DỰ ĐOÁN LAN TRUYỀN THÔNG TIN TRÊN MẠNG HỌC THUẬT: MỘT CÁCH TIẾP CẬN MỚI VỚI SỰ KẾT HỢP GIỮA CÁC YẾU TỐ BÊN NGOÀI VÀ NỘI TẠI

HỒ THỊ KIM THOA, LÊ PHƯỚC NAM HÀ*
NGUYỄN VĂN KHANG, PHAN MINH ĐỨC
Khoa Tin học, Trường Đại học Sư phạm, Đại học Huế
**Email: hothikimthoa@dhsphue.edu.vn*

Tóm tắt: Trong nghiên cứu này, chúng tôi đề xuất một cách tiếp cận mới để giải quyết bài toán dự đoán lan truyền chủ đề trên mạng học thuật với sự kết hợp của các yếu tố bên ngoài và nội tại. Chúng tôi sử dụng học máy để dự đoán sự lan truyền của một chủ đề với sự kết hợp của các đặc trưng khác nhau. Đầu tiên, chúng tôi đề xuất một phương pháp mới để tính *xác suất kích hoạt*, hay còn gọi là *xác suất lây nhiễm* từ một nút đã được *"kích hoạt*" sang một nút "*chưa được kích hoạt"* dựa vào thông tin meta-path và thông tin văn bản (text). Xác suất này được xem xét như một yếu tố tác động từ bên ngoài để một nút có thể bị kích hoạt. Bên cạnh đó, chúng tôi cũng khai thác sở thích của tác giả đối với chủ đề lan truyền, đây được xem xét như yếu tố nội tại của tác giả mà có thể dẫn đến việc "lây nhiễm" chủ đề. Cuối cùng, chúng tôi kết hợp giữa đặc trưng về xác suất lây nhiễm và sở thích của tác giả trong dự đoán lan truyền chủ đề. Các thực nghiệm được tiến hành trên các chủ đề khác nhau của các tập dữ liệu mạng học thuật và thu được các kết quả thỏa mãn.

Từ khóa: Mạng xã hội, mạng học thuật, mạng đa quan hệ, lan truyền thông tin, dự đoán lây nhiễm.

1. GIỚI THIỆU

Vấn đề lan truyền thông tin (information diffusion) đã được nghiên cứu rộng rãi với mục đích đó là mô phỏng và dự đoán quá trình lan truyền thông tin giữa các đối tượng trong một mạng khi chúng kết nối với nhau. Quá trình lan truyền thông tin được mô tả với hai trạng thái cơ bản: trạng thái *"kích hoạt" (active)* hay còn gọi là *"nhiễm"* và trạng thái *"chưa kích hoạt" (inactive)* hoặc còn gọi chưa nhiễm. Các nút mạng được xem xét ở trạng thái kích hoạt thông tin nếu như chúng đã thực hiện những hành động liên quan tới thông tin đó, ví dụ như một tác giả được gọi là *"kích hoạt"* với chủ đề "machine learning" kể từ khi tác giả đó đã nghiên cứu và xuất bản được các bài báo khoa học liên quan tới chủ đề "machine learning" hay trong chiến lược marketing về điện thoại iPhone, một khách hàng được đánh dấu là đã *"kích hoạt"* nếu khách hàng đó đã mua sản phẩm điện thoại iPhone. Đa số các nghiên cứu về lan truyền thông tin đều được thực hiện trên các mạng đồng nhất (homogeneous network), nơi chỉ có một kiểu đối tượng và một kiểu liên kết ở trong mạng, ví dụ như mạng đồng tác giả với các đối tượng là các tác giả và liên kết đồng tác giả hay trong mạng xã hội Twitter thì đối tượng là các user và liên kết đó là follow. Tuy nhiên, trong thực tế thì đa số các mạng đều ở dạng không đồng nhất (heterogeneous network),

Tạp chí Khoa học, Trường Đại học Sư phạm, Đại học Huế
ISSN 1859-1612, Số 3(63)/2022: tr.117-129
Ngày nhận bài: 29/6/2022; Hoàn thành phản biện: 27/7/2022; Ngày nhận đăng: 01/8/2022

trong đó có nhiều kiểu đối tượng và nhiều kiểu liên kết giữa các đối tượng, ví dụ như mạng học thuật (bibliographic network) là một mạng không đồng nhất, trong đó có nhiều kiểu đối tượng bao gồm như tác giả, bài báo, hội nghị, tổ chức liên kết,... và nhiều kiểu quan hệ như quan hệ đồng tác giả, quan hệ cùng đồng tác giả với một tác giả thứ ba, quan hệ cùng tham gia nộp bài ở một hội nghị,... Trong nghiên cứu này, chúng tôi tập trung và nghiên cứu vấn đề lan truyền thông tin trên mạng không đồng nhất.

Có hai cách tiếp cận chính để mô phỏng và dự đoán lan truyền thông tin trên mạng không đồng nhất. Thứ nhất, quá trình lan truyền thông tin được mô phỏng thông qua các mô hình lan truyền, bao gồm mô hình ngưỡng tuyến tính (linear threshold model (LT)) [5, 10], mô hình thác nước độc lập (independent cascade model (IC)) [4], mô hình thác nước giảm dần (decreasing cascade model) [8], mô hình ngưỡng tổng quát (general threshold model) [9], mô hình dựa vào khuếch tán nhiệt (heat diffusion-based model) [17],... Với cách tiếp cận này, một vài nút mạng đã kích hoạt sẽ tác động đến các hàng xóm chưa kích hoạt của chúng và có thể sẽ biến các hàng xóm đó chuyển sang trạng thái kích hoạt. Đã có một vài mô hình mở rộng từ mô hình IC như TextualHomo-IC (Homophily Independent Cascade Diffusion) [7] hay mô hình HPM-IC (Heterogeneous Probability Model – IC) [11]. Bên cạnh đó, cũng có một vài mô hình mở rộng từ mô hình LT, ví dụ như mô hình MLTM-R (Multi-Relational Linear Threshold Model - Relation Level Aggregation) [6] và mô hình HPM-LT (Probability Model-LT) [11]. Tất cả các mô hình trên đều đã đưa ra các phương pháp tính xác suất lây nhiễm của một nút ở trạng chưa kích hoạt dựa vào thông tin meta-path hoặc dựa vào thông tin văn bản. Do đó, các nghiên cứu trên chỉ mới xem xét sự ảnh hưởng từ bên ngoài, cụ thể là từ các hàng xóm, mà chưa xem xét đến yếu tố nội tại của các nút chưa lây nhiễm, ví dụ như mức độ thích thú của tác giả đối với chủ đề lan truyền hay tầm ảnh hưởng của tác giả,... Do đó, cách tiếp cận thứ hai xuất hiện với sự kết hợp của các đặc trưng trong dự đoán lan truyền thông tin.

Cách tiếp cận thứ hai trong dự đoán lan truyền thông tin trên mạng không đồng nhất đó là sử dụng học máy và học sâu, khai thác các đặc trưng khác nhau. Sự lan truyền của các tweet trên Twitter đã được nghiên cứu với phương pháp học có giám sát [16], Bên cạnh đó, sự lan truyền thông tin trên Github cũng đã được nghiên cứu sử dụng phương pháp học có giám sát [1]. Hơn nữa, học sâu cũng đã được sử dụng để nghiên cứu về dự đoán lan truyền thông tin trên mạng không đồng nhất [12]. Lan truyền thông tin trên mạng học thuật đã được nghiên cứu với cách tiếp cận thứ nhất thông qua nhiều mô hình lan truyền khác nhau. Bên cạnh đó, vấn đề này cũng đã được nghiên cứu ở cách tiếp cận thứ hai với phương pháp sử dụng học sâu [12]. Tuy nhiên, phương pháp học máy thì chưa được sử dụng trong nghiên cứu về dự đoán lan truyền thông tin trên mạng học thuật. Do đó, trong nghiên cứu này, chúng tôi sẽ tập trung vào sử dụng học máy cho dự đoán lan truyền các chủ đề trong mạng học thuật.

Trong nghiên cứu này, chúng tôi đề xuất một cách tiếp cận mới đó là kết hợp giữa yếu tố tác động bên ngoài và yếu tố nội tại trong việc dự đoán lan truyền chủ đề trong mạng học thuật. Đầu tiên, chúng tôi xem xét nhân tố ảnh hưởng từ các hàng xóm đã kích hoạt bằng cách đề xuất một phương pháp mới trong việc tính xác suất lây nhiễm từ một nút kích hoạt

sang một nút chưa kích hoạt dựa vào cả thông tin meta-path và thông tin văn bản. Thêm vào đó, chúng tôi xem xét đến yếu tố nội tại của các nút chưa kích hoạt dựa vào mức độ thích thú của tác giả đối với chủ đề lan truyền. Chúng tôi sử dụng học có giám sát để rút ra được các hệ số tốt nhất liên kết với các đặc trưng được trích rút từ dữ liệu. Kết quả thực nghiệm cho thấy rằng phương pháp tính xác suất lây nhiễm mới mang lại hiệu quả tốt hơn trong dự đoán lan truyền thông tin so với phương pháp tính cũ đó là chỉ sử dụng meta-path hoặc thông tin văn bản một cách riêng biệt. Bên cạnh đó, sự kết hợp giữa yếu tố tác động bên ngoài và yếu tối nội tại mang lại hiệu quả cao hơn so với chỉ sử dụng một trong hai một cách độc lập. Đặc biệt, sự kết hợp giữa yếu tố bên ngoài theo phương pháp tính xác suất lây nhiễm mới và yếu tố nội tại mang lại độ chính xác cao nhất.

2. CÁC NGHIÊN CỨU LIÊN QUAN

Lan truyền thông tin là quá trình một phần thông tin được lan truyền từ cá nhân hoặc cộng đồng này sang cá nhân hoặc cộng đồng khác trong một mạng lưới. Đã có nhiều nghiên cứu phân tích sự lan truyền thông tin, chủ yếu tập trung vào việc nghiên cứu thông tin nào được lan truyền nhanh nhất, các yếu tố nào ảnh hưởng đến sự lan truyền thông tin, và mô hình nào để mô phỏng và dự đoán sự lan truyền. Những câu hỏi này đóng một vai trò quan trọng trong việc tìm hiểu hiện tượng lan truyền thông tin. Chúng đã được trả lời bởi các kết quả trong các nhánh nghiên cứu nhỏ hơn bao gồm như các mô hình lây lan dịch bệnh, phân tích sự ảnh hưởng và mô hình dự đoán.

Phần lớn các nghiên cứu về lan truyền thông tin đã được nghiên cứu về mạng đồng nhất nơi chỉ tồn tại một loại đối tượng và một loại liên kết trong mạng lưới. Tuy nhiên, trong thế giới thực, phần lớn các mạng không đồng nhất khi có nhiều loại đối tượng khác nhau và nhiều quan hệ trong mạng, ví dụ, một mạng học thuật là một mạng không đồng nhất chứa nhiều đối tượng, bao gồm tác giả, bài báo, địa điểm, đơn vị liên kết, v.v., đồng thời tồn tại nhiều mối quan hệ giữa các tác giả như quan hệ đồng tác giả, mối quan hệ với đồng tác giả chung,... Nghiên cứu của chúng tôi tập trung vào thông tin lan truyền trên các mạng không đồng nhất.

Để nghiên cứu các mô hình dự đoán trên các mạng không đồng nhất, có hai các phương pháp tiếp cận để mô hình hóa và dự đoán sự lan truyền thông tin. Thứ nhất, quá trình lan truyền thông tin được mô phỏng thông qua các mô hình lan truyền, bao gồm mô hình ngưỡng tuyến tính (linear threshold model (LT)) [5, 10], mô hình thác nước độc lập (independent cascade model (IC)) [4], mô hình thác nước giảm dần (decreasing cascade model) [8], mô hình ngưỡng tổng quát (general threshold model) [9], mô hình dựa vào khuếch tán nhiệt (heat diffusion-based model) [17],... Với cách tiếp cận này, một nút mạng đã kích hoạt sẽ tác động đến các hàng xóm chưa kích hoạt của chúng và có thể sẽ biến các hàng xóm đó chuyển sang trạng thái kích hoạt. Ví dụ, trong mô hình IC, một nút mạng ở trạng thái kích hoạt có thể tác động và lây nhiễm một nút mạng khác ở trạng thái chưa kích hoạt với một xác suất nhất định. Trong mô hình LT, một nút ở trạng thái chưa kích hoạt sẽ bị tác động và chuyển sang trạng thái kích hoạt nếu như tổng trọng số giữa các nút hàng xóm đã được kích hoạt với nó lớn hơn một ngưỡng nào đó. Đã có một vài mô hình mở rộng từ mô hình IC như TextualHomo-IC (Homophily Independent Cascade

Diffusion) [7], trong đó xác suất lây nhiễm được ước lượng dựa vào mức độ đồng nhất về thông tin văn bản hay mô hình HPM-IC (Heterogeneous Probability Model – IC) [11] trong đó xác suất lây nhiễm được tính toán bằng xác suất điều kiện dựa vào thông tin meta-path. Bên cạnh đó, cũng có một vài mô hình mở rộng từ mô hình LT, ví dụ như mô hình MLTM-R (Multi-Relational Linear Threshold Model - Relation Level Aggregation) [6] và mô hình HPM-LT (Probability Model-LT) [11]. Tất cả các mô hình trên đều đã đưa ra các phương pháp tính xác suất lây nhiễm từ một nút mạng đã kích hoạt sang một nút khác chưa kích hoạt dựa vào thông tin meta-path hoặc dựa vào thông tin văn bản. Do đó, các nghiên cứu trên chỉ mới xem xét sự ảnh hưởng từ bên ngoài, cụ thể là sự ảnh hưởng từ các hàng xóm mà chưa xem xét đến yếu tố nội tại của các nút chưa lây nhiễm, ví dụ như mức độ thích thú của tác giả đối với chủ đề đó hay tầm ảnh hưởng của tác giả,... Do đó, cách tiếp cận thứ hai xuất hiện sự kết hợp của các đặc trưng khác nhau trong dự đoán lan truyền thông tin.

Học máy và học sâu là cách tiếp cận thứ hai trong dự đoán lan truyền thông tin trên mạng không đồng nhất trong đó khai thác và kết hợp các đặc trưng khác nhau. Sự lan truyền của các tweet trên Twitter đã được nghiên cứu với phương pháp học có giám sát [16], trong đó khai thác hai đặc trưng là sở thích của người dùng và độ tương tự về nội dung giữa người dùng đã kích hoạt và chưa kích hoạt. Thông tin văn bản được sử dụng để ước lượng các đặc trưng trên sử dụng kĩ thuật mô phỏng chủ đề, cụ thể là phân bổ Dirichlet tiềm ẩn. Bên cạnh đó, sự lan truyền thông tin trên Github cũng đã được nghiên cứu sử dụng phương pháp học có giám sát [1]. Hơn nữa, học sâu cũng đã được sử dụng để nghiên cứu về dự đoán lan truyền thông tin trên mạng không đồng nhất [12].

Lan truyền thông tin trên mạng học thuật đã được nghiên cứu với cách tiếp cận thứ nhất thông qua nhiều mô hình lan truyền khác nhau. Bên cạnh đó, vấn đề này cũng đã được nghiên cứu ở cách tiếp cận thứ hai với phương pháp sử dụng học sâu [12]. Tuy nhiên, phương pháp học máy thì chưa được sử dụng trước đó. Do đó, trong nghiên cứu này, chúng tôi sẽ tập trung vào sử dụng học máy cho dự đoán lan truyền các chủ đề trong mạng học thuật. Hơn nữa, đặc trưng về xác suất lây nhiễm trong các nghiên cứu trước đây đều chỉ được ước lượng dựa vào một trong hai thông tin meta-path hoặc thông tin văn bản. Do đó, trong nghiên cứu này, chúng tôi đề xuất kết hợp giữa hai thông tin gồm thông tin meta-path và thông tin văn bản để ước lượng đặc trưng về xác suất lây nhiễm thay vì chỉ sử dụng một trong hai thông tin như ở các nghiên cứu trước đây. Bên cạnh đó, chúng tôi còn khai thác thêm đặc trưng về sở thích của tác giả với chủ đề lan truyền, đây chính là đặc trưng thể hiện yếu tố nội tại. Chúng tôi muốn nghiên cứu và đánh giá sự hiệu quả của sự kết hợp giữa các thông tin trên trong dự đoán lan truyền thông tin so với trường hợp sử dụng chúng một cách riêng biệt.

3. KIẾN THỨC NỀN TẢNG

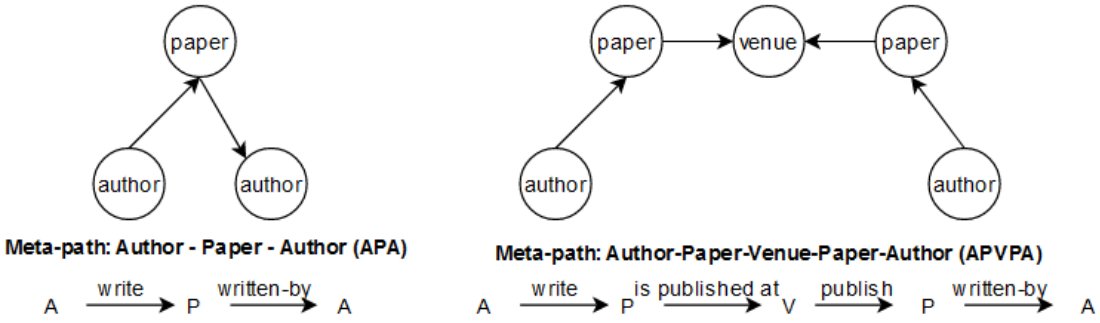
Meta-path P là một siêu đường dẫn được định nghĩa trên lược đồ tổng quát của mô hình mạng T_G= (A, G), trong đó A là tập các nút mạng và G là tập các quan hệ giữa chúng [6,

11, 15]. Meta-path được kí hiệu $A_1 \xrightarrow{R1} A_2 \xrightarrow{R2} A_3 \xrightarrow{R3} \ldots \xrightarrow{Rl} A_{l+1}$. Mối quan hệ tổng hợp thu được dưới dạng R = $R_1 o R_2 o \ldots R_l$ giữa A_1 và A_{l+1}, trong đó o chính là toán tử kết hợp.

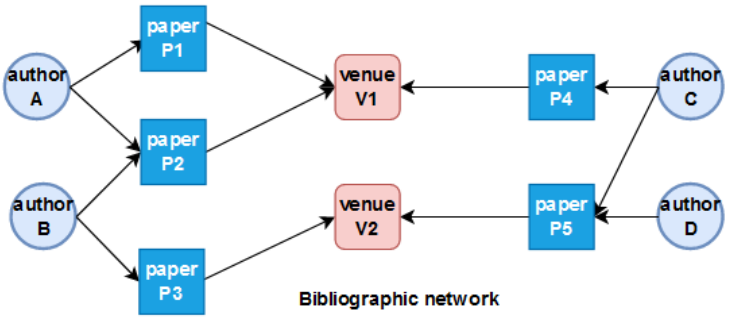
Độ dài của P chính là số lượng quan hệ ở trong P. Hơn nữa, một meta-path P được gọi là đối xứng nếu như các mối quan hệ R được định nghĩa trong P là đối xứng. Ví dụ như trong mạng học thuật, mối quan hệ đồng tác giả có thể được định nghĩa thông qua một meta-path đối xứng với độ dài bằng 2 như sau: $A \xrightarrow{writing} P \xrightarrow{written\ by} A$.

Một đường p = (a_1 a_2 ...a_{l+1}) giữa a_1 và a_{l+1} trong mạng G tuân theo cú pháp của meta-path P nếu $\forall i, \emptyset(a_i) = A_i$ và mỗi cạnh $e_i =< a_i a_{i+1} >$ thuộc mối quan hệ R_i trong meta-path P. Chúng ta gọi các đường p này là các thể hiện của P (path instance), được kí hiệu $p \in P$.

Ví dụ: Hai loại meta-path APA và APVPA được mô tả như ở Hình 1.



Hình 1. *Mô tả về hai loại meta-path APA và APVPA*



Meta-Path	Path instance
APA	A-P2-B C-P5-D
APVPA	A-P1-V1-P4-C A-P2-V1-P4-C B-P2-V1-P4-C B-P3-V2-P5-C B-P3-V2-P5-D

Hình 2. *Ví dụ về meta-path và path-instance trên mạng học thuật*

Ví dụ như ở Hình 2 thể hiện mô hình mạng học thuật với các nút mạng gồm có 4 tác giả, 5 bài báo, 2 hội nghị nộp bài. Trong mô hình mạng trên đồng thời thể hiện các mối quan hệ giữa các nút mạng. Chúng ta có thể thấy rằng trong mô hình mạng trên chứa 2 loại meta-path đó là APA và APVPA. Tương ứng với meta-path APA, chúng ta có 2 path-instance và có 5 path-instance đối với meta-path APVPA.

4. GIẢI PHÁP ĐỀ XUẤT

Trong phần này, chúng tôi mô tả chi tiết về một cách tiếp cận mới trong việc giải quyết bài toán dự đoán liệu rằng một tác giả trong mạng học thuật có bị "nhiễm" một chủ đề

lan truyền hay là không. Cách tiếp cận này bao gồm mô tả về phương pháp học có giám sát cho bài toán dự đoán lan truyền chủ đề và phương pháp trích rút các đặc trưng.

4.1. Học có giám sát cho bài toán dự đoán lan truyền chủ đề trên mạng học thuật

Chúng tôi sử dụng các phương pháp học có giám sát để dự đoán lan truyền chủ đề trên mạng học thuật. Đối với một chủ đề cụ thể, chúng tôi dự đoán liệu rằng một tác giả "*chưa kích hoạt/ chưa nhiễm (inactive)"* sẽ *"kích hoạt/nhiễm (active)"* chủ đề đó trong thời gian tương lai T2 hay không dựa vào các thông tin đã có sẵn của tác giả đó trong thời gian quá khứ T1. Tất cả các tác giả xuất bản các bài báo liên quan tới chủ đề lan truyền được gắn nhãn là *"đã kích hoạt/đã nhiễm (active)"* và ngược lại.

Trong bước huấn luyện, đầu tiên chúng tôi lấy mẫu một tập các tác giả "chưa từng kích *hoạt/nhiễm*" trong giai đoạn T1, và trích rút các đặc trưng. Sau đó, học máy sẽ được sử dụng để xây dựng mô hình huấn luyện để học các hệ số tốt nhất liên kết với các đặc trưng. Trong bước kiểm tra, chúng tôi áp dụng mô hình đã huấn luyện ở trên lên tập dữ liệu kiểm tra để so sánh kết quả độ chính xác với dữ liệu thực tế.

4.2. Đặc trưng xác suất kích hoạt/lây nhiễm

Đầu tiên, chúng tôi xem xét đến yếu tố tác động ảnh hưởng từ bên ngoài mà có thể dẫn đến sự kích hoạt/ lây nhiễm. Một tác giả có thể bị kích hoạt/lây nhiễm từ các hàng xóm của họ với một xác suất nhất định. Xác suất này gọi là xác suất kích hoạt hay còn gọi xác suất lây nhiễm. Xác suất kích hoạt từ một hàng xóm đã kích hoạt u đến một tác giả chưa từng kích hoạt v có thể được xác định bởi nhiều công thức khác nhau: dựa vào thông tin meta-path hoặc dựa vào thông tin văn bản (text). Chúng tôi sẽ tổng quát lại các phương pháp ước lượng xác suất kích hoạt ở các nghiên cứu trước đây, sau đó đề xuất một phương pháp ước lượng mới.

a. Xác suất kích hoạt dựa vào thông tin meta-path:

Xác suất kích hoạt được ước lượng dựa vào độ tương tự về thông tin meta-path thông qua các công thức sau:

PathSim: Đây là một độ đo tương tự dựa trên thông tin meta-path [6, 15]. Cho một meta-path đối xứng E_k tương ứng với mối quan hệ k, PathSim giữa hai đối tượng có thể được định nghĩa như ở công thức (1):

$$s^{E_k}(u,v) = \frac{2|P^{E_k}_{(u,v)}|}{\left|P^{E_k}_{(u,u)}\right| + |P^{E_k}_{(v,v)}|} \quad (1)$$

Trong đó $P^{E_k}_{(u,v)}$ là tập hợp các meta-path của mối quan hệ k, bắt đầu từ u và kết thúc tại v. |.| thể hiện kích thước của tập hợp.

Xác suất kích hoạt dựa vào thông tin meta-path và Bayesian [11]:

$$AP(MP) = P^k(u|v) = \frac{n^k_{v\to u}}{\sum_{r\in nei_v} n^k_{v\to r}} \quad (2)$$

$$P(u|v) = \frac{\sum_{k=1}^{m} \alpha_k \, n_{v\to u}^{k}}{\sum_{k=1}^{m} \alpha_k \, \sum_{r\in nei_v} n_{v\to r}^{k}} \quad (3)$$

$$P(u|\{v\}) = max_{M=1\ldots n}(P(u|v) \quad (4)$$

Công thức (2) thể hiện xác suất kích hoạt từ một nút đã kích hoạt v sang một nút chưa kích hoạt u trong meta-path k. $n_{v\to u}^{k}$ là số thể hiện đường đi theo meta-path k từ v qua u. Xác suất kích hoạt từ nút v qua u được thể hiện ở công thức (3) với sự tổng hợp của tất cả các loại meta-path k. Cuối cùng, công thức (4) thể hiện xác suất kích hoạt của một nút u bằng cách tối đa hóa các xác suất kích hoạt từ các nút hàng xóm đã kích hoạt v của u.

b. Xác suất kích hoạt dựa vào thông tin văn bản (AP(IS))

Xác suất kích hoạt cũng đã được ước lượng từ thông tin văn bản (text) [7] kể từ khi thông tin văn bản chứa đựng nhiều thông tin quan trọng của tác giả. Trong mạng học thuật, thông tin văn bản trong các bài báo xuất bản đóng vai trò quan trọng trong việc thể hiện chuyên ngành nghiên cứu của các tác giả. Do đó, thông tin văn bản vô cùng quan trọng trong công việc dự đoán lan truyền chủ đề giữa các nhà khoa học.

Đầu tiên, chúng ta có thể ước lượng thông tin của tác giả dựa trên thông tin văn bản sử dụng các kĩ thuật khai phá văn bản, bao gồm tần suất xuất hiện các từ - đảo ngược tần suất của văn bản (Term Frequency – Inverse Document Frequency, TFIDF) [14] hoặc mô phỏng chủ đề (Topic Modeling) [2, 13].

Sau khi khai thác thông tin văn bản, chúng ta có thể sử dụng các độ đó khoảng cách để đo lượng độ tượng tự giữa hai nút mạng. Đối với TF-IDF, khoảng cách cosin thường được sử dụng để đó lường độ tương tự ở công thức (5). Đối với mô phỏng chủ đề, đầu ra của mô phỏng chủ đề là các vector phân phối xác suất. Do đó, chúng ta có thể chọn một trong các độ đo khoảng cách liên quan đến khoảng cách vector, ví dụ như Euclidean hoặc Cosin, Jaccard,... Tuy nhiên, kết quả thực nghiệm ở nghiên cứu trước của chúng tôi [3] đã chứng minh rằng sẽ tốt hơn nếu chúng ta sử dụng độ đo khoảng cách liên quan tới phân phối xác suất như KullbackLeibler Divergence (Công thức 7), Jensen-Shannon divergence (Công thức 8), Hellinger distance (Công thức 6),...

$$IS(u,v) = Cos(T_u, T_v) = \frac{T_u . T_v}{||T_u||. || T_v||} \quad (5)$$

$$IS(u,v) = d_H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{K} (\sqrt{p_i} - \sqrt{q_i})^2} \quad (6)$$

$$IS(u,v) = d_{KL}(P|| Q) = \sum_{x\in X} P(x) \frac{P(x)}{Q(x)} \quad (7)$$

$$IS(u,v) = d_{JS}(P,Q) = \frac{1}{2} \sum_{i=1}^{K} p_i ln \frac{2p_i}{p_i + q_i} + \frac{1}{2} \sum_{i=1}^{K} q_i ln \frac{2q_i}{p_i + q_i} \quad (8)$$

Trong đó, where T_u, T_v là các vector TFIDF tương ứng của u and v; P, Q là phân phối xác suất của các chủ đề của hai nút u và v.

c. Xác suất kích hoạt tổng hợp dựa vào thông tin meta-path và văn bản

Thông tin về meta-path và văn bản đều đã được sử dụng để ước lượng xác suất kích hoạt, tuy nhiên chúng lại được sử dụng riêng lẻ. Do đó, trong nghiên cứu này, chúng tôi đề xuất phương pháp mới để ước lượng xác suất kích hoạt của một nút với sự kết hợp giữa hai thông tin là meta-path và văn bản, được đặt tên là xác suất kích hoạt tổng hợp *(aggregated activation probability, viết tắt AP(MP+IS)).*

$$AP(u,v) = (1-\sigma) * AP(MP) + \sigma * AP(IS) \quad (9)$$

$$AP(u,\{v\}) = max_{M=1..n}(AP(u,v)) \quad (10)$$

Công thức (9) thể hiện xác suất kích hoạt tổng hợp từ một nút đã kích hoạt v sang nút chưa được kích hoạt u dựa vào thông tin meta-path và văn bản. AP(MP) là xác suất kích hoạt được ước lượng dựa vào thông tin meta-path như ở công thức (3) và AP(IS) là độ tương tự về sở thích giữa hai nút u và v dựa vào thông tin văn bản, như ở công thức từ (5) đến (8). σ là tham số dùng để điều khiển tỉ lệ ảnh hưởng của xác suất kích hoạt dựa vào meta-path và độ tương tự về sở thích lên xác suất kích hoạt tổng hợp. σ ∈ [0,1], nếu σ lớn thì đồng nghĩa với việc chúng ta tập trung vào thông tin văn bản và ngược lại. Cuối cùng, chúng ta có thể định nghĩa xác suất kích hoạt tổng hợp của một nút chưa kích hoạt u bằng cách tối đa hóa các xác suất kích hoạt tổng hợp từ các hàng xóm đã kích hoạt v của u (ở công thức 10).

4.3. Đặc trưng sở thích của tác giả

Bên cạnh sự ảnh hưởng bên ngoài từ các hàng xóm, chúng tôi xem xét đến các yếu tố nội tại của các tác giả trong việc dự đoán lan truyền chủ đề trong mạng học thuật. Yếu tố nội tại có thể là tầm ảnh hưởng của tác giả, sở thích của tác giả,... Trong nghiên cứu này, chúng tôi tập trung vào xem xét yếu tố sở thích của tác giả đối với chủ đề được lan truyền. Bên cạnh sự tác động từ các hàng xóm, một tác giả có thể bị kích hoạt/lây nhiễm với một chủ đề phụ thuộc vào sở thích của họ đối với chủ đề đó. Sở thích của tác giả đối với một chủ đề có thể được hình thành thông qua các cách khác nhau như thông qua bạn bè giới thiệu, đọc các bài báo khoa học, hoặc có thể thông qua việc tham gia các hội thảo khoa học,... Các hội nghị khoa học là nơi cung cấp các diễn đàn học thuật để cho các nhà khoa học chia sẻ các kết quả nghiên cứu của họ. Tham gia các hội nghị khoa học chính là cơ hội để gặp gỡ, chia sẻ và có thể bắt đầu nghiên cứu một lĩnh vực, chủ đề mới. Do đó, để ước lượng mức độ thích thú của tác giả đối với các chủ đề lan truyền, chúng tôi đề xuất tính số lượng hội nghị liên quan tới chủ đề đó mà tác giả đã từng tham gia.

5. THỰC NGHIỆM VÀ KẾT QUẢ

5.1. Thực nghiệm

Tập dữ liệu: Chúng tôi sử dụng tập dữ liệu ”DBLP-SIGWEB.zip”, được tổng hợp từ 1995 đến 2015 từ cơ sở dữ liệu DBLP. Tập dữ liệu này chứa tất cả các thông tin liên quan tới các bài báo xuất bản và hồ sơ các tác giả từ 7 hội nghị ACM. Nó còn chứa các thông tin liên quan tới tác giả, cơ quan liên kết, các thông tin về các hội nghị,...

Cài đặt thực nghiệm: Chúng tôi xem xét quá trình lan truyền của một chủ đề T. Chúng tôi tiến hành thực nghiệm với 3 chủ đề: "*Data Mining*", "*Machine Learning*" và "*Social Network*". Đầu tiên, chúng tôi lấy mẫu các tác giả mà họ đã kích hoạt với chủ đề T và sử dụng tập tác giả này như là các nút với nhãn "active". Chúng tôi cũng lấy mẫu gồm một tập các tác giả chưa kích hoạt với chủ đề T và được gán nhãn là "inactive". Hai tập mẫu "active" và "inactive" có kích thước bằng nhau để tập dữ liệu huấn luyện cân bằng.

Chúng tôi sử dụng các phương pháp phân loại như là các mô hình dự đoán. Trong tập dữ liệu huấn luyện, một tác giả X đã kích hoạt với chủ đề T vào năm $Y_{XT,}$ chúng tôi sẽ trích xuất các đặc trưng của X trong giai đoạn $T_1 = [1995, Y_{XT} - 1]$. Một tác giả Y chưa kích hoạt, chúng tôi trích xuất các đặc trưng của Y trong giai đoạn $T_1 = [1995, 2015]$.

Bảng 1. *Sự kết hợp giữa các đặc trưng*

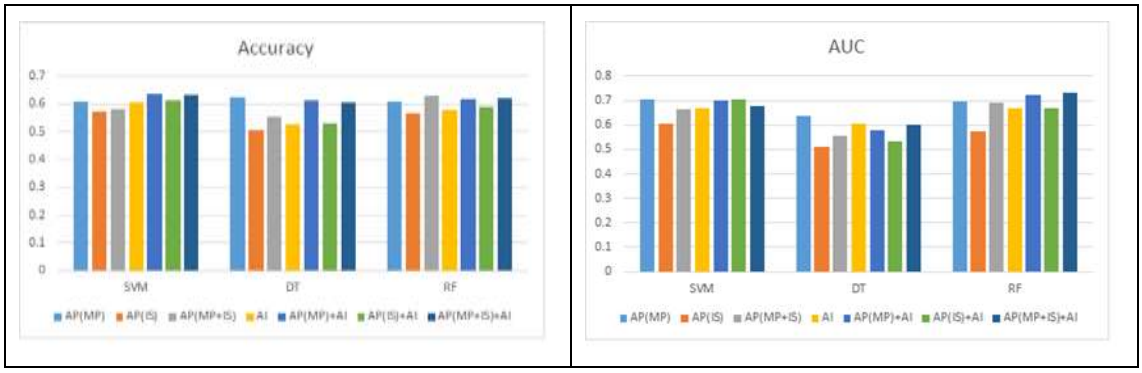
Số thứ tự	**Các đặc trưng**
1	Xác suất kích hoạt dựa trên meta-path (AP(MP))
2	Xác suất kích hoạt dựa trên thông tin văn bản (AP(IS))
3	Xác suất kích hoạt tổng hợp (AP(MP + IS))
4	Sở thích của tác giả đối với chủ đề T (AI)
5	AP(MP) + AI
6	AP(IS) + AI
7	AP(MP+IS) + AI

Chúng tôi tiến hành thực nghiệm trên các tập đặc trưng khác nhau để đánh giá sự cải tiến về hiệu suất. Sự kết hợp của các đặc trưng được thể hiện ở Bảng 1. Chúng tôi tiến hành tất cả 07 thực nghiệm với sự kết hợp giữa xác suất kích hoạt và sở thích của tác giả đối với chủ đề lan truyền. Thực nghiệm (1) và (2) thể hiện cho phương pháp tính xác suất kích hoạt cũ dựa trên thông tin meta-path và thông tin văn bản, trong khi đó (3) thể hiện phương pháp mà chúng tôi đề xuất. Thực nghiệm số (4) thể hiện sở thích của tác giả đối với chủ đề lan truyền T. Thực nghiệm số (5), (6), (7) là sự kết hợp giữa xác suất kích hoạt trong nhiều phương pháp tính khác nhau với đặc trưng sở thích của tác giả. Chúng tôi sử dụng kết quả của thực nghiệm (1, 2) như là cơ sở để so sánh với thực nghiệm (3), từ đó thấy được hiệu quả của phương pháp tính xác suất kích hoạt mới. Bên cạnh đó, các kết quả thực nghiệm của (1, 4), (2, 4) và (3, 4) làm cơ sở tương ứng cho các thực nghiệm (5), (6), (7), từ đó nhìn thấy tầm quan trọng của đặc trưng sở thích của tác giả cũng như hiệu quả của việc kết hợp giữa đặc trưng xác suất kích hoạt và sở thích của tác giả.

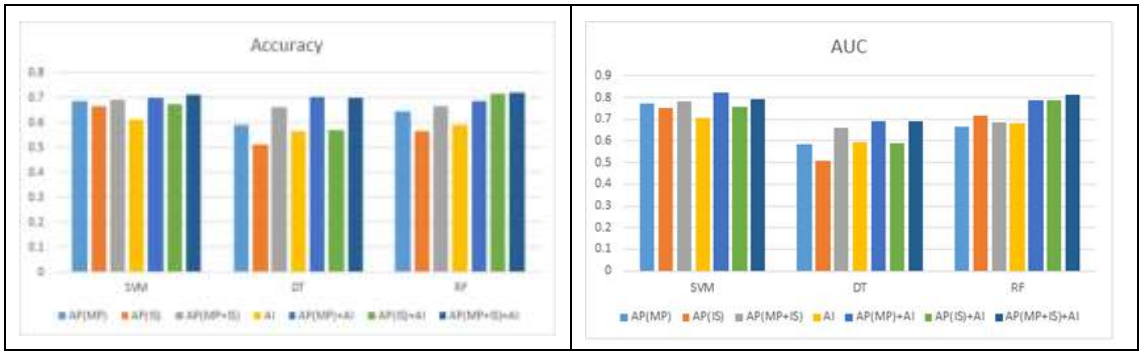
Đối với việc tính toán AP(MP), chúng tôi sử dụng hai loại meta-path đó là APA (Author-Paper-Author) và APAPA (Author-Paper-Author-Paper-Author). Để tính AP(IS), đầu tiên chúng tôi lấy các thông tin văn bản từ các từ khóa của các bài báo của các tác giả trong giai đoạn T_1, sau đó sử dụng công thức (5) để ước lượng độ tương tự về sở thích giữa hai tác giả. Cuối cùng, AP(MP+IS)(MP+IS) được tính với tham số $\sigma = 0.5$.

Trong nghiên cứu này, chúng tôi thực nghiệm với 03 thuật toán phân loại đó là Support Vector Machine (SVM, Linear Kernel), Decision Tree (DT) và Random Forest (RF). Để đánh giá độ chính xác của việc dự đoán lan truyền thông tin, chúng tôi sử dụng hai chỉ số đó là Accuracy và ROC-AUC để đánh giá hiệu quả.

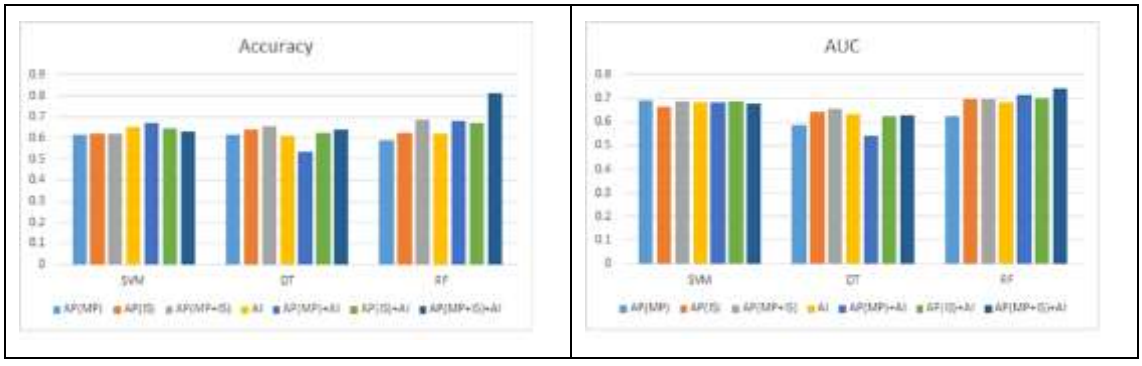
5.2. Kết quả



Hình 3. *Kết quả phân loại với chủ đề "Data Mining"*



Hình 4. *Kết quả phân loại với chủ đề "Machine Learning"*



Hình 5. *Kết quả phân loại với chủ đề "Social Network"*

Kết quả thực nghiệm thể hiện hiệu quả của việc phân loại sử dụng các tập đặc trưng khác nhau được trình bày ở Bảng 2, 3 và 4. Đầu tiên, chúng ta xem xét hiệu quả của phương pháp ước lượng xác suất kích hoạt mới mà chúng tôi đề xuất. Đối với chủ đề "Data Mining" (ở Hình 3), hiệu quả của AP(MP+IS) chỉ thể hiện ở phân loại Random Forest. Tuy nhiên, hiệu quả của AP(MP+IS) lại thể hiện rõ nét trong kết quả phân loại đối với hai chủ đề "Machine Learning" và "Social Network" (Hình 4 và 5). Đối với cả ba thuật toán phân loại, chúng ta có thể thấy rằng AP(MP+IS) mang lại độ chính xác cao hơn so

với các phương pháp ước lượng xác suất kích hoạt cũ (AP) đó là chỉ dựa vào một trong hai thông tin meta-path hoặc thông tin văn bản.

Sự kết hợp giữa hai thông tin meta-path và văn bản trong việc ước lượng xác suất kích hoạt đã mang lại hiệu quả cao hơn so với việc chỉ sử dụng một trong hai thông tin. Các kết quả này thể hiện tính hợp lý bởi vì cả hai loại thông tin này đều chứa đựng thông tin quan trọng trong dự đoán lan truyền thông tin : meta-path chứa đựng các thông tin liên quan tới cấu trúc của các nút mạng còn thông tin văn bản chứa nội dung của các nút mạng. Do đó, khi kết hợp hai loại thông tin này lại với nhau sẽ mang lại hiệu quả cao hơn so với việc sử dụng một cách riêng lẻ.

Bên cạnh đó, các kết quả phân loại cũng thể hiện tầm quan trọng của đặc trưng về sở thích của tác giả cũng như hiệu quả của việc kết hợp giữa yếu tố tác động bên ngoài và yếu tố nội tại trong dự đoán lan truyền chủ đề. Chúng ta có thể thấy rằng sự kết hợp giữa AP(MP), AP(IS) hoặc AP(MP+IS) với AI hầu hết đều mang lại độ chính xác cao hơn so với việc chỉ sử dụng chúng một cách riêng lẻ. Đặc biệt, thuật toán Random Forest với sự kết hợp giữa hai đặc trưng AP(MP+IS) và AI đạt được độ chính xác cao nhất đối với tất cả các chủ đề lan truyền xem xét.

Các kết quả này thể hiện tính hợp lý bởi vì với sự tác động từ bên ngoài kết hợp với sở thích của cá nhân thì một tác giả có thể dễ dàng kích hoạt một chủ đề so với trường hợp chỉ chịu sự tác động từ bên ngoài nhưng bản thân không thích chủ đề đó hoặc trong trường hợp tác giả thích chủ đề đó nhưng lại không có cộng đồng xung quanh kích hoạt.

Tóm lại, kết quả thực nghiệm đã chứng minh rằng phương pháp ước lượng xác suất kích hoạt tổng hợp (AP(MP+IS)) với sự kết hợp giữa thông tin meta-path và thông tin văn bản mang lại độ chính xác cao hơn so với các phương pháp ước lượng cũ chỉ dựa vào một trong hai thông tin một cách riêng lẻ. Hơn nữa, sự kết hợp giữa nhân tố tác động bên ngoài và nhân tố nội tại trong dự đoán lan truyền chủ đề trong mạng học thuật đạt được độ chính xác tốt hơn so với chỉ sử dụng đặc trưng đơn lẻ. Đặc biệt, sự kết hợp giữa AP(MP+IS) và AI trong thuật toán Random Forest đạt được độ chính xác cao nhất.

6. KẾT LUẬN VÀ CÔNG VIỆC TƯƠNG LAI

Trong nghiên cứu này, chúng tôi đã đề xuất một cách tiếp cận mới trong việc dự đoán liệu rằng một tác giả có bị kích hoạt/lây nhiễm một chủ đề trong mạng học thuật hay không. Chúng tôi đã phân tích dựa trên hai thông tin đó là meta-path và thông tin văn bản (text), từ đó đề xuất một phương pháp ước lượng xác suất kích hoạt mới dựa vào cả hai thông tin trên. Xác suất này được xem xét như là nhân tố tác động từ bên ngoài mà có thể làm cho một tác giả kích hoạt/lây nhiễm một chủ đề lan truyền. Bên cạnh đó, chúng tôi cũng khai thác đặc trưng về sở thích của tác giả đối với chủ đề lan truyền, đây chính là yếu tố nội tại. Hơn nữa, chúng tôi đề xuất kết hợp cả hai đặc trưng trên trong viêc dự đoán lan truyền chủ đề trong mạng học thuật. Chúng tôi sử dụng học máy với các thuật toán phân loại để dự đoán liệu rằng một tác giả có kích hoạt/lây nhiễm một chủ đề T hay không dựa vào các đặc trưng được trích rút từ dữ liệu. Kết quả thực nghiệm cho thấy rằng những phương pháp mà chúng tôi đề xuất đều cải tiến được độ chính xác trong dự đoán lan

truyền chủ đề trong mạng học thuật. Trong tương lai, chúng tôi sẽ tiến hành các thực nghiệm trên nhiều tập dữ liệu khác với tập dữ liệu văn bản lớn hơn và sử dụng mô phỏng chủ đề để ước lượng độ tương tự giữa các tác giả.

TÀI LIỆU THAM KHẢO

[1] Akula, R., Yousefi, N., Garibay, I. (2019). DeepFork: Supervised Prediction of Information Diffusion in GitHub p. 12.

[2] Blei, D.M., Ng, A.Y., Jordan, M.I. (2002). Latent Dirichlet Allocation. In: Dietterich, T.G., Becker, S., Ghahramani, Z. (eds.) Advances in Neural Information Processing Systems 14, pp. 601–608. MIT Press, http://papers.nips.cc/paper/2070-latent-dirichlet-allocation.pdf.

[3] Bui, Q.V., Sayadi, K., Amor, S.B., Bui, M. (2017). Combining Latent Dirichlet Allocation and K-Means for Documents Clustering: Effect of Probabilistic Based Distance Measures. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) Intelligent Information and Database Systems. pp. 248–257. Lecture Notes in Computer Science, Springer International Publishing, Cham.

[4] Goldenberg, J., Libai, B., Muller, E. (2001). Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. Marketing Letters 12(3), 211–223, https://doi.org/10.1023/A:1011122126881.

[5] Granovetter, M. (1978). Threshold Models of Collective Behavior. American Journal of Sociology 83(6), 1420–1443, https://www.journals.uchicago.edu/doi/abs/10.1086/226707.

[6] Gui, H., Sun, Y., Han, J., Brova, G. (2014). Modeling Topic Diffusion in Multi-Relational Bibliographic Information Networks. In: CIKM.

[7] Ho, T.K.T., Bui, Q.V., Bui, M. (2018). Homophily Independent Cascade Diffusion Model Based on Textual Information. In: Nguyen, N.T., Pimenidis, E., Khan, Z., Trawiński, B. (eds.) Computational Collective Intelligence. pp. 134–145. Lecture Notes in Computer Science, Springer International Publishing, Cham.

[8] Kempe, D., Kleinberg, J., Tardos, V. (2005). Influential Nodes in a Diffusion Model for Social Networks. In: Caires, L., Italiano, G.F., Monteiro, L., Palamidessi, C., Yung, M. (eds.) Automata, Languages and Programming. pp. 1127–1138. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg.

[9] Kempe, D., Kleinberg, J.M., Tardos, V. (2003). Maximizing the spread of influence through a social network. In: KDD.

[10] Macy, M.W. (1991). Chains of Cooperation: Threshold Effects in Collective Action. American Sociological Review 56(6), 730–747, https://www.jstor.org/stable/2096252.

[11] Molaei, S., Babaei, S., Salehi, M., Jalili, M. (2018). Information Spread and Topic Diffusion in Heterogeneous Information Networks. Scientific Reports 8(1), 1–14, https://www.nature.com/articles/s41598-018-27385-2.

[12] Molaei, S., Zare, H., Veisi, H. (2019). Deep learning approach on information diffusion in heterogeneous networks. Knowledge-Based Systems p. 105153, http://www.sciencedirect.com/science/article/pii/S0950705119305076.

[13] Rosen-Zvi, M., Griffths, T., Steyvers, M., Smyth, P. (2012). The Author-Topic Model for Authors and Documents. arXiv:1207.4169 [cs, stat], http://arxiv.org/abs/1207.4169, arXiv: 1207.4169.

[14] Salton, G., Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523, http://www.sciencedirect.com/science/article/pii/0306457388900210
[15] Sun, Y., Han, J., Yan, X., Yu, P.S., Wu, T. (2011). Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In: In VLDB' 11.
[16] Varshney, D., Kumar, S., Gupta, V. (2014). *Modeling Information Diffusion in Social Networks Using Latent Topic Information*. In: Huang, D.S., Bevilacqua, V., Premaratne, P. (eds.) Intelligent Computing Theory. pp. 137–148. Lecture Notes in Computer Science, Springer International Publishing, Cham.
[17] Yang, H. (2008). *Mining social networks using heat diffusion processes for marketing candidates selection. ACM*, https://aran.library.nuigalway.ie/handle/10379/4164.

Title: INFORMATION PROPAGATION PREDICTION ON BIBLIOGRAPHIC NETWORK: A NEW APPROACH WITH AMALGAMATION OF EXTERNAL AND INTERNAL ELEMENTS

Abstract: In this research, we propose a novel approach to solve the topic's propagation prediction problem on a bibliographic network with a combination of external and internal factors. We use machine learning to predict the spread of a topic with a combination of different features. Firstly, we propose a new method to calculate activation probability from an active node to an inactive node based on both meta-path and textual information. This activation probability is considered as an external factor. In addition, we also explore the author's preference for the topic, which is considered as an intrinsic factor of the author that can lead to the "infection" of the topic. Finally, we amalgamate the activation probability feature and the author's preference feature in the topic's spreading prediction. Experiments were conducted on dissimilar topics of the bibliographic network datasets and demonstrated satisfactory results.

Keywords: Social Network, Bibliographic Network, Multi-Relation Network, Information Diffusion, Activation Prediction.