ĐẠI HỌC QUỐC GIA TP. HCM
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

PHẠM THẾ ANH PHÚ

NGHIÊN CỨU MÔ HÌNH KHAI THÁC MẠNG THÔNG TIN KHÔNG ĐỒNG NHẤT VÀ ỨNG DỤNG

Chuyên ngành Công Nghệ Thông Tin
Mã ngành: 62.48.02.01

TÓM TẮT LUẬN ÁN TIẾN SĨ CÔNG NGHỆ THÔNG TIN

TP. Hồ Chí Minh, tháng 04/2021

Công trình này được hoàn thành tại: **Trường Đại học Công nghệ Thông tin (UIT), Đại học Quốc gia TP.HCM**.

Người hướng dẫn khoa học 1: **PGS. TS Đỗ Phúc**

Phản biện độc lập 1: ...

Phản biện độc lập 2: ...

Luận án đã được bảo vệ trước Hội đồng chấm luận án họp tại: ...

Vào lúc ... giờ ... ngày ... tháng ... năm

Có thể tìm luận án tại:

- Thư viện Quốc gia Việt Nam.
- Thư viện Trường Đại học Công nghệ Thông tin, ĐHQG-HCM.

MỤC LỤC

CHƯƠNG 1: TỔNG QUAN VỀ LUẬN ÁN

1.1. Dẫn nhập

A. Sự phổ biến & đa dạng của nhiều mạng thông tin hiện nay trên toàn cầu

B. Phân tích và khai phá mạng thông tin giúp đem lại nhiều tri thức hữu ích cho con người

Hình 1-1. Sự phổ biến & tầm quan trọng của việc phân tích và khai phá mạng thông tin

Phân tích & khai phá mạng thông tin (Information Network Analysis & Mining – INAM) [1] [2] là một trong các chủ đề nghiên cứu quan trọng và đóng vai trò ứng dụng trong nhiều lĩnh vực khác nhau, điển hình như: phân tích mạng xã hội (social network analysis), xây dựng các hệ khuyến nghị (recommedation system) dựa trên dữ liệu mạng thông tin, truy hồi dữ liệu trên mạng thông tin (networked data retrieval), hay phân tích các dạng dữ liệu có cấu trúc mạng thông tin như gene, protein (trong tin sinh học), cấu trúc & thành phần của phân tử (hóa học), v.v. Bên cạnh đó, lĩnh vực khai phá mạng thông tin còn đặc biệt được quan tâm trong thời gian gần đây vì nó được ứng dụng trong việc phân tích hành vi và xu thế của con người, thông qua sự tương tác của họ trên các mạng xã hội phổ biến hiện nay như: Facebook[1], Twitter[2], Weibo[3], Instagram[4], IMDb[5] (minh họa Hình 1-1).

Bản chất liên kết của dữ liệu và tầm quan trọng của khai phá mạng thông tin. Qua ví dụ trên, có thể cho thấy rằng tầm quan trọng của việc phân tích và khai phá mạng thông tin là hoàn toàn không thể phủ nhận được. Có thể thấy rằng hầu hết các dạng và cấu trúc dữ liệu mà chúng ta tiếp xúc mỗi ngày đều ít/nhiều tồn tại ở các dạng liên kết với nhau, điển hình như: mạng Internet (WWW), các trang mạng xã hội (Facebook, Twitter, MySpace, Weibo, v.v.), các mạng học thuật (DBLP, DBIS, v.v.), các bách khoa toàn thư mở (encyclopedia) ở dạng các đồ thị tri thức (Wikipedia, YAGO, v.v.), hay các diễn đàn, trang mạng tin tức, v.v. Và sự liên kết giữa các thực thể dữ liệu trong mạng thông tin giúp hỗ trợ và làm giàu thêm ngữ nghĩa cho chính nó cũng như các thực thể mà nó liên kết đến, ví dụ như sự liên kết/tham chiếu thông qua các siêu liên kết (hyperlink) giữa các website, các mối quan hệ giữa các người dùng với nhau trong mạng xã hội, các khái niệm

[1] Mạng XH Facebook: https://www.facebook.com/
[2] Mạng XH Twitter: https://twitter.com/
[3] Mạng XH Weibo: https://www.weibo.com
[4] Mạng XH hình ảnh Instagram: https://www.instagram.com/
[5] Mạng thông tin phim IMDb: https://www.imdb.com/

(concept) có các quan hệ tham chiếu lẫn nhau trong các bách khoa toàn thư, v.v. Bên cạnh đó, ta có thể thấy rằng bản chất của các "liên kết"/"cung"/"quan hệ" trong mạng thông tin không chỉ hỗ trợ làm giàu ngữ nghĩa cho các "thực thể"/"nút" trong mạng thông tin mà nó mà bản thân nó cũng mang nhiều thông tin quan trọng hàm chứa bên trong và làm cho nó khác biệt với các quan hệ khác.

Động lực của luận án. Kích thước lớn, tốc độ tăng trưởng nhanh và sự đa dạng trong cấu trúc được xem là các yếu tố thách thức nhưng cũng được coi là tiềm năng cho việc phát triển hữu ích cho con người trong nhiều lĩnh vực dựa trên việc phân tích và khai phá các tri thức của các mạng thông tin. Đặc biệt đối với sự đa dạng trong cấu trúc của các mạng thông tin hiện nay có thể được xem là một trong các thách thức lớn nhất cho lĩnh vực khai phá mạng thông tin. Sự đa dạng trong loại của các nút cũng như mối quan hệ giữa chúng khiến các mô hình khai phá truyền thống như P-PageRank, SimRank, v.v. không còn có thể áp dụng một cách hiệu quả nữa. Do đó một trào lưu mới trong khai phá mạng thông tin, được gọi là khai phá mạng thông tin không đồng nhất được ra đời.

1.2. Khái quát về bài toán và động lực của luận án

*1.2.1. Khai phá mạng thông tin đồng nhất (Homogeneous Information Network - HoIN) và không đồng nhất (*Heterogeneous *Information* N*etwork - HIN).*

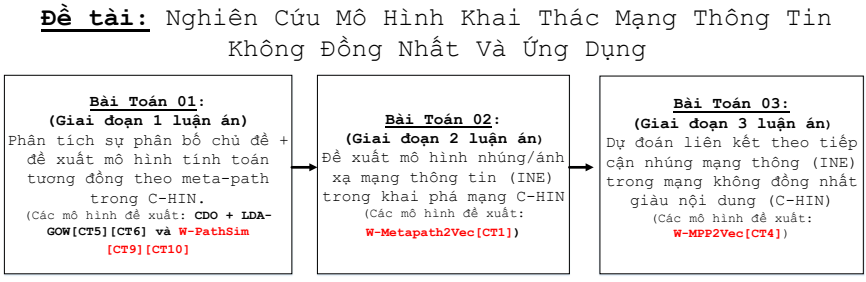
Với các mô hình phân tích và khai phá mạng thông tin truyền thống, thì việc đánh giá mức độ liên kết giữa các nút trong mạng thông tin có vai trò quan trọng hơn các yếu tố khác. Việc xác định mức độ tương đồng hay xếp hạng các nút sẽ hầu hết dựa trên số lượng/mật độ của các liên kết giữa nó với các nút khác (điển hình P-PageRank, SCAN và SimRank). Và hầu như các mô hình truyền thông trên bỏ qua sự khác biệt trong loại giữa các nút và mối quan hệ giữa chúng (chỉ có một loại nút và quan hệ duy nhất) – hướng tiếp cận này được gọi là khai phá mạng thông tin đồng nhất (homogeneous: đơn nhất/đồng nhất). Tuy nhiên trong thực tế thì cấu trúc của các mạng thông tin rất phức tạp và đa dạng với sự tham gia của nhiều loại nút và các liên kết khác nhau, ví dụ như các mạng học thuật (DBLP, DBIS, v.v.) thì có nhiều loại nút như: tác giả (author), bài báo (paper), hội nghị/tạp chí (venue/journal), v.v. các mạng xã hội (Facebook và Twitter), tin tức (VnEprress và BBC) thì có các loại nút: người dùng (user), bài viết (post), bình luận (comment) hay nhóm (group). Giữa các nút sẽ có một hay nhiều loại liên kết khác nhau, ví dụ như các mạng học thuật: tác_giả $\xrightarrow{\text{viết}}$ bài_báo, bài_báo $\xrightarrow{\text{nộp/xuất\_bản}}$ hội_nghị/tạp_chí, hay các mạng xã hội: người dùng $\xrightarrow{\text{bạn\_bè}}$ người dùng, người dùng $\xrightarrow{\text{tham\_gia}}$ nhóm. Sự đa dạng trong loại của nút và các mối quan hệ gây nhiều khó khăn cho việc áp dụng các mô hình phân tích và khai phá mạng thông tin truyền thống. Do đó, cần có một hướng tiếp cận mới, trong đó việc phân tích và khai phá mạng thông tin cần chú trọng đến sự khác biệt trong loại của các nút và mối quan hệ giữa chúng (*heterogeneous*: *đa dạng/đa tạp*), hướng tiếp cận này được gọi là phân tích và khai phá mạng thông tin không đồng nhất (HIN). Nền tảng và cơ sở lý thuyết về việc phân tích và khai phá mạng

thông tin không đồng nhất HIN – lần đầu được đề xuất bởi Sun Y. & J. Han (2011), trong đó các mô hình được đề xuất phải đảm bảo khả năng phân tách được sự khác biệt trong loại của các thực thể và liên kết cũng như đảm toàn được ngữ nghĩa của các mối quan hệ giữa các nút/thực thể.

1.2.2. Các hạn chế còn tồn tại và động lực thực hiện luận án

Trong hầu hết các hướng tiếp cận của bài toán phân tích và khai phá mạng thông tin không đồng nhất (HIN), bao gồm cả hướng tiếp cận mới nhất là học mô hình biểu diễn (NRL) thì hầu như chỉ chú trọng vào việc phân tích các mối quan hệ giữa các thực thể/nút trong HIN hơn là quan tâm đến sự ảnh hưởng của nội dung và chủ đề giữa các thực thể/nút trong các mạng thông tin có giàu nội dung hay còn gọi là: **Content-based HIN – C-HIN**. Có thể thấy trên thực tế thì hầu như tất cả các mạng thông tin phổ biến hiện hay như các mạng xã hội (Facebook hay Twitter), các mạng học thuật (DBLP hay DBIS) hay các diễn đàn (forum), tin tức online, v.v. đều chứa một lượng lớn các thực thể/nút ở dạng văn bản, và các thực thể/nút giàu nội dung này (content-based nodes) đóng vai trò rất phổ biến và xuất hiện trong hầu hết các quan hệ ngữ nghĩa (mô tả ở dạng các meta-paths) giữa các thực thể cùng loại. Sự tương đồng trong nội dung, chủ đề của các nút giàu ngữ liệu này cũng đóng vai trò quan trọng trong việc đánh giá sự tương đồng giữa các nút được xét thông qua các meta-path mà chúng xuất hiện.

1.3. Mục tiêu, phạm vi nghiên cứu của luận án



Hình 1-2. Tổng quan về nội dung và phạm vi nghiên cứu của luận án

Toàn bộ luận án sẽ được chia thành 3 bài toán chính và thực hiện trong từng giai đoạn của luận án, như sau (minh họa Hình 1-2):

1.3.1. Bài toán 1: Tính toán tương đồng trong mạng thông tin không đồng nhất giàu nội dung (C-HIN)

Trong giai đoạn đầu của luận án, NCS và GVHD tập trung vào việc xây dựng nền tảng lý thuyết cho việc khám phá sự phân bố của chủ đề trong mạng thông tin C-HIN, để từ đó kết hợp sự tương đồng trong chủ đề với mối quan hệ giữa các thực thể/nút nhằm đưa ra những mô hình cải tiến phù hợp cho việc khai phá mạng thông tin giàu nội dung thông qua mô hình chủ đề LDA, để hỗ trợ cho việc phân tích sự phân bố của các chủ đề có trong các nút giàu ngữ liệu của mạng thông tin.

Sự phân bố chủ đề của các nút ở dạng văn bản sau đó được sử dụng để xác định mức độ tương đồng trong chủ đề giữa các nút trong mạng thông tin dựa trên meta-path, với các mô hình cải tiến đề xuất, bao gồm: mô hình **W-PathSim** (công bố [**CT10**]) cùng với các mô hình mở rộng: DW-PathSim (công bố tại [**CT9**]), ComRank và TopCPathSim (công bố **[CT6]**).

1.3.2. Bài toán 2: Tiếp cận nhúng mạng thông tin (INE/NRL) trong ngữ cảnh mạng thông tin không đồng nhất giàu nội dung

Từ các kết quả nghiên cứu trong giai đoạn 1, NCS & GVHD đề xuất kết hợp với hướng tiếp cận nhúng/ánh xạ các thực thể/nút của mạng C-HIN về môi trường không gian vector, quá trình rút trích các đặc trưng của nút để huấn luyện mô hình học sẽ được áp dụng nguyên lý bước đi ngẫu nhiên dựa trên meta-path theo hướng tiếp cận tương đồng trong chủ đề (topic-driven meta-path-based random walk). Để hiện thực hóa các ý tưởng, NCS đã xây dựng và đề xuất mô hình **W-Metapath2Vec[CT1]** và **W-Metagraph2Vec[CT2]**, kế thừa chính từ ý tưởng của mô hình W-PathSim đã xây dựng trong giai đoạn 1. Mô hình W-Metapath2Vec hỗ trợ cho việc biểu diễn các nút tương đồng nhau, dựa trên hai tiêu chí là mức độ liên kết và tương quan trong chủ đề trên mạng C-HIN, về dạng các vector tương đương nhau. Từ đó hỗ trợ cho việc giải quyết các bài toán cốt lõi của INAM như tìm kiếm tương đồng (node similarity search), gom cụm (node clustering), phân lớp (node classification), v.v.

1.3.3. Bài toán 3: Tiếp cận nhúng mạng thông tin (INE/NRL) trong việc giải quyết bài toán dự đoán liên kết trên mạng không đồng nhất giàu nội dung (C-HIN)

Trong phần nội dung cuối của luận án này NCS & GVHD kết hợp các thành quả của hai phần nội dung trước để đề xuất một mô hình ứng dụng cho việc giải quyết bài toán dự đoán liên kết giữa các nút trong mạng C-HIN. Mô hình dự đoán liên kết được xây dựng dựa trên hướng tiếp cận INE nhằm hỗ trợ cho việc ánh xạ các nút có khả năng cao xuất hiện các liên kết trong mạng thông tin về các vector số thực tương đương nhau với số chiều quy định. Việc xây dựng và rút trích các đặc trưng giữa các cặp nút - dùng cho huấn luyện mô hình học biểu diễn các nút trong mạng thông – sẽ dựa trên việc đánh giá hình mẫu liên kết cũng như sự tương quan trong chủ đề giữa các thực thể. Để hiện thực hóa mô hình trên, NCS & GVHD đã đề xuất và xây dựng mô hình **W-MMP2Vec[CT4]** nhằm hỗ trợ cho việc giải quyết bài toán dự đoán liên kết trên mạng C-HIN. Ngoài ra NCS & GVHD cũng dựa và ý tưởng và các kết quả đã đạt được của các mô hình: **W-Metapath2Vec, W-Metagraph2Vec** và **W-MMP2Vec** để xây dựng mô hình W-Com2Vec (công bố **[CT3]**) nhằm giải quyết bài toán nhận diện & biểu diễn cộng đồng trên mạng thông tin không đồng nhất.

1.4. Bố cục của luận án

Nội dung của luận án sẽ được tổ chức thành 6 chương chính & các phần phụ khác, mỗi chương sẽ bao gồm các phần nội dung như sau:

- **Chương 1 - Tổng quan về luận án**: trong chương này, NCS trình bày tổng quan về luận án cũng như sơ nét về các hướng tiếp cận phổ biến trong khai phá mạng thông tin hiện nay. Để từ đó đưa ra các nhận định về các hạn chế còn tồn tại cần phải giải quyết. Thông qua đó xác định được đối tượng cũng như phạm vi nghiên cứu của luận án.
- **Chương 2 – Cơ sở lý thuyết & các mô hình liên quan**: trong nội dung chương này, NCS tập trung trình bày tổng quan về nền tảng lý thuyết của khai phá mạng thông tin, tập trung chuyên sâu về khai phá mạng thông tin không đồng nhất. NCS giới thiệu sơ nét về lịch sử phát triển của các mô hình/hướng tiếp cận phổ biến trong cả hai trào lưu khai phá mạng thông tin đồng nhất (HoIN) và khai phá mạng không đồng nhất (HIN).
- **Chương 3 – Tìm kiếm tương đồng trong mạng thông tin giàu nội dung (C-HIN), mô hình W-PathSim**: trong nội dung của chương 2, NCS trình bày về ý tưởng của việc áp dụng mô hình chủ LDA, trong việc phân tích sự phân bố các chủ đề tiềm ẩn của các thực thể giàu nội dung trong mạng thông tin không đồng nhất. Để từ đó làm nền tảng cho việc xây dựng mô hình tìm kiếm tương đồng theo hướng tiếp cận nội dung/tương quan chủ đề trong mạng thông tin không đồng nhất, với mô hình đề xuất là W-PathSim làm nền tảng có các mô hình đề xuất kết tiếp bao gồm: ComRank[**CT8**] và TopCPathSim[**CT7**].
- **Chương 4 - Nhúng mạng thông tin (INE/NRL) không đồng nhất giàu nội dung, mô hình W-Metapath2Vec**: Trong nội dung chương này, NCS trình bày về tổng quan về lịch sử phát triển cũng như hướng tiếp cận nhúng mạng thông tin (INE) trong khai phá mạng thông tin đồng nhất và không đồng nhất. Nội dung trọng tâm của chương 3 này sẽ là các trình bày chi tiết về ý tưởng cũng như cơ chế của hai mô hình cải tiến là W-Metapath2Vec và mô hình W-Metagraph2Vec được phát triển dựa trên ý tưởng của W-Metapath2Vec.
- **Chương 5 - Dự đoán liên kết trên mạng thông tin không đồng nhất dựa trên INE, mô hình W-MMP2Vec**: trong chương này, NCS trình bày về hướng tiếp cận giải quyết bài toán ứng dụng trên mạng thông tin, đó là dự đoán liên kết (link prediction), dựa trên tiếp cận INE. NCS đề xuất mô hình W-MMP2Vec**[CT4]** là một mô hình được kế thừa lại các kết quả của các mô hình trước đó, bao gồm: W-PathSim, W-Metapath2Vec và W-Metagraph2Vec. Trong nội dung chương này, NCS tập trung trình bày các ý tưởng cũng như cơ chế và phương pháp biểu diễn các nút trong mạng thông tin của mô hình W-MMP2Vec theo hướng tiếp cận tương quan trong chủ đề.
- **Chương 6 - Kết luận, hạn chế & hướng phát triển**: trong nội dung của chương này, NCS trình bày tổng quát và kết luận về các kết quả, đóng góp cũng như hướng phát triển của luận án.

CHƯƠNG 2: CƠ SỞ LÝ THUYẾT & CÁC MÔ HÌNH LIÊN QUAN

2.1. Cơ sở lý thuyết về khai phá mạng thông tin

2.1.1. Tổng quan về phân tích và khai phá mạng thông tin (INAM)

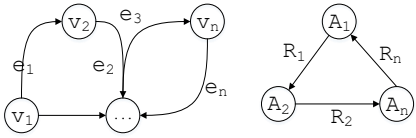
Đi cùng với sự phát triển của Internet thì phân tích và khai phá mạng thông tin (Information Network Analysis and Mining – INAM) [1] [2] [3] [4] được coi là một trong các lĩnh vực đóng vai trò then chốt trong hầu hết các nền tảng hệ thống cũng như các ứng dụng hỗ trợ cho những nhu cầu thiết yếu của con người. INAM có thể được coi là một trong các lĩnh vực con quan trọng của lĩnh vực khai phá dữ liệu (data mining), bởi bản chất của dữ liệu là luôn có sự gắn kết với nhau. Từng thực thể riêng biệt của một khối dữ liệu luôn có ít hay nhiều các mối quan hệ giữa chúng và hầu như không có thực thể nào tồn tại một cách độc lập và riêng biệt trên thực tế. Điển hình như dữ liệu mạng xã hội: Facebook, Twitter, v.v. với nhiều thực thể dữ liệu là người dùng, nhóm, v.v. liên kết với nhau, hay mạng lưới các website (WWW) được liên kết với nhau bởi các siêu liên kết (hyperlinks), v.v. Bởi do bản chất tự nhiên là kết nối của các thực thể trong tập dữ liệu, nên các mối quan hệ cũng sẽ đóng vai trò quan trọng nhất định cũng như chứa đựng những thông tin tri thức quý giá. Khởi thủy của INAM có thể được coi là một phân nhánh của lĩnh vực khai phá dữ liệu có liên kết (networked data mining), với hàng loạt các thuật toán khá nổi tiếng như: PageRank, HITS, SCAN, v.v. hỗ trợ cho việc khai phá dữ liệu hiệu một cách quả từ các CSDL có sự liên kết, điển hình như WWW, mạng xã hội (social networks), mạng trích dẫn (citation networks), v.v. Tuy nhiên càng về sau, thì độ phức tạp trong cấu trúc cũng như kích thước của các khối dữ liệu có sự liên kết càng trở nên quá lớn với số lượng các liên kết cũng như loại của các liên kết ngày càng đa dạng hơn, gây ra nhiều thách thức cho các thuật toán hiện hành. Các nhà khoa học đã thay đổi góc nhìn cũng như đánh giá tầm quan trọng đối với các khối dữ liệu có rất nhiều liên kết cũng như như không có cấu trúc nhất định và gọi chung các loại dữ liệu dạng này là "mạng thông tin" (Information Network - IN).

2.1.2. Tổng quan về khai phá mạng thông tin không đồng nhất (heterogeneous information network mining)

2.1.2.1. Lý thuyết về mạng thông tin và các khái niệm tổng quát

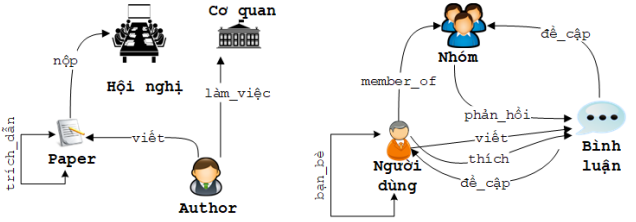
Lý thuyết về khai phá dữ liệu từ mạng thông tin lần đầu được đề cập đến trong các công trình nghiên cứu của L. Page & S. Brin (1999) [5] trong quá trình đề xuất thuật toán PageRank nổi tiếng hỗ trợ việc xếp hạng các websites thông qua đánh giá số lượng liên kết mà chúng được kết nối đến (hay còn gọi là vote). Tiếp nối các kết quả đạt được từ L. Page & S. Brin trong mô hình PageRank, hàng loạt các mô hình tính toán tương đồng và xếp hạng các nút trong mạng thông tin đã được đề xuất, điển hình như: HITS [6], Persionalized PageRank (P-PageRank) [7], SimRank [8], SCAN [9],v.v. đạt được nhiều bước tiến trong việc giải quyết các bài toán liên quan đến tính toán tương đồng (similarity measure) và xếp hạng (ranking) các nút trong mạng thông tin. Tuy nhiên các mô hình này chỉ phù hợp

cho việc khai phá các mạng thông tin đơn nhất (homogeneous), tức xem tất cả các loại của thực thể và quan hệ là một.



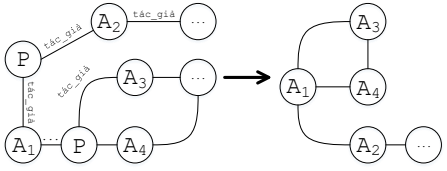
A. Minh họa mô hình tổng quát mạng thông tin (IN) B. Minh họa tổng quát về lược đồ liên kết (network schema)

Hình 2-1. Minh họa về mạng thông tin (IN) và lược đồ liên kết (network schema)



A. Minh họa lược đồ liên kết (NG) trong mạng DBLP B. Minh họa lược đồ liên kết (NG)trong mạng xã hội Facebook

Hình 2-2. Minh họa lược đồ liên kết (network schema) của một số mạng thông tin phổ biến



A. Mạng thông tin không đồng nhất (HIN) thể hiện mới quan hệ đồng tác giả B. Mạng thông tin đồng nhất (HoIN) thể hiện mới quan hệ đồng tác giả

Hình 2-3. Minh họa việc chuyển đổi từ mạng thông tin không đồng nhất sang đồng nhất (HIN2HoIN)

Định nghĩa 1. Mạng thông tin (information network) [2] [8]***:*** *được định nghĩa là một đồ thị có hoặc không hướng, được ký hiệu:* $G = (V, E)$*, với hai hàm ánh xạ (mapping function) là:* $\phi: V \to A$ *và* $\psi: E \to R$*, trong đó A là tập các loại của nút/thực thể và R là tập các loại quan hệ giữa các nút/thực thể trong mạng thông tin. Trong đó:*

- ***Định nghĩa 1-a. Mạng thông tin đồng nhất (homogeneous information network - HoIN):*** *là dạng mạng thông tin chỉ có duy nhất một loại thực thể* ($|A| = 1$) *và loại quan hệ* ($|R| = 1$).
- ***Định nghĩa 1-b. Mạng thông tin không đồng nhất (heterogeneous information network - HIN):*** *là dạng mạng thông tin có số lượng loại của thực thể và quan hệ luôn nhiều hơn 1* ($|A| > 1, |R| > 1$)*. Với mỗi thực thể/nút trong mạng thông tin, ký hiệu:* v, $v \in V$*, sẽ thuộc về một loại thực thể/nút cụ*

thể nào đó, với: $\phi(v) \in A$. Tương tự như vậy, với mỗi cung liên kết hai thực thể/nút - ký hiệu: $e, e \in E$, sẽ thuộc về một loại cụ thể nào đó, với: $\psi(e) \in R$.

Định nghĩa 2. Lược đồ liên kết (network schema - NG) [2] [8]***:*** *được dùng để mô tả cấu trúc của một mạng thông tin, $G = (V, E)$, ký hiệu: $T_G = (A, R)$. Tương tự như lược đồ liên kết và thực thể kết hợp ER (entity-relation) trong lý thuyết CSDL bảng quan hệ, lược đồ liên kết mô tả các hình mẫu liên kết giữa các loại thực thể và mối quan hệ giữa chúng. Lược đồ liên kết đóng vai trò quan trọng trong việc giúp người dùng có cái nhìn tổng quát về cấu trúc của mạng thông tin.*

Một mạng thông tin tổng quát sẽ bao gồm tập hợp của các nút/thực thể và mối quan hệ giữa chúng (minh họa Hình 2-1-A) tùy theo ngữ cảnh của việc định nghĩa mạng thông tin mà tập các loại của thực thể và loại quan hệ sẽ thay đổi. Một mạng thông tin sẽ có hai loại chính, bao gồm: mạng thông tin đồng nhất (HoIN) (xem [định nghĩa 1-a]) và không đồng nhất (HIN) (xem [định nghĩa 1-b]). Để mô tả cấu trúc của một mạng thông tin đặc biệt là các mạng không đồng nhất, ta dùng lược đồ liên kết (network schema) (minh họa Hình 2-1-B) để mô tả các mối liên kết giữa các loại thực thể và mối quan hệ giữa chúng với nhau (xem [định nghĩa 2]). Giữa hai loại thực thể cùng hay khác loại sẽ có một hay nhiều loại liên kết khác nhau (minh họa Hình 2-2), điển hình như giữa các người dùng trong mạng xã hội (Facebook, Twitter) sẽ có nhiều loại liên kết ví dụ: người_dùng $\xrightarrow{\text{bạn\_bè}}$ người_dùng, người_dùng $\xrightarrow{\text{người\_thân}}$ người_dùng, v.v. do đó việc định nghĩa lược đồ liên kết sẽ giúp người dùng có thể phân biệt được sự khác nhau giữa các loại liên kết cũng như ngữ nghĩa của chúng ngoài ra nó còn có tác dụng hỗ trợ cho việc định nghĩa các meta-path sau này phục vụ cho việc khai phá mạng thông tin.

2.1.2.2. Các hạn chế của hướng tiếp cận khai phá mạng thông tin đồng nhất (homogeneous information network mining)

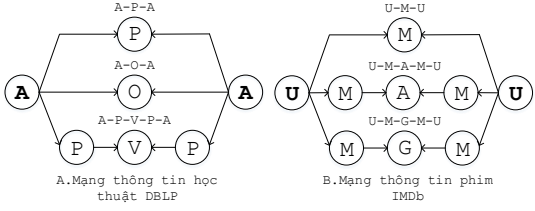
Như đã đề cập trong các phần trên về sự phức tạp cũng như đa dạng trong loại của các nút và quan hệ giữa chúng trong thực tế. Trong quá khứ, hầu hết các thuật toán được đề xuất để giải quyết các bài toán cơ bản của INAM đều không quan tâm đến sự khác biệt trong loại của các nút và mối quan hệ giữa chúng. Điều này dẫn đến các tranh cãi liên quan đến việc làm thế nào để có thể khai thác một các hiệu quả các mạng thông tin trong thực tế mà các thực thể thì có thể cùng hoặc hoàn toàn khác loại nhau, ví dụ: ta không thể so sánh tương đồng giữa người dùng (user) với các bình luận (comment) trong mạng XH, hoặc xếp hạng các tác giả (author) với các bài báo (paper) trong mạng học thuật DBLP. Ngoài ra, một số hướng giải pháp khác được đề xuất như chuyển đổi các mạng thông tin không đồng nhất thành dạng đồng nhất (HIN2HoIN) bằng cách xóa bỏ đi tất cả các loại nút và mối quan hệ khác trừ các quan hệ được xét để biến mạng thông tin chỉ còn một loại nút và mối quan hệ, sau đó các thuật toán cũ dành cho HoIN sẽ được áp dụng để giải quyết các bài toán INAM như cũ. Lấy một ví dụ về bài toán tìm các tác giả (author) tương đồng trong mạng thông tin học thuật DBLP thông qua mối quan hệ đồng tác giả (co-authorship). Ta sẽ xóa toàn bộ các loại nút khác như: bài báo (paper), hội nghị/tạp chí (venue/journal), v.v. cùng các mối quan hệ khác

ngoài quan hệ đồng tác giả, biểu diễn dạng meta-path là A-P-A (minh họa Hình 2-3). Tuy nhiên việc xóa bỏ đi các nút/thực thể và mối quan hệ khác giữa chúng để tiện cho việc áp dụng các thuật toán khai phá mạng thông tin cho HoIN, sẽ gây ra các hạn chế lớn, như sau:

- Việc xóa bỏ đi các nút và mối liên kết khác loại trong mạng thông tin sẽ ít nhiều gây ra việc mất mát dữ liệu cũng như ngữ nghĩa của các mối liên kết giữa các nút trong mạng thông tin. Hơn thế nữa việc xóa bỏ các nút cũng như mối quan hệ cũng phá vỡ tính đầy đủ và vẹn toàn trong cấu trúc của toàn mạng thông tin.
- Ngoài ra, việc chuyển đổi từ mạng thông tin không đồng nhất sang dạng đồng nhất, HIN2HoIN cũng sẽ gây ra các thách thức liên quan đến việc giảm thiểu độ chính xác của kết qua đầu ra do việc xóa bỏ đi rất nhiều các mối quan hệ của loại nút được xét. Ví dụ để tinh toán tương đồng giữa hai tác giả trong mạng học thuật DBLP thì ngoài các mối quan hệ đồng tác giả, ta còn hàng loạt các mối quan hệ cũng không kém phần quan trọng khác, ví dụ như quan hệ cùng xuất bản/công bố 1 công trình tại một số hội nghị/tạp chí (A-P-V-P-A), quan hệ đồng nghiệp (A-O[organization]-A), v.v. Do đó, nếu chỉ xét các mối quan hệ mục tiêu mà bỏ đi các loại quan hệ khác sẽ làm giảm chất lượng kết quả đầu ra của quá trình khai phá mạng thông tin.
- Quá trình chuyển đổi HIN2HoIN cũng sẽ gây tốn kém chi phí tính toán, lưu trữ cũng như thời gian thực thi cho các mô hình được khai phá. Hiển nhiên rằng, chúng ta sẽ phải dành một vùng lớn bộ nhớ tương đối lớn cũng như thời gian tính toán dài hơn cho việc sinh ra một mạng thông tin thứ 2 với chỉ các loại thực thể cùng với cac mối quan hệ được xét.

2.2. Tính toán tương đồng trên mạng không đồng nhất theo meta-path & các thuật toán phổ biến

Để giải quyết các thách thức liên quan đến sự đa dạng trong loại của các thực thể và quan hệ trong quá trình khai phá mạng thông tin không đồng nhất, Sun Y. & J. Han (2011) [2] [10] đã đề xuất một phương pháp mới trong khai phá dữ liệu mạng thông tin, đó là sử dụng meta-path. Meta-path (xem [định nghĩa 3]), hay còn gọi là "siêu liên kết" là một dạng hình mẫu được dùng để mô tả các mối quan hệ ngữ nghĩa giữa các nút trong mạng thông tin. Meta-path có thể được coi là một chuỗi của các nút cùng các mối quan hệ giữa chúng nhằm để chỉ mối quan hệ ngữ nghĩa giữa hai nút/thực thể được xét.



Hình 2-4. Minh họa về các meta-paths trong mạng thông tin DBLP và IMDb

Định nghĩa 3. Meta-path ($\mathcal{P}$) [2] [4]***:*** *là một hình mẫu liên kết giữa hai thực thể cùng loại, ký hiệu: $\mathcal{P}$, được định nghĩa dựa trên lược đồ liên kết $T_G = (A, R)$, hay nói cách khác thì một meta-path ($\mathcal{P}$) là một phần của lược đồ liên kết T_G. Một meta-path có chiều dài (l) có dạng: $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \dots \xrightarrow{R_l} A_{l+1}$, hoặc có thể viết là: $R_1 \circ R_2 \dots \circ R_l$, dùng để mô tả các loại quan hệ giữa hai thực thể cùng loại A_1 và A_{l+1}. Một đường đi giữa hai nút/thực thể cùng loại với nhau theo một meta-path ($\mathcal{P}$) sẽ được gọi là một "**path instance**".*

2.3. Giới thiệu về mô hình PathSim

Trong hầu hết các thuật toán tính toán tương đồng giữa các thực thể trong HIN dựa trên meta-path thì PathSim [10] được coi là một thuật toán nền tảng quan trọng và là nền tảng cho hầu hết các thuật toán tính toán tương đồng khác như: HeteSim [11], LSH-HeteSim [12], v.v. Để tính toán được độ tương đồng giữa hai thực thể cùng loại, điều tiên quyết là ta phải xác định được mật độ liên kết giữa các thực thể, Sun Y. & J. Han (2011) [10] đề xuất hướng tiếp cận để xác định trọng số liên kết giữa hai thực thể cùng loại (x) và (y), dựa trên meta-path $(\mathcal{P})$. Cho hai thực thể cùng loại (x) và (y) trong mạng thông tin: $G = (V, E)$, với $(x, y \in V)$ và $\phi(x) = \phi(y)$, độ tương đồng giữa hai thực thể (x) và (y) theo meta-path $(\mathcal{P})$, ký hiệu: $\text{PathSim}(x \rightsquigarrow y, \mathcal{P})$, được tính như sau (xem [công thức 2.1]):

$$\text{PathSim}(x \rightsquigarrow y, \mathcal{P}) = \frac{2 \times \text{PC}(x \rightsquigarrow y, \mathcal{P})}{\text{PC}(x \rightsquigarrow ., \mathcal{P}) + \text{PC}(y \rightsquigarrow ., \mathcal{P})} \quad (2.1)$$

Trong đó,

- $\text{PC}(x \rightsquigarrow y, \mathcal{P})$, là tổng trọng số của các path instances được xác định giữa hai thực thể cùng loại (x) và (y), theo meta-path $\mathcal{P}$.
- $\text{PC}(x \rightsquigarrow ., \mathcal{P})$, là tổng trọng số của các path instances được xác định giữa (x) và tất cả các thực thể cùng loại khác, theo meta-path $\mathcal{P}$.
- $\text{PC}(y \rightsquigarrow ., \mathcal{P})$, là tổng trọng số của các path instances được xác định giữa (y) và tất cả các thực thể cùng loại khác, theo meta-path $\mathcal{P}$.

2.3.1. So sánh ưu/nhược điểm của các mô hình tính toán tương đồng phổ biến áp dụng cho mạng HIN

Thông qua việc tìm hiểu các mô hình tính toán tương đồng trên HIN (từ 2011 → 2016), NCS đưa ra bảng nhận định so sánh tính năng (ưu/nhược điểm) các mô hình tính toán tương đồng theo meta-path áp dụng cho mạng không đồng nhất (HIN), như sau (xem Hình 2-5):

2.3.2. Các hạn chế của tiếp cận hướng liên kết (link-based) trong tính toán tương đồng trên HIN

Có thể thấy rằng, trong hướng tiếp cận áp dụng meta-path để giải quyết các bài toán khai phá mạng HIN, điển hình như tính toán tương đồng trong thuật toán PathSim thì hầu như mức độ tương đồng giữa các nút cùng loại với nhau trong HIN phụ thuộc chủ yếu vào số lượng các path instances ở giữa chúng và hầu như

không hề xét đến các yếu tố khác. Việc chỉ xét đến mức độ liên kết mà bỏ qua các thuộc tính quan trọng khác, điển hình như nội dung, chủ đề, v.v. có thể ảnh hưởng đến kết quả tìm kiếm tương đồng cũng như ý nghĩa của kết quả trả về.

	Phân biệt loại quan hệ	Phân biệt loại thực thể/nút	Không cần tập huấn luyện	Phù hợp cho meta-path có chiều dài lớn	Có cơ chế tỉa/loại bỏ bớt cặp ứng viên	Phù hợp cho mạng HIN kích thước lớn	Có phân tích trọng số quan hệ/ meta-path	Có quan tâm đến khía cạnh nội dung
PCRW [20] (2010): $h_{E_0,\mathcal{P}}(e)$	✓							
PathSim [11] (2011): PathSim (x ⇝ y, $\mathcal{P}$)	✓	✓	✓					
HeteSim [21] (2014): HeteSim(s, t|$\mathcal{P}$)	✓	✓	✓					
AvgSim [22] (2014): AvgSim(s, t|$\mathcal{P}$)	✓	✓	✓	✓		✓ (phiên bản AvgSim cài đặt trên Hadoop)		
NetSim [23] (2015): $S^i(u, v)$	✓	✓	✓		✓		✓	
RelSim [24] (2016): RS(r, r′)	✓	✓			✓		✓	

Hình 2-5. Phân tích các tính năng của các mô hình nổi bật áp dụng cho các bài toán khai phá dữ liệu trên HIN

CHƯƠNG 3: TÌM KIẾM TƯƠNG ĐỒNG TRONG MẠNG THÔNG TIN GIÀU NỘI DUNG, MÔ HÌNH W-PATHSIM

Trong nội dung của chương 3 này, NCS sẽ tập trung trình bày về nền tảng của bài toán tìm kiếm tương đồng trên mạng thông tin không đồng nhất (HIN) dựa trên meta-path, giới thiệu về thuật toán PathSim so sánh và nhận định các hạn chế của hướng tiếp cận của việc chỉ dựa hoàn toàn mối quan hệ giữa các nút trong việc đánh giá mức độ tương đồng giữa chúng trong mạng thông tin không đồng nhất giàu nội dung (C-HIN).

3.1. Tương đồng trong chủ đề giữa các thực thể giàu ngữ liệu văn bản

3.1.1. Áp dụng mô hình chủ đề LDA trong khám phá sự phân bố của chủ đề trong các thực thể ở dạng văn bản

Trong bài toán 1 của luận án, NCS đã có các nghiên cứu khảo sát cũng như đề xuất áp dụng mô hình chủ đề trọng việc phân tích sự phân bố của các chủ đề trong mạng thông tin, với các nghiên cứu & đề xuất cả tiến đã công bố tại các công trình **[CT5][CT6]**, dựa trên hướng tiếp cận kế thừa từ mô hình chủ đề LDA [13], (đề xuất bởi David M. Blie & cộng sự, 2003). Sự kết hợp của hai mô hình này nhằm mục đích giải quyết bài toán khám phá sự phân bố trong chủ đề giữa các thực thể giàu ngữ liệu văn bản trong mạng thông tin, điển hình như các bài báo khoa học trong mạng DBLP hay các bình luận, bài viết trong các trang mạng xã hội, v.v. Mô hình LDA sẽ giúp ước lượng sự phân bố của các chủ đề trên tập thực thể ở dạng văn bản, ký hiệu d, mỗi thực thể sẽ được đại diện bằng một vector số thực với số chiều bằng số lượng của chủ đề, ký hiệu: Z, ta có sự phân bố của các chủ đề trên mỗi thực thể, ký hiệu: $P(z_i|d_j) = \theta^{d_j}_{z_{(i,i\in|Z|)}}$, với: $(z_i : z_i \in Z)$ đại diện cho phân bố xác suất của chủ đề thứ (i) trên một thực thể ở dạng văn bản thứ (j). Vì

cứ mỗi thực thể sẽ được đại diện bởi một vector, ký hiệu: $\vec{d}$ với số chiều là $|Z|$, do đó ta có thể biểu diễn một thực thể ở dạng văn bản như sau (xem [công thức 3.1]):

$$\vec{d} = \begin{bmatrix} P(z_1|d_j) \\ \dots \\ P(z_{(i,i\in|Z|}|d) \end{bmatrix} = \begin{bmatrix} \theta^d_{z_1} \\ \dots \\ \theta^d_{z_{(i,i\in|Z|)}} \end{bmatrix} \quad (3.1)$$

Trong đó,

- $\vec{d}$, là một vector đại diện cho phân bố xác suất của tập các chủ đề (Z) tại một thực thể nhất định.
- $P(z_{(i,i\in|Z|}|d)$, là phân phối xác suất của chủ đề thứ (i) cho thực thể (d).

3.1.2. Tính toán tương đồng giữa các thực thể giàu nội dung

Để thực hiện tính toán tương đồng về mặt chủ đề giữa hai thực thể ở dạng văn bản, độ đo tương đồng cosine được sử dụng. Với hai thực thể ở dạng văn bản x và y, trong đó x và y là hai thực thể đối xứng với nhau trên một meta-path, ta có thể xác định độ tương đồng trong chủ đề giữa hai thực thể này, ký hiệu: $\text{top\_sim}(x, y)$, như sau (xem [công thức 3.2]):

$$\text{top\_sim}(x,y) = \frac{\vec{x}.\vec{y}}{\|\vec{x}\|.\|\vec{y}\|} = \frac{\sum_{i=1}^{|Z|}(\theta^x_{z_i} \times \theta^y_{z_i})}{\sqrt{\sum_{i=1}^{|Z|}(\theta^x_{z_i})^2} \times \sqrt{\sum_{i=1}^{|Z|}(\theta^y_{z_i})^2}} \quad (3.2)$$

Trong đó:

- Z, là tập các chủ đề phân bố trên các nút thực thể ở dạng văn bản trong mạng thông tin.
- $\theta^x_{z_i}$ và $\theta^y_{z_i}$, đại diện cho vector xác suất phân bố của các chủ đề trong hai thực thể ở dạng văn bản (x) và (y).

Xét một meta-path $(\mathcal{P})$ đối xứng, có cấu trúc: $A_s \to \cdots \to A_k \to \cdots \to A_e \leftarrow \cdots \leftarrow A_{k^-} \leftarrow \cdots A_{s^-}$, trong đó A_s và A_{s^-} là loại của hai thực thể đầu và cuối của meta-path cần tính toán tương đồng và A_e là loại của thực thể trung gian ở giữa chia meta-path làm hai phần bằng nhau. Do meta-path luôn có tính chất đối xứng, nên tại từng vị trí trên meta-path loại của các thực thể ở vế trái luôn luôn bằng với loại của thực thể ở vế phải, ta gọi tập các thực thể A_k và A_{k^-} là các thực thể ở dạng văn bản đối xứng với nhau qua meta-path $\mathcal{P}$.

3.2. Thuật toán W-PathSim: tương đồng theo meta-path có trọng số chủ đề

Từ đó các nền tảng đó, NCS đề xuất công thức xác định trọng số liên kết của meta-path mới dựa trên việc đánh giá sự tương quan trong chủ đề như sau: với meta-path $\mathcal{P}$, cho hai thực thể nguồn và đích của là (x) và (y), ta xác định trọng số liên kết, ký hiệu: $\text{W-PC}(x \rightsquigarrow y, P)$, như sau (xem [công thức 3.3]) (công bố tại [**CT9**][**CT10**]):

$$\text{W-PC}(x \rightsquigarrow y, \mathcal{P}) = \sum_{p,p\in|P|}^{|P|} W(p) \times \text{avg}[\text{top\_sim}(k, k^-)] \quad (3.3)$$

Trong đó:

- P, là tập các path instances được xác định giữa hai thực thể cùng loại (x) và (y), theo meta-path $\mathcal{P}$.
- w(p), là trọng số của một path instance cụ thể nào đó nối giữa (x) và (y), thông thường $W(p) = 1$ với các mạng thông tin không có trọng số.
- avg[top_sim(k, k⁻)], là trung bình trọng số tương đồng giữa các cặp thực thể ở dạng văn bản đối xứng nhau trong meta-path $\mathcal{P}$.

Từ công thức mới xác định trọng số liên kết của meta-path $\mathcal{P}$, giữa hai thực thể (x) và (y), NCS đề xuất công thức tính toán tương đồng giữa hai thực thể theo hướng chủ đề mới, ký hiệu: $\text{W-PathSim}(x \rightsquigarrow y, \mathcal{P})$, như sau (xem [công thức 3.4]) (công bố tại [**CT9**][**CT10**]):

$$\text{W-PathSim}(x \rightsquigarrow y, \mathcal{P}) = \frac{2 \times \text{W-PC}(x \rightsquigarrow y, \mathcal{P})}{\text{W-PC}(x \rightsquigarrow ., \mathcal{P}) + \text{W-PC}(y \rightsquigarrow ., \mathcal{P})} \quad (3.4)$$

Trong đó:

- $\text{W-PC}(x \rightsquigarrow y, \mathcal{P})$, là tổng trọng số liên kết của tất cả các path instances được xác định giữa hai thực thể (x) và (y), theo meta-path $\mathcal{P}$.
- $\text{W-PC}(x \rightsquigarrow ., \mathcal{P})$ và $\text{W-PC}(y \rightsquigarrow ., \mathcal{P})$, lần lượt là tổng trọng số liên kết của tất cả các path instances từ thực thể (x) và (y) đến các thực thể cùng loại khác trong mạng thông tin, theo meta-path $\mathcal{P}$.

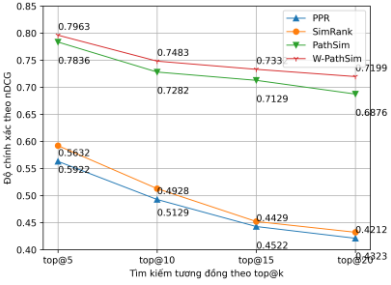
3.3. Thực nghiệm mô hình và đánh giá kết quả đạt được

Để chứng minh tính hiệu quả của mô hình W-PathSim đề xuất, các bước kiểm thử và thực nghiệm so sánh với các mô hình truyền thống, bao gồm các mô hình nguyên mẫu PathSim. Ngoài ra, W-PathSim cũng được cài đặt để thực nghiệm so sánh với các mô hình dành cho mạng thông tin đồng nhất bao gồm: Persionalized PageRank (PPR) và SimRank nhằm để có một đánh giá tổng quát hơn về hiệu xuất của các mô hình khai phá mạng thông tin trong các ngữ cảnh khai phá khác nhau. Để kiểm thử mô hình được đề xuất, trong thực nghiệm này mạng thông tin học thuật DBLP kết hợp với tập dữ liệu nội dung mở đầu (abstract) của các bài báo từ kho dữ liệu Aminer được sử dụng để làm dữ liệu kiểm thử. Hai tập dữ liệu thực nghiệm này bao gồm:

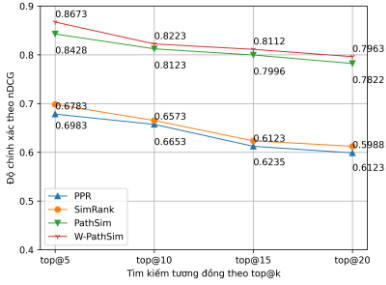
- **Mạng thông tin học thuật DBLP**[6]**:** với gần 2 triệu tác giả, 4.1 triệu bài báo và hơn 7K các hội nghị/tạp chí chuyên ngành. Tập dữ liệu DBLP là một mạng thông tin học thuật (bibliographical network) nguồn mở phổ biến nhất hiện nay.
- **Tập dữ liệu nội dung Aminer**[7]**:** với hơn 600K nội dung mở đầu (abstract) của các bài báo được chỉ mục trên mạng thông tin DBLP. Trong quá trình thực nghiệm, tập dữ liệu Aminer được sử dụng để rút trích sự phân bố của các chủ đề trong các bài báo khoa học trên mạng thông tin DBLP.

[6] Mạng thông tin DBLP: https://dblp.uni-trier.de/
[7] Aminer: https://aminer.org/

Việc đánh giá điểm tương đồng giữa các thực thể tác giả và hội nghị sẽ được dựa vào các chủ đề mà các thực thể này được chỉ mục dựa trên tập dữ liệu của ACM và Goolge Scholar Metric. Mức độ tương đồng trong tập các chủ đề chung mà các thực thể này được gán một tập nhãn chủ đề càng cao thì điểm xếp hạng tương quan giữa chúng sẽ sẽ càng cao. Việc đánh giá điểm cho các cặp thực thê. Các điểm đánh giá sau đó sẽ được dùng để tính toán kết quả cuối cùng cho độ chính xác của mỗi mô hình theo độ đo nDCG. Các mô hình được thực nghiệm với bài toán truy vấn tương đồng, bao gồm: xác định tập các tác giả tương đồng thông qua meta-path: A-P-V-P-A, và xác định tập các hội nghị/tạp chí tương đồng thông qua meta-path: V-P-A-P-V. Ở mỗi mô hình, quá trình thực nghiệm sẽ lựa chọn ngẫu nhiên 100 tác giả và 100 hội nghị/tạp chí để thực hiện tìm kiếm top-5, top-10, top-15 và top-20 tác giả và hội nghị/tạp chí tương đồng với truy vấn, sau đó lấy kết quả trung bình để làm kết quả đánh giá cuối cùng cho các mô hình. Dựa trên kết quả thực nghiệm (Hình 3-1 và Hình 3-2) có thể thấy mô hình W-PathSim đề xuất đạt độ chính xác cao hơn so với mô hình PathSim truyền thống, trong khoảng 2.39% cho cả hai bài toán tìm kiếm tác giả và hội nghị tương đồng. Đặc biệt so với hai mô hình dành cho mạng thông tin đồng nhất là PPR và SimRank, W-PathSim đạt độ chính xác vượt trội hơn tầm trung bình 42.78% so với PPR và 38.75% so với SimRank cho cả hai bài toán tìm kiếm tác giả và hội nghị tương đồng trên DBLP.



Hình 3-1. So sánh W-PathSim với các mô hình khác trong bài toán tìm kiếm tác giả tương đồng thông qua độ đo nDCG



Hình 3-2. So sánh W-PathSim với các mô hình khác trong bài toán tìm kiếm hội nghị/tạp chí tương đồng thông qua độ đo nDCG

CHƯƠNG 4: TIẾP CẬN NHÚNG MẠNG THÔNG TIN (INE) TRONG MẠNG C-HIN, MÔ HÌNH W-METAPATH2VEC

4.1. Tổng quan về ánh xạ/nhúng mạng thông tin (INE)

Về mặt Tổng quan, INE có thể được xem là một trong các hướng tiếp cận mới nhất hiện nay bao gồm trong lĩnh vực khai phá mạng thông tin (INAM) nói riêng và khoa học dữ liệu nói chung. INE là một ý tưởng được phát xuất từ một mô hình rất nổi tiếng trong lĩnh vực xử lý ngôn ngữ tự nhiên (natural language processing – NLP) được đề xuất bởi T. Mikolov & cộng sự (2013), mô hình Word2Vec. Kể

thừa từ các ý tưởng của mô hình Word2Vec, INE là sự kết hợp giữa phân tích, rút trích đặc trưng của các nút trong mạng thông tin, kết hợp với mô hình học và tối ưu thông qua một số kỹ thuật khác nhau, điển hình là huấn luyện mạng neuron (neural network) kết hợp với các kỹ thuật tối ưu mô hình như SGD, Adam, v.v. [3] [4] (ví dụ như: DeepWalk, LINE, Node2Vec, Metapath2Vec, v.v.) Một số mô hình INE còn áp dụng phương pháp phân tích ma trận thành nhân tử (matrix factorization) để huấn luyện mô hình biểu diễn mạng thông tin, điển hình như: M-NMF [14], GraRep [15], HOPE [16], v.v. Kết quả đầu ra của mô hình INE sẽ là một ma trận nhúng (embedding matrix) đại diện cho các nút trong mạng thông tin với kích thước: $|V| \times d$, trong đó $d \ll |V|$. Ở khía cạnh tổng quát, với một mạng thông tin, $G = (V, E)$, mục tiêu của một mô hình INE là tìm một hàm ánh xạ (f) để chuyển đổi tập các nút V thành các vector số thực với kích thước số chiều là (d) (xem [công thức 4.1]):

$$f: V \rightarrow \mathbb{R}^d \tag{4.1}$$

Trong đó:

- V là tập các nút/thực thể trong mạng thông tin được xét.
- $\mathbb{R}^d$, đại diện cho không gian vector số thực ở dạng ma trận với kích thước $|V| \times d$, với d là số chiều của vector được quy định trước.
- f là hàm ánh xạ các nút V về không gian $\mathbb{R}^d$.

4.2. Sơ nét về các mô hình INE/NRL phổ biến hiện nay & động lực

4.2.1. Tổng quan về cơ chế hoạt động của INE/NRL

	Phân biệt loại		**Mức độ bao quát cấu trúc mạng thông tin**		**Có cơ chế tối ưu quá trình huấn luyện**		**Mạng thông tin kích thước lớn**	**Quan tâm đến nội dung**	**Ứng dụng**				
	Quan hệ	Thực thể/nút	Cấu trúc nút/quan hệ	Cấu trúc mạng thông tin	Sinh tập huấn luyện	Huấn luyện mô hình học			Tương đồng	Gom cụm	Phân lớp	Nhận diện cộng đồng	Dự đoán liên kết
DeepWalk [28] (2014)			✓		✓				✓				
LINE [29] (2015)			✓				✓		✓				
Node2Vec [30] (2016)			✓	✓ (DFS/BFS)	✓				✓	✓			
HIN2Vec [31] (2017)		✓	✓		✓				✓	✓			
Metapath2Vec [32] (2017)	✓ (Metapath)	✓	✓	✓ (RW theo meta-path)	✓	✓ (Negative sampling cho HIN)			✓	✓	✓		

Hình 4-1. Phân tích các tính năng của các mô hình NRL/INE nổi bật

Tùy thuộc vào mục đích việc huấn luyện mô hình INE nhằm để khai thác các tri thức khác nhau từ mạng thông tin mà các hàm mục tiêu của mỗi mô hình sẽ được định nghĩa khác nhau kèm theo các cách thức tối ưu các tham số của mô hình tương ứng. Thông thường SGD và phương pháp huấn luyện mạng neuron sẽ được áp dụng để huấn luyện và tối ưu hóa các hàm mục tiêu của mô hình INE. Thông

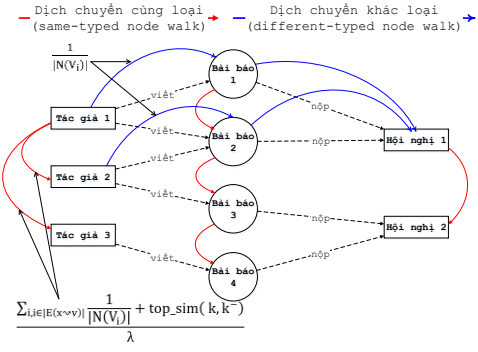
thường, đối với các mô hình INE dựa trên RW dành cho mạng thông tin không đồng nhất HIN sẽ có các cơ chế sinh tập các nút ngữ cảnh dựa thông qua cơ chế RW dựa trên meta-path, điển hình như Metapath2Vec. Ngoài ra, việc tối ưu hóa các hàm mục tiêu cũng sẽ phức tạp hơn các mô hình áp dụng cho HoIN do có sự tồn tại của nhiều loại nút cùng quan hệ trong mạng thông tin. So sánh các tính năng cũng như ưu/nhược điểm của các mô hình NRL/INE phổ biến hiện nay (xem Hình 4-1).

4.2.2. Các hạn chế của các mô hình INE hiện tại

Hầu hết các mô hình INE được đề xuất trong thời gian gần đây, điển hình như: DeepWalk, LINE, PTE, Node2Vec, v.v. đều chỉ có thể áp dụng cho các mạng thông tin đồng nhất (HoINs). Các mô hình này không thể phân tách được sự khác nhau trong loại giữa các thực thể cũng như quan hệ. Trong môi trường HIN, thì sự đa dạng trong loại của các thực thể sẽ ảnh hưởng khá nhiều đến việc xác định các nút ngữ cảnh hàng xóm (contextual neighborhood nodes). Để giải quyết cho sự đa dạng của các loại thực thể và quan hệ trong HIN trong quá trình ánh xạ các thực thể về miền không gian vector số thực liên tục, Dong & công sự (2017) đã đề xuất mô hình Metapath2Vec [17] để giải quyết bài toán node embedding trên HIN. Tuy nhiên, do phụ thuộc hầu hết vào các liên kết trong meta-path nên Metapath2Vec đã bỏ qua một số yếu tố quan trọng khác điển hình như sự tương quan trong chủ đề giữa các thực thể, đặc biệt là đối với các mạng thông tin giàu ngữ nghĩa như các mạng xã hội với hàng triệu bình luận, bài viết được đăng lên mỗi ngày hay các mạng thông tin học thuật như DBLP với hàng triệu các bài báo khoa học.

4.3. Mô hình W-Metapath2Vec: tiếp cận INE cho mạng thông tin không đồng nhất giàu nội dung (C-HIN)

4.3.1. Bước đi ngẫu nhiên dựa trên meta-path theo hướng chủ đề (topic-driven meta-path-based random walk)



Hình 4-2. Minh họa bước đi (node walk) giữa các thực thể cùng và khác loại dọc theo meta-path A-P-V-P-A

Mô hình W-Metapath2Vec đề xuất một cơ chế sinh tập các vectors đặc trưng cho mỗi thực thể trong HINs thông qua việc sửng dụng nguyên lý bước đi ngẫu nhiên dựa trên meta-path theo hướng tiếp cận của việc đánh giá sự tương quan trong chủ đề giữa các thực thể ở dạng văn bản dọc theo meta-path sử dụng. Để xác định được sự tương quan trong chủ đề giữa hai thực thể dọc theo meta-path, từng cặp các thực thể ở dạng văn bản $\langle k, k^{-} \rangle$ sẽ được tính toán mức độ tương đồng sau đó lấy giá trung bình để làm trọng số tương quan trong chủ đề. Hướng tiếp cận này được lấy ý tưởng từ mô hình W-PathSim trong việc tính trọng số tương đồng giữa hai thực thể dựa trên meta-path theo hướng tiếp cận chủ đề. Từ các ý tưởng trên, trong mô hình W-Metapath2Vec, NCS đề xuất công thức xác định trọng số dịch chuyển từ thực thể (x) đến một thực thể (v) bất kỳ, cùng loại $(\phi(x) = \phi(v))$, theo meta-path $\mathcal{P}$, ký hiệu $\pi_{x \rightsquigarrow v}, \mathcal{P}$, được tính theo công thức sau (xem [công thức 4.2]) (công bố tại **[CT1]**):

$$\pi_{x \rightsquigarrow v, \mathcal{P}} = \begin{cases} \text{với } e(x, v) \notin E \begin{cases} \dfrac{\sum_{\mathcal{P}_{x \rightsquigarrow v}} \sum_{i, i \in |E(x \rightsquigarrow v), \mathcal{P}|} \frac{1}{|N(V_i)|} + \text{top\_sim}(k, k^{-})}{\lambda}, \text{với } \phi(x) = \phi(v) (4.2a) \\ 0, \text{với } \phi(x) \neq \phi(v) \end{cases} \\ \text{với } e(x, v) \in E, \dfrac{1}{|N(x)|} (4.2b) \end{cases} \quad (4.2)$$

Trong đó:

- $N(x)$, là tập các nút hàng xóm lân cận của thực thể (x), hay nói cách khách là tập các thực thể liên kết trực tiếp với (x) và (x) là nút bắt đầu. Việc di chuyển từ thực thể (x) sang một thực thể (v) khác loại được gọi là dịch chuyển khác loại (different-typed node walk) Ví dụ: từ thực thể "tác giả 1" dịch chuyển qua các thực thể "bài báo 1" và "bài báo 2" (minh họa đường đi màu xanh dương Hình 4-2).
- $\frac{1}{|N(x)|}$, là xác suất dịch chuyển trực tiếp từ (x) sang (v) với loại của thực thể (x) có thể cùng hoặc khác với loại của thực thể (v) theo meta-path $\mathcal{P}$. Ví dụ "tác giả 1" sang "bài báo 1", "bài báo 1" sang "tạp chí 1", v.v. (minh họa Hình 4-2)
- $e(x, v) \in E$ và $e(x, v) \notin E$, lần lượt là có tồn tại một cung/cạnh nối giữa (x) và (v) và không tồn tại bất cứ cung/cạnh nào nối giữa (x) và (v) trong mạng thông tin được xem xét (G).
- $\sum_{i, i \in |E(x \rightsquigarrow v), \mathcal{P}|} \frac{1}{|N(V_i)|}$, là xác suất dịch chuyển từ thực thể (x) sang thực thể (v) được xác định bằng tổng trọng số đường đi giữa (x) và (v), theo meta-path: $\mathcal{P}$, với tập hợp các nút nằm trong đường đi giữa hai thực thể được xác định (x) và (v) là $|E(x \rightsquigarrow v), \mathcal{P}|$. Trường hợp này áp dụng cho khi (x) và (v) cùng loại với nhau $(\phi(x) = \phi(v))$. Đây được gọi là dịch chuyển cùng loại (same-typed node walk). Ví dụ ta có các dịch chuyển từ "tác giả 1" sang "tác giả 2", và từ "tác giả 2" sang "tác giả 3", v.v. (minh họa đường đi màu đỏ Hình 4-2).
- $\text{top\_sim}(k, k^{-})$, là trọng số tương đồng trong chủ đề giữa các thực thể ở dạng văn bản dọc theo meta-path $\mathcal{P}$.

Cơ chế bước đi ngẫu nhiên hướng chủ đề của mô hình W-Metapath2Vec dùng để sinh các vectors đặc trưng cho từng thực thể mục tiêu dưới dạng các thực thể hàng xóm dựa trên meta-path được định nghĩa trước. Cơ chế có hình thức dịch chuyển từ thực thể mục tiêu (x) đến các thực thể lân cận (v) bất kỳ là dịch chuyển cùng loại (same-typed node walk) và dịch chuyển khác loại (different-typed node walk) (minh họa Hình 4-2). Trong đó dịch chuyển cùng loại đóng vai trò quan trọng đối với mô hình W-Metapath2Vec vì nó giúp sinh ra các tập thực thể cùng loại với thực thể mục tiêu.

4.3.2. Áp dụng Skip-grams dành cho HIN trong mô hình W-Metapath2Vec

Sau khi hoàn tất quá trình sinh tập các vectors đặc trưng cho các thực thể trong HINs, bước tiếp theo sẽ là việc áp dụng mạng neuron để huấn luyện mô hình node embedding dựa trên tập các vectors đặc trưng đã được xây dựng. Tương tự như với mô hình Metapath2Vec của Dong & cộng sự (2017) đề xuất, mô hình W-Metapath2Vec cũng áp dụng phương pháp Skip-grams dành cho mạng thông tin không đồng nhất (HIN) để huấn luyện mô hình node embedding (xem [công thức 4.3] và [công thức 4.4]):

$$\underset{\theta}{argmax} \prod_{v \in V} \prod_{c \in N(v)} Prob(c|v;\theta) \tag{4.3}$$

$$\underset{\theta}{argmax} \sum_{v \in V} \sum_{t \in T_V} \sum_{c_t \in N_t(v)} Prob(c_t|v;\theta) \tag{4.4}$$

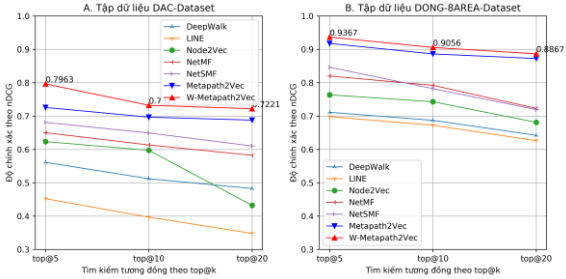
Trong đó:

- $N(v)$, là tập các thực thể lân cận của thực thể (v), không có sự phân biệt trong loại (có thể cùng hay khác loại).
- $N_t(v)$, là tập các thực thể lân cận của thực thể (v), và tập các thực thể này phải cùng một loại (t) với thực thể (v).
- $Prob(c|v;\theta)$, là xác suất xuất hiện của thực thể (v) trong tập các thực thể ngữ cảnh (c), không phụ thuộc vào loại của thực thể.
- $Prob(c_t|v;\theta)$, là xác suất xuất hiện của thực thể (v) trong tập các thực thể ngữ cảnh (c), phụ thuộc vào loại (t) của thực thể (v) và các thực thể của ngữ cảnh (c), với T_V là tập các loại thực thể của mạng thông tin.

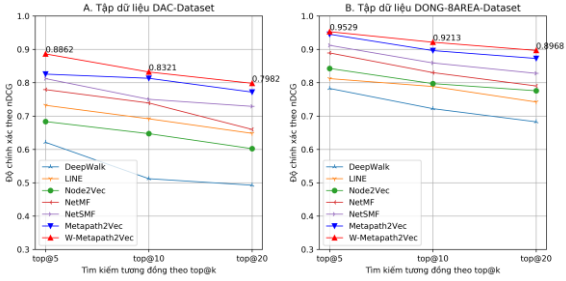
4.4. Thực nghiệm mô hình và đánh giá kết quả đạt được

Nhằm chứng minh tính hiệu quả của mô hình đề xuất W-Metapath2Vec, nội dung phần này sẽ tập trung mô tả các thực nghiệm so sánh giữa W-Metapath2Vec với các mô hình node embedding hiện tại bao gồm cho cả mạng thông tin không đồng nhất (Metapath2Vec) và mạng thông tin đồng nhất (Node2Vec, LINE, DeepWalk, NetMF và NetSMF). Các mô hình sẽ được thực nghiệm trên tập dữ liệu DBLP trong việc giải quyết ba bài toán chính của khai phá mạng thông tin, bao gồm: tìm kiếm tương đồng (similarity search), gom cụm (clustering) và phân đa lớp (classification). Trong phần thực nghiệm mô hình W-Metapath2Vec, tập dữ liệu DBLP và Aminer sẽ được sử dụng. Trong phần thực nghiệm này, kết quả đầu ra

dưới dạng tập các vectors đặc trưng của các thực thể được huấn luyện thông qua các mô hình node embedding sẽ được sử dụng để tính toán tương đồng giữa các thực thể, thông qua độ đo cosine (cosine similarity). Trong phần thực nghiệm này, các mô hình sẽ được áp dụng để giải quyết hai bài toán liên quan đến tìm tập top-5, top-10 và top-20 các tác giả và hội nghị/tạp chí tương đồng. Việc thực nghiệm cho mỗi bài toán tìm kiếm sẽ được thực hiện bằng cách lựa chọn ngẫu nhiên 100 tác giả và hội nghị/tạp chí sau đó thực hiện truy vấn tìm kiếm tương đồng. Các kết quả trả về của 100 trường hợp sẽ được đánh giá thông qua độ đo nDCG, sau đó lấy trung bình để làm kết quả đánh giá cuối cùng.



Hình 4-3. So sánh kết quả tìm kiếm tác giả tương đồng với hai datasets (DAC-Dataset) và (DONG-8AREA-Dataset)



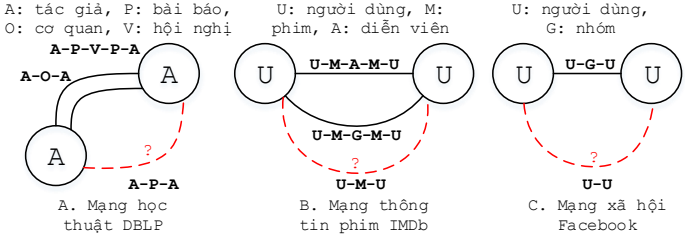
Hình 4-4. So sánh kết quả tìm kiếm hội nghị/tạp chí tương đồng với hai datasets (DAC-Dataset) và (DONG-8AREA-Dataset)

So sánh kết quả thực nghiệm giữa các mô hình (Hình 4-3 và Hình 4-4) cho thấy mô hình W-Metapath2Vec đạt độ chính xác cao hơn so với mô hình Metapath2Vec tầm 4.02%, và vượt trội hơn so với các mô hình truyền thống áp dụng cho HINs (NetSFM: 11.85%, NetFM: 14.54, Node2Vec: 29.65%, LINE: 55.9% và DeepWalk: 38.44%) cho cả hai bài toán tìm kiếm tác giả và hội nghị/tạp chí tương đồng.

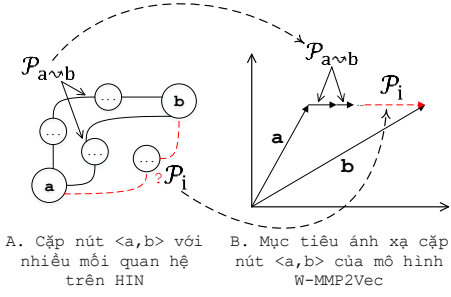
CHƯƠNG 5: DỰ ĐOÁN LIÊN KẾT TRÊN MẠNG C-HIN, MÔ HÌNH W-MMP2VEC

Trong nội dung của chương 5 này, NCS trình bày hướng tiếp cận xây dựng ứng dụng dự đoán liên kết (link prediction) trên mạng thông tin không đồng nhất giàu nội dung (C-HIN) theo hướng tiếp cận nhúng/ánh xạ mạng thông tin về môi trường vector (INE). Một phần nội dung của chương đã được công bố tại các công trình: **[CT3][CT4]**. Kế thừa từ các kết quả đạt được của mô hình INE, là W-Metapath2Vec áp dụng cho việc ánh xạ/nhúng các nút tương đồng nhau theo chủ đề trong C-HIN, NCS & GVHD đã đề xuất xây dựng mô hình W-MMP2Vec (công bố tại **[CT4]**) nhằm hỗ trợ cho việc huấn luyện mô hình biểu diễn các nút có khả năng cao xuất hiện các liên kết trong trong mạng C-HIN về các vector tương tự nhau.

5.1. Dự đoán sự tồn tại của liên kết mới dựa trên việc phân tích các liên kết khác loại hiện có theo meta-path



Hình 5-1. Minh họa về sự ảnh hưởng của các liên kết sẵn có trong việc hình thành các liên kết mới giữa các cặp nút trong các mạng thông tin khác nhau



Hình 5-2. Ý tưởng của mô hình W-MMP2Vec

Hầu hết các mô hình dự đoán liên kết theo hướng tiếp cận dựa trên meta-path truyền thống hay INE đều gặp các hạn chế là việc sử dụng dữ liệu và huấn luyện mô hình dự đoán đều chỉ dựa vào duy nhất một loại quan hệ được xét giữa hai nút trong mạng thông tin, dẫn đến kết quả dự đoán còn chưa đạt được độ chính xác cao nhất. Lấy lại ví dụ về bài toán dự đoán sự xuất hiện của quan hệ đồng tác giả (co-authorship) (A-P-A) giữa hai tác giả trong mạng thông tin DBLP, ta sẽ thấy

hầu như tất cả các tác giả có quan hệ đồng nghiệp (A-O-A) hay quan hệ cùng tham gia/nộp bài báo của họ cho một số hội nghị/tạp chí nhất định thường sẽ xuất hiện mối quan hệ đồng tác giả (A-P-A) (minh họa Hình 5-1-A). Điều này khá phù hợp với ý nghĩa trong thực tế, các tác giả sẽ có xu hướng và khả năng cao cùng cộng tác trong một công trình khoa học/bài báo nếu họ là đồng nghiệp hay thường gặp gỡ nhau tại một hội nghị khoa học nào đó. Lấy một ví dụ khác về việc hình thành mối quan hệ bạn bè (U-U) giữa hai người dùng trong mạng xã hội, ví dụ như Facebook (minh họa Hình 5-1-C). Ta sẽ thấy các người dùng cùng tham gia vào một hội nhóm, fanpage, v.v. (thể hiện qua meta-path U-G-U) hay đã từng bình luận về một bài viết nào đó (thể hiện qua meta-path U-C-P-C-U) sẽ có xu thế xuất hiện mối quan hệ bạn bè cao hơn so với các trường hợp chưa có bất cứ mối quan hệ nào.

5.2. W-MMP2Vec: mô hình dự đoán liên kết (link prediction) theo hướng tiếp cận hướng INE

5.2.1. Ý tưởng & các câu hỏi đặt ra trong quá trình nghiên cứu

Từ ý tưởng về sự phụ thuộc của các liên kết sẵn có trong việc dự đoán sự xuất hiện của các liên kết mới giữa các cặp nút trong mạng thông tin, NCS đã đề xuất một hướng tiếp cận mới thông qua việc kết hợp giữa hướng tiếp cận INE với việc xây dựng một mô hình huấn luyện đặc trưng trong đó mô hình sẽ dự đoán sự xuất hiện của liên kết mới (ở dạng meta-path) được xét giữa các cặp nút thông qua việc học các đặc trưng ở dạng các liên kết sẵn có (cũng ở dạng meta-path) và sự tương đồng trong chủ đề (ở dạng trọng số tương đồng trong chủ đề của các meta-path) giữa chúng. Ở mặt Tổng quan, mô hình dự đoán liên kết W-MMP2Vec được phát biểu một các có hệ thống, như sau:

- Cho một mạng thông tin không đồng nhất có cấu trúc ở dạng đồ thị $G = (V, E)$ với tập hàng loạt các quan hệ khác nhau ở dạng meta-path, $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2 \dots \mathcal{P}_n\}$.
- Cho một cặp nút cùng loại bất kỳ (a) và (b), ký hiệu $\langle a, b\rangle$, $\phi(a) = \phi(b)$.
- Giữa có sự xuất hiện của hàng loại các mối liên kết ở dạng meta-path, ký hiệu: $\mathcal{P}_{a\rightsquigarrow b}$, $\mathcal{P}_{a\rightsquigarrow b} \subseteq \mathcal{P}$, (minh họa Hình 5-2-A) ký hiệu: $\langle a, b, \mathcal{P}_{a\rightsquigarrow b}\rangle$.
- Mục tiêu của mô hình đặt ra là dự đoán sự xuất hiện của một liên kết cụ thể giữa (a) và (b) - ở dạng meta-path , ký hiệu: $\mathcal{P}_i$, $\mathcal{P}_i \subseteq \mathcal{P}$ và $\mathcal{P}_i \notin \mathcal{P}_{a\rightsquigarrow b}$.

Từ các yếu tố trên, mô hình học của W-MMP2Vec có nhiệm vụ phải cực đại hóa xác suất xuất hiện của $\mathcal{P}_i$ giữa cặp nút $\langle a, b\rangle$, với sự tồn tại của các mối quan hệ $\mathcal{P}_{a\rightsquigarrow b}$, ký hiệu $\langle a, b, \mathcal{P}_{a\rightsquigarrow b}\rangle$ như sau (xem [công thức 5.1]):

$$\text{Prob}(\mathcal{P}_i|\langle a, b, \mathcal{P}_{a\rightsquigarrow b}\rangle), \phi(a) = \phi(b), \mathcal{P}_i \in \mathcal{P}, \mathcal{P}_{a\rightsquigarrow b} \subseteq \mathcal{P}, \mathcal{P}_i \notin \mathcal{P}_{a\rightsquigarrow b} \quad (5.1)$$

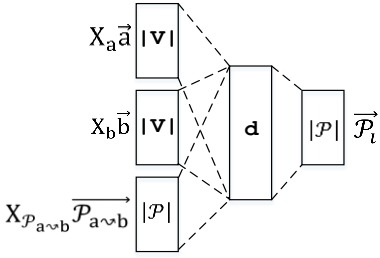
Trong đó,

- $\mathcal{P}_{a\rightsquigarrow b}$, là các mối quan hệ (ở dạng meta-path) đã xuất hiện giữa hai nút (a) và (b).

- $\mathcal{P}_i$, đại diện cho mối quan hệ chưa xuất hiện và cần dự đoán giữa hai nút (a) và (b) ở dạng meta-path và $\mathcal{P}_i \notin \mathcal{P}_{a\rightsquigarrow b}$.
- $\text{Prob}(\mathcal{P}_i|\langle a, b, \mathcal{P}_{a\rightsquigarrow b}\rangle)$ là xác suất xuất hiện của quan hệ $\mathcal{P}_i$ giữa hai nút (a) và (b) mà mô hình W-MMP2Vec cần phải cực đại hóa.

5.2.2. Hàm mục tiêu của mô hình W-MMP2Vec

Lấy ý tưởng từ các hướng tiếp cận của mô hình Trans-R, Trans-H và Trans-A trong việc chuyển đổi sự tương quan giữa cặp các nút và mối quan hệ giữa chúng trong mạng thông tin, Hình 5-2 minh họa Tổng quan về ý tưởng huấn luyện và mục tiêu tối ưu của mô hình W-MMP2Vec trong việc giải quyết bài toán dự đoán liên kết trên HIN theo hướng tiếp cận INE. Để biểu diễn cho các cặp nút: $\langle a, b\rangle$ với số chiều ánh xạ lên không gian vector là (d), ta sử dụng hai ma trận nhúng (embedding matrix), là X_a và X_b, với kích thước như nhau là: $|V| \times d$, trong đó mỗi hàng đại diện cho một nút của mạng thông tin.



Hình 5-3. Minh họa quá trình huấn luyện của mô hình W-MMP2Vec

Để biểu diễn cho các mối quan hệ giữa tập các cặp nút $\langle a, b\rangle$ ở dạng các meta-paths: $\mathcal{P}_{a\rightsquigarrow b}$, ta cũng sử dụng một ma trận nhúng $X_{\mathcal{P}_{a\rightsquigarrow b}}$ có kích thước: $|\mathcal{P}| \times d$. Để biểu diễn cho mối quan hệ cần dự đoán, ký hiệu: $\mathcal{P}_i$ giữa cặp nút $\langle a, b\rangle$, ta sử dụng một vector hàng có kích thước $1 \times |\mathcal{P}|$ ở dạng one-hot (gồm duy nhất một giá trị 1 tại vị trí của quan hệ $\mathcal{P}_i$ còn lại sẽ mang giá trị là 0). Hàm mục tiêu (objective function) của mô hình W-MMP2Vec được định nghĩa như sau (xem [công thức 5.2]):

$$X_a\vec{a} + X_{\mathcal{P}}\overrightarrow{\mathcal{P}_{a\rightsquigarrow b}} + \overrightarrow{\mathcal{P}_i} \approx X_b\vec{b} \tag{5.2}$$

Trong đó,

- $\vec{a}$ và $\vec{b}$ là vector hàng (row-vector) lần lượt đại diện cho nút hai nút (a) và (b), tương ứng trong ma trận nhúng X_a và X_b.
- $\overrightarrow{\mathcal{P}_{a\rightsquigarrow b}}$ là tập các vector hàng đại diện cho các quan hệ giữa hai nút (a) và (b), tương ứng trong ma trận nhúng $X_{\mathcal{P}}$.
- $\overrightarrow{\mathcal{P}_i}$, là vector hàng ở dạng one-hot vector biểu diễn cho quan hệ $\mathcal{P}_i$ giữa cặp nút được xét $\langle a, b\rangle$.

Về mặt tổng quan thì mô hình W-MMP2Vec sẽ được huấn luyện ở dạng bài toán phân lớp với tập các dữ liệu đầu vào là nút được xét (a) cùng với tập các mối quan hệ giữa chúng là $\mathcal{P}_{a\rightsquigarrow b}$ và mối quan hệ kỳ vọng sẽ xuất hiện là: $\mathcal{P}_i$ để kết quả đầu ra phải là nút (b) (minh họa Hình 5-2-B). Quá trình huấn luyện mô hình W-MMP2Vec sẽ được áp dụng kiến trúc mạng neuron để tối ưu các tham số của mô hình gồm: X_a, X_b và $X_{\mathcal{P}}$ thông qua áp dụng kỹ thuật tối ưu SGD. Mô hình W-MMP2Vec được đưa về dạng bài toán phân lớp với kết quả đầu ra của mạng neuron sẽ là xác suất xuất hiện của liên kết được xét ($\mathcal{P}_i$) giữa cặp nút (a) và (b), tại đầu ra của mạng neuron hàm softmax sẽ được áp dụng để bình thường hóa (normalized) và đưa tổng của các kết quả về giá trị 1. Ta có công thức suy diễn tiến (feed forward) của mô hình cho mỗi lần lặp như sau (xem [công thức 5.3]):

$$\text{Prob}(\mathcal{P}_i|\langle a, b, \mathcal{P}_{a\rightsquigarrow b}\rangle) = \text{softmax}\left(X_a\vec{a} \cdot X_b\vec{b} \cdot \sigma\left(X_{\mathcal{P}_{a\rightsquigarrow b}}\overrightarrow{\mathcal{P}_{a\rightsquigarrow b}}\right)\right) \quad (5.3)$$

Trong đó,

- $\sigma(.)$, là hàm sigmoid, với $\sigma\left(X_{\mathcal{P}_{a\rightsquigarrow b}}\overrightarrow{\mathcal{P}_{a\rightsquigarrow b}}\right) = \frac{1}{1+e^{-\left(X_{\mathcal{P}_{a\rightsquigarrow b}}\overrightarrow{\mathcal{P}_{a\rightsquigarrow b}} \cdot \overrightarrow{W_{\mathcal{P}_{a\rightsquigarrow b}}}\right)}}$.
- $\text{softmax}(.)$, là hàm softmax giúp đưa các giá trị đầu ra có tổng là 1.

Phương thức tối ưu các tham số mô hình của W-MMP2Vec thông qua việc cập nhất các tham số của mô hình ở mỗi lần lặp, gồm: suy diễn tiến (feedforward) và lan truyền ngược (back-propagation) sẽ được đề cập trong các phần tiếp theo.

5.2.3. Tương quan chủ đề trong bài toán dự đoán liên kết

Cuối cùng, để mô hình đạt độ tối ưu cao hơn cho các bài toán dự đoán liên kết trên các mạng thông tin giàu nội dung, NCS đưa thêm trọng số tương đồng trong chủ đề của các meta-path được xét, ký hiệu: $w\_topsim_{\mathcal{P}}$, được xác định bằng trọng số tương đồng W-PathSim (xem [công thức 3.4], mục 3.2) vào quá trình huấn luyện của mô hình W-MMP2Vec, nhằm làm tăng hiệu suất cũng như độ chính xác của mô hình dựa đoán. Với mỗi cặp nút $\langle a, b\rangle$ bất kỳ, ta sẽ có một hay nhiều các mối quan hệ, ở dạng meta-path, giữa chúng: $\mathcal{P}_{a\rightsquigarrow b}$. Các mối quan hệ này sẽ có hai loại trọng số được gán tùy thuộc vào cấu trúc của chúng, bao gồm:

- **Trong số dạng nhị phân (binary meta-path):** trọng số sẽ có hai giá trị là 0 hoặc 1, với trường hợp tồn tại liên kết giữa hai nút (a) và (b) sẽ có giá trị 1 và ngược lại sẽ là 0. Trường hợp trọng số dạng nhi phân sẽ được áp dụng khi quan hệ (meta-path) được xét không tồn tại bất cứ tập nút ở dạng văn (K) và (K^-) bản đối xứng nào.
- **Trong số tương đồng chủ đề (topic weighted meta-path):** nếu meta-path được xét có tồn tại các nút ở dạng văn bản đối xứng nhau (K) và (K^-) thì trọng số tương đồng trong chủ đề ($w\_topsim_{\mathcal{P}}$) (xem công thức 3.2) sẽ được dùng làm trọng số cho quan hệ (meta-path) được xét đó.

Để huấn luyện mô hình W-MMP2Vec ở dạng bài toán phân lớp thông qua kiến trúc mạng neuron với kết quả đầu ra là xác suất dự đoán sự xuất hiện của một liên

kết được kỳ vọng là ($\mathcal{P}_i$). Công thức suy diễn tiến của mô hình W-MMP2Vec sẽ được cải tiến lại thành như sau (xem [công thức 5.4]):

$$\text{Prob}(\mathcal{P}_i|\langle a, b, \mathcal{P}_{a\rightsquigarrow b}\rangle) = \text{softmax}\left(X_a\vec{a} \cdot X_b\vec{b} \cdot \sigma\left(X_{\mathcal{P}_{a\rightsquigarrow b}}\overrightarrow{\mathcal{P}_{a\rightsquigarrow b}} \cdot \overrightarrow{W_{\mathcal{P}_{a\rightsquigarrow b}}}\right)\right) \quad (5.4)$$

Trong đó,

- $\overrightarrow{W_{\mathcal{P}_{a\rightsquigarrow b}}}$, là vector hàng, có kích thước $1 \times |\mathcal{P}_{a\rightsquigarrow b}|$ đại diện cho trọng số của các mối quan hệ (ở dạng meta-path) $\mathcal{P}_{a\rightsquigarrow b}$, giữa hai nút (a) và (b).
- $\sigma(.)$ và $\text{softmax}(.)$, lần lượt là hàm sigmoid và softmax.

Việc đưa vào trọng số tương quan trong chủ đề của các mối quan hệ (ở dạng meta-path) $\mathcal{P}_{a\rightsquigarrow b}$ giữa các cặp nút $\langle a, b\rangle$, ký hiệu $W_{\mathcal{P}_{a\rightsquigarrow b}}$ là một hướng tiếp cận mới của mô hình W-MMP2Vec so với các mô hình INE khác như: HIN2Vec, Metapath2Vec hay PME vốn hầu như chỉ phụ thuộc vào sự xuất hiện của các liên kết ở dạng nhi phân mà không phân tích đến yếu tố trọng số của các mối quan hệ, hay cụ thể hơn là trọng số tương đồng trong nội dung/chủ đề giữa các thực thể. Trong phần tiếp theo của chương, NCS sẽ trình bày cac thức tối ưu các tham số của mô hình W-MMP2Vec dựa trên SGD thông qua việc áp dụng kiến trúc mạng neuron.

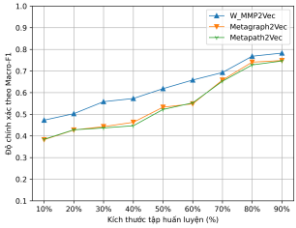
5.2.4. Thực nghiệm & đánh giá kết quả mô hình W-MMP2Vec

Trong nội dung của phần thực nghiệm tại chương 3 này, NCS sẽ trình bày về các phương pháp thực nghiệm, dữ liệu thu thập cũng như các nhận xét về các kết quả đạt được của mô hình W-MMP2Vec. Ngoài ra NCS cũng tiến hành so sánh hiệu suất của mô hình W-MMP2Vec với các mô hình INE phổ biến hiện này trong việc giải quyết bài toán dự đoán liên kết trong các loại mạng thông tin không đồng nhất (HIN) khác nhau cho cả mạng thông tin đồng nhất và không đồng nhất, bao gồm: DeepWalk, LINE, PTE, Node2Vec, Metapath2Vec, Metagraph2Vec và PME. Đối với mạng thông tin DBLP, tập dữ liệu huấn luyện (D_{train}) và kiểm thử (D_{test}) sẽ được chia dựa trên mốc thời gian, với:

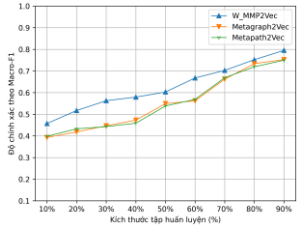
- Tập dữ liệu huấn luyện (D_{train}) sẽ bao gồm các nút và mối quan hệ giữa chúng trong khoản thời gian từ năm 1985 đến 2005 (căn cứ vào năm xuất bản của bài báo).
- Tập dữ liệu kiểm thử (D_{test}), sẽ bao gồm các nút và mối quan hệ giữa chúng trong khoản thời gian từ năm 2006 đến 2019 (hiện tại) (căn cứ vào năm xuất bản của bài báo).

Từ hai tập dữ liệu huấn luyện và kiểm thử được chia, các mô hình INE sau đó sẽ được áp dụng để hỗ trợ chuyển đổi các nút của từng tập dữ liệu sang dạng các vector với số chiều (d) quy định. Tập dữ liệu D_{train} sau đó sẽ được sử dụng để huấn luyện mô hình phân lớp LR và dự đoán sự xuất hiện của các liên kết kỳ vọng sẽ xuất hiện trong tập D_{test}. Kết quả trả về sau đó sẽ được đánh giá bằng hai độ đo là MAP và F-measure. Trong mang thông tin DBLP, NCS sẽ tiến hành thực nghiệm dự đoán sự xuất hiện của quan hệ đồng tác giả, với quan hệ kỳ vọng là $\mathcal{P}_i$: A-P-A thông qua việc xét đến các mối quan hệ có phụ thuộc khác, $P_{a\rightsquigarrow b}$: A-P-V-

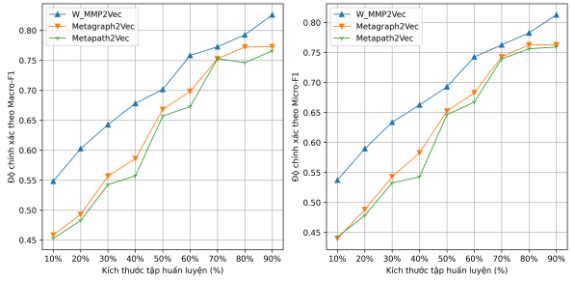
P-A (cùng xuất bản bài báo tại hội nghị/tạp chí) và A-O-A (quan hệ đồng nghiệp). Với các mô hình Metapath2Vec và PME sẽ được áp dụng quan hệ A-P-A.



Hình 5-4. So sánh W-MMP2Vec với các mô hình INE khác (Macro-F1)



Hình 5-5. So sánh W-MMP2Vec với các mô hình INE khác (Micro-F1)



Hình 5-6. Kết quả thực nghiệm cho bài toán dự đoán liên kết giữa các mô hình INE khác nhau trên mạng C-HIN - MovieLens100K

Kết quả thực nghiệm cho thấy mô hình W-MMP2Vec đạt hiệu suất cao hơn hẳn so với các mô hình INE dành cho HoIN (DeepWalk, LINE_1, LINE_2, PTE and Node2Vec), trung bình 12.03% và 23.27% tương ứng với độ đo MAP và F-1. So với các mô hình dành cho HIN, W-MMP2Vec cũng đạt độ chính xác nhỉnh hơn khoảng 15.5% và 3.37% (Metapath2Vec), 13.89% và 2.73% (Metagraph2Vec) với hai độ đo MAP và F-1. Qua kết quả thực nghiệm kiểm chứng mức độ ổn định của mô hình, cho thấy W-MMP2Vec đạt mức độ ổn định ở ngưỡng chấp nhận kích thước của tập dữ liệu huấn luyện khác nhau so với các mô hình INE dành cho HIN (Metapath2Vec, Metagraph2Vec và PME) (xem Hình 5-4 và Hình 5-5). Tương tự với kết quả thực nghiệm trên tập dữ liệu MovieLens100K (xem Hình 5-6), mô hình W-MMP2Vec cũng đạt độ chính xác cao hơn các mô hình INE/NRL khác trong bài toán dự đoán liên kết U-M-U.

CHƯƠNG 6: KẾT LUẬN & HƯỚNG PHÁT TRIỂN

6.1. Kết luận & các kết quả đạt được

Xuyên suốt nội dung 4 chương trong luận án, NCS đã trình bày tổng quan về các vấn đề nghiên cứu, phạm vị cũng như các bài toán chính đã đặt ra của luận án. Luận án được chia làm 3 nội dung chính với các kế hoạch cũng như hướng tiếp cận để giải quyết từng bài toán cụ thể. Các bài toán chính của luận án, bao gồm:

- Trong nội dung đầu của luận án, NCS tiến hành khảo sát tổng quan về các mô hình phân tích & khai phá mạng thông tin không đồng nhất. Phân tích các ưu/nhược điểm của các mô hình hiện tại để từ đó đưa ra các kết hoạch và phương hướng giải quyết cho từng bài toán cụ thể của luận án. Các công việc được tập trung được chú trọng giải quyết trong nội dung 1 của luận án bao gồm việc tiến hành cải tiến các mô hình hỗ trợ khám phá chủ đề trong mạng thông tin không đồng nhất giàu nội dung (C-HIN) theo hướng tiếp cận mô hình chủ đề LDA. Từ các kết quả của việc giải quyết bài toán khám phá sự phân bố của các chủ đề có trong C-HIN, NCS & GVHD đề xuất xây dựng mô hình W-PathSim (công bố tại [**CT9**][**CT10**]) nhằm hỗ trợ cho việc tìm kiếm tương đồng giữa các nút/thực thể cùng loại trong mạng C-HIN dựa trên meta-path theo hướng tiếp cận tương quan chủ đề.
- Trong nội dung 2, từ các kết quả đạt được trong nội dung 1 của luận án thông qua mô hình W-PathSim, NCS & GVHD tiếp tục đề xuất xây dựng mô hình cải tiến cho việc biểu diễn mạng thông tin giàu nội dung (C-HIN) theo hướng tiếp cận tương đồng trong chủ đề giữa các nút/thực thể. Mô hình W-Metapath2Vec (công bố tại **[CT1]**) là một mô hình ánh xạ/nhúng mạng thông tin (INE), hỗ trợ cho việc chuyển đổi các nút trong mạng C-HIN về dạng các vector với số chiều quy định, đảm bảo được việc bảo toàn cấu trúc cũng như sự tương đồng giữa các nút trong mạng thông tin. W-Metapath2Vec áp dụng cơ chế nguyên lý bước đi ngẫu nhiên dựa trên meta-path theo hướng tiếp cận tương đồng trong chủ đề giữa các cặp nút, hay còn gọi là topic-driven meta-path-based random walk mechanism, vốn được kế thừa từ mô hình W-PathSim đã đề xuất trước đó. Thực nghiệm trên các tập dữ liệu mạng thông tin thực tế như DBLP, MovieLens và BlogCatalog chứng minh tính hiệu quả của W-Metapath2Vec so với các mô hình INE phổ biến khác (DeepWalk, LINE, Node2Vec, Metapath2Vec, v.v.). Mô hình W-Metapath2Vec cũng là tiền đề để NCS & GVHD phát triển tiếp mô hình W-Metagraph2Vec (công bố tại **[CT2]**).
- Trong nội dung cuối của luận án, nội dung 3, NCS ứng dụng các thành quả đã đạt được trong các nghiên cứu của hai nội dung trước để giải quyết một bài toán ứng dụng quan trọng trong mạng thông tin là bài toán dự đoán liên kết (link prediction). Bài toán dự đoán liên kết giữa các cặp nút trong mạng thông tin không đồng nhất giàu nội dung (C-HIN) được xây dựng theo hướng tiếp cận INE, mô hình đề xuất được đặt tên là: W-MMP2Vec. Dựa trên hướng tiếp cận của INE và kết thừa từ W-Metapath2Vec, W-MM2PVec được phát triển dựa trên ý tưởng về sự tương đồng trong chủ đề và các mối quan hệ sẵn có giữa các thực thể sẽ ảnh hưởng lớn đến khả năng xuất hiện các liên kết mới giữa chúng. Mô hình W-MMP2Vec được chứng minh tính hiệu quả cũng như tính đúng đắn của các giải thuyết thông qua việc kiểm thử và thực nghiệm so sánh với hàng loạt các mô hình INE khác. W-MMP2Vec đã chứng minh được tính vượt trội cũng như khả năng hiệu quả trong việc giải quyết bài toán dự đoán liên kết trên HIN/C-HIN.

6.2. Các hạn chế còn tồn tại & hướng phát triển

NCS tập trung vào việc khắc phục các hạn chế của các mô hình khai phá liên quan đến việc phân tích sự tương đồng trong nội dung/chủ đề giữa các nút trong mạng thông tin không đồng nhất giàu nội dung, theo hướng tiếp cận nhúng mạng thông tin (NRL/INE) dựa trên meta-path. Tuy nhiên, luận án vẫn còn một số hạn chế còn tồn tại và đặt ra hướng nghiên cứu tiếp theo, như sau:

- Hướng tiếp cận của các mô hình NRL/INE đề xuất trong luận án hiện tại chỉ áp dụng nguyên lý huấn luyện mạng nơ-ron đơn giản với một tầng ẩn, do đó hiệu suất về độ chính xác của mô hình chưa thể đạt được hiệu quả cao nhất. Do đó, một trong các hướng cải tiến quan trọng trong tương lai của luận án là thay thế cơ chế huấn luyện bằng các kiến trúc mạng nơ-ron đa tầng của lĩnh vực học sâu (deep learning). Qua đó, có thể tăng cao hiệu suất về độ chính xác cho việc học mô hình biểu diễn mạng thông tin. Trong đó, kiến trúc mạng nơ-ron Graph Convolutional Network (GCN) [18] đa tầng là một trong các hướng cải tiến tiềm năng, được áp dụng trong nhiều lĩnh vực khác nhau [19] [20].
- Ngoài ra, các mô hình đề xuất trong luận án theo hướng tiếp cận NRL/INE (W-Metapath2Vec và W-MMP2Vec) chủ yếu dựa trên nguyên lý bước đi ngẫu nhiên để mô hình hóa cấu trúc mạng thông tin ở mức độ cục bộ và tương tự giữa các nút/thực thể trong mạng thông tin (local structure). Do đó, các mô hình đề xuất trong luận án hầu như chưa thể bảo toàn một cách hiệu quả cấu trúc toàn cục (global structure) của mạng thông tin. Các cải tiến trong tương lai sẽ tập trung vào việc kết hợp bảo toàn cấu trúc của mạng thông tin ở nhiều cấp độ khác nhau (cả local structure và global structure).
- Hướng tiếp cận cơ bản cho các bước xử lý các dữ liệu phi cấu trúc của mạng thông tin hiện tại được áp dụng trong luận án là mô hình chủ đề LDA. Tuy đạt được các hiệu quả nhất định trong việc giải quyết được bài toán xác định mức độ tương đồng giữa các nút/thực thể ở dạng nội dung trong mạng C-HIN, việc áp dụng mô hình LDA trong việc mô hình hóa các văn bản ở dạng các chủ đề ẩn vẫn còn gặp nhiều hạn chế. Điển hình là hướng tiếp cận bằng mô hình chủ đề sẽ gặp các hạn chế liên quan đến độ dài của văn bản và khả năng bảo toàn được cấu trúc ngữ nghĩa/thứ tự của các từ trong văn bản. Các hạn chế trên cũng phần nào làm giảm đi độ chính xác cho các mô hình đề xuất trong luận án. Đi cùng với sự phát triển của các kiến trúc học sâu mới thuộc lĩnh vực NLP như: auto-encoding/seq2seq [21] và attention [22], thì các thuật toán học mô hình biểu diễn của văn bản thuộc các trào lưu này, điển hình như: ELMo [23], BERT [24] là một trong các hướng cải tiến/thay thế cho mô hình LDA đầy tiềm năng. Hứa hẹn cho các cải thiện đáng kể về độ chính xác cho các mô hình đề xuất hiện tại của luận án.
- Xây dựng hệ khuyến nghị (recommendation) trên mạng thông tin không đồng nhất [25] [26] hiện là một trong những hướng tiếp cận phổ biến hiện nay với tính dụng cao cho nhiều bài toán ứng dụng đặc biệt trong lĩnh vực thương mại điện tử. Thông qua việc phân tích các tương tác (các bình luận, bài viết

hay like/share về các sản phẩm) và sự tương đồng giữa các nhím người dùng trên các trang mạng xã hội hay thương mại điện tử - hệ thống sẽ tìm kiếm và đưa ra các khuyến nghị sản phẩm phù hợp dựa trên sở thích của họ. Do đó, việc cải tiến các mô hình đề xuất trong luận án cho bài toán xây dựng các hệ khuyến nghị là một hướng cải tiến có tiềm năng và tính ứng dụng trong tương lai.

CÁC ĐỀ TÀI KHOA HỌC ĐÃ THAM GIA

Trong quá trình nghiên cứu và giải quyết các bài toán đặt ra của luận án, NCS Phạm Thế Anh Phú đã tham gia các đề tài khoa học, bao gồm:

- Đề tài NCKH "***Xây dựng và khai phá kho dữ liệu các bài báo trong lĩnh vực khoa học máy tính trên nền tính toán phân tán Hadoop hỗ trợ nghiên cứu khoa học***", có mã số: **B2017-26-02**, được tài trợ kinh phí bởi ĐHQG TP.HCM, do PGS.TS. Đỗ Phúc làm chủ nhiệm đã nghiệm thu đạt kết quả tốt (NCS là thành viên chính) (giai đoạn 2017-2019).
- Đề tài NCKH "***Phát triển hệ hỏi đáp ngôn ngữ tự nhiên các đồ thị tri thức lớn sử dụng nhúng đồ thị và học sâu***" có mã số: **DS2020-26-01**, được tài trợ kinh phí bởi ĐHQG TP.HCM, do PGS.TS. Đỗ Phúc làm chủ nhiệm (2020-2021).

DANH MỤC CÁC CÔNG TRÌNH ĐÃ CÔNG BỐ

Trong quá trình nghiên cứu và giải quyết các bài toán đặt ra của luận án, NCS Phạm Thế Anh Phú cùng GVHD là PGS.TS Đỗ Phúc đã đạt được một số kết quả, như sau:

Các bài báo tạp chí (chỉ mục SCIE) đã công bố (tổng: 4):

- **[CT1] PHAM, Phu**; DO, Phuc; "*W-MetaPath2Vec: the topic-driven meta-path-based model for large-scaled content-based heterogeneous information network representation learning*". In: *Expert Systems with Applications* (ISSN: 0957-4174) (SCIE indexed, IF: **5.452**), 2019, volume: 123, pp. 328-344.
(https://www.sciencedirect.com/science/article/pii/S0957417419300156)
- **[CT2] PHAM, Phu**; DO, Phuc; "*W-Metagraph2Vec: a novel approval of enriched schematic topic-driven heterogeneous information network embedding*". In: *International Journal of Machine Learning and Cybernetics* (ISSN: 1868-8071) (SCIE indexed, IF: **3.844**), 2020, volume: 11, issue: 8, pp. 1855-1874
(https://link.springer.com/article/10.1007%2Fs13042-020-01076-9)
- **[CT3] PHAM, Phu**; DO, Phuc; "*W-Com2Vec: a novel approach of topic-driven meta-path-based intra-community network embedding*". In: Intelligent Data Analysis (ISSN: 1571-4128) (SCIE indexed, IF: **0.860**), 2020, volume: 24 issue: 5, pp. 1207-1233
(https://content.iospress.com/articles/intelligent-data-analysis/ida194843)
- **[CT4] PHAM, Phu**; DO, Phuc; "*W-MMP2Vec: topic-driven network embedding model for link prediction in content-based heterogeneous information network*". In: Intelligent Data Analysis (ISSN: 1571-4128) (SCI indexed, IF: **0.860**), 2021, volume: 25, issue: 3.

Các bài báo tạp chí (chỉ mục Scopus) đã công bố (tổng: 4):

- **[CT5] PHAM, Phu**; DO, Phuc. "*Automatic topic labelling for text document using Ontology of graph-based concepts and dependency graph*". In: *International Journal of Business Information Systems* (ISSN: 1746-0972) (Scopus indexed), 2021, volume: 36, issue: 2, pp. 221-253.
(https://www.inderscienceonline.com/doi/abs/10.1504/IJBIS.2021.112826).
- **[CT6] PHAM, Phu**; DO, Phuc. "*The approach of using ontology as pre-knowledge source for semi-supervised labelled topic model by applying text dependency graph*". In: *International Journal of Business Intelligence and Data Mining* (ISSN: 1743-8187) (Scopus indexed)
(https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijbidm).

- [**CT7**] **PHAM, Phu**; DO, Phuc. "*Topic-driven top-k similarity search by applying constrained meta-path based in content-based schema-enriched heterogeneous information network*". In: *International Journal of Business Intelligence and Data Mining* (ISSN: 1743-8187) (Scopus indexed), 2020, volume: 17, issue: 3, pp. 349-376. (https://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijbidm).
- [**CT8**] **PHAM, Phu**; DO, Phuc. "*ComRank: community-based ranking approach for heterogeneous information network analysis and mining*". In: *International Journal of Business Intelligence and Data Mining* (ISSN: 1743-8187) (Scopus indexed), 2020, volume: 17, issue: 4, pp. 493-525. (https://www.inderscienceonline.com/doi/pdf/10.1504/IJBIDM.2020.110373).

Các bài báo tạp chí chuyên ngành đã công bố (tổng: 1):

- [**CT9**] DO, Phuc; **PHAM, Phu**. "*DW-PathSim: a distributed computing model for topic-driven weighted meta-path-based similarity measure in a large-scale content-based heterogeneous information network*". In: *Journal of Information and Telecommunication* (ISSN: 2475-1839), 2019, volume: 3, issue: 1, pp. 19-38.

  (https://www.tandfonline.com/doi/full/10.1080/24751839.2018.1516714).

Các bài báo hội nghị, đã công bố (tổng: 1):

- [**CT10**] **PHAM, Phu**; DO, Phuc; TA, Chien DC. "*W-PathSim: Novel Approach of Weighted Similarity Measure in Content-Based Heterogeneous Information Networks by Applying LDA Topic Modeling*". In: *Asian Conference on Intelligent Information and Database Systems. Springer, Cham*, 2018. p. 539-549. (https://link.springer.com/chapter/10.1007/978-3-319-75417-8_51)

TÀI LIỆU THAM KHẢO

[1] Shi, C., Li, Y., Zhang, J., Sun, Y., & Philip, S. Y., "A survey of heterogeneous information network analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 1, pp. 17-37, 2017.

[2] Sun, Y., & Han, J., "Mining heterogeneous information networks: principles and methodologies," *Synthesis Lectures on Data Mining and Knowledge Discovery*, vol. 3, no. 2, pp. 1-159, 2012.

[3] ZHANG, Daokun, et al., "Network representation learning: A survey," *IEEE transactions on Big Data*, 2018.

[4] CUI, Peng, et al., "A survey on network embedding," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 5, pp. 833-852, 2018.

[5] Page, L., Brin, S., Motwani, R., & Winograd, T., "The PageRank citation ranking: Bringing order to the web," *Stanford InfoLab*, 1999.

[6] KLEINBERG, Jon M., "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, pp. 604-632, 1999.

[7] Jeh, G., & Widom, J., "Scaling personalized web search," *Proceedings of the 12th international conference on World Wide Web*, pp. 271-279, 2003.

[8] Jeh, G., & Widom, J., "SimRank: a measure of structural-context similarity," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 538-543, 2002.

[9] Xu, X., Yuruk, N., Feng, Z., & Schweiger, T. A., "Scan: a structural clustering algorithm for networks," *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 824-833, 2007.

[10] Sun, Y., Han, J., Yan, X., Yu, P. S., & Wu, T., "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992-1003, 2011.

[11] Shi, C., Kong, X., Huang, Y., Philip, S. Y., & Wu, B., "HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2479-2492, 2014.

[12] Li, C., Sun, J., Xiong, Y., & Zheng, G., "An efficient drug-target interaction mining algorithm in heterogeneous biological networks," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 65-76, 2014.

[13] Blei, D. M., Ng, A. Y., & Jordan, M. I., "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993-1022, 2003.

[14] Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., & Yang, S., "Community preserving network embedding," *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[15] CAO, Shaosheng; LU, Wei; XU, Qiongkai, "Grarep: Learning graph representations with global structural information," *Proceedings of the 24th ACM international on conference on information and knowledge management. ACM*, pp. 891-900, 2015.

[16] OU, Mingdong, et al., "Asymmetric transitivity preserving graph embedding," *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM*, pp. 1105-1114, 2016.

[17] Dong, Y., Chawla, N. V., & Swami, A., "metapath2vec: Scalable representation learning for heterogeneous networks," *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM*, pp. 135-144, 2017.

[18] Kipf, T. N., & Welling, M., "Semi-supervised classification with graph convolutional networks," *5th International Conference on Learning Representations, ICLR*, 2017.

[19] Zitnik, M., Agrawal, M., & Leskovec, J., "Modeling polypharmacy side effects with graph convolutional networks.," *Bioinformatics*, vol. 34, no. 13, pp. i457-i466, 2018.

[20] He, X., Deng, K., Wang, X., Li, Y., Zhang, Y., & Wang, M., "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020.

[21] Bahdanau, D., Cho, K., & Bengio, Y., "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[22] Vaswani, Ashish, et al., "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[23] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L., "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2018.

[24] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K., "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

[25] Shi, C., Hu, B., Zhao, W. X., & Philip, S. Y., "Heterogeneous information network embedding for recommendation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 2, pp. 357-370, 2018.

[26] Zhao, Z., Zhang, X., Zhou, H., Li, C., Gong, M., & Wang, Y., "HetNERec: Heterogeneous network embedding based recommendation," *Knowledge-Based Systems*, vol. 204, p. 106218, 2020.