Machine Translated by Google

Một mô hình cơ học dịch tễ học của động lực phòng vé

Naghmeh Momeni1,*, Amir Tohidi Kalorazi2, Michael Rabbat1, và Babak Fotouhi3,4

1Khoa Kỹ thuật Điện và Máy tính, Đại học McGill, Montreal, Quebec, Canada
2Khoa Kỹ thuật Điện, Đại học Công nghệ Sharif, Tehran, Iran
3Chương trình Động lực học Tiến hóa, Đại học Harvard, Cambridge, MA, Hoa Kỳ
4Viện Khoa học Xã hội Định lượng, Đại học Harvard, Cambridge, MA, Hoa Kỳ
*naghmeh.momenitaramsari@mail.mcgilll.ca

TÓM TẮT

Trong bài báo này, chúng tôi đề xuất một mô hình cơ học liên kết các tương tác xã hội vi mô với các biến quan sát vĩ mô trong trường hợp khuếch tán các quyết định xem phim. Chúng tôi thiết kế một mô hình dịch bệnh tổng quát để nắm bắt sự tiến hóa theo thời gian của doanh thu phòng vé. Mô hình này bổ sung ảnh hưởng xã hội và hiệu ứng trí nhớ vào các mô hình dịch bệnh thông thường. Phù hợp mô hình với một tập dữ liệu thời gian chứa doanh thu hàng tuần trong nước của 5000 bộ phim có doanh thu cao nhất mọi thời đại tại Hoa Kỳ, chúng tôi thấy rằng một mô hình hai tham số có thể nắm bắt được một phần đáng kể của phương sai quan sát được trong dữ liệu. Sử dụng phân phối các tham số ước tính cho các thể loại khác nhau, sau đó chúng tôi trình bày một mô hình dự đoán cung cấp các ước tính a-priori hợp lý về doanh số bán hàng trong tương lai theo thời gian.

Giới thiệu

Hiểu được hành vi tập thể của con người là một chủ đề được nghiên cứu nhiều trong nhiều lĩnh vực khác nhau. Có các cách tiếp cận hồi cứu và các cách tiếp cận dự đoán. Trong cách tiếp cận trước, mục đích là hiểu cách thức (và lý do) một số phong trào xã hội tập thể xuất hiện. Ví dụ bao gồm các cuộc cách mạng1, 2 sự xuất hiện đột ngột của các thái độ cấp tiến trên toàn quốc3, 4 như xu hướng thời trang và mốt nhất thời7, 8, khủng hoảng tài chính5, 6, và thậm chí cả những chủ đề tầm thường hơn như Mức độ phức tạp của các vấn đề xã hội tập thể quá cao để cho phép có những câu trả lời đơn giản, điều này thể hiện ở thực tế là đối với nhiều vấn đề đã nói, vẫn chưa có khuôn khổ hội tụ và hiếm khi đạt được sự đồng thuận về cách thức (và đặc biệt là lý do) một số hiện tượng xã hội nhất định xuất hiện. Phương pháp tiếp cận dự đoán phổ biến hơn trong các ngành như chính sách công, tài chính, kinh tế và tiếp thị. Mục đích là dự báo cách mọi người sẽ phản ứng với một số chính sách, công nghệ, sản phẩm, v.v. nhất định với mức độ chính xác hợp lý. Hành vi của người tiêu dùng, hiệu ứng truyền miệng và hành vi bầy đàn là những chủ đề trung tâm trong nghiên cứu tài chính hành vi và tiếp thị9–12. Lý do chính rất rõ ràng; trong một hệ thống kinh tế xã hội có trọng tâm là thị trường tự do, các nhà đầu tư thường thích tăng hiệu quả đầu tư bằng cách dự đoán tương lai, ngay cả một phần. Mặc dù đã có tiến bộ, nhưng hiện trạng trong nhiều lĩnh vực vẫn còn lâu mới có thể dự đoán chắc chắn. Câu ngạn ngữ phổ biến “Một nửa số tiền tôi chi cho quảng cáo là lãng phí; vấn đề là tôi không biết đó là nửa nào”, thường được cho là của nhà tiên phong tiếp thị John Wanamaker13, vẫn còn phù hợp với tình trạng kiến thức hiện nay. Những lời kể phổ biến về thách thức dự đoán này cũng rất nhiều. Một ví dụ là việc Harry Potter bị 12 nhà xuất bản từ chối trước khi cuối cùng được xuất bản, tóm tắt khả năng hiện tại của ngay cả các chuyên gia trong việc dự đoán thành công của sản phẩm và hành vi của người tiêu dùng.

Trong bài báo này, chúng tôi tập trung vào một ví dụ cụ thể về tiêu dùng và cố gắng mô tả nó bằng một mô hình đơn giản nhất có thể. Chúng tôi tập trung vào phim ảnh. Dự báo hiệu suất phòng vé của phim là một chủ đề được nghiên cứu kỹ lưỡng15–21. Trong kỷ nguyên hậu Internet (đã biến đổi quy mô và phạm vi điển hình của

arXiv:1613.03085v1

Machine Translated by Google

(dữ liệu có sẵn), ngày càng có nhiều sự chú ý được dành cho số lượng tìm kiếm, nội dung blog và xu hướng truyền thông xã hội để hiểu và dự đoán hành vi của người tiêu dùng trong mọi bối cảnh tiếp thị, bao gồm cả doanh thu phim ảnh. Hầu hết các nghiên cứu này đều nhằm mục đích dự đoán doanh thu trong tương lai dựa trên dữ liệu trước/sau khi phát hành, chẳng hạn như các bài phê bình trước khi phát hành22, 23, các đề cập trên Twitter24, lượt xem trên YouTube25, các đề cập trên tin tức trước khi phát hành26, lượt truy cập và chỉnh sửa trang Wikipedia27 và dữ liệu truy vấn của công cụ tìm kiếm28. Mặc dù các nhóm nghiên cứu của hãng phim đang nâng cao kiến thức về những gì làm cho một bộ phim trở nên phổ biến, nhưng sự hiểu biết hiện tại về hành vi của người tiêu dùng trong bối cảnh này và khả năng dự đoán thành công phòng vé trong tương lai vẫn còn lâu mới chắc chắn.

Ngược lại với các cách tiếp cận của các nghiên cứu được đề cập ở trên, chúng tôi áp dụng cách tiếp cận cơ học, từ dưới lên trong bài báo này để nghiên cứu sự tiến hóa của mức độ phổ biến của phim. Chúng tôi quan tâm đến việc mô hình hóa các cơ chế liên cá nhân vi mô thúc đẩy động lực này và tìm cách nắm bắt sự xuất hiện của các kết quả quan sát được ở cấp độ vĩ mô (có lợi thế là có sẵn dữ liệu). Phim là một loại hàng hóa đặc biệt, với các đặc điểm riêng về mặt khuếch tán. Trong giai đoạn hiện tại của nền kinh tế tư bản, hầu hết các sản phẩm đều phải chịu sự gia tăng liên tục: thường xuyên, các phiên bản sản phẩm mới (ví dụ: điện thoại di động) xuất hiện và mọi người cập nhật tài sản của mình với tốc độ khác nhau (một số phải có iPhone mới nhất càng sớm càng tốt, một số phải đợi lâu hơn, một số bị đẩy qua sự lỗi thời theo kế hoạch, v.v.). Điều này ít nhiều đúng đối với một phần đáng kể các sản phẩm đang được sử dụng: điện thoại di động, TV, máy chơi trò chơi điện tử, máy tính xách tay và máy tính bảng, để nêu một vài ví dụ. Tuy nhiên, phim có động lực khác (chúng tôi xin nhấn mạnh rằng chúng tôi chỉ tập trung vào doanh thu phòng vé; chúng tôi không xem xét doanh thu tiếp theo thông qua việc phát hành DVD và các khoản tiền bản quyền khác Hầu như tất cả mọi người đều đi xem phim cùng nhau. Nghĩa là, xem phim là một hoạt động chung. Vì vậy, thành phần xã hội thúc đẩy quyết định xem phim là rất mạnh. Ngoài ra còn có yếu tố thời gian: phim không ở trên màn hình vô thời hạn và được thay thế tuần tự bằng những phim mới. Trong bài báo này, chúng tôi tìm cách xây dựng một mô hình tối thiểu nắm bắt được những đặc điểm này.

Chúng tôi thiết kế một mô hình cơ bản về ảnh hưởng xã hội bắt nguồn từ các mô hình dịch bệnh thông thường (xem29 để biết đánh giá kỹ lưỡng về các mô hình này) với một sự thay đổi. Bằng cách thêm hiệu ứng bộ nhớ vào một mô hình loại dịch bệnh cơ bản, chúng tôi thu được một mô hình hai tham số có thể phân tích được. Chúng tôi sử dụng dữ liệu về doanh số bán hàng trong nước (tại Hoa Kỳ) hàng tuần của 5000 bộ phim gần đây để phù hợp với các tham số mô hình và cho thấy rằng mô hình có thể nắm bắt được sự tiến hóa theo thời gian của doanh thu phòng vé một cách đáng tin cậy. Mặc dù mục đích chính của chúng tôi là làm sáng tỏ các cơ chế xã hội vi mô chứ không phải dự đoán, chúng tôi cho thấy rằng như một sản phẩm phụ của phân tích, các kết quả cũng có thể có tiện ích định hướng dự đoán đáng kể với độ chính xác đáng chú ý, mà chúng tôi phân tích nhưng không phải là chủ đề chính của bài báo.

Kết quả

mô hình

Chúng tôi biểu thị một cá nhân đã xem phim bằng I (bị nhiễm) và một cá nhân chưa xem phim bằng S (dễ bị nhiễm). Ảnh hưởng xã hội là kết quả của một người quan sát bạn bè, gia đình, đồng nghiệp, v.v. đã xem phim. Nếu tỷ lệ các mối quan hệ xã hội trong phạm vi gần của một người đã xem phim tăng lên, thì người đó có nhiều khả năng bắt kịp xu hướng. Đối với một mối quan hệ xã hội nhất định giữa một cá nhân S và một cá nhân I , chúng tôi biểu thị tỷ lệ 'lây truyền' bằng β. Nghĩa là, trong một khoảng thời gian nhỏ dt, xác suất cá nhân S bị nhiễm do mối quan hệ xã hội này là βdt. Điều đáng chú ý là trên thực tế, có hai ngăn riêng biệt đều tác động đến một cá nhân S để xem phim. Một ngăn bao gồm các mối quan hệ xã hội đã xem phim và ngăn còn lại bao gồm những người đã quyết định xem phim nhưng vẫn chưa xem. Vì mọi người hiếm khi đến rạp một mình nên nhóm sau được cho là có đóng góp đáng kể vào ảnh hưởng xã hội đối với một cá nhân. Bởi vì tính đơn giản của mô hình là ưu tiên hàng đầu của chúng tôi trong nghiên cứu này, chúng tôi ước tính tình hình bằng cách giả định rằng ngăn I bao gồm cả hai nhóm dân số phụ được đề cập ở trên và chúng tôi giả định rằng β là hệ số truyền trung bình cho

Machine Translated by Google

cả hai loại. Nói cách khác, chúng ta cho rằng "đã quyết định xem phim" và "đã xem phim" có thể hoán đổi cho nhau, và chúng ta đang bỏ qua những trư ờng hợp đư a ra quyết định như ng cuối cùng không thực hiện.

Một yếu tố trung tâm khác của động lực là trí nhớ. Có một dòng phim mới liên tục đư ợc phát hành. Khả năng nhận thức của con ngư ời là có hạn. 'Sự phấn khích' tạo ra bởi một bộ phim mới phát hành chắc chắn giảm dần theo thời gian. Một lựa chọn đơ n giản để mô hình hóa trí nhớ con ngư ời là hàm mũ e^{-Bt}, trong đó $B > 0$ là hằng số[30, 31]. Giá trị B càng lớn thì thời gian tồn tại của bộ phim trong trí nhớ tập thể của xã hội càng ngắn. Ngoài ra, chúng tôi kỳ vọng thời gian tồn tại của bộ phim (số tuần rạp chiếu phim) sẽ giảm theo B. Lý do là việc tăng B khiến bộ phim bị lãng quên nhanh hơ n và các rạp chỉ tiếp tục chiếu phim cho đến khi doanh thu tạo ra bằng với chi phí.

Hãy biểu thị doanh thu phòng vé tích lũy tại thời điểm t bằng G(t), và để ρt biểu thị tỷ lệ cá nhân trong ngăn I tại thời điểm t. Sự tiến hóa theo thời gian của ρ(t) theo chế độ trư ờng trung bình đư ợc đư a ra bởi

$$\dot{\rho}t = (1 - \rho t)\rho t\beta k e^{-Bt}, \quad (1)$$

trong đó k là số lư ợng trung bình các mối quan hệ xã hội của các cá nhân. Ký hiệu βk/B bằng A và lư u ý rằng G(t) tỷ lệ thuận với ρt , chúng ta thu đư ợc biểu thức sau cho G(t):

$$G(t) = G(0)e^{A(1 - e^{-Bt})}. \quad (2)$$

Điều đáng chú ý là A mô hình hóa ảnh hư ởng xã hội, nghĩa là A mạnh hơ n có nghĩa là tác động mạnh hơ n từ môi trư ờng xã hội lên quyết định của ngư ời xem phim. Giá trị của A có thể tăng theo hai cách: tăng khả năng kết nối của mọi ngư ời (thông qua k) hoặc tăng tốc độ truyền trên mỗi liên kết.

Một cách tiếp cận thay thế để mô hình hóa ảnh hư ởng xã hội là tính đến áp lực của nhóm, thay vì ảnh hư ởng của cá nhân. Ví dụ, trong mô hình cử tri thông thư ờng (là mô hình cơ bản về động lực ý kiến với nhiều tài liệu nghiên cứu và mở rộng[34]), khả năng chuyển sang trạng thái này từ trạng thái khác đư ợc xác định bởi tỷ lệ tư ơ ng ứng của các trạng thái ở vùng lân cận của cá nhân. Trong cách tiếp cận dịch bệnh, chúng tôi giả định rằng mỗi mối quan hệ xã hội đều tác động đến một cá nhân, ảnh hư ởng này không phụ thuộc vào số lư ợng mối quan hệ xã hội mà ngư ời đó có. Một ngư ời S có ba ngư ời bạn I sẽ nhận đư ợc cùng một lư ợng ảnh hư ởng cho dù ngư ời này có 20 ngư ời bạn hay chỉ có ba ngư ời bạn đó. Mô hình cử tri có cách tiếp cận khác và xem xét tỷ lệ thay vì số lư ợng tuyệt đối. Điều này có nghĩa là ngư ời S có ba ngư ời bạn I sẽ nhận đư ợc ảnh hư ởng tỷ lệ thuận với 3/20 nếu ngư ời đó có 20 ngư ời bạn và tỷ lệ thuận với 3/3 nếu ngư ời đó có ba ngư ời bạn. Có thể thiết kế một mô hình kiểu cử tri thay thế cho vấn đề đang xét, dẫn đến cùng một kết quả cuối cùng theo phép xấp xỉ trư ờng trung bình.

Nếu cá nhân x có trạng thái S có kx bạn bè và Ix trong số họ có trạng thái I, thì xác suất cá nhân x sẽ thay đổi trạng thái của mình thành I đư ợc đư a ra bởi $\alpha dt(Ix/kx)e^{-Bt}$. Hệ số bộ nhớ giống hệt với kịch bản trư ớc, vì về mặt bộ nhớ không có sự khác biệt giữa hai trư ờng hợp. Hệ số α chỉ là tốc độ sửa đổi chiến lư ợc, tức là tần suất trung bình mà các cá nhân cân nhắc cập nhật trạng thái của họ. Trong phần phư ơ ng pháp, chúng tôi chỉ ra rằng trong trư ờng hợp này, Sự tiến hóa theo thời gian của ρ(t) theo chế độ trư ờng trung bình đư ợc đư a ra bởi

$$\dot{\rho}t = (1 - \rho t)\rho t\alpha e^{-Bt}, \quad (3)$$

giống hệt (1), chỉ có tên của hệ số đư ợc thay đổi. Dạng tham số của phư ơ ng trình cho G(t) sẽ giống hệt với dạng thu đư ợc ở trên.

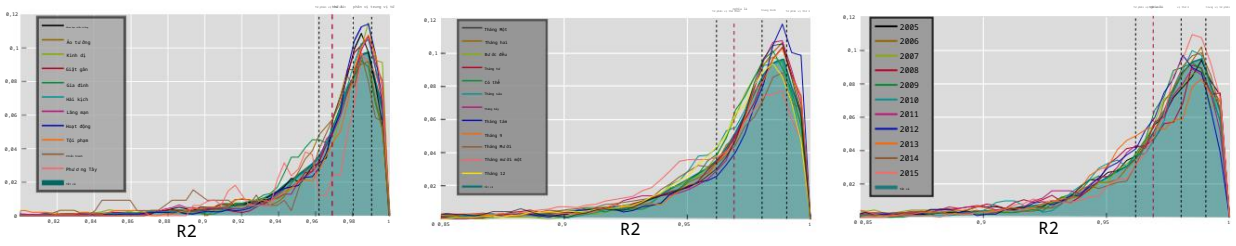
Machine Translated by Google

Ước tính tham số

Hãy xác định z(t) $\overset{\text{định nghĩa}}{=} \log G(t)/G(0)$. Để ước tính A và B trong (2) cho một bộ phim nhất định, về cơ bản, điều chúng ta cần làm là điều chỉnh mô hình sau cho phù hợp với dữ liệu chuỗi thời gian thực nghiệm về doanh số bán phim: $z(t) = A(1 - e^{-Bt})$. Việc giảm thiểu tổng bình phương lỗi (ký hiệu là E) thông qua các thuật toán thông thường như phương pháp giảm độ dốc hoặc phương pháp Newton sau đó trở nên đơn giản, đặc biệt là vì độ dốc và Hessian rất dễ tính toán:

$$E = \sum_t \left[2\left(1 - e^{-Bt}\right)z_t - A\left(1 - e^{-Bt}\right)\right]^2, \quad H = \sum_t \begin{bmatrix} 2te^{-2Bt} & 2te^{-Bt}(2Ae^{-Bt} - z_t) \\ 2te^{-Bt}(2Ae^{-Bt} - z_t) & 4At^2e^{-2Bt} - 2At^2e^{-Bt} + z_t t e^{-Bt} \end{bmatrix} \tag{4}$$

Nhưng để làm cho quy trình đề xuất khả thi hơn đối với lượng độc giả rộng hơn và nhấn mạnh thêm sức mạnh của mô hình, chúng tôi áp dụng một cách tiếp cận thậm chí còn đơn giản hơn. Với phép xấp xỉ, chúng tôi biến đổi bài toán hiện tại thành một bài toán hồi quy tuyến tính đơn giản. Chúng tôi đạt được sự đơn giản đáng kể với cái giá phải trả là mất đi một số độ chính xác, khá nhỏ (như sẽ được chứng minh bên dưới). Lưu ý rằng trong giới hạn khi $t \to \infty$, chúng ta có $z(t) \to A$. Điều này có nghĩa là nếu một bộ phim được phép chạy mãi mãi, z(t) sẽ tiến tới A. Là một phép xấp xỉ thời gian vô hạn, chúng tôi lấy doanh thu tuần trước của mỗi bộ phim làm ước tính của A. Biểu thị tuổi thọ của phim bằng L, điều này có nghĩa là chúng ta có thể ước tính A bằng z(L). Vì vậy, chúng ta cần ước tính B trong phương trình $z(t) = z(L)(1 - e^{-Bt})$, về cơ bản có nghĩa là với phép biến đổi y(t) $\overset{\text{định nghĩa}}{=} \log\left(1 - z(t)/z(L)\right)$, chúng ta có một bài toán hồi quy tuyến tính đơn giản trong đó y(t) là một hàm tuyến tính của t và B có thể dễ dàng ước tính cho từng phim. Biểu đồ histogram của hồi quy R^2 cho 5000 bộ phim trong tập dữ liệu được trình bày trong Hình 1 được phân tầng theo (a) thể loại, (b) tháng phát hành và (c) năm phát hành. Có thể dễ dàng thấy rằng sự gia tăng lỗi trong quá trình đơn giản hóa nói trên không đáng kể. Biểu đồ histogram ước tính cho A được trình bày trong Hình 2 theo phân tầng theo (a) thể loại, (b) tháng phát hành, (c) năm phát hành. Biểu đồ histogram cho B được trình bày trong Hình 3 với phân tầng tương tự.



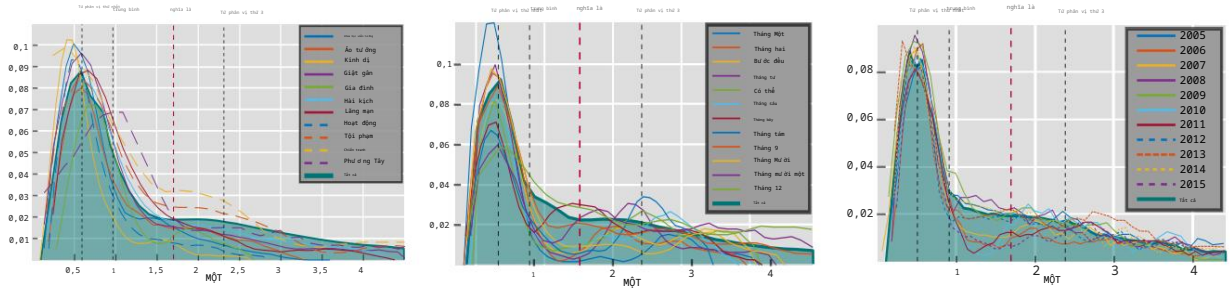
(Một) (b) (c)

Hình 1. Biểu đồ tần suất năm phát hành R đã điều ^{2} đối với phim được phân loại theo (a) thể loại, (b) tháng phát hành, (c) chỉnh.

Hình 4 mô tả A và B như một hàm số của số tuần phim vẫn chiếu tại rạp (L). Có thể thấy rằng A có mối quan hệ tích cực với L, và B có mối quan hệ tiêu cực, như mong đợi (vì tăng A—ảnh hưởng xã hội—có nghĩa là tăng doanh thu, và tăng B có nghĩa là suy giảm trí nhớ nhanh hơn).

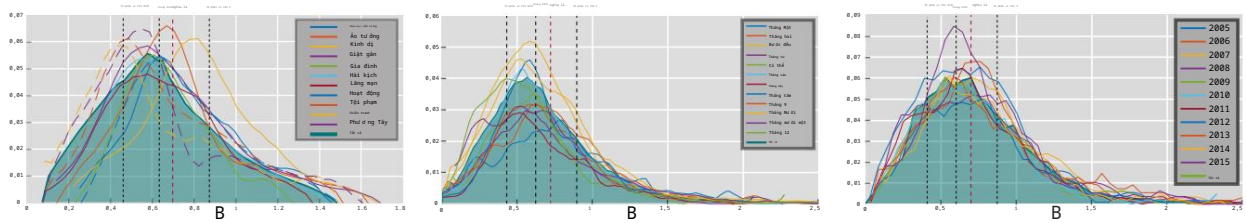
Biểu thị doanh thu ước tính từ quy trình trên bằng G(t), chúng ta có thể vẽ đồ thị G(t)/G(t) để xem dự đoán chính xác đến mức nào. Do thời gian sống khác nhau của các bộ phim khác nhau (tức là các giá trị L khác nhau do số tuần phim tồn tại trong rạp khác nhau), chúng ta vẽ đồ thị theo hàm của thời gian chuẩn hóa, được định nghĩa là t/L, để chúng ta có thể vẽ đồ thị tất cả các bộ phim trong cùng một khung hình. Hình 5a trình bày kết quả, được phân tầng theo

Machine Translated by Google



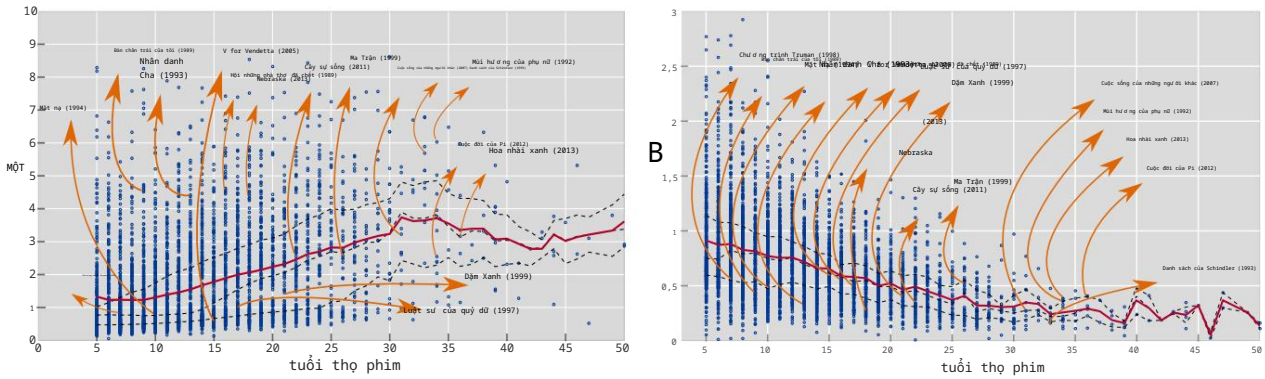
(Một) (b) (c)

Hình 2. Biểu đồ histogram của A (tham số của mô hình được đưa ra bởi phương trình (2)) cho các bộ phim được phân loại theo (a) thể loại, (b) tháng phát hành, (c) năm phát hành.



(Một) (b) (c)

Hình 3. Biểu đồ histogram của B (tham số của mô hình được đưa ra bởi phương trình (2)) cho các bộ phim được phân loại theo (a) thể loại, (b) tháng phát hành, (c) năm phát hành.



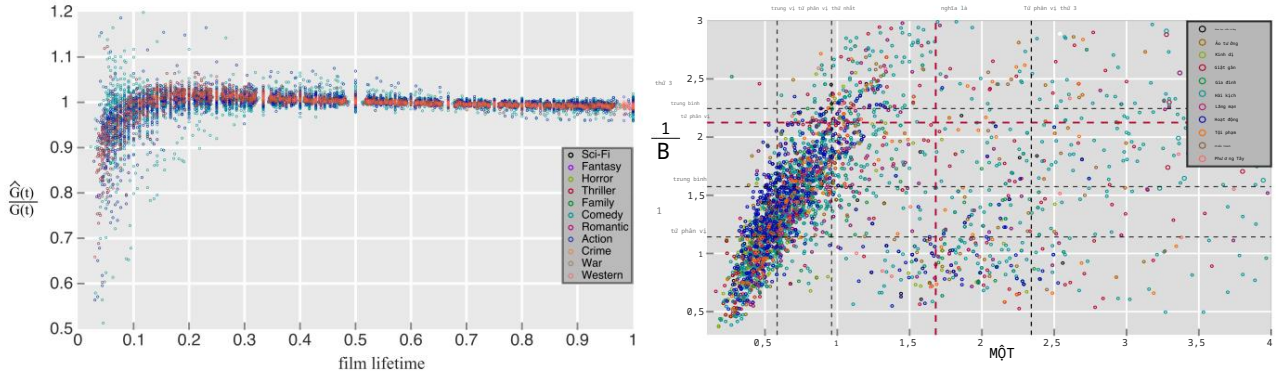
(Một) (b)

Hình 4. (a) và (b): Các tham số mô hình như một hàm của tuổi thọ màng phim, và (c): A như một hàm của $\frac{1}{B}$. Các đường màu đỏ liên quan đến giá trị trung bình, và các đường đứt nét liên quan đến phần trăm thứ 25, 50 và 75. Một số ví dụ phim được nêu ra với mục đích minh họa.

thể loại. Với sự đơn giản của mô hình, các lỗi khá thấp. Hơn nữa, không có sự phân biệt trực quan sự khác biệt giữa hiệu suất của các thể loại khác nhau.

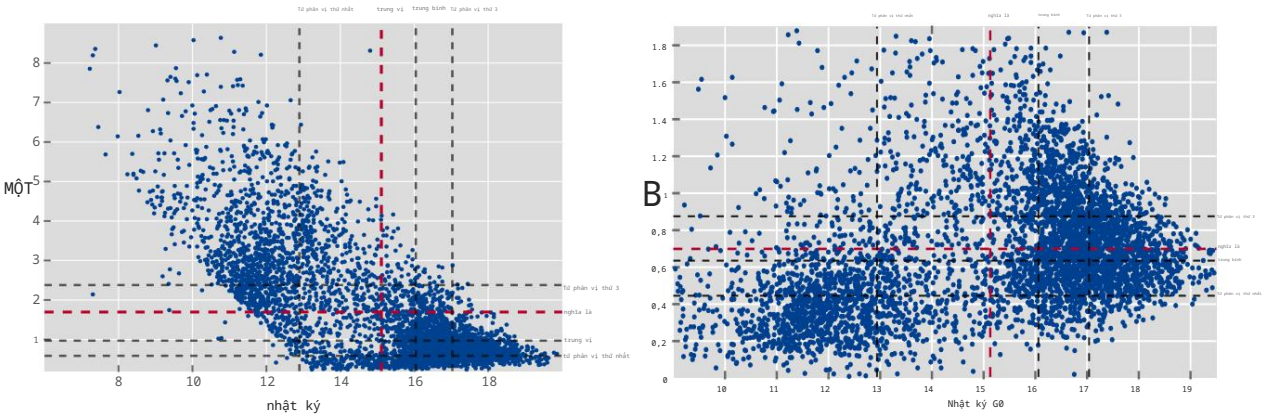
Chúng tôi cũng đã cố gắng đơn giản hóa mô hình hơn nữa bằng cách loại bỏ một trong các tham số bằng cách xấp xỉ nó như một hàm của tham số kia, để chúng tôi có một tham số thanh lịch. mô hình. Về mặt trực quan, chúng ta thấy rằng đối với một phần đáng kể của các bộ phim, mối quan hệ của A và 1/B có vẻ gần với tuyến tính, như được mô tả trong hình 5b. Tương quan của A và 1/B thực sự cao (ρ = 0,60). Nhưng

Machine Translated by Google



Hình 5. (a) tỷ lệ doanh thu dự kiến trên doanh thu thực tế cho tất cả các phim theo hàm thời gian chuẩn hóa, tức là t/L. (b) A theo hàm của $\frac{1}{B}$

vì mối quan hệ tuyến tính dường như chỉ đúng với một phần nhất định của dân số, thực sự là đa số nhưng không phải là mạnh, nên chúng tôi không thay thế hoàn toàn một trong các tham số theo tham số kia dựa trên mối quan hệ này. Chúng tôi cũng cố gắng đơn giản hóa như vậy bằng cách điều tra các mối liên kết có thể có giữa A và B với G0 (doanh thu phim tuần đầu tiên). Hình 6 trình bày các biểu đồ phân tán của A và B như một hàm của G0. Có một mối quan hệ tiêu cực rõ ràng giữa A và logG0 ($\rho = -0,59$ và $p < 10^{-5}$ đối với kiểm định t của giả thuyết liên kết tuyến tính). Cũng có một mối liên hệ tích cực nhỏ giữa B và logG0 ($\rho = 0,22$ và $p < 10^{-5}$ đối với kiểm định t của giả thuyết liên kết tuyến tính). Mối liên hệ này yếu vì B đang mô hình hóa trí nhớ của con người trong mô hình đơn giản của chúng tôi, do đó, nó không được mong đợi sẽ thay đổi nhiều với các tham số khác.

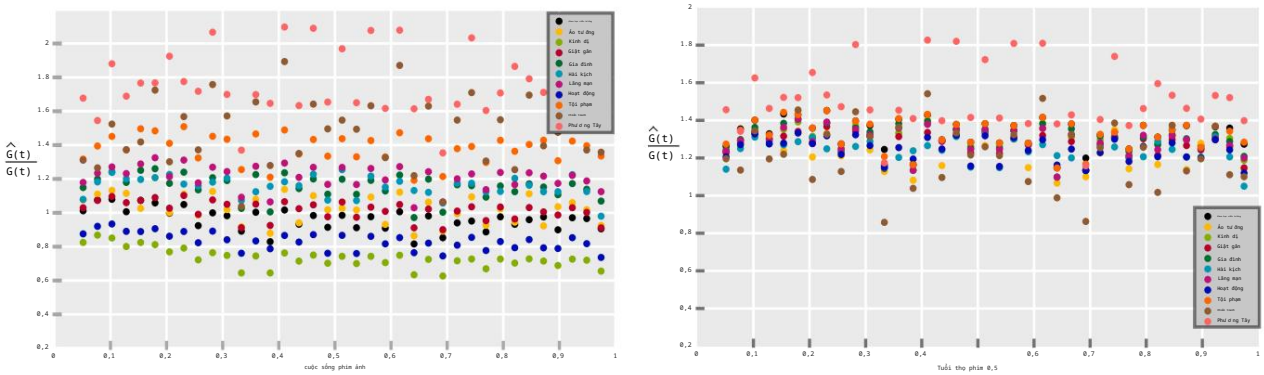


Hình 6. Các tham số của mô hình như một hàm của doanh thu trong tuần khai trương. Đường màu đỏ liên quan đến giá trị trung bình và các đường nét đứt liên quan đến phần trăm thứ 25, 50 và 75.

Như chúng tôi đã nhấn mạnh ở trên, trọng tâm chính của nghiên cứu này là thiết kế một mô hình cơ học liên kết các cơ chế vi mô của ảnh hưởng xã hội với các hiện tượng vĩ mô có thể quan sát được và dự đoán doanh số không phải là mục tiêu chính. Nhưng như một sản phẩm phụ, chúng ta có thể sử dụng các kết quả để thêm một chiều hướng dự đoán vào các kết quả. Chúng tôi đã ước tính một giá trị cho tham số A và một giá trị cho tham số B cho mỗi bộ phim. Để có được một công thức hoạt động như một 'hành vi trung bình' hợp lý cho mọi bộ phim, chúng ta có thể

Machine Translated by Google

thực hiện hai cách tiếp cận. Cách tiếp cận đầu tiên là gộp tất cả các ước tính A lại với nhau và tất cả các ước tính B lại với nhau, và lấy trung vị làm tham số cho mô hình dự đoán. Cách còn lại là đầu tiên phân tầng theo thể loại (hoặc bất kỳ thuộc tính nào khác, chúng tôi lấy thể loại ở đây làm ví dụ minh họa) và lấy trung vị của phân phối tham số chỉ dành cho phim trong thể loại đó để dự đoán doanh số bán phim trong từng thể loại. Hình 7 trình bày kết quả trong hai trường hợp này. Có thể thấy rằng quy trình gộp tạo ra kết quả tốt hơn so với phương pháp dành riêng cho thể loại. Cũng lưu ý rằng các phương pháp xác thực chéo thông thường không áp dụng được ở đây, vì khi ước tính các tham số cho từng phim, các phim khác chưa được sử dụng. Nói cách khác, các đơn vị dữ liệu của chúng tôi ở đây là chuỗi thời gian, không phải là điểm dữ liệu mà người ta sẽ khớp với đường cong. Vì vậy, xác thực chéo sẽ chỉ đơn giản là tác động vào quy trình lấy trung vị, không phải là quy trình học tham số.



(Một) (b)

Hình 7. Tỷ lệ giá trị dự đoán của G(t) so với giá trị thực nghiệm như một hàm của tuổi thọ phim được chuẩn hóa, sử dụng các giá trị trung bình trong phân phối của A và B trên (a) tất cả các phim được gộp lại với nhau, (b) chỉ những phim thuộc thể loại nhất định.

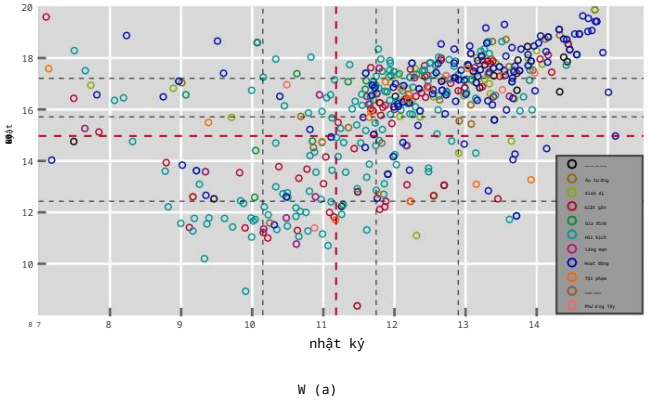
Chúng tôi cũng nghiên cứu xem liệu G0 có thể được thay thế bằng dữ liệu trước khi phát hành hay không. Như một ví dụ minh họa, chúng tôi sử dụng lượt truy cập trang Wikipedia (W) của các bộ phim trong tháng trước khi phát hành làm đại diện cho G0. Như Hình 8 trình bày, có một mối liên hệ tích cực giữa logG0 và logW (với ρ = 0,44). Dự đoán chính xác G(0) từ dữ liệu trước khi phát hành đã được thực hiện trong tài liệu với độ chính xác đáng kể (ví dụ: R sử dụng các $R^2 > 0,9$ lần chỉnh sửa và lượt truy cập trang Wikipedia27 và $R^2 > 0,95$ sử dụng các đề cập trên Twitter một ngày trước khi phát hành24). Vì vậy, đối với độc giả quan tâm chủ yếu đến dự đoán, có một cách để cải thiện khả năng dự đoán bằng cách kết hợp dữ liệu trước khi phát hành với các kết quả được trình bày trong bài báo này.

Cuộc thảo luận

Phân phối của A (được trình bày trong Hình 2) bị lệch. Theo tài liệu, doanh thu phim có phân phối đuôi nặng32, 33. Vì A lớn hơn có nghĩa là ảnh hưởng xã hội lớn hơn và điều này đến lượt nó có nghĩa là doanh thu nhiều hơn (và thời gian tồn tại lâu hơn trong rạp chiếu phim, như thể hiện trong Hình 4a), chúng tôi mong đợi rằng phân phối đuôi nặng như vậy cũng sẽ được phản ánh trong phân phối của A. Như chúng tôi đã thảo luận ở trên, B (đặc trưng cho sự suy giảm trí nhớ) mô hình hóa các đặc tính nội tại của hệ thống trí nhớ của con người và không được mong đợi là thay đổi nhiều giữa các bộ phim. Như Hình 3 minh họa, B khá cục bộ, phù hợp với kỳ vọng này.

Không giống như A, tham số B được kỳ vọng có mối quan hệ tiêu cực với doanh thu và do đó, với L (tuổi thọ phim). Điều này được xác nhận theo kinh nghiệm, như minh họa trong Hình 4b). Sự gia tăng dự kiến của L với A và sự giảm của nó với B đều có ý nghĩa thống kê (giả thuyết về mối quan hệ tuyến tính, chúng tôi có $p < 10^{-5}$ cho cả hai bài kiểm tra t).

Machine Translated by Google



Hình 8. Các tham số của mô hình như một hàm của doanh thu trong tuần khai trương

Như minh họa trong Hình 6, có một mối quan hệ tiêu cực rõ ràng giữa A và logG0 (ρ = 0,59 và p < 10 5 đối với kiểm định t của giả thuyết liên kết tuyến tính). Điều này được mong đợi vì G0 cao hơn có nghĩa là vai trò mạnh hơn cho các chiến dịch tiếp thị và ảnh hưởng trực tiếp mà các cá nhân nhận được khi gặp phải quảng cáo. Ví dụ, rất nhiều người đang chờ đợi riêng lẻ để phát hành bộ phim Star Wars: The Force Awakens (2016) gần đây, không chờ phản hồi của người khác để đưa ra quyết định.

Cũng có một mối liên hệ tích cực nhỏ giữa B và logG0 (ρ = 0,22 và p < 10 5 đối với kiểm định t của giả thuyết liên kết tuyến tính). Nguyên nhân của mối liên hệ tích cực này hoàn toàn là do cấu trúc. Điều đó không có nghĩa là những bộ phim có doanh thu ra mắt cao hơn sẽ bị lãng quên nhanh hơn. Nó chỉ phản ánh thực tế là năng lực thị trường điện ảnh có hạn. Nghĩa là, mọi người hoặc là 'người đi xem phim' hoặc không, và sự thay đổi đột ngột về thái độ đối với ngành công nghiệp điện ảnh không phải là điều phổ biến, đặc biệt là trong vòng đời của một bộ phim. Vì vậy , G0 cao hơn nhất thiết có nghĩa là bộ phim phải suy giảm nhanh hơn. Ví dụ, Finding Dory (2016) có doanh thu tuần đầu công chiếu là 231 triệu đô la và The Jungle Book (2016) có 130 triệu đô la, nhưng cả hai đều bán được 31 triệu đô la vào tuần thứ 4. Theo nghĩa này, chúng ta nói rằng bộ phim có G0 cao hơn phải suy giảm nhanh hơn.

Các dự đoán của mô hình tổng hợp chính xác hơn khi tất cả các phim được gộp lại (Hình 7a) so với khi sử dụng gộp theo thể loại cụ thể (Hình 7b). Độ chính xác khác nhau đối với các thể loại khác nhau. Cần lưu ý rằng các thể loại có nội dung tương tự nhau sẽ có thành tích tương tự nhau. Chiến tranh gần với phim Viễn Tây, Lãng mạn gần với Gia đình cũng như Hài kịch (chủ yếu là do các phim Hài lãng mạn đưa hai thể loại này lại gần nhau) và Khoa học viễn tưởng gần với Kỳ ảo. Ngoài ra, lưu ý rằng sự khác biệt về độ chính xác không phải do quy mô nhóm dân số khác nhau (tức là không phải trường hợp hiệu suất của Khoa học viễn tưởng tốt vì có nhiều phim Khoa học viễn tưởng trong tập dữ liệu, còn hiệu suất của Chiến tranh kém vì chỉ có một vài phim Chiến tranh), vì khi sử dụng gộp theo thể loại cụ thể (Hình 7b), các thành tích vẫn khác biệt. Điều này có nghĩa là, như người ta có thể trực quan mong đợi, tồn tại những khác biệt cố hữu giữa các bộ phim thuộc các thể loại khác nhau, đặc biệt là liên quan đến mô hình của chúng tôi.

Vì mô hình được trình bày trong bài báo này là tối thiểu và chúng tôi nhấn mạnh vào tính đơn giản của mô hình như là động lực chính của nghiên cứu này, nên có thể dễ dàng mở rộng mô hình hiện tại. Hưởng nào mang lại sự cải thiện nhiều nhất là một câu hỏi thú vị. Chúng tôi đã bỏ qua ảnh hưởng cá nhân nhận được từ các chiến dịch tiếp thị, có thể được thêm vào mô hình bằng một tham số mới. Theo cách đó, sẽ có hai nguồn ảnh hưởng riêng biệt. Một là xã hội, như đã đưa vào đây, và nguồn còn lại là cá nhân (thông qua tiếp xúc trực tiếp với quảng cáo trên phương tiện truyền thông, biển quảng cáo, v.v.). Nguồn sau sẽ thêm một tham số mới vào mô hình. Một phỏng đoán sẽ là khả năng dễ bị ảnh hưởng xã hội sẽ có mối tương quan cao với khả năng dễ bị tiếp thị trực tiếp, do đó, tham số mới có thể được xấp xỉ hợp lý như một hàm của các tham số đã tồn tại. Một cải tiến khác sẽ là sử dụng dữ liệu

Machine Translated by Google

từ phư ơ ng tiện truyền thông xã hội để kết hợp các tác động của vị trí mạng lư ới của cá nhân lên ảnh hư ởng xã hội của họ và các tác động của các đặc tính cấu trúc của các mạng lư ới xã hội cơ bản trong việc truyền bá các quyết định.

Phư ơ ng pháp

Phân tích mô hình

Chúng ta hãy biểu thị trạng thái của cá nhân x bằng sx, bằng 0 nếu cá nhân đó ở trạng thái S và bằng 1 nếu cá nhân đó ở trạng thái I. Đối với một mối quan hệ xã hội I cho trư ớc, xác suất rằng sau thời điểm dt, trạng thái sẽ đư ợc truyền cho cá nhân x đư ợc đư a ra bởi βdt, theo định nghĩa. Xác suất rằng sự lây truyền không xảy ra là 1 βdt. Xác suất mà cá nhân không bị lây nhiễm từ bất kỳ mối quan hệ xã hội hiện có nào là (1 βdt) ix xác suất mà cá nhân x sẽ bị lây nhiễm sau thời điểm dt là 1 (1 βdt) ix đến bậc một của dt đến βixdt., trong đó ix là số lư ợng các mối quan hệ xã hội của cá nhân x đang ở trạng thái I. Do đó, which Theo phép xấp xỉ trư ờng trung bình, trạng thái mong đợi của cá nhân x với tổng số kx mối ,có thể đư ợc mở rộng theo Taylor quan hệ xã hội sau thời điểm dt có thể đư ợc viết là E{sx(t +dt)} = sx + (1 sx)(βkxρt)dt. Lấy trung bình này trên tất cả các cá nhân, chúng ta có ρ˙t = ρt(1 ρt)βk, trong đó k là số lư ợng trung bình các mối quan hệ xã hội của các cá nhân. Nhân với hệ số trí nhớ và tích phân cả hai vế, chúng ta có Phư ơ ng trình (2).

Có thể suy ra các kết quả tư ơ ng tự cho mô hình thay thế kiểu cử tri đư ợc thảo luận trong văn bản. Trong trư ờng hợp này, chúng ta có E{sx(t +dt)} = sx + (1 sx)αdt(Ix/kx)e Bt, và nếu chúng ta cộng tổng này cho tất cả các nút, chúng ta sẽ có ρ˙t = ρt(1 ρt)α xy Axy/kx, trong đó Axy là ma trận kề của mạng xã hội cơ bản (bằng 1 nếu x đư ợc kết nối với y và bằng 0 nếu không). Nhân với hệ số bộ nhớ, tích của hai hệ số (α liên quan đến ảnh hư ởng xã hội và xyAxy/kx liên quan đến kết nối xã hội) có thể đư ợc hấp thụ thành một tham số mới duy nhất và sau khi tích hợp, chúng ta sẽ có đư ợc kết quả giống hệt với kết quả của mô hình dịch bệnh.

Mô tả dữ liệu Dữ liệu

phim Chúng tôi đã trích xuất bộ dữ liệu có sẵn công khai từ www.boxofficemojo.com. Chúng tôi giới hạn phân tích trong 5000 bộ phim có doanh thu cao nhất tại Hoa Kỳ. Chúng tôi loại trừ các bộ phim trư ớc năm 1980 để đảm bảo độ tin cậy của dữ liệu. Chúng tôi cũng loại trừ các phim imax không phải là phim truyện thông thư ờng và thời gian tồn tại của chúng dài bất thư ờng. Ví dụ, Space Station 3-D imax đư ợc phát hành vào năm 2002 và vẫn đang đư ợc chiếu. Chúng tôi giới hạn phân tích chỉ đối với các phim 'truyền thống'. Chúng tôi áp dụng ngư ỡng 70 tuần và loại bỏ các phim có thời gian tồn tại dài hơ n 70 tuần. Ngoài ra, chúng tôi loại trừ các phim có thời gian tồn tại ngắn hơ n 5 tuần. Chúng tôi cũng loại trừ các phim đư ợc phát hành vào năm 2016 để đảm bảo rằng không có phim nào trong tập dữ liệu vẫn còn trong rạp. Tập dữ liệu đã làm sạch với chuỗi thời gian bán hàng đi kèm với bài báo này.

Dữ liệu Wikipedia Chúng tôi đã trích xuất các lư ợt truy cập trang Wikipedia trong tháng phát hành từ bộ dữ liệu công khai do Wikimedia Foundation cung cấp. Mặc dù bộ dữ liệu lư ợt truy cập trang có từ năm 2007, chúng tôi đã loại trừ các bộ phim phát hành trư ớc năm 2010 để đảm bảo rằng chúng tôi đang xem xét giai đoạn mà Wikipedia đã trở nên phổ biến và đư ợc công chúng biết đến đủ để trở thành nguồn dữ liệu đáng tin cậy.

Tuyên bố đóng góp của tác giả

NM và BF xây dựng vấn đề. AT, NM và BF thực hiện nghiên cứu. AT, NM, BF và MR thảo luận về kết quả và đóng góp vào văn bản.

1 Tuyên bố về lợi ích tài chính cạnh tranh

Các tác giả tuyên bố không có xung đột lợi ích tài chính nào.

Machine Translated by Google

Tài liệu tham khảo

1. T. Skocpol, Các cuộc cách mạng xã hội trong thế giới hiện đại. Nhà xuất bản Đại học Cambridge, 1994.
2. T. Skocpol, Nhà nư ớc và cách mạng xã hội: Phân tích so sánh Pháp, Nga và Trung Quốc. Nhà xuất bản Đại học Cambridge, 1979.
3. H. Arendt, Eichmann ở Jerusalem. Penguin, 1963.
4. F. Neumann, H. Marcuse và O. Kirchheimer, Báo cáo bí mật về Đức Quốc xã: Frankfurt Đóng góp của trư ờng cho nỗ lực chiến tranh. Nhà xuất bản Đại học Princeton, 2013.
5. JE Stiglitz, Freefall: Nư ớc Mỹ, thị trư ờng tự do và sự sụp đổ của nền kinh tế thế giới. WW Norton & Công ty, 2010.
6. RG Rajan, Đư ờng đứt gãy: Những vết nứt ẩn giấu vẫn đe dọa nền kinh tế thế giới như thế nào. Đại học Princeton Báo chí, 2011.
7. S. Bikhchandani, D. Hirshleifer, và I. Welch, “Một lý thuyết về mốt nhất thời, thời trang, phong tục và thay đổi văn hóa như là chuỗi thông tin,” Tạp chí Kinh tế Chính trị, trang 992–1026, 1992.
8. F. Davis, Thời trang, văn hóa và bản sắc. Nhà xuất bản Đại học Chicago, 1994.
9. AV. Banerjee, “Một mô hình đơ n giản về hành vi bầy đàn,” Tạp chí Kinh tế Quý, tập 107, số 3, tr. 797 1992.
10. DS Scharfstein và JC Stein, “Hành vi bầy đàn và đầu tư ”, Tạp chí Kinh tế Hoa Kỳ tập 80, số 3, trang 465 1990. ,
11. PJ Peter, JC Olson và KG Grunert, Hành vi ngư ời tiêu dùng và chiến lư ợc tiếp thị. McGraw-Hill Luân Đôn, 1999.
12. MR Solomon, Hành vi của ngư ời tiêu dùng: Mua, sở hữu và tồn tại. Prentice Hall Engelwood Cliffs, New Jersey, 2014.
13. DJ Watts, Mọi thứ đều hiển nhiên:* Khi bạn biết câu trả lời. Crown Business, 2011.
14. J. Norton và FM Bass, “Mô hình lý thuyết khuếch tán về việc áp dụng và thay thế cho các thế hệ sản phẩm công nghệ cao tiếp theo”, Khoa học quản lý, tập 33, số 9, trang 1069–1086, 1987.
15. R. Neelamegham và P. Chintagunta, “Mô hình Bayesian để dự báo hiệu suất sản phẩm mới trên thị trư ờng trong nư ớc và quốc tế,” Marketing Science, tập 18, số 2, trang 115–136, 1999.
16. A. Ainslie, X. Dreze, và F. Zufryden, “Mô hình hóa vòng đời phim và thị phần,” Khoa học tiếp thị , tập 24, số 3, trang 508–517, 2005.
17. A. Elberse và J. Eliashberg, “Động lực cung và cầu đối với các sản phẩm phát hành tuần tự trên thị trư ờng quốc tế: Trư ờng hợp phim ảnh,” Marketing Science, tập 22, số 3, trang 329–354, 2003.
18. J. Eliashberg, A. Elberse và MA Leenders, “Ngành công nghiệp phim ảnh: Các vấn đề quan trọng trong thực tiễn, nghiên cứu hiện tại và các hư ớng nghiên cứu mới,” Khoa học tiếp thị, tập 25, số 6, trang 638–661 , 2006.
19. RE Krider và CB Weinberg, “Động lực cạnh tranh và sự ra đời của các sản phẩm mới: trò chơ i tính thời gian trong phim ảnh,” Tạp chí Nghiên cứu Tiếp thị, trang 1–15, 1998.

Machine Translated by Google

20. L. Einav, “Tính thời vụ trong ngành công nghiệp phim ảnh Hoa Kỳ,” Tạp chí kinh tế Rand, tập 38, số 1, trang 127–145, 2007.

21. SA Ravid, “Thông tin, phim bom tấn và các ngôi sao: Một nghiên cứu về ngành công nghiệp điện ảnh*,” Tạp chí Kinh doanh, tập 72, số 4, trang 463–492, 1999.

22. M. Joshi, D. Das, K. Gimpel và NA Smith, “Đánh giá phim và doanh thu: Một thí nghiệm về hồi quy văn bản”, trong Công nghệ ngôn ngữ của con ngư ời: Hội nghị thư ờng niên năm 2010 của Chi nhánh Bắc Mỹ của Hiệp hội ngôn ngữ học tính toán, trang 293–296, Hiệp hội ngôn ngữ học tính toán, 2010.

23. MDKN Smith, “Đánh giá phim và doanh thu: Một thí nghiệm về hồi quy văn bản,” Boston Globe, tập 461, số 154, tr. 116.

24. S. Asur và BA Huberman, “Dự đoán tư ơ ng lai bằng phư ơ ng tiện truyền thông xã hội,” trong Web Intelligence và Công nghệ tác nhân thông minh (WI-IAT), Hội nghị quốc tế IEEE/WIC/ACM năm 2010, tập 1, trang 492–499, IEEE, 2010.

25. KR Apala, M. Jose, S. Motnam, C.-C. Chan, KJ Liszka và F. de Gregorio, “Dự đoán doanh thu phòng vé của phim bằng phư ơ ng tiện truyền thông xã hội,” trong Advances in Social Networks Analysis and Mining (ASONAM), Hội nghị quốc tế IEEE/ACM năm 2013, tr. 1209–1214, IEEE, 2013.

26. W. Zhang và S. Skiena, “Cải thiện dự đoán doanh thu phim thông qua phân tích tin tức,” trong Biên bản Hội nghị chung quốc tế IEEE/WIC/ACM năm 2009 về Trí tuệ web và Công nghệ tác nhân thông minh - Tập 01, trang 301–304, IEEE Computer Society, 2009.

27. M. Mestyan, T. Yasseri, và J. Kert esz, “Dự đoán sớm về thành công phòng vé phim dựa trên dữ liệu lớn về hoạt động của Wikipedia,” PloS one, tập 8, số 8, tr. e71226, 2013.

28. S. Goel, JM Hofman, S. Lahaie, DM Pennock và DJ Watts, “Dự đoán hành vi của ngư ời tiêu dùng bằng tìm kiếm trên web,” Biên bản của Viện Hàn lâm Khoa học Quốc gia, tập 107, số 41, trang 17486–17490 , 2010.

29. R. Pastor-Satorras, C. Castellano, P. Van Mieghem, và A. Vespignani, “Các quá trình dịch bệnh ở “mạng lư ới phức hợp,” Tạp chí Vật lý hiện đại, tập 87, số 3, trang 925, 2015.

30. S. Sudman và NM Bradburn, “Ảnh hư ởng của thời gian và các yếu tố trí nhớ đến phản ứng trong các cuộc khảo sát,” Tạp chí Hiệp hội Thống kê Hoa Kỳ, tập 68, số 344, trang 805–815, 1973.

31. Z. Lu, S. Williamson và L. Kaufman, “Tuổi thọ hành vi của trí nhớ cảm giác thính giác của con ngư ời đư ợc dự đoán bằng các biện pháp sinh lý,” Khoa học , tập 258, trang 1668–1668, 1992.

32. S. Sinha và S. Raghavendra, “Phim bom tấn Hollywood và phân phối đuôi dài,” Tạp chí Vật lý Châu Â u B-Vật chất ngư ng tụ và Hệ thống phức tạp, tập 42, số 2, trang 293–296, 2004.

33. S. Sinha và RK Pan, “Phim bom tấn, bom tấn và phim kinh dị: Phân phối thu nhập của phim ảnh,” trong Kinh tế học về phân phối của cải, trang 43–47, Springer, 2005.

34. C. Castellano, S. Fortunato, và V. Loreto, “Vật lý thống kê về động lực xã hội,” Đánh giá về vật lý hiện đại, tập 81, số 2, trang 591 2009.