TÌM KIẾM CỘNG ĐỒNG MẠNG DỰA TRÊN TÍNH GÓC GIỮA HAI VÉC TƠ

Lại Văn Trung*, Nguyễn Thị Thanh Giang

Trường Đại học Công nghệ Thông tin và Truyền thông, Đại học Thái Nguyên, Việt Nam

ARTICLE INFORMATION

Journal: *Vinh University Journal of Science*
p-ISSN: *3030-4563*

Volume: 53
Issue: 1A
****Correspondence:*** *lvtrung@ictu.edu.vn*

Received: *08 December 2023*
Accepted: *18 January 2024*
Published: *20 March 2024*

Citation:
Lại Văn Trung, Nguyễn Thị Thanh Giang (2024). Tìm kiếm cộng đồng mạng dựa trên tính góc giữa hai véc tơ.
Vinh Uni. J. Sci.
Vol. 53 (1A), pp. 95-102
doi: 10.56824/vujs.2023a162

TÓM TẮT

Gần đây, nghiên cứu về bài toán tìm kiếm cộng đồng mạng đã thu hút sự quan tâm của nhiều nhà khoa học. Hầu hết các loại mạng như mạng máy tính, mạng sinh học và mạng xã hội đều có cấu trúc cộng đồng. Tìm kiếm cộng đồng mạng giúp hiểu rõ cấu trúc và tính chất của mạng thực đó. Đã có nhiều thuật toán với các hướng tiếp cận khác nhau, bao gồm cả tọa độ hóa các đỉnh và xây dựng khoảng cách phù hợp giữa chúng. Trong nghiên cứu này sử dụng bước đi ngẫu nhiên để tọa độ hóa các đỉnh của đồ thị và sử dụng cosin của góc giữa hai véc tơ để phát hiện cộng đồng mạng. Bài viết cũng trình bày hàm Modularity để đánh giá cho việc phân cụm đồ thị. Một số kết quả thực nghiệm trên đồ thị sinh ngẫu nhiên và đồ thị được sinh ra từ bộ dữ liệu thực Zachary's karate club network được trình bày và so sánh với thuật toán K-means++.

Từ khóa: Cộng đồng mạng; bước đi ngẫu nhiên; tọa độ; thuật toán Cosin; đồ thị vô hướng; Modularity.

1. Giới thiệu

Vấn đề nghiên cứu về cấu trúc của các mạng phức tạp như mạng xã hội, mạng sinh học, mạng truyền thông là vấn đề thời sự được nhiều nhà khoa học nghiên cứu trong những năm gần đây [1-6]. Trong đó phát hiện cấu trúc cộng đồng mạng là một trong những lĩnh vực quan trọng được rất nhiều nhà khoa học quan tâm. Nói một cách đơn giản, cộng đồng mạng là những nhóm đỉnh mà mật độ kết nối giữa các đỉnh trong nhóm đó dày đặc hơn so với mật độ kết nối giữa các đỉnh bên ngoài các nhóm đó. Phát hiện cộng đồng mạng là quá trình phân nhóm các đỉnh trong mạng thành các cộng đồng. Các cộng đồng tồn tại trong tất cả các loại mạng thực và đóng một vai trò quan trọng trong cấu trúc, động lực và sự phát triển cơ bản của mạng. Đã có rất nhiều thuật toán được đưa ra để giải quyết bài toán tìm kiếm cộng đồng mạng với nhiều hướng tiếp cận khác nhau. Một trong những hướng tiếp cận phổ biến đó là sử dụng khoảng cách dựa trên ý tưởng hai đỉnh cùng một cộng đồng thì có khoảng cách nhỏ, như thuật toán K-means [7] và một cải tiến của nó là thuật toán K-mean++ [8].

Thuật toán K-means [7] là một thuật toán đơn giản phân chia một tập dữ liệu gồm n đối tượng thành k cộng đồng với k là một số cho trước. Ý tưởng của thuật toán dựa trên việc cực tiểu hóa hàm mất mát

$$L = \sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} \left\| x_i - m_i \right\|^2,$$

trong đó $y_{ij} = \begin{cases} 1 \; khi \; i = j, \\ 0 \; khi \; i \neq j, \end{cases}$ $x_i = \left(x_{i_1}, x_{i_2}, ..., x_{i_n} \right)$ là điểm dữ liệu thứ i và $m_j = \frac{1}{n}\sum_{i=1}^{n} x_i$, $j = 1, 2, ..., k$. Thuật toán K-means với đầu vào là tập dữ liệu X và số cụm k, đầu ra là k cụm dữ liệu đã được phân chia. Các bước của thuật toán K-means:

- Khởi tạo k điểm dữ liệu trong tập X làm tâm.
- Lặp lại các bước sau đến khi hội tụ

Bước 1: Với mỗi điểm dữ liệu, tính khoảng cách của nó đến các tâm và gán nó vào cụm mà khoảng cách của nó đến tâm cụm là gần nhất.

Bước 2: Với mỗi cụm, xác định lại tâm của cụm bằng cách lấy trung bình cộng của các điểm dữ liệu trong cụm đó.

Thuật toán sẽ dừng khi việc chọn tâm của bước sau không thay đổi hoặc thay đổi không đáng kể so với các tâm được chọn vòng lặp trước đó.

Thuật toán K-means là một trong những thuật toán rất quan trọng, được sử dụng rộng rãi, thường tìm ra lời giải nhanh chóng và hợp lý trong việc phân cụng dữ liệu không gán nhãn. Tuy nhiên thuật toán Kmeans còn có một số nhược điểm trong đó có một nhược điểm lớn đó là phép tính gần đúng được tìm thấy có thể sai tùy ý đối với hàm mục tiêu so với phân cụm tối ưu. Điều này xảy ra do việc khởi tạo chọn ngẫu nhiên ban đầu không tốt. Nhược điểm này sẽ được khắc phục bằng một cải tiến của thuật toán K-means++.

1.2 Thuật toán K-means++

Ý tưởng của thuật toán K-means++ [8] là việc trải rộng k trung tâm cụm ban đầu: trung tâm cụm đầu tiên được chọn ngẫu nhiên một cách thống nhất từ các điểm dữ liệu đang được phân cụm, sau đó mỗi trung tâm cụm tiếp theo được chọn từ các điểm dữ liệu còn lại với xác suất tỷ lệ thuận với bình phương khoảng cách của nó đến tâm cụm gần nhất đã được chọn.

Các bước của thuật toán K-means++:

- ***Bước 1:*** Chọn ngẫu nhiên một trung tâm thống nhất trong số các điểm dữ liệu.
- ***Bước 2:*** Với mỗi điểm dữ liệu x chưa được chọn, hãy tính $d(x,C)$ là khoảng các giữa x và tâm gần nhất đã được chọn, tức là $d(x,C) = \min_{c \in C} d(x,c)$, với C là tập các tâm đã chọn.
- ***Bước 3:*** Chọn ngẫu nhiên một điểm dữ liệu mới làm tâm, sử dụng phân bố xác suất có trọng số trong đó điểm x được chọn với xác suất $p_x = \frac{d^2(x,C)}{\sum_{x \in X} d^2(x,C)}$.
- ***Bước 4:*** Lặp lại bước 2 và 3 cho đến khi k tâm được chọn.

• ***Bước 5:*** Khi đã chọn được *k* tâm ban đầu, sử dụng thuật toán K-means để phân cụm.

Thuật toán K-means++ cải thiện đáng kể cho thuật toán K-means gốc. Mặc dù việc lựa chọn ban đầu trong thuật toán mất thêm thời gian, phần K-means tự nó hội tụ rất nhanh sau lần chọn tâm này và do đó thuật toán thực sự làm giảm thời gian tính toán. Các tác giả trong [8] đã thử nghiệm phương pháp của họ với các bộ dữ liệu thực và tổng hợp thì nhận được sự cải thiện gấp 2 lần về tốc độ và đối với một số bộ dữ liệu nhất định, lỗi cải thiện gần 1000 lần. Trong các mô phỏng này, phương pháp mới hầu như luôn thực hiện tốt, cải thiện được đáng kể cả về tốc độ và sai số.

Để sử dụng được thuật toán K-means++ trong phân cụm cộng đồng mạng trên đồ thị, ta thường tìm cách tọa độ hóa các đỉnh rồi phân cụm theo thuật toán K-means++. Trong bài báo này, bước đi ngẫu nhiên được sử dụng để tọa độ hóa các đỉnh của đồ thị.

1.3 Bước đi ngẫu nhiên trên đồ thị

Xét đồ thị vô hướng $G=(V,E)$, trong đó $V=\{v_1,v_2,...,v_n\}$ là tập các đỉnh và E là tập các cạnh của đồ thị. Bước đi ngẫu nhiên trên đồ thị là một quá trình ngẫu nhiên mô tả một đường đi bao gồm một chuỗi các bước ngẫu nhiên trên đồ thị. Mỗi bước đi là việc di chuyển từ một đỉnh nào đó đến các đỉnh kề với nó một cách ngẫu nhiên với xác suất như nhau. Xác suất di chuyển từ đỉnh v_i sang đỉnh v_j là

$$p_{ij}=\frac{A_{ij}}{\deg(v_i)},(i,j=1,2,...,n),$$

trong đó A_{ij} là phần tử nằm trên giao của hàng i và cột j của ma trận kề A và $\deg(v_i)$ là bậc của đỉnh v_i. Ma trận $P=(P_{ij})_{n\times n}$ được gọi là ma trận chuyển của quá trình bước đi ngẫu nhiên trên đồ thị G sau một bước. Xác suất của việc di chuyển từ đỉnh i đến đỉnh j sau t bước được xác định là P_{ij}^t, với P_{ij}^t là phần tử nằm ở hàng i và cột j của ma trận P^t.

Trong [9], các tác giả đã chỉ ra rằng nếu hai đỉnh u và v thuộc cùng một cộng đồng các xác suất để đi ngẫu nhiên từ đỉnh u và đỉnh v đến đỉnh tuỳ ý w là gần như nhau, tức là $P_{uw}^t \simeq P_{vw}^t$, với mọi đỉnh w. Do đó chúng tôi nhận thấy khi hai đỉnh u và v cùng một cộng đồng thì $\left\| D^{-\frac{1}{2}}P_{u\bullet}^t - D^{-\frac{1}{2}}P_{v\bullet}^t \right\|$ sẽ nhỏ, với $\|\bullet\|$ là chuẩn Euclidean trong $\mathbb{R}^n$, ma trận D là ma trận bậc của đồ thị G và $P_{u\bullet}^t$ là hàng tương ứng với đỉnh u của ma trận P^t. Từ đó, chúng tôi đề xuất việc tọa độ hóa các đỉnh của đồ thị như sau:

Với mỗi đỉnh u của đồ thị G tương ứng với một véc tơ $D^{-\frac{1}{2}}P_{u\bullet}^t$, hay nói cách khác tọa độ của đỉnh u được xác định bởi:

$$Coord(u)=\left(\frac{P_{u_1}^t}{\sqrt{d_1}},\frac{P_{u_2}^t}{\sqrt{d_2}},...,\frac{P_{u_n}^t}{\sqrt{d_n}}\right), \quad (1)$$

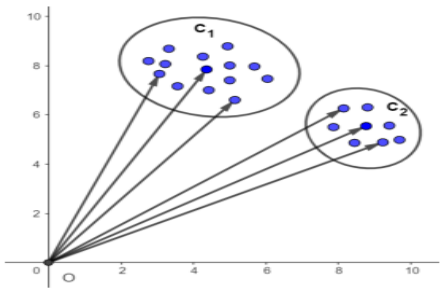
trong đó d_i là bậc của đỉnh v_i, với $i=1,2,...,n$.

Theo cách tiếp cận tọa độ hóa các đỉnh của đồ thị, đã có nhiều phương pháp được đưa ra, trong đó phải kể đến phương pháp tọa độ hóa theo véc tơ riêng của Shi và Malid hay Ng, Jordan và Weiss được trình bày trong [10]. Tuy nhiên với các mạng thực lớn, các phương pháp này thường gặp phải khó khăn trong việc tìm véc tơ riêng của ma trận Laplace chuẩn hóa. Trong bài báo này, chúng tôi sử dụng bước đi ngẫu nhiên để tọa độ hóa các đỉnh thông qua việc tính ma trận xác suất chuyển P^t, bằng cách này có thể giải quyết được nhanh chóng và chính xác với các mạng thực lớn.

Từ việc tọa độ hóa các đỉnh theo công thức (1) và dựa vào Cosin của góc giữa hai véc tơ chúng tôi đưa ra thuật toán tìm kiếm cộng đồng mạng theo thuật toán Cosin được trình bày trong phần sau.

2. Thuật toán Cosin tìm kiếm cộng đồng mạng

Dựa trên nhận đinh rằng, nếu hai đỉnh u và v cùng thuộc một cộng đồng thì góc giữa hai véc tơ $Coord(u)$ và $Coord(\text{v})$ sẽ khá nhỏ, hay $\cos in\left(Coord(u), Coord(\text{v})\right) \simeq 1$.



Hình 1: *Cộng đồng mạng và các đỉnh trong cộng đồng*

Dựa vào ý tưởng trên, chúng tôi đưa ra thuật toán Cosin tìm kiếm cộng đồng mạng với các bước như sau:

Dữ liệu đầu vào: Đồ thị vô hướng $G = (V, E)$ và số cụm k.

Đầu ra: k cụm cộng đồng đã được phân chia.

Bước 1: Tọa độ hóa các đỉnh của đồ thị $Coord(u) = \left(\frac{P^t_{u_1}}{\sqrt{d_1}}, \frac{P^t_{u_2}}{\sqrt{d_2}}, ..., \frac{P^t_{u_n}}{\sqrt{d_n}} \right)$.

Bước 2: Chọn k đỉnh làm tâm và mỗi tâm coi như một cộng đồng $C_1, C_2, ..., C_k$, sau đó lặp lại các bước sau đến khi hội tụ:

• Với mỗi đỉnh u, tính $C = \cos in(Coord(u), Center(C_i))$, (trong đó $Center(C_i)$ là tâm của cụm thứ i), nếu $C \geq \theta$ thì ta thêm đỉnh u vào cụm C_i.

• Với mỗi cụm, xác định lại tâm của cụm bằng cách lấy trung bình cộng của các điểm dữ liệu trong cụm đó.

Thuật toán sẽ dừng khi tâm được chọn ở bước lặp sau gần như không thay đổi so với bước lặp đó. Tham số θ thường được chọn trong ngưỡng $0.6 \leq \theta < 1$.

Để đánh giá được hiệu quả của thuật toán Cosin đã đề xuất, chúng tôi thực hiện thuật toán trên bộ dữ liệu thực tế Zachary's karate club network và dữ liệu là đồ thị sinh ngẫu nhiên rồi sử dụng hàm Modularity để so sánh với thuật toán K-means++. Kết quả của thuật toán sẽ được trình bày trong phần tiếp theo.

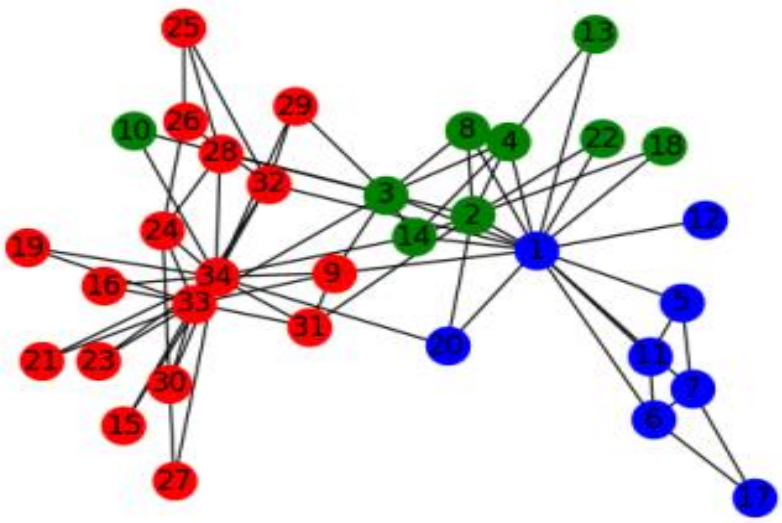
3. Kết quả thực nghiệm

Hàm Modularity là một đại lượng rất quan trọng trong bài toán tìm kiếm cộng đồng mạng. Có nhiều thuật toán tìm kiếm cộng đồng mạng nổi tiếng dựa vào hàm Modularity như Phương pháp phổ của Newman [11] hay thuật toán Louvain [12]. Ngoài ra hàm Modularity còn là thước đo để đánh giá chất lượng phân cụm của các thuật toán tìm kiếm cộng đồng mạng. Thuật toán phân cụm nào có giá trị modularity lớn hơn thì chất lượng phân cụm của thuật toán đó sẽ tốt hơn. Hàm Modularity trên đồ thị vô hướng được định nghĩa trong [13] như sau:

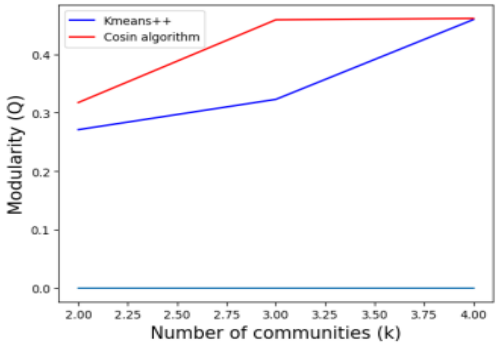
$$Q = \frac{1}{2m}\sum_{i=1,\, j=1}^{n}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta\left(C_i, C_j\right),$$

trong đó A_{ij} là phần tử nằm trên giao của hàng i cột j của ma trận kề A; k_i, k_j là bậc của đỉnh i và đỉnh j; C_i, C_j là các cộng đồng chứa đỉnh i và j tương ứng; $\delta\left(C_i, C_j\right) = \begin{cases} 1 \ khi \ C_i = C_j, \\ 0 \ khi \ C_i \neq C_j. \end{cases}$

Một trong các bộ dữ liệu thực tế phổ biến và quan trọng để thử nghiệm trong các thuật toán tìm kiếm cộng đồng mạng là bộ dữ liệu thực Zachary's karate club network. Bộ dữ liệu này được giới thiệu bởi Wayne W. Zachary [14]. Trong bài báo này, chúng tôi sử dụng thuật toán Cosin để tìm kiếm cộng đồng trên bộ dữ liệu thực Zachary's karate club network và thu được kết quả rất tốt, thể hiện ở Hình 2-3. Cụ thể, Hình 2 cho thấy, khi thử nghiệm thuật toán Cosin trên bộ dữ liệu Zachary's karate club network các đỉnh được phân thành ba cụm rất rõ ràng. Hình 3 cho thấy giá trị Modularity thu được từ thuật toán Cosin lớn hơn so với thuật toán Kmeans++. Điều đó chứng tỏ rằng chất lượng phân cụm của thuật toán Cosin mà chúng tôi đề xuất là tốt hơn.

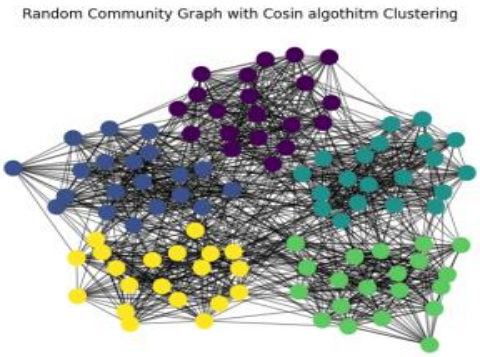


Hình 2: *Tìm kiếm cộng đồng mạng bằng thuật toán Cosin trên bộ dữ liệu Zachary's karate club network*

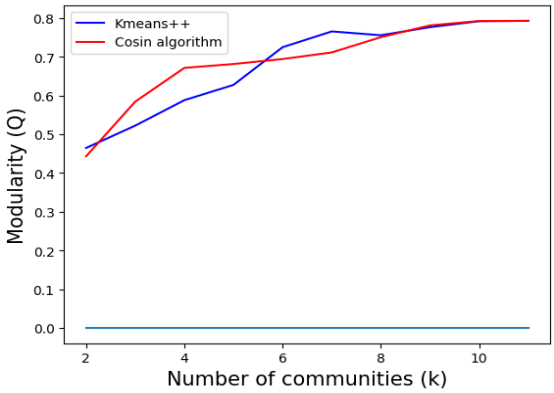


Hình 3: *So sánh Modularity giữa thuật toán Cosin và thuật toán Kmeans++ trên bộ dữ liệu Zachary's karate club network*

Ngoài ra, thuật toán Cosin còn được sử dụng để tìm kiếm cộng đồng trên đồ thị sinh ngẫu nhiên và cũng thu được kết quả rất khả quan, thể hiện ở Hình 4-5. Cụ thể Hình 4 cho thấy, khi thử nghiệm thuật toán Cosin trên đồ thị sinh ngẫu nhiên các đỉnh được phân thành năm cụm rất rõ ràng. Hình 5, cho thấy giá trị Modularity thu được từ thuật toán Cosin đa phần là cao hơn so với thuật toán Kmeans++. Điều đó chứng tỏ rằng chất lượng phân cụm của thuật toán Cosin mà chúng tôi đề xuất là tốt hơn trên bộ dữ liệu này.



Hình 4: *Tìm kiếm cộng đồng mạng bằng thuật toán Cosin trên trên đồ thị sinh ngẫu nhiên*



Hình 5: *So sánh Modularity giữa thuật toán Cosin và thuật toán Kmeans++ trên đồ thị sinh ngẫu nhiên*

Từ kết quả thực nghiệm trên cho thấy, thuật toán Cosin đề xuất có thể giải quyết tốt bài toán tìm kiếm cộng đồng mạng trên đồ thị vô hướng trên cả các dữ liệu thực và dữ liệu là đồ thị sinh ngẫu nhiên.

4. Kết luận

Bài báo đã trình bày việc sử dụng bước đi ngẫu nhiên trên đồ thị để tọa độ hóa các đỉnh, từ đó đề xuất thuật toán Cosin để phân cụm cộng đồng mạng. Thuật toán Cosin được đề xuất đã giải quyết tốt bài toán tìm kiếm cộng đồng mạng trên đồ thị vô hướng. Cụ thể, kết quả thực nghiệm cho thấy thuật toán Cosin tìm kiếm cộng đồng mạng có giá trị hàm modularity cao hơn thuật toán Kmeans++, do đó chất lượng phân cụm tốt hơn. Đây cũng là một kết quả quan trọng để tiếp tục nghiên cứu việc tìm kiếm cộng đồng mạng trên đồ thị có hướng và tìm kiếm cộng đồng mạng chồng chéo.

TÀI LIỆU THAM KHẢO

[1] M. E. J. Newman, "Assortative mixing in networks," *Phys. Rev. Lett*, 89, 8701, 2002. DOI: 10.1103/PhysRevLett.89.208701

[2] Newman M. E., "The structure and function of complex networks," *SIAM Rev, 45*, 167-256, 2003. DOI: 10.1137/S003614450342480

[3] Strogatz S. H, "Exploring complex networks," *Nature*, *410*, 268, 2001. DOI: 10.1038/35065725

[4] Jackson M. O., *Social and Economic Networks*, Princeton University Press: Princeton, 2010.

[5] Hossein Hajibabaei, Vahid Seydi and Abbas Koochari, "Community detection in weighted networks using probabilistic generative model," *Journal of Intelligent Information Systems*, vol. 60, pp. 119-136 , 2023. DOI: 10.1007/s10844-022-00740-6

[6] J. Zhu, C. Wang, C. Gao, F. Zhang, Zh. Wang and X. Li, "Community Detection in Graph: An Embedding Method," *IEEE Transactions on Network Science and Engineering*, vol. 9, pp. 689-702, 2022. DOI: 10.1109/TNSE.2021.3130321

[7] B. MacQueen, "Some Methods for classification and Analysis of Multi- variate Observations," *Proceedings of 5th Berkeley Symposium on Math- ematical Statistics and Probability*, University of California Press., vol. 1, pp. 281-297 ,1967.

[8] Arthur D. and Vassilvitskii, "K-Means++: The Advantages of Careful Seeding," *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, Philadelphia, PA, pp. 1027-1035, 2007.

[9] Luxburg V. U, "A tutorial on spectral clustering," *Stat Comput*, 17(4), pp. 395-416, 2007. DOI: 10.1007/s11222-007-9033-z

[10] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *Journal of Graph Algorithms and Applications*, vol. 10, no. 2, pp 191-218, 2006. DOI: 10.7155/jgaa.00124

[11] M. E. J. Newman, "Spectral methods for network community detection and graph partitioning," *Phys. Rev. E*, vol. 88, p. 042822, 2013. DOI: 10.1103/PhysRevE.88.042822

[12] D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre, "Fast unfold-ing of communities in large networks," *Journal of Statistical Mechanics, Theory and Experiment*, 10, P10008, 2008. DOI: 10.1088/1742-5468/2008/10/P10008

[13] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences of the United States of America*, 103 (23), 8577-8696, 2006. DOI: 10.1073/pnas.0601602103

[14] Zachary W. W., "An Information Flow Model for Conflict and Fission in Small Groups," *Journal of Anthropological Research*. 33 (4): 452-473, 1977. DOI: 10.1086/jar.33.4.3629752

ABSTRACT

NETWORK COMMUNITY DETECTION BASED ON THE ANGLE BETWEEN TWO VECTORS

Lai Van Trung, Nguyen Thi Thanh Giang
University of Information and Communication Technology,
Thai Nguyen University, Vietnam
Received on 08/12/2023, accepted for publication on 18/01/2024

Recently, the problem of community detection has attracted the attention of many scientists. Most types of networks such as computer networks, biological networks and social networks, have a community structure. Community detection helps to understand the structure and properties of that real network. There have been many algorithms with different approaches, including coordinating vertices and building appropriate distances between them. In this paper, a random walk has been used to coordinate the vertices of the graph and use the cosine of the angle between two vectors to detect network communities. The article also presents the Modularity function to evaluate graph clustering. Some experimental results on randomly generated graphs and graphs generated from the real data set Zachary's karate club network have been presented and compared with the K-means++ algorithm.

Keywords: Community network; random walk; coordinates; cosine algorithm; undirected graph; modularity.