**ĐO LƯỜNG SỰ TƯƠNG ĐỒNG GIỮA CÁC ĐIỂM CỦA ĐỒ THỊ:**

**ỨNG DỤNG CHO VIỆC TRÍCH XUẤT TỪ ĐỒNG NGHĨA VÀ TÌM KIẾM TRÊN WEB\***

∗ Bài báo này trình bày kết quả đã được thu thập trong một phần của nghiên cứu được thực hiện bởi AG, MH và PS dưới sự hướng dẫn của VB và PVD. Nghiên cứu này được hỗ trợ bởi Quỹ Khoa học Quốc gia dưới Hợp đồng số CCR 99-12415, bởi Đại học Louvain dưới dự án FSR2000 và bởi Chương trình nguyên tố liên đại của Bỉ, được khởi xướng bởi Nhà nước Bỉ, Văn phòng Thủ tướng cho Khoa học, Công nghệ và Văn hóa. Trách nhiệm khoa học nằm ở tác giả của nó.

**Tác giả:** VINCENT D. BLONDEL1, ANAH´I GAJARDO2, MAUREEN HEYMANS3, PIERRE SENELLART4, AND PAUL VAN DOOREN5

1. Khoa Toán Ứng dụng, Đại học Công giáo Louvain, 4 Ave. G. Lemaitre, B-1348 Louvain-la-Neuve, Bỉ ([blondel@inma.ucl.ac.be](mailto:blondel@inma.ucl.ac.be)), http://www.inma.ucl.ac.be/∼blondel.
2. Bộ môn Kỹ thuật Toán học, Đại học Concepción, Hộp thư 160-C, Concepción, Chile ([anahi@ing-mat.udec.cl](mailto:anahi@ing-mat.udec.cl)).
3. Google Inc., 2400 Bayshore Parkway, Mountain View, CA 94043, Hoa Kỳ ([maureen@google.com](mailto:maureen@google.com)).
4. Bộ môn Khoa học Máy tính, Trường Ecole normale supérieure, 45 rue d’Ulm, F-75230 Paris cedex 05, Pháp ([Pierre.Senellart@ens.fr](mailto:Pierre.Senellart@ens.fr)).
5. Bộ môn Toán Ứng dụng, Đại học Công giáo Louvain, 4 Ave. G. Lemaitre, B-1348 Louvain-la-Neuve, Bỉ ([vdooren@csam.ucl.ac.be](mailto:vdooren@csam.ucl.ac.be)), http://www.inma.ucl.ac.be/∼vdooren.

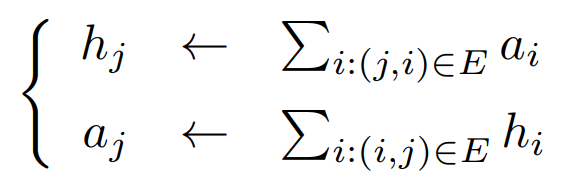
**Tóm tắt:** Chúng tôi giới thiệu một khái niệm về sự tương đồng giữa các đỉnh của đồ thị có hướng. Cho *GA* và *GB* là hai đồ thị có hướng với lần lượt nA và nB là đỉnh. Chúng tôi định nghĩa một ma trận tương đồng **S** kích thước nB × nA, trong đó mục nhập thực *sij* biểu thị mức độ tương đồng giữa đỉnh j (trong *GA*) và đỉnh i (trong *GB*): chúng tôi gọi *sij* là điểm tương đồng của chúng. Ma trận tương đồng có thể được tính toán bằng cách giới hạn của các lần lặp chuẩn hóa của *S(k+1) = BS(k)AT +BT S(k)A* trong đó A và B là ma trận kề của các đồ thị và *S(0)* là một ma trận mà tất cả các phần tử đều bằng một. Trong trường hợp đặc biệt khi *GA = GB = G*, ma trận S là vuông và điểm tương đồng *sij* là điểm tương đồng giữa các đỉnh i và j của *G*. Chúng tôi chỉ ra rằng phương pháp “hub và authority” của Kleinberg để xác định các trang web liên quan đến một truy vấn cụ thể có thể được xem như một trường hợp đặc biệt của định nghĩa của chúng tôi trong trường hợp một trong hai đồ thị có hai đỉnh và một cạnh có hướng duy nhất giữa chúng. Tương tự như Kleinberg, chúng tôi chỉ ra rằng các điểm tương đồng của chúng tôi được cho bởi các thành phần của một vector riêng ưu tiên của một ma trận không âm. Có rất nhiều ứng dụng tiềm năng của khái niệm tương đồng của chúng tôi. Chúng tôi minh họa một ứng dụng cho việc trích xuất tự động các từ đồng nghĩa trong một từ điển đơn ngôn ngữ.

**Từ khóa:** Thuật toán, thuật toán đồ thị, lý thuyết đồ thị, giá trị riêng của đồ thị Phân loại

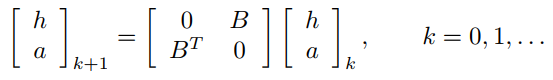
**AMS về chủ đề**: 05C50, 05C85, 15A18, 68R10

1. Tổng quát hóa về hubs và authorities. Các công cụ tìm kiếm web hiệu quả như Google thường dựa trên ý tưởng đặc trưng những đỉnh quan trọng nhất trong một đồ thị đại diện cho các kết nối hoặc liên kết giữa các trang trên web. Một phương pháp như vậy, được đề xuất bởi Kleinberg [18], xác định trong một tập hợp các trang web liên quan đến một truy vấn tìm kiếm, tập con các trang là hubs tốt hoặc tập con các trang là authorities tốt. Ví dụ, đối với truy vấn “university”, các trang chủ của các trường đại học như Oxford, Harvard và các trường khác là authorities tốt, trong khi các trang web trỏ đến các trang chủ này là hubs tốt. Hubs tốt là các trang web trỏ đến authorities tốt, và authorities tốt là các trang web được trỏ đến bởi các hubs tốt. Từ những mối quan hệ ngầm này, Kleinberg đề xuất một phương pháp lặp lại để gán một “điểm authority” và một “điểm hub” cho mỗi đỉnh của một đồ thị cụ thể. Những điểm này có thể được tính toán như giới hạn của một quá trình lặp hội tụ, mà chúng tôi sẽ mô tả sau đây.

Cho *G = (V, E)* là một đồ thị với tập đỉnh V và tập cạnh E, và cho *hj* và *aj* là điểm hub và điểm authority của đỉnh j. Chúng tôi cho phép các điểm này được khởi tạo bằng một số giá trị dương và sau đó cập nhật chúng đồng thời cho tất cả các đỉnh theo quan hệ tương hỗ như sau: điểm hub của đỉnh j được thiết lập bằng tổng của điểm authority của tất cả các đỉnh mà j trỏ đến và tương tự, điểm authority của đỉnh j được thiết lập bằng tổng của điểm hub của tất cả các đỉnh trỏ đến j:



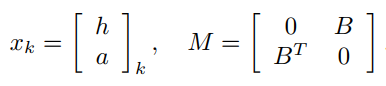
Gọi B là ma trận có phần nhập (i, j) bằng số cạnh giữa các đỉnh i và j trong G (ma trận kề của G), gọi h và a là các vector của điểm hub và điểm authority. Các phương trình cập nhật trên có dạng đơn giản như sau:



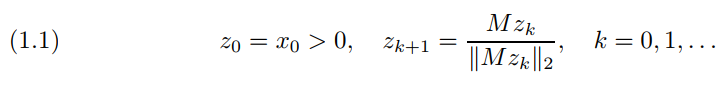
Mà chúng tôi biểu thị bằng hình thức gọn là:



Trong đó:



Lưu ý rằng ma trận M là đối xứng và không âm *(Một ma trận hoặc một vector Z sẽ được gọi là không âm (dương) nếu tất cả các thành phần của nó đều không âm (dương), chúng tôi viết Z ≥ 0 (Z > 0) để biểu thị điều này)*. Chúng tôi chỉ quan tâm đến các điểm số tương đối và vì vậy chúng tôi sẽ xem xét dãy vector được chuẩn hóa như sau:



trong đó ‖ · ‖2 là chuẩn vector Euclidean. Lý tưởng nhất là chúng ta muốn lấy giới hạn của dãy *zk* làm định nghĩa cho điểm hub và điểm authority. Tuy nhiên, có hai khó khăn với một định nghĩa như vậy.

Khó khăn đầu tiên là dãy *zk* không luôn luôn hội tụ. Trên thực tế, các dãy liên quan đến ma trận không âm M có cấu trúc khối như trên hầu như không bao giờ hội tụ mà thay vào đó luân phiên giữa các giới hạn.

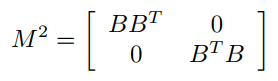


Chúng tôi chứng minh trong Định lý 2 rằng điều này là đúng nói chung đối với các ma trận đối xứng không âm và rằng hoặc dãy kết quả từ (1.1) hội tụ, hoặc nó không hội tụ và sau đó các dãy con chẵn và lẻ cũng hội tụ. Hãy xem xét cả hai giới hạn trong thời điểm này.

Khó khăn thứ hai là các vector giới hạn zeven và zodd nói chung phụ thuộc vào vector ban đầu z0 và không có sự lựa chọn tự nhiên nào cho z0 dường như có. Tập hợp tất cả các vector giới hạn được thu được khi bắt đầu từ một vector ban đầu dương được cho bởi:



và chúng tôi muốn chọn một vector cụ thể trong tập hợp đó. Vector *zeven* thu được cho *z0 =* **1** (chúng tôi biểu thị bằng **1** vector hoặc ma trận có tất cả các mục nhập bằng 1) có một số đặc điểm thú vị khiến nó trở thành một lựa chọn tốt: nó rất dễ tính toán, nó có một số tính chất đẹp (xem đặc biệt ở phần 4), và nó có tính chất cực đại, được chứng minh trong Định lý 2, là vector duy nhất trong Z có chuẩn 1 lớn nhất có thể (chuẩn 1 của một vector là tổng của tất cả các giá trị tuyệt đối của các mục nhập của nó). Bởi vì những đặc điểm này, chúng tôi lựa chọn hai vector con của *zeven(***1***)* như định nghĩa cho điểm hub và điểm authority. Trong trường hợp của ma trận M như trên, chúng ta có:



và từ phương trình này, ta có thể suy ra rằng, nếu các không gian con bất biến có liên quan đến *BBT* và *BTB* có kích thước bằng một, thì các điểm số hub và authority được chuẩn hóa đơn giản bằng các vector riêng ưu tiên được chuẩn hóa của *BBT* và *BTB*. Đây chính là định nghĩa được sử dụng trong [18] cho điểm authority và điểm hub của các đỉnh trong G. Việc chọn tùy ý *z0 =* **1** như trong [18] được chỉ ra ở đây có sự hợp lý từ góc độ chuẩn cực đại. Lưu ý rằng khi không gian con bất biến có kích thước bằng một, thì không có gì đặc biệt về vector bắt đầu 1 vì bất kỳ vector dương *z0* khác nào cũng sẽ cho kết quả giống nhau.

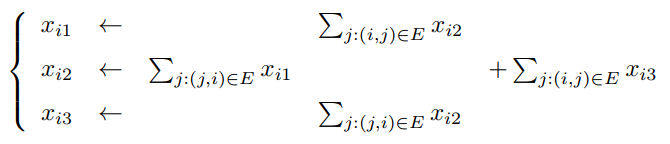
Bây giờ chúng ta sẽ tổng quát hóa xây dựng này. Điểm authority của đỉnh j trong G có thể được coi như một điểm tương đồng giữa đỉnh j của G và điểm authority của đồ thị:



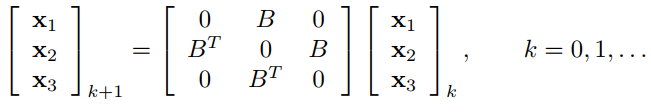
và tương tự, điểm hub của đỉnh j trong G có thể được coi là một điểm tương đồng giữa đỉnh j và đỉnh hub. Quá trình lặp lại tương hỗ lẫn nhau được sử dụng ở trên có thể được tổng quát hóa cho các đồ thị khác với đồ thị cấu trúc hub-authority. Ý tưởng của việc tổng quát hóa này dễ dàng nắm bắt hơn thông qua một ví dụ; chúng tôi minh họa nó trước tiên trên đồ thị đường với ba đỉnh và sau đó đưa ra định nghĩa cho các đồ thị tùy ý. Hãy cho G là một đồ thị có tập cạnh E và ma trận kề B và xem xét đồ thị cấu trúc:

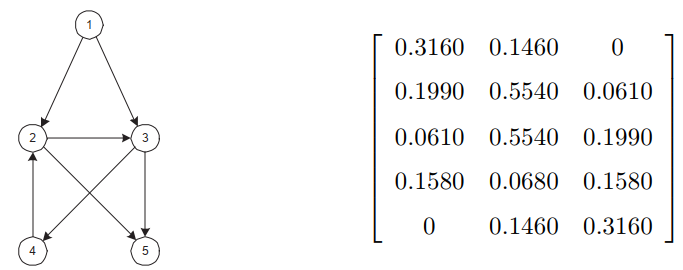


Với mỗi đỉnh j của G, chúng ta bây giờ liên kết ba điểm số *xi1, xi2* và *xi3*; một cho mỗi đỉnh của đồ thị cấu trúc. Chúng tôi khởi tạo các điểm số này bằng một giá trị dương nào đó và sau đó cập nhật chúng theo quan hệ tương hỗ lẫn nhau sau đây:



hoặc dưới dạng ma trận (chúng tôi biểu thị bằng vector cột *xj* với các mục nhập *xij*),

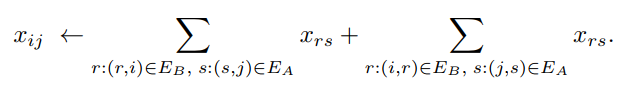




*Hình 1.1: Một đồ thị và ma trận tương đồng của nó với đồ thị cấu trúc 1 → 2 → 3. Điểm tương đồng của đỉnh 3 với đỉnh 2 của đồ thị cấu trúc bằng 0.5540.*

mà chúng ta biểu thị lại bằng *xk+1 = Mxk*. Tình hình bây giờ giống hệt với ví dụ trước đó và tất cả các lập luận về sự hội tụ được đề cập ở đó đều áp dụng ở đây. Ma trận M là đối xứng và không âm, các dãy lặp chẵn và lẻ được chuẩn hóa hội tụ và giới hạn *zeven(***1***)* trong tất cả các giới hạn có thể là vector duy nhất với chuẩn 1 lớn nhất có thể. Chúng tôi lấy ba thành phần của giới hạn cực đại này *zeven*(**1**) làm định nghĩa cho các điểm tương đồng *s1, s2* và *s3* và định nghĩa ma trận tương đồng bằng S = [*s1 s2 s3*]. Một ví dụ số học về một ma trận tương đồng như vậy được thể hiện trong Hình 1. Lưu ý rằng chúng tôi sẽ chứng minh trong Định lý 6 rằng điểm *s2* có thể được tính toán một cách trực tiếp từ B bằng cách tính toán vector riêng ưu tiên của ma trận *BBT +BTB*.

Bây giờ chúng ta đi đến phần mô tả trường hợp tổng quát. Giả sử chúng ta có hai đồ thị có hướng *GA* và *GB* với *nA* và *nB* đỉnh và tập cạnh *EA* và *EB*. Chúng tôi coi *GA* như một đồ thị cấu trúc đóng vai trò của các đồ thị hub → authority và 1 → 2 → 3 trong các ví dụ ở trên. Chúng tôi xem xét điểm số thực *xij* cho *i = 1, . . . ,* *nB* và *j = 1, . . . ,* *nA* và cập nhật tất cả các điểm số đồng thời theo các phương trình cập nhật sau đây:

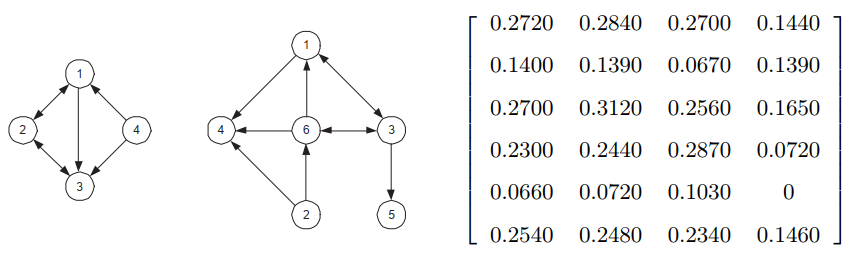


Phương trình này có thể được giải thích dưới góc độ của đồ thị tích của GA và GB. Đồ thị tích của GA và GB là một đồ thị có nA.nB đỉnh và có một cạnh giữa các đỉnh *(i1, j1)* và *(i2, j2)* nếu có một cạnh giữa i1 và i2 trong GA và có một cạnh giữa j1 và j2 trong GB. Phương trình cập nhật (1.1) sau đó tương đương với việc thay thế các giá trị của tất cả các đỉnh của đồ thị tích bằng các giá trị của các đỉnh ra và đỉnh vào trong đồ thị.

Phương trình (1.1) cũng có thể được viết dưới dạng ma trận gọn hơn. Đặt *Xk* là ma trận kích thước nB × nA với các mục nhập *xij* ở lần lặp *k*. Sau đó, các phương trình cập nhật có dạng đơn giản như sau:



trong đó A và B là các ma trận kề của GA và GB. Chúng tôi chứng minh ở Phần 3 rằng, giống như ví dụ ở trên, các dãy lặp chẵn và lẻ được chuẩn hóa của phương trình cập nhật này hội tụ, và giới hạn Zeven(**1**) trong tất cả các giới hạn có thể là duy nhất có chuẩn 1 lớn nhất. Chúng tôi lấy giới hạn này làm định nghĩa cho ma trận tương đồng. Một ví dụ về hai đồ thị và ma trận tương đồng của họ được hiển thị trong Hình 1.2.



*Hình 1.2: Hai đồ thị GA, GB và ma trận tương đồng của họ. Đỉnh của GA mà giống nhất với đỉnh 5 trong GB là đỉnh 3.*

Thật thú vị khi lưu ý rằng trong tài liệu cơ sở dữ liệu đã có các ý tưởng tương tự đã được đề xuất [20], [17]. Các ứng dụng trong đó liên quan đến truy xuất thông tin trong cơ sở dữ liệu lớn mà có thể được liên kết với một cấu trúc đồ thị cụ thể. Những ý tưởng được trình bày trong các bài báo hội nghị này rõ ràng liên quan đến những ý tưởng của bài báo hiện tại, nhưng để đảm bảo tính hội tụ của quá trình lặp được đề xuất đến một điểm cố định duy nhất, quá trình lặp phải được điều chỉnh một chút bằng cách sử dụng ví dụ như các hệ số trọng số cụ thể. Do đó, điểm hub và điểm authority của Kleinberg không phải là một trường hợp đặc biệt của định nghĩa của họ. Chúng tôi cảm ơn các tác giả của [17] đã gợi ý cho chúng tôi tham khảo những tài liệu này.

Phần còn lại của bài báo này được tổ chức như sau. Ở Phần 2, chúng tôi mô tả một số kết quả tiêu biểu Perron-Frobenius cho các ma trận không âm có ích trong phần còn lại của bài báo. Ở Phần 3, chúng tôi đưa ra một định nghĩa chính xác về ma trận tương đồng cùng với các định nghĩa thay thế khác nhau. Định nghĩa ngay lập tức chuyển thành một thuật toán xấp xỉ và chúng tôi thảo luận trong phần đó một số khía cạnh về độ phức tạp của thuật toán. Ở Phần 4, chúng tôi mô tả các ma trận tương đồng cho trường hợp mà một trong hai đồ thị có độ dài là 2 hoặc 3. Ở Phần 5, chúng tôi xem xét trường hợp đặc biệt *GA = GB = G* trong đó điểm sij là sự tương đồng giữa các đỉnh i và j trong một đồ thị duy nhất G. Phần 6 đề cập đến các đồ thị mà ma trận tương đồng của chúng có hạng là một. Chúng tôi chứng minh rằng nếu một trong hai đồ thị là đồ thị đều hoặc nếu một trong hai đồ thị là đồ thị vô hướng, thì ma trận tương đồng có hạng là một. Các đồ thị đều là các đồ thị mà tất cả các đỉnh của chúng có cùng bậc vào và cùng bậc ra; ví dụ, đồ thị chu kỳ là các đồ thị đều. Ở một phần cuối cùng, chúng tôi báo cáo các kết quả được thu được cho việc trích xuất từ đồng nghĩa tự động trong một từ điển bằng cách sử dụng điểm số trung tâm của một đồ thị. Một phiên bản ngắn gọn của bài báo này xuất hiện như một đóng góp hội nghị trong [6].

1. **Đồ thị và ma trận không âm:** Với bất kỳ đồ thị có hướng *G = (V, E)* nào, bạn có thể liên kết với một ma trận không âm thông qua việc đánh chỉ mục các đỉnh của nó. Các ma trận kề của G là ma trận *B ∈ Nn×n* có mục *bij* bằng số cạnh từ đỉnh i đến đỉnh j. Gọi B là ma trận kề của một số đồ thị G nào đó; mục *(Bk)ij* bằng số lượng đường đi có độ dài k từ đỉnh i đến đỉnh j. Từ đó suy ra rằng một đồ thị là mạnh liên thông nếu và chỉ nếu đối với mọi cặp chỉ số i và j có một số nguyên k sao cho *(Bk)ij* *> 0*. Các ma trận thỏa mãn tính chất này được gọi là không phân chia được (irreducible).

Tiếp theo, chúng ta sẽ cần đến khái niệm về phép chiếu vuông góc lên các không gian con vector. Đặt *V* là một không gian con tuyến tính của *Rn* và *v ∈ Rn*. Phép chiếu vuông góc của *v* lên *V* là vector trong *V* có khoảng cách Euclidean nhỏ nhất đến *v*. Một biểu diễn ma trận của phép chiếu này có thể được thu được như sau. Đặt {*v1, . . ., vm*} là một cơ sở trực giao cho *V* và sắp xếp các vector cột vi vào một ma trận *V*. Phép chiếu của *v* lên *V* được cho bởi Π*v = V VTv* và ma trận Π *= V VT* là bộ chiếu vuông góc lên *V*. Các bộ chiếu có tính chất *Π2 = Π*.

Lý thuyết Perron-Frobenius [14] xác định các tính chất thú vị về các vector riêng và giá trị riêng của các ma trận không âm. Đặt ρ(M) là giá trị tuyệt đối lớn nhất của các giá trị riêng của ma trận M (bán kính phổ của M). Theo Định lý Perron-Frobenius, bán kính phổ của một ma trận không âm M là một giá trị riêng của M (được gọi là gốc Perron), và tồn tại một vector không âm kết hợp *x ≥ 0 (x ∕= 0)* sao cho *Mx = ρx* (gọi là vector Perron). Trong trường hợp của ma trận đối xứng, có thể thu được kết quả cụ thể hơn.

*Định lý 1:* Cho M là một ma trận đối xứng không âm với bán kính phổ ρ. Sau đó, số lượng đại số và số lượng hình học của gốc Perron ρ bằng nhau; có một cơ sở không âm *X ≥ 0* cho không gian con bất biến liên quan đến gốc Perron; và các phần tử của bộ chiếu vuông góc Π trên không gian vector liên quan đến gốc Perron của M đều là số không âm.

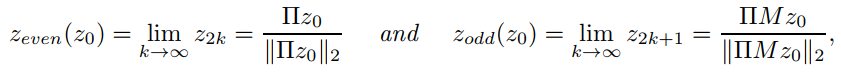
*Chứng minh:* Chúng ta sử dụng các sự thật rằng bất kỳ ma trận đối xứng không âm M nào có thể được sắp xếp thành một ma trận chéo khối với các khối không phân chia được *Mi* trên đường chéo [14, 3], và số lượng đại số của gốc Perron của một ma trận không âm không phân chia được bằng một. Từ những sự kết hợp này, ta có được rằng số lượng đại số và số lượng hình học của gốc Perron ρ của M bằng nhau. Hơn nữa, không gian con bất biến tương ứng của M được thu được từ các vector Perron đã được chuẩn hóa của các khối *Mi*, được đệm một cách thích hợp bằng số không. Cách thu được cơ sở *X* trong trường hợp đó cũng là không âm và trực giao.

Định lý tiếp theo sẽ được sử dụng để chứng minh định nghĩa ma trận tương đồng giữa hai đồ thị của chúng ta. Kết quả mô tả các vector giới hạn của các dãy liên quan đến các phép biến đổi tuyến tính đối xứng không âm.

*Định lý 2:* Cho M là một ma trận đối xứng không âm với bán kính phổ ρ. Đặt *z0 > 0* và xem xét dãy



*Có hai trường hợp hội tụ có thể xảy ra tùy thuộc vào việc −ρ có phải là giá trị riêng của M hay không. Khi −ρ không phải là giá trị riêng của M, thì dãy zk đơn giản hội tụ đến Πz0/*‖*Πz0*‖*2, trong đó Π là bộ chiếu vuông góc trên không gian con bất biến liên quan đến gốc Perron ρ. Khi −ρ là giá trị riêng của M, thì các dãy con z2k và z2k+1 hội tụ đến các giới hạn:*



*trong đó Π là bộ chiếu vuông góc trên tổng của các không gian con bất biến liên quan đến ρ và −ρ. Trong cả hai trường hợp, tập hợp tất cả các giới hạn có thể có được cho bởi:*

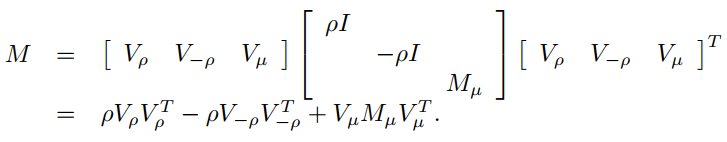


*và vector zeven*(**1**) *là vector duy nhất có tổng trị số tuyệt đối lớn nhất trong tập hợp đó.*

*Chứng minh:* Chúng tôi chỉ chứng minh trường hợp khi −ρ là giá trị riêng; trường hợp khác là một biến thể trivial. Hãy ký hiệu không gian con bất biến của M tương ứng với ρ, tương ứng với −ρ và tương ứng với phần còn lại của phổ bởi *Vρ, V−ρ* và *Vµ*. Giả sử rằng các không gian này không trivialis, và chúng ta có các cơ sở trực giao cho chúng:



trong đó *Mµ* là một ma trận vuông (đường chéo nếu *Vµ* là cơ sở của các vector riêng) có bán kính phổ µ nghiêm ngặt nhỏ hơn ρ. Phân rã giá trị riêng sau đó có thể được viết lại dưới dạng đường chéo khối:



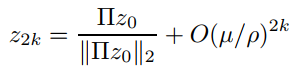
Sau đó, ta có:



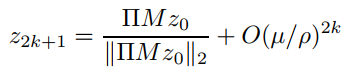
trong đó Π *:= VρVρT + V−ρVT−ρ* là bộ chiếu vuông góc vào không gian con bất biến *Vρ ⊕ V−ρ* của *M2* tương ứng với *ρ2*. Chúng ta cũng có:



và vì *ρ(Mµ) = µ < ρ*, nên từ việc nhân với *z0* và *Mz0*, ta có:



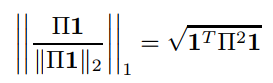
và



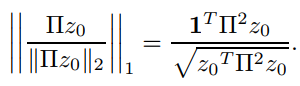
miễn là các vector ban đầu *z0* và *Mz0* có thành phần khác không trong các không gian con liên quan, tức là miễn là *Πz0* và *ΠMz0* không bằng không. Nhưng norm Euclidean của các vector này bằng *zT0* *Πz0* và *zT0MΠMz0* khi Π2 = Π. Cả hai norm này đều khác không vì *z0 > 0* và cả Π và *M*Π*M* đều không âm và khác không.

Kết luận từ tính không âm của *M* và công thức cho *zeven(z0)* và *zodd(z0)* là cả hai giới hạn đều nằm trong {*Πz/‖Πz‖2: z > 0*}. Bây giờ chúng ta hãy chứng minh rằng mọi phần tử zˆ0 ∈ {*Πz/‖Πz‖2: z > 0*} có thể được thu được như *zeven(z0)* cho một *z0 > 0*. Vì các phần tử của Π không âm, nên cũng như những phần tử của ˆz0. Tuy nhiên, vector này có thể có một số phần tử bằng không. Từ ˆz0, chúng ta xây dựng z0 bằng cách thêm ǫ vào tất cả các phần tử bằng không của ˆz0. Vector *z0 - zˆ0* rõ ràng vuông góc với *Vρ ⊕ V−ρ* và do đó sẽ biến mất trong lặp của *M2*. Do đó, chúng ta có *zeven(z0) = ˆz0* cho *z0 > 0*, như yêu cầu.

Bây giờ chúng ta chứng minh câu lệnh cuối cùng. Ma trận Π và tất cả các vector đều không âm và Π2 = Π, và vì vậy:



và cũng



Áp dụng bất đẳng thức Schwarz cho Π*z0* và Π**1**, ta có:



với sự bằng nhau chỉ khi Π*z0 = λ*Π**1** cho một số *λ ∈ C*. Nhưng vì Π*z0* và Π**1** đều là các số thực không âm, bằng chứng dễ dàng được chứng minh.

1. **Tương đồng giữa các đỉnh trong đồ thị:** Bây giờ chúng ta sẽ đưa ra một định nghĩa chính thức về ma trận tương đồng của hai đồ thị có hướng GA và GB. Công thức cập nhật cho ma trận tương đồng đã được đề cập trong phần giới thiệu và được biểu diễn thông qua phép ánh xạ tuyến tính:

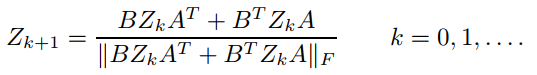


trong đó A và B là ma trận kề của GA và GB. Trong phương trình cập nhật này, các phần tử của *Xk+1* phụ thuộc tuyến tính vào các phần tử của *Xk*. Chúng ta có thể làm cho sự phụ thuộc này rõ ràng hơn bằng cách sử dụng phép toán ma trận thành vector, chuyển đổi một ma trận thành một vector bằng cách lấy các cột của nó một cách tuần tự. Phép toán này, được ký hiệu là vec, thỏa mãn tính chất cơ bản *vec(CXD) = (DT ⊗C) vec(X)* trong đó ⊗ biểu thị cho phép nhân Kronecker (còn được gọi là tích tensorial, trực tiếp hoặc hạng mục). Để xem chứng minh cho tính chất này, xem Bổ đề 4.3.1 trong [15]. Áp dụng tính chất này vào (3.1), chúng ta ngay lập tức thu được:



trong đó *xk = vec(Xk)*. Đây là định dạng được sử dụng trong phần giới thiệu. Kết hợp quan sát này với Định lý 2, chúng ta suy ra Định lý sau đây.

*Định lý 3: Cho GA và GB là hai đồ thị với các ma trận kề A và B, xác định một ma trận khởi đầu dương Z0 > 0 và định nghĩa:*



*Khi đó, các ma trận con Z2k và Z2k+1 hội tụ đến Zeven và Zodd. Hơn nữa, trong tất cả các ma trận thuộc tập hợp:*



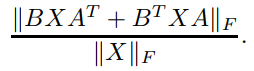
*ma trận Zeven*(**1**) *là ma trận duy nhất có norm 1 lớn nhất.*

Để đồng nhất với chuẩn vector xuất hiện trong Định lý 2, chuẩn ma trận *‖*.*‖*F mà chúng ta sử dụng ở đây là căn bậc hai của tổng bình phương của tất cả các phần tử (chuẩn này được gọi là chuẩn Euclid hoặc chuẩn Frobenius), và chuẩn 1-norm *‖*.*‖*1 là tổng giá trị tuyệt đối của tất cả các phần tử của nó. Chúng ta cũng có thể cung cấp một định nghĩa của tập Z dựa trên một trong những tính chất cực đại của nó.

*Định lý 4:* *Cho GA và GB là hai đồ thị có ma trận kề là A và B và xem xét các ký hiệu của Định lý 3. Tập hợp:*



*và tập hợp tất cả các ma trận dương mà làm tối đa hóa biểu thức:*



*là bằng nhau. Hơn nữa, trong tất cả các ma trận trong tập hợp này, có một ma trận duy nhất* ***S*** *có chuẩn 1 lớn nhất. Ma trận này có thể được thu được như sau:*



*khi Z0 =* **1***.*

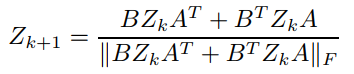
*Chứng minh*: Biểu thức trên cũng có thể được viết lại thành *‖*L(X)*‖*2/*‖X‖*2, là chuẩn 2 được tạo ra bởi phép ánh xạ tuyến tính *L* được định nghĩa bởi *L(X) = BXAT + BT XA*. Nó được biết [14] rằng mỗi vector riêng ưu dominan X của *L2* là một bộ tối ưu hóa của biểu thức này. Như đã được chứng minh ở trên, **S** là ma trận duy nhất có chuẩn 1 lớn nhất trong tập hợp đó.

Chúng ta lấy ma trận **S** xuất hiện trong Định lý này làm định nghĩa cho ma trận tương đồng giữa GA và GB. Lưu ý rằng theo định nghĩa này, ma trận tương đồng giữa GB và GA chính là ma trận chuyển vị của ma trận tương đồng giữa GA và GB.

Một biểu đồ trực tiếp của định nghĩa dẫn đến một thuật toán xấp xỉ để tính toán ma trận tương đồng của các đồ thị:

*1. Đặt Z0 = 1.*

*2. Lặp lại một số lần chẵn:*



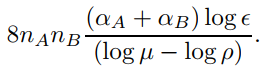
*và dừng lại khi Zk hội tụ.*

*4. Xuất ra* ***S****.*

Thuật toán này tương tự với phương pháp lũy thừa ma trận cổ điển (xem [14]) để tính một vector riêng ưu tiên của một ma trận. Phức tạp của thuật toán này dễ dàng ước tính. Giả sử GA, GB là hai đồ thị có số đỉnh là nA, nB và số cạnh lần lượt là eA, eB. Sau đó, tích *BZk* và *BT Zk* đòi hỏi ít hơn *2nA.eB* phép cộng và phép nhân mỗi phép tính, trong khi các tích sau này *(BZk)AT* và *(BT Zk)A* đòi hỏi ít hơn *2nB.eA* phép cộng và phép nhân mỗi phép tính. Tổng và tính chuẩn Frobenius đòi hỏi *2nA.nB* phép cộng và phép nhân, trong khi việc tỉ lệ đòi hỏi một phép chia và *nA.nB* phép nhân. Chúng ta định nghĩa:



Nếu chúng ta đặt αA và αB là số lượng trung bình các phần tử khác không trên mỗi hàng của ma trận A và B tương ứng, thì tổng số phép tính cho mỗi bước lặp là xấp xỉ bằng *4(αA + αB)nAnB* phép cộng và nhân. Như đã chứng minh trong Định lý 2, sự hội tụ của các lần lặp chẵn của phương trình tái phát trên là tuyến tính với tỷ số *(µ/ρ)2*. Số lần tính toán dấu phẩy động cần thiết để tính toán ma trận **S** với độ chính xác ǫ bằng phương pháp lũy thừa ma trận sẽ ở mức:



Các phương pháp xử lý ma trận thưa thớt khác có thể được sử dụng ở đây, nhưng trong bài báo này, chúng tôi không xem xét các khía cạnh thuật toán như vậy. Đối với các lớp cụ thể của ma trận kề, bạn có thể tính trực tiếp ma trận tương đồng **S** từ các không gian không đổi ưu dominan của các ma trận có kích thước tương đương với A hoặc B. Chúng tôi sẽ cung cấp các biểu thức cụ thể cho một số lớp như vậy trong phần tiếp theo.

1. **Hubs, authorities và các điểm trung tâm:** Như đã được giải thích trong phần giới thiệu, điểm hub và authority của một đồ thị có thể được biểu diễn dựa trên ma trận kề của nó.

*Định lý 5:* Gọi B là ma trận kề của đồ thị GB. Các điểm hub chuẩn hóa và điểm authority chuẩn hóa của các đỉnh trong GB được cho bởi các vector riêng độc lập chuẩn hóa của các ma trận BBT và BT B, miễn là giá trị riêng Perron tương ứng có bội là 1. Nếu không, chúng là phép chiếu chuẩn hóa của vector 1 lên các không gian con không đổi riêng độc lập tương ứng.

Điều kiện về bội của giá trị riêng Perron không thừa. Thực tế, ngay cả đối với các đồ thị liên thông, BBT và BT B có thể có nhiều giá trị riêng chủ đạo: ví dụ như đồ thị chu trình, cả BBT và BT B đều là ma trận đơn vị.

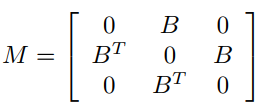
Một cấu trúc đồ thị khác thú vị là đồ thị đường dẫn có độ dài ba:



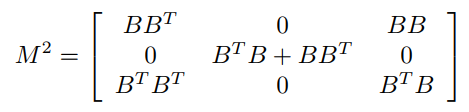
Về điểm hub và authority, chúng ta có thể đưa ra một biểu thức rõ ràng cho điểm tương đồng với đỉnh 2, một điểm số mà chúng ta sẽ gọi là điểm trung tâm. Điểm trung tâm này đã được sử dụng thành công để tự động trích xuất từ đồng nghĩa trong một từ điển. Ứng dụng này được mô tả chi tiết hơn trong Phần 7.

*Định lý 6:* Gọi B là ma trận kề của đồ thị GB. Các điểm số trung tâm chuẩn hóa của các đỉnh trong GB được cho bởi vector riêng độc lập chuẩn hóa của ma trận BT B + BBT, miễn là giá trị riêng Perron tương ứng có bội là 1. Nếu không, chúng là phép chiếu chuẩn hóa của vector 1 lên không gian con không đổi chủ đạo.

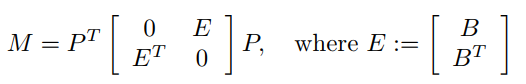
*Chứng minh:* Ma trận M tương ứng như sau:



và do đó



và kết quả sau đó được suy ra từ định nghĩa của các điểm tương đồng, miễn là ma trận trung tâm BT B + BBT có một giá trị riêng chủ đạo ρ2 của M2. Điều này có thể được thấy như sau. Ma trận M có thể được hoán vị để:



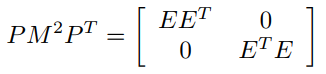
Giờ ta xét V và U là các cơ sở trực giao cho các không gian riêng độc lập chủ đạo phải và trái của E [14]:



rõ ràng V và U cũng là các cơ sở cho các không gian con không đổi chủ đạo tương ứng của ET E và EET, vì:



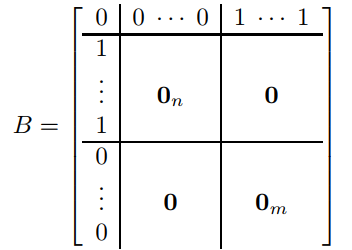
Hơn nữa,



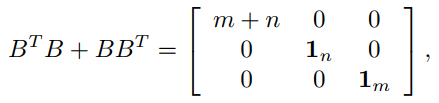
và các bộ chiếu liên quan đến các giá trị riêng chủ đạo của EET và ET E tương ứng là Πv := VVT và Πu := UUT. Bộ chiếu Π của M2 không gì khác ngoài *PTdiag*{Πv, Πu}*P* và do đó các vector con của Π**1** chính là các vector Πv**1** và Πu**1**, có thể tính từ các ma trận nhỏ hơn ET E hoặc EET. Vì ET E = BT B + BBT, vector trung tâm Πv**1** là vector ở giữa của Π**1**. Đáng chú ý rằng (4.1) cũng cho ta một mối quan hệ giữa hai bộ chiếu nhỏ hơn:



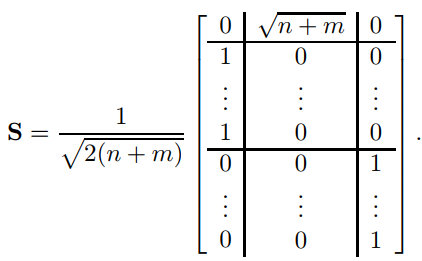
Để minh họa rằng đồ thị đường dẫn có độ dài 3 có thể có lợi thế hơn so với đồ thị cấu trúc hub-authority mà chúng ta đang xem xét, chúng ta xem xét trường hợp đặc biệt của “đồ thị hình cúp” GB được biểu diễn trong Hình 4.1. Nếu chúng ta đánh số nhãn cho đỉnh trung tâm trước, sau đó là nhãn cho m đỉnh bên trái và cuối cùng là nhãn cho n đỉnh bên phải, ma trận kề cho đồ thị này được cho bởi

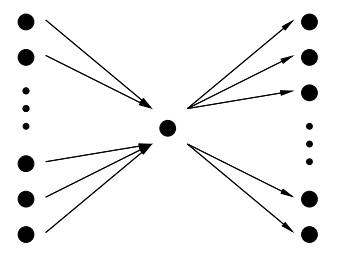


Ma trận BT B + BBT bằng với



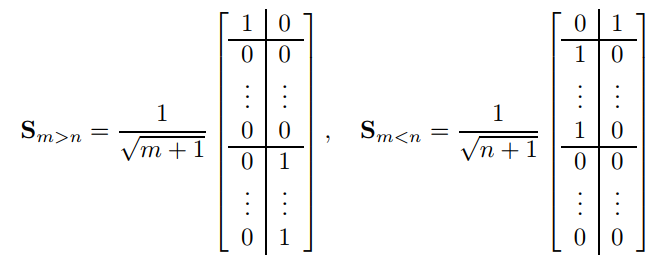
và, theo Định lý 6, giá trị riêng Perron của M bằng ρ = và ma trận tương đồng được cho bởi ma trận (1 + m + n) × 3 như sau:





*Hình 4.1: Một đồ thị hình cúp có hướng. Điểm hub của đỉnh trung tâm theo công thức của Kleinberg bằng 1/ nếu m > n và bằng 0 nếu m < n. Điểm trung tâm của đỉnh này bằng 1/ độc lập với giá trị tương đối của m và n.*

Kết quả này có hiệu lực bất kể giá trị tương đối của m và n. Hãy gọi ba đỉnh của đồ thị đường dẫn lần lượt là 1, trung tâm và 3. Một cách nhìn khác, một trung tâm có thể được coi là một đỉnh thông qua đó thông tin được truyền đi nhiều. Ma trận tương đồng S này cho thấy rằng đỉnh 1 của GB rất giống một trung tâm, các đỉnh bên trái của GB giống như 1 và các đỉnh bên phải của GB giống như 3. Tuy nhiên, nếu chúng ta phân tích đồ thị GB dựa trên đồ thị cấu trúc hub-authority của Kleinberg, các điểm tương đồng S sẽ khác nhau khi m < n và m > n:

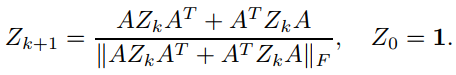


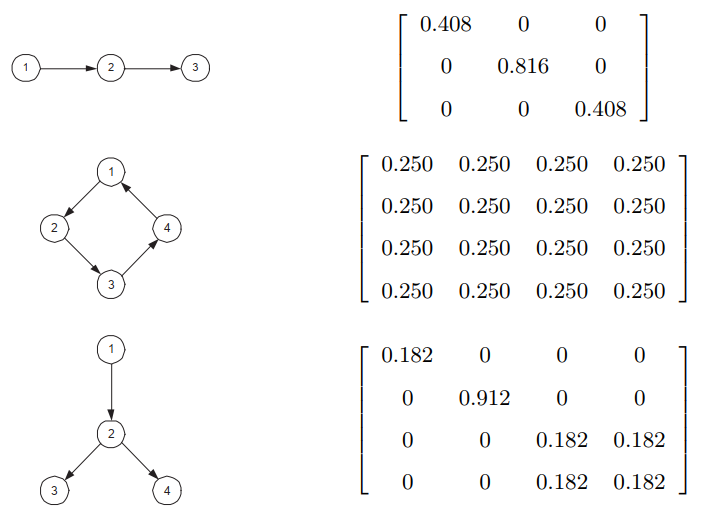
Điều này cho thấy một điểm yếu của đồ thị cấu trúc này, vì các đỉnh của GB mà xứng đáng được gán nhãn là “hub” hoặc “authority” hoàn toàn thay đổi giữa trường hợp m > n và trường hợp m < n.

1. **Ma trận tự tương đồng của một đồ thị:** Khi chúng ta so sánh hai đồ thị bằng nhau GA = GB = G, ma trận tương đồng S là một ma trận vuông có các phần tử là điểm tương đồng giữa các đỉnh của G; ma trận này là ma trận tự tương đồng của G. Các đồ thị khác nhau và ma trận tự tương đồng tương ứng của chúng được đại diện trong Hình 5.1. Nói chung, chúng ta mong đợi các đỉnh có điểm tương đồng cao với chính họ; tức là, chúng ta mong đợi các phần tử trên đường chéo chính của ma trận tự tương đồng là lớn. Chúng tôi chứng minh trong định lý tiếp theo rằng phần tử lớn nhất của ma trận tự tương đồng luôn xuất hiện trên đường chéo chính và rằng, trừ trường hợp trivial, các phần tử trên đường chéo chính của ma trận tự tương đồng không bằng không. Như có thể thấy từ các ví dụ cơ bản, tuy nhiên, không đúng rằng các phần tử trên đường chéo chính luôn áp đảo tất cả các phần tử trên cùng một hàng và cột.

*Định lý 7:* Ma trận tự tương đồng của một đồ thị là bán xác định dương. Đặc biệt, phần tử lớn nhất của ma trận xuất hiện trên đường chéo chính và nếu một phần tử trên đường chéo chính bằng không, thì hàng và cột tương ứng cũng bằng không.

*Chứng minh:* Vì A = B, việc lặp của các ma trận chuẩn hóa Zk bây giờ trở thành:





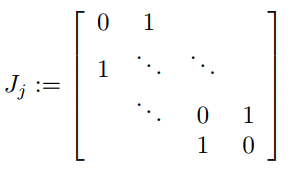
*Hình 5.1: Đồ thị và ma trận tự tương đồng tương ứng của chúng. Ma trận tự tương đồng của một đồ thị cung cấp một đo lường về mức độ tương đồng giữa các đỉnh với nhau.*

Vì tổng tỉ lệ của hai ma trận bán xác định dương cũng là một ma trận bán xác định dương, rõ ràng rằng tất cả các ma trận Zk sẽ là ma trận bán xác định dương. Hơn nữa, ma trận bán xác định dương là một tập đóng và do đó giới hạn S cũng sẽ là ma trận bán xác định dương. Các tính chất được đề cập trong câu đề của định lý là các tính chất rõ ràng của ma trận bán xác định dương.

Khi các đỉnh của đồ thị tương tự nhau, như trong đồ thị vòng tròn, chúng ta mong đợi có một ma trận tự tương đồng với tất cả các phần tử bằng nhau. Điều này thực sự là đúng như sẽ được chứng minh ở phần tiếp theo. Chúng ta cũng có thể tìm ra biểu thức rõ ràng cho ma trận tự tương đồng của các đồ thị đường dẫn.

*Định lý 8:* Ma trận tự tương đồng của một đồ thị đường dẫn là một ma trận đường chéo.

*Chứng minh:* Tích của hai đồ thị đường dẫn là một hợp của các đồ thị đường dẫn không giao nhau, và do đó ma trận M tương ứng với đồ thị này có thể được hoán vị thành một sắp xếp chéo khối của các ma trận Jacobi.



Các ma trận Jacobi có kích thước j = 1, ..., ℓ, trong đó ℓ là kích thước của đồ thị đường dẫn đã cho. Khối lớn nhất trong các khối này tương ứng với phần tử Perron ρ của ma trận M. Chỉ có một khối lớn nhất và các đỉnh của nó tương ứng với các phần tử trên đường chéo của **S**. Như được chỉ ra trong [19], ρ = 2 cos(π/(ℓ + 1)) nhưng M có cả giá trị riêng ±ρ và các vector tương ứng có các phần tử (±)j sin(jπ/(ℓ + 1)), j = 1, ..., ℓ, từ đó Π**1** có thể được tính toán một cách dễ dàng.

1. **Ma trận tương đồng của hạng một:** Trong phần này, chúng tôi mô tả hai lớp đồ thị dẫn đến ma trận tương đồng có hạng một. Chúng tôi xem xét trường hợp một trong hai đồ thị là đồ thị đều (một đồ thị được coi là đều nếu các bậc vào của các đỉnh của nó đều nhau và các bậc ra cũng đều nhau), và trường hợp ma trận kề của một trong hai đồ thị là ma trận chính quy (một ma trận A được coi là chính quy nếu nó thỏa mãn AAT = AT A). Trong cả hai trường hợp, chúng tôi chứng minh rằng ma trận tương đồng có hạng một. Các đồ thị không có hướng có ma trận kề đối xứng và ma trận đối xứng là ma trận chính quy, do đó các đồ thị không có hướng luôn tạo ra các ma trận tương đồng có hạng một.

*Định lý 9: Giả sử GA, GB là hai đồ thị có ma trận kề A và B và giả sử GA là đồ thị đều. Khi đó, ma trận tương đồng giữa GA và GB là một ma trận hạng một có dạng:*



*trong đó v = Π****1*** *là phép chiếu của* ***1*** *vào không gian con không đổi chủ đạo của (B + BT)2, và α là một hệ số tỷ lệ.*

*Chứng minh:* Được biết (xem, ví dụ như [4]) rằng một đồ thị đều GA có ma trận kề A với phần tử Perron có đa thức đại số bằng 1 và vector Perron tương ứng là **1** cho cả A và AT. Dễ dàng suy ra từ đó rằng mỗi ma trận Zk của quá trình lặp xác định ma trận tương đồng là ma trận hạng một và có dạng vk**1**T /, trong đó



Điều này rõ ràng hội tụ đến Π**1**/‖Π**1**‖2 nơi Π là phép chiếu vào không gian con không đổi chủ đạo của (B + BT)2.

Đồ thị vòng có ma trận kề A thỏa mãn AAT = I. Thuộc tính này tương ứng với việc trong một đồ thị vòng, tất cả các đường đi thuận-lùi từ một đỉnh quay trở lại đỉnh đó. Nói chung hơn, chúng ta xem xét trong định lý tiếp theo các đồ thị có ma trận kề A là ma trận chính quy, tức là có ma trận kề A sao cho AAT = AT A.

*Định lý 10: Giả sử GA và GB là hai đồ thị có ma trận kề A và B và giả sử một trong hai ma trận kề là chính quy. Khi đó, ma trận tương đồng giữa GA và GB có hạng một.*

*Chứng minh:* Đặt A là ma trận chính quy và α là phần tử Perron của nó. Khi đó, tồn tại một ma trận unitary U chéo hóa cả A và AT:



và các cột *ui, i = 1, ..., nA* của U là các vector riêng chung của chúng (lưu ý rằng ui chỉ là số thực nếu λi cũng là số thực). Do đó,



và các giá trị riêng của M là các giá trị riêng của các ma trận Hermitian.



mà rõ ràng được giới hạn bởi |λi|β trong đó β là phần tử Perron của (B + BT). Hơn nữa, nếu *v(i)j, j = 1, ..., nB* là các vector riêng của Hi thì các vector riêng của M được cho bởi



và chúng chỉ có thể là số thực nếu λi là số thực. Vì chúng ta muốn có các vector riêng thực tương ứng với các giá trị riêng cực đại của M, chúng ta chỉ cần xem xét các giá trị riêng thực lớn nhất của A, tức là ±α, trong đó α là phần tử Perron của A. Vì A là ma trận chính quy, chúng ta cũng có rằng các vector riêng thực của nó cũng là vector riêng của AT. Do đó,



Kết quả là:



và do đó Π := (Π+α + Π−α) ⊗ Πβ là phép chiếu của phần tử Perron chủ đạo α2β2 của M2. Áp dụng phép chiếu này vào vector 1 ta thu được vector



tương ứng với ma trận hạng một.



Khi một trong hai đồ thị GA hoặc GB là đồ thị đều hoặc có ma trận kề chính quy, ma trận tương đồng S kết quả có hạng một. Ma trận kề của các đồ thị đều và ma trận chính quy có thuộc tính rằng phép chiếu Π vào không gian con không đổi tương ứng với phần tử Perron của A cũng là phép chiếu vào không gian con không đổi của AT. Như kết quả, ρ(A + AT) = 2ρ(A). Trong ngữ cảnh này, chúng ta đưa ra giả thuyết sau đây.

*Giả thuyết 11: Ma trận tương đồng của hai đồ thị có hạng một nếu và chỉ nếu một trong hai đồ thị có thuộc tính rằng ma trận kề D của nó thỏa mãn ρ(D+DT) = 2ρ(D).*

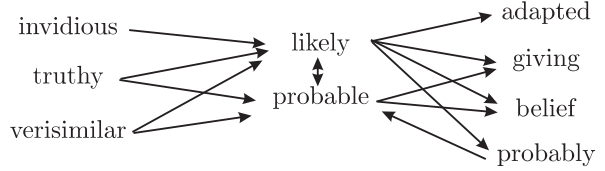
1. **Ứng dụng vào việc trích xuất tự động các từ đồng nghĩa:** Chúng tôi minh họa trong phần cuối này việc sử dụng điểm tương đồng trung tâm được giới thiệu trong Phần 4 để trích xuất tự động các từ đồng nghĩa từ một từ điển đơn ngữ. Phương pháp của chúng tôi sử dụng một đồ thị được xây dựng từ từ điển và dựa trên giả định rằng các từ đồng nghĩa có nhiều từ chung trong định nghĩa và xuất hiện cùng trong định nghĩa của nhiều từ khác. Chúng tôi tóm tắt ngắn gọn phương pháp của chúng tôi dưới đây và sau đó thảo luận về kết quả thu được từ từ điển Webster với bốn từ truy vấn. Ứng dụng này được trình bày trong phần này dựa trên [5], chúng tôi đề nghị độc giả quan tâm tham khảo để có một mô tả đầy đủ.

Phương pháp khá đơn giản. Bắt đầu từ một từ điển, chúng tôi đầu tiên xây dựng đồ thị từ điển G tương ứng; mỗi từ trong từ điển là một đỉnh của đồ thị và có một cạnh từ u đến v nếu v xuất hiện trong định nghĩa của u. Sau đó, liên quan đến một từ truy vấn cụ thể w, chúng tôi xây dựng một đồ thị hàng xóm Gw là đồ thị con của G mà các đỉnh được chỉ đến bởi w hoặc đang chỉ đến w (xem ví dụ, ví dụ hình 7).

Cuối cùng, chúng tôi tính toán điểm tương đồng của các đỉnh trong đồ thị Gw với đỉnh trung tâm trong đồ thị cấu trúc.



và xếp hạng các từ theo thứ tự giảm dần của điểm số. Vì cách mà đồ thị hàng xóm được xây dựng, chúng tôi mong đợi các từ có điểm số trung tâm cao nhất sẽ là những ứng cử viên tốt cho việc đồng nghĩa.



*Hình 7.1: Một phần của đồ thị hàng xóm liên quan đến từ “likely”. Đồ thị chứa tất cả các từ được sử dụng trong định nghĩa của “likely” và tất cả các từ sử dụng “likely” trong định nghĩa của chúng. Các từ đồng nghĩa được xác định bằng cách chọn trong đồ thị này các đỉnh có điểm số trung tâm cao nhất.*

Trước khi tiến vào mô tả các kết quả đã đạt được, chúng tôi tóm tắt ngắn gọn về đồ thị từ điển. Chúng tôi sử dụng Từ điển Tiếng Anh trực tuyến văn bản thuần túy [2], dựa trên "Dự án Gutenberg Etext của Từ điển không viết tắt Webster" dựa trên Từ điển không viết tắt Webster của Mỹ năm 1913. Từ điển bao gồm 27 tệp HTML (một cho mỗi chữ cái trong bảng chữ cái và một cho một số từ bổ sung). Những tệp này được cung cấp miễn phí từ trang web <http://www.gutenberg.net/>. Đồ thị kết quả có 112.169 đỉnh và 1.398.424 cạnh. Nó có thể được tải về từ trang web <http://www.eleves.ens.fr/home/senellar/>.

Để đánh giá chất lượng của phương pháp trích xuất từ đồng nghĩa của chúng tôi, chúng tôi đã so sánh kết quả được tạo ra với ba danh sách từ đồng nghĩa khác. Hai danh sách này (Distance và ArcRank) được tự động biên soạn bởi hai phương pháp trích xuất từ đồng nghĩa khác (xem [5] để biết chi tiết; phương pháp ArcRank được mô tả trong [16]), và một trong số chúng liệt kê các từ đồng nghĩa được thu thập từ nguồn tài nguyên WordNet do con người tạo ra, miễn phí trên WWW, [1]. Thứ tự xuất hiện của các từ trong nguồn cuối cùng này là tùy ý, trong khi nó được xác định rõ ràng cho ba phương pháp khác. Chúng tôi không bao gồm từ truy vấn trong danh sách từ đồng nghĩa, vì điều này không có nhiều ý nghĩa ngoại trừ phương pháp của chúng tôi, trong đó thú vị để nhận thấy là trong mọi ví dụ chúng tôi đã thử nghiệm, từ gốc xuất hiện đầu tiên trong danh sách; một điểm mà có xu hướng tạo điểm cho phương pháp của chúng tôi. Chúng tôi đã xem xét mười kết quả đầu tiên thu được cho bốn từ truy vấn được chọn vì tính đa dạng của chúng:

1. disappear: một từ có nhiều từ đồng nghĩa như vanish.

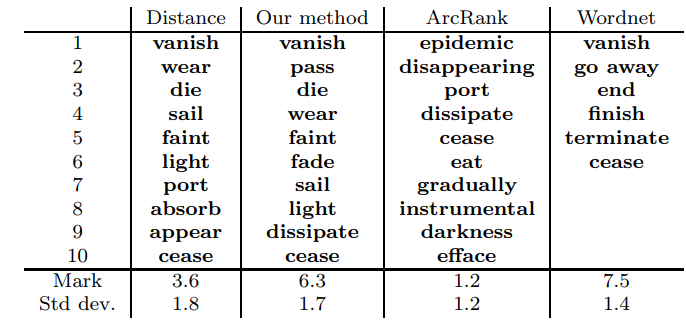
2. parallelogram: một từ rất cụ thể không có từ đồng nghĩa thực sự nhưng có một số từ tương tự: quadrilateral, square, rectangle, rhomb...

3. sugar: một từ thông dụng với các ý nghĩa khác nhau (trong hóa học, nấu ăn, dinh dưỡng...). Có thể mong đợi glucose là một ứng cử viên.

4. science: một từ thông thường và mơ hồ. Khó nói đoán từ đồng nghĩa là gì. Có thể knowledge là ứng cử viên tốt nhất.

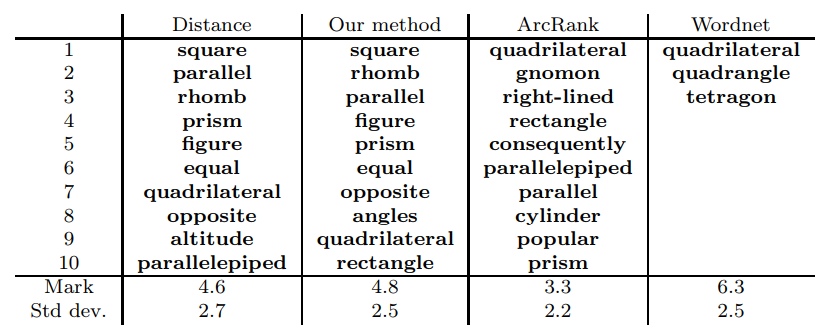
Để có một đánh giá khách quan về các phương pháp khác nhau, chúng tôi đã yêu cầu một mẫu gồm 21 người đánh giá (từ 0 đến 10) các danh sách từ đồng nghĩa, dựa trên mức liên quan của chúng đến từ đồng nghĩa. Danh sách được trình bày theo thứ tự ngẫu nhiên đối với mỗi từ. Kết quả thu được được cho trong Bảng 7.1, 7.2, 7.3 và 7.4. Hai dòng cuối cùng của mỗi bảng này cho biết điểm trung bình và độ lệch chuẩn.

Đối với từ **disappear**, phương pháp khoảng cách và phương pháp của chúng tôi đều làm tốt; **vanish, cease, fade, die, pass, dissipate, faint** là rất liên quan (không nên quên rằng động từ luôn xuất hiện mà không có giới từ đi kèm); **dissipate** hoặc faint cũng có ý nghĩa. Một số từ như **light** hoặc **port** hoàn toàn không liên quan, nhưng chúng chỉ xuất hiện ở vị trí thứ 6, thứ 7 hoặc thứ 8. Nếu chúng ta so sánh hai phương pháp này, chúng ta nhận thấy rằng phương pháp của chúng tôi tốt hơn: một từ đồng nghĩa quan trọng như **pass** được xếp hạng tốt, trong khi **port** hoặc **appear** không nằm trong danh sách 10 từ hàng đầu. Việc giải thích hiện tượng này là khó khăn.



*Bảng 7.1*

*Các từ đồng nghĩa được đề xuất cho từ disappear*



*Bảng 7.2*

*Các từ đồng nghĩa được đề xuất cho từ parallelogram*

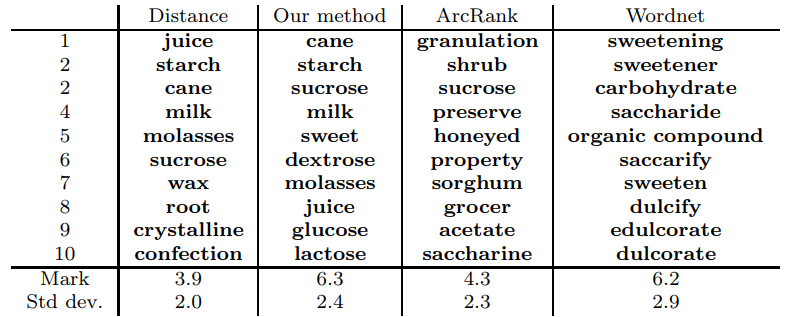
Nhưng chúng ta có thể nói rằng khía cạnh lẫn nhau của phương pháp của chúng tôi có hiệu quả tích cực. Trái ngược với điều này, ArcRank cho kết quả khá kém với các từ như eat, instrumental hoặc epidemic không liên quan.

Bởi vì đồ thị hàng xóm của parallelogram khá nhỏ (30 đỉnh), hai thuật toán đầu tiên cho kết quả tương tự nhau, và có ý nghĩa: square, rhomb, quadrilateral, rectangle, figure khá thú vị. Các từ khác ít liên quan hơn nhưng vẫn thuộc lĩnh vực ngữ nghĩa của parallelogram. ArcRank cũng hoạt động trên cùng một đồ thị con nhưng không cho kết quả chất lượng như vậy: consequently và popular rõ ràng không liên quan, nhưng gnomon là một bổ sung thú vị. Thú vị để nhận thấy rằng Wordnet ở đây ít phong phú hơn vì nó tập trung vào một khía cạnh cụ thể (quadrilateral).

Một lần nữa, kết quả được đưa ra bởi ArcRank cho từ sugar chủ yếu là không liên quan (property, grocer, ...). Phương pháp của chúng tôi lại tốt hơn phương pháp khoảng cách: starch, sucrose, sweet, dextrose, glucose, lactose là các từ có liên quan cao, ngay cả khi từ đồng nghĩa gần nhất được đưa ra (cane) không tốt bằng. Lưu ý rằng phương pháp của chúng tôi có các đánh giá thậm chí còn tốt hơn so với Wordnet.

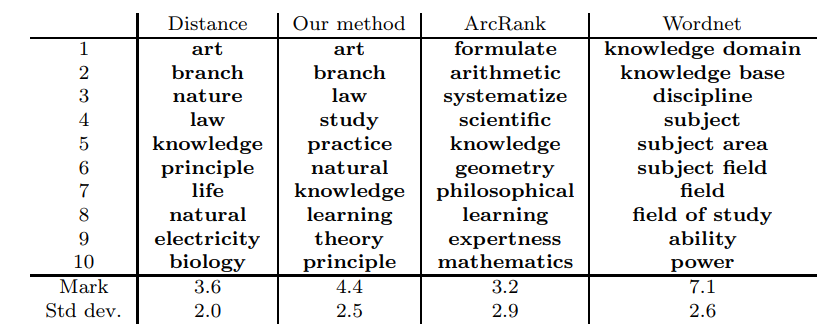
Kết quả cho từ science có thể là khó nhất để phân tích. Phương pháp khoảng cách và phương pháp của chúng tôi là tương đương. ArcRank có thể cho kết quả tốt hơn so với các từ khác nhưng vẫn kém hơn hai phương pháp khác.

Như một kết luận, hai thuật toán đầu tiên cho ra các từ thú vị và có liên quan, trong khi rõ ràng ArcRank không phù hợp cho việc tìm kiếm từ đồng nghĩa. Việc sử dụng của



*Bảng 7.3*

*Các từ đồng nghĩa được đề xuất cho từ sugar*



*Bảng 7.4*

*Các từ đồng nghĩa được đề xuất cho từ science*

điểm số trung tâm và mối quan hệ lẫn nhau củng cố của nó chứng tỏ sự ưu việt so với phương pháp khoảng cách cơ bản, ngay cả khi sự khác biệt không rõ ràng đối với tất cả các từ. Chất lượng kết quả thu được bằng các phương pháp khác nhau này vẫn khá khác biệt so với từ điển thủ công như Wordnet. Tuy nhiên, các kỹ thuật tự động này cho thấy sự quan tâm của chúng, vì chúng trình bày các khía cạnh hoàn thiện hơn của một từ so với từ điển thủ công. Chúng có thể được sử dụng một cách có lợi để mở rộng một chủ đề (xem ví dụ về parallelogram) và giúp việc biên soạn từ điển từ đồng nghĩa.

1. **Nhận xét cuối cùng:** Trong bài báo này, chúng tôi giới thiệu một khái niệm mới về ma trận tương đồng và giải thích cách liên kết một điểm số với sự tương đồng của các đỉnh của hai đồ thị. Chúng tôi chỉ ra cách tính điểm số này và chỉ ra cách nó mở rộng khái niệm điểm tâm và điểm uy lực được giới thiệu bởi Kleinberg. Chúng tôi chứng minh một số tính chất và minh họa sức mạnh và điểm yếu của khái niệm mới này. Các nghiên cứu về tính chất và ứng dụng của ma trận tương đồng của các đồ thị có thể được tiếp tục theo nhiều hướng khác nhau. Chúng tôi chỉ ra một số hướng nghiên cứu có thể.

Một sự mở rộng tự nhiên của khái niệm của chúng tôi là xem xét các mạng lưới thay vì các đồ thị; điều này tương đương với việc xem xét các ma trận kề với các phần tử thực tùy ý, không chỉ là số nguyên. Các định nghĩa và kết quả được trình bày trong bài báo này chỉ sử dụng tính chất rằng các ma trận kề có các phần tử không âm, vì vậy tất cả các kết quả vẫn còn hiệu lực đối với các mạng lưới có trọng số không âm. Sự mở rộng này đưa ra khả năng phân tích nhạy cảm: Ma trận tương đồng nhạy cảm đến trọng số trong mạng lưới như thế nào? Các thí nghiệm và lập luận chất lượng cho thấy, đối với hầu hết các mạng lưới, điểm tương đồng gần như liên tục trên hầu hết các phần tử của mạng lưới. Có thể phân tích điều này cho các mô hình đồ thị ngẫu nhiên như các mô hình xuất hiện trong [7] được không? Các câu hỏi này có thể có liên quan đến các tài liệu lớn về các giá trị riêng và không gian không đổi của đồ thị; xem, ví dụ, [8], [9] và [10].

Có vẻ như điều tự nhiên là nghiên cứu việc sử dụng ma trận tương đồng của hai đồ thị để phát hiện xem chúng có cùng cấu trúc hay không (Vấn đề đồng cấu đồ thị thuộc lớp phức tạp P hoặc NP-complete vẫn chưa được giải quyết). Nếu hai đồ thị đồng cấu, thì ma trận tương đồng của chúng có thể được làm đối xứng thông qua hoán vị cột (hoặc hàng). Việc kiểm tra xem một hoán vị như vậy có thể xảy ra trong thời gian đa thức và có duy nhất hay không (khi tất cả các phần tử của ma trận tương đồng đều khác nhau, nó chỉ có thể duy nhất). Trong trường hợp không tồn tại hoán vị như vậy hoặc khi chỉ có một hoán vị duy nhất có thể xảy ra, ta có thể kết luận ngay bằng cách trả lời phủ định hoặc kiểm tra hoán vị được đề xuất. Trong trường hợp nhiều hoán vị làm cho ma trận tương đồng đối xứng, tất cả các hoán vị đó phải được kiểm tra và điều này dẫn đến một số lượng hoán vị có thể lên đến cấp số mũ để xác minh. Có vẻ thú vị để xem cách heuristics này so sánh với các heuristics khác cho đồng cấu đồ thị và nghiên cứu xem các đặc điểm khác của ma trận tương đồng có thể được sử dụng để giới hạn số lượng hoán vị cần xem xét.

Cũng có những câu hỏi cụ thể hơn về ma trận tương đồng. Một vấn đề chưa được giải quyết là xác định các cặp ma trận tạo thành một ma trận tương đồng có hạng một. Cấu trúc của những cặp này được suy đoán ở cuối phần 6. Suy đoán này có đúng không? Một vấn đề đồ thị lâu đời cũng xuất hiện khi cố gắng xác định các đồ thị mà ma trận tương đồng của chúng chỉ có các phần tử dương. Các phần tử dương của ma trận tương đồng giữa các đồ thị GA và GB có thể được thu được như sau. Đầu tiên, xây dựng đồ thị tích, làm đối xứng nó, và sau đó xác định trong đồ thị kết quả các thành phần liên thông có giá trị Perron lớn nhất có thể. Chỉ số của các đỉnh trong đồ thị đó tương ứng chính xác với các phần tử khác không trong ma trận tương đồng của GA và GB. Các phần tử của ma trận tương đồng sẽ là dương nếu và chỉ nếu đồ thị tích đã được làm đối xứng là liên thông; tức là, nếu và chỉ nếu đồ thị tích của GA và GB là mạnh liên thông. Vấn đề xác định tất cả các cặp đồ thị có tích yếu liên thông đã được giới thiệu và phân tích vào năm 1966 trong [11]. Tài liệu đó cung cấp các điều kiện đủ để tích yếu liên thông. Mặc dù có nhiềuđóng góp sau đó về vấn đề này (xem ví dụ [12]), vấn đề xác định tất cả các cặp đồ thị có tích yếu liên thông một cách hiệu quả là một vấn đề mà, theo hiểu biết của chúng tôi, vẫn chưa được giải quyết.

Một chủ đề quan trọng khác là nghiên cứu cách các khái niệm được đề xuất ở đây có thể được sử dụng, có thể theo hình thức được sửa đổi, để đánh giá sự tương đồng giữa hai đồ thị, để gom cụm các đỉnh hoặc đồ thị, để nhận dạng mẫu trong đồ thị và cho mục đích khai thác dữ liệu.

**Cảm ơn.** Chúng tôi xin gửi lời cảm ơn đến Bert Coessens của KULeuven, Bỉ đã chỉ ra một sai sót trong phiên bản ban đầu của Theo-đề 8. Chúng tôi cũng xin cảm ơn các nhà đánh giá đã đọc bài báo một cách cẩn thận và giúp cải thiện chất lượng của bài báo.