**TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH**

**KHOA ĐÀO TẠO CHẤT LƯỢNG CAO**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_



**HỌC MÁY**

Đề tài:

**ỨNG DỤNG THUẬT TOÁN MÁY HỌC ĐỂ PHÁT TRIỂN “RECOMMENDER SYSTEMS” TRONG WEBSITE THƯƠNG MẠI**

|  |  |  |
| --- | --- | --- |
| **Giảng viên hướng dẫn:** | TS. Vũ Quang Huy |  |
|  |  |  |
| **Thành viên:** | Nguyễn Thị Cẩm Thu | 15110137 |
|  | Nguyễn Thái Học | 15110052 |
|  | Lâm Thành Tài | 15110121 |

**Tp. Hồ Chí Minh, tháng 10 năm 2018**

**NHẬN XÉT CỦA GIẢNG VIÊN:**

**MỤC LỤC**

[I. TÌM HIỂU VỀ THƯƠNG MẠI ĐIỆN TỬ 4](#_Toc526885751)

[II. TÌM HIỂU THUẬT TOÁN CONTENT-BASED SYSTEMS 4](#_Toc526885752)

[2.1. Giới thiệu 4](#_Toc526885753)

[2.2. Khái niệm 5](#_Toc526885754)

[2.3. Bản chất của thuật toán 5](#_Toc526885755)

[2.4. Các thành phần đặc trưng trong thuật toán 6](#_Toc526885756)

[2.4.1. Utility Matrix 6](#_Toc526885757)

[2.4.2. Item Profiles 6](#_Toc526885758)

[2.4.3. Loss Function 6](#_Toc526885759)

[2.5. Phương thức hoạt động 8](#_Toc526885760)

[III. TÌM HIỂU THUẬT TOÁN COLLABORATIVE FILTERING (CF) 8](#_Toc526885761)

[3.1. Giới thiệu 8](#_Toc526885762)

[3.2. Khái niệm về Neighborhood-based collaborative filtering 8](#_Toc526885763)

[3.3. Thuật toán 9](#_Toc526885764)

[3.3.1. user-user collaborative filtering 9](#_Toc526885765)

[3.3.2. item-item collaborative filtering 9](#_Toc526885766)

[IV. PHỤ LỤC 10](#_Toc526885767)

[4.1. Tài liệu tham khảo 10](#_Toc526885768)

# TÌM HIỂU VỀ THƯƠNG MẠI ĐIỆN TỬ

Thương mại điện tử là hình thức bán hàng hóa và dịch vụ trên các hệ thống điện tử như Internet và các mạng máy tính. Một số khái niệm được định nghĩa bởi các tổ chức uy tín thế giới như sau:

* Theo Tổ chức Thương mại thế giới (WTO), “Thương mại điện tử bao gồm việc sản xuất, quảng cáo, bán hàng và phân phối sản phẩm được mua bán và thanh toán trên mạng Internet, nhưng được giao nhận một cách hữu hình, cả các sản phẩm giao nhận cũng như những thông tin số hóa thông qua mạng Internet”.
* Theo Ủy ban Thương mại điện tử của Tổ chức Hợp tác kinh tế châu Á – Thái Bình Dương (APEC) định nghĩa: “Thương mại điện tử liên quan đến các giao dịch thương mại trao đổi hàng hóa và dịch vụ giữa các nhóm (cá nhân) mang tính điện tử chủ yếu thông qua các hệ thống có nền tảng dựa trên Internet”. Các kỹ thuật thông tin liên lac có thể là email, EDI, Internet và Extranet có thể được dùng để hỗ trợ thương mại điện tử.
* Theo Ủy ban châu Âu: “Thương mại điện tử có teher được định nghĩa chung là sự mua bán, trao đổi hàng hóa hay dịch vụ giữa các doanh nghiệp, gia đình, cá nhân hay tổ chức tư nhân bằng các giao dịch điện tử thông qua mạng Internet hay các mạng máy tính trung gian (thông tin liên lạc trực tuyến). Thuật ngữ bao gồm việc đặt hàng và dịch thông qua mạng máy tính, nhưng thanh toán và quá trình vận chuyển hàng hay dịch vụ cuối cùng có thể thực hiện trực tuyến hoặc bằng phương pháp thủ công”.

Tóm lại, thương mại điện tử chỉ xảy ra trong môi trường kinh doanh mạng Internet và các phương tiện điện tử giữa các nhóm (cá nhân) với nhau thông qua các công cụ, kỹ thuật và công nghệ điện tử.

# TÌM HIỂU THUẬT TOÁN CONTENT-BASED SYSTEMS

## Giới thiệu

Dựa theo các hiện tượng thực tế ngày nay thì thuật toán trên được áp dụng cho rất nhiều vấn đền ngoài xã hội nói chung và ngành công nghệ thông tin nói riêng. Hầu hết các trang web ứng dụng mua sắm, xem phim, đọc sách báo hay các web về thương mại điện tử đều áp dụng thuật toán này nhằm mục đích gợi ý cho người dùng về các mặt hàng, các sản phẩm, các loại phim mà hệ thống tự đề ra dựa theo những hoạt động mà họ đã thực hiện trước đó dựa theo mỗi sở thích của các khách hàng khác nhau.

Các thuật toán trên được gọi chung là “Recommender Systems” – tạm dịch là “Hệ thống gợi ý”. Có hai thực thể chính trong “Recommender Systems” (RS) là Users và Items:

* Users: ở đây là người dùng hay còn gọi là khách hàng (lấy ví dụ điển hình như các tài khoản trên các trang web bán hàng, xem phim hay các kênh trên youtube).
* Items: ở đây là các sản phẩm (lấy ví dụ như là một bộ phim, một quyển sách, một bài hát hay một video clip)

Recommender Systems là thuật toán có nhiệm vụ dự đoán mức độ quan tâm của người dùng đối với một sản phẩm nào đó, thông qua đó có thể đưa ra các gợi ý phù hợp với từng người dùng.

Thông thường thì các cửa hàng truyền thống sẽ bị ảnh hưởng bởi “Long Tail” - ở đây tức là việc tạo ra doanh thu trên tổng số các sản phẩm hiện có trong cửa hàng đó, nhưng đôi khi một số nhỏ sản phẩm có thể tạo ra lượng doanh thu lớn còn số lượng lớn các sản phầm còn lại trong cửa hàng có thể sẽ chỉ tạo ra một lượng doanh thu nhỏ cho cửa hàng. “Long Tail” ở đây muốn nhắc đến đó là vấn đề về sản phẩm kém phổ biến hơn, bán chậm hơn.

Vì thế mà các cửa hàng online được ra đời và giải quyết được vấn đề đó, do có thể chứa được khối lượng lớn các thông tin nên việc trưng bày toàn bộ sản phẩm lên một trang web có thể là một điều không quá khó. Việc bán hàng trên các cửa hàng online có thể sẽ tiện lợi hơn, chi phí thấp hơn nhưng doanh thu sẽ cao hơn, mang lại cho người dùng sự tiện lợi và giải quyết được vấn đề nhu cầu cho mỗi người dùng.

## Khái niệm

Content-based Systems là thuật toán nhằm đánh giá các đặc tính của một items được gợi ý. Lấy ví dụ về một sản phẩm như quần áo, có thể có nhiều khách hàng sẽ thích áo tay dài vậy nên dựa vào đó mà chúng ta có thể gợi ý các sản phẩm khác trong cửa hàng của chúng ta mà cùng tính chất với sản phẩm mà khách hàng đang muốn hướng đến.

Thuật toán này đòi hỏi phải có sự sắp xếp các sản phẩm vào từng nhóm hoặc phải đi tìm đặc trưng của các loại sản phẩm khác nhau. Tuy nhiên thì một vài sản phẩm có thể sẽ không biết thuộc nhóm nào hoặc việc tìm ra đặc trưng của sản phẩm đó đôi khi là bất khả thi.

## Bản chất của thuật toán

Về bản chất thì chúng ta có thể coi thuật toán Content-based Systems như một người môi giới. Đơn giản vì thuật toán có thể dự đoán được sở thích của người dùng và tìm kiếm các tiềm năng có chung sở thích đó để có thể gợi ý cho người dùng.

**Ví dụ:**

* Bạn thích xem phim hành động, mà bạn thì hoàn toàn không thể nào biết hết được tất cả các phim hành động khác nhau, vì thế mà bạn sẽ cần một hệ thống gợi ý cho bạn việc đó, nó có thể giúp bạn lọc ra các bộ phim hành động để cuốn hút bạn dựa theo sở thích của bạn trước đó.

Nói cách khác, khi bạn mua một sản phẩm nào đó thì sẽ có hệ thống gợi ý cho bạn các sản phẩm tương ứng và tính chất tương tự với sản phẩm mà bạn đã mua.

## Các thành phần đặc trưng trong thuật toán

### Utility Matrix

Do có hai thực thể chính trong thuật toán là users và items nên mỗi users sẽ có một mức độ quan tâm đối với mỗi items khác nhau. Các items được users quan tâm có thể sẽ được đánh dấu lại và ứng với mỗi items thì sẽ có các mức độ đánh giá khác nhau bởi các users khác nhau. Ma trận Utility Matrix được tạo nên thông qua sự đánh giá đó.

Ma trận Utility Matrix là một ma trận có chứa tất cả các dữ liệu đã biết và các dữ liệu chưa dự đoán được. Thông thường thì lượng items được đánh giá có thể sẽ là một phần nhỏ trong tổng số các items và lượng items chưa được đánh giá có thể sẽ chiếm phần lớn hơn, do mỗi users có thể chỉ đánh giá một lượng nhỏ các items và thậm chí là có những user không đánh giá bất kỳ items nào. Vấn đề trên sẽ tạo cho ma trận Utility Matrix của chúng ta chứa một lượng lớn các dữ liệu chưa được dự đoán và các dữ liệu đã được dự đoán có thể chỉ chiếm một lượng nhỏ trong số đó.

Ma trận Utility Matrix có thể sẽ giúp thuật toán hoạt động tốt hơn nếu như số ô dữ liệu đã được dự đoán có được nhiều hơn. Vì thế mà các trang web bán hàng luôn tích cực gửi lời mời đánh giá sản phẩm cho bên phía họ để họ có thể nhận biết được chất lượng sản phẩm của cửa hàng bên mình và cũng như biết được sở thích của người dùng mà từ đó có thể đưa ra các chiến lược quảng bá sản phẩm tốt hơn, đúng đối tượng hơn.

Ma trận Utility Matrix được cho là điều không thể thiếu để có thể xây dựng gợi ý cho người dùng, có nhiều cách để có thể xây dựng ma trận trên đó là dựa trên sự đánh giá sản phẩm của một user và dựa trên hành vi của một user.

### Item Profiles

Do Content – based Systems là thuật toán dựa trên nội dung của mỗi items nên việc xây dựng một “profile” cho một item là điều cần thiết – feature vector.

Từ ma trận Utility Matrix thì chúng ta có thể đơn giản hóa bài toán bằng cách xây dựng một feature vector hai chiều.

Nếu Utility Matrix có cấu trúc là một feature vector hai chiều thì chúng ta sẽ làm bài toán đi theo hướng Regression hoặc Classification tùy theo biến đánh giá của mỗi users ứng với mỗi items và tùy theo ứng dụng mà chọn thuật toán.

### Loss Function

Đặt số user là N, số item là M, Utility Matrix là **Y**, thành phần ở cột thứ m và hàng thứ n của ma trận **Y** là mức độ quan tâm của một user thứ n đối với một item thứ m. Vì thế mà ma trận **Y** sẽ bị khuyết rất nhiều chỗ trống – những dữ liệu chưa được dự đoán hoặc chưa được đánh giá bởi người dùng. Gọi **R** là ma trận gồm 1 và 0 (rating or not rating), cụ thể hơn thì nếu r­ij­ = 1 tức là item thứ i rated bởi user thứ j, và rij = 0 trong trường hợp item thứ i not rated bởi user thứ j.

Mức độ quan tâm của một user đối với một item có thể tính bằng một hàm tuyến tính:

ymn = xm + bn

Trong đó:

* xm là một vector hàng.
* là một vector cột.
* bn là bias.
* ymn là thành phần ở hàng thứ m và cột thứ n của ma trận **Y**.

Xét 1 user thứ n bất kì và training set được cho là tập hợp các thành phần đã được điền của yn , ta có thể xây dựng hàm mất mát như sau:

Trong đó:

* là một regularization term với λ là một tham số dương.
* là số lượng các item mà user thứ n đã rated, là tổng các phần tử trên cột thứ n của ma trận **R**.

Do biểu thức hàm mất mát chỉ phụ thuộc vào các item được đánh giá nên ta có thể rút ngắn nó lại bằng cách:

* Đặt là vetor con của , hay có thể hiểu là các thành phần chưa được đánh giá ở cột thứ n trong ma trận **Y**.
* Đặt là ma trận con của ma trận đặc trưng X, là ma trận chứa các thành phần dữ liệu đã được đánh giá bởi user thứ n.

Ta sẽ có biểu thức viết gọn của hàm mất mát như sau:

Trong đó:

* là vector cột gồm các thành phần là 1.
* sẽ được xác đinh nếu như user thứ n đánh giá ít nhất một item.

## Phương thức hoạt động

Như đã tìm hiểu các thành phần cũng như bản chất của thuật toán trên nên việc xác định phương thức hoạt động là điều khả thi.

Thuật toán Content – based Systems thực chất là một thuật toán dùng để gợi ý các items cho các users thông qua nội dung, đặc tính và thông tin của item đó.

Thuật toán thiết lập nên một ma trận bao gồm các thành phần đã được đánh giá và chưa được đánh giá của các item bởi các user và áp dụng thuật toán để có thể hoàn thiện các thành phần bị khuyết trong ma trận đó (các dữ liệu chưa được đánh giá).

Cách tạo ra ma trận sẽ là bước khởi đầu cho việc gợi ý items cho user nhằm mục đích quảng bá đúng người và đúng sản phẩm họ cẩn thiết. Ngoài ra thì thuật toán có thể dựa vào số lần truy cập hoặc tìm hiểu thông tin về một item của một user để có thể dựa vào đó mà gợi ý cho user những item có tính chất tương ứng.

# TÌM HIỂU THUẬT TOÁN COLLABORATIVE FILTERING (CF)

## Giới thiệu

Collaborative Filtering là một phương pháp gợi ý item với ý tưởng chính dựa trên hành vi của các user tương tự khác lên cùng một item. Việc suy ra này được thực hiện dựa trên ma trận similarity để đo độ giống nhau giữa các user.

Để tính ma trận similarity, trước tiên ta cần chuẩn hóa dữ liệu. Phương pháp phổ biến nhất là mean offset, tức trừ các ratings đi giá trị trung bình mà một user đưa ra cho các item.

Một hướng tiếp cận khác là thay vì đi tìm các user gần giống với một user (user – user Collaborative Filtering), người ta đi tìm các item gần với một item cho trước (item – item Collaborative Filtering). Trên thực tế, item – item Collaborative Filtering thường cho kết quả chính sác hơn so với user – user Collaborative Filtering.

## Khái niệm về Neighborhood-based collaborative filtering

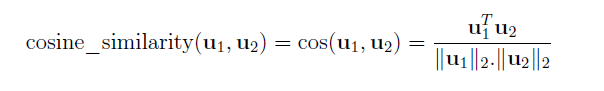
Neighborhood-based collaborative filtering là xác định mức độ quan tâm của một user tới một item dựa trên hành vi của các user khác gần giống với user này. Việc gần giống nhau giữa các user có thể được xác định thông qua mức độ quan tâm của các user này tới các item khác mà hệ thống đã biết.

Việc xác định mức độ quan tâm của mỗi user tới một item dựa trên mức độ quan tâm của user tương tự tới item đó còn được gọi là *user-user collaborative filtering*. Có một hướng tiếp cận khác được cho là làm việc hiệu quả hơn là *item-item collaborative filtering*. Trong hướng tiếp cận này, thay vì xác định sự giống nhau giữa các user, hệ thống sẽ xác định sự giống nhau giữa các item. Từ đó, hệ thống gợi ý những item gần giống với những item mà user đó có mức độ quan tâm cao.

## Thuật toán

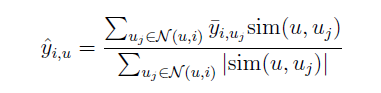
### user-user collaborative filtering

* **Hàm xác định độ giống nhau:**



Trong đó: U1,2 là các vector tương ứng với user 1 và user 2.

* **Điền các giá trị khuyết trong ma trận utility:**



Trong đó: là tập hợp user gần giống nhất, tức có­­­­­­­­­ similarity cao nhất của đã đánh giá .

### item-item collaborative filtering

Có một các tiếp cận khác, thay vì tìm sự giống nhau giữa các user, ta có thể tìm sự giống nhau giữa các item. Từ đó nếu một user thích một item thì hệ thống nên gợi ý các item tương tự với user đó.

Quy trình dự đoán các đánh giá bị khuyết cũng tương tự như trong user-user Collaborative Filtering, chỉ khác là bây giờ ta cần tính độ giống nhau giữa các hàng.

Về mặt tính toán, item-item Collaborative Filtering có thể nhận được từ user-user Collaborative Filtering bằng cách chuyển vị ma trận utility, và coi như item đang đánh giá ngược user. Sau khi tính ra kết quả cuối cùng ta lại chuyển vị một lần nữa để thu được kết quả.

# PHỤ LỤC

## Tài liệu tham khảo