TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

PHÂN TÍCH MẠNG XÃ HỘI

ĐỒ ÁN CUỐI KÌ

PHÂN TÍCH SỰ PHÁT TRIỂN VÀ DỰ ĐOÁN LIÊN KẾT TRÊN MẠNG WIKIPEDIA

Giảng viên hướng dẫn:

Văn Chí Nam

Phan Thị Phương Uyên

Mã đề tài: SNA2006

Nhóm thực hiện: 387_399

1. Nguyễn Nhật Duy – 1712387

2. Nguyễn Quí Em – 1712399

TP. Hồ Chí Minh, tháng 01 năm 2021

Mục lục

1. Giới thiệu

Wikipedia được biết đến như là một cơ sở tri thức, một bách khoa toàn thư mã nguồn mở trực tuyến lớn nhất. Wikipedia chứa hàng triệu bài viết, các bài viết kết nối với nhau thông qua các siêu liên kết mà chúng tôi gọi là *wikilinks* tạo thành một mạng lưới chứa đựng tri thức khổng lồ của nhân loại. Hiểu được sự phát triển của mạng Wikipedia không chỉ giúp cho Wikipedia cải thiện nội dung và đề xuất, mà còn giúp chúng ta hiểu được sự phát triển mạnh mẽ của tri thức con người theo thời gian.

Trong báo cáo này, chúng tôi dựa theo sự hướng dẫn từ bài báo của Zecheng Zhang et al. [1], chúng tôi tập trung vào hai phần chính. Phần một, chúng tôi phân tích sự phát triển của mạng Wikipedia trên ngôn ngữ tiếng Anh dựa trên các bản chụp nhanh (snapshot) được ghi lại. Trong phần này chúng tôi sẽ phân tích chặt chẽ các thuộc tính và cách chúng thay đổi theo thời gian để khám phá các đặc tính thú vị của đồ thị cơ sở tri thức liệu có giống với mô hình đồ thị thế giới thật hay không. Phần hai, chúng tôi giải quyết bài toán dự đoán liên kết trên chính đồ thị này, chúng tôi chủ yếu tập trung vào các độ đo tương đồng giữa các cặp đỉnh, cũng như cấu trúc của đồ thị để giải quyết bài toán. Đồng thời chúng tôi cũng sử dụng một số kiến thức cơ bản của đồ thị để áp dụng vào mô hình của mình nhằm cải thiện hiệu suất đạt được so với chỉ dùng các phương pháp cơ bản.

2. Bài toán đặt ra và các định nghĩa

Trong đồ án này, chúng tôi của yếu tập trung vào các kỹ thuật để phân tích một đồ thị có cấu trúc và quy mô lớn như mạng Wikipedia thành các thông tin hữu ích có thể khai thác được để áp dụng vào bài toán của chúng tôi mà ở đây có 2 tác vụ chính đó là khám phá sự phát triển và dự đoán liên kết trong đồ thị.

Chúng tôi giải quyết các vấn đề ở trên dựa vào ba giải thuyết dưới đây:

- **Giả thuyết 1:** Mạng Wikipedia có các thuộc tính vĩ mô tương tự như các mạng thế giới thực khác: độ lệch phân phối bậc (degree distribution) cao, hệ số phân cụm (clustering coefficient) cao so với đồ thị ngẫu nhiên (random graph) và có một thành phần liên thông yếu rất lớn.
- **Giả thuyết 2:** Sự phát triển của mạng Wikipedia tuân theo luật mũ (power law) và có hiện tượng co đường kính (shrinking diameter): bậc trung bình của đồ thị tăng dần theo thời gian và tuân theo luật mũ, đồng thời đường kính hiệu dụng giảm dần theo thời gian.
- **Giả thuyết 3:** Các đặc trưng cấu trúc mạng cục bộ của Wikipedia có khả năng dự đoán về sự hình thành liên kết giữa các trang.

Chúng tôi sử dụng giả thuyết 1 và 2 cho tác vụ thứ nhất là khám phá sự phát triển của mạng Wikipedia để xem quá trình phát triển của mạng có tuân theo các quy luật trong giả thuyết hay không. Thay vì phân tích cho 18 năm từ 2001 đến 2018 với nhiều loại mạng Wikipedia ở các ngôn ngữ khác nhau như trong [1], thì chúng tôi chỉ thực hiện cho 6 năm từ 2001 đến 2006 với mạng Wikipedia tiếng Anh (gọi là *enwiki*) vì lí do bộ nhớ phần cứng có giới hạn. Giả thuyết thứ 3 chúng tôi sử dụng để giải quyết bài toán dự đoán liên kết như đã giới thiệu ở mục 1, với dữ liệu enwiki năm 2002 dùng để huấn luyện (training) và dữ liệu năm 2003 dùng để kiểm tra (testing).

Trong phần này chúng tôi cũng phát biểu một số định nghĩa cần thiết và liên quan đến bài toán để tránh sự nhập nhằng khó hiểu về khái niệm hay công thức của các độ do.

Coi Wikipedia là một mạng động (mạng mà có sự phát triển về đỉnh và cạnh theo giời gian) biểu diễn bởi một đồ thị có hướng $G = (N_t, E_t)$ với N_t, E_t lần lượt là số đỉnh và số cạnh của đồ thị tại thời điểm $t,\ t = 1, 2, \dots, 6$ ứng với các mốc thời gian 2001, 2002, …2006. Mỗi đỉnh của đồ thị là một trang trên Wikipedia đang hoạt động tại thời điểm đó, được biểu diễn bằng một định danh

(ID) và tiêu đề tương ứng. Mỗi cạnh trong đồ thị là cạnh có hướng $(i, j)_t$ thể hiện cho việc tồn tại một siêu liên kết từ trang i đến trang j tại thời điểm t.

Chúng tôi nhắc lại một số khát niệm cũng như công thức mà mà chúng tôi sử dụng trong bài này:

- **Định nghĩa 1: Bậc trong (in-degree)** của một đỉnh trong đồ thị là số lượng kết nối trỏ vào đỉnh đó từ các đỉnh láng giềng của nó, **bậc ngoài (out-degree)** là số lượng kết nối từ nó đi ra các đỉnh láng giềng của nó.
- **Định nghĩa 2: Hệ số phân cụm (clustering coefficient)** được đỉnh nghĩa bởi công thức:

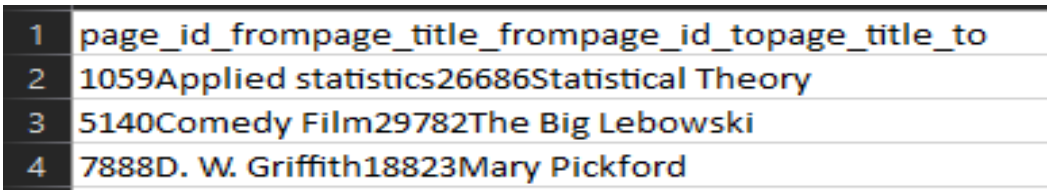
$$C_i = \frac{2e_i}{k_i(k_i - 1)},$$

  với e_i là số cạnh giữa các đỉnh láng giềng của đỉnh i và k_i là bậc của đỉnh i (bao gồm cả bậc trong và bậc ngoài).
- **Định nghĩa 3: Thành phần liên thông yếu lớn nhất (largest weakly connected component – WCC)** là thành tập hợp lớn nhất chứa các đỉnh mà ở đó giữa hai đỉnh bất kì luôn có một đường đi (không xét yếu tố về hướng).
- **Định nghĩa 4: Đường kính hiệu dụng (effective diameter)** của đồ thị là một số nguyên d sao cho 90% các cặp đỉnh có đường đi với nhau với độ dài đường đi ngắn nhất nhỏ hơn hoặc bằng d.

3. Bộ dữ liệu về mạng Wikipedia

Bộ dữ liệu WikiLinkGraphs [2] chúng tôi sử dụng trong đồ án này có nguồn gốc từ bài báo của Cristian Consonni et al. [3]. Đây là bộ dữ liệu chứa thông tin các bài viết trên Wikipedia với nhiều ngôn ngữ khác nhau từ năm 2001 đến năm 2018. Bộ dữ liệu của chúng tôi được rút gọn lại từ bộ dữ liệu này từ năm 2001 đến năm 2006 trên ngôn ngữ tiếng Anh (en-wiki).

Thông tin trong dữ liệu được tổ chức dưới dạng CSV, gồm các trường "page_id_from", "page_title_from", "page_id_to", "page_title_to" tương ứng với định danh và tên tiêu đề từ trang nguồn đến trang đích.

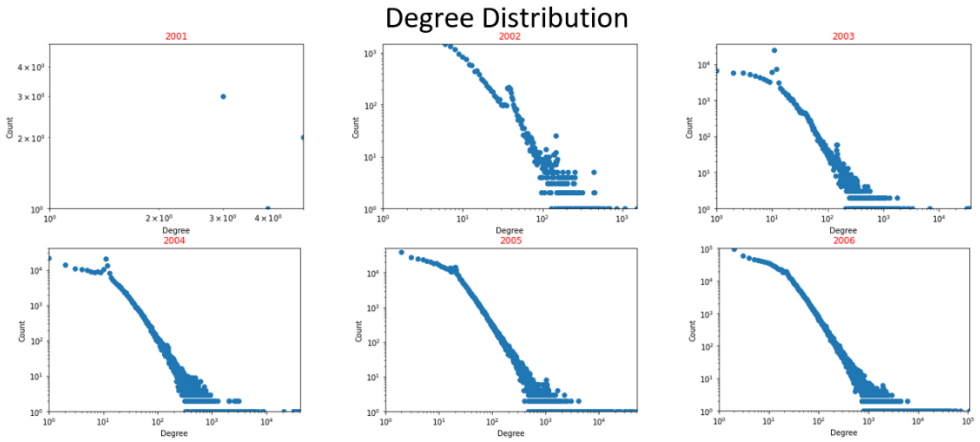


*Hình 1. Một phần dữ liệu lưu trong tệp *.csv*

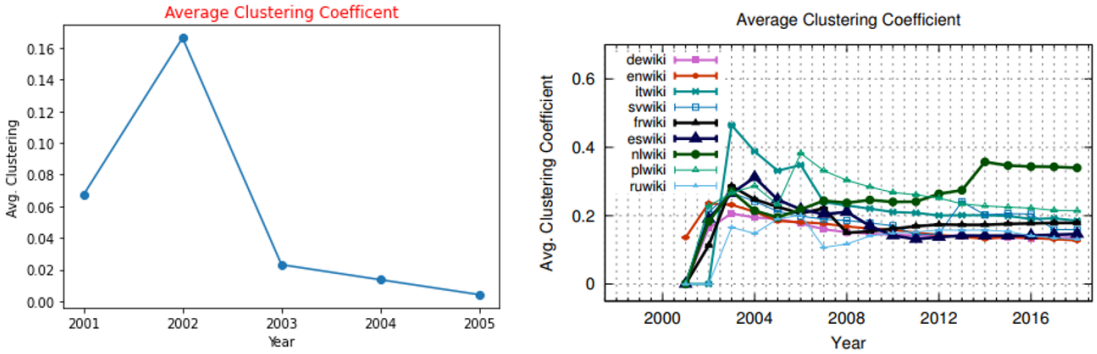
4. Sự phát triển của Wikipedia

Mục đích của việc phân tích sự phát triển mạng Wikipedia là để hiểu rõ các quy luật mà mạng tuân theo để khám phá cách tổ chức, lưu trữ, kết nối,… thông tin tri thức của nhân loại.

Như đã đề cập ở mục 2, chúng tôi thực hiện phân tích sự phát triển thông qua *giả thuyết 1* và *giả thuyết 2*. Ở giả thuyết 1, chúng tôi thực hiện tính phân phối bậc của các đồ thị dựa theo dữ liệu các năm từ 2001 đến 2006. Tiếp đến, chúng tôi thực hiện tính hệ số phân cụm trung bình (average clustering coefficient) từ 10,000 đỉnh ngẫu nhiên của mỗi đồ thị từ năm 2001 đến năm 2005. Đối với đồ thị có số đỉnh ít hơn 10,000, chúng tôi sẽ sử dụng toàn bộ số đỉnh của đồ thị đó để tính toán. Lý do chỉ chọn ra 10,000 đỉnh và bỏ năm 2006 là vì việc tính hệ số phân cụm rất mất thời gian nếu số đỉnh quá lớn, đồng thời vì thiết bị phần cứng của chúng tôi có giới hạn về mặt bộ nhớ. Để kiểm tra xem mạng Wikipedia có tồn tại một thành phần liên thông yếu rất lớn hay không, chúng tôi tính tỉ lệ số đỉnh thuộc thành phần liên thông yếu lớn nhất trên tổng số đỉnh của đồ thị. Một số kết quả phân tích do thực hiện trên số năm quá ít nên không thể hiện được rõ rệt ý nghĩa của nó. Do đó, chúng tôi sẽ dùng thêm kết quả tham khảo trong [1] ở một số tác vụ. Bên dưới là các kết quả phân tích mà chúng tôi thu được cho giả thuyết 1.

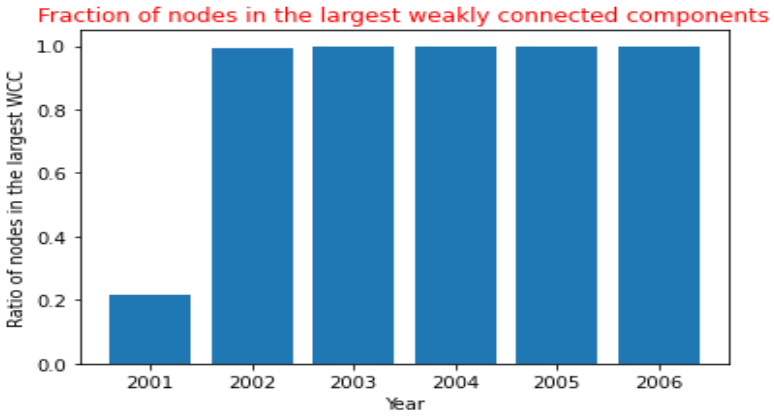


Hình 2. Phân phối bậc của các đồ thị Wikipedia từ năm 2001 đến năm 2006



Hình 3. Hệ số phân cụm của các đồ thị Wikipedia.

Bên trái: từ năm 2001 đến 2005; Bên phải: từ năm 2001 đến 2018 [1]



Hình 4. Tỉ lệ đỉnh trong thành phần liên thông yếu lớn nhất của các đồ thị từ năm 2001 đến 2006.

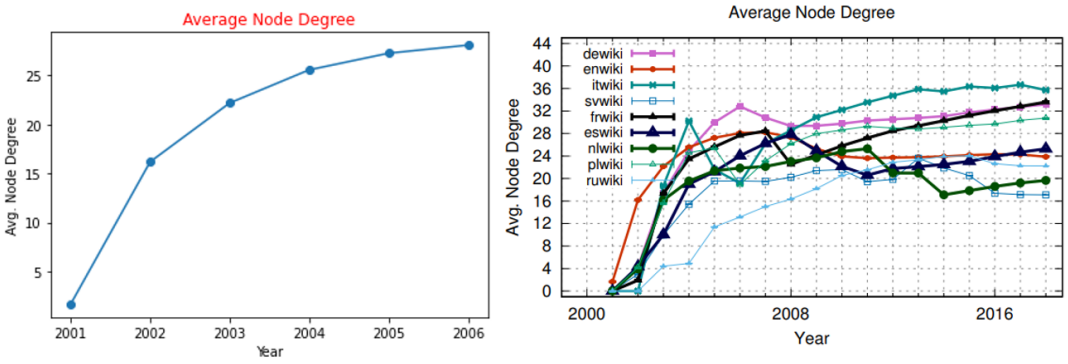
Nhận xét 1:

- Hình 2 cho thấy phân phối bậc của các đồ thị có lệch cao, tức là số đỉnh có bậc cao thì rất ít, còn số đỉnh có bậc thấp thì rất nhiều.
- Hình 3 cho thấy hệ số phân cụm trung bình có sự dao động ở những năm đầu tiên và bắt đầu ổn định ở các năm sau đó. Hệ số phân cụm này lớn hơn so với đồ thị ngẫu nhiên.
- Hình 4 cho thấy tỉ lệ các đỉnh thuộc thành phần liên thông yếu lớn nhất trên tổng số đỉnh của đồ thị là 1 bắt đầu từ năm 2002 trở đi. Điều này nói lên rằng mạng Wikipedia có một thành phần liên thông yếu rất lớn.

Kết luận 1:

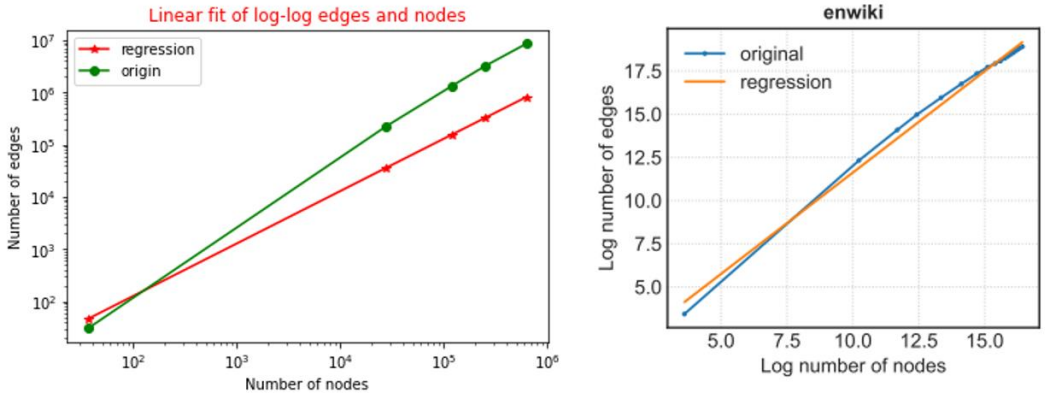
- Các kết quả thể hiện rằng giả thuyết 1 đưa ra là đúng: Phân phối bậc của các đồ thị có độ lệch cao, hệ số phân cụm cao so với đồ thị ngẫu nhiên, mạng có một thành phần liên thông yếu rất lớn.
- Cho thấy mạng Wikipedia – hay mạng liên kết tri thức của con người – cũng có các thuộc tính giống như các mạng thế giới thực.

Tiếp theo, chúng tôi sẽ trình bày các kết quả phân tích cho giả thuyết 2.

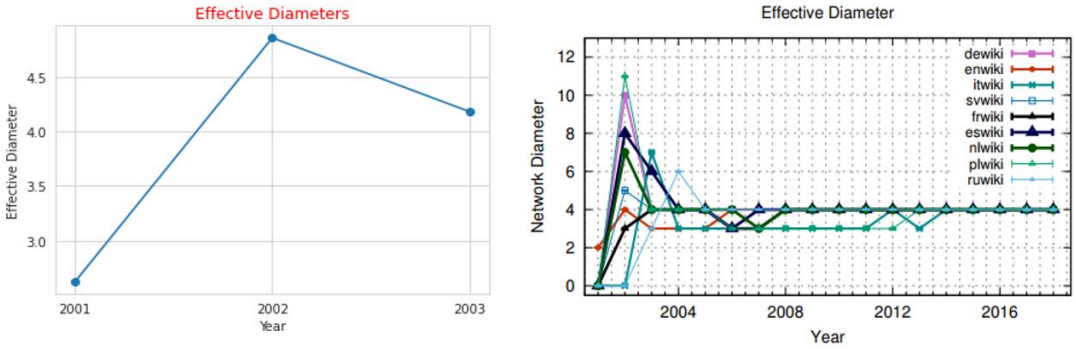


Hình 5. Bậc trung bình đỉnh của các đồ thị.

Bên trái: Từ năm 2001 đến 2006. Bên phải: Từ năm 2001 đến 2018 [1].



Hình 6. Đường hồi quy log-log số cạnh theo số đỉnh của các đồ thị enwiki. Bên trái: Từ năm 2001 đến 2005. Bên phải: Từ năm 2001 đến 2018 [1]



Hình 7. Đường kính hiệu dụng của các đồ thị.

Bên trái: Từ năm 2001 đến 2003. Bên phải: Từ năm 2001 đến 2018 [1].

Nhận xét 2:

- Hình 5 cho thấy bậc trung bình của đỉnh tăng dần theo thời gian, chứng tỏ số cạnh tăng nhiều hơn số đỉnh.
- Đường hồi quy ở Hình 6 cho thấy số cạnh và số đỉnh ngày càng tăng, số cạnh tăng nhiều hơn số đỉnh.
- Hình 7 cho thấy có hiện tượng co đường kính, xảy ra rất sớm từ năm 2002. Tuy nhiên, sau năm 2007 (đồ thị bên phải), độ dài đường kính hiệu dụng gần như ít thay đổi. Hay nói cách khác là hiện tượng co đường kính bắt đầu diễn ra chậm hơn.

Kết luận 2:

- Sự phát triển của mạng Wikipedia tuân theo luật mũ với luỹ thừa dương, trong đó nếu số đỉnh tăng gấp đôi thì số cạnh tăng **nhiều hơn** gấp đôi (được thể hiện ở hai nhận xét đầu trong ***Nhận xét 2***).
- Mạng Wikipedia có hiện tượng co ngắn đường kính. Điều này xảy ra là do cơ chế **tam giác đóng** (triadic closure) thường thấy trong các mạng xã hội khác (*Tam giác đóng: "Bạn của bạn trở thành bạn của mình"*). Tuy nhiên, từ năm 2007 hiện tượng co đường kính xảy ra chậm hơn. Nguyên nhân xuất phát từ việc mạng Wikipedia là mạng tri thức nên những thay đổi của nó biểu thị sự thay đổi về nhận thức của nhân loại, điều này vốn diễn ra chậm hơn các kết nối xã hội.

Tác vụ phân tích sự phát triển của mạng Wikipedia ở phần này cho thấy các thuộc tính vĩ mô của nó cũng giống như các mạng xã hội thế giới thực, như: phân phối bậc có độ lệch cao, hệ số gom cụm cao so với đồ thị ngẫu nhiên, có một thành phần liên thông yếu rất lớn, sự phát triển đỉnh và cạnh tuân theo luật mũ,…Bên cạnh đó, Wikipedia cũng có những đặc trưng rất riêng của một mạng tri thức, biểu hiện rõ nhất là hiện tượng co đường kính diễn ra chậm.

5. Dự đoán liên kết trên mạng Wikipedia

Trong phần này chúng tôi sẽ tập trung giải quyết bài toán dự đoán liên kết trong đồ thị WikiLinkGraphs. Mục tiêu của chúng tôi là cố gắng biểu diễn một cặp đỉnh có liên kết với nhau thành một vector đặc trưng bằng cách sử dụng các độ đo tương đồng rồi dùng các phương pháp học máy như Hồi quy Logistic (Logistic Regression) để dự đoán xác suất có thể xảy ra liên kết. Tiếp theo đó chúng tôi mở rộng vấn đề hơn bằng cách áp dụng một số phương pháp cơ bản trên đồ thị như nhúng đỉnh để biểu diễn thông tin đồ thị rồi sử dụng các

mô hình học máy để dự đoán với hi vọng cải thiện được hiệu suất dự đoán liên kết.

Vấn đề dự đoán liên kết trong đồ thị là một vấn đề khá lớn, đem lại nhiều ứng dụng trong thực tế như hệ thống gợi ý trong các công ty, doanh nghiệp. Để tiến hành dự đoán liên kết trong mạng Wikipedia, chúng tôi bắt đầu từ những điều cơ bản nhất:

Định nghĩa bài toán: Cho một đồ thị tại thời điểm t với các đỉnh và các cạnh. Gọi (i, j) là một cạnh của đồ thị tại thời điểm đó, bài toán đặt ra là liệu (i, j) có phải là một cạnh của đồ thị tại thời điểm $t + 1$ hay không, tức là:

$$f(v(i,j)_t) \rightarrow IsEdge(i,j)_{t+1}$$

với $v(i, j)$ là vector đặc trưng của cạnh (i, j), f là một hàm biến đổi vector đặc trưng v sao cho trả lời được câu hỏi nếu tại thời điểm $t + 1$ giữa i và j có một cạnh nối thì $f(v) = 1$ và ngược lại.

Các độ đo tương đồng trong dự đoán liên kết: Độ đo tương đồng ở đây là sự giống nhau giữa hai đỉnh trong đồ thị, chúng tôi cho rằng hai đỉnh càng có nhiều điểm chung thì càng có khả năng liên kết với nhau. Chúng tôi bắt đầu từ những độ đo đơn giản sau:

(1) **Jaccard Coefficient** (JC): Số láng giềng chung chuẩn hóa cho tổng số láng giềng.

$$JC(x,y) = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|}$$

(2) **Common Neighbors** (CN): Đếm số láng giềng chung.

$$CN(x,y) = |\Gamma(x) \cap \Gamma(y)|$$

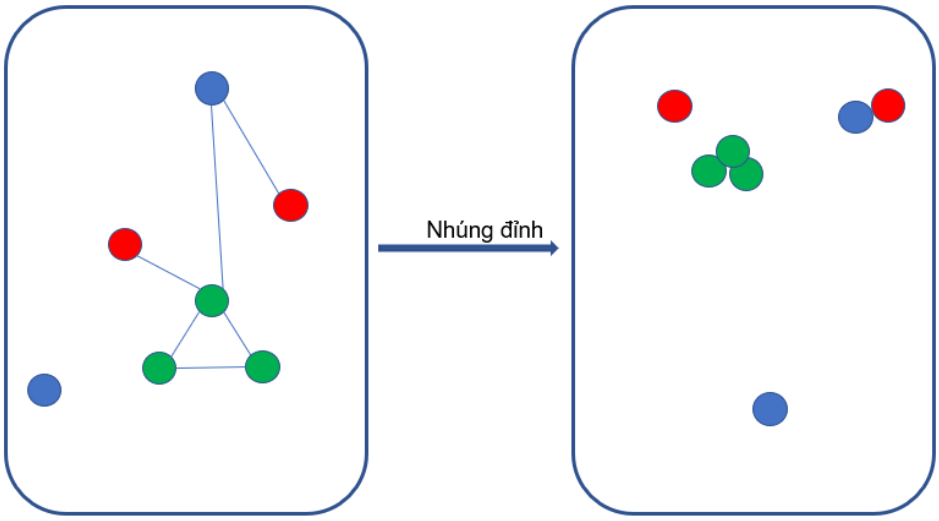
(3) **Preferential Attachment** (PA): Dựa trên nguyên lí gắn kết ưu tiên, nếu một đỉnh có bậc cao thì khi thêm cạnh mới hay thêm đỉnh vào đồ thị sẽ ưu tiên kết nối đến đỉnh đó

$$PA(x,y) = |\Gamma(x)|.|\Gamma(y)|$$

Chúng tôi sử dụng Hồi quy Logistic để làm mô hình học cho dự đoán liên kết với đầu vào là các độ đo tương đồng giữa cặp đỉnh, ngoài ra chúng tôi

cũng dựa vào ý tưởng của tác giả [1] bằng cách ghép các độ đo lại thành một vector duy nhất $v(x,y) = concat[JC(x,y), CN(x,y), PA(x,y)]$ rồi đưa vào mô hình học để so sánh kết quả.

Phần tiếp theo trong đồ án chúng tôi thử áp dụng một số lý thuyết đơn giản để tăng hiệu suất cho mô hình. Chúng tôi nhận thấy yếu tố **cộng đồng** (community) có thể ảnh hưởng đến hiệu suất dự đoán. Cộng đồng trong đồ thị là một tập hợp các đỉnh có kết nối dày đặc với nhau. Vì thế chúng tôi cho rằng xác suất mà hai đỉnh trong cùng một cộng đồng có khả năng liên kết với nhau sẽ cao hơn xác suất khi mà hai đỉnh đó nằm trên hai cộng đồng khác nhau. Chúng tôi mở rộng ý tưởng này bằng cách nhúng đỉnh của đồ thị Wiki về không gian Euclide sao cho các đỉnh có kết nối với nhau sẽ nằm gần nhau, hình bên dưới minh họa cho kiến trúc nhúng đỉnh đồ thị của chúng tôi



Hình 8. Minh hoạ tác vụ nhúng đỉnh đồ thị

Với kết quả thu được sau phép nhúng, chúng tôi dùng một mô hình học máy để thực hiện tiếp tác vụ dự đoán. Ở đây chúng tôi chọn mô hình GraphSAGE [4] vì nó là mô hình được thiết kế để có thể học tốt các tác vụ dự đoán liên kết.

6. Thực nghiệm và đánh giá dự đoán liên kết

Trong phần này chúng tôi cài đặt các thuật toán cũng như chạy thử nghiệm các mô hình đã nêu ra ở phần trước. Do giới hạn về mặt phần cứng, chúng tôi chỉ sử dụng dữ liệu của WikiLinkGraph [2] trong 2 năm (2002 và 2003). Chúng tôi cũng giới hạn lại mô hình của mình bằng cách chỉ xét đến trường hợp đồ thị có sự tăng trưởng về cạnh mà không tăng thêm về đỉnh.

Phân chia tập huấn luyện và tập kiểm tra

Chúng tôi sử dụng dữ liệu đồ thị tại thời điểm năm 2002 làm tập huấn luyện, dữ liệu đồ thị năm 2003 làm tập kiểm tra. Trong đồ thị năm 2003 chúng tôi tiến hành xử lí xóa tất cả các đỉnh mới xuất hiện (đồ thị năm 2002 không có các đỉnh này), các cạnh mới phát sinh trên đồ thị đã được xử lí này sẽ là cái mà mô hình chúng tôi cần phải dự đoán.

Tạo mẫu sai (Negative Sampling)

Kĩ thuật tạo mẫu sai cũng là một phần quan trọng trong bài toán của chúng tôi. Vì dữ liệu các cạnh của đồ thị được đưa vào trong quá trình học luôn là mẫu đúng dẫn đến việc mô hình học chỉ nhìn thấy được sự thật rằng "tất cả các đỉnh đều có khả năng kết nối với nhau" nên mô hình sẽ học sao cho kết quả đưa ra luôn trả lời có. Vì thế để cho việc mô hình có thể học và phân biệt được đâu là cặp đỉnh thật sự có kết nối và đâu là cặp đỉnh không có kết nối, chúng tôi tiến hành tạo các mẫu sai và đưa vào cùng với các mẫu đúng để mô hình học cùng một lúc.

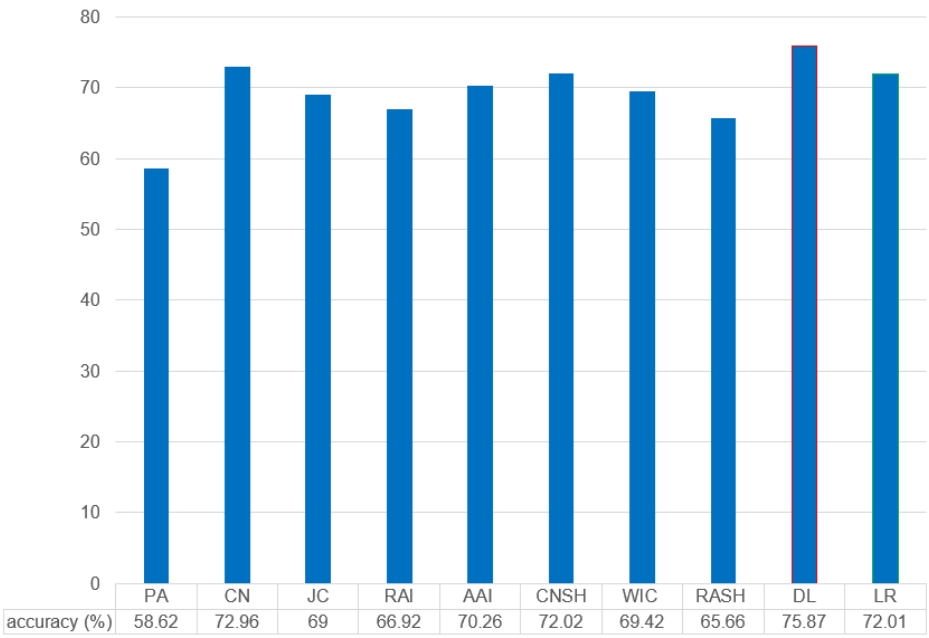
Có nhiều cách để tạo mẫu sai, tuy nhiên ở đây chúng tôi làm theo cách cơ bản nhất là cho một cặp đỉnh có kết nối, ta chọn mẫu sai cho cặp đỉnh này bằng cách giữ lại một đỉnh và thay thế bằng một đỉnh khác sao cho cặp đỉnh đó hiện tại chưa có kết nối. Việc làm này có thể vô tình dẫn đến việc chọn nhầm mẫu sai (vì chúng ta không biết thế nào là mẫu sai thật sự). Tuy nhiên, với cách chọn ngẫu nhiên chúng tôi hy vọng xác suất chọn mẫu sai đó là một

mẫu sai thật sự sẽ lớn hơn xác suất mẫu sai đó sẽ trở thành mẫu đúng trong tương lai.

Cách đánh giá

Chúng tôi sử dụng độ đo *accuracy score* để làm độ đo đánh giá cho mô hình của chúng tôi. Hiệu suất dự đoán được tính bằng cách lấy tổng số dự đoán đúng chia cho tổng số sự đoán. Điểm số này càng cao tức là mô hình dự đoán càng chính xác.

Kết quả đạt được



	PA	CN	JC	RAI	AAI	CNSH	WIC	RASH	DL	LR
accuracy (%)	58.62	72.96	69	66.92	70.26	72.02	69.42	65.66	75.87	72.01

Hình 9. Kết quả đạt được cho tác vụ dự đoán liên kết

Kết quả ở Hình 9 cho thấy, về tổng thể chúng tôi thấy được ngay cả với những đọ đô tương đồng đơn giản như common neighbors (CN) cũng đã đem đến một hiệu suất sự đoán liên kết đủ tốt (hơn 70%). Mô hình dùng deep learning (DL) của chúng tôi về cơ bản chỉ tốt hơn các mô hình cơ sở (sử dụng độ đo tương đồng)

một ít vì thế chúng tôi cho rằng ngay chính cấu trúc mạng của Wikipedia đã đủ đem lại thông tin dự đoán liên kết với hiệu suất đủ tốt.

7. Kết luận

Thông qua việc phân tích sự phát triển và dự đoán liên kết đã trình bày bên trên, chúng tôi kết luận rằng mạng Wikipedia có những thuộc tính vĩ mô tương tự như các mạng thế giới thực, đồng thời cũng có những đặc tính riêng của một mạng liên kết tri thức. Ngoài ra, các đặc trưng cấu trúc của nó hoàn toàn có đầy đủ thông tin cho tác vụ dự đoán liên kết với hiệu suất đủ tốt. Tất cả những kết quả này sẽ giúp ích rất lớn cho Wikipedia trong việc tổ chức, lưu trữ tri thức, đề xuất bài viết,…để phục vụ cộng đồng học thuật ngày một tốt hơn.

8. Tài liệu tham khảo

[1] Zecheng Zhang, Yuan Shi, Xinwei He. "*Evolution and Link Prediction of the Wikipedia Network*", 2019, URL: web.stanford.edu/class/cs224w/project/26424675 .pdf

[2] WikiLinkGraphs dataset. URL: https://consonni.dev/datasets/

[3] Cristian Consonni, David Laniado, Alberto Montresor. "*WikiLinkGraphs: A Complete, Longitudinal and Multi-Language Dataset of the Wikipedia Link Networks*". URL: https://arxiv.org/pdf/1902.04298.pdf

[4] GraphSAGE, "*Inductive Representation Learning on Large Graphs*", W.L. Hamilton, R. Ying, and J. Leskovec arXiv:1706.02216 [cs.SI], 2017.