LÀM GIÀU BIỂU DIỄN NỘI DUNG BẰNG TIÊN NGHIỆM THỂ LOẠI: MÔ HÌNH GAGE CHO HỆ THỐNG GỢI Ý

Dương Tấn Nghĩa[1*], Trần Tiến Dũng[2]
[*]Tác giả liên hệ, email: nghia.duongtan@hust.edu.vn. ORCID: 0000-0002-2442-6263

Ngày tòa soạn nhận được bài báo: 15/01/2026
Ngày phản biện đánh giá: 17/03/2026
Ngày bài báo được duyệt đăng: 14/04/2026

DOI: 10.59266/houjs.2026.1165

Tóm tắt: *Biểu diễn nội dung dạng dày đặc cho phim đóng vai trò nền tảng trong nhiều hệ gợi ý hiện đại, đặc biệt khi mục tiêu không chỉ là dự đoán sở thích mà còn là giải thích được vì sao một mục nội dung được đề xuất. Tag Genome của MovieLens cung cấp một ma trận mức liên quan liên tục giữa phim và nhãn mô tả, nhờ đó cho phép đo độ tương tự nội dung ở mức tinh hơn so với cách gắn nhãn rời rạc truyền thống. Tuy nhiên, các điểm liên quan này vẫn chịu ảnh hưởng của nhiễu, của thiếu bằng chứng quan sát và của bất nhất trong tri thức cộng đồng. Bài báo đề xuất mô hình Genre-Aware Genome Enrichment (GAGE), một cơ chế hiệu chỉnh điểm liên quan dựa trên tiên nghiệm theo thể loại, nhằm khử nhiễu và làm giàu biểu diễn nội dung trước khi đưa vào mô hình gợi ý dựa trên độ tương tự. Phương pháp gồm bốn giai đoạn: xác định ngưỡng thích ứng theo từng nhãn, ước lượng xác suất tiên nghiệm P(nhãn|thể loại), tổng hợp tiên nghiệm đa thể loại cho từng phim, và cập nhật phi tuyến điểm liên quan với cơ chế bảo vệ biên để tránh làm méo các giá trị đã có độ chắc chắn cao. Trên dữ liệu MovieLens 20M sau tiền xử lý với 10.133 phim và 916 nhãn, thực nghiệm với phương pháp lân cận gần nhất k, độ tương tự cosin và sai số căn phương trung bình cho thấy khi tăng cường độ can thiệp theo thể loại, sai số dự đoán có xu hướng tăng so với mốc đối chứng. Kết quả này chỉ ra rằng việc khuếch tăng hai chiều theo thể loại có thể làm đồng nhất hóa các vectơ nội dung và bơm dương tính giả. Từ đó, bài báo đề xuất một định hướng thận trọng hơn: ưu tiên khử nhiễu một chiều thay vì đồng thời vừa thưởng vừa phạt, và đánh giá mô hình theo nhiều mục tiêu gợi ý thay vì chỉ dựa trên một chỉ số duy nhất.*

Từ khóa: *hệ thống gợi ý, biểu diễn nội dung, tiên nghiệm thể loại, độ tương tự cosin, khử nhiễu*

I. Đặt vấn đề

Các hệ gợi ý đã trở thành một hạ tầng không thể thiếu của kinh tế số, từ nền tảng thương mại điện tử, dịch vụ xem phim trực tuyến cho tới cổng thông tin công cộng. Về bản chất, một hệ gợi ý hiệu quả không chỉ cần dự đoán đúng người dùng sẽ thích gì, mà còn phải bảo toàn được cấu trúc ngữ nghĩa của nội dung để hỗ trợ diễn giải, khám phá và kiểm soát

[1] Đại học Bách Khoa Hà Nội, Hà Nội, Việt Nam
[2] Trường Đại học Mở Hà Nội, Hà Nội, Việt Nam

thiên lệch. Trong dòng nghiên cứu này, bộ dữ liệu MovieLens được xem là một chuẩn tham chiếu có ảnh hưởng sâu rộng, bởi nó phản ánh lịch sử lâu dài của một hệ thống gợi ý vận hành thực tế, đồng thời cung cấp nhiều lớp dữ liệu từ đánh giá, gắn nhãn tới siêu dữ liệu mô tả phim (Harper & Konstan, 2015).

Trong số các thành phần dữ liệu của MovieLens, Tag Genome giữ một vị trí đặc biệt vì nó chuyển bài toán gắn nhãn rời rạc thành bài toán lượng hóa liên tục mức liên quan giữa phim và nhãn. Thay vì chỉ ghi nhận một nhãn có xuất hiện hay không, Tag Genome ước lượng mức độ phù hợp của từng nhãn đối với từng phim trên thang liên tục từ 0 đến 1, nhờ đó tạo ra một biểu diễn nội dung dày đặc, thuận lợi cho việc tính toán độ tương tự và thiết kế giao diện tương tác theo ngữ nghĩa (Vig và cộng sự, 2012). Cách tiếp cận này mở ra khả năng xử lý phim như một cấu trúc giàu thuộc tính thay vì một thực thể chỉ được nhận diện qua tiêu đề và thể loại.

Tuy nhiên, chính vì Tag Genome được hình thành từ tri thức cộng đồng và từ các mô hình suy diễn mức liên quan, nên ma trận điểm liên quan không phải là chân lý tuyệt đối. Một số nhãn có thể nhận điểm dương nhỏ dù trên thực tế không mang ý nghĩa đối với phim; ngược lại, có những nhãn thật sự quan trọng nhưng bị cho điểm thấp do thiếu tín hiệu quan sát hoặc do dữ liệu gắn nhãn ban đầu không đủ bao quát. Kotkov và cộng sự (2021) chỉ ra rằng bài toán dự đoán mức liên quan gắn nhãn chịu ảnh hưởng đáng kể của bất nhất giữa người đánh giá, đặc biệt đối với các nhãn mang tính chủ quan. Vì vậy, nếu sử dụng trực tiếp các điểm liên quan này để xây dựng độ tương tự nội dung, mô hình gợi ý có nguy cơ học theo cả tín hiệu hữu ích lẫn nhiễu.

Từ nhận định trên, bài báo đặt ra câu hỏi cốt lõi: liệu có thể dùng tri thức ở mức siêu dữ liệu ổn định hơn, cụ thể là thể loại phim, làm tiên nghiệm để hiệu chỉnh lại điểm liên quan gắn nhãn hay không. Trực giác ở đây là đơn giản nhưng có sức gợi mạnh. Nếu một nhãn rất hiếm khi gắn với một thể loại nhất định, một giá trị liên quan cao cho nhãn đó có thể đáng ngờ. Ngược lại, nếu một nhãn xuất hiện lặp đi lặp lại trong một thể loại, thì một điểm số quá thấp có thể là biểu hiện của việc tín hiệu chưa được bộc lộ đầy đủ.

Trên cơ sở đó, nghiên cứu đề xuất mô hình Genre-Aware Genome Enrichment, viết tắt là GAGE. Mô hình này xem thể loại như một nguồn tiên nghiệm thống kê để điều chỉnh mức liên quan nhãn - phim theo hướng vừa bảo tồn thông tin hiện hữu, vừa giảm ảnh hưởng của các giá trị phi lý. Đóng góp chính của bài báo gồm bốn điểm: thứ nhất, xây dựng một quy trình hiệu chỉnh bốn giai đoạn có thể diễn giải được; thứ hai, kết nối cách hiệu chỉnh này với bài toán gợi ý dựa trên nội dung bằng độ tương tự cosin; thứ ba, phân tích thực nghiệm tác động của cường độ can thiệp tới sai số dự đoán; và thứ tư, rút ra hàm ý phương pháp luận rằng khử nhiễu một chiều có thể phù hợp hơn khuếch tăng hai chiều trong các mô hình lân cận gần nhất.

II. Cơ sở lý thuyết

2.1. Hệ thống gợi ý dựa trên nội dung

Lọc dựa trên nội dung (CBF) xây dựng hồ sơ người dùng và hồ sơ sản phẩm dựa trên các thuộc tính nội dung có thể trích xuất được, sau đó tính toán độ tương đồng giữa chúng để đưa ra gợi ý (Pazzani & Billsus, 2007). Trong bối cảnh gợi ý phim, các thuộc tính nội dung thường bao gồm thể loại, diễn viên, đạo diễn, mô tả cốt truyện và các thẻ ngữ

nghĩa (semantic tags). Ưu điểm của CBF là khả năng gợi ý cho người dùng mới (cold-start user) và không bị ảnh hưởng bởi vấn đề thưa thớt ma trận xếp hạng. Tuy nhiên, nhược điểm lớn của phương pháp này là xu hướng over-specialization - chỉ gợi ý các mục tương tự những gì người dùng đã từng tương tác (Lops và cộng sự, 2011).

2.2. Genome Scores và Tag-Based Recommendation

Vig và cộng sự (2012) đã giới thiệu khái niệm Tag Genome trong MovieLens - một không gian tag dày đặc nơi mỗi phim được biểu diễn bởi vector điểm relevance so với 1.128 tags ngữ nghĩa. Genome Scores được ước lượng bằng mô hình học máy kết hợp dữ liệu gán tag của người dùng, đánh giá sao, và các đặc trưng nội dung. Mặc dù Tag Genome đã được chứng minh là biểu diễn ngữ nghĩa hiệu quả hơn các phương pháp truyền thống, nó vẫn mang theo hạn chế cơ bản: chất lượng phụ thuộc vào mật độ đóng góp của cộng đồng người dùng, dẫn đến sự không đồng đều giữa các phim phổ biến và phim ít được đánh giá (long-tail problem).

2.3. Ứng dụng suy luận Bayes trong hệ thống gợi ý

Suy luận Bayes cung cấp nền tảng lý thuyết vững chắc để kết hợp tri thức tiên nghiệm (prior knowledge) với bằng chứng quan sát (likelihood) nhằm cập nhật niềm tin (posterior belief). Trong hệ thống gợi ý, Breese và cộng sự (1998) đã đề xuất áp dụng mạng Bayes để mô hình hóa sở thích người dùng. Gần đây hơn, các phương pháp kết hợp tri thức miền (domain knowledge) vào quá trình học đã trở nên phổ biến, đặc biệt trong các bài toán có dữ liệu thưa. Bài báo này kế thừa tư tưởng này bằng cách sử dụng thể loại phim như một tiên nghiệm Bayes để hiệu chỉnh Genome Score.

III. Phương pháp, vật liệu nghiên cứu

3.1. Dữ liệu và tiền xử lý

Thực nghiệm được tiến hành trên bộ dữ liệu MovieLens 20M (Harper & Konstan, 2015) với 10.133 phim và 916 tags genome. Quá trình tiền xử lý bao gồm: loại bỏ các phim không có thông tin thể loại, chuẩn hóa ma trận Genome Score theo phân vị Q3 từng tag (Giai đoạn 1), và xây dựng ma trận xác suất tiên nghiệm P từ 19 thể loại tiêu chuẩn theo phân loại MovieLens.

3.2. Phương pháp đánh giá

Đánh giá được thực hiện theo giao thức tiêu chuẩn: sử dụng thuật toán kNN với k=20, độ đo tương đồng Cosine Similarity để xác định láng giềng trong không gian vector genome, và RMSE (Root Mean Squared Error) trên tập kiểm tra để đo chất lượng dự đoán xếp hạng. Các tham số cố định: $\beta = \text{mean}(P) \approx 0.08$, $\sigma = \text{std}(P) \approx 0.15$. Biến thực nghiệm duy nhất là α, được thử nghiệm tại sáu mức độ: {0.0, 0.1, 0.2, 0.4, 0.8, 1.2}. Giá trị $\alpha = 0.0$ tương đương Baseline (không áp dụng GAGE).

IV. Thuật toán GAGE

Thuật toán GAGE được thiết kế theo kiến trúc bốn giai đoạn tuần tự, mỗi giai đoạn thực hiện một bước biến đổi toán học rõ ràng trên dữ liệu đầu vào là ma trận Genome Score gốc , trong đó M là số lượng phim và T là số lượng tags.

4.1. Giai đoạn 1: Chuẩn hóa ngữ nghĩa bằng ngưỡng cục bộ

Thay vì áp dụng một ngưỡng toàn cục cố định (Global Threshold) để quyết định một tag có 'tồn tại' trong phim hay không - phương pháp này bị chỉ trích vì

không tính đến sự phân phối không đồng đều của các tag - GAGE sử dụng Phân vị thứ ba (Q3, tức phân vị 75%) của riêng từng tag làm ngưỡng thích nghi:

$$Threshold_i = Q3(S_{:,i}) \quad (1)$$

Kỹ thuật này được gọi là *Adaptive Sensitivity*: với các tag phổ biến (như "Drama"), ngưỡng cao giúp lọc bớt nhiễu; với các tag hiếm (như "Aliens"), ngưỡng thấp bảo toàn tín hiệu yếu, tránh bỏ sót thông tin có giá trị.

4.2. Giai đoạn 2: Xây dựng ma trận xác suất tiên nghiệm

Giai đoạn này xây dựng ma trận P ∈ ℝ^(19×916) - được gọi là 'Bản đồ tri thức' - biểu diễn xác suất có điều kiện của tag T_i khi biết thể loại Gj:

$$P(G_j) = \Sigma_{m \in M_{G_j}} I(S_{m,i} > Q3_i) / \left|M_{G_j}\right| \quad (2)$$

Trong đó I(·) là hàm chỉ thị, M_{G_j} là tập hợp tất cả phim thuộc thể loại G_j. Ma trận P mã hóa tri thức thể loại như tiên nghiệm Bayes, ví dụ: P(Spaceship | Sci-Fi) ≈ 0.85 trong khi P(Spaceship | Romance) ≈ 0.01. Điều này cho phép phát hiện các điểm Genome Score bất thường - khi một phim Romance có điểm cao cho tag Spaceship, đây nhiều khả năng là nhiễu thống kê.

4.3. Giai đoạn 3: Tính toán trọng số ngữ cảnh

Phần lớn phim trong MovieLens thuộc nhiều thể loại đồng thời (multi-label). Để xác định mức độ phù hợp của tag i với phim m trong bối cảnh đa thể loại, GAGE sử dụng hàm MAX (logic hợp - union logic):

$$W_{m,i} = max_{g \in Genres(m)} P(g) \quad (3)$$

Lựa chọn hàm MAX dựa trên nguyên tắc 'Strongest Advocate': một tag tồn tại trong phim chỉ cần có một thể loại biện hộ cho sự hiện diện của nó. Nếu sử dụng trung bình cộng, các thể loại không liên quan sẽ làm pha loãng tín hiệu (Signal Dilution), gây mất mát thông tin. Ví dụ, với phim Action|Sci-Fi|Romance và tag Gun: W = max(0.75, 0.40, 0.02) = 0.75, trong khi trung bình cộng chỉ cho giá trị 0.39.

4.4. Giai đoạn 4: Công thức điều chỉnh phi tuyến

Điểm Genome Score mới S_{new} được tính bằng cách cộng thêm một lượng điều chỉnh Δ vào điểm gốc S_{old}:

$$S_{new} = S_{old} + \alpha \cdot S_{old}(1 - S_{old}) \cdot \tanh((W_{m,i} - \beta)/\sigma) \quad (4)$$

Trong đó:

(a) Tham số α (Learning Rate): Kiểm soát cường độ can thiệp. Giá trị α = 0.4 được lựa chọn như mức thận trọng (conservative), chỉ tạo ra sự thay đổi khoảng 10% so với điểm gốc, nhằm bảo toàn đặc trưng riêng của từng phim.

(b) Cụm $S_{old}(1 - S_{old})$: Đây là đạo hàm của hàm sigmoid $f'(x) = f(x)(1 - f(x))$. Cơ chế này đạt cực đại tại S_{old}=0,5 (vùng không chắc chắn) và tiến về 0 khi $S_{old} \to 0$ hoặc $S_{old} \to 1$ (vùng đã xác định rõ). Điều này bảo vệ những điểm số đã có độ tin cậy cao khỏi sự can thiệp không cần thiết của thuật toán.

(c) Hàm tanh (·): Quyết định hướng điều chỉnh dựa trên hiệu số (W - β)/σ. Khi W >> β (tag phù hợp với thể loại), tanh cho giá trị dương → tăng điểm (Boosting). Khi W << β (tag không phù hợp), tanh cho giá trị âm → giảm điểm (Denoising). Hàm tanh đảm bảo |Δ| bị giới hạn trong khoảng hợp lý.

(d) Tham số β và σ: β là giá trị trung bình của ma trận P, đóng vai trò 'điểm cân bằng' (pivot point) phân định tag 'hợp lý' và tag 'nhiễu'. σ là độ lệch chuẩn của P, đảm bảo hàm tanh hoạt động trong vùng nhạy cảm và tránh bão hòa.

Bảng 1. Hệ số bảo vệ biên $S_{old}(1 - S_{old})$ theo giá trị điểm gốc S_{old}

S_{old}	$S_{old}(1 - S_{old})$	Mức độ cho phép thay đổi
0.01	0.0099	Rất thấp - điểm đã xác định rõ
0.20	0.1600	Trung bình
0.50	0.2500	Cực đại - vùng không chắc chắn nhất
0.80	0.1600	Trung bình
0.99	0.0099	Rất thấp - điểm đã xác định rõ

V. Kết quả và thảo luận

Để đánh giá ảnh hưởng của cơ chế Genre-Aware Boosting trong mô hình GAGE, chúng tôi tiến hành thí nghiệm thay đổi tham số α, là hệ số điều chỉnh mức độ tăng cường điểm liên quan của các tag phù hợp với thể loại phim. Tham số này kiểm soát cường độ can thiệp của tri thức tiên nghiệm thể loại vào vector biểu diễn nội dung.

Trong thí nghiệm, số lượng láng giềng được cố định ở k=20 và độ đo tương tự được sử dụng là Cosine Similarity, nhằm đảm bảo tính nhất quán với cấu hình baseline của hệ thống gợi ý. Giá trị được thay đổi từ 0.0 đến 1.2, trong đó α=0.0 tương ứng với mô hình baseline không áp dụng GAGE.

Hiệu năng của mô hình được đánh giá thông qua chỉ số Root Mean Square Error (RMSE) trên tập kiểm thử. Ngoài giá trị RMSE tuyệt đối, chúng tôi cũng báo cáo độ chênh lệch (Δ) so với baseline để quan sát mức độ cải thiện hoặc suy giảm của từng cấu hình. Kết quả thực nghiệm được trình bày trong Bảng 2.

Bảng 2. Kết quả RMSE theo giá trị tham số α (k=20, Cosine Similarity)

Alpha (α)	RMSE (k=20)	Delta so với Baseline	Đánh giá
0.0 (Baseline)	0.81672	0.00000	Mốc chuẩn (Baseline)
0.1	0.81701	+0.00029	Suy giảm nhẹ
0.2	0.81742	+0.00070	Suy giảm
0.4	0.81839	+0.00167	Suy giảm rõ
0.8	0.82084	+0.00412	Suy giảm mạnh
1.2	0.82331	+0.00659	Tệ nhất

Kết quả trong Bảng 2 cho thấy xu hướng đơn điệu: RMSE tăng dần theo α, nghĩa là mọi mức độ can thiệp đều làm giảm chất lượng gợi ý so với Baseline (RMSE = 0.81672). Mức tăng nhỏ nhất là +0.00029 tại α = 0.1 và lớn nhất là +0.00659 tại α = 1.2. Kết quả này cho thấy thuật toán GAGE trong cấu hình hiện tại chưa đạt được mục tiêu cải thiện RMSE.

Nguyên nhân chính dẫn đến sự suy giảm chất lượng là hiện tượng Genre Homogenization: cơ chế tăng điểm (Boosting) dựa trên thể loại đã làm cho các bộ phim trong cùng một thể loại trở nên quá giống nhau về mặt vector genome. Điều này phá vỡ tính đặc thù của từng tác phẩm - ví dụ, cả 'The Dark Knight' lẫn một bộ phim Action bình thường đều nhận được mức tăng điểm tương tự cho tag Gun, mặc dù chúng có phong cách nghệ thuật rất khác nhau. Hệ quả là kNN không còn tìm được láng giềng 'cùng gu' (taste-similar) mà chỉ tìm được láng giềng 'cùng loại' (genre-similar) - một dạng suy thoái tương tự vấn đề over-specialization trong CBF.

Vấn đề thứ hai là False Positive Injection: khi W > β, thuật toán tăng điểm cho tất cả tag phù hợp với thể loại, bất kể

điểm gốc là bao nhiêu. Điều này vô tình cộng điểm cho các tag không tồn tại trong phim cụ thể đó. Ví dụ điển hình: một bộ phim võ thuật thuần túy (Action, không sử dụng súng) vẫn nhận được tăng điểm cho tag Gun chỉ vì nó thuộc thể loại Action. Kết quả là thêm nhiễu mới vào dữ liệu - nghịch lý khi mục tiêu ban đầu là khử nhiễu.

Độ đo Cosine Similarity nhạy cảm với cả hướng lẫn độ lớn tương đối của vector. Việc cộng $\Delta > 0$ (chủ yếu xảy ra trong trường hợp boosting) làm thay đổi độ dài vector một cách không đồng đều giữa các phim, ảnh hưởng tiêu cực đến kết quả tính khoảng cách trong không gian kNN. Một số công trình gợi ý rằng trong không gian vector genome nhiều chiều và có nhiễu, các độ đo thay thế như Pearson Correlation hay Weighted Jaccard có thể ổn định hơn (Breese và cộng sự, 1998).

VI. Kết luận

Bài báo đã trình bày thuật toán GAGE - một nỗ lực có cơ sở lý thuyết vững chắc trong việc kết hợp tri thức chuyên gia (Genre) vào quá trình làm giàu dữ liệu Genome Score. Kiến trúc bốn giai đoạn được thiết kế cẩn thận với ngưỡng thích nghi, tiên nghiệm Bayes, trọng số MAX và điều chỉnh phi tuyến. Tuy nhiên, kết quả thực nghiệm cho thấy phương pháp tăng cường hai chiều (Boosting + Denoising) với độ đo Cosine gây ra phản tác dụng trên bộ dữ liệu MovieLens 20M, với RMSE tăng dần theo cường độ can thiệp α. Phân tích chỉ ra rằng vấn đề cốt lõi không nằm ở nền tảng lý thuyết mà ở cơ chế thực thi - đặc biệt là nguy cơ dương tính giả và đồng nhất hóa thể loại.

Tài liệu tham khảo

Adamopoulos, P., & Tuzhilin, A. (2014). On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)* (pp. 153-160). ACM. https://doi.org/10.1145/2645710.2645752

Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98)* (pp. 43-52). Morgan Kaufmann.

Harper, F. M., & Konstan, J. A. (2015). The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems, 5*(4), Article 19. https://doi.org/10.1145/2827872

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5-53. https://doi.org/10.1145/963770.963772

Kaminskas, M., & Bridge, D. (2017). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems, 7*(1), Article 2. https://doi.org/10.1145/2926720

Kotkov, D., Maslov, A., & Neovius, M. (2021). Revisiting the tag relevance prediction problem. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)* (pp. 1768-1772). ACM. https://doi.org/10.1145/3404835.3463019

Lops, P., de Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems*

handbook (pp. 73-105). Springer. https://doi.org/10.1007/978-0-387-85820-3_3

Park, Y.-J., & Tuzhilin, A. (2008). The long tail of recommender systems and how to leverage it. In *Proceedings of the 2008 ACM Conference on Recommender Systems (RecSys '08)* (pp. 11-18). ACM. https://doi.org/10.1145/1454008.1454012

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (Lecture Notes in Computer Science, Vol. 4321, pp. 325-341). Springer. https://doi.org/10.1007/978-3-540-72079-9_10

Rendle, S., Freudenthaler, C., Gantner, Z., & Schmidt-Thieme, L. (2009). BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI '09)* (pp. 452-461). AUAI Press.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01)* (pp. 285-295). ACM. https://doi.org/10.1145/371920.372071

Vargas, S., & Castells, P. (2011). Rank and relevance in novelty and diversity metrics for recommender systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)* (pp. 109-116). ACM. https://doi.org/10.1145/2043932.2043955

Vargas, S., Baltrunas, L., Karatzoglou, A., & Castells, P. (2014). Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (RecSys '14)* (pp. 209-216). ACM. https://doi.org/10.1145/2645710.2645743

Vig, J., Sen, S., & Riedl, J. (2012). The tag genome: Encoding community knowledge to support novel interaction. *ACM Transactions on Interactive Intelligent Systems, 2*(3), Article 13. https://doi.org/10.1145/2362394.2362395

Zhang, S., Yao, L., Sun, A., & Tay, Y. (2019). Deep learning-based recommender system: A survey and new perspectives. *ACM Computing Surveys, 52*(1), Article 5. https://doi.org/10.1145/3285029

Zhang, Y., & Chen, X. (2020). Explainable recommendation: A survey and new perspectives. *Foundations and Trends in Information Retrieval, 14*(1), 1-101. https://doi.org/10.1561/1500000066

CONTENT REPRESENTATION ENRICHMENT VIA GENRE PRIORS: THE GAGE MODEL FOR RECOMMENDER SYSTEMS

Duong Tan Nghia[1], Tran Tien Dung[2]

Abstract: *Dense content representations for movies constitute a fundamental component in many modern recommender systems, particularly when the objective extends beyond preference prediction to providing interpretable explanations for recommended items. The Tag Genome dataset from MovieLens provides a continuous relevance matrix between movies and descriptive tags, enabling finer-grained measurement of content similarity than traditional discrete tagging approaches. However, these relevance scores are still affected by noise, limited observational evidence, and inconsistencies within community-generated knowledge. This paper proposes the Genre-Aware Genome Enrichment (GAGE) model, a mechanism that adjusts tag relevance scores using genre-based priors in order to denoise and enrich content representations before they are used in similarity-based recommendation models. The proposed approach consists of four stages: (1) determining adaptive thresholds for each tag, (2) estimating prior probabilities P(tag|genre), (3) aggregating multi-genre priors for each movie, and (4) applying a nonlinear update to relevance scores with boundary-protection mechanisms to avoid distorting values that already exhibit high certainty. Experiments conducted on the MovieLens 20M dataset, after preprocessing (10,133 movies, 916 tags), evaluate the approach using the k-nearest neighbors algorithm, cosine similarity, and root mean square error (RMSE) as the evaluation metric. The results show that increasing the level of genre-based intervention tends to increase prediction error compared with the baseline model. This finding suggests that bidirectional amplification based on genre priors may homogenize content vectors and introduce false positive signals. Based on these observations, the paper proposes a more cautious direction: prioritizing one-directional denoising rather than simultaneously applying reward and penalty adjustments, and evaluating recommender models under multi-objective recommendation criteria rather than relying solely on a single accuracy metric.*

Keywords: *recommender systems, content representation, genre prior, cosine similarity, denoising*

---

[1] Hanoi University of Science and Technology, Hanoi, Vietnam
[2] Hanoi Open University, Hanoi, Vietnam