Recommendation systems: Principles, methods and evaluation

# Abstract

Trong hệ thống Internet ngày nay, khi mà sự lựa chọn trở nên quá nhiều, chúng cần phải được lọc, sắp xếp thứ tự ưu tiên và các thông tin có liên quan phải được cung cấp một cách hiệu quả nhằm giảm bớt sự quá tải thông tin. Điều này đã tạo nên một vấn đề tiềm ẩn cho người sử dụng Internet.

Các hệ thống kiến nghị sẽ giải quyết được vấn đề này bằng cách tìm kiếm thông qua một lượng lớn các thông tin động và cung cấp cho người dùng những nội dung và dịch vụ mang tính cá nhân hóa.

Bài báo cáo này sẽ khảo sát các đặc điểm và tiềm năng của các kỹ thuật dự đoán khác nhau trong hệ thống kiến nghị.

# Introduction

Sự bùng nổ của lượng thông tin số và lượng người dùng Internet đã tạo ra một thử thách tiềm năng về tình trạng quá tải thông tin, điều nay gây cản trở sự tức thời mà Internet mang lại. Các hệ thống truy xuất thông tin chẳng hạn như Google, DevilFinder hay Altavista đã giải quyết một phần nào đó vấn đề này, nhưng tính cá nhân hóa và tối ưu hóa vẫn còn đang rất sơ sài. Điều này đã làm tăng nhu cầu trong việc sử dụng các hệ thống kiến nghị hơn bao giờ hết.

Hệ thống kiến nghị là các hệ thống lọc thông tin để đối phó với tình trạng quá tải thông tin bằng cách lọc phần thông tin quan trọng ra khỏi lượng lớn thông tin sẵn có, dựa trên sở thích hoặc là các hành vi quan sát được từ người dùng. Hệ thống kiến nghị có khả năng dự đoán liệu một người dùng cụ thể có quan tâm tới một nội dung nào đó hay không dựa trên hồ sơ lưu trữ của người dùng đó.

Hệ thống kiến nghị mang lại lợi ích cho cả nhà cung cấp lẫn người dùng sử dụng dịch vụ. Chúng giảm thiểu chi phí giao dịch trong việc tìm kiếm và lựa chọn các mặt hang ở môi trường mua sắm trực tuyến. Hệ thống kiến nghị còn chứng minh được khả năng cải thiện trong quá trình đưa ra quyết định. Trong các thư viện khoa học, hệ thống kiến nghị còn hỗ trợ người dùng bằng cách cho phép họ bằng những tìm kiếm vượt ra ngoài cả những danh mục thường dùng.

# Related work

Hệ thống kiến nghị được định nghĩa như một chiến thuật đưa ra quyết định cho người dùng dưới một môi trường thông tin phức tạp. Hệ thống kiến nghị được định nghĩa từ góc nhìn của thương mại điện tử như một công cụ giúp người dùng tìm kiếm thông qua bộ cơ sở dữ liệu liên quan tới sở thích và thói quen tiêu thụ của chính họ. Nó còn là một phương tiện để giúp đỡ và tăng cường tính tương tác của các mạng xạ hội bằng cách đồng thời sử dụng các khuyến nghị của người khác để đưa ra lựa chọn cuối cùng. Hệ thống kiến nghị xử lý các vấn đề quá tải về thông tin mà người dùng thường gặp phải bằng cách cung cấp cho họ những nội dung và dịch vụ mang tính cá nhân hóa.

Gần đây, các phương pháp tiếp cận khác nhau cho hệ thống khuyễn nghĩ đã được phát triển: lọc cộng tác, lọc sửa trên nội dung hoặc là lọc hỗn hợp

# Phase of recommendation process

## Information collection phase

Đây là bước để thu thập các thông tin liên quan của người dùng để tạo ra các hồ sơ hay mô hình của họ phục vụ cho các nhiệm vụ dự đoán. Bao gồm đặc tính, hành vi hoặc là nội dung của các nguồn tài nguyên mà người dùng đã và thường truy cập. Một tác tử khuyến nghĩ sẽ không thể hoạt động chính xác cho đến khi hồ sơ/mô hình về người dùng được xây dựng. Hệ thống cần phải được thu thập thông tin càng nhiều càng tốt từ người dùng và từ đó cung cấp những khuyến nghị có cơ sở từ trên nền tảng thông tin đó. Hệ thống khuyến nghị dựa vào các loại thông tin đầu vào khác nhau chẳng hạn như những phản hồi mang tính hài lòng rõ ràng từ người dùng, hoặc các loại thông tin ẩn dựa vào sự suy diễn thông qua quá trình quan sát người dùng. Các phản hồi hỗn hợp cũng có thể thu được thông qua sự kết hợp của cả 2 dạng phản hồi trên. Trong nền tảng E-learning, một hồ sơ người dùng là một tập hợp của các thông tin mang tính cá nhân của người đó, chẳng hạn như kĩ năng nhận thức, khả năng trí tuệ, phong cách học, sự thích thú và sự tương tác với hệ thống. Hồ sơ này sẽ được dùng để truy xuất các thông tin cần thiết để xây dựng mô hình của người dùng. Sự thành công của bất cứ một hệ thống khyến nghị nào cũng phụ thuật rất lớn vào khả năng thể hiện mô hình người dùng của chính nó.

## Explicit feedback

Hệ thống sẽ thường xuyên nhắc nhở thông qua giao diện người dùng để nhận đánh giá về sản phẩm của họ với mục đích cải thiện mô hình. Sự chính xác của sự kiến nghị phụ thuộc vào số lượng đánh giá của người dùng. Vấn đề còn lại là phương pháp này đòi hỏi sự hợp tác từ người dùng, và họ cũng thường không cung cấp đủ các thông tin cần thiết. Mặc dù điều này cần sự nỗ lực lớn từ người dùng, nhưng đó là một nguồn dữ liệu đáng tin cậy để từ đó đưa ra các khuyến nghị xác đáng.

## Implicit feedback

Hệ thống sẽ tự động lấy các số liệu của người dùng bằng cách giám sát nhiều hoạt động khác nhau của họ, chẳng hặn như lịch sử mua sắp, lịch sử chuyển hướng, và thời gian mà họ dành cho một trang web, những đường dẫn hay email có liên quan. Phương pháp này giảm thiếu gánh nặng cho người dùng bởi vì nó hoạt động độc lập dựa theo hành vi của họ. Bên cạnh đó, còn nhiều sự tranh cãi về dữ liệu mà phương pháp này mang lại. Một số ý kiến cho rằng nó chính xác khách quan hơn bởi vì nó không bị ảnh hưởng dựa theo phong trào về mặt xã hội và một số cách thức để kiếm rating khác.

## Hybrid feedback

Sức mạnh của cả EF và IF có thể tổng hợp lại thành hệ thống hỗn hợp để từ đó giảm thiểu những thiếu sót và có được một hệ thống với hiệu năng tốt nhất. Điều này có thể đạt được bằng cách sử dụng dữ liệu IF để kiểm tra thông số EF hoặc chỉ cho phép người dùng cung cấp phản hồi EF một khi họ đã thể hiện sự quan tâm đó thông qua IF.

## Learning phase

Trong bước này chúng ta sẽ áp dụng thuật toán học để lọc và khai thác các thuộc tính của người dùng có được từ bước trước đó.

## Prediction/recommendation phase

Bước này sẽ khuyến nghị hoặc dự đoán những loại hàng mà người dùng có thể thích thú hơn. Điều này có thể được thực hiện trực tiếp dựa trên bộ dữ liệu được thu thập trong giai đoạn thu thập thông tin.



# Recommendation filtering techniques

Việc sử dụng các kỹ thuật kiến nghị hiệu quả và chính xác là rất quan trọng đối với một hệ thống, nó sẽ cung cấp các kiến nghị tốt và tối ưu cho chính những người dùng đó. Điều này giải thích tầm quan trọng của việc hiểu các tính năng và tiềm năng của các kỹ thuật đề xuất khác nhau. Hình dưới cho thấy sự khác nhau cơ bản của các kỹ thuật lọc đề xuất.



## Collaborative recommendation systems

* Hệ thống gợi ý dựa theo lọc cộng tác (Collaborative recommendation systems): là phương pháp gợi ý được triển khai rộng rãi nhất và thành công nhất trong thực tế.
* Hệ thống theo lọc công tác phân tích và tổng hợp các điểm số đánh giá của các đối tượng, nhận ra sự tương đồng giữa những người sử dụng trên cơ sở các điểm số đánh giá của họ và tạo ra các gợi ý dựa trên sự so sánh này. Hồ sơ (profile) của người sử dụng điển hình trong hệ thống lọc cộng tác bao gồm một vector các đối tượng (item) và các điểm số đánh giá của chúng, với số chiều tăng lên liên tục khi người sử dụng tương tác với hệ thống theo thời gian.
* Một số hệ thống sử dụng phương pháp chiết khấu dựa trên thời gian (time-based discounting) để tính toán cho yếu tố “trượt” đối với sự quan tâm của người sử dụng. Trong một số trường hợp điểm số đánh giá (rating) có thể là nhị phân (thích/không thích) hoặc các giá trị số thực cho thấy mức độ ưu tiên.
* Thế mạnh lớn nhất của kỹ thuật gợi ý theo lọc cộng tác là chúng hoàn toàn độc lập với sự biểu diễn của các đối tượng đang được gợi ý, và do đó có thể làm việc tốt với các đối tượng phức tạp như âm thanh và phim. Schafer, Konstan & Riedl (1999) gọi lọc cộng tác là “tương quan giữa người – với – người” (people-to-people correlation).

## Content-based recommendation systems

* Hệ thống gợi ý dựa theo nội dung (Content-based recommendation systems): là sự kế thừa và mở rộng của lĩnh vực nghiên cứu lọc thông tin.
* Trong hệ thống thì các đối tượng được biểu diễn bởi các đặc điểm liên quan tới chúng.
* Ví dụ, hệ thống gợi ý văn bản như hệ thống lọc tin NewsWeeder sử dụng những từ của các văn bản như các đặc điểm.
* Một số hệ thống gợi ý dựa trên nội dung học một hồ sơ cá nhân về sở thích của người sử dụng dựa trên các đặc điểm xuất hiện trong chính các đối tượng người sử dụng đã đánh giá (rated). Schafer, Konstan & Riedl gọi gợi ý theo nội dung là “tương quan đối tượng với đối tượng” (item-to-item correlation). Hồ sơ người sử dụng của một hệ thống gợi ý theo nội dung phụ thuộc vào phương pháp học máy được dùng.
* Cây quyết định (Decision trees), mạng noron (neural nets) và biểu diễn dựa theo vector (vector-based representations) đều có thể được sử dụng để học hồ sơ người dùng. Cũng giống như trong lọc cộng tác, hồ sơ người dùng trong gợi ý dựa theo nội dung là những dữ liệu lâu dài và được cập nhật theo thời gian.

## Knowledge-based recommenders systems

* Hệ thống gợi ý dựa trên cơ sở tri thức (Knowledge-based recommenders systems): gợi ý các đối tượng dựa trên các suy luận về nhu cầu và sở thích của người dùng. Theo một nghĩa nào đó, tất cả các kỹ thuật gợi ý có thể mô tả như là làm một số suy luận. Phương pháp tiếp cận dựa trên cơ sở tri thức được phân biệt ở chỗ: chúng có kiến thức làm thế nào một đối tượng cụ thể đáp ứng nhu cầu một người dùng cụ thể, và do đó có thể lập luận về mối quan hệ giữa nhu cầu và các gợi ý cụ thể.
* Sử dụng miền tri thức rõ ràng, có liên quan tới mối quan hệ giữa yêu cầu của người dùng và sản phẩm cụ thể. Ban đầu người ta đưa ra 3 dạng tri thức: tri thức về danh mục (tri thức về sản phẩm được gợi ý), tri thức người sử dụng (tri thức về các yêu cầu của người sử dụng), tri thức về các chức năng (tri thức để ánh xạ các yêu cầu của người sử dụng tới các sản phẩm thoả mãn các yêu cầu đó).
* Phương pháp này không dựa trên tiểu sử người sử dụng nên không gặp phải khó khăn về sản phẩm mới và người dùng mới. Gợi ý trên cơ sở tri thức có khả năng suy diễn, khả năng suy diễn phụ thuộc vào độ phù hợp của yêu cầu người sử dụng với các thuộc tính của sản phẩm.
* Mọi hệ thống dựa trên cơ sở tri thức đều là mối quan hệ thu nhận tri thức. Thực tế, chất lượng của các phương án gợi ý tùy thuộc vào độ chính xác của cơ sở tri thức. Đây cũng là hạn chế lớn nhất của phương pháp này.

# Comparing

|  |  |  |
| --- | --- | --- |
|  | Adv | DisAdv |
| Collaborative | * Khả năng đa hạng mục * Không cần tri thức miền * Chất lượng tăng theo thời gian * Đủ thông tin phản hồi không tường minh | * Vấn đề người dùng mới * Vấn đề sản phẩm/đối tượng mới * Vấn đề “Gray sheep” * Chất lượng phụ thuộc vào độ lớn dữ liệu lịch sử thao tác của người sử dụng * Vần đề về tính bền vững và mềm dẻo |
| Content-based | * Không cần tri thức miền * Chất lượng tăng theo thời gian * Đủ thông tin phản hồi không tường minh | * Vấn đề người dùng mới * Chất lượng phụ thuộc vào độ lớn dữ liệu lịch sử thao tác của người sử dụng * Vần đề về tính bền vững và mềm dẻo |
| Knowledge-based | * Có thể ánh xạ giữa nhu cầu người dùng và sản phẩm/đối tượng | * Cần phải thu thập tri thức. |

# Conclusion

Các hệ thống kiến nghị ngày nay đã mở ra nhiều cơ hội mới để thu thập thông tin cá nhân trên Internet. Nó cũng giúp giảm bớt vấn đề về quá tải thông tin – một hiện tượng rất phổ biến với các hệ thống truy xuất thông tin. Trên đây đã giới thiệu một cái nhìn sơ lược về nội dung cũng như các kỹ thuật về hệ thống kiến nghị.

References:

* + <https://viblo.asia/p/gioi-thieu-ve-he-thong-goi-y-recommender-systems-hoac-recommendation-systems-maGK78yOZj2>
  + <http://www.sciencedirect.com/science/article/pii/S1110866515000341>
  + <http://infolab.stanford.edu/~ullman/mmds/ch9.pdf>
  + <https://www.coursera.org/learn/machine-learning/supplement/gXdW5/lecture-slides>
  + <https://www.coursera.org/learn/machine-learning/supplement/gXdW5/lecture-slides>

# Next steps

Tiến hành học và nghiên cứu về Recommendation System, khóa học Machine Learning ở Coursera – Andrew Ng.