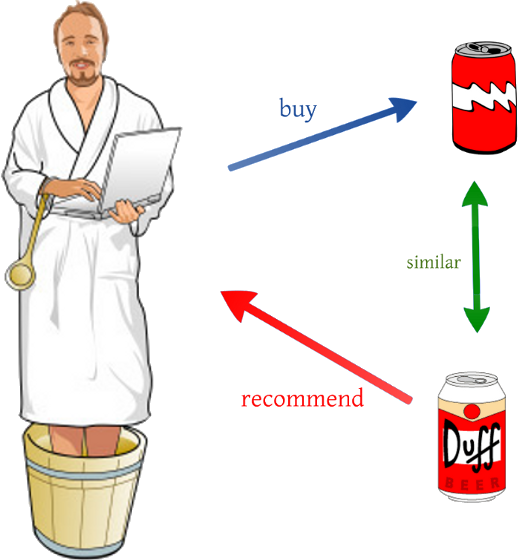
**Tìm hiểu về bài toán Recommender System:**

* Đầu tiên, chúng ta cần hiểu, Recommender System hay RS là gì? Recommender System được hiểu đơn giản là một hệ thống gợi ý, trong đó nó sẽ hỗ trợ chúng ta trong việc ra quyết định dựa trên những quyết định trước đó của bản thân chúng ta. Nó dự đoán dựa trên sở thích, thói quen,… để có thể gợi ý cho chúng ta. Lấy một ví dụ đơn giản hơn, giả sử bạn là một người độc thân, vui tính. Bạn cực kỳ hứng thú với những cô gái hoạt bát, xinh xắn, năng động. Nhưng, bạn không thể tìm hiểu được hết tất cả những cô gái như vậy bởi trên thế giới này có rất nhiều người con gái có cùng những tính chất như vậy. Vậy nên RS như một cứu cánh giúp cho bạn có thể được gợi ý những sự lựa chọn tốt nhất dựa trên những mong muốn của bạn. Đó chính là một ví dụ minh họa, còn trong thực tế, RS đã và đang được áp dụng với chính những ngành nghề của chúng ta, đó chính là thương mại điện tử.



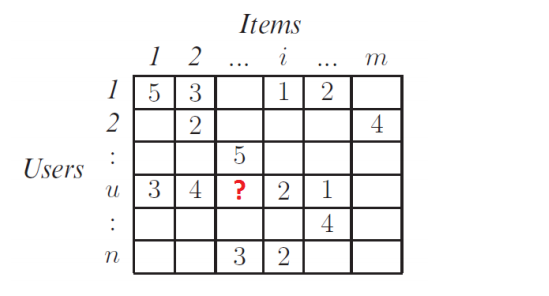
* Đây là một hình ảnh minh họa dễ hiểu về thương mại điện tử sự dụng RS. Khi bạn mua hoặc có hứng thú với một sản phẩm, hệ thống sẽ thu thập các thông tin về sản phẩm bạn đang có hứng thú hoặc đã mua, đó là số lần họ click vào sản phẩm, thời gian họ chốt đơn hàng, đánh giá về sản phẩm của họ nếu đã mua. Từ đó, các hệ thống thương mại điện tử sẽ dự đoán được người dùng và đưa cho họ những gợi ý phù hợp về sản phẩm khác.

**Các thành phần cơ bản của một RS bao gồm:**

Với mỗi hệ thống thông tin mới, chúng ta sẽ phải xây dựng và định hình được những thành phần cấu tạo nên chúng. Và không ngoại trừ RS, để xây dựng lên một hệ thống gợi ý thì chúng ta sẽ phải chọn tiếp cận theo phương pháp nào, và ở đây chúng ta sẽ chọn Machine Learning. Sẽ có ba đặc điểm mà chúng ta cần phải lưu ý như sau:

* Đầu tiên, thứ chúng ta cần quan tâm tất nhiên là người dùng (user). Đây chính là mục đích mà RS ra đời, bởi nếu không có người dùng, RS sẽ phục vụ ai bây giờ?
* Thứ hai, chúng ta sẽ quan tâm đến các “mục tin”(items). Mục tin là những thông tin đa dạng từ các sản phẩm trên các trang bán hàng, trang thương mại điện thử, thông tin tìm kiếm từ các hệ thống tìm kiếm, các bài hát trên các trang nghe nhạc trực tuyến hay chỉ là một bài viết trên các diễn đàn điện tử. Mục tin là một thứ chúng ta cần phải quan tâm tới bởi đây chính là phương thức giúp cho RS hỗ trợ được cho người dùng.
* Thứ ba chính là “phản hồi”(feedback) từ chính các user đối với mục tin của hệ thống. Nó có thể là điểm đánh giá, lời nhận xét, mức độ quan tâm,… Đây chính là một thông số giúp cho đánh giá sự hiệu quả của hệ thống đối với người dùng về mục tin.

Sau khi chúng ta đã có những thông tin của hệ thống sau khi thu thập, chúng ta sẽ biểu diễn các thông tin này dưới dạng mà máy tính có thể tính toán được. Thông thường, sự lựa chọn mà chúng ta hay sử dụng là biến thông tin đã thu thập trở thành dạng ma trận. Một ma trận ra đời thể hiện độ ưa thích của từng user lên các item tương ứng.



Đây chính là một ví dụ về ma trận thông tin phục vụ tính toán. Trong ma trận này, có những ô được điền những trọng số để thể hiện được mức độ ưa thích của mỗi user đối với các item. Những ô chưa điền trọng số là đang thể hiện user chưa hề tiếp cận tới các item đó. Đây chính là thể hiện vai trò của hệ thống gợi ý, đó chính là dựa trên những thông tin đã từng được biết, để có thể đưa ra gợi ý cho người dùng đó về những thông tin mà người dùng chưa được biết. Tóm gọn lại hơn chính là dự đoán các giá trị tại ô còn trống trong ma trận, sắp xếp thứ tự độ ưa thích và đưa ra sự gợi ý cho người dùng.

**Phân loại hệ thống gợi ý:**

Để dễ hiểu hơn, chúng ta sẽ đến với một ví dụ thực tế như sau. Khi lướt Shopee, chúng ta đang cần mua một chiếc áo Phông cổ tròn, thì hệ thống sẽ lập tức gợi ý cho chúng ta về những áo Phông tương tự như là áo phông cổ gấp, áo phông cổ lọ,.. Và khi chúng ta lướt Zing MP3, chúng ta có thể xem được những bài nhạc trẻ hot từ chính những đề xuất của hệ thống từ những thành viên có cùng sở thích nhạc trẻ hot.

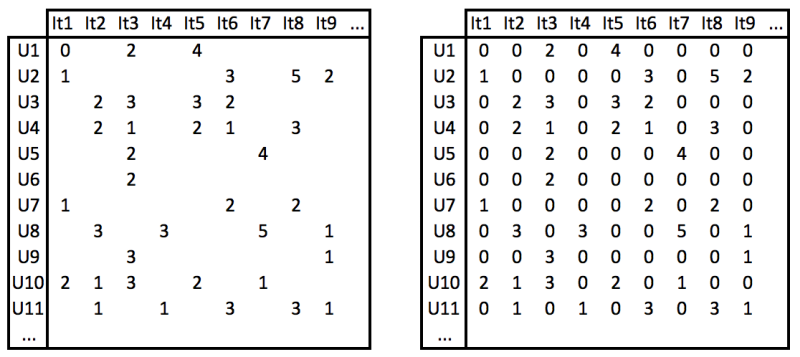
Từ hai ví dụ trên, chúng ta sẽ phân thành hai loại hệ thống gợi ý:

* Hệ thống gợi ý dựa trên nội dung - “Content based recommender systems” : Đây là một loại hệ thống sẽ quan tâm tới nội dung của item mà người dùng đang hướng tới, đánh giá đặc điểm, nội dung của item đó. Sau đó, gợi ý cho người dùng các item tương tự như item mà người dùng đang tìm kiếm. Đó chính là ví dụ đầu tiên của ví dụ thực tế.
* Hệ thống gợi ý dựa trên các user – và lọc cộng tác – “Collaborative filtering recommender systems” : Đây là một loại hệ thống sẽ phân tích các user có cùng item hiện tại. Sau đó sẽ đưa ra danh sách các item khác cũng được đánh giá tốt từ các user này, sau đó gợi ý tới các user đó. Đây là một phương pháp dựa trên thuyết tương đồng, khi mà đa phần những người dùng chung item sẽ có cùng sở thích để đưa ra gợi ý.

Với hai phương pháp trên, chúng ta nhận thấy phương pháp gợi ý dựa trên nội dung đòi hỏi chúng ta cần phải có rất nhiều thông tin từ các mục tin tương tự. Ở đây, chúng ta sẽ phải xác định xem một mục tin nào sẽ có tương thích với một mục tin hiện tại, điều đó chúng ta sẽ phải phân tích các dữ liệu đã thu thập, sau đó xử lý toàn một mục tin trong cơ sở dữ liệu hiện có. Tuy nhiên, với phương pháp lọc cộng tác, chúng ta không cần quá nhiều thông tin. Đơn giản hóa qua phương pháp này, chúng ta chỉ cần truy vấn tới item\_id của item hiện tại và các feedback item đó. Vậy nên, phương pháp lọc cộng tác sẽ được phổ biến và ưu tiên sử dụng trên các hệ thống gợi ý.

**Vậy các bước cần làm để xây dựng một hệ thống gợi ý là:**

* Việc đầu tiên, chúng ta cần phải xem xét xem, hệ thống của chúng ta sẽ cần gì? Đa phần với các hệ thống RS hiện nay, chúng ta sẽ có thể sử dụng điểm rating của user để có thể xử lý tới các bước tiếp theo. Tuy nhiên, tùy vào trường hợp, chúng ta sẽ có một vài chỉ số có thể thay thế điểm rating như là: số lần click vào item, thời gian xem xét và thao tác với item, số lần chia sẻ, lưu và thích item,…Sau khi thu thập chúng ta sẽ có dữ liệu từ các phiên làm việc khác nhau tương ứng với mỗi user với item đó. Sau khi xử lý xong dữ liệu, việc chúng ta sẽ là thu lại những cặp chỉ số user-item và đến ngay bước tiếp theo là chuẩn hóa dữ liệu.



* Ban đầu, ma trận dữ liệu của chúng ta sẽ là ma trận thưa, bởi vì số lượng dữ liệu trống còn rất nhiều. Điều này sẽ khó khăn cho chúng ta khi sử dụng ma trận này nếu không chuẩn hóa nó. Chúng ta chuẩn hóa được ma trận thì chúng ta mới áp dụng được các thuật toán Machine Learning lên nó. Sau khi chuẩn hóa dữ liệu, chúng ta sẽ xây dựng mô hình học máy để tính toán độ tương tự của các user.
* Qua những bước trên, bước tiếp theo là chúng ta sẽ chọn mô hình phù hợp. Tiếp theo đó chúng ta sẽ chạy mô hình và chọn ra top item sử dụng để gợi ý tới những người dùng.
* Cuối cùng, chúng ta sẽ đánh giá mô hình dựa trên tập dữ liệu kiểm tra.

**Tìm hiểu phương pháp giải quyết bài toán trên Content-Based Recommendation:**

**1. Item Profiles:**

- Trong các hệ thống content-based, tức là chúng ta sẽ dựa trên nội dung của mỗi item để phân loại thì việc đầu tiên chúng ta cần làm là xây dựng một bộ hồ sơ (profile) cho mỗi item. Profile này được biểu diễn dưới dạng toán học là một feature vector. Trong một số trường hợp đơn giản hơn thì feature vector trích xuất trực tiếp từ item. Với một ví dụ đơn giản, chúng ta sẽ xem xét các features của một món đồ bán trên trang thương mại điện tử shopee mà có thể sử dụng được trong hệ thống RS:

+ Nhà sản xuất: có thể đều là bánh mì, tuy nhiên có những khách hàng lại thích của NSX Kinh Đô, có những khách hàng lại thích của Komeco.

+ Chất liệu: Cùng là bánh mì bơ, tuy nhiên bánh mì bơ được làm bằng bột nếp thì được một số người ưa thích, một số người lại ưu tiên dùng bột mì.

+ Hình dáng: có những người cực kì thích bánh mì dài vì dễ xé, dễ chấm, nhưng có một số người lại thích bánh tròn vì dễ mang đi mang lại, và ăn vừa miệng, không quá nhiều.

- Đây chỉ là một số yếu tố cơ bản của một sản phẩm bánh mì được sử dụng đến trong RS. Tuy nhiên, một số yếu tố khó định nghĩa không nên đưa vào trong hệ thống, còn lại thì nên được sử dụng và xác định rõ ràng.

**2.Xây dựng hàm mất mát:**

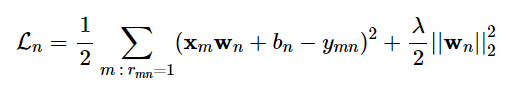
Với số users được gán là N, số items được gán là M, ma trận của chúng ta sẽ được mô tả bởi ma trận Y. Thành phần ở hàng thứ m, cột thứ n của Y là mức độ quan tâm(rating) của user thứ n lên sản phẩm thứ m của hệ thống đã ghi nhận. Ma trận Y khuyết rất nhiêu thành phần tương ứng với các giái trị mà hệ thống cần dự đoán. Thêm nữa, gọi R là ma trận rated or not thể hiện việc một user đã rated một item hay là chưa. Cụ thể, rij bằng 1 nếu item thứ I đã rate bởi user thứ j, bằng 0 nếu chưa rate.

- Mô hình tuyến tính: Giả sử rằng ta có thể tìm được một mô hình cho mỗi user, minh hoạ bởi vector cột hệ số wi và bias bn sao cho mức độ quan tâm của một user tới một item có thể tính được bằng một hàm tuyến tính:

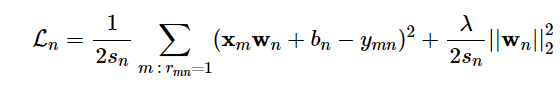
**Ymn = xmwn + bn (1)**

(\*Xm là một vector hàng, wn là một vector cột)

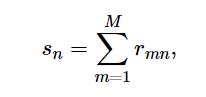
- Xét một user thứ n bất kỳ, nếu ta coi training set là tập hợp các thành phần đã được điền của yn, ta sẽ xây dựng được hàm mất mát như sau:



- Trong đó, thành phần thứ hai là regularization term và λ là một tham số dương. Chú ý rằng regularization thường không được áp dụng lên bias bn. Trong thực hành, trung bình cộng của lỗi thường được dùng, và mất mát Ln được viết lại thành:



- Trong đó sn là số lượng các items mà user thứ n đã rated. Nói cách khác:



là tổng các phần tử trên cột thứ n của ma trận rated or not R.

- Với những công thức trên, biểu thức hàm mất mát của mô hình cho user thứ n được viết gọn là:

