# CHƯƠNG 2: KHẢO SÁT MỘT VÀI PHƯƠNG PHÁP KHUYẾN NGHỊ ĐÃ CÓ

## 

Trong lĩnh vực xây dựng các hệ thống khuyến nghị trong quá khứ, người ta đã làm việc và nghiên cứu khá nhiều. Hầu hết công việc chủ yếu tập trung phát triền những phương pháp gợi ý những những đối tượng ưa thích đến cho người dùng. Ví dụ như những trang web gợi ý những bộ phim, gợi ý những quyển sách mà người dùng có thể yêu thích.

Những hệ thống khuyến nghị truyền thống ban đầu, người ta quan tâm đến 2 thực thể là người dùng và đối tượng cần được được khuyến nghị đến cho người dùng (ở đây chúng tôi xin phép gọi tắt là đối tượng). Quá trình khuyến nghị bắt đầu bằng một tập hợp các chỉ số đánh giá của người dùng đối với các đối tượng. Các chỉ số này được cung cấp một cách trực tiếp, tường minh từ người dùng hoặc được suy ra bởi hệ thống dựa vào một số thuật toán nào đó.

Ví dụ ta có một danh sách các chỉ số đánh giá như sau:

* Người A đánh giá phim “Người nhện” với chỉ số là 4 (trên tổng mức 5).
* Người A đánh giá phim “Siêu nhân” với chỉ số là 5 (trên tổng mức 5).
* Người A đánh giá phim “Dị nhân” với chỉ số là 3 (trên tổng mức 5).
* Người B đánh giá phim “Người nhện” với chỉ số là 4 (trên tổng mức 5).
* Người B đánh giá phim “Siêu nhân” với chỉ số là 4 (trên tổng mức 5).

Từ đó, người ta xây dựng một hàm R dùng để dự đoán các chỉ số đánh giá còn chưa biết. Ví dụ ta cần dự đoán người C sẽ đánh giá phim “Dị nhân” với chỉ số là bao nhiêu.

*R: Users* x *Items 🡪 Ratings*

Trong đó Users là các người dùng, Items là các đối tượng và Ratings là các chỉ số đánh giá. Sau khi hàm R dự đoán được các các chỉ số đánh giá trong toàn bộ không gian (Users x Items), hệ thống khuyến nghị sẽ có thể chọn ra k đối tượng i có chỉ số đánh giá dự đoán cao nhất và gợi ý chúng đến người dùng u.

Trong thực tế, người ta không nhất thiết phải dự đoán dựa trên toàn bộ không gian (Users x Items), vì chi phí cho việc tính toán trên một miền lớn như vậy rất tốn kém. Thay vào đó, người ta nghiên cứu để tìm ra những giải pháp hữu hiệu nhằm thu nhỏ không gian dự đoán để giảm thiểu chi phí tính toán. Nhìn chung, những hệ thống khuyến nghị thường được phân thành 3 loại: collaborative, content-based và hybrid. Chúng tôi sẽ lần lượt nói qua một cách vắn tắt về chúng ngay sau đây.

**2.1. Hệ thống khuyến nghị dựa trên nội dung (Content-based).**

Trong những phương pháp khuyến nghị dựa trên nội dung, chỉ số đánh giá dự đoán *R(u,i)* của người dùng *u* đối với đối tượng *i* thường được ước lượng dựa vào những chỉ số dự đoán *R(u,i’)* của người dùng *u* đó đối với những đối tượng tương tự với đối tượng *i*. Sự tương tự giữa hai đối tượng *i* và *i’* được tính toán tùy theo nội dung của chúng.

Ví dụ trong hệ thống khuyến nghị phim dựa trên nội dung, để gợi ý những bộ phim cho người dùng *u*, hệ thống cố gắng tìm hiểu những sở thích của người dùng bằng cách phân tích những điểm tương đồng về mặt nội dung của những bộ phim mà người dùng u đã từng đánh giá trong quá khứ. Khi đó, chỉ những bộ phim nào có độ tương tự cao, phù hợp với sở thích của người dùng mới được hệ thống gợi ý.

Nói một cách hình thức, ta gọi *Content(i)* là một tập hợp các thuộc tính nói lên đặc điểm của đối tượng *i*.Phần đặc tính của đối tượng *i* này thường được tính toán bằng cách trích ra một phần trong nội dung của nó và dùng phần trích xuất đó để xác định những yếu tố cần thiết phục vụ cho mục đích khuyến nghị. Có nhiều hệ thống khuyến nghị được thiết kế để gợi ý những đối tượng dựa vào văn bảng (*text-based*). Chúng được dùng để gợi ý trong những trang web, những bảng thông tin mà phần nội dung là những thông tin ở dạng văn bảng, được mô tả bằng những từ khóa (*keywords*). Để xác đinh mức độ quan trọng của những từ khóa dùng gợi ý, người ta cần xác định một độ đo. Độ đo đó là những trọng số được tính toán dựa vào những thông tin thu thập được. Ví dụ như độ đo *term-frequency/inverse document frequency (TF-IDF)*.

Hàm dự đoán chỉ số đánh giá *R(u,i)* trong những hệ thống khuyến nghị dựa trên nội dung thường được định nghĩa như sau:

*R(u,i) = score(ContentBasedProfile(u), Content(i))*

Trong đó, *ContentBasedProfile(u)* và *Content(i))* được định nghĩa như là những vector , mô tả những đặc tính của người dùng *u* và đối tượng *i*. Và chỉ số đánh giá được dự đoán cho người dùng *u* đối với đối tượng *i* được tính toán dựa trên độ tương tự giữa hai vector, ví dụng như độ tương tự Cosine.

Ngoài những phương pháp truyền thống hầu hết dựa trên việc trích xuất thông tin, những kỹ thuật khác cũng được sử dụng, như phân lớp Bayesian, hay những kỹ thuật máy học gồm có gom cụm, cây quyết định, mạng Neuron nhân tạo... Những kỹ thuật này khác so với kỹ thuật trích xuất thông tin ở chỗ chúng tính toán những chỉ số dự đoán không phải dựa trên heuristic formula, mà dựa trên những mô hình (model) được học từ tập dữ liệu bằng cách dùng những phương pháp học thống kê hoặc những kỹ thuật máy học.

Ví dụ dựa vào một tập những trang web đã được người dùng đánh giá và xếp thành hai loại là “phù hợp” và “không phù hợp”. Người ta dùng thuật toán phân lớp Bayesian để phân lớp cho những trang web chưa được đánh giá xếp loại.

Những hệ thống khuyến nghị dựa trên nội dung có những giới hạn nhất định. Đặc biệt là giới hạn về khả năng phân tích phần nội dung. Chúng chỉ hoạt động tốt trên những miền nội dung nơi mà thông tin có thể được trích xuất tự động (thông tin dạng văn bảng) hoặc được cung cấp một cách thủ công (thông tin của các bộ phim được mô tả cũng ở dạng văn bảng). Còn đối với những thông tin dạng đa phương tiện (hình ảnh, âm thanh …) thì thật sự rất khó khăn. Thông thường, những hệ thống khuyến nghị gợi ý những đối tượng tương tự với những đối tượng mà người dùng đã đánh giá trước đó. Tuy nhiên trong một số trường hợp đặc biệt, đối tượng không nên được gợi ý vì chúng có độ tương tự gần như tuyệt đối, nói cách khác là chúng quá tương tự với những thứ người dùng vừa mới xem. Ví dụ như nhiều mục tin tức khác nhau cùng nói về một sự kiện người dùng vừa xem qua ở mục tin tức này, khi đó người dùng sẽ không quan tâm đến những mục tin tức cùng sự kiện kia, hệ thống cũng không nên gợi ý. Thêm một bất cập nữa, là người dùng phải có đánh giá cho những đối tượng trước khi hệ thống có thể hiểu được sở thích và gợi ý cho họ những đối tượng khác. Như vậy, hệ thống sẽ gặp vấn đề đối với những người dùng mới, họ chưa cung cấp hoặc cung cấp rất ít những chỉ số dự đoán, hệ thống không đủ dữ liệu ban đầu của người dùng đó để có thể đưa ra những lời gợi ý chính xác dành cho họ.

**2.2. Hệ thống khuyến nghị Collaborative.**

Theo truyền thống, những hệ thống khuyến nghị Collaborative dự đoán những chỉ số đánh giá đối với một đối tượng cho một người dùng nào đó dựa trên những chỉ số đánh giá trước đây của những người dùng khác đối với đối tượng đang xem xét. Nói đơn giản, chỉ số *R(u,i)* dự đoán đánh giá của người dùng *u* đối với đối tượng *i* dựa trên những chỉ số *R(u’,i)* của những người dùng *u’* (có tính chất tương tự với người dùng *u*) đối với đối tượng *i*.

Những thuật toán khuyến nghi kiểu Collaborative có thể được phân thành hai nhóm là phương pháp đánh giá sử dụng bộ nhớ (memory-based hay còn gọi là heuristic-based) và phương pháp đánh giá dựa trên mô hình (model-based).

**2.2.1. Phương pháp đánh giá sử dụng bộ nhớ**

Phương pháp đánh giá dựa trên bộ nhớ dự đoán những chỉ số đánh giá dựa vào tập hợp tất cả những đối tượng đã được đánh giá bởi những người dùng trước đó. Giá trị cần dự đoán *r(u,i)* của người dùng *u* đối với đối tượng *i* thường được tính toán bằng một hàm kết tụ (aggregation) của những chỉ số đánh giá từ những người dùng khác đối với đối tượng *i* (ở đây người ta xét đến những người dùng tương tự với người dùng *u*, có thể là tất cả người dùng, hoặc chọn ra tập hợp n người dùng tương tự nhất):

Trong đó, *U* là tập hợp n người dùng tương tự với người dùng *u* (n có giá trị nhỏ nhất là 1, lớn nhất là toàn bộ tập hợp những người dùng) và họ đã có đánh giá trên đối tượng *i*. Một vài ví dụ về công thức tính r(u,i) như sau:

Trong đó:

k là tác nhân chuẩn hóa, thường được tính bằng công thức

là chỉ số đánh giá trung bình của người dùng u đối với các đối tượng.được định nghĩa bằng công thức với

sim(u,u’) là độ tương tự (similarity) giữa hai người dùng *u* và *u’*. Những người dùng có độ tương tự càng cao thì khả năng sở thích của họ giống nhau càng tăng, mức độ ảnh hưởng của những chỉ số đánh giá *r(u’,i)* trong công thức càng lớn. Điều này giúp chỉ số dự đoán r(u,i) cho người dùng *u* có độ chính xác cao hơn.Có nhiều cách để tính độ tương tự này, trong hầu hết các cách tính toán, người ta dựa vào những chỉ số đánh giá trên những đối tượng mà cả người dùng *u* cũng như *u’* đã từng đánh giá. Có hai công thức được dùng phổ biến đó là hệ số tương quan Pearson và hệ số tương quan Cosine.

*Hệ số tương quan Pearson:*

*Hệ số tương quan Cosine:*

Trong đó, r(x,s) và r(y,s) là những chỉ số đánh giá đối với đối tượng s lần lượt của người dùng *x* và *y*, , là các chỉ số đánh giá trung bình của người dùng *x* và *y* vừa được nhắc đến ở trên, là tập hợp các đối tượng cùng được đánh giá bởi người dùng *x* và *y*, là tích vô hướng của 2 vector và .

Người ta cũng đã tìm ra nhiều cách để cải tiến, mở rộng cách tính các hệ số tương quan. Hầu hết các hệ số tương quan này đều dùng để tính độ tương tự giữa những người dùng. Ngoài ra, những hệ số này cũng được dùng để tính độ tương tự giữa những đối tượng với nhau. Như vậy, trong phương pháp khuyến nghị dựa trên ký ức, có thể phân thành hai nhánh nhỏ là phương pháp dựa trên người dùng (user-based) và phương pháp dựa trên đối tượng (item-based). Cũng có những nghiên cứu cho rằng phương pháp dựa trên đối tượng trong một số trường hợp có thể cho hiệu năng cao hơn, chất lượng dự đoán cũng cao hơn [62] G. Karypis, “Evaluation of item-based top-N recommendation algorithms,” in Proceedings of the International Conference on Information and KnowledgeManagement (CIKM’01), pp. 247–254, Atlanta, Ga, USA, November 2001

**2.2.2. Phương pháp khuyến nghị dựa trên mô hình**

Ngoài ra, còn có những kỹ thuật khác như là mô hình hồi quy tuyến tính, các giải thuật gom cụm … kết hợp với các kỹ thuật rút trích, cắt giảm số chiều của ma trận cũng được nghiên cứu. Người ta cũng đã tiến hành so sánh và ghi nhận phương pháp khuyến nghị dựa trên mô hình trong một vài trường hợp cho hiệu năng cao hơn, độ chính xác cao hơn so với phương pháp khuyến nghị dựa trên ký ức.

[Note: Bổ sung về ưu khuyết điểm của CF trong paper Survey, mục 2 trang 2]

Cũng như hệ thống khuyến nghị dựa trên nội dung, hệ thống khuyến nghị collaborative cũng có những giới hạn, khiếm khuyết nhất định. Vẫn là vấn đề người dùng mới, hệ thống không đủ dữ liệu để có thể đưa ra những lời gợi ý cho người mới. Giờ đây thêm vấn đề đối tượng mới chưa được người dùng nào đánh giá, hệ thống cũng không đủ dữ liệu để tính toán cho ra những lời gợi ý (có thể thấy rõ khó khăn này ở bước tính độ tương tự giữa hai đối tượng trong phương pháp dựa trên đối tượng vừa được nói đến ở phần trên). Và một vấn đề quan trọng khác hệ thống khuyến nghị collaborative phải đối mặt là độ thưa thớt của dữ liệu đánh giá. Những đối tượng nổi bật được quan tâm sẽ có nhiều chỉ số đánh giá từ nhiều người dùng hơn. Trong khi đó, có một vài đối tượng hiếm khi được người dùng chú ý, dẫn đến không có dữ liệu đánh giá cho các đối tượng này. Ví dụ nhưng trong hệ thống khuyến nghị phim, những phim ăn khách sẽ được nhiều người quan tâm đánh giá, cũng có nhiều phim ít được quan tâm, chỉ nhận được rất ít đánh giá từ người dùng. Điều này dẫn đến khi khuyến nghị, những phim ít được quan tâm cũng sẽ hiếm khi được gợi ý, mặc dù chúng có thể rất phù hợp với sở thích của một số ít đối tượng người dùng nào đó.

**Hệ thống khuyến nghị Hybrid.**

Để hạn chế những khiếm khuyết trên, người ta đã nghiên cứu tìm cách kết hợp phương pháp dựa trên nội dung với phương pháp collaborative với nhau, dẫn đến sự ra đời của hệ thống khuyến nghị lai (hybrid). Ví dụ như người ta lập trình cho hệ thống học và lưu giữ những thông tin cá nhân người dùng (user profiles) bằng những phương pháp rút trích, phân tích nội dung (content-based), sau đó tiến hành so sánh những thông tin cá nhân đó để xác định được những người dùng có tính chất, sở thích tương tự nhau để đưa ra gợi ý bằng phương pháp collaborative. Ví dụ như tùy vào độ tuổi, giới tính, sở thích phim ảnh khác nhau, người ta có xu hướng xem những loại phim khác nhau. Khi tìm được những người có độ tuổi, giới tính, sở thích tương tự, thì khả năng họ có cùng xu hướng xem những thể loại phim giống nhau là rất cao. Một số cách kết hợp khác như người ta triển khai hai hệ thống riêng biệt, một dùng phương pháp dựa trên nội dung, một dùng phương pháp collaborative, chỉ số dự đoán sau cùng được đưa ra sau khi tổ hợp tuyến tính chỉ số của hai hệ thống riêng biệt. Hoặc tùy từng thời điểm mà người ta chọn dùng hệ thống này hay hệ thống kia tùy thuộc vào chất lượng thông tin gợi ý của hệ thống nào tốt hơn.

Người ta cũng đã nghiên cứu và nhận xét rằng hệ thống khuyến nghị lai sẽ cho ra kết quả gợi ý tốt hơn so với từng hệ thống khuyến nghị riêng lẻ dùng phương pháp dựa trên nội dung hoặc phương pháp collaborative.