ĐÁNH GIÁ HIỆU QUẢ CỦA CÁC ĐỘ ĐO TRONG MÔ HÌNH KẾT HỢP GIỮA PHÂN BỔ DIRICHLET TIỀM ẨN VÀ K-MEANS CHO BÀI TOÁN PHÂN CỤM TÀI LIỆU

Bùi Quang Vũ*, Trần Thiện Thành, Ngô Nhân Đức, Nguyễn Hoàng Hà, Nguyễn Dũng

Trường Đại học Khoa học, Đại học Huế

*Email: buiquangvu@hueuni.edu.vn

Ngày nhận bài: 01/12/2022; ngày hoàn thành phản biện: 08/12/2022; ngày duyệt đăng: 20/12/2022

TÓM TẮT

Bài báo này là một nghiên cứu thực nghiệm nhằm mục đích đánh giá hiệu quả của các độ đo khoảng cách khi sử dụng mô hình kết hợp giữa LDA và K-means cho bài toán phân cụm tài liệu. Kết quả thực nghiệm cho thấy rằng các độ đo khoảng cách dựa trên xác suất tốt hơn so với các độ đo dựa vào véc tơ khi sử dụng trong bài toán phân cụm tài liệu trong không gian các chủ đề. Với việc chọn độ đo độ tương tự dựa trên xác suất, K-means kết hợp với mô hình phân bổ Dirichlet tiềm ẩn (LDA) cho kết quả tốt hơn so với sử dụng LDA+ Naive và mô hình không gian véc tơ.

Từ khóa: Mô phỏng chủ đề, phân bổ Dirichlet tiềm ẩn, phân cụm tài liệu, K-means, độ đo khoảng cách dựa vào xác suất.

1. GIỚI THIỆU

Phân loại một tập hợp tài liệu là một bài toán điểm hình được giải quyết trong học máy và xử lý ngôn ngữ tự nhiên. Trong tác vụ này, văn bản được gán cho một hoặc nhiều nhãn lớp được xác định trước (tức là danh mục) thông qua một quy trình cụ thể trong đó bộ phân loại được xây dựng, đào tạo trên một tập hợp các đặc trưng và sau đó áp dụng để gắn nhãn văn bản đến trong tương lai. Một số thuật toán Học máy đã được áp dụng để phân loại văn bản như thuật toán Rocchio, N-Nearest Neighbors, Naive Bayes, Cây quyết định, Rừng ngẫu nhiên, Máy véc tơ hỗ trợ (SVM), v.v.

Trong trường hợp không có các nhãn được xác định trước, tác vụ được gọi là tác vụ phân cụm và được thực hiện trong khuôn khổ học tập không giám sát. Phân cụm là một trong những thuật toán khai phá dữ liệu phổ biến nhất và đã được nghiên cứu rộng rãi trong ngữ cảnh của văn bản. Phân cụm là nhiệm vụ tìm kiếm các nhóm tài

liệu giống nhau trong một tập hợp các tài liệu. Độ tương tự được tính bằng cách sử dụng một hàm tương tự. Có nhiều thuật toán phân cụm có thể được sử dụng trong ngữ cảnh của dữ liệu văn bản. Tài liệu văn bản có thể được biểu diễn dưới dạng vectơ nhị phân, tức là xem xét sự hiện diện hoặc vắng mặt của từ trong tài liệu. Hoặc chúng ta có thể sử dụng các biểu diễn tinh chỉnh hơn liên quan đến các phương pháp trọng số như ma trận Tf-idf (tần số xuất hiện từ vựng (tf) và tần số tài liệu nghịch đảo (idf)). Trong Tfidf, mỗi tài liệu được biểu diễn dưới dạng véc tơ của các lần xuất hiện từ. Điều này đặt tên cho mô hình là Mô hình không gian vectơ (VSM).

Có hai loại điển hình của thuật toán phân cụm: phân vùng và kết tụ. K-means và phân cụm phân cấp (hierarchical clustering) lần lượt là đại diện của hai loại này. Có nhiều so sánh giữa k-means và phân cụm phân cấp. Trong công việc của mình, chúng tôi chọn thuật toán k-mean để phân cụm tài liệu vì nó nhanh hơn nhiều. Đối với phân nhóm tài liệu sử dụng mô hình VSM, chúng ta phải đối mặt với hai thách thức: "lời nguyền của chiều không gian" [14] vì số thuật ngữ (term) thường rất lớn và làm thế nào để chọn một thước đo khoảng cách "tốt" để có được các cụm chính xác nhất?

Trong nghiên cứu này, chúng tôi đã giảm số chiều của ma trận bằng cách sử dụng mô hình phân bổ dirichlet tiềm ẩn (Latent Dirichlet Allocation, LDA) [2]. Mỗi tài liệu sẽ được biểu diễn dưới dạng một phân phối xác suất các chủ đề và mỗi chủ đề lại được đặc trưng bởi một phân phối xác suất của các thuật ngữ (term). Chúng tôi sử dụng phân phối xác suất của các chủ đề như là đầu vào cho các thuật toán phân cụm như K-means. Cách tiếp cận này gọi là LDA + K-means đã được đề xuất trong nghiên cứu [3, 17]. Tuy nhiên, trong [17], LDA + K-means chỉ sử dụng độ đo Euclidean nên chưa thể hiện được hiệu quả của cách tiếp cận này.

Đối với vấn đề thứ hai, chúng tôi tiến hành thực nghiệm và so sánh hiệu quả trên các độ đo khác nhau [5]. Các độ đó này được dựa trên hai cách tiếp cận: (i) Tiếp cận dựa vào véc tơ (VBA) bao gồm khoảng cách Euclidean, Sørensen, Cosine; (ii) Tiếp cận dựa vào xác suất (PBM) bao gồm Bhattacharyya, Jensen-Shannon, Taneja divergence.

Để đưa ra được các kết luận đúng đắn, chúng tôi đã tiến hành thực nghiệm với sáu độ đo khoảng cách theo phân nhóm có gán nhãn. Chúng tôi đã đánh giá việc phân cụm với 2 tiêu chí: Adjusted Rand Index (ARI) [9] và Adjusted Mutual Information (AMI) [16]. Chúng tôi đã sử dụng hai tập dữ liệu phổ biến trong xử lý ngôn ngữ tự nhiên (NLP): tập dữ liệu *20NewsGroup* chứa nhóm tin các bài đăng và *WebKB* chứa các văn bản được trích xuất từ các trang web.

Bài báo này được tổ chức như sau: phần tiếp theo sẽ trình bày về các phương pháp phân cụm tài liệu: phân cụm tài liệu với mô hình không gian véc tơ được trình bày ở phần 2, phân cụm tài liệu sử dụng mô phỏng chủ đề được trình bày trong phần

3, việc khảo sát vai trò của độ đo tương tự trên không gian xác suất được trình bày ở phần 4 và phần kết luận được trình bày ở phần 5.

2. PHÂN CỤM TÀI LIỆU VỚI MÔ HÌNH KHÔNG GIAN VECTOR (VECTOR SPACE MODEL)

2.1. Mô hình không gian véc tơ

Để phân cụm văn bản, trước tiên cần chuyển văn bản sang dạng dữ liệu thích hợp. Phương pháp phổ biến được sử dụng để biểu diễn văn bản là BoW (Bag-of-Words). Trong phương pháp này, một văn bản d được thể hiện dưới dạng một véc tơ tần suất các từ (term-frequence). Cụ thể, cho $D = \{d_1, d_2, \dots, d_n\}$ là tập các văn bản và $T = \{t_1, t_2, \dots, t_m\}$ là tập các thuật ngữ (terms) duy nhất trong tập dữ liệu. Một văn bản sau đó được biểu diễn bởi một véc tơ $\vec{t}_d$ trong không gian m chiều: $\vec{t}_d = (tf(d,t_1), tf(d,t_2), ..., tf(d,t_m))$ trong đó $tf(d,t_i)$ chính là tần suất xuất hiện của thuật ngữ $t_i \in T$ trong văn bản $d \in D$. Mô hình này có tên gọi là mô hình không gian vector (VSM) [15]. Trong mô hình này, tập dữ liệu $D = \{d_1, d_2, \dots, d_n\}$ và tập các thuật ngữ xuất hiện trong tập dữ liệu $T = \{t_1, t_2, \dots, t_m\}$ được biểu diễn dưới dạng ma trận A_{nxm}. Trong đó văn bản d_i được biểu diễn là dòng thứ i và cột j là thuật ngữ (term) thứ j xuất hiện trong tập dữ liệu, và $a_{i,j}$ là tần suất xuất hiện của thuật ngữ j trong văn bản d_i.

Văn bản	đây	là	con	mèo	ngồi	trên	cái	mũ	chó	và
đây là con mèo	1	1	1	1	0	0	0	0	0	0
con mèo ngồi	0	0	1	1	1	0	0	0	0	0
con mèo ngồi trên cái mũ	0	0	1	1	1	1	1	1	0	0
con mèo và con chó	0	0	2	1	0	0	0	0	1	1

Hình 1. Ví dụ về mô hình Vecto Space Model

2.2. Một số độ đo độ tương tự trong không gian véc tơ

Trước khi phân cụm, một độ đo tương tự hoặc khoảng cách cần được xác định. Phép đo phản ánh mức độ gần gũi hoặc tách biệt của các đối tượng mục tiêu và phải tương ứng với các đặc điểm được cho là để phân biệt các cụm được nhúng trong dữ liệu. Trong nhiều trường hợp, những đặc điểm này phụ thuộc vào dữ liệu hoặc bối cảnh bài toán và không có biện pháp nào là tốt nhất cho tất cả các bài toán phân cụm. Sau đây, chúng ta trình bày ba khoảng cách giữa hai vectơ thường được sử dụng trong phân nhóm văn bản: khoảng cách Euclidean, khoảng cách Cosine, khoảng cách

Sørensen. Từ phần này trở về sau, chúng ta ký hiệu $A=(a_1,a_2,\ldots,a_k)$ và $B=(b_1,b_2,\ldots,b_k)$ là hai vector với số chiều là k.

2.2.1. Khoảng cách Euclidean

Khoảng cách Euclidean, còn được gọi là chuẩn L2 trong họ thước đo khoảng cách Minkowski, là một thước đo tiêu chuẩn cho các bài toán hình học. Khoảng cách Euclidean được sử dụng rộng rãi trong các bài toán phân cụm, bao gồm phân cụm văn bản và cũng là thước đo khoảng cách mặc định được sử dụng với thuật toán K-means. Khoảng cách Euclidean giữa hai điểm A và B với bậc k là được tính như sau:

$$d_{Euc}(A,B)=\sqrt{\sum_{i=1}^{k}|a_i-b_i|^2}$$

2.2.2. Khoảng cách Sørensen

Khoảng cách Sørensen là một trong những khoảng cách trong họ L1, chính xác hơn là sự khác biệt tuyệt đối. Cho hai véc tơ A và B, khoảng cách Sørensen giữa A và B được xác định là:

$$d_{Sor}(A,B)=\frac{\sum_{i=1}^{k}|a_i-b_i|}{\sum_{i=1}^{k}(a_i+b_i)}$$

2.2.3. Khoảng cách Cosin

Độ tương tự cosine là độ đo sự tương quan giữa các véc tơ trong không gian. Độ tương tự cosine chính là cosin của góc giữa hai vectơ. Cho hai véc tơ A và B, khoảng cách cosin giữa A và B được xác định là

$$d_{Cos}(A,B)=1-Sim_{Cos}(A,B)=1-\frac{\sum_{i=1}^{k}a_i.b_i}{\sqrt{\sum_{i=1}^{k}{a_i}^2}\sqrt{\sum_{i=1}^{k}{b_i}^2}}$$

2.3. Thuật toán K-means

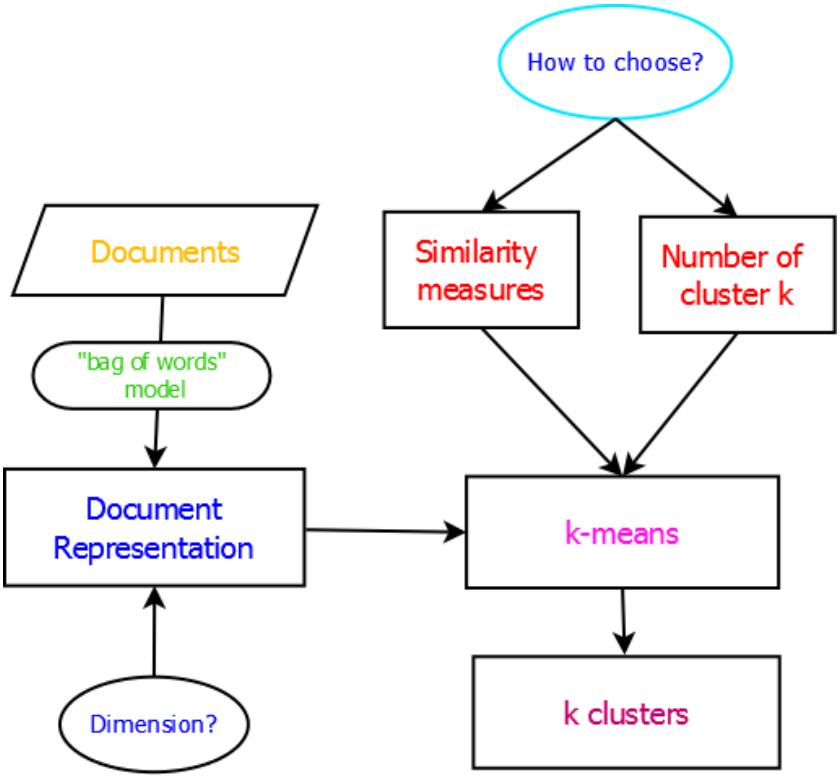
K-means được đề xuất bởi Forgy [6], là một trong những thuật toán phân cụm phổ biến nhất. Thuật toán này đơn giản và dễ sử dụng, dùng để phân loại các đối tượng vào trong k cụm cố định trước. Ý tưởng của thuật toán này đó là định nghĩa k tâm cụm (centroid), sau đó gán các đối tượng vào trong các cụm gần nó nhất. Thuật toán gồm các bước như sau:

- Bước 1: Khởi tạo K điểm dữ liệu trong bộ dữ liệu và tạm thời coi nó là tâm của các cụm dữ liệu của chúng ta.
- Bước 2: Gán nhãn cụm cho mỗi điểm dữ liệu trong bộ dữ liệu với tâm cụm của nó sẽ được xác định là 1 trong K tâm cụm gần nó nhất.

- Bước 3: Sau khi tất cả các điểm dữ liệu đã có tâm, tính toán lại vị trí của tâm cụm để đảm bảo tâm của cụm nằm ở chính giữa cụm.
- Lặp lại Bước 2 và Bước 3 cho tới khi vị trí của tâm cụm không thay đổi hoặc tâm của tất cả các điểm dữ liệu không thay đổi.

2.4. Phân cụm văn bản với mô hình không gian vec tơ

Sau khi biểu diễn các văn bản bằng mô hình VSM, chúng ta có dữ liệu là ma trận A_{nxm} và do đó chúng ta có thể sử dụng ma trận này làm đầu vào cho các thuật toán phân cụm tiêu chuẩn. Chúng ta có thể sử dụng các thuật toán phân cụm thường sử dụng như phân cụm có thứ bậc, k-means, k-medoids, …Trong nghiên cứu này, chúng tôi chọn thuật toán K-means để phân cụm văn bản. Hình 2 cho thấy một lược đồ phân cụm văn bản với VSM sử dụng thuật toán K-means.



Hình 2. Mô hình VSM cho phân cụm văn bản

3. PHÂN CỤM VĂN BẢN SỬ DỤNG MÔ PHỎNG CHỦ ĐỀ

3.1. Phân bổ Dirichlet tiềm ẩn (Latent Dirichlet Allocation)

Phân bổ Dirichlet tiềm ẩn (LDA) [2] là một mô hình sinh xác suất để khám phá chủ đề từ một bộ các văn bản. Trong LDA, mỗi văn bản có thể được coi là một phân phối xác suất của các chủ đề khác nhau và mỗi chủ đề được đặc trưng bởi một phân

phối xác suất trên một lượng từ vựng hữu hạn. Mô hình sinh văn bản của LDA được mô tả bằng mô hình đồ họa xác suất trong Hình 3, tiến hành các bước như sau:

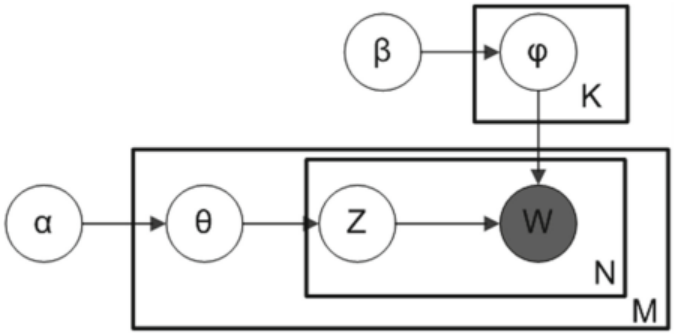
Bước 1: Chọn một phân phối theo các chủ đề θ_i từ phân phối Dirichlet với tham số α cho mỗi văn bản.

Bước 2: Chọn một phân phối theo các từ $\emptyset_k$ từ phân phối Dirichlet với tham số β cho mỗi chủ đề.

Bước 3: Với mỗi vị trí từ i, j:

Bước 3.1: Chọn một chủ đề $z_{i,j}$ từ phân phối đa thức (Multinomial distribution) với tham số θ_i

Bước 3.2: Chọn 1 từ $w_{i,j}$ từ phân phối đa thức với tham số $\emptyset_{z_{i,j}}$

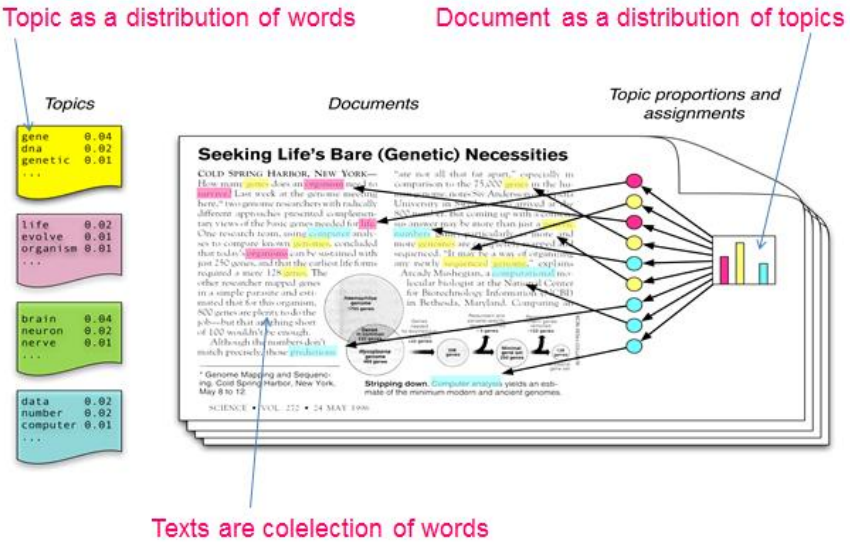


Hình 3. Mô hình xác suất của LDA

Để suy ra hậu nghiệm, chúng ta cần giải phương trình:

$$p(\theta, \emptyset, z|w, \alpha, \beta) = \frac{p(\theta, \emptyset, z,\ w|\ \alpha, \beta)}{p(w|\ \alpha, \beta)}$$

Có một số thuật toán để ước lượng các tham số của mô hình như variational inference [2] hoặc Gibbs Sampling [1]. Ví dụ minh họa cho mô hình LDA được thể hiện ở Hình 4.



Hình 4. LDA Example [1]

3.2. LDA + Naïve

Cách tiếp cận đầu tiên mà chúng tôi gọi là LDA + Naïve sử dụng trực tiếp các mô hình mô phỏng chủ đề. Giả định cơ bản là mỗi chủ đề tương ứng với một cụm, vì vậy số lượng chủ đề trong các mô hình chủ đề phù hợp với số lượng cụm. Sau khi ước lượng các thông số, mỗi văn bản là một phân phối xác suất trên các chủ đề $\theta = (\theta_1, \theta_2, ..., \theta_K)$ với K là số chủ đề. Khi đó văn bản sẽ được gán cho cụm tương ứng với chủ đề có xác suất cao nhất:

$$x = \arg\max_j \theta_j$$

3.3. Kết hợp LDA và k-means

Cách tiếp cận thứ hai sử dụng mô hình chủ đề chẳng hạn như LDA để ánh xạ biểu diễn với số chiều khá cao của văn bản (đối tượng thuật ngữ (term)) sang biểu diễn có số chiều thấp hơn (đối tượng chủ đề) và sau đó áp dụng thuật toán phân cụm tiêu chuẩn như K-means trong không gian đối tượng mới. Trong công việc của mình, chúng tôi đã sử dụng phân phối chủ đề của mỗi văn bản (document-topic distributions) θ được trích xuất từ LDA làm đầu vào cho các thuật toán phân cụm K-means. Chúng tôi gọi cách tiếp cận này là LDA + K-means.

Một vấn đề đặt ra ở đây là vì đầu vào của chúng ta là một phân phối xác suất, do đó các độ đo trên không gian véc tơ như Euclidean hay Cosine có thể không cho kết quả tốt. Vì vậy cần nghiên cứu việc chọn các độ đo phù hợp trong đó quan tâm đến các độ đo liên quan đến xác suất.

4. VAI TRÒ CỦA ĐỘ ĐO ĐỘ TƯƠNG TỰ LIÊN QUAN ĐẾN PHÂN PHỐI XÁC SUẤT TRONG MÔ HÌNH LDA + K-MEANS

4.1. Một số độ đo độ tương tự liên quan đến phân phối xác suất

Vì LDA biểu diễn cho các văn bản dưới dạng phân phối xác suất, chúng ta cần xem xét cách "tốt" để chọn một thước đo khoảng cách hoặc độ tương tự để so sánh hai phân phối xác suất. Có hai cách tiếp cận trong các phép đo khoảng cách / độ tương tự của hàm mật độ xác suất (pdf): véc tơ và xác suất. Đối với phương pháp tiếp cận theo xác suất, việc tính toán khoảng cách giữa hai pdf có thể được xem giống như tính toán xác suất Bayes (hoặc minimum misclassification). Điều này tương đương với việc đo lường sự chồng chéo giữa hai pdf dưới dạng khoảng cách. Đối với cách tiếp cận này, f-divergence $D_f(P||Q)$ thường được sử dụng để đo sự khác biệt giữa hai phân phối xác suất P và Q. Một cách trực quan, chúng ta có thể xem divergence như một giá trị trung bình, có trọng số bởi hàm f, của tỷ số chênh lệch do P và Q đưa ra.

Định nghĩa 1: Cho $P = \{p_i | i = 1,2, \dots, d)$ và $Q = \{q_i | i = 1,2, \dots, d)$ là hai phân phối xác suất trên không gian xác suất Ω sao cho P là liên tục tuyệt đối với Q. Khi đó, đối với một hàm lồi f sao cho f(1) = 0, f-divergence của Q từ P được định nghĩa:

$$D_f(P||Q) = \sum_{i=1}^{d} q_i f(\frac{p_i}{q_i})$$

Nhiều divergence thông dụng như Kullback-Leibler (KL) divergence ($f(t) = tlnt, t = \frac{p_i}{q_i}$) , Hellinger distance ($f(t) = (\sqrt{t} - 1)^2$, χ^2 divergence ($f(t) = (t-1)^2$ là các trường hợp đặc biệt của f-divergence tương ứng với việc chọn cụ thể hàm f.

Để phân tích hiệu quả của các thước đo khoảng cách dựa trên xác suất, chúng tôi đã chọn ba khoảng cách hoặc divergence được mô tả dưới đây:

4.1.1. Khoảng cách Jensen-Shannon

Đây là một độ đo dựa trên Kullback-Leibler (KL) divergence, liên quan tới khái niệm Shannon về sự không chắc chắn hoặc 'entropy' $H(P) = \sum_{i=1}^{d} p_i \, lnp_i$

$$d_{JS}(P,Q) = \frac{1}{2}\sum_{i=1}^{d} p_i \ln\left(\frac{2p_i}{p_i + q_i}\right) + \frac{1}{2}\sum_{i=1}^{k} q_i \ln(\frac{2q_i}{p_i + q_i})$$

4.1.2. Khoảng cách Bhattacharyya

Đây là độ đo dạng phân kỳ, được định nghĩa như sau:

$$d_{Bhat}(P,Q) = -ln\sum_{i=1}^{d}\sqrt{p_i q_i}$$

4.1.3. Taneja Divergence

Độ đo này là sự kết hợp giữa KL divergence và Bhattacharyya, sử dụng KL-divergence với $p_i = \frac{p_i+q_i}{2}$ và $q_i = \sqrt{p_i q_i}$

$$d_{TJ}(P,Q) = \sum_{i=1}^{d} \frac{p_i + q_i}{2}\ ln\frac{p_i + q_i}{2\sqrt{p_i q_i}}$$

4.2. Phương pháp đánh giá hiệu quả

Với mỗi tập dữ liệu, chúng tôi thu được một kết quả phân cụm từ thuật toán K-means. Để đánh giá hiệu quả của việc phân cụm, chúng tôi sử dụng hai chỉ số đánh giá sau: Adjusted Rand Index (ARI) [9] và Adjusted Mutual Information (AMI) [16], đây là hai chỉ số được sử dụng rộng rãi trong đánh giá hiệu quả của các thuật toán học không giám sát.

4.2.1. Adjusted Rand Index (ARI)

Adjusted Rand Index (ARI) [9] là dạng điều chỉnh của Rand Index (RI), được định nghĩa như sau:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{n_{io}}{2}\sum_j \binom{n_{jo}}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{io}}{2} + \sum_j \binom{n_{jo}}{2}\right] - \left[\sum_i \binom{n_{io}}{2}\sum_j \binom{n_{jo}}{2}\right]/\binom{n}{2}} \quad (1)$$

trong đó n_{ij}, n_{io}, n_{jo} là các giá trị từ Bảng 1.

4.2.2. Adjusted Mutual Information (AMI)

Adjusted Mutual Information (AMI) [16] là một dạng điều chỉnh của mutual information (MI), được định nghĩa như sau:

$$AMI(P,Q) = \frac{MI(P,Q) - E\{MI(P,Q)\}}{\max\{H(P), H(Q)\} - E\{MI(P,Q)\}} \quad (2)$$

trong đó:

$$H(P) = -\sum_{i=1}^{k}\frac{n_{io}}{n} log\frac{n_{io}}{n}; MI(P,Q) = \sum_{i=1}^{k}\sum_{j=1}^{l}\frac{n_{ij}}{n}\log\frac{n_{ij}/n}{n_{io}n_{jo}/n^2}$$

Bảng 1:. Bảng tương quan $n_{ij} = |P_i \cap Q_j|$

P\Q	Q_1	Q_2	...	Q_l	Sum
P_1	n_{11}	n_{12}	...	n_{1l}	n_{1o}
P_2	n_{21}	n_{22}	...	n_{2l}	n_{20}
....	...	...	...	...	...
P_k	n_{k1}	n_{k2}	...	n_{kl}	n_{k0}
Sum	n_{o1}	n_{o2}	...	n_{ol}	$\sum_{ij} n_{ij}$

Cả ARI và AMI đều có cận trên là 1. Giá trị ARI và AMI càng lớn thì thể hiện sự phân vùng càng chính xác. Chúng ta có thể tham khảo các tài liệu [9], [16] để hiểu chi tiết hơn.

4.3. Thực nghiệm và kết quả

4.3.1 Tập dữ liệu

Phương pháp đề xuất được đánh giá trên hai tập dữ liệu khác nhau được cộng đồng NLP sử dụng rộng rãi cho bài toán phân cụm văn bản. Bảng 2 mô tả một vài thống kê về hai tập dữ liệu được sử dụng. Tập dữ liệu *20Newsgroup* bao gồm 18821 văn bản được phân phối trên 20 thể loại tin tức khác nhau. Mỗi văn bản tương ứng với một bài báo gồm tiêu đề, chủ đề, và văn bản được trích dẫn. Tập dữ liệu *WebKB* chứa 8230 trang web từ các khoa khoa học máy tính của các trường đại học (ví dụ: Texas, Wisconsin, Cornell, ...)

Dataset	#Docs	#Classes	< Class	> Class
20NewsGroups	18821	20	628	999
WebKB	8230	4	504	1641

Bảng 2. Thống kê về các tập dữ liệu. #Docs là số lượng văn bản trong tập dữ liệu; #Classes là số cụm văn bản trong một tập dữ liệu; <Class và >Class thể hiện số văn bản nhỏ nhất và lớn nhất trong các cụm.

4.3.2 Cài đặt

Trong các thực nghiệm, chúng tôi so sánh 6 loại độ đo khoảng cách sử dụng cho LDA+K-means được chia làm hai thể loại: Độ đo khoảng cách dựa vào véc tơ (VBM) và Độ đo khoảng cách dựa vào xác suất (PBM). Chúng tôi cài đặt LDA với thuật toán Gibbs sampling sử dụng gói topicmodels trong ngôn ngữ R. Các tham số tiền nghiệm α và β tương ứng với 0.1 và 0.01. Các tham số này được chọn tương ứng với các tiêu chuẩn mới nhất [7]. Số vòng lặp trong thuật toán Gibbs sampling được thiết lập là 5000. Số lượng chủ đề đối với tập dữ liệu *20NewsGroups* được chọn là 30 và đối với *WebKb* là 8. Số lượng chủ đề được xác định thông qua thực nghiệm với việc kiểm tra nhiều giá trị khác nhau. Mỗi độ đo khoảng cách, chúng tôi chạy K-means 20 lần với số vòng lặp lớn nhất là 1000. Chúng tôi tính các chỉ số ARI và AMI dựa trên giá trị trung bình từ các kết quả của mỗi vòng lặp K-means.

4.3.3 Kết quả

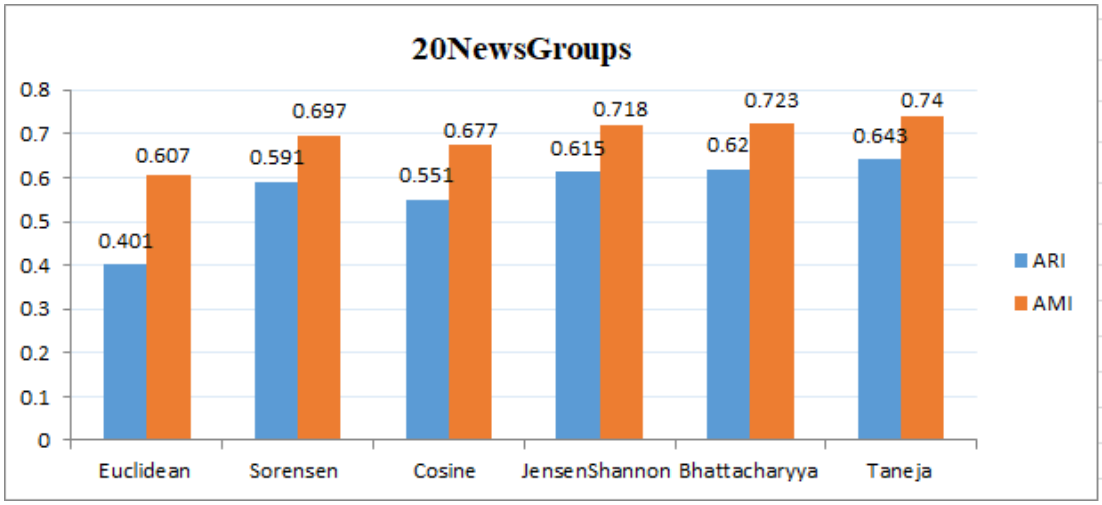
a) So sánh hiệu quả của các độ đo trong việc sử dụng kết hợp LDA + K-means

Các giá trị trung bình của ARI và AMI được trình bày ở Bảng 3. Các giá trị trung bình của ARI và AMI cho nhóm PBM cho kết quả tốt hơn so với nhóm VBM. Chúng ta có thể thấy rằng khoảng cách Euclidean cho kết quả ARI và AMI tệ nhất.

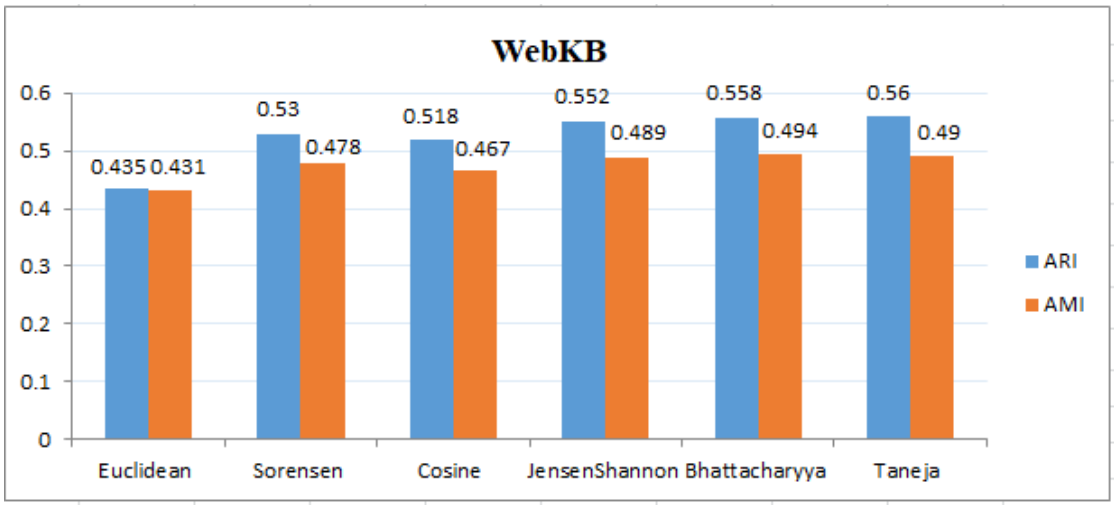
Trong nhóm PBM, mức trung bình tốt nhất đạt được bởi sử dụng hai khoảng cách Bhattacharyya và Taneja. Do đó, chúng tôi đề xuất nên sử dụng hai độ đo này cho LDA + K-means. Kết quả được thể hiện rõ nét hơn thông qua biểu đồ ở Hình 5 đối với tập dữ liệu 20NewsGroup và Hình 6 với tập dữ liệu WebKB.

Distances	20NewsGroups		WebKB	
	ARI	AMI	ARI	AMI
Euclidean	0,401	0,607	0,435	0,431
Sorensen	0,591	0,697	0,530	0,478
Cosine	0,551	0,677	0,518	0,467
JensenShannon	0,615	0,718	0,552	0,489
Bhattacharyya	0,620	0,723	0,558	0,494
Taneja	0,643	0,740	0,560	0,490
VSM	0,127	0,371	0,267	0,334
LDA + Naive	0,435	0,591	0,170	0,198

Bảng 3. Các giá trị trung bình ARI và AMI cho VSM, LDA Naive, LDA + K-means với sáu độ đo khoảng cách cho hai tập dữ liệu 20NewsGroups, WebKB



Hình 5. Các giá trị trung bình ARI và AMI cho LDA + K-meansvới sáu độ đo khoảng cách đối với tập dữ liệu 20NewsGroups



Hình 6. Các giá trị trung bình ARI và AMI cho LDA + K-meansvới sáu độ đo khoảng cách đối với tập dữ liệu WebKB.

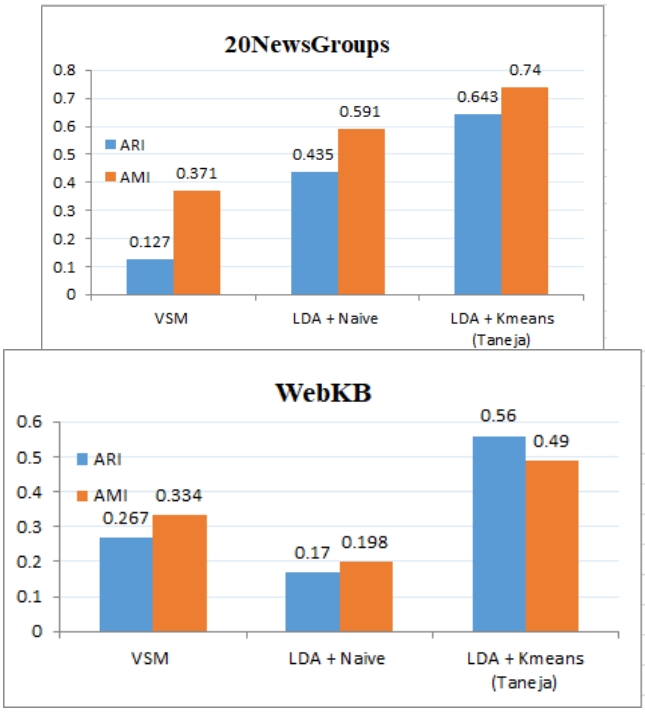
b) So sánh 03 mô hình: LDA + K-Means, LDA + Naive, VSM

Để thể hiện được hiệu quả của mô hình mô phỏng chủ đề, chúng tôi tiến hành so sánh 3 phương thức phân cụm văn bản:

i) Mô hình không gian vector (VSM): sử dụng vector tần suất xuất hiện của các từ $\vec{t}_d = (tf(d,t_1), tf(d,t_2), ..., tf(d,t_m))$ làm đầu vào cho K-means.

ii) LDA +Naïve: một văn bản được gán vào cụm x nếu x = argmax x_j θ_j trong đó $\theta = (\theta_1, \theta_2, ..., \theta_K)$ là phân phối xác suất của văn bản (document-topic distributions) được trích xuất từ LDA

iii) LDA + K-means: dùng kết quả phân phối xác suất của văn bản (document-topic distributions) được trích xuất từ LDA làm đầu vào cho k-means. Ở đây chúng tôi sử dụng khoảng cách Taneja từ kết quả khảo sát ở phần 4.3.3a.



Hình 7. AMI và ARI cho 3 phương thức phân cụm tài liệu: VSM + K-means, LDA + Naive và LDA+K-meansvới khoảng cách Taneja cho 02 bộ dữ liệu.

Kết quả được thể hiện ở biểu đồ Hình 7. Chúng ta thấy rằng LDA + K-means cùng với khoảng cách Taneja đạt được kết quả trung bình AMI và ARI tốt nhất trên cả hai tập dữ liệu.

5. KẾT LUẬN

Trong bài báo này, chúng tôi đã so sánh hiệu quả của việc sử dụng sáu loại độ đo khoảng cách trong bài toán phân cụm văn bản sử dụng kết hợp giữa LDA và K-means. Các thực nghiệm trên hai tập dữ liệu và hai chỉ số đánh giá thể hiện rằng các độ đo dựa vào phân phối xác suất cho kết quả phân cụm tốt hơn so với các độ đo dựa vào véc tơ, bao gồm độ đo Euclidean. Bên cạnh đó, về việc so sánh kết quả của LDA + K-means, LDA + Naive, VSM, các kết quả thực nghiệm cho thấy rằng nếu chúng ta chọn số lượng chủ đề phù hợp cho LDA và một độ đo khoảng cách dựa vào xác suất phù hợp cho thuật toán K-means thì sự kết hợp giữa LDA và K-means sẽ mang lại hiệu quả tốt cho bài toán phân cụm văn bản.

TÀI LIỆU THAM KHẢO

[1]. Blei, D.M.: Probabilistic topic models. Commun. ACM 55(4), 77–84 (2012)

[2]. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022 (2003)

[3]. Bui, Q.V., Sayadi, K., Bui, M.: A multi-criteria document clustering method based on topic modeling and pseudoclosure function. Informatica 40(2), 169–180 (2016)

[4]. Buntine, W.: Estimating likelihoods for topic models. In: Zhou, Z.-H., Washio, T. (eds.) ACML 2009. LNCS (LNAI), vol. 5828, pp. 51–64. Springer, Heidelberg (2009). doi:10.1007/978-3-642-05224-8 6

[5]. Cha, S.-H.: Comprehensive survey on distance/similarity measures between probability density functions. City 1(2), 1 (2007)

[6]. Gordon, A.: Classification. Chapman & Hall/CRC Monographs on Statistics &Applied Probability, 2nd edn. CRC Press, Boca Raton (1999)

[7]. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. U.S.A. 101(Suppl 1), 5228–5235 (2004)

[8]. Huang, A.: Similarity measures for text document clustering. In: Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC

[9]. 2008), Christchurch, New Zealand, pp. 49–56 (2008)

[10]. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. 2(1), 193–218 (1985)

[11]. Lu, Y., Mei, Q., Zhai, C.: Investigating task performance of probabilistic topic models: an empirical study of PLSA and LDA. Inf. Retrieval 14(2), 178–203 (2010)

[12]. Maher, K., Joshi, M.S.: Effectiveness of different similarity measures for text classification and clustering. Int. J. Comput. Sci. Inf. Technol. 7(4), 1715–1720 (2016)

[13]. Manning, C.D., Raghavan, P.: An Introduction to Information Retrieval. Cambridge University Press, Cambridge (2009)

[14]. Modha, D.S., Spangler, W.S.: Feature weighting in k-means clustering. Mach.Learn. 52(3), 217–237 (2003)

[15]. Pestov, V.: On the geometry of similarity search: dimensionality curse and concentration of measure. Inf. Process. Lett. 73(1), 47–51 (2000)

[16]. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5), 513–523 (1988)

[17]. Vinh, N.X., Epps, J., Bailey, J.: Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. J. Mach.Learn. Res. 11, 2837–2854 (2010)

[18]. Xie, Pengtao and Xing, Eric P. : Integrating Document Clustering and Topic Modeling, UAI'13: Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, August 2013, Pages 694–703, 2013.

APPLICATION OF TOPIC MODELING IN DOCUMENT CLUSTERING

Bui Quang Vu, Tran Thien Thanh, Ngo Nhan Duc, Nguyen Hoang Ha, Nguyen Dung

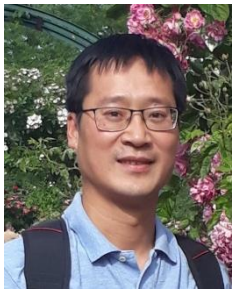
University of Sciences, Hue University

*Email: buiquangvu@hueuni.edu.vn

ABSTRACT

This paper is an experimental study aimed at evaluating the effectiveness of distance measures when using the combined model of LDA and K-means for clustering document. Our experimental results indicate that the probabilistic-based distance measures are better than the vector based distance measures including Euclidian when it comes to cluster a set of documents in the topic space. By choosing the probabilistic-based distance measures, K-means combined to the results of the Latent Dirichlet Allocation allows us to have better results than the LDA + Naïve and Vector Space Model.

Keywords: Topic Modeling, Latent Dirrichlet Allocation, Document clustering, K-means probabilistic-based distance measures.



Bùi Quang Vũ sinh ngày 28/08/1980 tại Thừa Thiên Huế. Ông tốt nghiệp Cử nhân Toán tại trường Đại học Khoa học, Đại học Huế năm 2002. Năm 2007, ông tốt nghiệp thạc sĩ chuyên ngành Lý thuyết Xác suất và Thống kê Toán học tại trường Đại học Khoa học, Đại học Huế. Năm 2018, ông nhận học vị tiến sĩ tại Université Paris Sciences et Lettres (PSL) chuyên ngành Khoa học máy tính, Thống kê và Nhận thức. Hiện nay ông đang công tác tại Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Xác suất Thống kê, Học máy, Trí tuệ nhân tạo

Trần Thiện Thành sinh ngày 01/01/1983 tại Thừa Thiên Huế. Ông tốt nghiệp Cử nhân Toán tại trường Đại học Khoa học, Đại học Huế năm 2005. Năm 2008, ông tốt nghiệp thạc sĩ chuyên ngành Lý thuyết Xác suất và Thống kê Toán học tại trường Đại học Khoa học, Đại học Huế. Hiện nay ông đang công tác tại Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Xác suất Thống kê, Khoa học dữ liệu.

Ngô Nhân Đức sinh ngày 15/12/1980 tại Thừa Thiên Huế. Ông tốt nghiệp Cử nhân Toán tại trường Đại học Khoa học, Đại học Huế năm 2002. Năm 2008, ông tốt nghiệp thạc sĩ chuyên ngành Lý thuyết Tối ưu tại trường Đại học Khoa học, Đại học Huế. Hiện nay ông đang công tác tại Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Lý thuyết tối ưu, Trí tuệ nhân tạo

Nguyễn Hoàng Hà sinh ngày 22/11/1976 tại Thăng Bình, Quảng Nam. Năm 1999, ông tốt nghiệp đại học ngành Công nghệ Thông tin tại trường Đại học Khoa học, ĐH Huế; ông nhận bằng thạc sỹ chuyên ngành Khoa học Máy tính tại Trường Đại học Khoa học, Đại học Huế năm 2005. Năm 2017, ông tốt nghiệp tiến sĩ chuyên ngành Khoa học máy tính tại trường Đại học Khoa học, Đại học Huế. Hiện nay, ông công tác tại Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Xử lý song song và phân tán, tính toán lưới và tính toán đám mây.

Nguyễn Dũng sinh ngày 13/06/1988 tại Thừa Thiên Huế. Ông tốt nghiệp cử nhân Tin học tại trường Đại học Khoa học, Đại học Huế năm 2010. Năm 2013, ông tốt nghiệp thạc sĩ chuyên ngành Khoa học máy tính tại trường Đại học Khoa học, Đại học Huế. Hiện nay ông đang công tác tại Trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Công nghệ phần mềm, trí tuệ nhân tạo, học máy, học sâu, cơ sở dữ liệu