Hệ gợi ý

Nguyễn Tuấn Đạt
Đặng Quang Trung

Ngày 8 tháng 1 năm 2017

Nội dung

Mô tả dữ liệu

Giới thiệu bộ dữ liệu

- Bộ dư liệu: ml-20m (size: 190MB).
- Download: `http://grouplens.org/datasets/movielens/`.
- Bộ dữ liệu mô tả đánh giá 1 - 5 sao của một dịch vụ giới thiệu phim từ MovieLens.
- Bộ dữ liệu chứa:
  - 20000263 rating và 465564 tag của 27278 bộ phim.
  - Dữ liệu đươc tạo bởi 138493 users.
- Các users được thu thập ngẫu nhiên. Tất cả users đã đánh giá ít nhất 20 bộ phim.
- Các file dữ liệu chứa trong 6 tập, genome-scores.csv, genome-tags.csv, links.csv, movies.csv, ratings.csv và tags.csv.
  - Các file đã sử dụng: movies.csv và ratings.csv.

Mô tả dữ liệu

- Cấu trúc file dữ liệu ratings:
  - Tất cả đánh giá của người dùng điều chứa trong file ratings.csv
  - Mỗi dòng của file sau dòng header có định dạng (userId,movieId,rating,timestamp).
  - Các rating thực hiện trên thang điểm 5 sao, với gia số (0.5 sao - 5.0 sao).
- Cấu trúc file dữ liệu Movies:
  - Thông tin của các movies chứa trong file moives.csv.
  - Mỗi dòng sau dòng header có định dạng (movieId,title,genres).
  - genres là danh sách thể loại được lựa chọn: Action, Adventure, Animation , Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western, (no genres listed)

Xử lý dữ liệu

- Sử dung Java để xử lý dữ liệu ban đầu(chuẩn hóa dữ liệu). Các file sử dụng gồm có: ratings.csv, movies.csv.
- File ratings.csv
  - Đưa ra file biểu diễn theo định dạng (userId,movieId,rating).
  - File biểu diễn mà trận rating của các user.
- File movies.csv
  - Đưa ra file dưới dạng ma trận
  - Hàng biểu diễn cho movie và Côt biểu diễn cho danh sách các đặc tính

$$(i,j) = \begin{cases} 1 & \text{nếu movie i có đặc tính j} \\ 0 & \text{nếu movie i ko đặc tính j} \end{cases}$$

Content-based

Ý tưởng: Giới thiệu các movie đến user x với các movie có đặc tính gần với các movie mà user x đã đánh giá cao trước đó.

- user x sẽ có một vector người dùng

	Movie					
	0	1	2	3	4	5
user x	3.5	4.0	5.0	0	2.0	4.5

Hình: vector người dùng

- Chọn ngưỡng đánh giá rating của người dùng α

$$(x, i) = \begin{cases} 1 & \text{nếu rating(x,i)} \geq \alpha \\ -1 & \text{nếu rating(x,i)} \leq \alpha \end{cases}$$

Content-based

- $\alpha = 4.0$ Ta có vector người dùng

Movie

		0	1	2	3	4	5
user	x	-1	1	1	-1	-1	1

Hình: vector người dùng sau khi đánh giá

- Tìm đặc tính của user ta thực hiện:
  - $V_{feature} = V_{user} * M_{movie}$.
    Trong đó: $V_{feature}$ vector đặc tính người dùng, V_{user} vector người dùng, M_{movie} ma trận movie và các đặc tính.

Content-based

- Chuẩn hóa lại $V_{feature}$

$$V_{feature}(i) = \begin{cases} 1 & \text{nếu } V_{feature}(i) > 0 \\ 0 & \text{nếu } V_{feature}(i) < 0 \end{cases}$$

- Sử dụng độ đo Cosin để tình khoảng cách giữa vector đặc tính của người dùng và movie
- Gợi ý những movie gần với vector đặc tính của người dùng.

Tài liệu tham khảo