọi là nhân tử (factors).Thu thập dữ liệu đánh giá người dùng – sản phẩm vào ma trận $\hat{R} = (\hat{r}_{ui})$. Đầu tiên, thu thập các vector người dùng vào 1 ma trận X^T, và các vector sản phẩm vào ma trậnY^T.Sau đó,có thể biểu diễn mô hình trên:

$$\hat{R} = XY^T \tag{2.12}$$

Mô hình này giả sử có thể được xấp xỉ bởi hạng N nhân tử:

$$R \sim \tilde{R} = XY^T \tag{2.13}$$

2.5.2. Alternating Least Squares

Một phương pháp phổ biến để tìm kiếm ma trận X, Y được biết đến là *alternating least squares.* Ý tưởng là tìm kiếm các tham số x_u^j và y_i^j (các đầu vào của X và Y) mà tối thiểu hóa hàm chi phí:

$$C = \sum_{u,i \epsilon \text{ các đánh giá quan sát được}} (r_{ui} - x_u^T y_i)^2 + \lambda \left(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2 \right) \tag{2.14}$$

Hằng số λ được gọi là tham số chính quy. Nó có một hiệu quả thậm chí còn quan trọng hơn trong bối cảnh này, mặc dù:

Nếu giữ các vectơ sản phẩm Y cố định, C là một hàm bậc hai của các thành phần của X. Tương tự, nếugiữ các vector người dùng X cố định, thì C là một hàm bậc hai của các thành phần của Y.

Vì vậy, để tối thiểu hóa C:

1. Giữ vectơ người dùng cố định và giải phương trình bậc hai cho y_i^j. Đây có lẽ sẽ không phải là mức tối thiểu toàn cục của C vì chưa đụng tới các biến số x_u^j, nhưng ít nhất cũng giảm C.
2. Giữ vectors sản phẩm cố định và giải phương trình bậc hai cho x_u^j.
3. Lặp lại 1 và 2.

Các thuật toán trên được gọi là *alternating least squares.*

2.5.3. Xu hướng người dùng và xu hướng sản phẩm

Đối với mô hình đánh giá rõ ràng, một cách để tính xu hướng của người dùng là mô hình đánh giá người dùng thực tế như:

$$r_{ui} \sim \beta_u + \hat{r}_{ui} \tag{2.15}$$

Với β_u là xu hướng người dùng và $\hat{r}_{ui}$ là mô hình của bất cứ điều gì còn lại; ví dụ, $\hat{r}_{ui} = x_u^T y_i$ nếu sử dụng mô hình phân tích ma trận như đã đề cập trước đó. Có thể tính toán xu hướng của sản phẩm theo cách tương tự:

$$r_{ui} \sim \beta_u + \gamma_i + \hat{r}_{ui} \tag{2.16}$$

ở đây, γ_i là xu hướng của sản phẩm i.

- ALS với xu hướng (Biased ALS)

Xu hướng của người dùng và sản phẩm có thể được tích hợp trực tiếp vào thuật toán ALS. Mô hình hóa ma trận đánh giá người dùng – sản phẩm thành

$$\boldsymbol{r_{ui} \sim \beta_u + \gamma_i + x_u^T y_i} \tag{2.17}$$

Và tối thiểu hóa hàm chi phí:

$$C^{biased} = \sum_{u,i\epsilon \text{ các đánh giá quan sát được}} (r_{ui} - x_u^T y_i)^2 + \lambda\left(\sum_u (\|x_u\|^2 + \beta_u^2) + \sum_i (\|y_i\|^2 + \gamma_i^2)\right) \quad (2.18)$$

Một lần nữa, vì sự chuẩn hoá, có thể giữ các biến người dùng cố định và tối thiểu hóa các biến sản phẩm. Sau đó, có thể giữ các biến sản phẩm được cố định và tối thiểu hóa các biến người dùng.

Tuy nhiên, có thể viết lại hàm chi phí ở mỗi bước để nó giống như một mô hình không xu hướng, và sau đó chỉ sử dụng các công thức tương tự tìm thấy ở trên cho ALS không xu hướng. Bí quyết là xác định các vectơ mới bao gồm các xu hướng như các thành phần một cách đúng đắn.

Giả sử β là vector của xu hướng của người dùng (với n_users thành phần) và γ là vector của xu hướng của item (với n_itemsthành phần).

Đây là thuật toán ALS với xu hướng:

1. Khởi tạo các vectơ người dùng ngẫu nhiên và thiết lập tất cả các xu hướng thành số 0 (hoặc khởi tạo chúng ngẫu nhiên, nó không quá quan trọng).

2. Đối với mỗi sản phẩm i, xác định 3 vectơ mới:

$$r_i^{\beta} = r_i - \beta$$

Với các thành phần $r_{ui}^{\beta} := r_{ui} - \beta_u$ (chú ý rằng 2 vector có n_user thành phần),

$$\tilde{x}_u^T := (1, x_u^T)$$

Và

$$\tilde{y}_i^T := (y_i, y_i^T)$$

Sau đó $C^{biased} = \sum\left(r_{ui}^{\beta} - \tilde{x}_u^T \tilde{y}_i\right) + \lambda(\sum_u(\|\tilde{x}_u\|^2) + \sum_i(\|\tilde{y}_i\|^2))$. Do đó,thấy rằng xu hướng item và vector có thể được tính toán như sau:

$$\tilde{y}_i := \begin{pmatrix} \gamma_i \\ y_i \end{pmatrix} = \left(\tilde{X}^T \tilde{X} + \lambda I\right)^{-1} \tilde{X}^T r_i^{\beta}$$

(I là ma trận dạng $(N+1) \times (N+1)$, và $\tilde{X}$và $\tilde{Y}$ là các ma trận mà có các cột là các vector $\tilde{x}_u$và $\tilde{y}_i$ như thường lệ).

3. Bây giờ, với mỗi người dùng u, xác định ba vectơ mới:

$$r_u^{\gamma} := r_u - \gamma,$$
$$\tilde{y}_i^T := (1, y_i),$$

Và

$$\tilde{x}_u^T := (\beta_u, x_u)$$

Xu hướng người dùng và vector có thể được tính toán như sau:

$$\tilde{x}_u := \begin{pmatrix} \beta_u \\ x_u \end{pmatrix} = \left(\tilde{Y}^T \tilde{Y} + \lambda I\right)^{-1} \tilde{Y}^T r_u^{\gamma}$$

4. Lặp lại 2 và 3 cho đến khi hội tụ.

2.5.4. Kết luận

Như vậy, thuật toán đã sử dụng một trong những yếu tổ ẩn là xu hướng người dùng và sản phẩm. Với ALS, tác giả đã đưa ra một cách giải quyết bài toán khuyến nghị không quá phức tạp nhưng lại cho kết quả dự đoán với độ chính xác khá cao.

CHƯƠNG 3: THỰC NGHIỆM, ĐÁNH GIÁ VÀ XÂY DỰNG HỆ THỐNG KHUYẾN NGHỊ PHIM TỰ ĐỘNG

3.1. Thực nghiệm mô hình, thuật toán

Do phương pháp khuyến nghị dựa trên nội dung và lọc cộng tác là 2 phương pháp khuyến nghị cơ bản ban đầu, chúng có độ chính xác thấp hơn 3 phương pháp còn lại là PMF, BPMF và ALS rất nhiều nên luận văn chỉ tiến hành thực nghiệm trên 3 phương pháp là PMF, BPMF và ALS.

3.1.1. Giới thiệu tập dữ liệu thử nghiệm MovieLens

Hầu hết các bộ dữ liệu đều bao gồm các thành phần chính là các file: users, ratings, movies.

a. Users

File users chứa thông tin người dùng, thường có định dạng: ID người dùng::Giới tính::Tuổi::Nghề nghiệp::Mã bưu chính.

b. Ratings

File ratings chứa thông tin đánh giá của người dùng, thường có định dạng: ID người dùng::ID phim::Đánh giá::Mốc thời gian.

c. Movies

File movies chứa thông tin về phim, thường có định dạng: ID phim::Tiêu đề::Thể loại.

3.1.2. PMF

Thực nghiệm với mô hình PMF [11], [20]. Để so sánh, cần huấn luyện một loạt các mô hình PMF tuyến tính sử dụng MAP (maximum a posteriori probability - tối đa một xác suất hậu nghiệm), chọn các tham số định chuẩn của chúng bằng cách sử dụng bộ xác nhận. Ngoài các mô hình PMF tuyến tính, cũng cần huấn luyện các mô hình PMF logistic.

3.1.3. BPMF

Thực nghiệm với mô hình BPMF [11], [20]. Khởi tạo bộ lấy mẫu Gibbs bằng cách thiết lập các tham số mô hình U và V với các ước tính MAP của họ bằng cách huấn luyện một mô hình PMF tuyến tính. Thiết lập $\mu_0 = 0,\ v_0 = D$, và W_0cho ma trận đơn vị, cho cả tiên nghiệm người dùng và phim. Độ chính xác nhiễu quan sátα được đặt là 2. Phân bố dự đoán đã được tính bằng cách sử dụng công thức số 10, bằng cách chạy bộ lấy mẫu Gibbs với các mẫu $\left\{U_i^{(k)}, V_j^{(k)}\right\}$ thu thập sau mỗi bước Gibbs đầy đủ.

3.1.4. ALS

Thực nghiệm thuật toán ALS với các tham số:

- Vòng lặp: 10
- Tham số chính quy: 0.1
- Hạng của ma trận nhân tử lần lượt là rank = [6, 8, 10, 12, 14].

3.2. Kết quả thực nghiệm

3.2.1. PMF và BPMF

Bảng 3.1. Kết quả thực nghiệm RMSE của PMF và BPMF.

Số chiều D	PMF	BPMF
30	0.92142	0.90482
40	0.92016	0.89955
60	0.91685	0.89668
100	0.92056	0.89409
150	0.92201	0.89347

Có thể quan sát thấy, BPMF đã có sự cải thiện độ chính xác khá nhiều so với PMF.

3.2.2. ALS

Bảng 3.2. Kết quả thực nghiệm ALS.

rank	RMSE	Thời gian (giây)
6	0.87974	14.00735
8	0.87816	16.29091
10	0.87399	18.96353
12	0.87248	20.47909
14	0.87577	21.63209

Thời gian thực nghiệm của thuật toán được tính trên máy tính có:

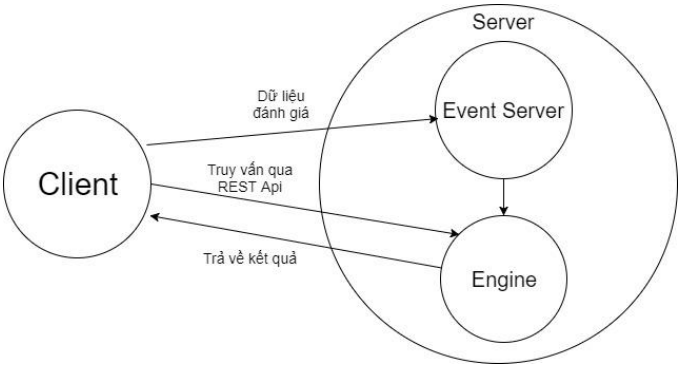
- Hệ điều hành: Linux Mint 64-bit.
- Vi xử lý: Intel Core i5-2520M 2.5GHz.
- RAM: 8GB.

3.3. Đánh giá kết quả

Dựa trên chỉ số RMSE, dễ dàng nhận thấy, thuật toán ALS cho độ chính xác là tốt hơn so với PMF và BPMF. Tuy kết quả tốt nhất của ALS là RMSE = 0.87248 lớn hơn so với giải pháp Bellkor 2008 (RMSE = 0.8643) nhưng giải pháp Bellkor là quá phức tạp để cài đặt. Vì vậy, luận văn xin đề xuất sử dụng thuật toán ALS với hạng của ma trận nhân tử là 8 (vì đảm bảo kết quả tốt và thời gian huấn luyện vừa phải).

3.4. Áp dụng xây dựng hệ thống

3.4.1.. Thiết kế hệ thống



Hình 3.1. Sơ đồ hệ thống khuyến nghị.

Trong đó:

- Client: bao gồm các trang web, ứng dụng gửi dữ liệu người dùng và phim về phía server để tính toán.

- Event Server: Nhận dữ liệu từ phía Client và lưu trữ dữ liệu.

- Engine: Xây dựng mô hình dự đoán với thuật toán ALS, sử dụng dữ liệu từ Event Server để huấn luyện. Sau đó được triển khai thành một web service dạng REST Api, nó lắng nghe các truy vấn từ Client và trả về kết quả dự đoán lập tức.

3.4.2.. Xây dựng hệ thống khuyến nghị phim.

a. Tập dữ liệu huấn luyện:

Hệ thống sử dụng bộ dữ liệu MovieLens 1 triệu đánh giá làm dữ liệu huấn luyện cho hệ thống (dữ liệu được mô tả trong phần 2.1).

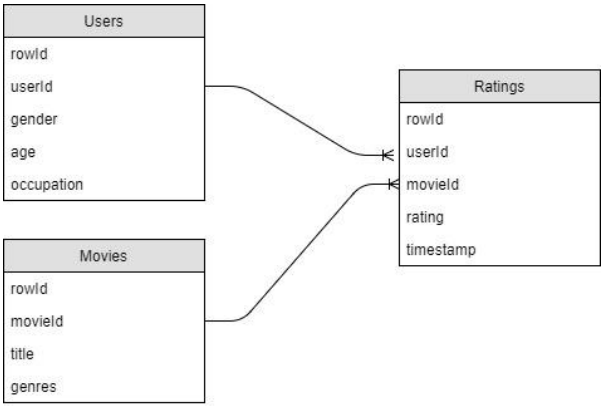
b. Event Server:

Ngôn ngữ lập trình: Python.

Phương thức giao tiếp với Client: REST Api

Cơ sở dữ liệu sử dụng: HBase.

Các bảng trong cơ sở dữ liệu:



Hình 3.5. Quan hệ giữa các bảng trong cơ sở dữ liệu.

c. Engine

Phương thức giao tiếp với Client: REST Api

Thuật toán sử dụng: ALS

KẾT LUẬN

1. Những đóng góp của luận văn

Luận văn đã trình bày tổng quan về bài toán khuyến nghị cũng như vai trò của bài toán trong xã hội hiện nay. Hai nghiên cứu nổi tiếng để giải quyết bài toán đó được giới thiệu là "Giải pháp BellKor 2008 cho giải Netflix" và "Học theo thời gian và mô hình tuần tự cho một hệ thống giới thiệu việc làm".

Tiếp theo, luận văn đã trình bày và thực nghiệm ba trong số rất nhiều thuật toán được sử dụng trong vấn đề khuyến nghị. Kết quả thu được sau khi thực nghiệm với bộ dữ liệu MovieLens với độ chính xác của dự đoán là khá tốt. Đặc biệt là thuật toán ALS. Dựa vào đó, luận văn đã xây dựng một hệ thống khuyến nghị phim. Hệ thống giao tiếp với Client thông qua REST Api, sử dụng cơ sở dữ liệu HBase, thuật toán huấn luyện ALS.

2. Hướng phát triển của luận văn

Trong tương lai, luận văn cần cải thiện tốc độ huấn luyện của thuật toán cũng như các giải pháp xung quanh việc xây dựng hệ thống để đảm bảo tốc độ đáp ứng của hệ thống theo thời gian thực và đảm bảo tính bảo mật của hệ thống.

TÀI LIỆU THAM KHẢO

Tài liệu Tiếng Anh

[1] D. Goldberg, D. Nichols, B. M. Oki, D. Terry, Using Collaborative Filtering to Weave an Information Tapestry, Comm. ACM 35.

[2] Graves, A. and Schmidhuber, J. (2009). Offline handwriting recognition with multi dimensional recurrent neural networks. In Advances in neural information processings ystems, pages 545–552.

[3] John S. Breese, David Heckerman, and Carl Kadie, Empirical Analysis of Predictive Algorithms for Collaborative Filtering, 1998 Archived 19 October 2013 at the Wayback Machine.

[4] Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 426–434. ACM.

[5] Koren, Y., Bell, R., Volinsky, C., et al. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8):30–37.

[6] Kuan Liu, Xing Shi, Anoop Kumar, Linhong Zhu, Prem Natarajan. "Temporal Learning and Sequence Modeling for a Job Recommender System".

[7] Luong, M.-T., Sutskever, I., Le, Q. V., Vinyals, O., and Zaremba, W. (2014). Addressing the rare word problem in neural machine translation. arXiv preprint arXiv:1410.8206.

[8] Recommender Systems - The Textbook | Charu C. Aggarwal | Springer

[9] Rendle, S., Freudenthaler, C., Gantner, Z., and Schmidt-Thieme, L. (2009). Bpr: Bayesian personalized ranking from implicit feedback. In Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence, pages 452–461. AUAI Press.

[10] Robert Bell; Yehuda Koren; Chris Volinsky (2008-12-10). "The BellKor 2008 Solution to the Netflix Prize"

[11] Ruslan Salakhutdinov, Andriy Mnih. ""Bayesian Probabilistic Matrix Factorization using Markov Chain Monte Carlo.

[12] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In

Proceedings of the 10th international conference on World Wide Web, pages 285–295. ACM.
[13] Sutskever,I.,Vinyals,O.,andLe,Q.V.(2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems, pages 3104–3112.
[14] Terveen, Loren; Hill, Will (2001). "Beyond Recommender Systems: Helping People Help Each Other" (PDF). Addison-Wesley. p. 6. Retrieved 16 January 2012.
[15] Usunier, N., Buffoni, D., and Gallinari, P. (2009). Ranking with ordered weighted pairwise classification. In Proceedings of the 26th annual international conference on machine learning, pages 1057–1064. ACM.
[16] Weimer, M., Karatzoglou, A., Le, Q. V., and Smola, A. (2007). Maximum margin matrix factorization for collaborative ranking. Advances in neural information processing systems, pages 1–8.
[17] Weston, J.Bengio, and Usunier, N.(2010). Large scale image annotation: learning to rank with joint word-image embeddings. Machine learning, 81(1):21–35.
[18] Xiaoyuan Su, Taghi M. Khoshgoftaar, A survey of collaborative filtering techniques, Advances in Artificial Intelligence archive, 2009.
[19] Y. Koren, "Factorization Meets the Neighborhood: a Multifaceted Collaborative Filtering Model", Proc. 14th ACM Int. Conference on Knowledge Discovery and Data Mining (KDD'08), ACM press, 2008
[20] J. Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, The YouTube Video Recommendation System.

Trang web

[21] https://www.cs.toronto.edu/~rsalakhu/BPMF.html
[22]http://www.businessinsider.com/how-the-netflix-recommendation-algorithm-works-2016-2