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

## 

**Mô hình khuyến nghị truyền thống (2 chiều)**

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 khuyến nghị sử dụng bộ nhớ**

Phương pháp khuyến nghị 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**

Việc thiết kế và mô phỏng các mô hình (máy học và các thuật toán khai phá dữ liệu) cho phép các hệ thống học và nhận ra các mô hình (pattern) phức tạp thông qua bước huấn luyện dữ liệu, và sau đó có thể đưa ra những dự đoán thông minh trong hệ thống khuyến nghị. Các thuật toán khuyến nghị dựa trên mô hình như mô hình Bayes (Bayesian models), mô hình gom cụm (clustering model), mạng phụ thuộc (dependency network) đã được đưa ra để giải quyết những thiếu sót của phương pháp khuyến nghị sử dụng bộ nhớ [12, 13]. Thông thường các thuật toán phân lớp được sử dụng trong các hệ thống khuyến nghị nếu đánh giá của người dùng là tuyệt đối, trong khi đó các mô hình hồi qui và phương thức SVD (singular value decomposition) áp dụng cho các đánh giá bằng số liệu. Trong giới hạn đề tài (các đánh giá của người dùng là số liệu) chúng tôi tiến hành khảo sát các mô hình hồi qui, có thể tóm gọn như sau:

Quay trở lại phương pháp khuyến nghị sử dụng bộ nhớ đã đề cập ở trên, có thể nhận ra trong nhiều trường hợp hai vector có thể có khoảng cách Euclide xa nhau nhưng chúng lại cũng có thể rất gần nhau với độ đo cosine hay Pearson. Sử dụng mô hình hồi qui trong hệ thống khuyến nghị sẽ giải quyết tốt hơn những vấn đề trên. Hồi qui là phương pháp sử dụng một phép toán xấp xỉ để dự đoán đánh giá của người dùng. Ta có thể xem tập X = (X1, X2, …, Xn) là tập biến số đại diện cho sự ưa thích của một người dùng đối với những item khác nhau. Mô hình hồi qui tuyến tính có thể được phát biểu như sau:

Trong đó α là một ma trận n × k, β = (β1, β2, …, βn) đại diện cho độ nhiễu trong lựa chọn của người dùng, Y là một ma trận n × m với Yij là đánh giá của người dùng i cho đối tượng j, X là một ma trận k × m với mỗi cột là một đánh giá của một người dùng trong không gian đánh giá k chiều. Ta có thể dễ dàng thấy được ma trận Y ở đây rất thưa thớt. Để giải quyết vấn đề này, Canny [14] đề xuất một kỹ thuật phân tích (sparse factor analysis) giúp thay thế những ô trống này bằng những giá trị mặc định (giá trị trung bình của một vài ô có giá trị, trung bình theo cột, theo dòng, hoặc tất cả).

[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.

**Mô hình khuyến nghị đa chiều**

Trên đây là cái nhìn tổng quan về các hệ thống khuyến nghị truyền thống. Ta có thể nhận thấy điểm chung ở các hệ thống trên là chúng chỉ quan tâm đến người dùng và đối tượng (chỉ hai chiều) chứ không quan tâm đến những thông tin ngữ cảnh bên ngoài. Có nhiều cách định nghĩa khái niệm thông tin ngữ cảnh. Định nghĩa được dùng nhiều nhất và phù hợp nhất trong tình huống này là của Dey (2001): “Thông tin ngữ cảnh là những thông tin có thể mô tả được hoàn cảnh của một thực thể. Thực thể ở đây có thể là người, là vật hoặc là đối tượng có liên quan đến sự tương tác giữa người dùng và ứng dụng, bao gồm cả bản thân người dùng và ứng dụng đó.” Ví dụ dễ hiểu hơn, thời gian, nơi chốn, thời tiết, tâm trạng … là những thông tin ngữ cảnh. Chúng có thể ảnh hưởng đến các chỉ số đánh giá của người dùng đối với đối tượng, từ đó kéo theo sự ảnh hưởng của những gợi ý trong hệ thống khuyến nghị. Từ nhu cầu thực tế đó, dẫn đến sự ra đời của các hệ thống khuyến nghị đa chiều (ngoài hai chiều là người dùng và đối tượng, mở rộng thêm các chiều khác được quan tâm như các chiều biểu thị các thông tin của điều kiện ngữ cảnh).

Trong hệ thống khuyến nghị hai chiều, ta có hàm R dùng để dự đoán các chỉ số đánh giá chưa biết như sau:

*R: Users* x *Items 🡪 Ratings*

Giờ đây, với hệ thống khuyến nghị nhiều chiều, hàm R được bổ sung thêm thông tin ngữ cảnh và trở thành:

*R: Users* x *Items* x *Contexts🡪 Ratings*

Ví dụ trong hệ thống khuyến nghị du lịch, ta có người dùng là những người cần được hệ thống gợi ý những điểm du lịch, đối tượng là những điểm du lịch, và ngữ cảnh là thời gian đi (buổi sáng/buổi tối, các tháng, các mùa), bạn đồng hành (đi một mình, đi với bạn trai bạn gái, đi với gia đình, trẻ nhỏ …). Khi đó những chỉ số đánh giá cho một điểm du lịch bởi một người sẽ phụ thuộc vào những ngữ cảnh đó. Ví dụ vào buổi tối mùa thu đi cùng gia đình thì địa điểm X là hấp dẫn nhất.

Để dễ hình dung khái niệm nhiều chiều, người ta thường dùng mô hình dữ liệu nhiều chiều dựa trên OLAP. Giả sử ta có các chiều là D1, D2, D3, …, Dn. Trong đó có hai chiều như đã biết là chiều “người dùng” và chiều “đối tượng”. Còn lại là các chiều “ngữ cảnh”. Mỗi chiều là một tập con của một tập hơp tích Descartes gồm nhiều thuộc tính. Di Ai1 x Ai2 x … x Aik , trong đó mỗi thuộc tính Aik định nghĩa một miền giá trị. Thêm nữa, một hoặc nhiều thuộc tính tạo thành một khóa để phân biệt duy nhất. Trong một số trường hợp, một chiều có để được định nghĩa bằng một thuộc tính đơn lẻ (khi đó k =1 trong Aik).

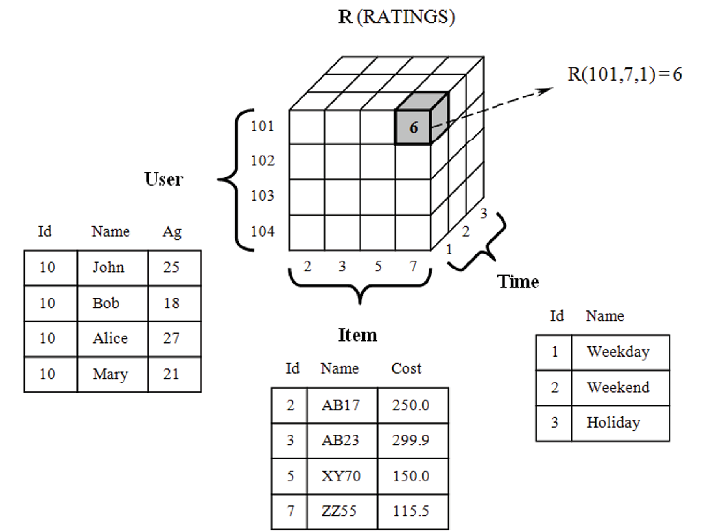
Ví dụ ta có không gian khuyến nghị ba chiều là Người dùng x Đối tượng x Thời gian. Trong đó:

*Người dùng Tên người dùng x Địa chỉ x Thu nhập x Tuổi.*

*Đối tượng Tên đối tượng x Thể loại x Giá cả*

*Thơi gian Năm x Tháng x Ngày*

Khi đó hàm *R: Users x Items x Contexts 🡪 Ratings* sẽ trở thành *R: Users x Items x Times 🡪 Ratings* có ý nghĩa rằng một người dùng *u* Users thích đối tượng *i* Items vào thời điểm *t* Times với mức độ thích thể hiện bằng một chỉ số đánh giá nào đó.



# Không gian khuyến nghị ba chiều có thể được mô tả trong khối lập phương ở hình trên. Ô tô đậm cho biết chỉ số R(101,7,1) = 6 có ý nghĩa là người dùng có mã số 101 đánh giá đối tượng có mã số 7 trong điều kiện thời gian có mã số 1 với chỉ số đánh giá là 6. Trong khối lập phương trên, không phải ô nào cũng có giá trị do người dùng chưa tiến hành đánh giá. Mục tiêu của hệ thống khuyến nghị là dự đoán giá trị tại những ô còn thiếu đó, từ đó đưa ra lời gợi ý đến với người dùng. Chúng ta cũng cần xem xét rằng không phải tất cả các điểu kiện ngữ cảnh đều cần thiết cho mục đích gợi ý. Ví dụ trường hợp gợi ý người dùng mua sách. Xét các điều kiện ngữ cảnh sau: mục đích mua sách (để giải trí, để học …), thời gian sẽ đọc quyển sách đó (đầu tuần, cuối tuần …), nơi sẽ đọc sách (ở trường, ở nhà, trên máy bay …), giá cổ phiếu ở thị trường chứng khoán vào thời điểm mua sách. Ở đây, có thể thấy giá cổ phiếu ít có ảnh hưởng đển quyết định mua sách người dùng hơn là mục đích mua sách.

# Trên thế giới đã có những nghiên cứu về vấn đề xây dựng hệ thống khuyến nghị có điều kiện ngữ cảnh kèm theo, hiện tại người ta phân thành hai nhóm: một là dùng điều kiện ngữ cảnh rồi tiến hành truy vấn, tìm kiếm những nội dung phù hợp cho việc gợi ý, hai là gợi ý dựa vào suy luận và dự đoán những sở thích của người dùng (có liên quan đến điều kiện ngữ cảnh, sở thích bị điều kiện ngữ cảnh tác động). Nhóm một được sử dụng rộng rãi trong các hệ thống khuyến nghị du lịch. Những hệ thống này thường dùng những điểu kiện ngữ cảnh hiện tại được cung cấp trực tiếp từ người dùng (sở thích, tâm trạng …) hoặc từ môi trường (thời gian, thời tiết, vị trí hiện tại …) sau đó truy vấn, tìm kiếm những nguồn tài nguyên thích hợp nhất để đưa ra gợi ý. Ví dụ như hệ thống sẽ chỉ gợi ý những nhà hàng gần nhất với vị trí hiện tại của người dùng. Còn với nhóm hai, người ta dùng kỹ thuật để mô hình hóa và học những thói quen, sở thích của người dùng bằng cách theo dõi sự tương tác giữa người dùng với hệ thống hoặc suy ra từ những đánh giá của người dùng đối với những đối tượng trước đây. Nghĩa là người dùng không cẩn cung cấp thông tin ngữ cảnh một cách trực tiếp mà hệ thống sẽ tự động thu thập, phân tích và suy luận rồi đưa ra lời gợi ý phù hợp nhất với người dùng.

# Vậy chúng ta sẽ sử dụng những thông tin ngữ cảnh trong quá trình đưa ra gợi ý vào lúc nào? Hình dưới đầy mô tả ba trường hợp:

# 

# Trường hợp a: điều kiện ngữ cảnh được sử dụng ở pha chọn lọc không gian dữ liệu hệ thống dùng để gợi ý. Dữ liệu ban đầu gồm người dùng, đối tượng, ngữ cảnh và những chỉ số đánh giá tương ứng. Điều kiện ngữ cảnh hiện tại *c* của người dùng *u* được dùng để chọn ra tập dữ liệu có liên quan đến đúng ngữ cảnh đó. Như vậy, sau khi chọn, ta có thể bỏ qua điểu kiện ngữ cảnh *c*. Bài toán trở về bài toán khuyến nghị hai chiều. Từ đây, ta có thể dùng bất cứ thuật toán khuyến nghị hai chiều nào để tiến hành dự đoán các chỉ số đánh giá.

# Ví dụ cho trường hợp này là hướng tiếp cận dựa trên việc cắt giảm số chiều (ở đây là chiều ngữ cảnh) *(reduction-based)*. Lợi ích của việc này là có thể tái sử dụng tất cả những kỹ thuật khuyến nghị hai chiều sau khi chiều ngữ cảnh được cắt giảm. Để dễ hình dung, ta tiếp tục ví dụ gợi ý phim với không gian gợi ý gồm người dùng, đối tượng và thời gian. Quá trình cắt giảm được minh họa như sau:

Ở đây, ta cần gợi ý cho người dùng *u* những bộ phim *i* mà người đó có thể thích xem vào thời điểm *t.* Trước hết, hệ thống tiền hành chọn ra tập những chỉ số đánh giá có thời điểm đánh giá là *t*. Khi đó, số chiều D của không gian khuyến nghị sẽ được cắt giảm chỉ còn người dùng, đối tượng với các chỉ số đánh giá tương ứng. Từ đây, sẽ tiến hành các thuật toán khuyến nghị hai chiều (người dùng và đối tượng) để đưa ra gợi ý.

# Trường hợp b: điều kiện ngữ cảnh được sử dụng ở pha sau khi đã có lời gợi ý. Ban đầu, vẫn là không gian dữ liệu có kèm thông tin ngữ cảnh. Người ta lờ đi xem như những dữ liệu đánh giá đó không bị ảnh hưởng bởi những thông tin ngữ cảnh. Sau đó dùng những thuật toán khuyến nghị hai chiều để tiến hành dự đoán những chỉ số đánh giá và gợi ý trên toàn bộ tập dữ liệu ban đầu. Sau khi có kết quả gợi ý cho người dùng *u*, họ lọc ra những lời gợi ý phù hợp với điều kiện ngữ cảnh *c* rối mới chuyển đến người dùng.

# Ví dụ tương tự như với trường hợp a. Hệ thống dự đoán những bộ phim người dùng có thể thích mà không quan tâm ngữ cảnh. Sau khi có danh sách các bộ phim đó, hệ thống sẽ dùng những điều kiện ngữ cảnh để lọc lại danh sách phim sao cho phù hợp. Ví dụ nếu hệ thống biết rằng người dùng *u* chỉ thích xem phim kinh dị vào những ngày cuối tuần thì trong danh sách phim trên, hệ thống sẽ chỉ chọn ra những phim nào thuộc thể loại kinh dị và được những người dùng quan tâm xem lúc cuối tuần.

# Trường hợp c: những điều kiện ngữ cảnh được dùng một cách trực tiếp trong kỹ thuật mô hình hóa ở ngay pha dự đoán chỉ số đánh giá. Sau đó đưa ra những lời gợi ý. Những thuật toán được dùng ở đây nhìn chung phức tạp hơn nhiều so với những thuật toán khuyến nghị hai chiều truyền thống. Chúng cũng được phân thành hai nhóm là nhóm dựa trên ký ức, và nhóm dựa trên mô hình. Người ta nghiên cứu và tính toán những trọng số giữa những chỉ số dự đoán trong không gian nhiều chiều thay vì chỉ tính toán độ tương tự giữa người dùng với người dùng, đối tượng với đối tượng. Hoặc xây dựng những mô hình tính toán hồi quy tuyến tính phục vụ cho việc dự đoán những chỉ số đánh giá còn thiếu.

# Do tính chất phức tạp cũng như những giới hạn trong quá trình thực hiện luận văn này, chúng tôi xin phép không đi vào chi tiết những kỹ thuật này. Phần khảo sát sơ bộ những hệ thống, những kỹ thuật khuyến nghị đã được những người đi trước thực hiện xin được phép dừng lại tại đây. Chương tiếp theo sẽ là phần trình bày chi tiết thuật toán chúng tôi nghiên cứu, tìm hiểu và sử dụng trong hệ thống khuyến nghị du lịch mà chúng tôi đang xây dựng.