Phát hiện đánh giá thư rác với Mạng phù hợp bằng biểu đồ

Ao Li *
Alibaba Group
Hangzhou, China
jianzhen.la@alibaba-inc.com

Zhou Qin *
Tập đoàn Alibaba
Hàng Châu, Trung
Quốc qinzhou.qinzhou@alibaba-inc.com

Runshi Liu
Tập đoàn Alibaba
Hàng Châu, Trung
Quốc runninghi.lrs@alibaba-inc.com

Yiqun Yang
Alibaba Group
Bắc Kinh, Trung
Quốc yiqun.yyq@alibaba-inc.com

Dong Li
Alibaba Group
Hangzhou, China
shiping@taobao.com

TRỪU TƯỢNG

Khách hàng thực hiện rất nhiều đánh giá trên các trang web mua sắm trực tuyến mỗi ngày, ví dụ: Amazon và Taobao. Trong khi đó, các bài đánh giá ảnh hưởng đến quyết định mua hàng của khách hàng, thu hút rất nhiều người gửi thư rác nhằm vào những người mua hàng kém chất lượng. Xianyu, ứng dụng bán đồ cũ lớn nhất ở Trung Quốc, bị đánh giá spam. Hệ thống chống thư rác của Xianyu phải đối mặt với hai thách thức lớn: khả năng mở rộng dữ liệu và các hành động đối nghịch do những kẻ gửi thư rác thực hiện. Trong bài báo này, chúng tôi trình bày các giải pháp kỹ thuật để giải quyết những thách thức này. Chúng tôi đề xuất một phương pháp chống thư rác quy mô lớn dựa trên mạng phức hợp đồ thị (GCN) để phát hiện các quảng cáo spam tại Xianyu, có tên là mô hình Chống thư rác (GAS) dựa trên GCN. Trong mô hình này, một đồ thị không đồng nhất và một đồ thị neous đồng nhất được tích hợp để nắm bắt bối cảnh cục bộ và bối cảnh toàn cầu của một nhận xét. Thử nghiệm ngoại tuyến cho thấy rằng phương pháp được đề xuất vượt trội hơn so với mô hình cơ sở của chúng tôi, trong đó thông tin về các bài đánh giá, tính năng của người dùng và các mặt hàng đang được đánh giá được sử dụng. Hơn nữa, chúng tôi triển khai hệ thống của mình để xử lý dữ liệu quy mô hàng triệu hàng ngày tại Xianyu. Hiệu suất trực tuyến cũng thể hiện tính hiệu quả của phương pháp đề xuất.

1. GIỚI THIỆU

Đánh giá của các trang web mua sắm trực tuyến cung cấp thông tin có giá trị, chẳng hạn như chất lượng sản phẩm và dịch vụ hậu mãi. Những đánh giá này, ảnh hưởng trực tiếp đến quyết định mua hàng của khách hàng [19], đã trở thành mục tiêu cho những kẻ gửi thư rác công bố thông tin độc hại. Trong trường hợp của chúng tôi, Xianyu, ứng dụng hàng cũ lớn nhất ở Trung Quốc, tạo điều kiện bán hàng ngày hơn 200.000 sản phẩm và đạt được Tổng khối lượng hàng hóa (GMV) hàng năm hơn 13 tỷ đô la từ tháng 8 năm 2017 đến tháng 7 năm 2018, cũng đang bị spam đánh giá. Những đánh giá spam này cần phải được làm sạch vì chúng không chỉ làm suy yếu trải nghiệm của người dùng mà còn tạo ra một ổ cắm cho gian lận trên internet.

Đánh giá tại Xianyu khác với đánh giá trên các trang web thương mại điện tử khác ở một số khía cạnh. Ví dụ, tại Amazon hoặc Taobao, đánh giá thường được thực hiện bởi những khách hàng đã mua sản phẩm , do đó, hành động đánh giá thường xảy ra sau khi mua hàng. Ngược lại, người dùng không biết gì về chất lượng và mức giá thấp nhất có thể của hàng cũ. Do đó, các đánh giá tại Xianyu hoạt động như một công cụ giao tiếp cho người mua và người bán (ví dụ: truy vấn để biết thông tin chi tiết và giảm giá) và hành động đánh giá thường xảy ra trước khi mua hàng, như thể hiện trong Hình 1 và Hình 2. Vì vậy, thay vì xem xét, thuật ngữ nhận xét sẽ được sử dụng trong phần còn lại của tờ giấy để gạch chân những điều cần thiết sự khác biệt của các loại thư rác tại Xianyu. Nói chung, có hai loại nhận xét spam chính tại Xianyu: nhận xét thô tục và quảng cáo spam. Với thực tế là các quảng cáo spam chiếm phần lớn các bình luận spam, chúng tôi tập trung vào việc phát hiện các quảng cáo spam trong công việc này ,.

Những thách thức chính của việc phát hiện quảng cáo spam là:

- Khả năng mở rộng: Dữ liệu quy mô lớn của Xianyu với hơn 1 tỷ đồ cũ được xuất bản bởi hơn 10 triệu người dùng. • Hành động bất lợi: Tương tự như hầu hết các hệ thống kiểm soát rủi ro, hệ thống chống thư rác bị suy giảm hiệu suất theo các hành động bất lợi do những người gửi thư rác thực hiện.

Những kẻ gửi thư rác thường thực hiện hai thủ đoạn đối nghịch sau đây để phá vỡ hệ thống chống thư rác:

- Ngụy trang: Sử dụng các cách diễn đạt khác nhau với giá trị trung bình tương tự . Ví dụ: “Hãy quay số này để nhận công việc bán thời gian” và “Bạn muốn kiếm thêm tiền trong thời gian rảnh rỗi? Liên hệ với tôi ” đều là quảng cáo spam với mục đích hướng dẫn mọi người đến các hoạt động ngoại tuyến rủi ro.
- Làm biến dạng các bình luận: Những kẻ gửi thư rác thay thế một số từ chính trong các bình luận bằng các ký tự Trung Quốc hiếm khi được sử dụng hoặc cố tình mắc lỗi chính tả. Ví dụ: “Thêm vx của tôi", "Thêm v của tôi" và "Thêm wx của tôi" đều có nghĩa là "Thêm tài khoản WeChat1 của tôi ".

Những thủ thuật này mang lại một số bất tiện nhưng vẫn có thể hiểu được cho người đọc. Ngược lại, một thách thức lớn đối với hệ thống chống thư rác là nhận ra các mẫu khác nhau được thiết kế bởi những kẻ gửi thư rác.

Đồng thời, người ta nhận thấy rằng tác động của các đối thủ cạnh tranh có thể được giảm bớt bằng cách giới thiệu bối cảnh của các nhận xét. Chúng tôi xác định bối cảnh thành hai loại: bối cảnh địa phương và bối cảnh toàn cầu. Ngữ cảnh cục bộ đề cập đến thông tin từ pub lisher và mục có liên quan, trong khi ngữ cảnh toàn cầu đề cập đến thông tin được cung cấp bởi phân phối tính năng của tất cả các nhận xét.

Trong nghiên cứu này, chúng tôi trình bày một phương pháp chống thư rác có khả năng mở rộng cao dựa trên mạng phức hợp đồ thị (GCN), được gọi là phương pháp Chống thư rác dựa trên GCN (GAS).

Tóm lại, những đóng góp của tác phẩm được liệt kê dưới đây:

(1) Chúng tôi đề xuất thuật toán phát hiện spam đồ thị không đồng nhất dựa trên GCN, thuật toán này hoạt động trên biểu đồ lưỡng phân với các thuộc tính cạnh tại Xianyu. Mô hình được thiết kế ra ngoài đáng kể thực hiện mô hình cơ sở và có thể dễ dàng khái quát hóa

---

* Hai tác giả đầu tiên đóng góp như nhau.

1WeChat là ứng dụng nhắn tin và truyền thông xã hội lớn nhất ở Trung Quốc và là một trong những ứng dụng di động độc lập lớn nhất thế giới bởi người dùng hoạt động hàng tháng.

Hình 1: Ứng dụng Xianyu: trang chính (bên trái) và khu vực bình luận (bên phải). Khu vực bình luận được đánh dấu bằng dấu gạch ngang cung cấp công cụ giao tiếp chính của Xianyu.

đến thuật toán GCN không đồng nhất dựa trên meta-path [23] cho các đồ thị và ứng dụng không đồng nhất khác nhau.

(2) Bên cạnh biểu đồ không đồng nhất sử dụng bối cảnh địa phương của các nhận xét, chúng tôi sử dụng bối cảnh toàn cầu và đề xuất GAS, giúp cải thiện kết quả hơn nữa.

(3) Chúng tôi triển khai mô hình chống thư rác được đề xuất với khung Tensorflow phân tán để xử lý hàng triệu nhận xét hàng ngày tại Xianyu. Theo các thử nghiệm ngoại tuyến và đánh giá trực tuyến, hệ thống của chúng tôi xác định đáng kể nhiều nhận xét spam hơn và giảm bớt tác động của các hành động đối nghịch trong khi vẫn đáp ứng yêu cầu về hiệu quả.

Phần còn lại của bài báo được tổ chức như sau. Phần 2 liệt kê các công việc liên quan. Trong Phần 3, chúng tôi đã trình bày rõ hơn về mô hình GAS được đề xuất . Các thí nghiệm ngoại tuyến và trực tuyến được trình bày trong Phần 4. Chúng tôi giới thiệu việc triển khai và triển khai hệ thống tại Xianyu trong Phần 5. Công việc được tóm tắt trong Phần 6.

2 CÔNG VIỆC LIÊN QUAN

Hầu hết các phương pháp phát hiện spam hiện có đều tập trung vào việc trích xuất các tính năng được thiết kế mạnh mẽ từ nội dung đánh giá hoặc hành vi của người đánh giá. [7] đã nghiên cứu việc trùng lặp nội dung đánh giá để phát hiện các lượt xem lại spam. Họ thu thập các tính năng tập trung vào đánh giá, tập trung vào người đánh giá và trung tâm của sản phẩm, và đưa chúng vào mô hình hồi quy logistic. [17] chỉ tập trung vào nội dung của một bài phê bình. Các tác giả đã tiếp cận vấn đề bằng cách sử dụng ba chiến lược như là các tính năng trong Naive Bayes và trình phân loại SVM. [13] đã tóm tắt các tính năng của chuyên gia miền cho khai thác opin ion, sau đó một tập hợp các tính năng được thiết kế phức tạp được sử dụng cho nhiệm vụ phân loại đánh giá. Các phương pháp tập trung vào tính năng này bỏ qua mối quan hệ giữa người đánh giá, hàng hóa và nhận xét. Tuy nhiên, dựa trên quan sát của chúng tôi, các mối quan hệ cũng đóng một vai trò quan trọng trong việc phát hiện thư rác. Ví dụ, các quảng cáo spam thường được xuất bản bởi những người gửi spam trong các nhóm.

Dựa trên những quan sát tương tự, một số học giả bắt đầu sử dụng thông tin biểu đồ. Phương pháp phát hiện thư rác dựa trên biểu đồ đầu tiên đã được trình bày trong [26]. Họ đã xây dựng "biểu đồ đánh giá" với ba loại nút - người đánh giá, cửa hàng và bài đánh giá. Sau đó, cung cấp lại sự tin cậy của người đánh giá, lưu trữ độ tin cậy và tính trung thực của đánh giá theo cách giống như HITS [10] . Liang và cộng sự [15] Sử dụng hai biểu đồ: một là biểu đồ không đồng nhất được đề cập ở trên, một là biểu đồ thể hiện mối quan hệ hỗ trợ hoặc xung đột giữa những người đánh giá. Soliman [22] đã đề xuất một kỹ thuật dựa trên đồ thị mới để phát hiện thư rác bằng cách sử dụng phân nhóm đồ thị trên một đồ thị tương tự của người dùng được xây dựng để mã hóa các mẫu hành vi của người dùng trong cấu trúc liên kết của nó. Khung NetSpam [21] đã xác định các loại đường dẫn meta khác nhau trên đồ thị đánh giá và sử dụng chúng trong phân loại.

Những năm gần đây đã chứng kiến mối quan tâm ngày càng tăng trong việc phát triển các thuật toán dựa trên học sâu trên đồ thị, bao gồm các phương pháp không giám sát [5, 12, 18] và phương pháp có giám sát [6, 9, 11, 25]. Một trong những tiến bộ nổi bật nhất được biết đến là GCN [9], trong đó các tính năng của các nút được tổng hợp từ các vùng lân cận địa phương. Toán tử "tích chập đồ thị" được định nghĩa là tập hợp đặc trưng của các lân cận một bước. Thông qua sự tích lũy lặp đi lặp lại, thông tin truyền đi nhiều bước trong biểu đồ. GCN đạt được những cải tiến đáng kể so với các phương pháp khai thác đồ thị trước đây như DeepWalk [18]. Sau đó, rất nhiều nhà nghiên cứu đã tham gia vào lĩnh vực này. William và cộng sự [6] đề xuất GraphSAGE, một khuôn khổ quy nạp tận dụng kỹ thuật lấy mẫu nút và tổng hợp tính năng để tạo hiệu quả nhúng nút cho dữ liệu không nhìn thấy, giúp phá vỡ giới hạn của việc áp dụng GCN trong cài đặt chuyển đổi. Mạng lưới chú ý đồ thị (GAT) [25] kết hợp cơ chế chú ý vào GCN. Bằng cách tính toán hệ số chú ý giữa các nút, GAT cho phép mỗi nút tập trung vào những người hàng xóm có liên quan nhất để đưa ra quyết định.

Hầu hết các phương pháp đồ thị tập trung vào đồ thị đồng nhất, trong khi trong nhiều ứng dụng trong thế giới thực, dữ liệu có thể bị coi là đồ thị không đồng nhất một cách tự nhiên. Các biểu đồ không đồng nhất hiếm khi được nghiên cứu trước đây và ngày nay thu hút sự quan tâm ngày càng tăng. EAGCN [20] tính toán nhúng nút không đồng nhất bằng cách sử dụng cơ chế chú ý. Mô hình này tập trung vào trường hợp nhiều loại liên kết kết nối các nút trong một đồ thị. Tác giả đề xuất sử dụng "đa chú ý" - mỗi hàm chú ý chỉ xem xét các bors lân cận được xác định bởi một loại liên kết cụ thể. Tương tự, GEM [14] tập trung vào trường hợp có nhiều loại nút. Tác giả đề xuất một cơ chế chú ý để tìm hiểu tầm quan trọng của các loại nút khác nhau. Cụ thể, họ chia đồ thị thành các đồ thị con theo các loại nút và tính toán đóng góp của mỗi đồ thị con cho toàn hệ thống dưới dạng hệ số chú ý.

Phương pháp biểu đồ đã được áp dụng trong nhiều lĩnh vực, ví dụ, hệ thống gợi ý [4, 27, 30, 31], dự đoán đặc tính hóa học [20], chăm sóc sức khỏe [2], phát hiện tài khoản độc hại [14] , v.v. Trong bài báo này, phương pháp dựa trên GCN lần đầu tiên được áp dụng cho vấn đề phát hiện xem xét thư rác, theo hiểu biết tốt nhất của chúng tôi.

3 PHƯƠNG PHÁP ĐỀ XUẤT

Trong phần này, trước tiên, chúng tôi trình bày nội dung sơ bộ của các mạng cấu trúc đồ thị, sau đó chúng tôi minh họa vấn đề chống thư rác tại Xianyu. Cuối cùng, chúng tôi sẽ chứng minh phương pháp GAS của chúng tôi theo hai cách : đầu tiên chúng tôi giới thiệu cách mở rộng thuật toán GCN cho đồ thị khác nhau và sau đó minh họa GAS bằng cách kết hợp thêm bối cảnh toàn cầu.

3.1 Sơ bộ

Công việc trước đây [6, 9, 25] tập trung chủ yếu vào đồ thị thuần nhất. Gọi G = (V, E) là một đồ thị thuần nhất với nút v V, cạnh R d0 với v V trong đó của nút). Trạng thái ẩn của nút = h d0 (v, v biểu thị chiều đặc trưng

v được học bởi lớp thứ l của mô hình được ký hiệu là h biểu thị R dl , dl thứ nguyên của trạng thái ẩn tại lớp thứ l.

Các phương pháp dựa trên GCN tuân theo một người đàn ông truyền bá theo lớp khôn ngoan . Trong mỗi lớp lan truyền, tất cả các nút cập nhật đồng thời. Như đã tóm tắt trong [28, 29], một lớp lan truyền có thể được tách thành hai lớp con: tập hợp và kết hợp. Nói chung, đối với GCN có các lớp L , các lớp con tổng hợp và tổ hợp ở lớp thứ l (l = 1, 2, · · · L) có thể được viết là:

$$h_v^l = \sigma\left(W^l \cdot \mathrm{AGG}\left(\{h_{v'}^{l-1}, \forall v' \in N(v)\}\right)\right) \quad (1)$$

$$h_v^l = \mathrm{COMBINE}\left(h_v^{l-1}, h_{N(v)}^l\right) \quad (2)$$

trong đó N (v) là tập hợp các nút liền kề với v, AGG là một hàm được sử dụng để nhúng tổng hợp từ các hàng xóm của nút v, hàm này có thể được tùy chỉnh bằng các mô hình cụ thể, ví dụ: gộp tối đa, gộp trung bình [6] hoặc chú ý tổng kết dựa trên trọng số [25]. W là một ma trận có thể đào tạo được chia sẻ giữa tất cả các nút tại layerl. σ là một phi tuyến tính biểu thị fea tổng hợp chức năng kích hoạt, ví dụ, Relu. $h_{N(v)}^l$ N (v) mật độ lân cận của nút v tại lớp thứ l. Hàm COMBIN E được sử dụng để kết hợp tự nhúng và nhúng tổng hợp của các lân cận, đây cũng là một thiết lập tùy chỉnh cho các mô hình đồ thị khác nhau, ví dụ như nối như trong GraphSAGE [6].

Trong GCN [9] và GAT [25], không có lớp con kết hợp rõ ràng nào. Thông tin bản thân của v được giới thiệu bằng cách thay thế N (v) bằng N˜ (v) trong phương trình (1), trong đó N˜ (v) = v N (v). Do đó bước COMBIN E thực sự xảy ra bên trong bước AGG.

3.2 Thiết lập vấn đề Mục

đích của chúng tôi là xác định các bình luận spam tại Xianyu, có thể được xây dựng cho một bài toán phân loại cạnh trên biểu đồ lưỡng phân có hướng với các nút và cạnh được phân bổ.

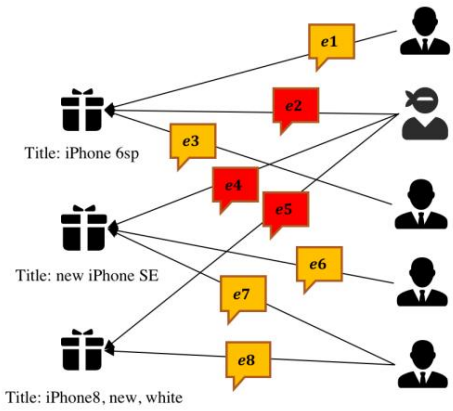
Các nhận xét về Xianyu có thể được biểu diễn một cách tự nhiên dưới dạng đồ thị lưỡng phân. G (U, I, E) trong đó U là tập các nút người dùng (đỉnh), I là tập các nút mục (đỉnh) và E là tập các chú thích (các cạnh).

Một cạnh e E từ người dùng u U đến một mục i I tồn tại nếu u đưa ra nhận xét e với i. Ngoài ra, với một đỉnh v I , U , N (v) là tập các đỉnh trong các lân cận một bước của nút v, tức là N (v) = {v I U | (v, v) E}. Trong trường hợp đồ thị lưỡng phân của Xianyu, N (i) U và N (u) I. E (v) biểu thị các cạnh được kết nối với v. Gọi U (e) và I (e) biểu thị nút người dùng và nút mục của cạnh e. Đồ thị lưỡng phân này được đặt tên là Đồ thị Xianyu. Xem Hình 2 để biết ví dụ từ thực tế.

3.3 Mạng chuyển đổi đồ thị không đồng nhất trên đồ thị Xianyu

Như đã giới thiệu trong Phần 3.1, trong nhiệm vụ phân loại nút dựa trên GCN trên đồ thị đồng nhất, việc nhúng nút từ lớp cuối cùng được sử dụng làm đầu vào của bộ phân loại.

Thay vào đó, chúng tôi sử dụng phép nhúng cạnh từ lớp lan truyền cuối cùng cùng với việc nhúng hai nút mà cạnh này liên kết với. Chúng tôi kết hợp ba cách nhúng này để phân loại cạnh



e1: more discount?
e2: get an iPhone 8plus, $130, contact me: #1
e3: hello, I want it
e4: get an iPhone 8plus for $130, contact my wechat #1
e5: get an iPhone 8plus for $130, contact my wechat #1
e6: how large is the storage space ?
e7: hi, is it still on sell?
e8: is it a complete new one?

Hình 2: Bản thu nhỏ của Đồ thị Xianyu. Trong cài đặt này, spam mer muốn đánh lừa người mua đến các giao dịch ngoại tuyến, vì vậy anh ta đã đăng một bình luận bắt mắt nói rằng anh ta có một chiếc điện thoại rẻ hơn để bán với nhiều mặt hàng liên quan. I, E, U lần lượt đại diện cho các nút mục, các chú thích, các nút người dùng. Ở đây # 1 là viết tắt của ID tài khoản WeChat cụ thể.

nhiệm vụ như thể hiện trong Hình 3, trong đó ze , zu và zi biểu thị cạnh, người và nhúng mục, tức là ze = $h_{U(e)}^L$ e , zu = $h_{I(i)}^L$ U (e) dùng và zi = h_i^L I (i) .

Chúng tôi sẽ trình bày cụ thể cách điều chỉnh GCN tiêu chuẩn cho biểu đồ lưỡng phân với các thuộc tính cạnh. Điểm mấu chốt là giảm thiểu lớp con tổng hợp và lớp con kết hợp trong phương trình (1) và phương trình (2).

3.3.1 Lớp con tổng hợp. Lớp con tổng hợp trong GCN xử lý tất cả các loại nút như nhau và bỏ qua các thuộc tính cạnh. Để phù hợp với khuôn khổ chung Phương trình (1) với Đồ thị Xianyu, ba hàm tổng hợp cho từng loại thực thể (người dùng, mục, nhận xét) được xác định.

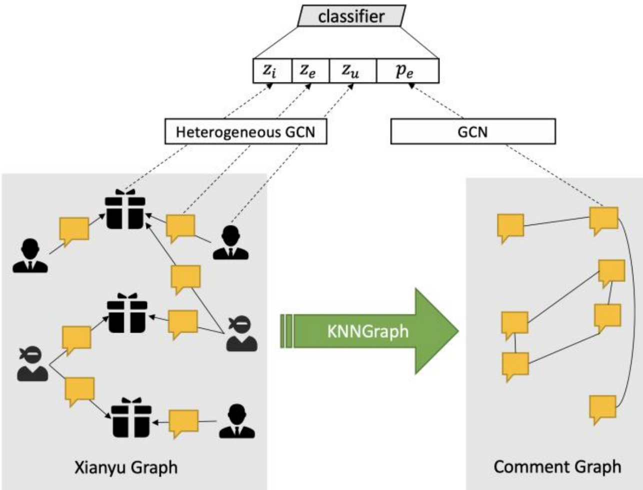
Đối với một nhận xét, tức là một cạnh, trạng thái ẩn được cập nhật dưới dạng nối các trạng thái ẩn trước đó của chính cạnh đó và hai nút mà nó liên kết đến. Vì vậy, lớp con tổng hợp được định nghĩa là

$$h_e^l = \sigma\left(W_E^l \cdot \mathrm{AGG}_E^l\left(h_e^{l-1}, h_{U(e)}^{l-1}, h_{I(e)}^{l-1}\right)\right) \quad (3)$$

ở đâu

$$\mathrm{AGG}_E^l\left(h_e^{l-1}, h_{U(e)}^{l-1}, h_{I(e)}^{l-1}\right) = \mathrm{concat}\left(h_e^{l-1}, h_{U(e)}^{l-1}, h_{I(e)}^{l-1}\right) \quad (4)$$

Đối với nút người dùng u U và nút mục i I, ngoài thông tin từ các nút lân cận, các thuộc tính của các cạnh được kết nối với chúng cũng được thu thập. Hàng xóm tổng hợp nhúng $h_{N(u)}^l$



Hình 3: Hình minh họa GAS kết hợp hai mô hình đồ thị. GCN không đồng nhất hoạt động trên Đồ thị Xianyu cung cấp cho người dùng, mục, nhúng nhận xét, tương ứng. GCN hoạt động cung cấptrênhĐồnthịhbìnhgthhúnthúchzi , ze neous đồng nhất

Trong GAS, các phần nhúng này được nối với nhau như một phần trong bộ phân loại: y = classi f ier (concat (zi , zu, ze , pe)).

l và h được tính là
N (i)

$$h^l_{N(u)} = \sigma\left(W^l_U \cdot AGG^l_U\left(H^{l-1}_{IE}\right)\right)$$
$$h^l_{N(i)} = \sigma\left(W^l_I \cdot AGG^l_I\left(H^{l-1}_{UE}\right)\right) \quad (5)$$

ở đâu

$$H^{l-1}_{IE} = \{concat(h^{l-1}_i, h^{l-1}_e), \forall e = (u, i) \in E(u)\}$$
$$H^{l-1}_{UE} = \{concat(h^{l-1}_u, h^{l-1}_e), \forall e = (u, i) \in E(i)\} \quad (6)$$

Hai loại nút duy trì các tham số khác nhau (W và các hàm tổng hợp W^l_U, W^l_I) khác nhau (AGG^l_U, AGG^l_I).

Đối với các dạng cụ thể của AGG^l_U và AGG^l_I chúng tôi điều chỉnh cơ chế chú ý:

$$AGG^l_U\left(H^{l-1}_{IE}\right) = ATT_U\left(h^{l-1}_u, H^{l-1}_{IE}\right)$$
$$AGG^l_I\left(H^{l-1}_{UE}\right) = ATT_I\left(h^{l-1}_i, H^{l-1}_{UE}\right) \quad (7)$$

ATT N ở đây là một hàm f : hkey × Hval hval ánh xạ một vectơ đặc trưng hkey và tập hợp các vectơ đặc trưng của ứng cử viên Hval thành tổng trọng số của các phần tử trong Hval . Trọng số của tổng, tức là giá trị chú ý được tính bằng chú ý của sản phẩm được chia tỷ lệ [24].

3.3.2 Lớp con kết hợp. Sau khi tổng hợp thông tin của những người hàng xóm , chúng tôi thực hiện theo một chiến lược kết hợp trong [6] cho người dùng và các nút mục dưới dạng

$$h^l_u = concat\left(V^l_U \cdot h^{l-1}_u, h^l_{N(u)}\right)$$
$$h^l_i = concat\left(V^l_I \cdot h^{l-1}_i, h^l_{N(i)}\right) \quad (8)$$

trong đó V^l_U và V^l_I biểu thị ma trận trọng lượng có thể đào tạo cho nút người dùng và mục, h^{l-1}_u và h^{l-1}_i là trạng thái ẩn của người dùng và mục i ở lớp thứ l - 1.

Toàn bộ thuật toán được mô tả trong Thuật toán 1. Lưu ý rằng phương pháp này thực sự có thể được tổng quát hóa thành thuật toán mạng phức hợp đồ thị không đồng nhất dựa trên siêu đường dẫn cho các đồ thị không đồng nhất khác nhau với các thuộc tính cạnh. Về chi tiết, cho một meta đường dẫn P ở dạng $A_0 \xrightarrow{R_1} A_1 \xrightarrow{R_2} \cdots \xrightarrow{R_L} A_L$, trong đó A_l tương ứng với kiểu nút và R_l là loại cạnh thứ l trên P, tập hợp và kết hợp cho một nút v thuộc loại A_l trên P. Đối với một nút v thuộc loại, một quá trình kết hợp có thể được viết như sau:

$$H^{l-1}_{A_{l-1}E} = \{concat(h^{l-1}_{v'}, h^{l-1}_e), \forall e = (v, v') \in E_{R_{l-1}}(v)\}$$
$$h^l_{N_{A_{l-1}}(v)} = \sigma\left(W^l_{A_l} \cdot AGG^l_{A_l}\left(H^{l-1}_{A_{l-1}E}\right)\right) \quad (9)$$
$$h^l_v = concat\left(V^l_{A_l} \cdot h^{l-1}_v, h^l_{N_{A_{l-1}}(v)}\right)$$

trong đó $E_{R_{l-1}}(v)$ biểu thị các cạnh liên kết với v với loại cạnh R_{l-1} và $N_{A_{l-1}}(v)$ biểu thị các nút lân cận của v với loại nút A_{l-1}. Trong của chúng tôi Trường hợp E và cài đặt 2 lớp, hai đường dẫn meta ngầm định là $U \xrightarrow{E} I \xrightarrow{E} U$ và $I \xrightarrow{E} U \xrightarrow{E} I$.

3.3.3 Chiến lược lấy mẫu liên quan đến thời gian. Với lớp con tổng hợp và lớp con kết hợp được đề xuất, có thể tiến hành chiến lược huấn luyện toàn bộ hoặc chiến lược huấn luyện theo lô nhỏ.

Việc đào tạo toàn bộ lô, cần cập nhật tất cả các thực thể trong một lần lặp, là không thực tế đối với dữ liệu lớn do thời gian hạn chế . Xem xét quy mô của Đồ thị Xianyu, chiến lược đào tạo theo lô nhỏ là phù hợp hơn. Đối với mỗi nút mục / người dùng, chúng tôi lấy mẫu một số lượng hàng xóm cố định để tạo thành một ma trận cung cấp hàng loạt nhỏ như [30] . Khác với chiến lược lấy mẫu ngẫu nhiên của họ, chúng tôi tận dụng thông tin thời gian và đề xuất lấy mẫu liên quan đến thời gian như thể hiện trong Hình 4.

Chúng tôi tóm tắt chiến lược lấy mẫu như sau:

- Khi số lượng thí sinh nhiều hơn số mẫu, tức là M, ta chọn M ý kiến gần nhất về mặt thời gian.
- Khi số lượng ứng cử viên ít hơn M, chúng tôi chèn chúng bằng trình giữ chỗ và bỏ qua tất cả các tính toán liên quan đến các trình giữ chỗ này.

Chiến lược lấy mẫu của chúng tôi hợp lý hơn lấy mẫu ngẫu nhiên ở hai khía cạnh. Đầu tiên, việc chọn các nhận xét gần nhất sẽ hợp lý hơn là lấy mẫu con ngẫu nhiên vì các nhận xét gần nhất có liên quan nhiều hơn đến nhận xét được xác định. Trong khi đó, padding hợp lý hơn là resampling vì các bình luận được đăng dưới hàng hóa cũ thường rất thưa thớt. Padding tránh thay đổi phân phối vùng lân cận so với lấy mẫu lại. Bằng cách này, chúng tôi đạt được kết quả có thể so sánh được với M nhỏ , do đó tiết kiệm thời gian đào tạo cũng như giảm tiêu thụ bộ nhớ.

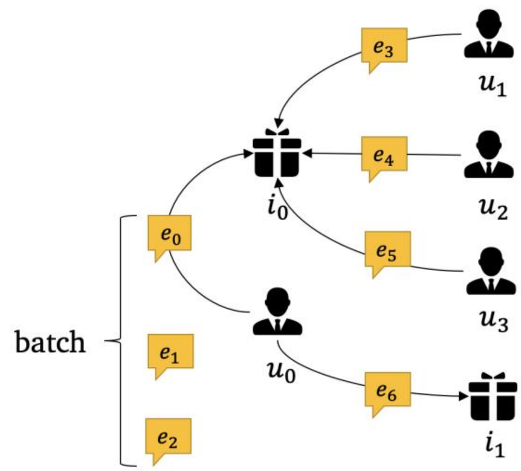
Thuật toán 1: Mạng chuyển đổi đồ thị không đồng nhất trên đồ thị Xianyu.

Đầu vào: Tập hợp các cạnh Eb E, số lớp L, chức năng
U (Eb) và I (Eb) ánh xạ Eb tới các nút người dùng và các nút mục
Eb được liên kết tương ứng. Đồ thị Xianyu G (U, I, E)

Đầu ra: Các trạng thái ẩn của lớp thứ L, bao gồm các trạng thái ẩn
của các cạnh: ze , e dùng Eb , zu các trạng thái ẩn của người dùng
Tôi (Eb)
zi ,

. bắt đầu
E l Eb ;
U l U (Eb);
Tôi l I (Eb);
// Lấy mẫu ; cho
l = L, · · · 1 , 1 làm
U l 1 l U;
Tôi l 1 l 1 I l ;
l với u U do l
U l 1 U l 1 LẤY MẪU (N (u));
chấm dứt
tôi vì tôi mà tôi làm
Tôi l 1 l 1 l 1 I LẤY MẪU (N (i));
chấm dứt
chấm dứt

// Đi qua mạng nơron; cho l = 1, · · · , Tôi làm
l cho e E do
tôi e l = σ W E · AGGl E (h e l 1 l 1 l , h, h); U (e) I E)
chấm dứt
l cho u U làm
Hl-1 I E n concat (h l 1 l 1 l , h e), e = (u, i) E (u) o ;
ll h σ W N (u) Bạn · AGGl U (Hl I E l);
là bạn concat V ll U u h = l · , h l N (u) ;
chấm dứt
tôi vì tôi mà tôi làm
Hl-1 UE n concat (h l 1 l 1 l , h u e), e = (u, i) E (i) o ;
ll h σ W N (i) · AGGl Tôi (Hl UE l);
Tôi sẽ 1 h · h = concat V i l , h N (i) ;
chấm dứt
chấm dứt
vì e Eb làm
ze = h e L
zu = h; U (e)
L zi = h; I E)
chấm dứt
chấm dứt



Hình 4: Chiến lược lấy mẫu liên quan đến thời gian. Giả sử {e0, e1, e2} tạo thành một loạt các nhận xét được xác định. Để cập nhật bộ chăn ga gối đệm e0, trước tiên cần tính toán các lần nhúng của mục i0 và u0 của người dùng. Không mất tính tổng quát, giả sử chúng ta đặt số lượng mẫu tối đa M = 2. Đối với mặt hàng, 2 xu có thời gian xuất bản gần nhất với thời gian xuất bản của e0 sẽ được chọn từ {e3, e4, e5}, giả sử e3 , e4 được chọn, sau đó {e3, u1} và {e4, u2} sẽ được tổng hợp thành i0. Tương tự, đối với phía người dùng, ví dụ: {e6, i1} với trình giữ chỗ có đệm sẽ được chọn để tổng hợp thành u0.

3.3.4 Kết hợp Mạng Đồ thị với Mô hình Phân loại Văn bản.

Văn bản trong nhận xét phải được chuyển đổi thành một bản nhúng trước khi được hợp nhất với các tính năng của người dùng và mục. TextCNN [8] là một mô hình phân loại văn bản thỏa đáng cân bằng giữa hiệu lực và hiệu quả. Do đó, chúng tôi sử dụng mô hình TextCNN để nhúng nhận xét và tích hợp nó vào mô hình mạng nơron đồ thị của chúng tôi như một khung phân loại end-to-end.

Cụ thể, chúng tôi sử dụng tính năng nhúng từ được đào tạo trước bởi word2vec [16] làm đầu vào của TextCNN. Đầu ra của TextCNN sau đó được sử dụng làm phần nhúng của nhận xét. Chi tiết,

$$h_e^0 = \text{TextCNN}(w0, w1, w2, \cdots, wn) \quad (10)$$

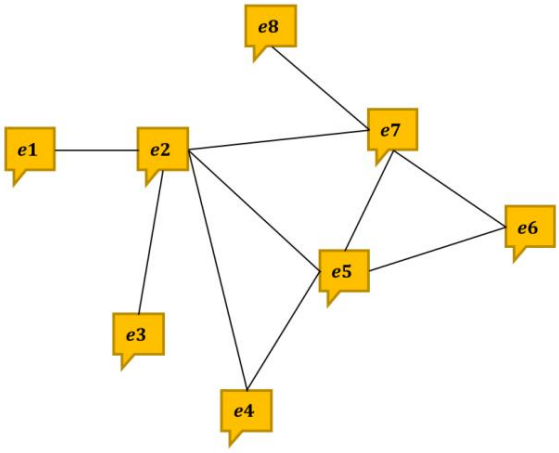
trong đó wi đại diện cho phép nhúng từ của từ nhận xét thứ i e, 0 và h là lần nhúng đầu tiên của nhận xét e trong phương trình (3). Do đó , các tham số của TextCNN được huấn luyện cùng với các tham số khác trong mô hình được mô tả trong Phần 3.3.

3.4 Mô hình Chống thư rác dựa trên GCN Các bình luận

spam tại Xianyu thường bị biến dạng bởi những người gửi thư rác độc hại như một biện pháp đối phó với hệ thống phát hiện thư rác của chúng tôi. Ví dụ, các lỗi chính tả và viết tắt có chủ ý được sử dụng để phá vỡ sự dò xét của chúng tôi . Những thư rác này có tác động nhỏ đến việc con người đọc được nhưng thường gây nhầm lẫn cho mô hình NLP của chúng ta. Đặc biệt là khi thư rác được đăng bởi những người dùng khác nhau và dưới các mục khác nhau. Rõ ràng, bối cảnh địa phương không thể giúp đỡ trong tình huống này. Theo trực giác, đối với loại thư rác này, chúng tôi muốn tìm một số thông tin bổ sung từ toàn bộ biểu đồ như “Có bao nhiêu nhận xét tương tự trên toàn bộ trang web và ý nghĩa của chúng là gì?”.

Người ta nhận thấy rằng ngay cả khi việc tăng số lượng lớp lan truyền giúp các nút nắm bắt được bối cảnh toàn cầu, thì tiếng ồn được đưa vào vẫn không thể bị bỏ qua, như các nhà nghiên cứu khác đã báo cáo [9, 11]. Các thí nghiệm trong Phần 4.1 cho thấy rằng hiệu suất khó có thể được hưởng lợi từ việc tăng số lượng lớp lan truyền trong trường hợp của chúng tôi.

Do đó, chúng tôi sử dụng một cách tắt để nắm bắt ngữ cảnh toàn cầu của các nút. Cụ thể hơn, chúng tôi xây dựng một đồ thị thuần nhất có tên là Đồ thị Nhận xét bằng cách kết nối các nhận xét có nội dung tương tự. Bằng cách này, các cạnh (chú thích) trong Đồ thị Xianyu không đồng nhất bây giờ trở thành các đỉnh trong Đồ thị Chú thích. Xem Hình 5 để biết trường hợp thực tế của Đồ thị Nhận xét.



e1: for more certified products, please add #2
e2: for more styles, please add #3
e3: need more styles and discounts, please add v #4
e4: more styles please add wecha #5
e5: people who like it please add my wechat, different styles
e6: if you like it, leave me a message, add v
e7: I have others styles, if you like it please add: #6
e8: anything you like? please add #5

Hình 5: Bản thu nhỏ của Biểu đồ nhận xét, trong đó "wechat", "wecha" và "v" đều có nghĩa là tài khoản WeChat và # 2, # 3, # 4, # 5, # 6 đại diện cho các ID tài khoản khác nhau, tất cả các nhận xét trong tiểu đoạn này là những quảng cáo spam.

Như đã trình bày trong [11], GCN trên đồ thị đồng nhất có thể được xem như một dạng đặc biệt của phép làm mịn Laplacian. Các nhiệm vụ phân loại nút có thể được hưởng lợi từ lý thuyết này dựa trên giả định rằng các nút có cùng nhãn thường được nhóm lại với nhau trong biểu đồ, các tính năng của các nút có thể được làm mịn bởi các nút lân cận của nó, do đó làm cho nhiệm vụ phân loại dễ dàng hơn. Do đó, một thuật toán GCN quy nạp [6] được thực hiện trên Đồ thị chú thích thuần nhất đã xây dựng để học cách nhúng chú thích.

Bằng cách kết hợp GCN quy nạp hoạt động trên Đồ thị nhận xét với GCN không đồng nhất hoạt động trên Đồ thị Xianyu được mô tả trong Phần 3.3, chúng tôi đề xuất một thuật toán có tên là Chống thư rác (GAS) dựa trên GCN để đào tạo mô hình theo cách thức end-to-end . Xem Hình 3 để biết toàn bộ cấu trúc của GAS.

Phép nhúng nhận xét của e E được GCN học được từ Đồ thị Com ment được ký hiệu là pe . Nhúng cuối cùng của GAS là nối pe và các cách nhúng khác được học từ Đồ thị Xianyu ,

$$y = classi\ f\ ier\ (concat\ (zi\ ,\ zu,\ ze\ ,\ pe\ )), \tag{11}$$

trong đó ze , zu và zi lần lượt biểu thị các phép nhúng của e, U (e) và I (e) được học theo mô hình GCN không đồng nhất được đề xuất.

Một vấn đề không nhỏ cần được thảo luận là làm thế nào để tạo Biểu đồ Nhận xét, cụ thể là, làm thế nào để xác định các nhận xét tương tự. Điều này có thể được thực hiện một cách đơn giản bằng cách quét tất cả các nhận xét, tìm k đồng nghiệp gần nhất của mỗi nhận xét. Tuy nhiên, điều này là không thực tế $O(|E|^2)$ đối với độ phức tạp thời gian . Trong thực tế, chúng tôi sử dụng thuật toán Đồ thị KNN gần đúng [3] để xây dựng một đồ thị dựa trên K lân cận gần nhất của các nút.

Về chi tiết, Biểu đồ Nhận xét được xây dựng như sau:

- Loại bỏ tất cả các bình luận trùng lặp để tránh giải pháp nhỏ nhặt, tức là hai bình luận có cùng nội dung luôn giống nhau nhất. • Tạo các nhúng nhận xét bằng phương pháp được mô tả trong 1].
- Thu được các cặp nhận xét tương tự bằng cách sử dụng thuật toán Đồ thị KNN gần đúng. • Xóa các cặp nhận xét được đăng bởi cùng một người dùng hoặc được đăng không cùng mục, vì bối cảnh địa phương đã được xem xét trên Đồ thị Xianyu.

Theo cách này, chúng tôi giả định rằng các đánh giá thư rác khác nhau có thể được làm trơn tru bằng cách tích hợp các tính năng của những người hàng xóm của họ. Hình ảnh trực quan của một tập hợp con các mẫu huấn luyện được cung cấp để cho thấy rằng các nhận xét có thể phân tách rõ ràng hơn sau quá trình suôn sẻ, xem Hình 6 để biết chi tiết.

Hình 6: Hình dung không gian mẫu. Các điểm màu xanh lá cây đại diện cho các bình luận không phải là spam và các điểm màu đỏ thể hiện số lượng thư rác. Bên trái: nhận xét ban đầu nhúng trực tiếp vào không gian 2-D bởi PCA; Đúng: nhúng com ment được làm mịn (bằng cách lấy trung bình các tính năng của bors tự và hàng xóm) được PCA chiếu vào không gian 2-D.

Tính năng mô hình	Điểm AUC	F1
Nhúng thô	0.9342	0,8332
Nhúng mịn	0.9373	0,8448

Bảng 1: Phân tích định lượng ảnh hưởng của việc nhúng làm mịn tập hợp con các mẫu huấn luyện. Nó cho thấy rõ ràng rằng các mẫu có nhúng được làm mịn sẽ phù hợp hơn cho nhiệm vụ của chúng tôi.

Một phân tích định lượng cũng được thực hiện để chứng minh rằng các nhận xét có thể tách rời hơn sau quá trình suôn sẻ. Hai mô hình hồi quy logistic được đào tạo và thử nghiệm trên nhúng thô và nhúng mịn trong Hình 6. AUC và F1-score được báo cáo trong Bảng 1. Kết quả chứng minh rằng khả năng phân tách tuyến tính của các mẫu được cải thiện sau quá trình trơn tru. Cải tiến này chỉ ra rằng bộ phân loại được xây dựng dựa trên tính năng nhúng được làm mịn sẽ hoạt động tốt hơn.

4 THÍ NGHIỆM

Trong phần này, chúng tôi đánh giá hiệu suất của phương pháp đề xuất của chúng tôi trên tập dữ liệu Xianyu thực sự. Đầu tiên, chúng tôi so sánh phương pháp của mình với một số mô hình sử dụng tập dữ liệu ngoại tuyến, sau đó báo cáo hiệu suất trực tuyến trên ứng dụng Xianyu. Cuối cùng, chúng tôi trình bày một số trường hợp trong thế giới thực để cung cấp cái nhìn sâu sắc về các phương pháp được đề xuất.

4.1 Đánh giá ngoại tuyến

4.1.1 Tập dữ liệu. Để đánh giá phương pháp được đề xuất tại Xianyu, chúng tôi xây dựng một biểu đồ không đồng nhất với tổng thể 37.323.039 nhận xét do 9.158.512 người dùng đăng trên 25.107.228 mục trong một khoảng thời gian. Chúng tôi thu thập 1.725.438 nhận xét thường xuyên cùng với 74.213 nhận xét spam được đánh dấu bởi các chuyên gia về con người.

Tập hợp đào tạo, xác nhận và kiểm tra được chia ngẫu nhiên với tỷ lệ 6: 1: 3.

4.1.2 Phương pháp so sánh. Để so sánh phương pháp của chúng tôi với phương pháp khai thác đánh giá theo từng bậc, chúng tôi làm theo hướng dẫn trong [13]. Cụ thể, chúng tôi thiết kế rất nhiều tính năng làm thủ công cho nhận xét, mục và người dùng (ví dụ: độ dài nhận xét, cho dù đó là nhận xét đầu tiên của mục, liệu đó có phải là nhận xét duy nhất của mục hay không, tính tương đồng của nhận xét và các tính năng của mục, giá mặt hàng, số lượng bình luận của người dùng, v.v.). Để mã hóa nội dung bình luận, chúng tôi tính toán thông tin lẫn nhau của từng từ. 200 từ hàng đầu có giá trị thông tin lẫn nhau lớn nhất được chọn ra và sau đó được sử dụng để tạo vectơ giá trị nhị phân cho mỗi nhận xét. Mỗi thực thể của vectơ này cho biết liệu một từ có xuất hiện hay không. Cuối cùng, chúng tôi kết hợp các tính năng này làm đầu vào của mô hình GBDT. Chúng tôi gọi mô hình này là GBDT dưới dạng viết tắt.

Thay vì mã hóa giá trị nhị phân dựa trên thông tin lẫn nhau, mô hình TextCNN cũng được sử dụng để trích xuất nhúng nhận xét. Sau đó, nhúng nhận xét được nối với các tính năng của người dùng và mục được mô tả ở trên như là đầu vào của mô hình MLP (Multilayer Perceptron) 2 lớp. Đây là mô hình được triển khai trực tuyến tại Xianyu, do đó nó được coi là mô hình cơ sở, được đặt tên là TextCNN + MLP.

Để chứng minh tính hiệu quả của bối cảnh toàn cầu do Đồ thị chú thích giới thiệu, chúng tôi cũng so sánh mô hình chỉ sử dụng Đồ thị Xianyu bối cảnh cục bộ như trong Phần 3.3. Chúng tôi gọi mô hình này là GAS-local.

Tóm lại, cấu hình thử nghiệm được trình bày chi tiết bên dưới:

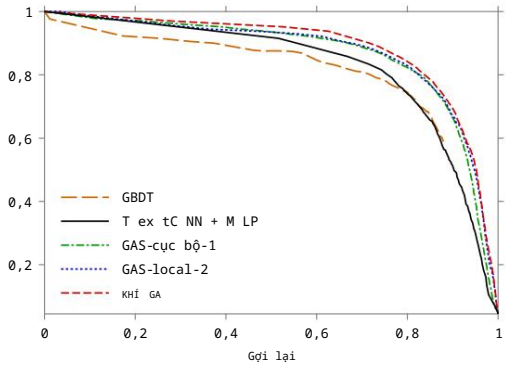
- GBDT: Các tính năng của chuyên gia tên miền với mô hình GBDT.
- TextCNN + MLP (đường cơ sở): TextCNN + tính năng người dùng + mục tính năng với mô hình MLP 2 lớp.
- Lớp lan truyền GAS-local-1: 1 trên Đồ thị Xianyu (tức là 1-hop người hàng xóm).
- GAS-local-2: 2 lớp lan truyền trên Đồ thị Xianyu.
- GAS: Mô hình GAS với 2 lớp lan truyền trên Đồ thị Xianyu và 1 lớp lan truyền trên Đồ thị Nhận xét.

Mô hình GBDT được đào tạo sử dụng 100 cây với tỷ lệ học tập là 0,05. Đối với các mô hình khác, cấu trúc TextCNN được sử dụng bởi tất cả các phương pháp đều có chung siêu tham số, ví dụ: kích thước bộ lọc được đặt thành {3, 4, 5} với 128 bộ lọc cho mỗi bộ. Tất cả các phương pháp ngoại trừ GBDT được đào tạo trong 8 kỷ nguyên. Số lượng mẫu tối đa M là 16 cho Đồ thị Xianyu và số lượng mẫu tối đa cho Đồ thị Nhận xét được đặt thành 64. Mô hình TextCNN + MLP được đào tạo với chương trình Tensorflow độc lập và tốc độ học là 0,001, trong khi các phương pháp được đề xuất là tất cả đều được đào tạo theo cách phân tán với 8 GPU. Tốc độ học tập được đặt thành 0,005 và kích thước lô được đặt thành 128.

4.1.3 Phân tích kết quả. Chúng tôi đánh giá các phương pháp này theo AUC, F1-score và thu hồi với độ chính xác 90%. Tỷ lệ thu hồi số liệu ở 90% được chọn trước vì các đánh giá spam được phát hiện sẽ được xử lý trong thực tế. Chúng tôi phải đảm bảo độ chính xác của mô hình cao để tránh làm phiền người dùng bình thường. Trong trường hợp của chúng tôi, độ chính xác trên 90% là điều kiện thiết yếu để triển khai một mô hình. vì vậy việc thu hồi ở độ chính xác 90% trở thành một tiêu chí thiết yếu để so sánh các mô hình khác nhau.

phương pháp	Điểm AUC	F1 thu hồi @ độ chính xác 90% 0,9649
GBDT	0,7686 50,55%	
Văn bảnCNN + MLP 0,9750 0,7784	GAS-	54,86%
local-1 GAS-local-2	0,9806 0,8138	66,90%
	0,9860 0,8143	67,02%
KHÍ GA	0,9872 0,8217	71,02%

Bảng 2: So sánh kết quả của các thí nghiệm ngoại tuyến về AUC, điểm F1 và thu hồi ở độ chính xác 90% , được ký hiệu là độ chính xác @ 90%.



Hình 7: Biểu đồ PR của đánh giá ngoại tuyến. GAS, GAS-local 1 và GAS-local-2 hoạt động tốt hơn đáng kể và GAS còn cải thiện hơn nữa so với GAS-local-2.

Kết quả được thể hiện trong Bảng 2 và các đường cong PR được hiển thị trong Hình 7. Chúng ta có thể thấy rằng GAS-local-1, GAS-local-2 và GAS hoạt động tốt hơn mô hình GBDT và TextCNN + MLP trong tập dữ liệu của chúng tôi. Điều này chứng tỏ tính ưu việt của các phương pháp đề xuất. So sánh giữa GAS-local-1 và TextCNN + MLP cho thấy rằng hiệu suất tăng đáng kể là do sự giới thiệu của bối cảnh địa phương. Độ chính xác thu hồi @ 90% được cải thiện từ 54,86%

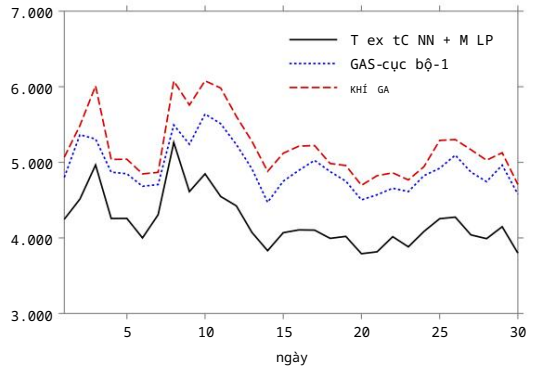
lên 66,90%, nghĩa là có thể phát hiện thêm 12,04% thư rác và bị xóa khỏi ứng dụng. Khi so sánh GAS-local-2 và GAS local-1, sự cải thiện không nổi bật, điều này cho thấy nâng cao hiệu suất bằng cách kết hợp bối cảnh địa phương 1-hop chiếm ưu thế. Nó có thể là do bên cạnh những thông tin được giới thiệu bởi Bối cảnh địa phương 2-hop, tiếng ồn cũng được giới thiệu. Như nhiều tác giả đã đã báo cáo [9, 11], hiệu suất tăng giảm đáng kể khi bước nhảy tăng lên. Khi so sánh GAS và GAS-local-2, chúng ta có thể thấy một sự cải tiến hơn nữa chứng tỏ hiệu quả của kết hợp bối cảnh toàn cầu. Khi độ chính xác được cố định thành 90%, so với GAS-local-2, chúng tôi phát hiện thêm 4% nhận xét spam.

Nhìn chung, phương pháp đề xuất của chúng tôi hoạt động tốt hơn so với phương pháp cơ sở đã triển khai hệ thống với thang điểm 4,33 F1. Điều quan trọng, dưới sự xử lý cố định ngưỡng chính xác 90%, phương pháp của chúng tôi mang lại thêm hiệu quả 16,16%.

4.2 Hiệu suất trực tuyến

Chúng tôi tiến hành các thử nghiệm trực tuyến trên nền tảng của mình như được giới thiệu trong Phần 5. Mục tiêu của thử nghiệm là so sánh số các nhận xét spam được phát hiện bởi các kiểu máy khác nhau với độ chính xác 90% được kiểm tra bởi các chuyên gia con người. Các bình luận spam được phát hiện sẽ sạch sẽ trong sản xuất.

Chúng tôi triển khai TextCNN + MLP, GAS-local-1 và GAS trong môi trường đấu giá chuyên nghiệp hàng ngày của mình và so sánh hiệu suất của chúng. Như mô tả trong Hình 8, GAS-local-1 và GAS hoạt động tốt hơn TextCNN + MLP nhất quán về số lượng bình luận spam được phát hiện. Trên mặt khác, GAS liên tục vượt trội hơn GAS-local-1, vốn chứng minh thêm rằng hệ thống được hưởng lợi từ văn bản lừa đảo toàn cầu do Biểu đồ nhận xét giới thiệu.



Hình 8: Kết quả đánh giá trực tuyến. Các meth ods được đề xuất luôn hoạt động tốt hơn GBDT và TextCNN + MLP mô hình tại Xianyu.

4.3 Nghiên cứu điển hình

Các nhận xét spam được phát hiện bằng các phương pháp khác nhau là thủ công đã kiểm tra.

4.3.1 GAS-local-1 so với TextCNN + MLP. Đầu tiên chúng ta so sánh cái sai mẫu phủ định của TextCNN + MLP (được ký hiệu là TextCNN + MLPF N) và các mẫu trong TextCNN + MLPF N được GAS-local-1 thu hồi (được ghi chú là TextCNN + MLPF N ∩GAS-local-1T P. Kết quả được hiển thị trong Bàn số 3.

mẫu	#spams trung bình trong vòng 1 bước lên Đồ thị Xianyu
TextCNN + MLPF N	2,60
TextCNN + MLPF N ∩GAS-local-1T P	3,24

Bảng 3: So sánh TextCNN + MLP và GAS-local 1. Rõ ràng, các mẫu được GAS-local-1 thu hồi từ con địa phương có thêm "hàng xóm spam" trong TextCNN + MLPF N chữ.

Chúng tôi phân tích các mẫu thư rác bổ sung được bao gồm bởi GAS-local-1, phát hiện rằng chúng hầu hết là các quảng cáo tương tự được xuất bản bởi cùng một người hoặc dưới cùng một mục. Ví dụ: một lời đề cập điển hình về spam com là "kiểm tra ảnh hồ sơ của tôi để xem có bất ngờ không", trong đó spam thông tin được ẩn trong ảnh hồ sơ (Thông tin hình ảnh ảnh hồ sơ không được sử dụng ở đây vì ảnh hồ sơ không có trong nhận xét và chi phí thời gian xử lý hình ảnh cao. Thông tin hình ảnh sẽ được giới thiệu trong tác phẩm sau này). Chỉ những quảng cáo này không chứa từ khóa cụ thể và không được TextCNN + MLP công nhận. Nhưng GAS-local-1 đã xác định chính xác những quảng cáo này bằng cách liên kết nhận xét với nhận xét được xuất bản bởi người dùng này.

4.3.2 GAS so với GAS-local-1. Tương tự như vậy, chúng tôi so sánh các mẫu nega tive sai của GAS-local-1 (được ký hiệu là GAS-local-1F N) và các mẫu ở GAS-local-1F N được GAS thu hồi (ký hiệu là GAS-local 1F N ∩GAST P. Kết quả được thể hiện trong Bảng 4.

mẫu	#spams trung bình trong vòng 1 bước lên Đồ thị Xianyu	#spams trung bình trong vòng 1 bước lên Biểu đồ nhận xét
GAS-cục bộ-1F N	2,23	17,23
GAS-cục bộ-1F N ∩GAST P	3.53	36,68

Bảng 4: So sánh GAS và GAS-local-1. Rõ ràng, các mẫu do GAS thu hồi từ GAS-local-1F N có nhiều hơn "Hàng xóm spam" trong bối cảnh toàn cầu.

Chi tiết, chúng tôi phân tích kết quả và phát hiện ra rằng hai loại thư rác nhận xét ủng hộ GAS hơn so với GAS-local-1:

- quảng cáo đối địch được xuất bản bởi những kẻ gửi thư rác Loại quảng cáo spam này bị biến dạng bởi những kẻ gửi thư rác với ý nghĩa tương tự (xem Bảng 5 để biết ví dụ điển hình). Vì hầu hết các đánh giá spam này không được xuất bản bởi cùng một tài khoản hoặc dưới cùng một mục, nhưng được kết nối với nhau trên Đồ thị Nhận xét. Với việc thải bỏ cố định ngưỡng chính xác 90%, chúng không được phát hiện bởi GAS local-1 nhưng bị GAS bắt được, lợi dụng của Biểu đồ bình luận giới thiệu bối cảnh toàn cầu. Này thư rác có thể được xuất bản bởi một nhóm người gửi thư rác, họ có thể thu thập nhiều tài khoản và xuất bản một số quảng cáo sử dụng từng tài khoản.

• tin nhắn phiếu thưởng được xuất bản bởi những người dùng khác nhau
Loại tin nhắn phiếu giảm giá này nhằm mục đích dẫn mọi người đến một ứng dụng khác. Khi ai đó sử dụng mã mời trong tin nhắn, người xuất bản nhận xét sẽ được trả tiền. Không giống như người gửi thư rác độc hại xuất bản nhiều nhận xét giống nhau dưới các mục khác nhau, loại nhận xét spam này được xuất bản bởi nhiều người khác nhau, đó có thể không phải là người gửi thư rác độc hại . Nhưng những tin nhắn phiếu giảm giá này thực sự làm phiền những người viết thư khác . Loại nhà xuất bản này không xuất bản quá nhiều đánh giá như những kẻ gửi thư rác độc hại, và sẽ không tập hợp dưới một mục cụ thể. Điều đó khiến bạn khó có thể nhận ra chỉ với ngữ cảnh địa phương thông qua Đồ thị Xianyu. Bằng cách giới thiệu Com ment Graph, loại bình luận tương tự này sẽ nhóm lại với nhau, sẽ được GAS công nhận.

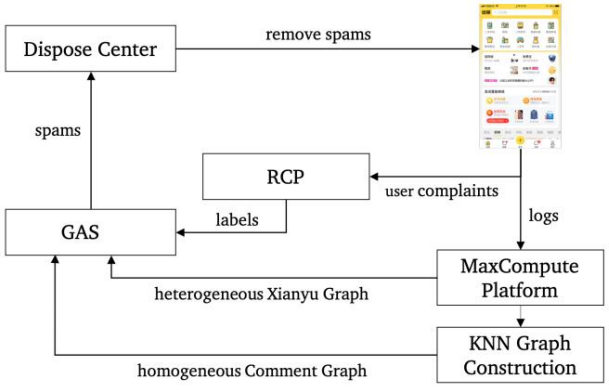
user_id item_id	bình luận
8737661 12381953	Đây là phần thưởng tôi nhận được tại Taobao, hãy liên hệ với tôi và tôi sẽ hướng dẫn bạn cách làm điều đó
8737661 26771502	Nhận tiền thưởng từ Taobao, tôi có thể dạy bạn 420310
27063522	Liên hệ với tôi để tìm hiểu cách nhận tiền thưởng từ Taobao
653613 20374180	Hướng dẫn bạn cách nhận tiền thưởng: + V xxxxxx
8806574 20634558	Vx: xxxxxx để dạy bạn nhận tiền thưởng

Bảng 5: Ví dụ về các thư rác bổ sung được GAS phát hiện so với GAS-local, trong đó “+ V xxxxxx” và “Vx: xxxxxx” có nghĩa là tài khoản ứng dụng WeChat.

5 TRIỂN KHAI HỆ THỐNG VÀ KHAI THÁC

Trong phần này, chúng tôi giới thiệu việc triển khai và triển khai phương pháp chống thư rác dựa trên GCN được đề xuất tại Xianyu. Trước tiên , chúng tôi giới thiệu tổng quan về hệ thống và sau đó xây dựng chi tiết về các mô-đun liên quan đến phương pháp của chúng tôi, đặc biệt là mô hình Tensorflow phân tán.

5.1 Tổng quan về hệ thống



Hình 9: Tổng quan về hệ thống.

Trong Hình 9, chúng tôi cho thấy kiến trúc của nền tảng chống thư rác tại Xianyu. Quy trình làm việc được minh họa như sau:

• Khi người dùng nhận xét về Ứng dụng Xianyu, nhật ký sẽ được lưu trữ trên nền tảng MaxCompute, một nền tảng xử lý dữ liệu chuyên nghiệp để lưu trữ dữ liệu quy mô lớn. Trong thực tế , chúng tôi chọn các bản ghi trong tháng gần đây để xây dựng biểu đồ không đồng nhất mỗi ngày. • Dựa trên nhật ký, Biểu đồ KNN được xây dựng hàng ngày trên nền tảng MaxCompute. • Việc triển khai Tensorflow phân tán của GAS được sử dụng để phát hiện các thư rác. • Các thư rác được phát hiện sẽ bị xóa khỏi ứng dụng và mali
tài khoản gian dối có thể bị vô hiệu hóa.
• RCP (Nền tảng kiểm tra rủi ro) được sử dụng để kiểm tra các khiếu nại từ người dùng bị trừng phạt. Các khiếu nại được các chuyên gia về con người xem xét và hỗ trợ sẽ dẫn đến hình phạt nhẹ và bị coi là sai lầm của mô hình. Kết quả của RCP sẽ được sử dụng làm các mẫu được gắn nhãn để tối ưu hóa hơn nữa mô hình của chúng tôi.

5.2 Thực hiện Chúng tôi tiến hành triển khai phân tán phương pháp được đề xuất.

Xem xét dữ liệu quy mô lớn ở Xianyu, tức là, hàng tỷ mục, hàng triệu người dùng và hàng tỷ nhận xét, kiến trúc máy chủ tham số của Tensorflow được sử dụng để cung cấp giải pháp phân tán cho việc lưu trữ, tìm nạp dữ liệu, đào tạo và dự đoán. Cụ thể, chúng tôi sử dụng 8 máy chủ tham số cùng với 8 công nhân, mỗi công nhân được trang bị một card GPU Nvidia V100, 6 nhân CPU, bộ nhớ 32GB. Máy chủ tham số có 2 lõi CPU với bộ nhớ 300 GB mỗi lõi.

5.2.1 lưu trữ. Đầu tiên, dữ liệu biểu đồ phải được lưu trữ và sẵn sàng khi mô hình cần lấy dữ liệu từ biểu đồ.
Đồ thị Xianyu là rất lớn và do đó không thực tế để được lưu trong một máy, vì vậy cấu trúc đồ thị cũng như các đặc điểm của đỉnh và cạnh được lưu trong máy chủ tham số. Lưu ý rằng cấu trúc đồ thị được lưu trữ dưới dạng danh sách kề để tăng hiệu quả cho bộ nhớ.

Một bước tốn thời gian là tải dữ liệu, cả hai giới hạn CPU (phân tích cú pháp) và I / O (tìm nạp). Để đẩy nhanh quá trình tải, chúng tôi chia danh sách kề và ma trận tính năng thành nhiều phần, rải đều chúng đến các máy chủ tham số để thực hiện tác vụ tải phân tán. Cụ thể, mỗi công nhân chịu trách nhiệm tải một phần cụ thể của bảng để điền vào ma trận con tương ứng trong máy chủ pa rameter. Bằng cách này, đạt đến gia tốc tuyến tính O (#workers) để tải dữ liệu.

Phân tán danh sách kề và ma trận tính năng đồng đều đến các máy chủ tham số có một ưu điểm khác: nó giúp cân bằng tải đọc / ghi cho quá trình đào tạo sắp tới.

5.2.2 Tìm nạp dữ liệu. Trong suốt thời gian tính toán, trước tiên nhân viên sẽ tìm kiếm thông tin cần thiết trên máy chủ tham số, tìm nạp chúng, sau đó thực hiện tính toán cục bộ. Tham số-server ar chitecture tránh tràn bộ nhớ trong khi dẫn đến kém hiệu quả hơn. Máy chủ tham số và công nhân được kết nối thông qua mạng nhưng thông lượng mạng chậm hơn nhiều so với truy cập bộ nhớ. Trong thử nghiệm của chúng tôi, trên mỗi nhân viên, có khoảng 41% thời gian bị lãng phí vào việc tìm nạp thông tin từ máy chủ. Để đẩy nhanh việc tra cứu

giai đoạn, chúng tôi sử dụng cơ chế bộ nhớ cache trên mỗi công nhân, tức là mỗi công nhân sẽ lưu vào bộ nhớ cache các tính năng và danh sách kề cục bộ khi thực hiện tìm kiếm trên máy chủ tham số. Kỹ thuật cache tiết kiệm khoảng 30% thời gian đào tạo.

Danh sách kề và ma trận tính năng được lưu trữ trên máy chủ param eter. Ngay cả khi chúng tôi lưu trữ bộ nhớ cache trên mỗi nhân viên, thường xuyên truy cập thông tin vùng lân cận và tính năng của các nút từ bộ nhớ CPU là không đủ cho GPU. Vì vậy, chúng tôi thực hiện theo phương pháp được giới thiệu trong [30], phương pháp này thu thập tất cả thông tin có liên quan đến lô nhỏ hiện tại, sau đó đưa nó vào bộ nhớ GPU cùng một lúc. Bằng cách này, giao tiếp CPU-GPU trong quá trình computa tion bị loại bỏ. Cùng với cơ chế nhà sản xuất-người tiêu dùng được sử dụng trong [30], việc sử dụng GPU được cải thiện đáng kể.

Cuối cùng, thời gian đào tạo của thử nghiệm ngoại tuyến được mô tả trong Mục 4.1 được giảm xuống còn 2 giờ đối với GAS.

6 KẾT LUẬN

Vấn đề phát hiện thư rác tại Xianyu phải đối mặt với hai thách thức chính: khả năng mở rộng và các hành động đối nghịch. Để giải quyết hai thách thức này, chúng tôi đã đề xuất thuật toán Chống thư rác (GAS) dựa trên GCN dựa trên GCN kết hợp bối cảnh địa phương và bối cảnh toàn cầu của com ments. Đánh giá ngoại tuyến và hiệu suất trực tuyến chứng minh hiệu quả của phương pháp của chúng tôi tại Xianyu. Các trường hợp trong thế giới thực được nghiên cứu để chứng minh thêm tác động của các bối cảnh khác nhau mà phần giới thiệu đưa ra nhằm giảm bớt tác động của các hành động đối nghịch. Cuối cùng, chúng tôi xây dựng chi tiết về việc thực hiện, triển khai và quy trình làm việc của phương pháp được đề xuất tại Xianyu.

7 LỜI CẢM ƠN

Chúng tôi muốn cảm ơn Yuhong Li, Jun Zhu, Leishi Xu đã hỗ trợ về thuật toán KNN Graph, và cảm ơn Huan Zhao đã đánh giá và thảo luận.

NGƯỜI GIỚI THIỆU

[1] Sanjeev Arora, Yingyu Liang và Tengyu Ma. 2017. Đơn giản nhưng khó đánh bại Cơ sở cho Nhúng câu. (2017).

[2] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart và Jimeng Sun. 2017. GRAM: mô hình chú ý dựa trên biểu đồ để học đại diện chăm sóc sức khỏe. Trong Kỷ yếu của Hội nghị Quốc tế ACM lần thứ 23 về Khám phá Tri thức và Khai thác Dữ liệu, SIGKDD. ACM, 787–795.

[3] Wei Dong, Charikar Moses, và Kai Li. 2011. Xây dựng đồ thị lân cận k-gần nhất hiệu quả cho các biện pháp tương tự chung. Trong Kỷ yếu của hội nghị liên quốc gia lần thứ 20 trên World wide web, WWW. ACM, 577–586.

[4] Mihajlo Grbovic và Haibin Cheng. 2018. Cá nhân hóa thời gian thực bằng cách sử dụng Nhúng ding cho Xếp hạng Tìm kiếm tại Airbnb. Trong Kỷ yếu của Hội nghị Quốc tế ACM lần thứ 24 về Khám phá Tri thức và Khai thác Dữ liệu, SIGKDD. ACM, 311–320.

[5] Aditya Grover và Jure Leskovec. 2016. node2vec: Học tính năng có thể mở rộng cho các mạng. Trong Kỷ yếu hội nghị quốc tế ACM lần thứ 22 về Khám phá tri thức và khai thác dữ liệu, SIGKDD. ACM, 855–864.

[6] Will Hamilton, Zhitao Ying và Jure Leskovec. 2017. Học biểu diễn quy nạp trên đồ thị lớn. Với những tiến bộ trong hệ thống xử lý thông tin thần kinh, NIPS. 1024–1034.

[7] Nitin Jindal và Bing Liu. 2008. Spam ý kiến và phân tích. Trong Kỷ yếu của Hội nghị Quốc tế về Tìm kiếm Web và Khai thác Dữ liệu năm 2008, WSDM. ACM, 219–230.

[8] Yoon Kim. 2014. Mạng nơ-ron hợp pháp để phân loại câu. Trong Kỷ yếu của Hội nghị 2014 về Phương pháp Thực nghiệm trong Xử lý Ngôn ngữ Tự nhiên, EMNLP. Hiệp hội Ngôn ngữ học Tính toán, 1746–1751.

[9] Thomas N Kipf và Max Welling. 2017. Phân loại bán giám sát với mạng chập đồ thị. Trong Hội nghị Quốc tế lần thứ 5 về Đại diện Học tập , ICLR.

[10] Jon M Kleinberg. 1999. Trung tâm, chính quyền và cộng đồng. Khảo sát điện toán ACM CSUR 31, 4es (1999), 5.

[11] Qimai Li, Zhichao Han và Xiao-Ming Wu. 2018. Thông tin chi tiết sâu hơn về Mạng lưới biểu đồ cho học tập bán giám sát. Trong Kỷ yếu của Hội nghị AAAI lần thứ 30 về Trí tuệ nhân tạo, AAAI. 3538–3545.

[12] Lizi Liao, Xiangnan He, Hanwang Zhang, và Tat-Seng Chua. 2018. Nhúng mạng xã hội thuộc tính. Giao dịch IEEE về Kiến thức và Kỹ thuật Dữ liệu, TKDE 30, 12 (2018), 2257–2270.

[13] Bing Liu và Lei Zhang. 2012. Một cuộc khảo sát về khai thác ý kiến và phân tích tình cảm. Trong Khai thác dữ liệu văn bản. Springer, 415–463.

[14] Ziqi Liu, Chaochao Chen, Xinxing Yang, Jun Zhou, Xiaolong Li và Le Song. 2018. Mạng Neural Đồ thị Không đồng nhất để Phát hiện Tài khoản Độc hại. Trong Kỷ yếu của Hội nghị Quốc tế ACM lần thứ 27 về Quản lý Thông tin và Tri thức, CIKM. ACM, 2077–2085.

[15] Marcin Luckner, Michał Gad, và Paweł Sobkowiak. 2014. Phát hiện spam web ổn định bằng cách sử dụng các tính năng dựa trên các mục từ vựng. Máy tính và Bảo mật 46 (2014), 79–93.

[16] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado và Jeff Dean. 2013. Các đại diện phân tán của các từ và cụm từ và cấu tạo của chúng. Trong Những tiến bộ trong hệ thống xử lý thông tin thần kinh, NIPS. 3111–3119.

[17] Myle Ott, Yejin Choi, Claire Cardie và Jeffrey T Hancock. 2011. Tìm thư rác ý kiến lừa dối bằng mọi trí tưởng tượng. Trong Kỷ yếu của Hội nghị Thường niên lần thứ 49 của Hiệp hội Ngôn ngữ học Tính toán, ACL. 309–319.

[18] Bryan Perozzi, Rami Al-Rfou và Steven Skiena. 2014. Deepwalk: Học trực tuyến về các đại diện xã hội. Trong Kỷ yếu hội nghị quốc tế ACM lần thứ 20 về Khám phá tri thức và khai thác dữ liệu, SIGKDD. ACM, 701–710.

[19] Amani K Samha, Yuefeng Li, và Jinglan Zhang. 2014. Trích xuất ý kiến dựa trên khía cạnh từ đánh giá của khách hàng. arXiv bản in trước arXiv: 1404.1982 (2014).

[20] Chao Shang, Qinqing Liu, Ko-Shin Chen, Jiangwen Sun, Jin Lu, Jinfeng Yi và Jinbo Bi. 2018. Mạng biểu đồ đa quan hệ dựa trên sự chú ý của Edge. arXiv bản in trước arXiv: 1802.04944 (2018).

[21] Saeedreza Shehnepoor, Mostafa Salehi, Reza Farahbakhsh và Noel Crespi. 2017. NetSpam: một khung phát hiện spam dựa trên mạng để đánh giá trên các phương tiện truyền thông xã hội trực tuyến. Giao dịch IEEE về Pháp y và Bảo mật Thông tin 12, 7 (2017), 1585–1595.

[22] Amira Soliman và Sarunas Girdzijauskas. 2017. Nhịp điệu bí danh dựa trên biểu đồ thích ứng để phát hiện thư rác trong mạng xã hội. Trong Hội nghị Quốc tế lần thứ 5 về Hệ thống Mạng, NETYS. 338–354.

[23] Yizhou Sun và Jiawei Han. 2012. Khai thác mạng thông tin không đồng nhất: nguyên tắc và phương pháp luận. Tổng hợp Bài giảng Khai phá dữ liệu và Khám phá tri thức 3, 2 (2012), 1–159.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser và Illia Polosukhin. 2017. Chú ý là tất cả những gì bạn cần. Những tiến bộ trong hệ thống xử lý thông tin thần kinh, NIPS. 5998–6008.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò và Yoshua Bengio. 2018. Mạng chú ý đồ thị. Hội nghị Quốc tế về Đại diện Học tập, ICLR (2018).

[26] Guan Wang, Sihong Xie, Bing Liu, và Philip S Yu. 2012. Xác định những kẻ gửi thư rác đánh giá cửa hàng trực tuyến thông qua biểu đồ đánh giá xã hội. Giao dịch ACM trên Hệ thống và Công nghệ Thông minh TIST (2012).

[27] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, và Dik Lun Lee. 2018. Đề xuất Nhúng hàng hóa quy mô tỷ cho thương mại điện tử ở Alibaba. Trong Kỷ yếu của Hội nghị Quốc tế ACM lần thứ 24 về Khám phá Tri thức và Khai thác Dữ liệu, SIGKDD. ACM, 839–848.

[28] Keyulu Xu, Weihua Hu, Jure Leskovec và Stefanie Jegelka. 2018. Mạng Neural Đồ thị mạnh đến mức nào ?. Trong Hội nghị Quốc tế về Đại diện Học tập , ICLR.

[29] Keyulu Xu, Chengtao Li, Yonglong Tian, Tomohiro Sonobe, Ken-ichi Kawarabayashi và Stefanie Jegelka. 2018. Học về Biểu diễn trên Đồ thị với Mạng Kiến thức Nhảy. Trong Kỷ yếu của Hội nghị Quốc tế lần thứ 35 về Học máy, ICML. 5449–5458.

[30] Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L Hamilton, và Jure Leskovec. 2018. Đồ thị Mạng Neural Hợp pháp cho Hệ thống Đề xuất Quy mô Web. Trong Kỷ yếu của Hội nghị Quốc tế ACM lần thứ 24 về Khám phá Tri thức và Khai thác Dữ liệu, SIGKDD. 974–983.

[31] Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song, và Dik Lun Lee. 2017. Tổng hợp khuyến nghị dựa trên biểu đồ meta trên các mạng thông tin không đồng nhất. Trong Kỷ yếu của Hội nghị Quốc tế ACM lần thứ 23 về Khám phá Tri thức và Khai thác Dữ liệu, SIGKDD. ACM, 635–644.