Tạp chí Khoa học Đại học Cần Thơ
Phần A: Khoa học tự nhiên, Công nghệ và Môi trường
ISSN 1859-2333 | e-ISSN 2815-5599

DOI:10.22144/ctujos.2025.035

PHÂN CỤM CÁC BÀI BÁO TOÁN HỌC THEO NHÓM DỰA TRÊN TỪ KHÓA BẰNG THUẬT TOÁN SVD VÀ THUẬT TOÁN K-MEANS

Phạm Bích Như*, Nguyễn Thị Tú Trinh, Lê Thị Huỳnh Như, Lưu Minh Thư và Huỳnh Lan Thanh
Khoa Khoa học Tự nhiên, Trường Đại học Cần Thơ, Việt Nam
**Tác giả liên hệ (Corresponding author): pbnhu@ctu.edu.vn*

Thông tin chung (Article Information)

Nhận bài (Received): 12/09/2024
Sửa bài (Revised): 01/10/2024
Duyệt đăng (Accepted): 12/12/2024

Title: *Clustering mathematical articles into groups based on keywords using the SVD algorithm and the K-means algorithm*

Author(s): *Pham Bich Nhu*, Nguyen Thi Tu Trinh, Le Thi Huynh Nhu, Luu Minh Thu and Huynh Lan Thanh*

Affiliation(s): *College of Natural Sciences, Can Tho University, Viet Nam*

TÓM TẮT

Bài báo này trình bày phương pháp phân cụm các bài báo toán học theo nhóm dựa trên từ khóa bằng cách sử dụng thuật toán SVD và K-means. Đầu tiên, các từ khóa được biểu diễn bằng TF-IDF, sau đó áp dụng SVD để giảm số chiều, giữ lại các đặc trưng quan trọng. Tiếp theo, thuật toán K-means được sử dụng để phân cụm các bài báo theo mức độ tương đồng của từ khóa, từ đó các tác giả có cùng chủ đề nghiên cứu được nhóm lại với nhau. Việc giảm chiều thông qua SVD giúp các thuật toán học máy như K-means hoạt động tốt hơn nhờ tập trung vào các yếu tố quan trọng nhất của dữ liệu.

Từ khóa: *Giá trị kì dị, phương pháp TF-IDF, thuật toán K-means, thuật toán SVD*

ABSTRACT

This paper presents a method for clustering mathematical articles into groups based on keywords using the SVD and K-means algorithms. First, the keywords are represented using TF-IDF, and then SVD is applied to reduce the dimensionality, retaining the important features. Next, the K-means algorithm is used to cluster the articles based on the similarity of keywords, thereby grouping authors with similar research topics. Dimensionality reduction through SVD helps machine learning algorithms like K-means perform better by focusing on the most important features of the data.

Keywords: *K-means algorithm, singular value, singular value decomposition algorithm (SVD), Term Frequency-Inverse Document Frequency method*

1. GIỚI THIỆU

Từ những năm đầu của thế kỷ 19, các nhà toán học như Carl Gustav, Jacob Jacobi, Eugenio Beltrami và Camille Jordan là những người tiên phong trong việc nghiên cứu bài toán phân rã ma trận. Đặc biệt, nhà toán học Jacobi là người đề xướng việc tính toán các giá trị riêng và véctơ riêng của ma trận đối xứng. Những nghiên cứu này đã đặt nền móng cho sự phát triển của thuật toán SVD (singular value decomposition). Năm 1936, Eckart và Gale Young (1936) đã mô tả phương pháp xấp xỉ một ma trận bằng một ma trận có hạng thấp hơn bằng cách sử dụng SVD. Đến những năm 60 của thế kỷ 20, các nhà toán học đã bắt đầu nhận ra giá trị của SVD trong các ứng dụng thực tế. Những tính toán SVD trong các công trình của Golub và Kahan (1965) đã đặt nền móng cho các thuật toán hiệu quả

để tính toán SVD bằng công cụ máy tính sau này. Giai đoạn đầu, SVD được sử dụng rộng rãi trong các lĩnh vực như giảm chiều dữ liệu, xử lý tín hiệu, phân tích dữ liệu, thống kê,... Đến những năm 90, SVD trở thành công cụ quan trọng trong khoa học máy tính, cụ thể như giải các bài toán tìm kiếm và truy hồi thông tin, phân loại văn bản và nhận dạng mẫu (Stewart, 1993). Đến cuối thế kỷ 20, SVD tiếp tục được phát triển và ứng dụng rộng rãi trong nhiều lĩnh vực, đặc biệt là trong lĩnh vực xử lý ngôn ngữ tự nhiên, nhận dạng ảnh (Turk & Pentland, 1991). Vào đầu thế kỷ 21, với sự phát triển mạnh mẽ của trí tuệ nhân tạo và học máy, SVD vẫn chứng tỏ mình là công cụ cơ bản trong nhiều thuật toán và mô hình. SVD được sử dụng trong việc giảm số chiều và tối ưu hóa ma trận (Koren et al., 2009; Kaloorazi, 2018). Sự phát triển liên tục của thuật toán và ứng dụng của SVD trong nhiều lĩnh vực khoa học và kỹ thuật đã biến nó trở thành một công cụ không thể thiếu trong phân tích và xử lý dữ liệu hiện đại. Bên cạnh thuật toán SVD truyền thống thì các biến thể của nó như Truncated SVD, Randomized SVD và Incremental SVD được sử dụng rộng rãi trong việc phân tích và xử lý dữ liệu lớn (big data), học máy và thị giác máy tính.

Cho A là một ma trận cấp $m \times n$, SVD biểu diễn ma trận A như sau:

$$A = U \sum V^T, \qquad (1)$$

trong đó, U là một ma trận trực giao cấp $m \times m$; $\sum$ là một ma trận đường chéo cấp $m \times n$ với các giá trị trên đường chéo là không âm và được sắp theo thứ tự giảm dần $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0 = 0 = \cdots = 0$, với σ_i bằng căn bậc hai dương của các giá trị riêng của ma trận $A^T A$, các giá trị này gọi là giá trị kỳ dị của A; V là một ma trận trực giao cấp $n \times n$ và V^T là ma trận chuyển vị của V.

Thuật toán SVD giúp giảm số chiều của dữ liệu mà vẫn giữ lại các đặc trưng quan trọng nhất, giúp xử lý dữ liệu lớn hiệu quả hơn và giảm độ phức tạp khi tính toán.

Bên cạnh đó, bài toán tối ưu rời rạc đóng một vai trò quan trọng trong nhiều lĩnh vực của đời sống và khoa học kỹ thuật (Stanimirovic, 2020). Một số ứng dụng thực tế có thể kể đến như bài toán phân loại văn bản, bài toán quản lý chuỗi cung ứng, bài toán vận tải, bài toán thiết kế và quản lý mạng, bài toán quản lý tài nguyên, bài toán lập lịch, bài toán phân tích cộng đồng (Ramponi et al, 2019); ứng dụng trong khoa học máy tính và trí tuệ nhân tạo, điện toán lượng tử (Zha et al., 2001; Miettinen et al., 2008). Thuật toán SVD đóng góp một cách trực tiếp cũng như gián tiếp trong việc giải các bài toán tối ưu rời rạc (Sarkar & Dong, 2011; Chicco & Masseroli, 2013; Li et al., 2021).

Như đã giới thiệu ở trên, thuật toán SVD có nhiều ứng dụng. Bài viết này trình bày một số kết quả của việc sử dụng hai thuật toán SVD và K-means để phân nhóm tác giả thuộc lĩnh vực toán học dựa trên các từ khóa trong các bài báo khoa học của họ đã được công bố trên Tạp chí Khoa học Đại học Cần Thơ từ năm 2020 đến năm 2024.

2. THUẬT TOÁN VÀ PHƯƠNG PHÁP NGHIÊN CỨU

Phân cụm là việc chia một tập hợp các đối tượng thành các nhóm sao cho các đối tượng trong nhóm có tính chất tương tự nhau. Bài toán này thường được ứng dụng trong phân tích dữ liệu, nhận dạng mẫu, học máy, ...

2.1. Phương pháp TF-IDF

TF-IDF (Term Frequency-Inverse Document Frequency) là một phương pháp phổ biến trong phân tích văn bản để chuyển đổi dữ liệu văn bản thành các đặc trưng số. Phương pháp TF-IDF là công cụ hữu ích cho các thuật toán học máy như phân cụm và phân loại.

Áp dụng TF-IDF cho danh sách các bài báo (từ khóa), văn bản đã được chuyển đổi thành ma trận, trong đó mỗi hàng tương ứng với một bài báo và mỗi cột tương ứng với một thuật ngữ. Mỗi ô trong ma trận chứa điểm TF-IDF cho một thuật ngữ cụ thể trong một bài báo. Biểu diễn số này sau đó có thể được sử dụng cho phân cụm, phân loại hoặc các loại phân tích khác.

Tần suất thuật ngữ (TF) đo lường tần suất xuất hiện của một từ (thuật ngữ) trong một tài liệu. Từ xuất hiện càng nhiều trong một tài liệu thì điểm số TF càng cao.

$$TF(t,d) = \frac{m}{n},$$

trong đó m là số lần xuất hiện thuật ngữ t trong tài liệu d và n là tổng số thuật ngữ xuất hiện trong tài liệu d.

Nghịch đảo tần suất tài liệu (IDF) giúp giảm tầm quan trọng của những từ thông dụng xuất hiện trong nhiều tài liệu và làm tăng trọng số của các thuật ngữ hiếm gặp trong toàn bộ tập tài liệu

$$IDF(t) = \log\left(\frac{s}{l}\right),$$

trong đó, s là tổng số tài liệu và l là số tài liệu chứa thuật ngữ t.

Điểm TF-IDF cuối cùng cho mỗi thuật ngữ trong một tài liệu là tích của tần suất thuật ngữ (TF) và nghịch đảo tần suất tài liệu (IDF)

$$\text{TF-IDF}(t,d) = TF(t,d).IDF(t).$$

2.2. Thuật toán SVD

2.2.1. Giới thiệu về SVD

Phân tích giá trị riêng (chéo hóa) là một công cụ mạnh mẽ được sử dụng trong nhiều lĩnh vực khác nhau từ toán học thuần túy đến các ứng dụng thực tiễn. Tuy nhiên, việc chéo hóa một ma trận không phải lúc nào cũng thực hiện được và chỉ áp dụng được cho ma trận vuông. Trong thực tế có rất nhiều bộ dữ liệu được biểu diễn dưới dạng các ma trận không vuông. Từ yêu cầu thực tiễn đó, cần một công cụ phân rã ma trận mạnh mẽ hơn và phân tích SVD đã ra đời. Yêu cầu đặt ra, làm thế nào để phân tích một ma trận A cấp $m \times n$ thành tích của ba ma trận có dạng (1).

Xét ma trận

$$A^TA = (U\sum V^T)^T(U\sum V^T) = V\sum^T U^T U\sum V^T \quad (2)$$

$$AA^T = (U\sum V^T)(U\sum V^T)^T = U\sum V^T V\sum^T U^T \quad (3)$$

Do U, V là hai ma trận trực giao nên $U^TU = I_m$ và $V^TV = I_n$, nghĩa là $U^T = U^{-1}$ và $V^T = V^{-1}$. Thay vào (2) và (3) ta được

$$A^TA = V\sum^T\sum V^{-1} \quad (4)$$

và

$$AA^T = U\sum\sum^T U^{-1} \quad (5)$$

Từ (4) suy ra V là ma trận chéo hóa được ma trận A^TA và $\sum^T\sum$ là dạng chéo của ma trận A^TA.

Tương tự, từ (5) suy ra U là ma trận chéo hóa được ma trận AA^T và $\sum\sum^T$ là dạng chéo của ma trận AA^T.

Khi đó, U, V là các ma trận mà các cột của chúng lần lượt là các véctơ riêng độc lập tuyến tính của ma trận AA^T và A^TA.

Hơn nữa, vì hai ma trận AA^T và A^TA là hai ma trận nửa xác định dương nên các giá trị riêng của chúng là không âm.

Định lý 2.1 Hai ma trận AA^T và A^TA có các giá trị riêng dương bằng nhau.

Chứng minh:

Giả sử ma trận A cấp $m \times n$. Gọi σ_i là các giá trị kỳ dị của ma trận A, λ và λ' lần lượt là giá trị riêng của ma trận AA^T và A^TA . Khi đó, ta có

Nếu $m > n$ thì

$$\sum = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

Nếu $m < n$ thì

$$\sum = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & 0 \\ 0 & 0 & \cdots & \sigma_n & \cdots & 0 \end{pmatrix}.$$

Đặt $r = \min(m,n)$. Khi đó, ta có

$$\det(AA^T - \lambda I_m) = \det(U\sum\sum^T U^{-1} - \lambda I_m)$$
$$= \det(U(\sum\sum^T - \lambda I_n)U^{-1}) = \det(\sum\sum^T - \lambda I_m)$$
$$= (\sigma_1^2 - \lambda)(\sigma_2^2 - \lambda)\cdots(\sigma_r^2 - \lambda)(-\lambda)^{m-r}.$$

Mặt khác,

$$\det(A^TA - \lambda' I_n) = \det(V\sum^T\sum V^{-1} - \lambda' I_n)$$
$$= \det(V(\sum^T\sum - \lambda' I_n)V^{-1}) = \det(\sum^T\sum - \lambda' I_n)$$
$$= (\sigma_1^2 - \lambda')(\sigma_2^2 - \lambda')\cdots(\sigma_r^2 - \lambda')(-\lambda')^{n-r}.$$

Từ kết quả trên ta suy ra điều phải chứng minh. ■

Nhận xét: Từ kết quả trên ta suy ra $\lambda = \lambda' = {\sigma_i}^2$.

Hệ quả 1: Giả sử A là một ma trận vuông đối xứng nửa xác định dương. Khi đó, các giá trị kì dị của A chính là các giá trị riêng của nó.

Chứng minh:

Vì A là một ma trận vuông đối xứng nên $A = A^T$.

Gọi λ là giá trị riêng của A và v là một véctơ riêng tương ứng với giá trị riêng λ và $\|v\| = 1$.

Vì A là ma trận nửa xác định dương nên suy ra $\lambda \geq 0$.

Ta có,

$$Av = \lambda v.$$

Xét $(A^TA)v = A^T(Av) = A\lambda v = \lambda(Av) = \lambda^2 v.$

Suy ra, λ^2 là giá trị riêng của ma trận A^TA, nghĩa là λ là giá trị kỳ dị của ma trận A. ■

2.2.2. Các bước để tính SVD

Bước 1: Tính giá trị riêng, véctơ riêng của A^TA để tìm hai ma trận V, $\sum$; tiếp đó tính giá trị riêng, véctơ riêng của AA^T để tìm ma trận U.

Bước 2: Xây dựng ma trận $U, \sum$ và V, trong đó ma trận Σ chứa các căn bậc hai dương của các giá trị riêng (không âm) của A^TA hoặc AA^T; các cột của U là các véctơ riêng đã được chuẩn hóa của AA^T và các cột của V là các véctơ riêng chuẩn hóa của A^TA.

Bước 3: Biểu diễn ma trận A=$U\sum V^T$.

Ví dụ: Cho ma trận $A = \begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 1 & 1 \end{pmatrix}$. Phân tích ma trận A thành $U\sum V^T$.

Giải:

Xét

$$\det(A^TA - \lambda I_2) = \begin{vmatrix} 2-\lambda & 3 \\ 3 & 6-\lambda \end{vmatrix} = \lambda^2 - 8\lambda + 3 = 0,$$

suy ra $\lambda_{1,2} = 4 \pm \sqrt{13}$.

Với $\lambda_1 = 4 + \sqrt{13}$, ta có

$$A^TA - \lambda_1 I_2 = \begin{pmatrix} -2-\sqrt{13} & 3 \\ 3 & 2-\sqrt{13} \end{pmatrix} \to \begin{pmatrix} 1 & \frac{2-\sqrt{13}}{3} \\ 0 & 0 \end{pmatrix}.$$

Suy ra véctơ riêng của A^TA tương ứng với giá trị riêng $\lambda_1 = 4 + \sqrt{13}$ là $v_1 = \begin{pmatrix} -2+\sqrt{13} \\ 3 \end{pmatrix}$.

Chuẩn của véctơ riêng

$$\|v_1\| = \sqrt{\left(-2+\sqrt{13}\right)^2 + 3^2}.$$

Với $\lambda_2 = 4 - \sqrt{13}$, ta có

$$A^TA - \lambda_2 I_2 = \begin{pmatrix} -2+\sqrt{13} & 3 \\ 3 & 2+\sqrt{13} \end{pmatrix} \to \begin{pmatrix} 1 & \frac{2+\sqrt{13}}{3} \\ 0 & 0 \end{pmatrix}.$$

Suy ra véctơ riêng của A^TA tương ứng với giá trị riêng $\lambda_2 = 4 - \sqrt{13}$ là $v_2 = \begin{pmatrix} 2+\sqrt{13} \\ 3 \end{pmatrix}$.

Chuẩn của véctơ riêng

$$\|v_2\| = \sqrt{\left(2+\sqrt{13}\right)^2 + 3^2}.$$

Khi đó, ma trận $V = \begin{pmatrix} \frac{v_1}{\|v_1\|} & \frac{v_2}{\|v_2\|} \end{pmatrix}$.

Mặt khác,

$$\det(AA^T - \lambda I_3) = \begin{vmatrix} 5-\lambda & 2 & 3 \\ 2 & 1-\lambda & 1 \\ 3 & 1 & 2-\lambda \end{vmatrix} = -\lambda^3 + 8\lambda^2 - 3\lambda = 0.$$

Suy ra, $\lambda_{1,2} = 4 \pm \sqrt{13}, \lambda_3 = 0$.

Với $\lambda_1 = 4 + \sqrt{13}$, ta có

$$AA^T - \lambda_1 I_3 \to \begin{pmatrix} 1 & 0 & \frac{-3-\sqrt{13}}{4} \\ 0 & 1 & \frac{1-\sqrt{13}}{4} \\ 0 & 0 & 0 \end{pmatrix}.$$

Suy ra véctơ riêng của A^TA tương ứng với giá trị riêng $\lambda_1 = 4 + \sqrt{13}$ là $u_1 = \begin{pmatrix} 3+\sqrt{13} \\ -1+\sqrt{13} \\ 4 \end{pmatrix}$.

Với $\lambda_2 = 4 - \sqrt{13}$, ta có

$$AA^T - \lambda_2 I_3 \to \begin{pmatrix} 1 & 0 & \frac{-3+\sqrt{13}}{4} \\ 0 & 1 & \frac{1+\sqrt{13}}{4} \\ 0 & 0 & 0 \end{pmatrix}.$$

Suy ra véctơ riêng của A^TA tương ứng với giá trị riêng $\lambda_2 = 4 - \sqrt{13}$ là $u_2 = \begin{pmatrix} 3-\sqrt{13} \\ -1-\sqrt{13} \\ 4 \end{pmatrix}$.

Với $\lambda = 0$, ta có

$$AA^T - \lambda_3 I_3 \to \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Suy ra véctơ riêng của A^TA tương ứng với giá trị riêng $\lambda_3 = 0$ là $u_3 = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$.

Khi đó, ma trận $U = \begin{pmatrix} \frac{u_1}{\|u_1\|} & \frac{u_2}{\|u_2\|} & \frac{u_3}{\|u_3\|} \end{pmatrix}$.

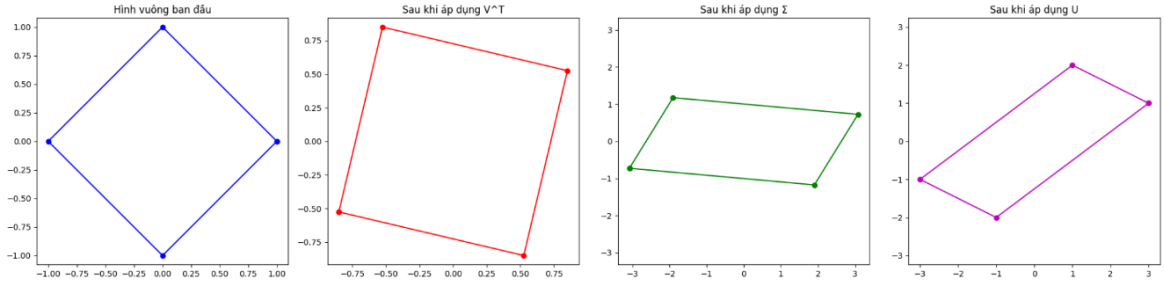
Giá trị kỳ dị của A là $\sigma_1 = \sqrt{\lambda_1}$, $\sigma_2 = \sqrt{\lambda_2}$. Suy ra, ma trận

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \sqrt{4+\sqrt{13}} & 0 \\ 0 & \sqrt{4-\sqrt{13}} \\ 0 & 0 \end{pmatrix}.$$

Vậy ma trận $A = U\sum V^T$.

2.2.3. Ý nghĩa hình học của phân tích SVD

Phân tích SVD là việc thực hiện ba phép biến hình liên tiếp, bao gồm phép quay tương ứng với ma trận trực giao V^T, phép vị tự tương ứng ma trận chéo $\sum$ và phép quay tương ứng với ma trận trực giao U (Hình 1).



Hình 1. Biểu diễn hình học của phân tích SVD

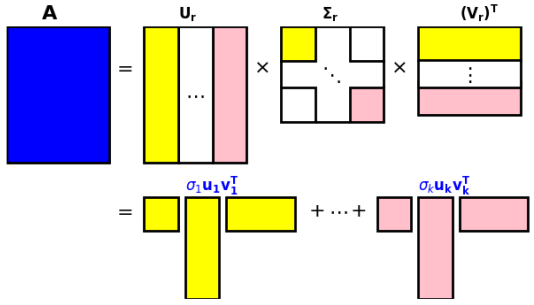
2.3. Giảm chiều dữ liệu bằng SVD (Truncated SVD)

Cho ma trận A là một ma trận cấp $m \times n$, thuật toán SVD phân tích ma trận A thành dạng

$$A = U \sum V^T.$$

Ở đây, $\sum$ là một ma trận đường chéo cấp $m \times n$ với các giá trị trên đường chéo là không âm và được sắp theo thứ tự giảm dần $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0 = 0 = \cdots = 0$. Khi đó, ta xấp xỉ ma trận A bằng tổng của k ma trận có hạng bằng 1 (với $k < n$).

$$A \approx A_k = U_k \sum_k (V_k)^T = \sum_{i=1}^{k} \sigma_i u_i {v_i}^T.$$



Hình 2. Truncated SVD (Vu, 2018)

2.4. Thuật toán phân cụm K-means

Trong thực tế, một yêu cầu quan trọng trong xử lý dữ liệu là phải phân chia dữ liệu đó. Ví dụ, một công ty cần đưa ra chính sách chăm sóc khách hàng, chiến lược kinh doanh thì công ty cần phân chia các nhóm khách hàng phù hợp với tiêu chí đặt ra. Phân chia dữ liệu sẽ giúp hiểu rõ hơn cấu trúc dữ liệu, làm giảm độ phức tạp của dữ liệu, là công cụ tiền xử lý cho các thuật toán khác cũng như ứng dụng thực tiễn vào nhiều lĩnh vực khác. Do đó, phân cụm dữ liệu không những làm cho các công đoạn khác trở nên đơn giản hơn mà còn tăng hiệu quả cũng như thời gian xử lý công việc, khả năng ra quyết định. Một số thuật toán phân cụm dữ liệu tiêu biểu như K-means, Hierarchical, DBSCAN, Mean Shift, GMM,... Mỗi phương pháp đều có những ưu điểm và nhược điểm. Trong phạm vi bài báo này, tác giả trình bày thuật toán phân cụm K-means bởi vì nó đơn giản nhưng hiệu quả với các tập dữ liệu lớn.

Thuật toán phân cụm K-means thực hiện như sau:

Bước 1: Từ dữ liệu đầu vào ta thực hiện việc lựa chọn số nhóm cần tìm.

Bước 2: Mỗi cụm chọn một điểm đại diện (centroid).

Bước 3: Phân mỗi dữ liệu vào nhóm bằng cách tính khoảng cách của một điểm bất kỳ đến các centroid. Điểm đó gần centroid nào hơn thì sẽ thuộc về nhóm chứa centroid đó.

Bước 4: Cập nhật lại các centroid cho từng cụm bằng cách lấy trung bình cộng của tất cả các điểm dữ liệu của nhóm.

Bước 5: Lặp lại Bước 3.

Việc lựa chọn khoảng cách nào phụ thuộc vào đặc điểm dữ liệu và mục đích cụ thể của bài toán.

2.5. Một số loại khoảng cách

Việc đo khoảng cách giữa các điểm dữ liệu thường được thực hiện dựa trên các chuẩn (khoảng cách). Dưới đây là một số khoảng cách phổ biến thường hay sử dụng.

2.5.1. Khoảng cách Euclide

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}.$$

Khoảng cách này thường được sử dụng trong K-means, K-nearest neighbors (KNN).

2.5.2. Khoảng cách Manhattan

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|.$$

Khoảng cách này phù hợp với các dữ liệu có cấu trúc lưới, chẳng hạn như dữ liệu hình ảnh hoặc các bài toán tối ưu tuyến tính.

Ngoài ra còn một số loại khoảng cách khác như khoảng cách Chebyshev, khoảng cách Minkowski, khoảng cách Cosine, khoảng cách Hamming.

3. KẾT QUẢ VÀ THẢO LUẬN

Thuật toán SVD đã mang lại nhiều ứng dụng. Bài báo này trình bày việc phân cụm các bài báo khoa học theo nhóm tác giả hoặc hướng nghiên cứu thông qua các từ khóa trong các công trình đã công bố.

3.1. Sự cần thiết phải phân cụm dữ liệu

Để khai phá dữ liệu một cách hiệu quả thì việc nhóm các đối tượng hoặc các điểm dữ liệu lại với nhau là cần thiết theo nguyên tắc các dữ liệu có nhiều điểm tương đồng sẽ nhóm thành một nhóm. Việc phân cụm mang lại nhiều lợi ích cũng như ứng dụng trong nhiều lĩnh vực, cụ thể như: Trong phân tích thị trường, việc phân cụm khách hàng dựa vào thói quen mua sắm hoặc đặc điểm cá nhân (giới tính, nghề nghiệp, sở thích,...) có thể giúp các doanh nghiệp xây dựng các chiến lược marketing phù hợp; trong phân tích văn bản, việc nhóm các tài liệu hoặc bài báo tương tự nhau có thể xác định được định hướng nghiên cứu của các nhóm tác giả để từ đó có thể xây dựng những nhóm nghiên cứu liên ngành, đa lĩnh vực hiệu quả hơn. Hơn nữa, việc phân cụm có thể hỗ trợ tốt trong việc phân tích và nhận diện các nhóm điểm ảnh hoặc đối tượng trong một hình ảnh, phân cụm các cấu trúc gene, protein dựa trên chức năng hoặc cấu trúc của chúng. Phân cụm dữ liệu giúp khám phá cấu trúc ẩn trong dữ liệu, phát hiện các dị thường cũng như làm giảm đáng kể số lượng mẫu cần phân tích, từ đó làm tăng hiệu quả của việc xử lý dữ liệu. Nhờ đó, ta có thể hiểu rõ hơn về dữ liệu và đưa ra các quyết định tốt hơn trong việc khai thác thông tin từ dữ liệu đó.

3.2. Phân nhóm các bài báo khoa học dựa trên từ khóa bằng cách kết hợp thuật toán SVD và K-means

Việc phân nhóm các bài báo khoa học dựa trên từ khóa bằng cách kết hợp thuật toán SVD và K-means có thể được thực hiện như sau:

Bước 1: Chuẩn bị dữ liệu (lập danh sách các từ khóa của một số bài báo cần phân nhóm).

Bước 2: Xây dựng ma trận dữ liệu bằng cách sử dụng phương pháp TF-IDF.

Bước 3: Sử dụng thuật toán SVD để giảm chiều dữ liệu (Truncated SVD).

Bước 4: Thực hiện phân cụm bằng thuật toán K-means.

3.3. Sử dụng Python để thực hiện phân cụm các nhóm bài báo bằng cách kết hợp thuật toán SVD và K-means

Nội dung mục này là minh họa cho thuật toán được đề xuất ở Mục 3.2. Yêu cầu bài toán: Sử dụng Python để phân cụm 20 bài báo được đăng trên Tạp chí Khoa học Đại học Cần Thơ từ năm 2020 đến năm 2024 thuộc lĩnh vực toán học của 4 Nhóm tác giả dựa trên từ khóa bằng cách kết hợp thuật toán SVD và K-means.

3.3.1. Chuẩn bị dữ liệu

Thực hiện việc thống kê từ khóa của 20 bài báo thuộc lĩnh vực toán học của 4 nhóm tác giả.

Nhóm tác giả thứ nhất có 09 bài báo khoa học (Vo et al., 2020a, 2020b, 2021, 2022a, 2022b, 2022c, 2022d, 2024; Truong et al., 2024).

Nhóm tác giả thứ hai có 04 bài báo khoa học (Lam, 2021; Lam et al., 2021, 2024a, 2024b).

Nhóm tác giả thứ ba có 03 bài báo khoa học (Nguyen, 2020; Nguyen & Dao, 2022; Tran & Nguyen, 2024).

Nhóm tác giả thứ tư có 04 bài báo khoa học (Dinh et al., 2022; Nguyen et al., 2023; Nguyen et al., 2023; Tran et al., 2023).

Danh sách các từ khóa được trích từ 20 bài báo trên được thống kê trong Bảng 1.

Bảng 1. Thống kê từ khóa của 20 bài báo thuộc lĩnh vực toán học

Bài báo số	Nhóm tác giả	Từ khóa
1	1	Clustering, distance, extracting images, probability density function
2	1	Interval time series, forecasting model, fuzzy relationship, cluster analysis
3	2	Random walk, Jacob-Bernoulli's formula, moment, Markov operator
4	3	Mordukhovich subdifferential, optimal control, coderivative, marginal function, elliptic partial differential equation
5	2	Markov operator, Poisson equation, Random walk, variance
6	4	Ekeland's variational principle, interval-valued functions, outer semicontinuity
7	4	Weierstrass theorem, semicontinuity, interval valued function
8	1	Cluster analysis, distance, discrete data, genetic algorithm
9	4	Interval-valued functions, Ekeland's variational principle, weakly boundedness from below, inner semicontinuity
10	1	Forecasting, fuzzy time series, future, original data
11	1	Classification, distance, extracting image, probability density function
12	2	Law of large number, unfair game model, method of moment, Markov operator
13	2	Central limit theorem, fair game model, Markov operator, method of moment, random walk
14	1	Algorithm, cluster analysis, distance, similar index
15	1	Classification, image data, priori probability, overlap distance
16	3	Boundary control, distributed control, existence of solution, full Lipschitzian stability, optimality condition
17	1	Cluster analysis, forecasting model, interval data, time series
18	3	Marginal function, objective function, optimal control, regular subdifferential (Fréchet subdifferential), solution map
19	1	Clustering algorithm, image data, interval data, overlap distance
20	4	Ekeland's variational principle, lower semicontinuity, set perturbation

Thư viện Pandas là một công cụ trong Python được sử dụng để xử lý và phân tích dữ liệu dưới dạng bảng. Pandas cung cấp các cấu trúc dữ liệu như DataFrame và Series để quản lý và thao tác với dữ liệu dễ dàng hơn.

Thư viện Pandas trong Python được sử dụng để nhập dữ liệu bằng lệnh

```
import pandas as pd

# Danh sách các bài báo với từ khóa

articles = [

 "Clustering, distance, extracting images, probability density function",

"Interval time series, forecasting model, fuzzy relationship, cluster analysis",

"Random walk, Jacob-Bernoulli's formula, moment, Markov operator",

"Mordukhovich subdifferential, optimal control, coderivative, marginal function, elliptic partial differential equation",

"Markov operator, Poisson equation, Random walk, variance",

"Ekeland's variational principle, interval-valued functions, outer semicontinuity",

"Weierstrass theorem, semicontinuity, interval valued functions",

"Cluster analysis, distance, discrete data, genetic algorithm",

"Interval-valued functions, Ekeland's variational principle, weakly boundedness from below, inner semicontinuity",

"Forecasting, fuzzy time series, future, original data",

"Classification, distance, extracting image, probability density function",

"Law of large number, unfair game model, method of moment, Markov operator",

"Central limit theorem, fair game model, Markov operator, method of moment, random walk",

"Algorithm, cluster analysis, distance, similar index",
```

"Classification, image data, priori probability, overlap distance",

"Boundary control, distributed control, existence of solution, full Lipschitzian stability, optimality condition",

"Cluster analysis, forecasting model, interval data, time series",

"Marginal function, objective function, optimal control, regular subdifferential (Fréchet subdifferential), solution map",

"Clustering algorithm, image data, interval data, overlap distance",

"Ekeland's variational principle, lower semicontinuity, set perturbation"

]

3.3.2. Xây dựng ma trận dữ liệu

Ma trận TF-IDF là một dạng biểu diễn văn bản dưới dạng số liệu thống kê để đánh giá mức độ quan trọng của các từ trong văn bản, đồng thời giảm ảnh hưởng của các từ thông dụng (stop words). Mỗi văn bản được biểu diễn dưới dạng véctơ trong không gian từ vựng. Để chuyển đổi dữ liệu văn bản thành ma trận TF-IDF ta sử dụng lệnh TfidfVectorizer từ thư viện scikit-learn trong Python. Đây là một công cụ phổ biến được sử dụng trong xử lý dữ liệu và học máy.

```
from sklearn.feature extraction.text import TfidfVectorizer

# Biến đổi các từ khóa thành véctơ TF-IDF

vectorizer = TfidfVectorizer ()

X = vectorizer.fit_transform(articles)
```

3.3.3. Giảm chiều dữ liệu bằng thuật toán SVD

Lớp TruncatedSVD từ thư viện scikit-learn được sử dụng để giảm số chiều của dữ liệu sau khi đã được chuyển thành véctơ, giúp giữ lại những thành phần quan trọng nhất của dữ liệu và giảm độ phức tạp khi thực hiện các thuật toán phân cụm. Trong xử lý văn bản, nó có thể giúp giảm số lượng từ đại diện mà vẫn giữ được nhiều thông tin.

```
from sklearn.decomposition import TruncatedSVD

# Giảm chiều dữ liệu bằng SVD

svd = TruncatedSVD(n_components=2) # Giảm xuống 2 chiều để dễ trực quan hóa

X_svd = svd.fit_transform(X)
```

3.3.4. Phân cụm dữ liệu bằng thuật toán K-means

Thuật toán K-means được sử dụng để phân cụm các văn bản sau khi đã biểu diễn chúng dưới dạng véctơ, giúp xác định các nhóm văn bản có nội dung tương tự dựa trên khoảng cách Euclide.

```
from sklearn.cluster import KMeans

# Thực hiện phân cụm K-means

k = n  # Số cụm mong muốn

kmeans = KMeans(n_clusters=k, random_state=0).fit(X_svd)

labels = kmeans.labels_
```

3.3.5. Kết quả phân cụm

Gói plt từ thư viện matplotlib.pyplot trong Python được dùng để vẽ đồ thị và trực quan hóa kết quả phân cụm, ví dụ như vẽ biểu đồ thể hiện sự phân bố của các cụm hoặc biểu đồ thể hiện số chiều được giữ lại sau khi giảm số chiều bằng SVD.

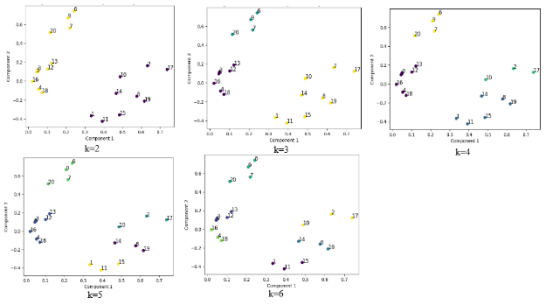
```
import matplotlib.pyplot as plt

# Vẽ biểu đồ phân cụm

plt.scatter(X_svd[:, 0], X_svd[:, 1], c=labels, cmap='viridis')

for i, article in enumerate(articles):

    plt.text(X_svd[i, 0], X_svd[i, 1], str(i+1), fontsize=12)

plt.xlabel('Component 1')

plt.ylabel('Component 2')

plt.show()
```

Dưới đây là kết quả thực hiện phân nhóm trong các trường hợp $k = 1, 2, 3, 4, 5, 6$. Khi đó, kết quả phân cụm các bài báo của 4 Nhóm tác giả dựa trên từ khóa của các bài báo khoa học của họ bằng cách sử dụng Python để tính toán được cho bởi Bảng 2 với số lượng cụm khác nhau.

Bảng 2: Kết quả phân cụm với $k = 1, 2, 3, 4, 5, 6$

Cụm	Bài báo số				
1	1, 2, 8, 10, 11, 14, 15, 17,19	1, 2, 8, 10, 11, 14, 15, 17,19	1, 8, 11, 14, 15, 19	1, 11, 15	1, 11, 15
2	3, 4, 5, 6, 7, 9, 12, 13, 16, 18, 20	3, 4, 5, 12, 13, 16, 18	3, 4, 5, 12, 13, 16, 18	3, 4, 5, 12, 13, 16, 18	3, 5, 12, 13
3		6, 7, 9, 20	6, 7, 9, 20	6, 7, 9, 20	6, 7, 9, 20
4			2, 10, 17	2, 10, 17	2, 10, 17
5				8, 14, 19	8, 14, 19
6					4, 16, 18

Hình 3 là hình ảnh minh họa cho kết quả phân cụm 20 bài báo khoa học của 4 Nhóm tác giả trên không gian 2 chiều được cho bởi Bảng 1.



Hình 3. Kết quả phân cụm của bài báo dựa trên từ khóa bằng SVD và K-Means

Kết quả trên cho thấy việc kết hợp thuật toán SVD và thuật toán phân cụm K-means cho kết quả khá chính xác các công trình nghiên cứu của các tác giả. Ví dụ, với trường hợp $k = 6$, ta thấy các cụm 1, 4, 5 là các bài báo thuộc về nhóm tác giả thứ nhất; cụm 2 là các bài báo thuộc về nhóm tác giả thứ hai; cụm 3 là các bài báo thuộc về nhóm tác giả thứ tư; cụm 6 là các bài báo thuộc về nhóm tác giả thứ ba. Nhóm tác giả thứ nhất do có nhiều hướng nghiên cứu khác nhau nên được phân thành các cụm nhỏ hơn phù hợp với các hướng nghiên cứu đó.

3.3.6. Đánh giá hiệu quả của thuật toán được đề xuất

Việc phân cụm các bài báo khoa học được dựa trên các từ khóa nên thuật toán được đề xuất rõ ràng không phụ thuộc vào nguồn dữ liệu được chọn từ tạp chí nào. Tuy nhiên, độ chính xác của kết quả thuật toán phụ thuộc vào nhiều yếu tố khác nhau. Một số yếu tố quan trọng ảnh hưởng đến kết quả có thể kể đến như:

Thứ nhất, dữ liệu về các từ khóa của các bài báo khóa học được chọn đã mang tính đại diện cho hướng nghiên cứu hay bài báo khoa học đã được công bố chưa. Vì đôi khi, các từ khóa trong các bài báo chưa thực sự phản ánh đúng nội dung nghiên cứu hoặc quá thông dụng (stop words).

Thứ hai, một nhóm tác giả có thể theo đuổi nhiều hướng nghiên cứu chuyên sâu khác nhau.

Thứ ba, các nhóm tác giả khác nhau có thể có một số công trình có hướng nghiên cứu tương tự nhau.

Thứ tư, số lượng nhóm được chọn để phân cụm chưa phù hợp.

Thứ năm, dữ liệu có thể phân tán, không tập trung vào một số ít các hướng nghiên cứu chủ đạo, nghĩa là dữ liệu được chọn thuộc quá nhiều hướng nghiên cứu hoặc quá nhiều nhóm tác giả nhưng mỗi nhóm hoặc mỗi hướng nghiên cứu có quá ít bài báo.

Do đó, trong trường hợp nguồn dữ liệu lớn để nâng cao tính chính xác của thuật toán ta có thể áp dụng thuật toán theo phương pháp phân tầng, chia nhỏ dữ liệu dần dần. Đầu tiên, thuật toán sẽ phân chia dữ liệu thành một số lượng cụm phù hợp, sau đó tiếp tục áp dụng thuật toán cho từng cụm kết quả của bước trước. Nói một cách tường minh, giả sử yêu cầu bài toán là phân chia dữ liệu thành n cụm. Nếu ta chỉ áp dụng thuật toán trên đề xuất một lần, tức là chọn $k = n$ thì ta có thể bắt đầu với $k = m$ cụm (với $m < n$), tiếp tục phân chia từng cụm trong m thành các m_i cụm con, cứ thực hiện như vậy cho đến khi đạt số lượng cụm mong muốn. Số lượng cụm con của từng cụm ở mỗi bước tiếp theo không nhất thiết phải bằng nhau. Phương pháp này tạo nên sự linh hoạt trong quá trình phân tích, giảm dung lượng bộ nhớ, thời gian xử lý dữ liệu và tăng tính chính xác của kết quả.

4. KẾT LUẬN

Phân cụm các bài báo toán học theo nhóm dựa trên từ khóa bằng thuật toán SVD và K-means là một phương pháp hiệu quả trong việc tổ chức và quản lý thông tin khoa học. Bằng cách sử dụng TF-IDF để mã hóa từ khóa và áp dụng SVD để giảm số chiều, phương pháp này giúp tập trung vào các đặc trưng

quan trọng nhất, loại bỏ nhiễu và tăng cường khả năng phát hiện các nhóm bài báo có chủ đề tương tự. Sau khi giảm số chiều, thuật toán K-means phân cụm các bài báo dựa trên mức độ tương đồng của từ khóa, giúp các nhà nghiên cứu dễ dàng nhận diện các nhóm tác giả có chung hướng nghiên cứu. Phương pháp này không chỉ cải thiện khả năng phân tích và xử lý dữ liệu lớn, mà còn đóng góp vào việc tối ưu hóa việc tìm kiếm và phân loại các bài báo khoa học trong lĩnh vực toán học. Dựa trên kết quả này, trong thời gian tới nhóm tác giả có thể tiếp tục nghiên cứu việc giải bài toán này bằng cách kết hợp thuật toán K-means và phân tích thành phần chính (PCA) để so sánh, đánh giá kết quả cũng như độ tin cậy của phương pháp đã đề xuất.

LỜI CẢM TẠ

Đề tài này được tài trợ bởi Trường Đại học Cần Thơ, Mã số: TSV2024-25.

TÀI LIỆU THAM KHẢO (REFERENCES)

Chicco, D., & Masseroli, M. (2013, November). A discrete optimization approach for SVD best truncation choice based on ROC curves. *In 13th IEEE International Conference on BioInformatics and BioEngineering* (pp. 1-4). IEEE. https://doi.org/10.1109/BIBE.2013.6701705

Dinh, Q. N., Do, D. H., & Ha, A. N. H. (2022). Ekeland's variational principle for bifunctions involving set perturbations. *Can Tho University Journal of Science*, *58*, 121-128 *(in Vietnamese)*. https://doi.org/10.22144/ctu.jvn.2022.106

Eckart, C., & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211-218. https://doi.org/10.1007/BF02288367

Golub, G., & Kahan, W. (1965). Calculating the singular values and pseudo-inverse of a matrix. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis, 2*(2), 205-224. https://doi.org/10.1137/0702016

Kaloorazi, M. F. (2018). *Low-Rank Matrix Approximations and Applications* (PhD Thesis). Pontifícia Universidade Católica do Rio de Janeiro.

Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer, 42*(8), 30-37.

https://doi.org/10.1109/MC.2009.263Li, X., Pang, Y., Zhao, C., Liu, Y., & Dong, Q. (2021). A new multi-level algorithm for balanced partition problem on large scale directed graphs. *Advances in Aerodynamics, 3*(1), 23. https://doi.org/10.1186/s42774-021-00074-x

Lam, C. H. (2021). Law of large numbers in the unfair game model. *Can Tho University Journal of Science*, *57*(2), 44-48 *(in Vietnamese)*.

Lam, C. H., Tran, L. P., La, K. M., & Duong, Tn. T. (2021). Central limit theorem in the fair game model. *Can Tho University Journal of Science*, *57*(2), 39-43 *(in Vietnamese)*.

Lam, Chuong. H., Trinh, Nghiem. H., & Le, Nhan. H (2024a). Higher order moment for random walks in discrete state space. *Can Tho University Journal of Science*, *60*, 58-62 *(in Vietnamese)*.

Lam, Chuong. H., Nguyen, Truong. V., Nguyen, Nhu. T. H., Phan, Hang. T. M., & Nguyen, Nhiem. C. (2024b). Limit of variance for random walk in space $a\mathbb{Z}$. *Can Tho University Journal of Science*, *60*(2), 36-40 *(in Vietnamese)*.

Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., & Mannila, H. (2008). The discrete basis problem. *IEEE transactions on knowledge and data engineering*, *20*(10), 1348-1362. https://doi.org/10.1109/TKDE.2008.53

Nguyen, P. T., Huynh, D. T., Pham, D. C., Do, D. H., & Dinh, Q. N. (2023). On generalized Ekeland's variational principle for interval-valued functions based on the inner semicontinuity. *Can Tho University Journal of Science*, *59*(5), 17-24 *(in Vietnamese)*.

Nguyen, Q. T. (2020). Distributed and boundary control problems governed by semilinear elliptic partial differential equations. *Can Tho University Journal of Science*, *56*, 1-7 *(in Vietnamese)*.

Nguyen, Q. T., & Dao, P. D. (2022). Generalized differentiation of marginal functions in parametric optimal control governed by elliptic partial differential equations. *Can Tho University Journal of Science*, *58*(1), 87-94 *(in Vietnamese)*.

Nguyen, T. C., Tran, D. V., Huynh, D. T., Nguyen, P. T., & Dinh, Q. N. (2023). Weierstrass theorem for interval valued functions. *Can Tho University Journal of Science*, *59*(5), 55-63 *(in Vietnamese)*.

Ramponi, G., Brambilla, M., Ceri, S., Daniel, F., & Di Giovanni, M. (2019). Vocabulary-based community detection and characterization. *In Proceedings of the 34th ACM/SIGAPP symposium on applied computing* (pp. 1043-1050). https://doi.org/10.1145/3297280.3297384

Sarkar, S., & Dong, A. (2011). Community detection in graphs using singular value decomposition. *Physical Review E, 83*(4), 046114. https://doi.org/10.1103/PhysRevE.83.046114

Stanimirovic, I. (2020). *Applications of Graph Theory*. Arcler Press.

Stewart, G. W. (1993). On the early history of the singular value decomposition. *Society for Industrial and Applied Mathematics review, 35*(4), 551-566. https://doi.org/10.1137/1035134

Tran, A. S. H., & Nguyen, Q. T. (2024). Mordukhovich subdifferential of marginal functions in parametric optimal control with equilibrium constraints. *Can Tho University Journal of Science*, *60*, 176-184 *(in Vietnamese)*.

Tran, D. V., Ha, A. N. H., Do, D. H., & Dinh, Q. N. (2023). Ekeland's variational principle for interval-valued functions based on the outer semicontinuity. *Can Tho University Journal of Science*, *59*(5), 10-16 *(in Vietnamese)*.

Truong, L. M., Nguyen, N. K., Nguyen, C. H., Phan, H. N., & Vo, T. V. (2024). The clustering algorithm for images based on extracted color pixels. *Can Tho University Journal of Science*, *60*, 98-107 *(in Vietnamese)*.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of cognitive neuroscience, 3*(1), 71-86. https://doi.org/10.1162/jocn.1991.3.1.71

Vo, T. V., Nguyen, Q. V., Huynh, H. V., Trang, K. T. M., & Nguyen, D. T. H. (2020a). An improved fuzzy time series forecasting model. *Can Tho University Journal of Science*, *56*(1), 86-94 *(in Vietnamese)*.

Vo, T. V., Nguyen, T. T., Huynh, H. V., Tran, T. T., & Chau, T. N. (2020b). Improving the cluster analysis algorithm for discrete elements. *Can Tho University Journal of Science*, *56*(2), 30-36 *(in Vietnamese)*.

Vo, T. V., Tu, T. N., & Tran, H. N. N. (2021). Building the time series forecasting model for interval data based on cluster analysis problem. *Can Tho University Journal of Science, 57*(5), 94-103 *(in Vietnamese)*.

Vo, T. V., Le, C. T. K., & Chau, T. N. (2022a). Building clusters for image data from the extracted two dimensional interval. *Can Tho University Journal of Science*, *58*(5), 22-30 *(in Vietnamese)*.

Vo, T. V., Nguyen, T. H., Phan, T. N. N., Tang, K. X., & Tran, T. Đ. (2022b). Genetic algorithm in building cluster for discrete data and applying for image. *Can Tho University Journal of Science*, *58*(3), 107-114 *(in Vietnamese)*.

Vo, T. V., Nguyen, T. T. H., Dang, T. T. P., & Tran, H. N. (2022c). Classify images based on the extracted interval features from the gray level co-occurrence matrix. *Can Tho University Journal of Science*, *58*(5), 31-38 *(in Vietnamese)*.

Vo, T. V., Tran, H. N., & Huynh, N. V. (2022d). Classifying for image based on the extracted probability density function. *Can Tho University Journal of Science*, *58*(6), 43-50 *(in Vietnamese)*.

Vo, T. V., Nguyen, L. H., Danh, T. N., Tang, K. M. & Le, N. D. (2024). Building a forecasting model for interval time series based on point series. *Can Tho University Journal of Science*, *60,* 150-158 *(in Vietnamese)*.

Vu, T. H. (2018). *Basic Machine Learning*. https://machinelearningcoban.com

Zha, H., He, X., Ding, C., Simon, H., & Gu, M. (2001). Bipartite graph partitioning and data clustering. *In Proceedings of the tenth international conference on Information and knowledge management* (pp. 25-32). https://doi.org/10.2172/816202