ĐẠI HỌC QUỐC GIA HÀ NỘI
**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ**

**NGUYỄN THẠC ĐAN THANH**

# KHAI PHÁ DỮ LIỆU VẾT DUYỆT WEB CHO TƯ VẤN CÁ NHÂN HÓA

Ngành: Hệ thống thông tin
Chuyên ngành: Hệ thống thông tin
Mã số: 60480104

**LUẬN VĂN THẠC SĨ HỆ THỐNG THÔNG TIN**

**NGƯỜI HƯỚNG DẪN KHOA HỌC: PGS. TS. Hà Quang Thụy**

**Hà Nội - 2016**

# Lời cảm ơn

Trước tiên, tôi xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới thầy giáo, Phó Giáo sư Tiến sĩ Hà Quang Thụy, người đã tận tình chỉ bảo và hướng dẫn tôi trong suốt quá trình thực hiện đề tài luận văn.

Tôi xin gửi lời cảm ơn tới Phòng Thí nghiệm DS&KTLab và Đề tài QG.15.22, các thầy, cô giáo trong Khoa Công nghệ thông tin nói riêng và trong trường Đại học Công nghệ - Đại học Quốc Gia Hà Nội nói chung, đã tận tình giảng dạy và truyền đạt kiến thức quý báu cho tôi trong suốt quá trình học tập.

Cuối cùng, tôi muốn gửi lời cảm ơn vô hạn tới gia đình và bạn bè, những người đã luôn bên cạnh và động viên tôi trong suốt quá trình học tập cũng như quá trình thực hiện đề tài.

Tôi xin chân thành cảm ơn!

Học viên

Nguyễn Thạc Đan Thanh

# Tóm tắt nội dung

Hệ tư vấn (recommender system) đã trở thành một trong những chủ đề nghiên cứu quan trọng được ứng dụng cao trong thực tế. Hệ tư vấn ra đời nhằm đáp ứng nhu cầu tư vấn sản phẩm của thương mại điện tử (e-commerce), và ngày càng được ứng dụng rộng rãi trong hầu hết các miền ứng dụng đa dạng như mạng xã hội, các trang tin tức, giải trí, du lịch,… Một vài ứng dụng nổi tiếng như: hệ tư vấn sách, CDs của Amazon, hệ tư vấn phim của Netflix, MovieLens, gợi ý kết bạn của Facebook,...Gợi ý nội dung phù hợp cho người dùng trên một website cũng là một vấn đề đáng quan tâm của các nhà quản lý trang web hiện nay, đặc biệt là ở Việt Nam, khi mà hệ tư vấn vẫn chưa thực sự phổ biến hoặc còn khá thô sơ trên hầu hết các website. Luận văn hướng tới xây dựng một mô hình hệ tư vấn nội dung trên các trang web tiếng Việt, đưa ra gợi ý các URL (trang web thành phần) có nội dung được coi là phù hợp với từng cá nhân người dùng nhất, dựa trên phân tích vết duyệt web của người dùng.

Luận văn đề xuất một mô hình hệ tư vấn cộng tác (collaborative recommendation) cho các website tạp chí ở Việt Nam dựa trên phương pháp biểu diễn nội dung trang web theo mô hình chủ đề ẩn (Latent Dirichlet Allocation - LDA [1]). Nội dung các trang web từ vết duyệt web (“mối quan tâm trong quá khứ”) của người dùng được so sánh với nội dung các trang web hiện thời và sau đó hệ thống đưa ra gợi ý các trang web hiện thời (qua URL) phù hợp với quan tâm của người dùng. Thực nghiệm ban đầu của hệ thống cho kết quả khả quan.

**Từ khóa**: *recommender system, collaborative, LDA*

# Lời cam đoan

Tôi xin cam đoan mô hình hệ tư vấn nội dung trên website và thực nghiệm được trình bày trong luận văn là do tôi đề ra và thực hiện dưới sự hướng dẫn của PGS. TS Hà Quang Thụy.

Tất cả các tài liệu tham khảo từ các nghiên cứu liên quan đều có nguồn gốc rõ ràng từ danh mục tài liệu tham khảo trong luận văn. Trong luận văn, không có việc sao chép tài liệu, công trình nghiên cứu của người khác mà không chỉ rõ về tài liệu tham khảo.

Hà Nội, ngày tháng năm 2016

Học viên

Nguyễn Thạc Đan Thanh

# Mục lục

# Danh sách bảng

# Danh sách hình vẽ

# Danh sách từ viết tắt

| STT | Tên viết tắt | Cụm từ đầy đủ |
|---|---|---|
| **1** | CF | Collaborative Filtering |
| **2** | LDA | Latent Dirichlet Allocation |
| **3** | pLSA | Probabilistic Latent Semantic Analysis |
| **4** | RMSE | Root Mean Square Error |
| **5** | MAE | Mean Absolute Error |

# Lời mở đầu

Internet mang đến cho con người nhiều tiện ích khác nhau, bạn có thể tìm kiếm bất cứ thông tin hoặc sản phẩm mình cần thông qua Internet. Tuy nhiên, đối mặt với tình trạng bùng nổ thông tin hiện nay, càng ngày bạn càng mất nhiều thời gian trong việc lựa chọn thông tin hay sản phẩm nào phù hợp với mình. Cùng với sự phát triển của thương mại điện tử (e-commerce), hệ tư vấn xuất hiện với vai trò vô cùng quan trọng trong việc hỗ trợ người dùng lựa chọn sản phẩm phù hợp đồng thời tăng giá trị kinh doanh cho doanh nghiệp. Và cho đến nay, hệ tư vấn được ứng dụng rộng khắp trên nhiều lĩnh vực khác như mạng xã hội, các trang tin tức, giải trí, du lịch,…, với bất cứ thông tin nào người dùng quan tâm thì chúng ta đều có thể thấy sự xuất hiện của hệ tư vấn.

Hệ tư vấn có thể nói đã thay đổi cách thức mà người dùng giao tiếp với các trang web, thay vì người dùng phải chủ động tìm kiếm và lựa chọn thông tin mình cần thì nhờ có hệ tư vấn, website có thể giới thiệu, gợi ý những sản phẩm, thông tin được cho là cần thiết, phù hợp nhất với người dùng dựa trên profile của họ. Profile của người dùng có thể là lịch sử giao dịch mua bán sản phẩm trên các trang bán hàng trực tuyến, những đánh giá hay các tương tác của người dùng với các trang web. Mặc dù vai trò và lợi ích của một hệ tư vấn là rất lớn, tuy nhiên ở Việt Nam, hệ thống này vẫn chưa thực sự phổ biến và còn khá thô sơ. Đa phần các trang web Việt Nam hiện nay chưa có một hệ thống gợi ý hiệu quả dựa trên profile của người dùng, mà chỉ sử dụng các phương pháp đơn giản như gán nhãn tay (thẻ catagoried tags), thống kê để gợi ý những thông tin, sản phẩm liên quan với sản phẩm đang được xem, hay gợi ý những thông tin nổi bật nhiều người quan tâm.

Chính vì vậy, luận văn mong muốn xây dựng một mô hình hệ tư vấn tự động trên các website tạp chí tiếng việt, nhằm mục đích gợi ý những nội dung liên quan tới sở thích của từng cá nhân người dùng, dựa trên lịch sử duyệt web của họ trên website đó (vết duyệt web).

Nội dung của luận văn bao gồm những nội dung sau:

Chương 1. Hệ tư vấn và bài toán tư vấn: Trình bày những nội dung cơ bản về hệ tư vấn bao gồm mô tả bài toán tư vấn, ứng dụng và các hệ thống nổi tiếng, phân loại các kĩ thuật tư vấn.

Chương 2. Khai phá dữ liệu vết duyệt web của người dùng cho hệ tư vấn: Phân loại dữ liệu profile người dùng, ưu nhược điểm của từng loại dữ liệu và một số nghiên cứu, phương pháp ứng dụng trên các miền dữ liệu này. Giới thiệu về hệ tư vấn nội dung website dựa trên vết duyệt web được xây dựng trong luận văn.

Chương 3. Mô hình hệ tư vấn nội dung trên website dựa trên vết duyệt web: Trình bày mô hình tư vấn nội dung trên một website do chúng tôi đề xuất, là mô hình tư vấn cộng tác kết hợp phương pháp ước lượng hạng giả định theo mô hình chủ đề ẩn LDA.

Chương 4. Thực nghiệm và đánh giá: Thử nghiệm và đánh giá mô hình hệ thống với dữ liệu thực tế từ trang web http://www.otoxemay.vn/ và http://www.emdep.vn/

Phần kết luận tổng kết nội dung chính của luận văn, các vấn đề còn tồn tại và định hướng phát triển của hệ thống.

# Chương 1 Hệ tư vấn và bài toán tư vấn

## 1.1 Giới thiệu về hệ tư vấn

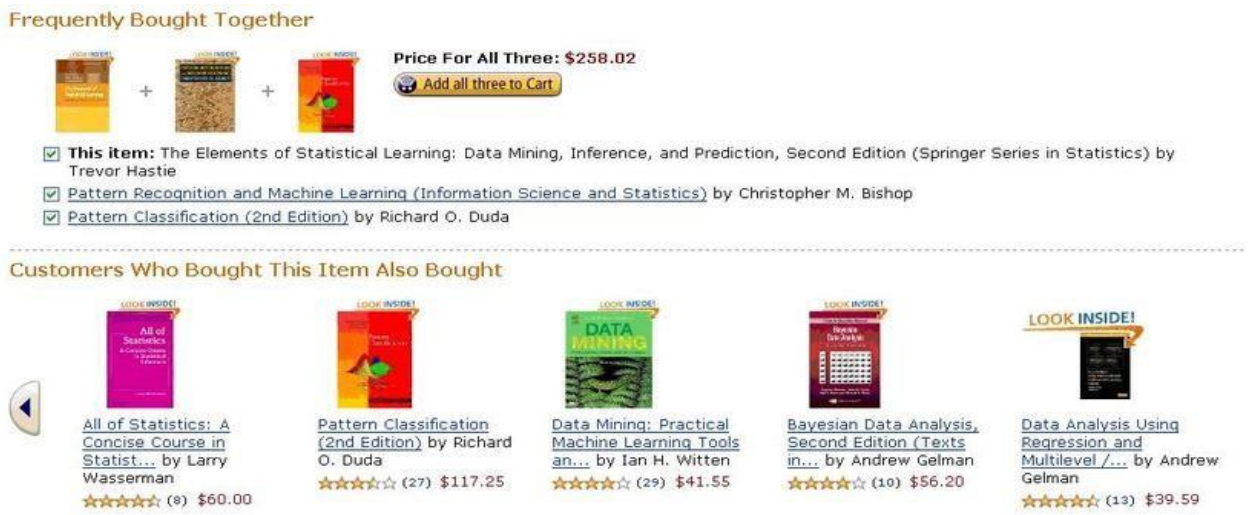
Hệ tư vấn (recommender system, còn được gọi là hệ gợi ý) là công cụ phần mềm và kỹ thuật cung cấp các tư vấn về các mục (item; ví dụ phim, CD, nhà hàng,…) cho một người dùng [2]. Item là thuật ngữ chung để chỉ những gì mà hệ thống muốn tư vấn cho người dùng. Một hệ tư vấn truyền thống thường tập trung tư vấn một mục nhất định để đạt được hiệu quả tối đa cho từng loại mục cụ thể. Hệ tư vấn thường hướng tới cá nhân người dùng, tức là với mỗi người dùng khác nhau sẽ nhận được một danh sách mục tư vấn khác nhau. Hệ thống này đưa ra gợi ý dựa trên những gì người dùng đã làm trong quá khứ, hoặc dựa trên tổng hợp ý kiến của những người dùng khác. Hệ tư vấn phát triển lên như một lĩnh vực nghiên cứu độc lập vào giữa thập niên 90. Trong những năm gần đây, sự quan tâm về hệ tư vấn đã tăng lên đáng kể, được minh chứng qua các sự kiện sau [2]:

- Các hội nghị, hội thảo chuyên nghiên cứu về lĩnh vực này đã được tổ chức. Đặc biệt là ACM Recommender Systems (RecSys), thành lập năm 2007 và giờ đây là sự kiện được tổ chức thường niên vào đầu mỗi năm trong nghiên cứu công nghệ tư vấn và các ứng dụng liên quan. Ngoài ra, các buổi trao đổi dành riêng cho hệ tư vấn thường được đề cập trong các hội nghị truyền thống trong lĩnh vực cơ sở dữ liệu, hệ thống thông tin và hệ thống thích nghi. Trong số các hội nghị, đáng được nhắc đến nhất là hội nghị về các nhóm lĩnh vực đặc biệt quan tâm trong truy hồi thông tin (ACM SIGIR Special Interest Group on Information Retrieval - SIGIR), hội nghị về mô hình hóa, thích ứng và cá nhân hóa người dùng (User Modeling, Adaptation and Personalization - UMAP), nhóm vấn đề đặc biệt chú ý của ACM trong quản lý dữ liệu (ACM’s Special Interest Group on Management of Data - SIGMOD)
- Tại các tổ chức giáo dục đại học trên khắp thế giới, đại học và sau đại học có các khóa học được tập trung hoàn toàn vào hệ tư vấn; hướng dẫn về hệ tư vấn rất phổ biến tại các hội nghị khoa học máy tính; và nhiều cuốn sách giới thiệu các kỹ thuật tư vấn đã được xuất bản, chẳng hạn [2], [3], [4].
- Đã có một số công bố đặc biệt trong tạp chí khoa học bao gồm các nghiên cứu và phát triển trong lĩnh vực hệ tư vấn. Trong số các tạp chí có những công trình chuyên về hệ tư vấn như: hệ truyền thông AI (2008), hệ thống thông minh IEEE (2007), tạp chí quốc tế về thương mại điện tử (2006), tạp chí quốc tế về khoa học và ứng dụng (2006), giao dịch trên máy tính ACM tương tác người – máy (2005), và giao dịch ACM trên hệ thống thông tin.

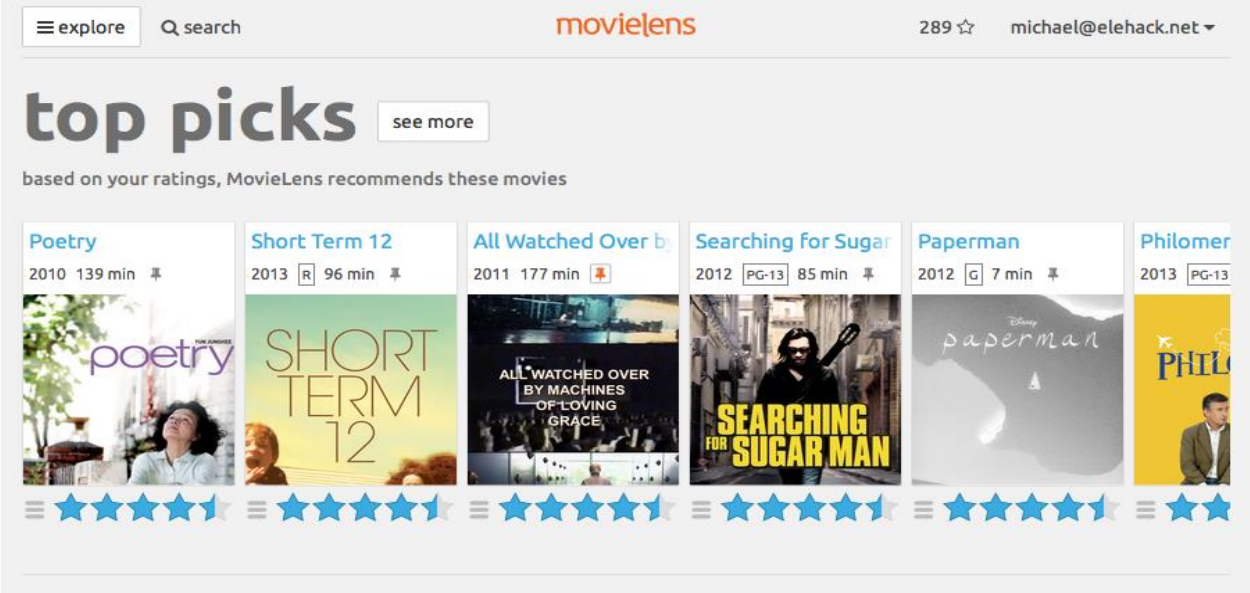
- Hiện nay, hệ tư vấn đóng vai trò rất quan trọng trong nhiều các trang web được đánh giá cao như Amazon.com, Youtube, Netflix,…Một số ứng dụng hệ tư vấn nổi tiếng trên thế giới được giới thiệu trong bảng 1.1

Bảng 1.1 Ví dụ một số hệ tư vấn nổi tiếng **[3]**

| **Hệ tư vấn** | **Mục tư vấn** |
|---|---|
| Amazon.com | Sách, CD, và một số sản phẩm khác |
| Netflix | DVD, streaming video |
| GroupLens | Tin tức |
| MovieLens | Phim ảnh |
| Google News | Tin tức |
| Facebook | Bạn bè, quảng cáo |
| Pandora | Âm nhạc |
| YouTube | Video trực tuyến |
| Tripadvisor | Sản phẩm về du lịch (nhà hàng, khách sạn, …) |




Hình 1.1 Hệ tư vấn sách của Amazon




Hình 1.2 Hệ tư vấn phim của MovieLens

## 1.2 Bài toán tư vấn

Theo Adomavicius và Tuzhilin trong [5], trong hầu hết các trường hợp, bài toán tư vấn được coi là bài toán ước lượng trước hạng của các item chưa được người dùng xem xét. Việc ước lượng này thường dựa trên những đánh giá đã có của chính người dùng đó hoặc những người dùng khác. Những item có hạng cao nhất sẽ được dùng để tư vấn.

Một cách hình thức, bài toán tư vấn được mô tả như sau:

Gọi $C$ là tập tất cả người dùng; $S$ là tập tất cả các item có thể tư vấn. Tập $S$ có thể rất lớn, từ hàng trăm ngàn (video, phim,…) đến hàng triệu (như website). Tập $C$ trong một số trường hợp cũng có thể lên tới hàng triệu. Mỗi người dùng trong không gian $C$ được xác định bởi một hồ sơ (profile). Profile này có thể gồm rất nhiều loại thông tin: nghề nghiệp, giới tính, sở thích, … hoặc có thể chỉ gồm một trường mã số người dùng (user id) duy nhất. Tương tự, mỗi item trong không gian $S$ cũng được xác định bởi một tập các đặc trưng. Ví dụ, trong hệ tư vấn sách, đặc trưng của mỗi quyển sách có thể là: tên, thể loại, tác giả, năm xuất bản, nhà xuất bản, chủ đề chính, mục lục,…

Hàm $u(c,s)$ đo độ phù hợp (hay hạng) của item $s$ với user $c$: $u: C \times S \rightarrow R$. Với mỗi người dùng $c \in C$, cần tìm sản phẩm $s' \in S$ sao cho hàm $u(s', c)$ đạt giá trị lớn nhất: $\forall c \in C, s'_c = arg\ max\ u(c, s'), s' \in S$

Vấn đề chính của hệ tư vấn là các giá trị hàm $u$ chưa có được trên toàn không gian $R$ mà chỉ trên một miền nhỏ của không gian đó, các giá trị đó có thể được xác định bởi người dùng hoặc được tính toán bởi hệ thống từ những thông tin về người dùng cho trước. Điều này dẫn tới việc hàm $u$ phải được ngoại suy trong không gian $R$. Thông

thường, mật độ của ma trận đánh giá trong hệ tư vấn thường rất thưa, điều đó cho thấy còn rất nhiều đánh giá chưa biết trong không gian *R*. [6] Sarwar và cộng sự nhận định rằng mật độ của ma trận đánh giá trong hệ thống thường ít hơn 1%. Nhiệm vụ của hệ tư vấn là ngoại suy, dự đoán hạng mà người dùng $c_i$ đánh giá một item $s_j$ chưa được đánh giá, từ đó đưa ra danh sách các item có hạng cao nhất với người dùng $c_i$.

## 1.3 Các kĩ thuật tư vấn

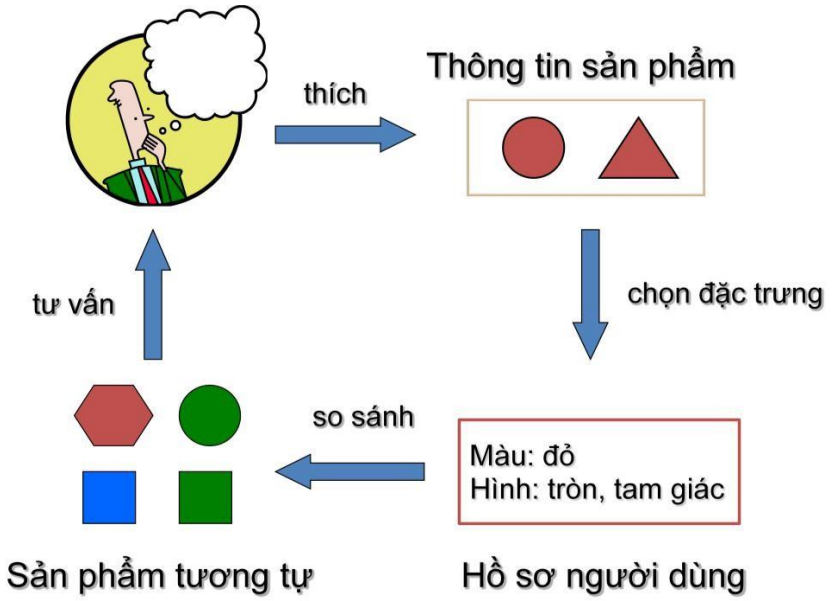
Có rất nhiều cách để dự đoán, ước lượng hạng cho các sản phẩm, theo [4] các hệ thống tư vấn thường được phân thành bốn loại dựa trên kỹ thuật tư vấn:

- Dựa trên nội dung (content-based): người dùng được gợi ý những sản phẩm tương tự như các sản phẩm từng được họ đánh giá cao.
- Cộng tác (collaborative): người dùng được gợi ý những sản phẩm được đánh giá cao bởi những người có cùng sở thích với họ.
- Dựa trên tri thức (knowledge-based): người dùng được gợi ý các sản phẩm đáp ứng với các yêu cầu đặt ra của họ.
- Lai ghép (hybrid): kết hợp các phương pháp trên.

### 1.3.1 Kĩ thuật tư vấn dựa trên nội dung

Hệ tư vấn dựa trên nội dung đưa ra các tư vấn dựa trên phỏng đoán rằng một người có thể thích các sản phẩm có nhiều đặc trưng tương tự với các sản phẩm mà họ đã từng ưa thích. Theo [5], với phương pháp tư vấn dựa trên nội dung, độ phù hợp $u(c, s)$ của sản phẩm $s$ với người dùng $c$ được đánh giá dựa trên độ phù hợp $u(c, s_i)$, trong đó $s_i \in S$ và "tương tự" như $s$. Ví dụ, để gợi ý một bộ phim cho người dùng $c$, hệ thống tư vấn sẽ tìm các đặc điểm của những bộ phim từng được c đánh giá cao (như diễn viên, đạo diễn…); sau đó chỉ những bộ phim tương đồng với sở thích của $c$ mới được giới thiệu.

Hướng tiếp cận dựa trên nội dung bắt nguồn từ những nghiên cứu về thu thập thông tin (IR - information retrieval) và lọc thông tin (IF - information filtering). Do đó, rất nhiều hệ thống dựa trên nội dung hiện nay tập trung vào tư vấn các đối tượng chứa dữ liệu text như văn bản, tin tức, website… Những tiến bộ so với hướng tiếp cận cũ của IR là do việc sử dụng hồ sơ về người dùng (chứa thông tin về sở thích, nhu cầu…). Hồ sơ này được xây dựng dựa trên những thông tin được người dùng cung cấp trực tiếp (khi trả lời khảo sát) hoặc gián tiếp (do khai phá thông tin từ các giao dịch của người dùng).




Hình 1.3 Tư vấn dựa trên nội dung **[7]**

Để cụ thể hơn, đặt *Content(s)* là tập thông tin (hay tập các đặc trưng) về sản phẩm *s*. Do hệ thống dựa trên nội dung được thiết kế chủ yếu dành cho các sản phẩm là text, nên nội dung sản phẩm thường được biểu diễn bởi các từ khóa (keyword): *Content(s)* = $(w_{1s}, \ldots w_{ks})$, *với* $w_{1s}, \ldots, w_{ks}$ là trọng số của các từ khóa từ 1 tới k (có thể được tính bằng TF-IDF).

Đặt *Profile(c)* là hồ sơ về người dùng *c*, bao gồm các thông tin về sở thích của *c*. Những thông tin này có được bằng cách phân tích nội dung của các sản phẩm từng được *c* đánh giá trước đó. Phương pháp được sử dụng thường là các kĩ thuật phân tích từ khóa của IR, do đó, *Profile(c)* cũng có thể được định nghĩa như một vector trọng số:

*Profile(c)* = $(w_{1c}, \ldots, w_{kc})$ với $w_{ic}$ biểu thị độ quan trọng của từ khóa i với người dùng *c*.

Trong hệ thống tư vấn dựa trên nội dung, độ phù hợp *u(c,s)* được xác định bởi công thức: *u(c,s) = score(Profile(c), Content(s))*, với *score* là một hàm được xây dựng để đo độ tương đồng giữa *Content(s)* và *Profile(c)*

Cả *Profile(c)*, *Content(s)* đều có thể được biểu diễn bằng vector trọng số từ TF-IDF (tương ứng là $\vec{w_c}, \vec{w_s}$) nên có thể đo độ tương đồng của chúng bằng độ đo cosine: *u(c,s)* = $cos(\vec{w_c}, \vec{w_s})$

Ví dụ, nếu người dùng *c* đọc nhiều bài báo thuộc lĩnh vực thời trang thì các từ khóa liên quan tới thời trang (như bộ sưu tập, thiết kế, mẫu,…) trong *Profile(c)* sẽ có trọng số cao. Hệ quả là với các bài báo *s* cũng thuộc lĩnh vực này sẽ có độ phù hợp *u(c,s)* cao hơn với người dùng *c*.

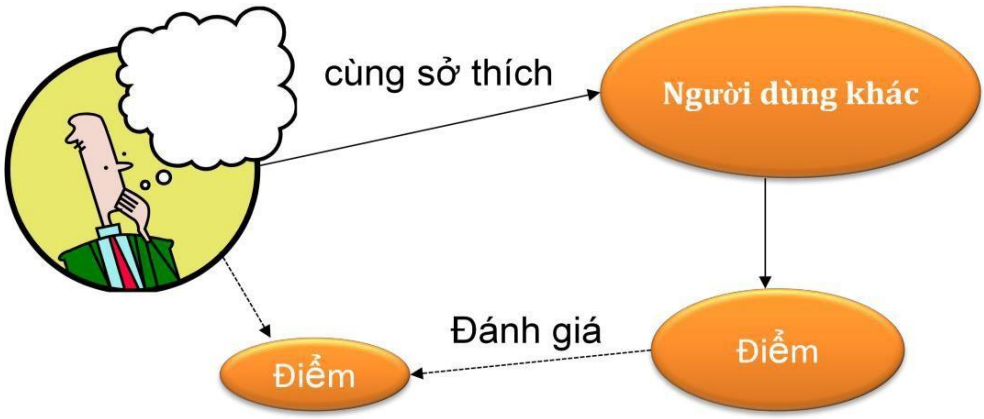
Bên cạnh các phương pháp IR, hệ tư vấn dựa trên nội dung còn sử dụng nhiều phương pháp học máy khác như: phân lớp Bayes, cây quyết định, mạng nơron nhân tạo,… Các phương pháp này khác với các phương pháp của IR ở chỗ nó dựa trên các mô hình học được từ dữ liệu nền. Ví dụ, dựa trên tập các trang web đã được người dùng đánh giá là có nội dung "hay" hoặc "không hay" có thể sử dụng phân lớp Bayes để phân loại các trang web chưa được đánh giá.

Phương pháp tư vấn theo nội dung làm việc khá hiệu quả với các tài liệu văn bản và đã có nhiều ứng dụng trên thực tế như hệ thống lọc email, thư rác,… phương pháp này vẫn được khá nhiều hệ thống tư vấn sử dụng do tính dễ cài đặt, và hiệu quả trong xử lý dữ liệu là văn bản. Nhược điểm chính của phương pháp này là gặp khó khăn trong vấn đề trích chọn đặc trưng với kiểu dữ liệu không phải là văn bản.

### 1.3.2 Kĩ thuật tư vấn cộng tác

Theo Adomavicius và cộng sự [5], không giống như phương pháp tư vấn dựa trên nội dung, hệ thống cộng tác dự đoán độ phù hợp *u(c,s)* của một sản phẩm *s* với người dùng *c* dựa trên độ phù hợp $u(c_j, s)$ giữa người dùng $c_j$ và *s*, trong đó $c_j$ là người có cùng sở thích với *c*. Ví dụ, để gợi ý một bộ phim cho người dùng *c*, đầu tiên hệ thống cộng tác tìm những người dùng khác có cùng sở thích phim ảnh với *c*. Sau đó, những bộ phim được họ đánh giá cao sẽ được dùng để tư vấn cho *c*.

Có rất nhiều hệ thống cộng tác đã được phát triển như: Grundy, GroupLens (tin tức), Ringo (âm nhạc), Amazon.com (sách, CD), Phoaks (web)… Các hệ thống này có thể chia thành hai loại: dựa trên kinh nghiệm (heuristic-based hay memory-based) và dựa trên mô hình (model-based).




Hình 1.4 Tư vấn dựa trên cộng tác **[7]**

**Hệ thống cộng tác dựa trên kinh nghiệm**

Các thuật toán dựa trên kinh nghiệm dự đoán hạng của một sản phẩm dựa trên toàn bộ các sản phẩm đã được đánh giá trước đó bởi người dùng. Lọc cộng tác dựa trên kinh nghiệm có thể được phân thành hai loại:

- Lọc cộng tác theo người dùng (user-based): người dùng tương đồng có thể đánh giá hạng cho một sản phẩm tương tự nhau, nghĩa là, hạng của sản phẩm *s* với người dùng *u (r(u,s))* được tổng hợp từ đánh giá của những người dùng *u'* khác về *s* (*u'* là người có sở thích tương đồng nhất với *u*).
- Lọc cộng tác theo item (item-based): một người dùng có thể đánh giá hạng cho các sản phẩm tương đồng một cách tương tự, nghĩa là, hạng của sản phẩm *s* với người dùng *u (r(u,s))* được tổng hợp từ đánh giá của người dùng *u* với các sản phẩm *s'* (*s'* là các sản phẩm tương đồng với *s*).

Phương pháp được thực hiện theo hai bước: Tính toán mức độ tương tự và bước tạo nên dự đoán:

- Tính toán mức độ tương tự *sim(u, u')*: Mô tả khoảng cách, sự liên quan, hay trọng số giữa hai người dùng *u* và *u'* (hoặc giữa hai sản phẩm *s* và *s'* vày).
- Dự đoán (predict): Đưa ra dự đoán cho người dùng cần được tư vấn bằng cách xác định tập láng giềng của người dùng này. Tập láng giềng của người dùng cần tư vấn được xác định dựa trên mức độ tương tự giữa các cặp người dùng hoặc sản phẩm.

**Hệ thống cộng tác dựa trên mô hình**

Mặc dù tiếp cận lọc cộng tác dựa trên kinh nghiệm về lý thuyết thì chính xác hơn bởi vì toàn bộ dữ liệu đánh giá được sử dụng cho việc tư vấn, tuy nhiên những hệ thống như vậy sẽ gặp phải vấn đề về không gian xử lý khi đối mặt với dữ liệu gồm hàng chục triệu người dùng và hàng triệu sản phẩm. Khác với phương pháp dựa trên kinh nghiệm, phương pháp dựa trên mô hình (model-based) sử dụng kĩ thuật thống kê và học máy trên dữ liệu nền (các đánh giá đã biết) để xây dựng nên các mô hình. Mô hình này sau đó sẽ được dùng để dự đoán hạng của các sản phẩm chưa được đánh giá.

Giải thưởng Netflix [8] cho thấy hiệu quả của việc sử dụng mô hình ma trận hệ số hay mô hình hệ số ẩn (matrix factorization/ latent factor model) nhằm tăng độ chính xác cho hệ tư vấn cộng tác. Bell và cộng sự [9] đã giành được giải thưởng này với thuật toán Alternating least squares (ALS), một hình thức của phương pháp ma trận hệ số. Phương pháp SVD (Singular value decomposition) là mô hình ma trận hệ số nhằm giảm số chiều của ma trận đánh giá, được áp dụng nhiều trong các hệ tư vấn như [10], [11]. Ngoài ra còn nhiều hướng tiếp cận khác như mô hình thống kê, mô hình bayes, mô hình hồi quy tuyến tính, mô hình entropy cực đại…

Hệ thống tư vấn cộng tác khắc phục được nhiều nhược điểm của hệ thống dựa trên nội dung. Một điểm quan trọng là nó có thể xử lý mọi loại dữ liệu và gợi ý mọi loại sản phẩm, kể cả những sản phẩm mới, khác hoàn toàn so với những gì người dùng đã từng xem nhờ vào tham khảo được ý kiến của những người dùng khác cùng sở thích đối với các sản phẩm, do đó có thể hiệu quả hơn đối với những sản phẩm không có những mô tả rõ ràng về đặc trưng nội dung. Tuy nhiên, hệ thống lọc dựa trên cộng tác vẫn gặp một số vấn đề như vấn đề dữ liệu thưa hay vấn đề về sản phẩm mới.

### 1.3.3 Kĩ thuật tư vấn dựa trên tri thức

Với những miền dữ liệu đặc thù khác, ví dụ như điện máy, đây là miền dữ liệu bao gồm phần lớn các sản phẩm chỉ được mua một lần trong một khoảng thời gian dài, đồng thời có những yêu cầu khá chi tiết với các sản phẩm này. Điều đó có nghĩa là hệ thống không thể dựa trên lịch sử mua bán của người dùng, phương pháp cộng tác hay dựa trên nội dung không đáp ứng trong trường hợp này. Tuy nhiên, nhiều thông tin nội dung chi tiết về đặc trưng của sản phẩm có thể có ích bao gồm thông số kĩ thuật và đặc trưng chất lượng. Ví dụ, một hệ thống tư vấn sản phẩm máy ảnh số có thể giúp người dùng tìm ra được mẫu máy phù hợp với các tiêu chí, yêu cầu của người mua đặt ra. Hệ thống như vậy xây dựng *Profile(c)* không còn là lịch sử giao dịch của người dùng mà là những yêu cầu của họ về sản phẩm, và tập *Content(s)* là các đặc trưng của sản phẩm. Theo [4], trong hầu hết các kỹ thuật tư vấn dựa trên tri thức, hệ thống đều cần thêm thông tin được cung cấp bởi khách hàng là các yêu cầu của người mua đối với sản phẩm, từ đó đưa ra tư vấn thỏa mãn yêu cầu của người dùng. Hệ tư vấn ràng buộc (Constraint-based recommender) là một ví dụ về hệ tư vấn như vậy. Một số hệ tư vấn ràng buộc như hệ tư vấn của Felfernig và Burke [12], của Zanker và cộng sự [13].

Trong ví dụ hệ tư vấn máy ảnh số, hệ tư vấn ràng buộc sử dụng các tri thức về máy ảnh như độ phân giải, khối lượng, giá tiền,… làm đặc trưng sản phẩm tư vấn. Những ràng buộc có thể được đề cập trực tiếp từ thông tin khách hàng đưa ra (như việc lựa chọn độ phân giải tối thiểu, cân nặng tối đa, giá tối đa,…) hoặc được mô tả trong một ngữ cảnh mà trong đó có đề cập đến yêu cầu đặc tính của máy ảnh, ví dụ một chiếc máy ảnh với độ phân giải cao là ưu điểm nếu như khách hàng có sở thích rửa và phóng ảnh,… Hệ tư vấn dựa trên tri thức thường được xây dựng phục vụ riêng với từng miền sản phẩm độc lập, khai thác tối đa các đặc trưng của sản phẩm, và xây dựng các giao diện thích hợp dễ dàng tương tác với người dùng, giúp thu thập được yêu cầu của người dùng một cách hiệu quả, để có thể thỏa mãn tối đa nhu cầu của khách hàng.

Các hệ thống tư vấn dựa trên tri thức có ưu điểm là hoạt động tốt ngay từ lúc đầu triển khai, không phụ thuộc dữ liệu học như các phương pháp cộng tác hay dựa trên nội dung. Tuy nhiên, đây cũng chính là nhược điểm của hệ thống này, vì không khai thác

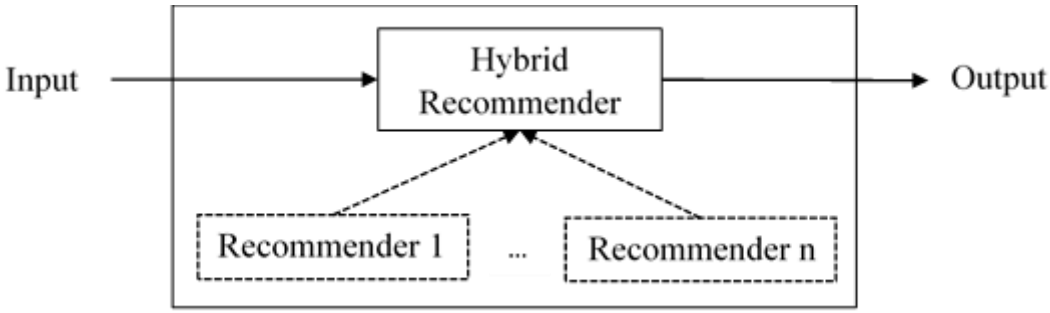
được lịch sử tương tác của con người với máy tính, do đó hạn chế về sản phẩm cũng như đối tượng tư vấn.

### 1.3.4 Phương pháp lai ghép

Các phương pháp tư vấn trên khai thác những nguồn dữ liệu khác nhau, tùy thuộc vào miền dữ liệu cũng như mục đích tư vấn để xây dựng một hệ tư vấn hiệu quả. Trong khi tư vấn cộng tác khai thác kiểu dữ liệu đánh giá hạng của người dùng cộng đồng, thì tư vấn dựa trên nội dung thường sử dụng dữ liệu nội dung của sản phẩm tư vấn trong một văn bản mô tả, mặt khác, thuật toán dựa trên tri thức lại xây dựng một kiểu hệ tư vấn phụ thuộc vào mô hình tri thức về một miền sản phẩm. Mỗi một tiếp cận trên đều có những ưu và nhược điểm riêng. Các hệ tư vấn cố gắng tận dụng thế mạnh của mỗi phương pháp, do đó kết hợp chúng cho ra một hệ tư vấn lai. Phương pháp lai ghép có thể kết hợp hai hoặc nhiều hơn các phương pháp tư vấn, nhưng nhìn chung có thể phân thành ba cách kết hợp như sau [4]:

- Xây dựng mô hình khối hợp nhất (monolithic hybridization): sử dụng kết hợp đặc trưng của các phương pháp cho đặc trưng của mô hình
- Xây dựng mô hình song song (parallelized hybridization): cài đặt các phương pháp riêng rẽ rồi kết hợp kết quả dự đoán của chúng
- Xây dựng mô hình nối tiếp (pipelined hybridization): đầu ra của phương pháp này là đầu vào của phương pháp kia.

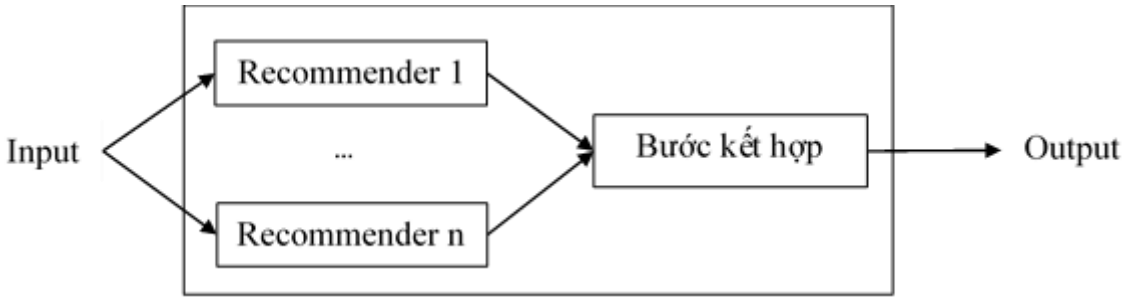
**Mô hình khối hợp nhất:**




Hình 1.5 Thiết kế của hệ tư vấn lai hợp nhất **[4]**

Phương pháp lai này hướng đến việc xây dựng một bộ trích chọn đặc trưng của nhiều kiểu dữ liệu đầu vào khác nhau đưa vào mô hình hợp nhất thuật toán. Có khá nhiều nghiên cứu về mô hình hợp nhất: Zanker và Jessenitschnig [14] đề xuất kết hợp dữ liệu đánh giá rõ ràng (explicit feedback) và đánh giá ẩn (implicit feedback) của người dùng đưa về một kiểu dữ liệu đánh giá chung cho mô hình hợp nhất của họ. Meville và cộng sự [15] đưa ra mô hình trong đó công thức dự đoán cho lọc cộng tác có tính đến trọng số của dự đoán dựa trên nội dung.

**Mô hình song song:**

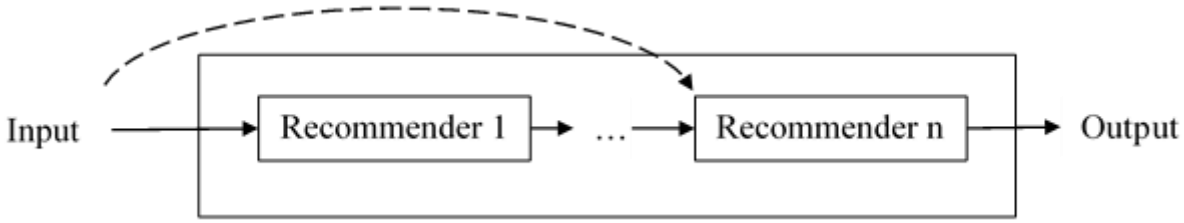



Hình 1.6 Thiết kế của hệ tư vấn lai song song **[4]**

Có hai kịch bản cho trường hợp này:

- Cách 1: Kết hợp kết quả của cả hai phương pháp thành một kết quả chung duy nhất, sử dụng cách kết hợp trộn lẫn *(mixed)* hoặc đánh trọng số kết quả của mỗi phương pháp *(weighted)*
- Cách 2: Tại mỗi thời điểm, chỉ chọn phương pháp cho kết quả tốt hơn *(switching)* dựa trên một số độ đo chất lượng tư vấn nào đó.

**Mô hình nối tiếp:**




Hình 1.7 Thiết kế của hệ tư vấn lai nối tiếp **[4]**

Hệ thống kết hợp các phương pháp tư vấn theo một trật tự nhất định, kết quả của phương pháp trước làm đầu vào của phương pháp sau. Một kết hợp nối tiếp giữa phương pháp cộng tác và dựa trên tri thức trên hệ tư vấn nhà hàng EntreeC được đề cập trong [16].

# Chương 2 Khai phá dữ liệu vết duyệt web của người dùng cho hệ tư vấn

## 2.1 Phân loại dữ liệu profile người dùng

Để xây dựng hệ tư vấn cá nhân hóa cần phải thiết lập profile của người dùng. Trong quá trình sử dụng internet như: lướt web, mua sắm online, xem phim, nghe nhạc,… người dùng sẽ thực hiện rất nhiều tác vụ khác nhau, những tác vụ này đều được ghi lại trên máy chủ của website mà người dùng vừa truy cập. Người làm khai phá dữ liệu có thể thu thập lại và khai thác những dữ liệu này để phân tích qua đó có thể tối ưu trải nghiệm của website và phục vụ người dùng tốt hơn. Những tác vụ này được chia thành hai loại là thông tin đánh giá ẩn (implicit feedback) và thông tin đánh giá rõ ràng (explicit feedback).

### 2.1.1 Thông tin đánh giá rõ ràng

Thông tin đánh giá rõ ràng là những tác vụ của người dùng đem lại dữ liệu một cách trực tiếp cho người muốn thu thập dữ liệu. Ví dụ như:

- Người dùng bấm nút like và dislike trên các mạng xã hội như Facebook, các trang chia sẻ video trực tuyến, nghe nhạc như Youtube, Pandora
- Hệ thống vote theo thang điểm (1,2,3,4 hay 5 sao) trên các trang xem phim, đọc sách,…
- Những thông tin cụ thể được người dùng cung cấp thông qua việc trả lời câu hỏi như sở thích, công việc,... trên trang mạng xã hội.
- …

Những dữ liệu này có thể được dùng trực tiếp để đánh giá về thói quen hay nhu cầu của họ, giúp hệ thống hiểu người dùng hơn và phục vụ họ tốt hơn bằng cách gợi ý những nội dung có liên quan. Có khá nhiều hệ tư vấn nổi tiếng sử dụng dữ liệu đánh giá rõ ràng như hệ tư vấn phim của Netflix, MovieLens dựa trên việc thu thập ý kiến đánh giá trực tiếp của người dùng (vote theo thang điểm).

Tuy nhiên loại thông tin đánh giá rõ ràng có một số nhược điểm:

- Khó để thu thập: khi người dùng xem xong một bản nhạc hay, mua một món đồ ưa thích, hay đọc một cuốn sách, chúng ta thường mong chờ họ đánh giá xem nội dung này thế nào đối với họ. Nhưng đừng mong họ làm thế, ít nhất là với số lượng lớn người dùng internet.
- Người dùng thường không quay lại để update những đánh giá họ đã làm trước đó: sở thích có thể thay đổi theo thời gian. Một người thích một ban nhạc nhưng sau đó có thể họ sẽ không thích ban nhạc đó nữa. Giả sử trong quá khứ người đó luôn

vote cho những bài hát của ban nhạc A. Nhưng sau này anh ta chỉ thích nghe nhạc của ca sĩ B, thì chắc chắn là anh ta cũng chẳng bao giờ quay lại website nhạc kia để update những vote đã thực hiện cho ban nhạc A.

### 2.1.2 Thông tin đánh giá ẩn

Thông tin đánh giá ẩn là những tác vụ mà người dùng thực hiện trên các website nhưng không trực tiếp yêu cầu họ phải đánh giá hay làm gì đó lên nội dung mà họ vừa xem. Nói cách khác, hệ thống chỉ quan sát xem user làm gì mà thôi, ví dụ như:

- Lịch sử duyệt web của người dùng: người dùng A khi vào một website thì chỉ xem các nội dung của mục X, và Y
- Lịch sử giao dịch mua bán trên trang bán hàng trực tuyến: người dùng B mua một vài món đồ, xem một vài sản phẩm
- Thông tin tìm kiếm trên máy tìm kiếm
- Hay phức tạp hơn như: quan sát xem người dùng có nghe hết một bài hát, nghe bài hát đó bao nhiêu lần,…

Và còn rất nhiều tác vụ khác có liên quan tới đánh giá của người dùng trên một sản phẩm, hay nội dung mà hệ thống có thể khai thác để đưa ra profile của người dùng và qua đó xây dựng hệ tư vấn phục vụ họ. Hệ thống không yêu cầu người dùng phải tác động gì lên nội dung, mà chỉ quan sát thói quen họ hay làm gì trên website, sau một thời gian sẽ đưa ra được một profile của người dùng và qua đó xây dựng hệ thống tốt hơn để phục vụ họ. Ví dụ như Yifan Hu và cộng sự [17] đưa ra một mô hình sử dụng ma trận hệ số cho hệ tư vấn chương trình truyền hình (TV shows), sử dụng thông tin đánh giá ẩn là số lần xem và tỉ lệ xem hết chương trình đó của người dùng. [18], [19], [20] đều nghiên cứu về việc xây dựng hệ tư vấn sách điện tử dựa trên đánh giá ấn từ hành vi đọc sách online của người dùng, đó là những thông tin như thời điểm dừng đọc và tỉ lệ trang đã đọc của sách.

Nhược điểm lớn nhất của thông tin đánh giá ẩn là thông tin của người dùng đôi lúc hoàn toàn sai lệch:

- Cô A mua vài món đồ trên Amazon không hẳn là cô ta cần hoặc thích nó. Mà có thể mua hộ ai đó hoặc mua làm quà
- Anh B mở một list bài hát lặp đi lặp lại nhưng lại ngủ quên và không nghe nó.
- Khó để biết được liệu bộ phim A được xem nhiều lần, thì sẽ được yêu thích hơn bộ phim B chỉ được xem một lần, hay đơn giản là bộ phim B nhiều tập hơn.

Thông tin đánh giá ẩn ưu điểm là dễ thu thập và thu thập được đa dạng thông tin, nhưng để sử dụng nó thì cần phải nghiên cứu và xem xét cẩn thận, thì mới xây dựng được hệ tư vấn hiệu quả, phù hợp với mục đích của website.

## 2.2 Sử dụng mô hình chủ đề ẩn LDA trên dữ liệu vết duyệt web

Dữ liệu vết duyệt web là một kiểu thông tin đánh giá ẩn, không thể hiện rõ ràng người dùng thích hay không thích nội dung url đã đọc, tuy nhiên với một lịch sử truy cập đủ lâu, vết duyệt web sẽ định hình được sở thích của người đọc, đặc biệt là với những website theo hướng tạp chí. Mô hình chủ đề ẩn LDA có thể xác định được độ phù hợp giữa một url với sở thích đọc của người dùng, để giả định đánh giá của người dùng cho url đó.

### 2.2.1 Khái quát mô hình chủ đề ẩn LDA

Mô hình chủ đề ẩn là mô hình xác suất phân phối các chủ đề ẩn trên mỗi tài liệu. Chúng được xây dựng dựa trên ý tưởng rằng mỗi tài liệu có một xác suất phân phối vào các chủ đề, và mỗi chủ đề là sự phân phối kết hợp giữa các từ khóa. Hay nói cách khác, ý tưởng cơ bản là dựa trên việc coi tài liệu là sự pha trộn của các chủ đề. Biểu diễn các từ và tài liệu dưới dạng phân phối xác suất có lợi ích rất lớn so với không gian vector thông thường.

Ý tưởng của các mô hình chủ đề ẩn là xây dựng những tài liệu mới dựa theo phân phối xác suất. Trước hết, để tạo ra một tài liệu mới, cần chọn ra một phân phối những chủ đề cho tài liệu đó, điều này có nghĩa tài liệu được tạo nên từ những chủ đề khác nhau, với những phân phối khác nhau. Tiếp đó, để sinh các từ cho tài liệu ta có thể lựa chọn ngẫu nhiên các từ dựa vào phân phối xác suất của các từ trên các chủ đề. Một cách hoàn toàn ngược lại, cho một tập các tài liệu, có thể xác định một tập các chủ đề ẩn cho mỗi tài liệu và phân phối xác suất của các từ trên từng chủ đề.

Sử dụng mô hình chủ đề ẩn để biết được xác suất các chủ đề ẩn trong nội dung văn bản đang xét. Xác suất đó được biểu diễn theo vectơ thể hiện sự phân bố nội dung của văn bản trên các chủ đề theo xác suất. Từ đó, sử dụng vecto này làm đặc trưng nội dung để so sánh sự tương đồng giữa hai văn bản.

Hai phân tích chủ đề sử dụng mô hình ẩn là Probabilistic Latent Semantic Analysis (pLSA) và Latent Dirichlet Allocation (LDA):

- pLSA là một kỹ thuật thống kê nhằm phân tích những dữ liệu xuất hiện đồng thời [21]. Phương pháp này được phát triển dựa trên LSA [1], mặc dù pLSA là một bước quan trọng trong việc mô hình hóa dữ liệu văn bản, tuy nhiên nó vẫn còn chưa hoàn thiện ở chỗ chưa xây dựng được một mô hình xác suất tốt ở mức độ tài liệu. Điều đó dẫn đến vấn đề gặp phải khi phân phối xác suất cho một tài liệu nằm ngoài tập dữ liệu học, ngoài ra số lượng các tham số có thể tăng lên một cách tuyến tính khi kích thước của tập dữ liệu tăng.

- LDA là một mô hình sinh xác suất cho tập dữ liệu rời rạc dựa trên phân phối Dirichlet, xây dựng dựa trên ý tưởng mỗi tài liệu là sự trộn lẫn của nhiều chủ đề (topic), được David M. Blei và cộng sự phát triển vào năm 2003 [1], và được nhiều nghiên cứu ứng dụng sau đó như [22]. Nhiều nghiên cứu kết hợp lọc cộng tác trên mô hình chủ đề ẩn cũng được đề xuất như hệ tư vấn bài báo, tài liệu khoa học của Chong Wang và David M. Blei [23], hệ tư vấn địa điểm du lịch của Zhiqiang He và cộng sự [24].

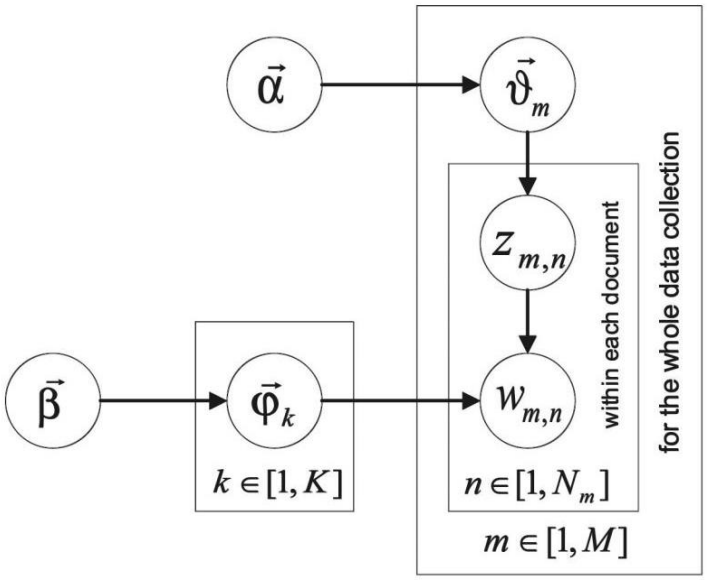
LDA là một mô hình hoàn thiện hơn so với pLSA, do đó, chúng tôi chọn loại mô hình chủ đề ẩn này để sử dụng trong việc mô hình hóa profile của người dùng (vết duyệt web) cho hệ tư vấn trong luận văn. Về bản chất, LDA là một mô hình Bayes phân cấp ba mức (mức kho ngữ liệu, mức tài liệu và mức từ ngữ). Mỗi tài liệu trong tập hợp được coi là một hỗn hợp xác định trên tập cơ bản các chủ đề. Mỗi chủ đề là một hỗn hợp không xác định trên tập cơ bản các xác suất chủ đề. Về khía cạnh mô hình hóa văn bản, các xác suất chủ đề là một biểu diễn cụ thể, rõ ràng cho một tài liệu. Dưới đây là những nét cơ bản về mô hình sinh trong LDA.

**Mô hình sinh trong LDA**

Cho trước tập M tài liệu $D = \{d_1, d_2 \ldots d_M\}$, trong đó tài liệu thứ m gồm $N_m$ từ, từ $w_i$ được rút ra từ tập các thuật ngữ $\{t_1, t_2 \ldots t_V\}$, V là số các thuật ngữ.

Quá trình sinh trong mô hình LDA diễn ra như sau:

- Mô hình LDA sinh các từ $w_{m,n}$ có thể quan sát, các từ này được phân chia về các tài liệu.
- Với mỗi tài liệu, một tỉ lệ chủ đề $\vec{\vartheta}_m$ được chọn từ phân bố Dirichlet *(Dir($\vec{\alpha}$))*, từ đó, xác định các từ thuộc chủ đề cụ thể.
- Sau đó, với mỗi từ thuộc tài liệu, chủ đề của từ đó được xác định là một chủ đề cụ thể bằng cách lấy mẫu từ phân bố đa thức *(Mult($\vec{\vartheta}_m$))*.
- Cuối cùng, từ phân bố đa thức (*Mult($\vec{\varphi}z_{m,n}$)*), một từ cụ thể $w_{m,n}$ được sinh ra dựa trên chủ đề đã được xác định. Các chủ đề $\vec{\varphi}z_{m,n}$ được lấy mẫu một lần trong toàn kho ngữ liệu.




Hình 2.1 Mô hình biểu diễn của LDA **[22]**

Các khối vuông trong hình trên biểu diễn các quá trình lặp.

Các tham số đầu vào bao gồm:

- $\alpha$ và $\beta$ : tham số mức tập hợp kho ngữ liệu
- $\vec{\vartheta}_m$: phân bố chủ đề trên tài liệu m (tham số mức tài liệu)
- Và $\Theta = \{\vec{\vartheta}_m\}_{m=1}^{M}$: ma trận M x K
- $z_{m,n}$: chỉ số chủ đề của từ thứ n trong tài liệu m (biến mức từ ngữ)
- $\vec{\varphi} z_{m,n}$: phân bố thuật ngữ trên chủ đề cụ thể $z_{m,n}$
- Và $\Phi = \{\vec{\varphi}_k\}_{k=1}^{K}$: ma trận K x V
- $w_{m,n}$: từ thứ n của văn bản n (biến mức từ ngữ)
- $M$: số lượng các tài liệu
- $N_m$: số lượng từ trong tài liệu m (độ dài của văn bản sau khi đã loại bỏ stop word – từ không mang nhiều ý nghĩa)
- $K$: số lượng các chủ đề ẩn
- $DirvàMult$: phân bố Dirichlet và phân bố đa thức

Vì $w_{m,n}$ phụ thuộc điều kiện vào phân bố $\vec{\varphi}_k$ $và z_{m,n}$ phụ thuộc vào phân bố $\vec{\vartheta}_m$, xác suất để một chỉ mục chủ đề $w_{m,n}$ là một từ $t$ nằm trong phân bố chủ đề trên tài liệu $\vec{\vartheta}_m$ và phân bố từ trên chủ đề (Φ) là:

$$p(w_{m,n=t}|\vec{\vartheta}_m, \Phi) = \sum p(w_{m,n} = t|\vec{\varphi}_k)\, p(z_{m,n} = k|\vec{\vartheta}_m)$$

Với xác suất của mỗi thuật ngữ, ta có thể xác định được xác suất chung của tất cả các biến đã biết và biến ẩn với các tham số Dirichlet cho trước:

$$p(\vec{d}_m, \vec{z}_m, \vec{\vartheta}_m, \Phi|\vec{\alpha}, \vec{\beta}) = p(\Phi|\vec{\beta}) \prod_{n=1}^{N\_m} p(w_{m,n}|\vec{\varphi} z_{m,n}) p(z_{m,n}|\vec{\vartheta}_m)\, p(\vec{\vartheta}_m|\vec{\alpha})$$

Tính tích phân trên $\vec{\vartheta}_m$, Φ và tổng trên $\vec{z}_m$, ta xác định được xác suất của tài liệu $\vec{d}_m$. Khi đã có xác suất của mỗi tài liệu $p(\vec{d}_m|\vec{\alpha}, \vec{\beta})$, xác suất của cả kho ngữ liệu D = {$d_1$, $d_2$, …, $d_M$} là tích của tất cả các xác suất của tất cả các tài liệu nằm trong đó:

$$p(D|\vec{\alpha}, \vec{\beta}) = \prod_{n=1}^{M} p(\vec{d}_m|\vec{\alpha}, \vec{\beta})$$

### 2.2.2 Mô hình LDA trong việc ước lượng hạng giả định cho mô hình tư vấn

Với mô hình chủ đề ẩn LDA, cho trước một tập các văn bản, LDA tìm xem topic model nào đã sinh ra tập các văn bản trên. Bao gồm:

- Tìm phân phối xác suất trên tập từ đối với mỗi topic
- Tìm phân phối topic của mỗi tài liệu

Mô hình sử dụng phân phối chủ đề của mỗi tài liệu được tìm ra bởi LDA để làm đặc trưng nội dung cho việc xây dựng vector đặc trưng người dùng và vector đặc trưng cho url đã đọc.

Giả sử chúng ta xác định được K chủ đề ẩn từ tập dữ liệu học. Với mỗi tài liệu *t*, mô hình tính các xác suất để tài liệu t thuộc vào topic *i* là *pt(i)*, với *i=1,…,k.* Từ đó xác định được vector đặc trưng nội dung từ mô hình chủ đề ẩn LDA là :

$$\vec{t} = (pt_1, pt_2, ..., pt_k)$$

Vector đặc trưng của người dùng chính là phân phối xác suất các chủ đề ẩn trên nội dung của tất cả các url mà người dùng đã đọc. Tương tự, vector đặc trưng cho url là phân phối xác suất của các chủ đề ẩn trên nội dung của url. Từ đó vết duyệt web của mỗi người dùng sẽ được biểu diễn dưới dạng tập các vector đặc trưng trong không gian chủ đề ẩn:

$$p_i = \{ \vec{u}_i , \vec{i}_1 , \vec{i}_2 , …., \vec{i}_k \}$$

Trong đó: $\vec{u}_i$ là vector đặc trưng cho người dùng $u_i$, $\vec{i}_j$ là vector đặc trưng của url $i_j$ đã đọc.

Để ước lượng hạng giả định cho từng cập người dùng – url, ta so sánh sự giống nhau của hai phân phối xác suất của chúng. Có thể sử dụng khoảng cách *cosine (1),* một độ đo cơ bản trong không gian vector, hay độ đo *Jensen–Shannon (2)* là thước đo độ tương đồng trong không gian phân phối xác suất. Các độ đo tương đồng này có giá trị từ 0 đến 1, với ý nghĩa giá trị càng lớn thì độ tương đồng giữa hai vector càng lớn.

$$(1): \cos(\vec{A}, \vec{B}) = \frac{A.B}{|A||B|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

$$(2): JSD_{\pi_1,\pi_2}(P_1, P_2) = H(\pi_1 P_1 + \pi_2 P_2) - (\pi_1 H(P_1) + \pi_2 H(P_2))$$

Trong đó $\pi_1, \pi_2$ là trọng số được lựa chọn cho phân phối xác suất $P_1, P_2$ và $H(P)$ là Shannon entropy của phân phối P: $H = -\sum_i p_i \log_b p_i$ ($p_i$ là các giá trị xác suất của phân phối xác suất P). Thông thường chọn $\pi_1 = \pi_2 = \frac{1}{2}$ và b = 2.

## 2.3 Bài toán tư vấn nội dung trên một website dựa trên vết duyệt web của người dùng

Với các trang web Việt Nam hiện nay, nhìn chung hệ thống gợi ý còn khá đơn giản, chưa hướng tới cá nhân người dùng, và chưa đạt hiệu quả cao. Các nội dung gợi ý thường là tin mới, tin được nhiều người xem, tin liên quan tới bài viết đang đọc. Và như vậy, với bất kì người dùng nào họ cũng đều nhận được danh sách gợi ý như nhau, hơn nữa còn bị trùng lặp các nội dung đã đọc, và các nội dung gợi ý chưa chắc đã phù hợp với sở thích cá nhân người xem. Do vậy, nhu cầu xây dựng một hệ tư vấn cá nhân hóa cho các website Việt Nam là rất cần thiết, và hiện nay đang là một vấn đề được nhiều người quan tâm.

Vì những lí do trên, luận văn sẽ hướng đến việc xây dựng mô hình hệ tư vấn cá nhân hóa trên các website ở Việt nam dựa trên vết duyệt web của người dùng trên từng website đó. Phương pháp được sử dụng là tư vấn dựa trên cộng tác kết hợp giả định hạng của các item (url của website) thông qua mô hình chủ đề ẩn LDA.

### 2.3.1 Phát biểu bài toán

Với một website, ta sẽ có tập W gồm tất cả nội dung của các url trên website, và tập U gồm tất cả người dùng website đó.

$W = \{c_1, c_2, \ldots, c_n\}$

$U = \{u_1, u_2, \ldots, u_m\}$

Tập dữ liệu vết duyệt web của người dùng, $P = \{p_1, p_{2,} \ldots . p_m\}$

Trong đó: $p_i$ là vết duyệt web của người dùng $u_i$ trên trang web đang xét

$p_i = \{(t_1, i_1), (t_2, i_2), \ldots (t_k, i_k)\}$, $k <= n$; $t_k$ là thời gian truy cập url $i_k$ của người dùng $u_i$

- Input: người dùng $u_i$, tập các url trên trang web mà người dùng $u_i$ chưa đọc
- Output: các url phù hợp với người dùng $u_i$

### 2.3.2 Hướng giải quyết

Phương pháp tư vấn cộng tác là phương pháp phổ biến được nhiều hệ tư vấn sử dụng. Bản chất của phương pháp này chính là hình thức tư vấn truyền miệng tự động. Trong phương pháp này, hệ thống sẽ so sánh, tính toán độ tương tự nhau giữa những người dùng hay sản phẩm, từ đó người dùng sẽ được tư vấn những thông tin, sản phẩm được ưa chuộng nhất bởi những người dùng có cùng thị hiếu. Các hệ tư vấn này có khả năng tư vấn phong phú trên toàn bộ sản phẩm. Do vậy, luận văn đề xuất sử dụng phương pháp cộng tác cho mô hình tư vấn. Với mục tiêu nhằm vào các website tiếng việt mang khuynh hướng tạp chí (các website với nội dung theo từng chuyên mục, lĩnh vực, nội dung ít bị lỗi thời ví dụ như các tạp chí làm đẹp, phụ nữ, xe cộ,…), để thu thập được các đánh giá cụ thể like hay dislike, đánh giá theo điểm là việc khó có thể thực hiện được, do vậy mô hình sẽ sử dụng vết duyệt web là thông tin đánh giá ẩn phục vụ cho mục đích tư vấn.

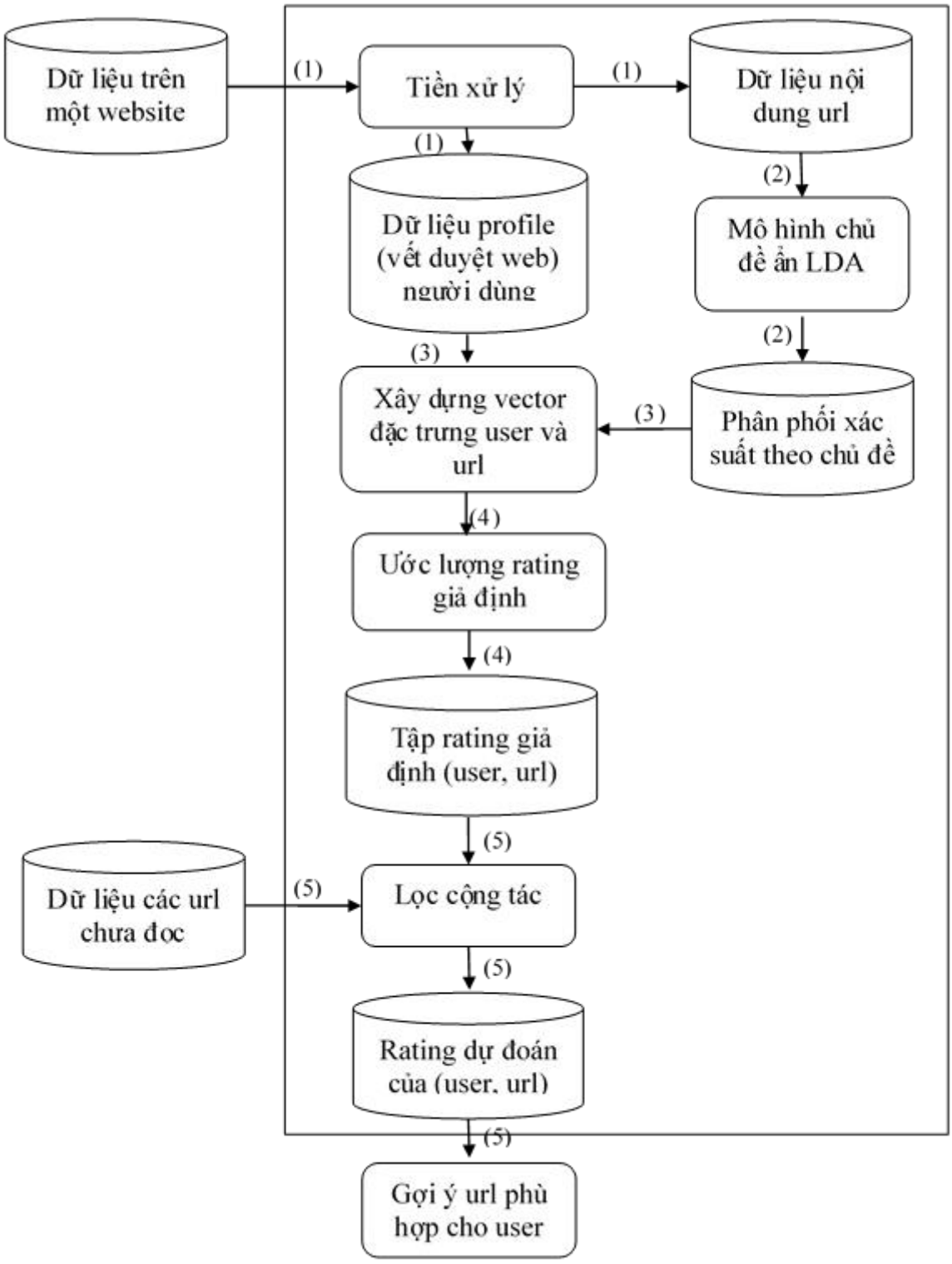
Dữ liệu vết duyệt web không thể hiện rõ ràng người dùng thích hay không thích nội dung url đã đọc, tuy nhiên với một lịch sử truy cập đủ lâu, vết duyệt web sẽ định hình được sở thích của người đọc. Khác với các website tin tức, khi mà người dùng thường đọc những tin mới có tính cập nhật, và thường ít theo một định hướng sở thích riêng, thì với những website theo hướng tạp chí, họ thường đọc những nội dung phù hợp với sở thích của mình hơn. Do vậy, việc kết hợp giữa phương pháp lọc cộng tác và một phương pháp dựa trên nội dung sẽ giúp hệ tư vấn trên website này vừa đảm bảo sự đa dạng và tính cập nhật của nội dung tư vấn, mà còn phù hợp với sở thích riêng của người đọc.

Với phương pháp tư vấn cộng tác truyền thống, hệ tư vấn thường cần có dữ liệu đánh giá của người dùng đối với các item. Trong khi đó, dữ liệu vết duyệt web không thể hiện điều đó. Vì vậy, luận văn đề xuất xây dựng dữ liệu hạng giả định dựa trên độ phù hợp của url với sở thích của người đọc bằng mô hình chủ đề ẩn LDA. Nghĩa là url nào càng gần về nội dung với lịch sử đọc của người dùng thì giả định hạng càng cao.

Tóm lại, hệ tư vấn nội dung trên một website sử dụng mô hình chủ đề ẩn LDA để xây dựng tập dữ liệu hạng giả định của người dùng cho mỗi url đã đọc, sau đó sử dụng phương pháp cộng tác để dự đoán các url chưa đọc phù hợp với người dùng.

# Chương 3 Mô hình hệ tư vấn nội dung trên website dựa trên dữ liệu vết duyệt web

## 3.1 Sơ đồ mô hình tư vấn




Hình 3.1 Mô hình hệ tư vấn nội dung website

Sơ đồ mô hình hệ tư vấn nội dung trên một website dựa trên dữ liệu vết duyệt web của người dùng được mô tả trong hình 3.1. Mô hình là sự kết hợp giữa mô hình lọc cộng tác truyền thống với việc đưa thêm mô hình LDA vào để tính toán hạng giả định cho mô hình cộng tác.

Mô hình tư vấn bao gồm các bước xử lý chính:

❖ **Bước 1: Tiền xử lý dữ liệu**

Tiền xử lý là bước xử lý dữ liệu trên tập dữ liệu ban đầu để trích xuất ra được vết duyệt web của người dùng và dữ liệu mô tả nội dung các url, bao gồm 2 nhiệm vụ chính:

- Đưa ra tập profile người dùng: lọc ra tập dữ liệu vết duyệt web của người dùng, bao gồm các vết duyệt web có độ dài lịch sử truy cập lớn hơn 5 (ít nhất 5 url đã được đọc trước đó)
- Trích chọn ra được tập từ điển gồm các token phân biệt trên toàn bộ dữ liệu nội dung url, và tập dữ liệu corpus là tần suất của các tokens trong mỗi url, bao gồm các công việc: tách từ tiếng việt, chuẩn hóa từ, loại bỏ các từ ít mang thông tin ngữ nghĩa (stop word).

❖ **Bước 2: Mô hình hóa chủ đề cho nội dung website**

Sử dụng mô hình chủ đề ẩn LDA để mô hình hóa nội dung các url trên website thành phân phối xác suất theo chủ đề của các từ. Mỗi chủ đề sẽ bao gồm tập các từ xây dựng nên chủ đề cùng với xác suất của từ khóa đó:

$$[Topic\ 1: (token_{11}, p_{11}), (token_{21}, p_{21}), \ldots, (token_{i1}, p_{i1});$$

$$Topic\ 2: (token_{12}, p_{12}), (token_{22}, p_{22}), \ldots, (token_{i2}, p_{i2}); \ldots;$$

$$Topic\ N: (token_{1n}, p_{1n}), (token_{2n}, p_{2n}), \ldots, (token_{in}, p_{in})]$$

Mô hình xây dựng phân phối xác suất trên tập 50 chủ đề, N = 50

❖ **Bước 3: Xây dựng vector đặc trưng user và url**

Xây dựng vector đặc trưng người dùng và đặc trưng của url từ phân phối xác suất theo chủ đề ở bước 2 (chi tiết ở mục 3.2.1)

❖ **Bước 4: Ước lượng hạng giả định**

So sánh vector đặc trưng của người dùng và vector đặc trưng url đã đọc để ước lượng hạng giả định, thu được ma trận hạng giả định (chi tiết ở mục 3.2.2)

❖ **Bước 5: Sử dụng mô hình tư vấn cộng tác gợi ý url cho người dùng**

Với tập url chưa đọc, hệ thống sử dụng mô hình tư vấn cộng tác với hạng giả định để tư vấn các url chưa đọc cho người dùng.

Ma trận đánh giá hạng giả định sẽ được đưa vào huấn luyện cho mô hình cộng tác, giống như với hạng thật mà người dùng đánh giá. Mô hình tính toán độ tương tự giữa những người dùng và giữa các url (theo mục 1.3.2 Kĩ thuật tư vấn cộng tác), đưa ra dự đoán hạng của người dùng cho các url chưa đọc (các vị trí còn thiếu trong ma trận hạng), từ đó gợi ý những url có nội dung phù hợp nhất (có hạng cao nhất) với người đọc.

## 3.2 Phương pháp ước lượng hạng giả định bằng mô hình chủ đề ẩn LDA

Như đã trình bày ở nội dung 1.3.2 Kĩ thuật tư vấn cộng tác, hệ thống tư vấn cộng tác truyền thống sẽ dự đoán hạng của một sản phẩm dựa trên các đánh giá trước đó bởi người dùng. Với dữ liệu vết duyệt web của hệ thống không có thông tin đánh giá của người dùng, hệ thống sẽ sử dụng thông tin đánh giá ẩn là hạng giả định được tính toán bởi mô hình chủ đề ẩn LDA.

### 3.2.1 Xây dựng vector đặc trưng người dùng và vector đặc trưng của url

Sau khi mô hình LDA huấn luyện tập dữ liệu học là nội dung của tất cả các url có trên website, mô hình sẽ cho ra một phân phối xác suất trên K chủ đề của các từ (K là tham số của mô hình). Với mỗi tài liệu *t*, mô hình tính các xác suất để tài liệu t thuộc vào topic *i* là *pt(i)*, với *i=1,...,k.*

Từ đó xác định được vector đặc trưng nội dung từ mô hình chủ đề ẩn LDA là :

$$\vec{t} = (pt_1, pt_2, ..., pt_k)$$

Dựa trên khái quát bài toán như mục 1.4.1, ta biểu diễn:

$w_i = \{c_{i1}, c_{i2}, \ldots, c_{ik}\}$ là tập nội dung các url mà người dùng $u_i$ đã đọc.

Để tính vector đặc trưng cho người dùng $u_i$, tài liệu *t* sẽ được xem là tổng nội dung của các url mà người dùng $u_i$ đã đọc, $t = c_{i1} + c_{i2} + \ldots + c_{ik}$, ta tính được vector $\vec{u_i}$ từ mô hình LDA

Tương tự, để tính vector đặc trưng cho url $i_k$, tài liệu *t* sẽ được xem là nội dung của url $i_k$, ta sẽ được vector $\vec{i_k}$

Từ đó, vết duyệt web của người dùng $u_i$ được biểu diễn bằng tập vector đặc trưng trong không gian xác suất của LDA:

$$p_i = \{ \vec{u_i}, \vec{i_1}, \vec{i_2}, \ldots., \vec{i_k} \}$$

### 3.2.2 Xây dựng ma trận hạng giả định

Sau khi xây dựng được vết duyệt web của người dùng bao gồm tập vector đặc trưng của người dùng và các vector url đã đọc trên không gian phân phối xác suất chủ đề ẩn, mô hình sẽ tính độ tương tự giữa hai vector đặc trưng bằng khoảng cách Jensen-Shannon, giá trị từ 1 đến 0, với 1 nghĩa là hai phân phối bằng nhau. Sau đó hạng giả định được tính bằng công thức: $r(u_i, i_{k'}) = JSD(\vec{u_i}, \vec{i_{k'}}) / \max(JSD_i)$, với k'=1,2,…,k. $\max(JSD_i)$ là giá trị Jensen-Shannon lớn nhất của người dùng $u_i$. Nghĩa là giá trị JSD càng lớn thì hạng càng cao.

Ma trận hạng giả định là một ma trận thưa, với mỗi hàng là một người dùng, mỗi cột là một url trong toàn bộ người dùng, và url có trên website. Ma trận được mô tả dưới đây, tại đó dấu × biểu thị cho hạng giả định mà người dùng $u_i$ đánh giá url $i_k$, kí hiệu ∘ là người dùng chưa đọc url đó.

| $user/url$ | $i_1$ | $i_2$ | $i_3$ | $i_4$ | ... | $i_n$ |
|---|---|---|---|---|---|---|
| $u_1$ | × | ∘ | ∘ | × | . | × |
| $u_2$ | × | ∘ | × | ∘ | . | ∘ |
| $u_3$ | ∘ | × | × | ∘ | . | × |
| ... | . | . | . | . | . | . |
| $u_m$ | × | × | ∘ | ∘ | . | ∘ |

## 3.3 Phương pháp ước lượng hạng giả định bằng tần suất từ

Để so sánh với phương pháp ước lượng hạng bằng LDA được đề xuất cho hệ tư vấn website, chúng tôi đưa ra một phương pháp cơ sở khác để so sánh hiệu quả, đó là phương pháp ước lượng hạng giả định bằng tần suất từ

Với phương pháp này, mô hình tính tần suất xuất hiện của các từ riêng biệt trong tất cả các url người dùng đã đọc, và trong url cần đánh giá hạng.

Vết duyệt web của người dùng *u* được định nghĩa là: $Profile(u) = (w_{1u}, \ldots, w_{mu})$ với $w_{iu}$ biểu thị tần suất của từ khóa *i* trong dữ liệu duyệt web của *u*.

Dữ liệu nội dung url cần ước lượng hạng là: $Content(s) = (w_{is}, \ldots, w_{ns})$ với $w_{is}$ biểu thị tần suất của từ khóa *i* trong nội dung url *s*.

Độ phù hợp giữa url *s* với người dùng *u* được tính bằng điểm p(u,s):

$$p(u,s) = \sum (w_{iu} * w_{is}) \text{ với mỗi từ khóa } i \text{ (token i) trong url } s$$

Giá trị p(u,s) thể hiện ý nghĩa là những từ xuất hiện nhiều trong url *s* mà cũng xuất hiện nhiều trong profile(u), thì đồng nghĩa với việc url *s* có độ tương đồng cao với profile(u). Hạng giả định được quy đổi sang thang điểm 0-1 theo công thức:

$$r(u,s) = p(u,s) / p_{max}$$

với $p_{max}$ là giá trị điểm cao nhất trong tập các $p(u, s_k)$ của người dùng u.

## 3.4 Đánh giá kết quả tư vấn

Việc đánh giá chất lượng của các tư vấn trả về bởi hệ thống là một bài toán khó, vì không có một độ đo ngữ nghĩa đánh giá chính xác được sự phù hợp giữa người dùng và các tư vấn hệ thống trả lại. Chất lượng của hệ tư vấn thường được xem xét tương quan giữa các mục tiêu đạt được. [3] đề cập đến một số tiêu chí đánh giá hệ tư vấn trên thực tế:

Mức độ phù hợp (relevance): mục đích chính trước tiên của một hệ tư vấn là đưa ra danh sách các mục tư vấn có liên quan tới người dùng. Bên cạnh đó, các yếu tố tiếp theo dưới đây không phải là tiêu chí quan trọng nhất, nhưng nó đủ để ảnh hưởng tới một hệ tư vấn.

Tính mới (novelty): hệ tư vấn là thực sự hữu ích khi mục đề nghị là một cái gì đó mà người dùng đã không nhìn thấy trong quá khứ. Ví dụ, bộ phim nổi tiếng của một thể loại ưa thích sẽ hiếm khi có tính mới cho người dùng. Lặp đi lặp lại đề xuất của mặt hàng phổ biến cũng có thể dẫn đến giảm tính đa dạng trong bán hàng.

Tính bất ngờ (serendipity): các mặt hàng được tư vấn là bất ngờ, trái ngược với các khuyến nghị rõ ràng. Serendipity khác với novelty trong các khuyến nghị thực sự gây ngạc nhiên cho người sử dụng, chứ không phải chỉ đơn giản là một cái gì đó họ không biết trước. Nó thường là trường hợp một người dùng thường tiêu thụ các mặt hàng cụ thể, mặc dù mối quan tâm tiềm ẩn trong các mặt hàng khác có thể tồn tại nhưng người dùng phải tự tìm ra. Các sản phẩm bất ngờ có xu hướng không liên quan trực tiếp đến các mặt hàng họ từng mua. Tăng tính bất ngờ cho hệ tư vấn có thể xây dựng được lợi ích lâu dài cho hệ thống, khi chuyển hướng tập trung của khách hàng sang một mặt hàng mới, nhưng cần cân đối trong lợi ích chung của hệ tư vấn

Mức độ đa dạng (diversity): các hệ thống tư vấn thường đề nghị một danh sách các mục top-k. Khi tất cả các mặt hàng được đề nghị là rất giống nhau, nó làm tăng nguy cơ mà người sử dụng có thể không thích bất kỳ sản phẩm của loại mặt hàng này. Mặt khác, khi danh sách đề nghị chứa các sản phẩm của các chủng loại khác nhau, có một cơ hội lớn hơn mà người dùng có thể muốn ít nhất một trong các mặt hàng này. Sự đa dạng có lợi ích đảm bảo rằng người dùng không cảm thấy nhàm chán bởi đề nghị lặp đi lặp lại của các mặt hàng tương tự.

Mỗi một tiêu chí đều đóng một vai trò nhất định trong hệ tư vấn, cần cân bằng các tiêu chí để hệ tư vấn phù hợp với xu hướng khách hàng tiềm năng và phù hợp với sản phẩm tư vấn hướng đến.

Trong nội dung khuôn khổ thực nghiệm mô hình, chúng tôi không đưa ra các đánh giá về mặt thực tế, thay vào đó để đánh giá khả năng đúng đắn của mô hình, chúng tôi sử dụng thước đo căn bậc hai trung bình bình phương các sai số RMSE (root mean square error – độ lệch chuẩn) và sai số trung bình (mean absolute error) để so sánh độ lệch giữa dự đoán hạng của mô hình với hạng giả định.

Công thức tính sai số RMSE và MAE như sau:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Trong đó: n là số lượng hạng giả định (bằng số lượng hạng dự đoán của mô hình)

$y_i, \hat{y}_i$ lần lượt tương ứng là giá trị hạng từ phương pháp ước lượng hạng giả định, và giá trị hạng dự đoán từ mô hình.

# Chương 4 Thực nghiệm và đánh giá

## 4.1 Môi trường thực nghiệm

### 4.1.1 Cấu hình phần cứng

Thực nghiệm được tiến hành trên máy tính có thông số như bảng dưới đây.

Bảng 4.1 Bảng thông số cấu hình phần cứng

| Thành phần | Chỉ số |
|---|---|
| **CPU** | Intel Core i7-6700HQ 2.6Ghz |
| **RAM** | 16GB |
| **HDD** | 500GB |
| **OS** | Ubuntu 14.04 (64bit) |

### 4.1.2 Công cụ phần mềm

Trong quá trình thực nghiệm, chúng tôi sử dụng một số công cụ mã nguồn mở được liệt kê trong bảng dưới đây:

Bảng 4.2 Danh sách công cụ sử dụng trong thực nghiệm

| STT | Tên công cụ | Tác giả | Nguồn |
|---|---|---|---|
| 1 | Mã nguồn mở thuật toán CF: python-recsys | Ocelma | https://github.com/ocelma/python-recsys |
| 2 | Gensim – thư viện topic modelling | Radim Řehůřek | https://radimrehurek.com/gensim/ |
| 3 | vnTokenizer | Lê Hồng Phương | http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer |

Thực nghiệm được chúng tôi xây dựng bằng ngôn ngữ Python, có sử dụng một số API của các công cụ hỗ trợ trên để mô hình hóa chủ đề ẩn, huấn luyện mô hình cộng tác và đưa ra gợi ý cho người dùng.

**Thư viện Gensim – topic modeling:**

Đây là công cụ mã nguồn mở được cài đặt trên ngôn ngữ Python, mô hình hóa ngôn ngữ thành các không gian vector. Gensim cài đặt một số mô hình như TF-IDF, mô hình deep learning, Latent semantic analysis (LSA) và Latent Dirichlet Allocation (LDA),…

Trong thực nghiệm trên, chúng tôi sử dụng mô hình chủ đề ẩn LDA để mô hình hóa profile của người dùng và nội dung của url thành vector user và vector item.

Ví dụ một số API được sử dụng:

```
# build mô hình LDA
lda_model = gensim.models.ldamodel.LdaModel(corpus=mm, id2word=dictionary, num_topics=NUM_TOPICS, minimum_probability=0.0)
# liệt kê danh sách các token trong dữ liệu contents đầu vào
user_bow= dictionary.doc2bow(user_contents.split())
# chuyển đổi thành vector trong không gian LDA
user_vec = lda_model[user_bow]
```

**Thư viện python-recsys:**

Là thư viện Python cài đặt thuật toán SVD (Singular Value Decompostion - thuật toán nhằm mục đích giảm số chiều cho mô hình CF). Thư viện hỗ trợ:

- tính độ tương tự giữa hai sản phẩm
- gợi ý những sản phẩm giống với sản phẩm cho trước
- dự đoán hạng mà một người dùng có thể đánh giá cho sản phẩm
- gợi ý các sản phẩm phù hợp với người dùng
- đưa ra những người dùng thích với một sản phẩm cho trước

Trong đó, chúng tôi đã sử dụng tính năng dự đoán hạng, và gợi ý sản phẩm cho người dùng. Ví dụ một số API được sử dụng:

```
# tính mô hình SVD
    svd = SVD()
    svd.compute(k=k,min_values=1, pre_normalize=None, mean_center=True, post_normalize=True, savefile='../Data/datamodel')
# dự đoán hạng
```

```
pred_rating = svd.predict(item_id, user_id)
```

# gợi ý các items cho user_id

```
recommend_list = svd.recommend(int(user_id),
n=10, is_row=False)
```

**Công cụ vnTokenizer:**

Là một công cụ tách từ tự động cho văn bản tiếng Việt (mã hóa bằng bảng mã Unicode UTF-8). Công cụ chạy dưới dạng dòng lệnh:

```
vnTokenizer.sh -i <tệp-input> -o <tệp-output> [<các-tùy-chọn>]
```

Thực nghiệm xây dựng trên dữ liệu website tiếng việt, nên cần sử dụng công cụ để tách văn bản thành các từ (token), để xây dựng tập từ điển và corpus

## 4.2 Dữ liệu thực nghiệm

Dữ liệu thực nghiệm là dữ liệu thực tế trên hai trang web http://www.otoxemay.vn/ và http://www.emdep.vn/. Dữ liệu bao gồm lịch sử duyệt web của tất cả người dùng và nội dung của tất cả url trên mỗi trang web.

Bảng 4.3 Dữ liệu thực nghiệm

| Dữ liệu | otoxemay.vn | emdep.vn |
|---|---|---|
| Thời gian | 06/09/2016 – 06/10/2016 | 01/09/2016 – 01/11/2016 |
| Số lượng người dùng | 1496 | 12356 |
| Số lượng url | 3504 | 24655 |

Với mỗi trang web, dữ liệu được chia thành 2 file với nội dung và định dạng cụ thể như sau:

- File user_profiles chứa vết duyệt web của người dùng

  Định dạng: mỗi dòng trong file là vết duyệt web của một người dùng

  *user_idtimestamp1,item_id1 timestamp2,item_id2 ... timestampN,item_idN* (khoảng cách là một dấu tab \t)

  *user_id*: định danh người dùng (int)

  *timestamp1*: thời điểm đọc item_id1 (timestamp)

  *item_id1*: định danh của url (int)

- File item_contents chứa nội dung của các url

  Định dạng: mỗi dòng của file là một url

  *Item_id content* (khoảng cách là một dấu tab \t)

  *Item_id*: định danh của url (int)

  *Content:* nội dung tiếng việt của url (string-utf8)

## 4.3 Thực nghiệm

Chúng tôi xây dựng mô hình thực nghiệm trên phương pháp giả định hạng của người dùng với url bằng mô hình chủ đề ẩn LDA, đồng thời cũng xây dựng một mô hình thực nghiệm trên phương pháp giả định hạng khác để làm cơ sở so sánh hiệu quả (phương pháp sử dụng tần suất xuất hiện của các từ để tính toán sự phù hợp giữa nội dung url cần đánh giá với nội dung của các url người dùng đã đọc).

### 4.3.1 Mô hình tư vấn khi sử dụng phương pháp giả định hạng bằng LDA

Sơ đồ thực nghiệm mô hình được mô tả trong hình 4.1

Các pha trong sơ đồ như sau:

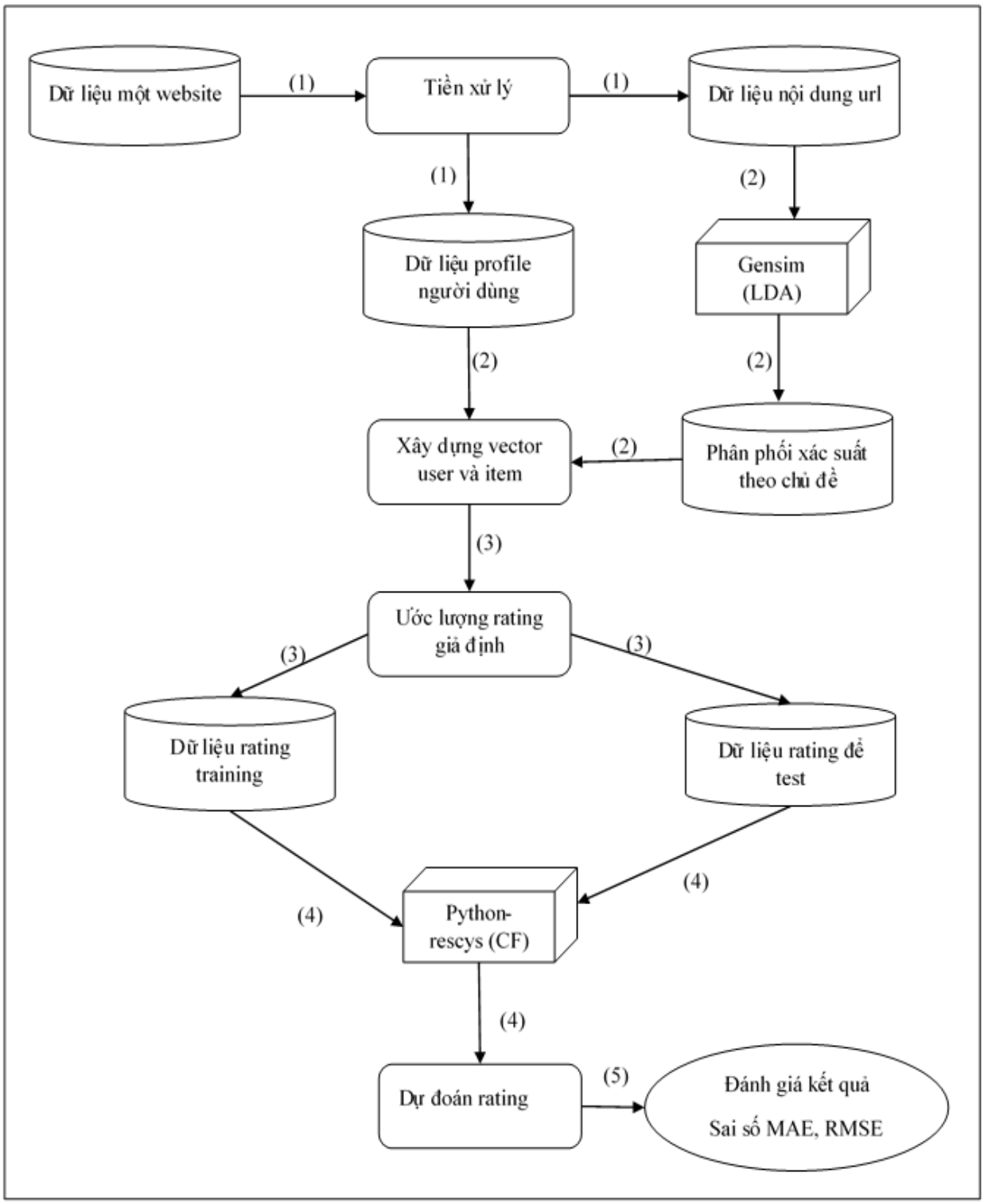
- (1): Tiền xử lý

- (2): Xây dựng vector đặc trưng cho người dùng và url đã đọc

- (3): Ước lượng hạng giả định cho từng cặp user-item

- (4): Huấn luyện mô hình cộng tác, dự đoán hạng

- (5): Đánh giá độ lệch của mô hình




Hình 4.1 Sơ đồ thực nghiệm với hạng giả định bằng LDA

**Công việc 1: Tiền xử lý**

Tiền xử lý là bước xử lý dữ liệu trên tập dữ liệu ban đầu gồm vết duyệt web của người dùng và nội dung các url, bao gồm 2 nhiệm vụ chính:

- Đưa ra tập profile người dùng: Dựa vào dữ liệu thu được của mỗi website, lọc ra tập dữ liệu vết duyệt web trên từng trang web. Mỗi vết duyệt web của người dùng đưa vào thực nghiệm là những vết duyệt web có lịch sử truy cập lớn hơn 5.
- Trích chọn ra được tập từ điển gồm các token phân biệt trên toàn bộ dữ liệu nội dung url, và tập dữ liệu corpus là tần suất của các tokens trong mỗi url.

+ Sử dụng file input item_contents, tách từ bằng công cụ vnTokenizer

+ Đưa về chữ thường, loại bỏ số, kí tự đặc biệt

+ Loại bỏ các từ có khả năng mang ít thông tin ý nghĩa bằng cách loại bỏ 10% các từ có tần suất xuất hiện cao nhất và thấp nhất.

+ Đưa ra tập từ điển gồm các token (từ) phân biệt và tập corpus là tần suất của các từ trong mỗi url

**Công việc 2: Xây dựng vector đặc trưng cho người dùng và url đã đọc**

Chúng tôi sử dụng thư viện gensim để xây dựng mô hình chủ đề ẩn LDA với tập dữ liệu từ điển và corpus đã xây dựng ở trên. Mô hình tìm sự phân phối xác suất trên 50 chủ đề. Ví dụ về đặc trưng của dữ liệu huấn luyện được minh họa trong bảng dưới đây:

Bảng 4.4 Minh họa đặc trưng dữ luyện huấn luyện trên trang web emdep.vn

| Chủ đề | Từ khóa đại diện và xác suất của từ khóa |
|---|---|
| Topic 1 | 0.008*giải_khát + 0.007*tráng_miệng + 0.006*thực_đơn + 0.005*bếp |
| Topic 2 | 0.005*đồ_hiệu + 0.004*đồng + 0.003*thẩm_mỹ + 0.003*xu_hướng |
| Topic 3 | 0.005*cồn + 0.005*phái_mạnh + 0.004*cầu_thủ + 0.004*nam_giới |
| Topic 4 | 0.006*mụn + 0.005*mặt + 0.004*khô + 0.004*lotion |

Sau đó, chúng tôi tính vector đặc trưng của người dùng với dữ liệu là nội dung của tất cả các url mà người dùng đã từng đọc, và vector đặc trưng của url là nội dung của url. Vector đặc trưng của người dùng và url trên không gian xác suất của mô hình LDA là vector 1 chiều gồm 50 giá trị xác suất phân phối trên 50 chủ đề.

**Công việc 3: Ước lượng hạng giả định**

Ở bước này, chúng tôi sẽ tính cả khoảng cách cosine và khoảng cách Jensen-Shannon (để so sánh độ chính xác) giữa hai vector user và item, để giả định hạng của người dùng user với item url, tức là độ tương đồng của hai vector càng lớn thì độ phù hợp của url với người dùng càng cao, tương đương điểm càng cao (thang điểm từ 0 đến 1). Kết quả sẽ được lưu vào file user_rating với định dạng mỗi dòng của file là cặp 3 giá trị <*user_id, item_id, rating_value*>

Dữ liệu hạng trên sẽ được chia thành 2 phần: dữ luyện huấn luyện và kiểm tra, training:testing với tỉ lệ 4:1

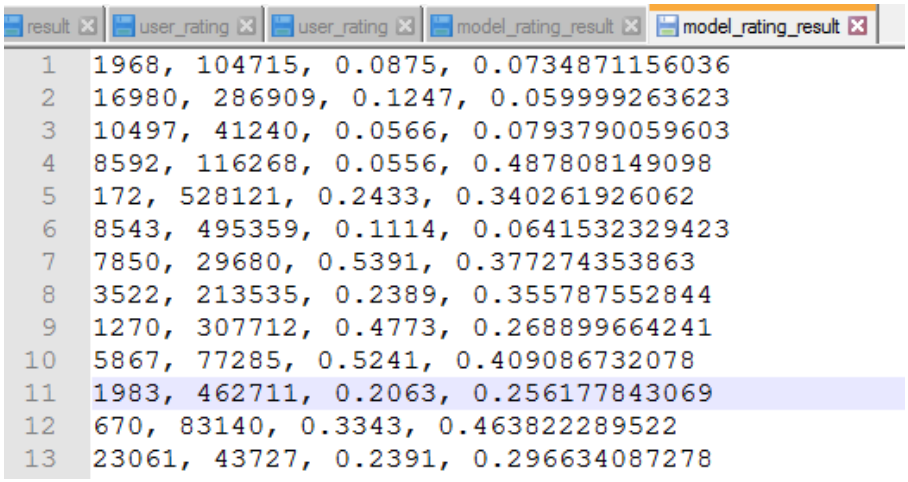
**Công việc 4: Huấn luyện mô hình cộng tác và dự đoán hạng**

Thực nghiệm sử dụng thư viện python-recsys để huấn luyện mô hình cộng tác với dữ liệu hạng giả định trong tập dữ liệu huấn luyện.

Sau đó, mô hình sẽ tính toán đưa ra dự đoán hạng của các url chưa đọc.

**Công việc 5: Đánh giá độ lệch của mô hình**

Như đã trình bày ở mục 3.4 (Đánh giá hệ tư vấn), việc đánh giá mô hình tư vấn có hiệu quả hay không phụ thuộc vào rất nhiều thước đo. Trong khuôn khổ của luận văn, để đánh giá mô hình, chúng tôi tính sai số RMSE (căn bậc hai trung bình bình phương sai số) và sai số MAE (sai số trung bình). Để tính toán độ lệch này, chúng tôi sử dụng mô hình CF trong thư viện python-recsys để dự đoán hạng cho từng cặp user-item trong dữ liệu test, và sau đó tính sai số giữa hạng dự đoán của model với hạng giả định. Hình 4.3 mô tả kết quả dự đoán hạng của mô hình với định dạng *<user_id, item_id, rating_test, rating_model>*




Hình 4.2 Ví dụ về kết quả dự đoán hạng

### 4.3.2 Mô hình tư vấn khi sử dụng phương pháp giả định hạng bằng tần suất của từ

Sơ đồ thực nghiệm được mô tả trong hình 4.3. Các pha trong sơ đồ như sau:

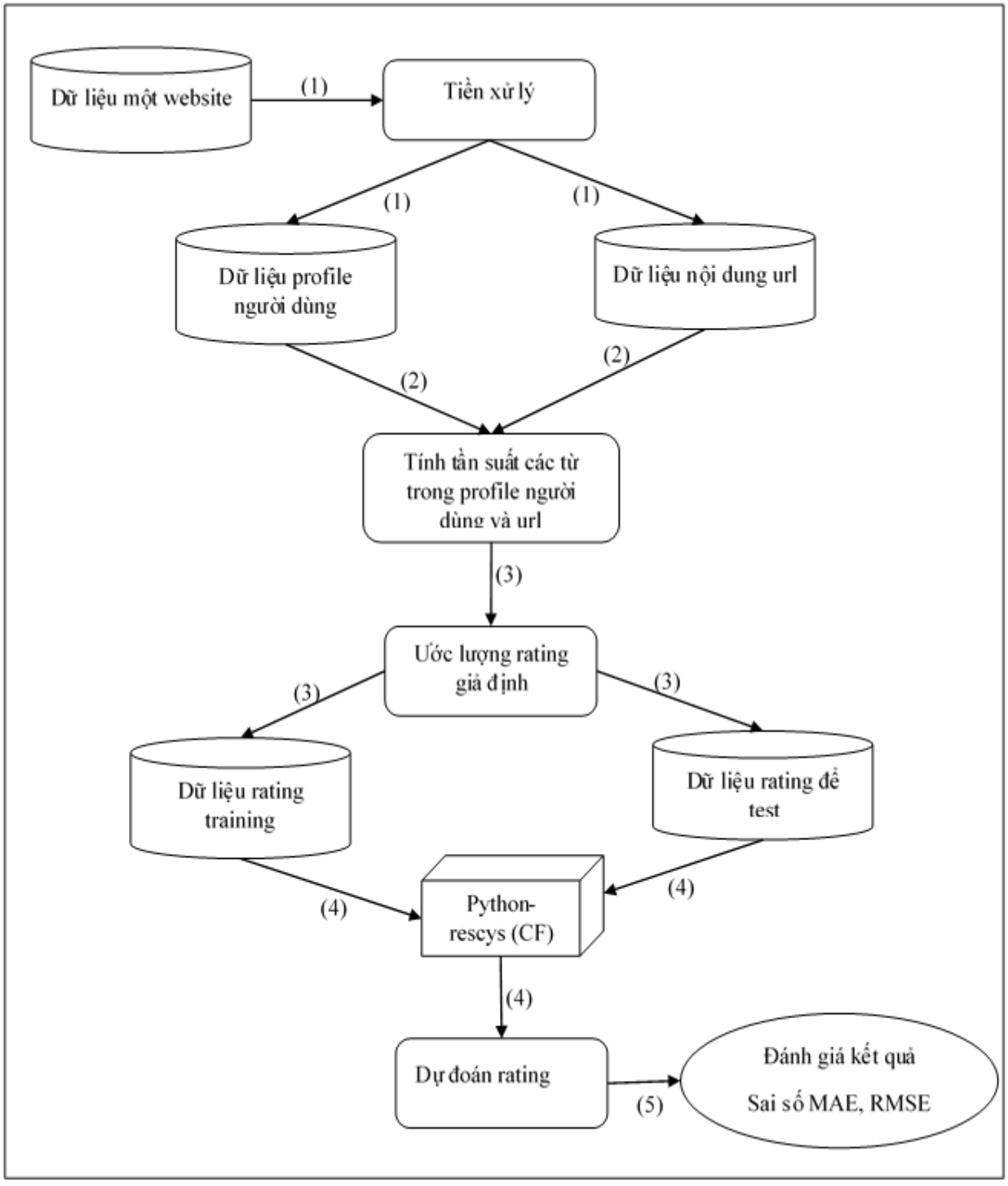
- (1): Tiền xử lý

- (2): Tính tần suất xuất hiện các từ

- (3): Ước lượng hạng giả định cho từng cặp user-item

- (4): Huấn luyện mô hình cộng tác, dự đoán hạng

- (5): Đánh giá độ lệch của mô hình




Hình 4.3 Sơ đồ thực nghiệm với hạng giả định là tần suất từ

Công việc tiền xử lý, huấn luyện mô hình cộng tác, dự đoán hạng, và đánh giá độ lệch mô hình (1, 4, 5) tương tự như mô hình tư vấn khi kết hợp với giả định hạng bằng LDA

**Công việc 2: Tính tần suất xuất hiện các từ**

Dựa trên tập từ điển gồm các token (từ) phân biệt và tập corpus là tần suất của các từ trong mỗi url từ bước 1, ta tính được tần suất xuất hiện của các từ trong dữ liệu duyệt web của người dùng (tổng nội dung của các url đã đọc của người dùng), và dữ liệu trên mỗi url. Ta được:

*Profile(c)* = $(w_{1c}, \ldots, w_{mc})$ với $w_{ic}$ biểu thị tần suất của từ khóa *i* trong dữ liệu duyệt web của *c*.

*Content(s)* = $(w_{is}, \ldots, w_{ns})$ với $w_{is}$ biểu thị tần suất của từ khóa *i* trong nội dung url *s*

**Công việc 3: Ước lượng hạng giả định**

Hạng giả định được tính bằng công thức:

$$r(u,i) = p(c,s) / p_{max}$$

Trong đó: $p(c,s) = \sum (w_{ic} * w_{is})$ với mỗi từ khóa *i (token i)* trong url *s*, $p_{max}$ là giá trị điểm cao nhất trong tập các $p(c, s_k)$ của người dùng u. Kết quả sẽ được lưu vào file user_rating với định dạng mỗi dòng là <*user_id, item_id, rating_value*>.

## 4.4 Kết quả và đánh giá

Kết quả của thực nghiệm được thể hiện ở bảng 4.5. Trong đó:

**(1)**: mô hình tư vấn khi dữ liệu hạng giả định từ mô hình LDA với khoảng cách Jensen-Shannon

**(2)**: mô hình tư vấn khi dữ liệu hạng giả định từ mô hình LDA với khoảng cách cosine

**(3)**: mô hình tư vấn khi dữ liệu hạng giả định từ phương pháp tần suất từ

Bảng 4.5 Kết quả thực nghiệm

| **Kết quả** | **Dữ liệu otoxemay.vn** | | | **Dữ liệu emdep.vn** | | |
|---|---|---|---|---|---|---|
| | **(1)** | **(2)** | **(3)** | **(1)** | **(2)** | **(3)** |
| **Số lượng hạng** | 19588 | | | 256123 | | |
| **Số lượng người dùng** | 1496 | | | 12356 | | |
| **Số lượng item** | 3504 | | | 24655 | | |
| **Sai số RMSE** | **0.11** | 0.16 | 0.16 | **0.09** | 0.12 | 0.13 |
| **Sai số MAE** | **0.08** | 0.13 | 0.12 | **0.07** | 0.09 | 0.09 |

Kết quả trên cho thấy, mô hình tư vấn khi sử dụng giả định hạng bằng mô hình LDA với khoảng cách Jensen-Shannon cho kết quả cao nhất so với các mô hình còn lại trên tập dữ liệu thực nghiệm. Qua đó cũng cho thấy, mô hình luận văn xây dựng nhìn chung có kết quả khả quan trên dữ liệu thực nghiệm, và có tính khả thi. Tuy nhiên, muốn đánh giá được chính xác hiệu quả của mô hình hệ tư vấn, cần đưa mô hình áp dụng vào chạy thực tế trên website. Và đây cũng là định hướng tiếp theo của nhóm nghiên cứu.

# Kết luận và định hướng nghiên cứu tiếp theo

Qua quá trình tìm hiểu về hệ tư vấn và các phương pháp tư vấn, luận văn đã đề xuất ra mô hình hệ tư vấn cho các website tạp chí ở Việt Nam sử dụng kĩ thuật lọc cộng tác và mô hình chủ đề ẩn LDA.

Luận văn đạt được một số kết quả sau đây:

- Giới thiệu hệ tư vấn, và các kĩ thuật sử dụng trong bài toán tư vấn, nghiên cứu về việc ứng dụng hệ tư vấn cho các website tại Việt Nam
- Phân tích hướng tiếp cận giải quyết vấn đề dữ liệu đánh giá ẩn của người dùng cho bài toán tư vấn
- Đề xuất mô hình hệ tư vấn website dựa trên khai phá dữ liệu vết duyệt web của người dùng, mô hình đã đưa thêm mô hình chủ đề ẩn LDA vào phương pháp cộng tác truyền thống để ước lượng hạng giả định của người dùng với url.
- Thực nghiệm mô hình hệ tư vấn đề xuất trên tập dữ liệu thực tế từ trang web http://www.otoxemay.vn/ và trang web http://www.emdep.vn/, đồng thời cũng thực nghiệm với một mô hình cơ sở (mô hình tư vấn khi kết hợp ước lượng hạng giả định bằng tần suất từ) để so sánh hiệu quả. Qua thực nghiệm, kết quả cho thấy mô hình mà luận văn đề xuất có tính khả thi.

Tuy nhiên, do hạn chế về thời gian nên luận văn vẫn tồn tại những hạn chế như: dữ liệu thực nghiệm còn chưa phong phú, cần có thêm một vài tập dữ liệu ở một số website khác để đánh giá, đồng thời cần có giải pháp đánh giá trên hiệu quả thực tế

Trong thời gian tới, chúng tôi sẽ thực hiện với dữ liệu ở nhiều website đa dạng hơn, và sẽ hướng tới việc tích hợp mô hình trên website để đánh giá hiệu quả thực tế.

# Tài liệu tham khảo

[1] David M. Blei, Andrew Y. Ng, Michael I. Jordan: “Latent Dirichlet Allocation”. Journal of Machine Learning Research (JMLR) 3:993-1022, 2003.

[2] Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor: “Recommender systems handbook”, Springer, 2011.

[3] Charu C. Aggarwal: “Recommender Systems” textbook, Springer, 2016.

[4] Dietmar Jannach, Alexander Felfernig, Gerhard Friedrich, and Markus Zanker: “Recommender Systems An introduction” book, Cambridge University Press, 2010.

[5] G.Adomavicius, A.Tuzhilin: “Towards the Next Generation of Recommender Systems. A Survey of the State-of-the-Art and Possible Extensions”. IEEE Transactions on Knowledge and Data Engineering, 2005.

[6] B. Sarwar, G. Karypis, J. Konstan, J. Riedl: “Item-based collaborative filtering recommendation algorithms”. Proceedings of the 10th international conference on World Wide Web, 2001, pages 285-295.

[7] HB.Deng: “Introduction to Recommendation System”. China University of Hongkong seminar, 2006.

[8] Netflix prize http://www.netflixprize.com/.

[9] R. M. Bell, Y. Koren, C. Volinsky: “The BellKor 2008 Solution to the Netflix Prize”. http://www.netflixprize.com/assets/ProgressPrize2008_BellKor.pdf.

[10] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl: “Incremental singular value decomposition algorithms for highly scalable recommender systems”. Proceedings of the 5th International Conference on Computer and Information Technology (ICCIT ’02), 2002.

[11] H. Polat and W. Du: “SVD-based collaborative filtering with privacy”. ACM symposium on Applied Computing, 2005, pp. 791-195.

[12] A. Felfernig and R. Burke: “Constraint-based recommender systems: technologies and research issues”. Proceedings of the 10th International Conference on Electronic Commerce (ICEC ’08) (Innsbruck, Austria), ACM, 2008, pp. 1–10.

[13] M. Zanker, M. Jessenitschnig, and W. Schmid: “Preference Reasoning with Soft Constraints in Constraint-Based Recommender Systems”. Constraints 15 (2010), no. 4, 574–595.

[14] M. Zanker and M. Jessenitschnig: “Collaborative feature-combination recommender exploiting explicit and implicit user feedback”. Proceedings of the 2009 IEEE Conference on Commerce and Enterprise Computing (CEC ’09) (Vienna), IEEE Computer Society, pp. 49-56, 2009.

[15] P. Melville, R. J. Mooney, and R. Nagarajan: “Content-Boosted Collaborative Filtering for Improved Recommendations”, Proceedings of the 18th National Conference on Artificial Intelligence (AAAI) (Edmonton, Alberta, Canada), 2002, pp. 187–192.

[16] R. Burke, P. Brusilovsky and A. Kobsa and W. Nejdl: “Hybrid web recommender systems”. The Adaptive Web: Methods and Strategies of Web Personalization, Springer, Heidelberg, Germany, 2007, pp. 377–408.

[17] Y. Hu, Y. Koren, C. Volinsky: “Collaborative Filtering for Implicit Feedback Datasets”. Proceeding of the 8th IEEE International Conference on Data Mining, 2008, pp. 263 -- 272.

[18] E. R. Nuez-Valdz, J. M. Cueva Lovelle, O. Sanjun Martnez, V. Garca-Daz, P. Ordoez de Pablos, C. E. Montenegro Marn: “Implicit feedback techniques on recommender systems applied to electronic book”. Computers in Human Behavior, 2012, pp. 1186-1193.

[19] E. R. Nuez-Valdz, J. M. Cueva Lovelle, G. Infante Hernandez, A. Juan Fuente, J. E. Labra-Gayo: “Creating recommendations on elictronic books”. Computers in Human Behavior, 2015, pp. 1320-1330.

[20] Megharani V. Misal, Pramod D. Ganjewar: “Electronic Books Recommender System Based on Implicit Feedback Mechanism and Hybrid Methods”.

International Journal of Advanced Research in Computer Science and Software Engineering, 2016, pp. 681-686.

[21] Thomas Hofmann, “Probabilistic Latent Semantic Analysis”. UAI 1999, pp. 289-196, 1999.

[22] Xuan-Hieu Phan, Cam-Tu Nguyen, Dieu-Thu Le, Le-Minh Nguyen, Susumu Horiguchi, Senior Member, IEEE and Quang-Thuy Ha. “A Hidden Topic-Based Framework toward Building Applications with Short Web Documents”. TKDE vol. 23 NO. 7, July 2011.

[23] Chong Wang, David M. Blei: “Collaborative topic modeling for recommending scientific articles”. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, 2011, pp. 448-456.

[24] Zhiqiang He, Zhongyi Wu, Bochong Zhou, Lei Xu, Weifeng Zhang: “Tourist routs recommendation based on Latent Dirichlet Allocation Model”. Web Information System and Application Conference (WISA), 2015.