TÓM TẮT HIỆN TRẠNG HỆ THỐNG RECOMMENDER SYSTEM

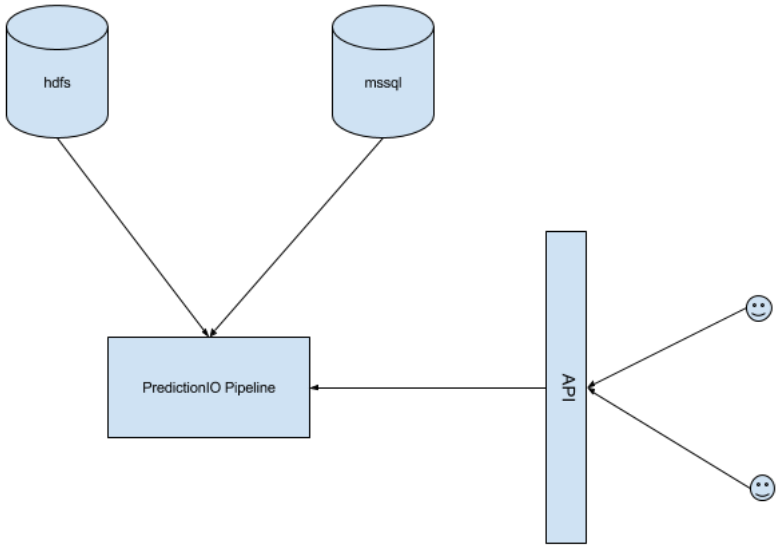
PBCQUOC

MỤC LỤC

DANH SÁCH HÌNH VẼ

DANH SÁCH BẢNG

* Department of Biology, University of Examples, London, United Kingdom
[1] Department of Chemistry, University of Examples, London, United Kingdom



Hình 1: Kiến trúc mô hình

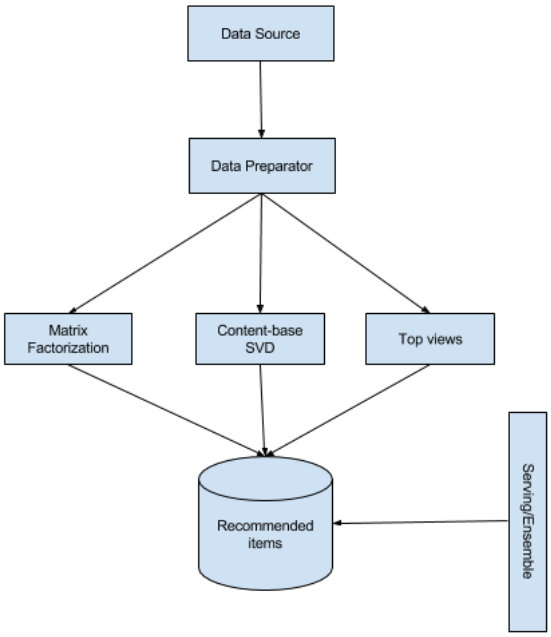
GIỚI THIỆU

Văn bản này tóm tắt cài đặt chi tiết của mô hình Collaborative Filtering(CF) và Content-Based SVD (CB).

KIẾN TRÚC HỆ THỐNG

Dữ liệu log được đưa vào định kì trên hệ thống từ hdfs 172.20.2.157 (xem job) và meta của moives được lấy từ 172.20.2.110 (xem job). Sau đó, dữ liệu được xử lý, tính toán theo mô hình chuẩn được định nghĩa trong PredictionIO. Luồng xử lý tính toán của PredictionIO được thực hiện như sau:

- DataSource đọc dữ liệu và trả về TrainingData
- Preparator nhận TrainingData, sau đó xử tiền xử lý và trả về PreparedData
- Các thuật toán lần lượt nhận cùng nhận vào PreparedData và thực hiện tính toán, do đó nếu sử dựng nhiều mô hình thì dữ liệu phải được đọc lên toàn bộ cùng một lần
- Các mô hình sau khi tính toán sau phải lưu trữ kết quả lại để phục vụ cho mỗi lần predict. Nên thực hiện tính toán sẵn kết quả để hạn chế tính toán quá nhiều trong lúc predict làm chậm hệ thống
- Sau khi predict, Serving sẽ nhận kết quả, và kết hợp các kết quả lại trong trường hợp có nhiều mô hình và trả về cho người dùng dưới dạng json



Hình 2: Kiến trúc mô hình 2

CÁC MÔ HÌNH

Collaborative Filtering

Phương pháp

Mô hình được lựa chọn cài đặt được công bố trong [1], và được hộ trợ bởi Spark. Mô hình được định nghĩa như sau:

$$\min_{x^*,y^*} \sum_{u,i} c_{ui}(p_{ui} - x_u^T y_i)^2 + \lambda(\sum_u \|x_u\|^2 + \sum_i \|y_i\|^2) \tag{1}$$

Trong đó:

- x_u là vector biểu diễn của user u
- y_i là vector biểu diễn của item i

Trong Spark, mô hình được cài đặt trong đối tượng ALS. Sau quá trình huấn luyện mô hình, chúng ta tính user preference như sau

$$\hat{p}_{ui} = y_i^T x_u \tag{2}$$

Để gợi ý cho người dùng, chúng ta có thể chọn những items có $\hat{p}$ cao nhất.

Content-Based SVD

Các đặc trưng

Đặc trưng	mã hóa
producers	one-hot encoding
directors	one-hot encoding
actors	one-hot encoding
genres	one-hot encoding
publishCountry	one-hot encoding
mpa	one-hot encodig
desc	tf-idf
year	giá trị liên tục
duration	giá trị liên tục

Đặc trưng cuối cùng được tổng hợp bằng cách nối các đặc trưng đã được liệt kê ở trên với trọng số bằng nhau thành một vector duy nhất

Phương Pháp

Chúng ta sử dụng Singular Value Decomposition(SVD) để giảm số chiều của ma trận đặc trưng, từ đó sử dụng cosine similarity để tính toán độ tương tự giữa 2 vectors.

Giả sử ta có ma trận A có kích thước mxn, phép phân rã ma trận SVD được định nghĩa như sau:

$$A = U\Sigma V^T \tag{3}$$

Với

- U là orthonormal matrix, với các cột là left singular vectors.
- Σ là ma trận đường chéo, các phần từ trên đường chéo là singular values.
- V là orthonormal matrix, với các cột là right singular vectors.

Để thực hiện giảm chiều ta chọn top k singular vectors có singular values có giá trị lớn nhất. Gọi A' là ma trận sau khi giảm chiều:

$$A' = U_k\Sigma_k \tag{4}$$

Các dòng của ma trận A' là các vectors sau khi đã được giảm chiều, chúng ta có thể sử dụng công thức đo khoảng cách bất kì để tìm ra những phim gần nhau nhất.

RANKING

Gọi u, p_i là vector biểu diễn của user và item sau quá trình huấn luyện. Để gợi ý cho người dùng, chúng ta cần xếp hạng mối quan hệ giữa user và item. Một số công thức có thể dụng để ranking

- Tính độ relevance giữa user với item, hay item vs item

$$rel(u, i) = \overrightarrow{u} * \overrightarrow{p}_i \tag{5}$$

- Tính độ tương tự như trên nhưng bổ sung thêm các tiêu chí như độ phổ biến của item, hay sự khác biệt của item đó với những item họ đã xem vì người dùng cũng có nhu cầu xem những phim họ khác thể loại.

$$score(i|u, S) = \min_{j \in S} pop(i)^{\beta} * rel(u, i) * div(i, j) \tag{6}$$

  Trong đó

  - S là tập tự định nghĩa theo một số tiêu chí như, nên khác với những item đã giới thiệu, ...
  - $p(i)$ là mức đó phổ biến, #plays, #clicks,
  - $div(i, j) = \|p_i - p_j\|^2$

Trong hệ thống hiện tại độ đo 5 được sử dụng, tuy nhiên, cần thay đổi sang độ đo 6

LOGIC HIỂN THỊ HIỆN TẠI

Nội dung gợi ý được hiển thị trên 3 ứng dụng Phim Truyện, Giải Trí, Thiếu Nhi. Các logic sau được áp dụng cho cả 3 ứng dụng

- Hiển thị 7 items đầu tiên trong danh sách gợi ý tại màn hình Home, và 28 items còn lại trong tab đầu tiên của các ứng dụng
- Phim bộ sẽ được hiển thị lại trong vòng 3 ngày kể từ lần xem gần nhất

RESULTS AND DISCUSSION

REFERENCES

[1] Yifan Hu, Yehuda Koren, and Chris Volinsky. Collaborative filtering for implicit feedback datasets. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08, pages 263–272, Washington, DC, USA, 2008. IEEE Computer Society.