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?

#### //1.3.1. 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

--------------- Hieu

>> >> 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

------------- Thao

Việc đưa ra gợi ý cho người dùng luôn phải được đảm bảo độ chính xác, tin cậy. Và để đảm bảo điều đó, chúng ta cần phải có được một nguồn dữ liệu lớn, có độ tin cậy cao. Đây là một vấn đề mà các trang đa số lấy dữ liệu từ người dùng (như Youtube, Wiki, Facebook) còn đang mắc phải.

Ví dụ: Hiện tại, Youtube đang sử dụng những video từ người dùng upload lên để làm nguồn dữ liệu cho việc so sánh và gợi ý. Điều đó làm nguồn dữ liệu không được đảm bảo (nội dung rác, không chính xác...), đặc biệt Youtube chỉ có nguồn dữ liệu với các video và một số ít dữ liệu dạng văn bản (như tựa đề, mô tả…). Điều này dẫn đến việc đưa ra các gợi ý trong nhiều trường hợp còn chưa đáp ứng được nhu cầu của người dùng.

Do đó, trong đồ án này, tôi đưa ra phương pháp giải quyết vấn đề về độ tin cậy của nguồn dữ liệu bằng cách ghép một số nguồn với nhau. Cụ thể, tôi sẽ ghép nguồn từ Youtube và Wiki vì hai nguồn này có lượng dữ liệu giao nhau tương đối ít: Youtube tập trung chủ yếu vào video trong khi Wiki chủ yếu có nguồn dữ liệu văn bản. Từ đó, chúng ta sẽ có một nguồn dữ liệu khá phong phú, tăng độ tin cậy của nguồn dữ liệu cũng như kết quả của việc so sánh và gợi ý.

-------------------- Anh

Một trong những câu hỏi vui vẻ và có chút hài hước như thế này, “Bạn đã xem BigHero 6, vậy bạn đã từng xem BigHero 1,2,3… hay chưa?”, bạn cũng sẽ ngơ ngác với câu hỏi trên nếu chỉ xem bộ phim trên mà không có chút thông tin gì. Bạn đi tìm kiếm hi vọng rằng tìm được các phần còn lại của bộ phim.

Xin trả lời rằng, BigHero 6 là mang tên của 6 vị anh hùng giải cứu thế giới, không phải là phần 6 trong một chuỗi các phim nào đó như trong suy nghĩ của bạn. Vậy từ đâu mà ta biết được thông tin đó, liệu xem youtube có cho ta hiểu biết đầy đủ các thông tin trên? Wiki giúp bạn hiểu thêm về bộ phim ?

Chúng ta được thấy rằng, youtube là một trong những nguồn video hay clip tốt nhất mà chính người dùng đưa lên cho những người dùng khác có thể khai thác. Tuy nhiên, chính vì vậy mà không phải video hay clip nào đưa lên cũng đầy đủ những thông tin cần thiết (đặc biệt là thông tin dưới dạng văn bản), việc thiếu thông tin khiến hệ thống tìm kiếm hay gợi ý gặp không ít những vấn đề, độ chính xác cũng không phải là tốt nhất, đối với người dùng khó khăn ở đây khi vừa phải xem trên youtube, vừa phải tìm kiếm thêm thông tin từ google, wiki để có chi tiết.

Từ những yếu điểm mà người dùng gặp phải ở trên, chúng ta cần có những giải pháp cải thiện hệ thống, ghép giữa hai nguồn lại với nhau thành một nguồn duy nhất, làm giàu dữ liệu để tối ưu việc cung cấp thông tin cho người sử dụng.

#### //1.3.2. 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

----------- Thao

Với lượng dữ liệu khổng lồ, việc so sánh và đưa ra gợi ý thường mất rất nhiều thời gian và công sức. Vấn đề này tuy đã được xử lý nhưng còn chưa đủ tối ưu. Thường thì các trang chỉ có thể chia nhỏ thành các bộ dữ liệu dựa trên một số thông số (như thể loại, năm phát hành…) một cách thủ công.

Chính vì thế, trong đồ án này, tôi sẽ sử dụng phương pháp phân cụm tự động dựa trên thông số truyền vào (ví dụ phân loại theo thể loại, đạo diễn, quốc gia sản xuất…). Bằng phương pháp này, chúng ta có thể giảm thời gian so sánh, tìm kiếm và gợi ý.

Ví dụ: Bạn vừa mới xem X-men: Apocalypse và bạn đang muốn tìm những bộ phim có nội dung tương tự, với quốc gia sản xuất là Mỹ thì chúng ta có thể chia nhỏ nguồn dữ liệu của chúng ta thành các cụm nhỏ có nội dung tương tự X-men, và cụm nhỏ đó chúng ta lại chia nhỏ hơn dựa trên quốc gia sản xuất. Từ đó, chúng ta đã loại bỏ được rất nhiều dữ liệu không cần thiết cũng như tăng độ phù hợp với yêu cầu của người dùng.

--------------- Anh

Bạn vừa xem “Fast and furious 7”, một trong những bộ phim đình đám nhất trong năm 2016, bạn cảm thấy nghiền về mê mẩn những bộ phim hành động đầy kịch tính như vậy, bạn cảm thấy nó hợp với phong cách xem phim của bạn và muốn tìm những bộ phim tương tự để xem. Dựa vào vốn dữ liệu sẵn có từ wiki, chúng tôi sẽ phân loại các bộ phim theo đúng sở thích của bạn, giúp bạn dễ dàng tìm kiếm những phần còn lại của “Fast and furious” hay những bộ phim hành động mang tính chất tương tự.

#### //1.3.3. 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?

------------ Hieu

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

------------- Thao

Trong việc đưa ra gợi ý thì lượng dữ liệu của người dùng là một yếu tốt hết sức quan trọng, vì từ những dữ liệu đó chúng ta có thể cụ thể hoá được yêu cầu của người dùng. Và một trong những nguồn dữ liệu nhiều nhất đó là nguồn dữ liệu người dùng chia sẻ trên Facebook. Mặc dù với lượng dữ liệu khổng lồ (hơn 1.65 tỷ người dùng vào quý 1 năm 2016), Facebook hiện chưa sử dụng được nguồn dữ liệu đó để đưa ra gợi ý cho người dùng một cách hợp lý.

Vì vậy, trong đồ án này tôi đưa ra một phương pháp gợi ý những bộ phim mà người dùng có thể thích dựa trên lượng dữ liệu được chia sẻ. Bằng cách kết hợp giữa lượng dữ liệu từ Wiki và những yêu cầu của người dùng từ Facebook, chúng ta có thể đưa ra được gợi ý cho người dùng.

------------- Anh

Facebook là một trong những trung tâm dữ liệu người dúng lớn nhất (với hơn 2tỉ người đã từng tham gia và sử dụng facebook), mỗi lần post hay upload là một lần thu thập được thông tin từ của facebook, làm giàu dữ liệu người dùng. Tuy nhiên, với lượng thông tin lớn và mạnh như vậy nhưng facebook lại không chú trọng tới việc giới thiệu hay tìm kiếm những thông tin hữu ích tới người dùng. Ví dụ, khi người dùng chia sẻ một bộ film hay lên dòng thời gian của mình, không có quá nhiều gợi ý từ facebook tới bộ phim đó, người dùng không biết thêm thông tin gì khác, nhưng người bạn khác từ facebook cũng chỉ biết những thông tin được post lên, những lời đánh giá bình luận từ người post mà không có một cái nhìn nhận tổng quan nhất từ bộ phim.

Chúng tôi có thể làm tốt hơn nữa, với mỗi bộ phim mà người dùng chia sẻ, và thêm những dữ liệu được thu thập từ wiki, chúng tôi sẽ cung cấp tới người dùng khác những sự gợi ý tốt nhất.

Một ví dụ cho bạn thấy: Một người dùng chia sẻ về bộ phim “Fast and furious 7” với vài lời bình luận “Bộ film này thật tuyệt vời, những cảnh quay thật là ấn tượng”, và chúng tôi có thể cải tiến lời bình luận đó bằng những dữ liệu mà mình có về hình ảnh, tác giả, hay diễn viên… để bạn thấy rằng đó không chỉ là đánh giá người dùng từ một phía, hay bộ phim đó có thực sự đúng như những gì mà người xem bình luận.

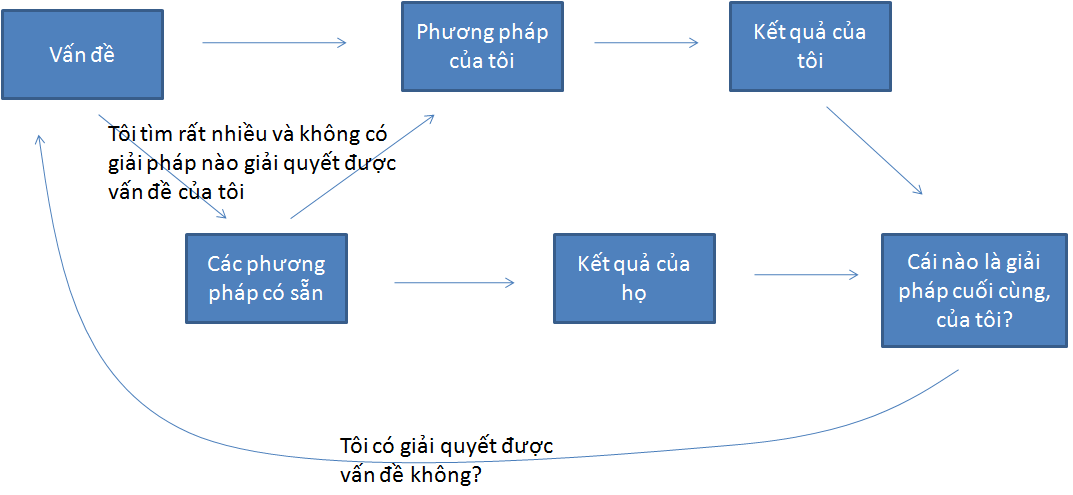
# Chương 2, tổng quan về so sánh (literature review)

//So sánh hai tập hợp

**//Nhìn lại các phương pháp trong quá khứ làm cơ sở cho phương pháp của mình**

**//Lời dẫn trước khi đi vào phương pháp của mình**

**//Trước khi nghĩ ra giải pháp thì đi tìm giải pháp. Nếu không có giải pháp nào giải quyết được vấn đề của mình, thì mới phải nghĩ ra / thiết kế giải pháp mới.**



//Chương 2 tập trung vào phương pháp / kết quả có sẵn.

|  |  |  |
| --- | --- | --- |
| Thao muốn thiết kế một movie recommendation system | -> Các hệ thống movie recommendation systems có sẵn (imdb, youtube, amazon, netflix, rotten tomatoes): vấn đề của họ là dựa trên người dùng, thiếu thông tin | -> Hệ thống mới mà chúng tôi đề xuất là join giữa youtube+wiki do đó đầy đủ thông tin, và chính xác hơn. |
| So sánh hai kết quả, và có giải quyết được vấn đề hay không???  Lựa chọn danh sách các giải pháp và có chọn giải pháp tốt nhất (không làm đồng nát) | Kết quả chụp hình của trang imdb  Kết quả chụp hình của youtube  + x-men (action, scifi)  + finding dory (family)  + tarzan (adventure)  + the conjuring (horror)  + mr bean (comedy) | Kết quả của hệ thống mới của chúng tôi (wetube)  + x-men (action, scifi)  + finding dory (family)  + tarzan (adventure)  + the conjuring (horror)  + mr bean (comedy) |

### Kỹ năng 1: Đọc thêm trước khi viết

[tkde]http://pages.stern.nyu.edu/~atuzhili/pdf/TKDE-Paper-as-Printed.pdf

users.csc.calpoly.edu/~dekhtyar/students/Lee-**survey**.pdf

www.mysmu.edu/faculty/fdzhu/paper/SocInfo'12\_57.pdf

[wiki-rec] https://en.wikipedia.org/wiki/Recommender\_system

**------ các phương pháp gợi ý**

**------------ phương pháp gợi ý dựa trên nội dung (content-based)**

**---------------------- so sánh bằng jaccard index**

------------ phương pháp gợi ý lọc cộng tác (collaborative filtering)

Lọc cộng tác: https://en.wikipedia.org/wiki/Collaborative\_filtering

### Kỹ năng 2: Viết thử ở dạng bảng (nói chung có ... trong đồ án này ...)

Chúng ta có thể nhìn tổng quan tất cả các phương pháp hiện có trong tài liệu [tkde] [wiki-rec] trong đó chủ yếu được chia thành ba nhánh lớn.

...

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Một số bài báo điển hình? | Một số hệ thống điển hình (ảnh chụp) |
|  |  | (academic) | (industrial) |
| Content-based | Nghĩa là gì?  Ví dụ? |  | Amazon  Youtube  facebook |
| Collaborative-ba | Nghĩa là gì? |  | imdb |
| hybrid | Nghĩa là gì?  Ví dụ? |  | netflix |

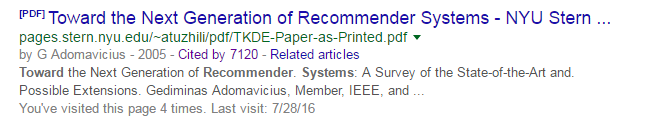
* Frank Meyer. (2012), [Recommender systems in industrial contexts. ArXiv e-prints](http://arxiv.org/abs/1203.4487) >> cho cột instrial

Trong đồ án này, chúng tôi tập trung vào gợi ý dựa trên nội dung (content-based) và chúng tôi so sánh hai phương pháp dùng mô hình (cluster-based recommendation) và sử dụng láng giềng gần nhất (NN-based recommendation).

https://www.google.com.vn/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=towards+recommendation+system

### Kỹ năng 3: Tìm bài báo mới hơn

Ấn vào link "Cited by " để tìm bài mới hơn >> chọn năm để focus



### ---------------- Đức Anh

Trước khi bắt đầu vào việc xử lý những vấn đề còn thiếu sót nêu ở mục trên, chúng tôi tìm hiểu tổng quan về so sánh, những kỹ thuật phổ biến trên thế giới hiện nay

So sánh là cách nhìn nhận vào hai hay nhiều đối tượng để từ đó thấy được điểm giống nhau hay khác nhau giữa chúng. Đưa ra các điểm tương đồng, hệ số tương đồng (Jarrcard Index) giữa các đối tượng.

Tại sao ta cần tìm hiểu về hệ số tương đồng? Hệ số tương đồng hay Jaccard Index [jacind] cho phép ta phân loại hay ghép các đối tượng vào thành một nhóm, được Paul Jaccard, giáo sư thực vật học và sinh lý thực vật học, phát triển và xuất bản vào năm 1901. Việc so sánh hai tập hợp là cách chúng ta đưa ra những điểm giống nhau, đánh dấu sự giống nhau đó bằng một đơn vị là Hệ số tương đồng.

Jaccard đã sử dụng tỷ lệ phần trăm giữa những phần tử chung trong tổng số các phần tử của hai đối tượng. Với A,B là 2 tập hợp, hệ số tương đồng (%) giữa các đối tượng được tính theo công thức:

|  |
| --- |
|  |
| Phương trình A1: độ đo jaccard giữa hai tập hợp |

Nếu A và B đều trống rỗng, chúng ta quy ước J(A,B) = 1, còn trong trường hợp tổng quát, ta có

|  |
| --- |
|  |
| Phương trình A2: khoảng cách |

Chúng tôi biểu diễn thuật toán về mặt lập trình như sau (ngôn ngữ Java) :

|  |
| --- |
| double jaccardIndex(Set<String> s1, Set<String> s2)  {  int c = 0;    for(String x: s1)  if(s2.contains(x)) c++;    int s = s1.size() + s2.size() - c;    return c/(double)(s==0 ? 1 : s);  } |
| Bảng A1: triển khai jaccard index trong java |

### ---------------- Thao

Mở đầu, chúng ta cần có một cách nhìn tổng quan về vấn đề so sánh cũng như một số phương pháp để thực hiện việc so sánh giữa hai hay nhiều đối tượng với nhau.

So sánh, được biết đến đơn giản là việc đánh giá độ tương đồng dựa trên một số thuộc tính của đối tượng, từ đó đưa ra được một giá trị nằm trong độ đo được quy định, nhằm mục đích phân loại, sắp xếp, gom nhóm… Ví dụ: so sánh 2 bộ phim, chúng ta sẽ đánh giá sự giống hay khác nhau của một số thuộc tính như tên phim, tên đạo diễn…

Hiện nay, có khá nhiều chuẩn so sánh được đề ra. Và trong đồ án này, tôi sẽ sử dụng phương pháp so sánh với độ đo Jaccard.

//So sánh hai tập với số lượng

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

### ---------------- Đức Anh

Từ việc so sánh hai chuỗi văn bản đơn giản ở trên, chúng ta có thể triển khai và mở rộng với việc so sánh hay hay nhiều tập thông tin không gian (n-gram). Hiện nay trên thế giới,

//Khoảng cách bin-bin

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

//Khoảng cách cross-bin

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

//Kết hợp hai khoảng cách

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

# 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/

[jacind] https://en.wikipedia.org/wiki/Jaccard\_index

# Phụ lục (appendix)

//Để cho Đức Anh show code