<https://docs.google.com/document/d/1BmROktrRQcZwxRoDsyHO16c39rTATeLd7WoV5pz9W68/edit#>

Trang bìa

//Tên trường

//Tên tác giả: Lê Đức Anh | Trương Hữu Thao

//Tên đồ án:

Các độ đo giữa hai bộ phim và ứng dụng của chúng trong các hệ thống tìm kiếm, gợi ý, phân cụm, và phân loại

(Movie similarity functions and their applications in retrieval, recommendation, classification, clustering)

# Chương 1, giới thiệu

## //1.1. Tại sao vấn đề này quan trọng?

~~T: Trong xã hội hiện nay,~~ với lượng thông tin khổng lồ trên mạng Internet thì việc lựa chọn dữ liệu cần thiết trong đó là một vấn đề không mấy dễ dàng. Theo đó, xuất hiện nhu cầu đặt phải tìm được cách chọn lọc dữ liệu phù hợp với yêu cầu đề ra theo hướng đơn giản, nhanh chóng và chính xác.

Và với lượng thông tin người dùng được lấy từ các nguồn được người dùng chia sẻ như Facebook, Google+..., chúng ta hoàn toàn có thể gợi ý những dữ liệu phù hợp nhất với thói quen cũng như sở thích của người dùng đó. Ví dụ: Bạn thích phim khoa học viễn tưởng, phim trinh thám… và Facebook của bạn thường xuyên chia sẻ các thông tin về các bộ phim đó thì chúng ta có thể gợi ý cho bạn thêm một số bộ phim có nội dung cũng như thể loại phù hợp với bạn.

Để làm được điều đó, chúng ta cần phải có một phương thức so sánh dữ liệu người dùng (các bộ phim, diễn viên yêu thích) với lượng dữ liệu có sẵn, từ đó có thể đưa ra gợi ý tốt nhất cho người dùng.

---------------------------------------

@Andy: Sự phát triển và tiến bộ vượt bậc của nền công nghệ dẫn theo sự phát triển thông tin dữ liệu trên Internet về số lượng và nội dung ngày càng lớn dần. Nhu cầu sử dụng dữ liệu cũng theo đó mà tăng, bằng những thao tác đơn giản, chúng ta có thể nhận về hàng tỉ dữ liệu trong một thời gian tương đối. Google, Youtube hay Amazon là những ứng dụng điển hình phục vụ con người trong việc khai thác dữ liệu.

Tuy nhiên chính sự dễ dàng trong việc tìm kiếm dữ liệu khiến việc chắt lọc thông tin gặp không ít khó khăn trong việc tiếp thu những tri thức mới. Vì vậy, sự cải tiến về công nghệ để mang lại những hiệu quả tối ưu phục vụ con người - ví dụ, chúng ta có thể cải tiến kỹ thuật phân loại trong việc phân loại tài liệu, doanh thu của các bộ phim phục vụ cho việc báo cáo, thống kê trong nền công nghiệp điện ảnh hay các loại mặt hàng, sự yêu thích, sự theo dõi của con người về loại mặt hàng nào đó rồi phân loại thành các nhóm người để dễ dàng phục vụ họ, cũng là giải pháp tốt cho ngành kinh tế marketing.

---------------------------------------

*Khi lượng thông của mỗi trang cá nhân (google+, facebook, instagram) trở nên phong phú do xu hướng dùng điện thoại và máy tính, lượng thông tin đó trở thành hiệu quả để phục vụ cho chính cá nhân đó trong việc tìm kiếm và giới thiệu. Ví dụ, tôi thích 10 bộ phim hành động và 2 bộ phim hài, thì bộ phim tiếp theo mà tôi nên xem sẽ là gì?*

*Đây là vấn đề so sánh độ tương tự giữa một người (lịch sử xem phim) với một bộ phim mới hoặc có sẵn trong cơ sở dữ liệu. Vấn đề này đòi hỏi ta cần có một độ đo giữa người dùng (user) và một sản phẩm (item).*

## //1.2. Các giải pháp hiện có? Và vấn đề của chúng?

**+ IMDB recommendation: giới thiệu dựa trên người dùng (không dựa trên nội dung)**

**>> Chụp vài trường hợp để thấy chẳng liên quan**

**+ Youtube recommendation**

**+ RottenTomatoes recommendation**

+ Facebook recommendation

+ Amazon recommendation

## //1.3. Trong đồ án này, chúng tôi tập trung vào giải quyết cái gì? giải pháp của chúng tôi là gì? Và có gì khác biệt so với cái hiện có?

>> Mở rộng thông tin dựa trên wiki (youtube+wiki sẽ tốt hơn???): giả định rằng có nhiều thông tin sẽ tốt hơn, nhưng còn phải chờ

|  |  |
| --- | --- |
| H | >> >> tải hai nguồn về  >> >> phải ghép hai nguồn với nhau (matching title? Matching year)  >> >> recommend dựa trên nguồn duy nhất (unified) đã ghép  Foundation: youtube kém vì nó ít dữ liệu >> Nếu ta làm giàu dữ liệu, ta sẽ có kq tốt hơn |
| T |  |
| A |  |

>> Recommending dựa trên các cụm (hierarchical clustering): giả định rằng nếu mình cây khung với các phim gần nhất, mình có thể giới thiệu tốt hơn

|  |  |
| --- | --- |
| H |  |
| T |  |
| A |  |

>> Recommend dựa trên thông tin đã post của facebook (các phim đã thích).  
Nếu lấy được các film đã xem ở trên fb + các thông tin ở wiki -> có giới thiệu tốt hơn?

|  |  |
| --- | --- |
| H | Foundation: thông tin của fb nghèo, nhưng mọi người lại hay upload lên  Khi join hai nguồn (fb, wiki) ta có thêm thông tin recommend |
| T |  |
| A |  |

# Chương 2, tổng quan về so sánh

//So sánh hai tập hợp

//So sánh hai tập với số lượng

//So sánh hai tập với thông tin không gian (n-gram)

//Khoảng cách bin-bin

//Khoảng cách cross-bin

//Kết hợp hai khoảng cách

# Chương 3, tìm kiếm (retrieval) và gợi ý (recommendation)

# Chương 4, tìm kiếm dựa trên phân cụm (clustering) và phân loại (classification)

# Chương 4b, tìm phim dựa trên post (Thao) triển khai một hệ thống gợi ý phim dựa trên độ tương tự X

+ Ý tưởng

|  |  |
| --- | --- |
| Mình muốn xem Người Rừng trong tháng bảy này  Bởi Thao  Tháng 7 năm 2016 | Người rừng = tarzan  Tháng này = (7/2016)  + tarzan (2016)  + tarzan (1994)  + jungle book (2016)  + jungle book (1970) |
| Post >> q >> R  B1: từ post - q, giả sử làm được (NLP)  B2: từ q - R có làm được không? (IRR) |  |

+ Phương pháp làm?

q = { các từ khóa quan trọng của câu post } = parse(thaoPost);

S = { tất cả các phim trong kho của mình }

R = { tất cả các phim sẽ show ở dưới post của thao } = filter(S, q) = retrieval(S, q)

+ Thông thường phải có độ đo giữa sk trong S và q do đó cần score = simf(sk, q) và do đó chúng ta cần định nghĩa

# Chương 5, kết luận

# Tham khảo (references)

[irbook] http://nlp.stanford.edu/IR-book/

[elsml] http://statweb.stanford.edu/~tibs/ElemStatLearn/

# Phụ lục (appendix)

//Để cho Đức Anh show code