Gom Nhóm Văn Bản và Thuật Toán Áp Dụng

Nguyễn Lương Điền

Trường Đại học Khoa học Tự nhiên

TP HCM, Việt Nam

[tom.nguyenluong@gmail.com](mailto:tom.nguyenluong@gmail.com)

Mai Khải Huy

Trường Đại học Khoa học Tự nhiên

TP HCM, Việt Nam

[maikhaihuy@gmail.com](mailto:maikhaihuy@gmail.com)

Tóm tắt – Gom nhóm văn bản nhằm mục đích khám phá bản chất của các thông tin được ẩn chứa trong văn bản. Dữ liệu kiểu văn bản không giống như dữ liệu số hay hình ảnh, thường phải trải qua giai đoạn tiền xử lý rất công phu và cần chuyển đổi sang dạng thức phù hợp để khai thác, thường là Vector Space Model. Có 3 phương pháp phân cụm thường được sử dụng là connectivity-based clustering, centroid-based clustering, density-based clustering. Mỗi phương pháp đều có những thế mạnh riêng, mang lại nhiều giá trị ứng dụng trong thực tiễn.

Chỉ mục – K-means, Agglomerative, OPTICS, tf-idf, Vector Space Model.

# Giới thiệu:

Gom nhóm là một bài toán điển hình trong Khai thác dữ liệu, thuộc về nhóm phương pháp *học không giám sát* (Unsupervised Learning) trong lĩnh vực Học máy. Khác với dữ liệu số và hình ảnh, dữ liệu dạng văn bản mang nhiều đặc trưng của con người hơn, do đó trong máy học phải có những phương pháp tiếp cận đặc biệt để xử lý và khai thác. Vector space model cho ta một phương thức chuyển đổi văn bản từ phi số sang dạng đặc trưng số, từ đó quy về các bài toán gom cụm truyền thống, tận dụng được các thành tựu nghiên cứu khoa học trong suốt quá trình dài trước đó, vừa làm tăng được tính tổng quát, vừa giúp giải quyết được bài toán một cách hiệu quả.

Trong bối cảnh tầm ảnh hưởng của học máy ngày càng trở nên to lớn trên nhiều mặt của đời sống xã hội, khoa học về xử lý và khai thác thông tin hứa hẹn cũng sẽ có được những bước phát triển vượt bậc, góp phần làm cho cuộc sống của con người ngày một tốt đẹp hơn.

# Gom nhóm văn bản:

## Mục tiêu:

Mục tiêu của gom nhóm văn bản để phân nhóm các đoạn văn bản không cấu trúc. Mỗi văn bản có các đặc trưng về ngữ nghĩa, các văn bản trong cùng một nhóm sẽ có đặc trưng giống (tương tự) nhau. Sau quá trình gom nhóm, các đặc trưng này sẽ trở nên cụ thể hơn, từ đó làm cơ sở cho việc phân tích nội dung trên các văn bản khác.

Khi dữ liệu các văn bản được phân nhóm, quá trình tra cứu, quản lý số lượng rất nhiều các văn bản được lưu trữ lớn trên các hệ thống trở nên dễ dàng, nhanh chóng.

## Vấn đề xử lý ngôn ngữ tự nhiên

Dân số thế giới hiện nay vào khoảng 7 tỷ người, phân bố trên 5 châu lục, bao gồm hơn 200 quốc gia. Hầu hết mỗi quốc gia đều có một ngôn ngữ riêng cho mình, chưa kể đến sự phân hóa ngôn ngữ theo vùng miền trong chính quốc gia đó, kéo theo số lượng vô cùng lớn, lên đến hàng nghìn thứ tiếng. Hầu hết trong số chung được sử dụng một cách rất hạn chế, gần như bị tiêu vong. Phần lớn trong số chúng thậm chí còn chưa có chữ viết chính thức để sử dụng.

Chỉ có khoảng 500 ngôn ngữ đã được nghiên cứu một cách tương đối đầy đủ. Tuy nhiên đây cũng vẫn là một số lượng lớn. Mỗi ngôn ngữ có các đặc trưng riêng về từ ngữ và ngữ pháp, đòi hỏi những phương pháp tiếp cận đặc thù. Ví dụ như trong tiếng Anh, thường thấy các hiện tượng đảo ngữ, các biến thể của từ gốc tùy thuộc vào ngữ cảnh; trong tiếng Việt, các từ đơn đôi lúc lại không mang một ý nghĩa cụ thể chính xác nào, cần phải xem xét trong mối quan hệ với từ đơn khác để hiểu rõ được ý nghĩa. Điều này đặt ra thách thức không nhỏ trong khâu xử lý đối với khoa học máy tính.

Trong phạm vi tài liệu này, ngôn ngữ được sử dụng để khai thác là tiếng Anh. Đây là một ngôn ngữ vô cùng phổ biến và đã được nghiên cứu vô cùng sâu sắc. Để sử dụng được dữ liệu tiếng Anh vào khai thác dữ liệu, người ta đề xuất ra một số bước xử lý như sau, đã được kiểm chứng là có hiệu quả tốt:

* Loại bỏ stopword: stopword là những từ có tần số xuất hiện rất lớn trong tiếng Anh và thường có giá trị thấp về mặt ngữ nghĩa. Chúng có thể là các mạo từ, trạng từ, đại từ, giới từ, etc. Việc lại bỏ các từ này giúp giảm khối lượng công việc, làm tăng sự hiệu quả cho khâu khai thác.
* Stemming: Các từ tiếng Anh thường có rất nhiều biến thể, được sử dụng tùy theo ngữ cảnh khác nhau, tuy nhiên về mặt ý nghĩa thì lại tương tự nhau. Việc chuyển đổi các biến thể này về cùng một dạng gốc (origin) giúp việc xem xét sự phân bố và đóng góp về mặt nội dung của chúng được đầy đủ và chính xác hơn.

Ngoài ra, một số phương pháp xử lý tổng quát khác như tách từ, loại bỏ dấu câu và các kí tự đặc biệt cũng được sử dụng cho bước tiền xử lý văn bản.

## Vec-tơ đặc trưng:

Văn bản phi cấu trúc mang các tính chất gần với tự nhiên, gần với con người hơn, do đó việc áp dụng các phương pháp học máy để xử lý trên những đặc trưng nguyên thủy này là rất khó khăn, và thường mang lại hiệu quả không cao. Do đó, chuyển đổi văn bản phi cấu trúc sang một dạng đặc trưng khác trước khi thực hiện quá trình khai thác là một công việc vô cùng cần thiết. Để phân nhóm cho các văn bản, đầu tiên cần rút ra được các “mẫu đặc trưng” (từ khóa) trong các văn bản. Các từ khóa này thường là các từ đơn có tần số xuất hiện nhiều (từ 3% - 50%), trong đó không bao gồm các stopword. Tập hợp các từ khóa này được gọi là Bag of Words.

Các văn bản sau đó sẽ được biểu diễn theo danh sách các từ khóa này. Có hai mô hình phổ biến thường được sử dụng:

* Mô hình Boolean: mỗi văn bản được tổ chức theo tập hợp các từ theo đúng vị trí trong Bag of Words.

Gọi *V* = {*t*0, *t*1,..., *tm*} là tập hợp các từ của tập các văn bản *D*.

Gọi *di* = {*w*0, *w*1…*wm*} (*di* thuộc *D*) là tập hợp các trọng số của văn bản *di*. *wk* có giá trị 0 khi văn bản *di* không chứa *tk*, có giá trị 1 nếu văn bản có chứa ­*tk*.

Mô hình boolean chỉ có giá trị về mặt định tính (*có* hoặc *không* tương ứng với 1 hoặc 0), do đó không thể hiện được giá trị định lượng về sự đóng góp vào nội dung của từ trong văn bản (tần số), làm cho việc xử lý vẫn gặp rất nhiều khó khăn và thiếu chính xác, vì thế mô hình này hầu như không được áp dụng vào thực tế.

* Vector space: tương tự như mô hình Boolean, tuy nhiên, các trọng số là các giá trị liên tục. Các giá trị trọng số thường được sử dụng là TF và TF-IDF:
  + TF: giá trị thể hiện cho tần số xuất hiện của mẫu *t* trong văn bản *d*.
  + TF-IDF: giá trị được tính bằng tích của 2 hệ số TF và IDF. IDF thể hiện cho tần số văn bản *d* có chứa mẫu *t* trong tập hợp tất cả các văn bản được khai thác. Tích giữa 2 giá trị này là trọng số để biểu diễn văn bản trong không gian vector đặc trưng.
  + Ngoài cách tính trọng số như trên, người ta còn sử dụng nhiều biến thể tùy vào các mục đích khác nhau, được liệt kê trong bảng sau:

|  |  |
| --- | --- |
| **Term frequency (TF)** | **Document frequency (DF)** |
| tf*t*, *d* | n (no of document) |
|  |  |
|  |  |
|  |  |
|  |  |

Bảng 1 – các biến thể trong tính toán trọng số

## Các độ đo tương tự:

Để xác định 2 văn bản có cùng thuộc một nhóm hay không, ta cần một giá trị định lượng thể hiện cho mối quan hệ tương quan giữa chúng. Giá trị này được gọi là độ tương tự giữa 2 văn bản.

Sau khi đã biểu diễn văn bản dưới dạng vector hệ số đặc trưng tf-idf, một phương pháp tính sẽ được áp dụng dựa trên những giá trị này để xác định độ tương tự giữa 2 văn bản. Trong không gian, mối quan hệ giữa 2 vector thường được định lượng bằng khoảng cách hoặc góc giữa 2 vector đó.

1. Khoảng cách

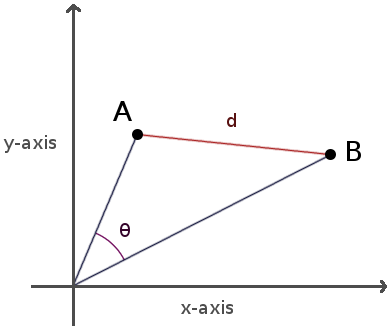
Độ đo thường được sử dụng là độ đo Euclidean. Khoảng cách giữa 2 vector được đo bằng độ dài khoảng cách giữa 2 điểm cuối tương ứng.



Phương pháp này chỉ có hiệu quả khi số chiều không gian không quá lớn. Trong trường hợp gom cụm văn bản, không gian vector thường có số chiều rất lớn, các điểm phân bố thưa thớt và cách xa nhau, dẫn đến khoảng cách giữa chúng cũng rất lớn, cho dù giữa chúng chỉ có sự khác biệt trên một số ít chiều nhất định.

1. Góc

Việc sử dụng khoảng cách không mang lại hiệu quả cao trong gom nhóm văn bản, thay vào đó người ta thường sử dụng góc giữa 2 vector để đo độ tương tự giữa chúng.



Trong trường hợp 2 vector có cùng phương (chỉ khác nhau về một thành phần nào đó), khoảng cách Euclidean giữa chúng có thể rất lớn trong khi chúng gần như tương tự nhau. Khi đó, sử dụng độ đo góc (trong trường hợp này có giá trị 0) giúp ta xác định được độ tương tự chính xác hơn hẳn. Độ đo thường được sử dụng là cosine.



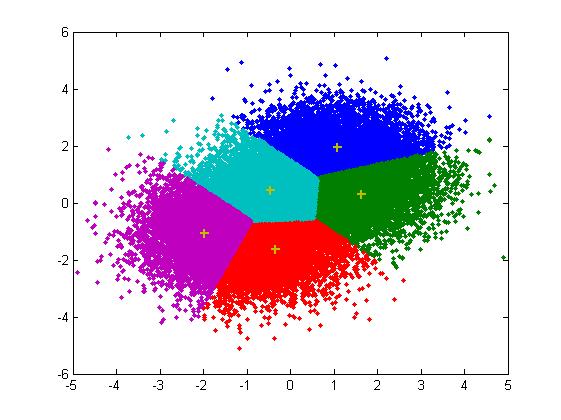
Ngoài ra, còn có một số độ đo khác như Jaccard (theo xác suất), Hamming, Tanimoto nhưng ít được sử dụng.

# Các thuật toán gom nhóm:

## Chameleon:

## K-medoids:

K-medoids thuộc nhóm phương pháp gom nhóm dựa theo trọng tâm. Đây là một phương pháp khá đơn giản và được sử dụng nhiều, có hiệu quả tốt nhất trong trường hợp dữ liệu phân bố dạng hình cầu.



Hình 3.2 – Minh họa thuật toán k-means

Tư tưởng chính của thuật toán là phân chia tập dữ liệu thành k nhóm sao cho các thành phần thuộc cùng một nhóm có độ tương tự cao và các thành phần khác nhóm có độ tương tự thấp.

1. Thuật toán:

Với k cho trước, các nhóm được đại diện bởi một trung tâm (thường mang giá trị trung bình hoặc trung vị).

* Bước 1: Chọn ngẫu nhiên k điểm bất kì làm trung tâm của mỗi nhóm.
* Bước 2: Sử dụng độ đo tương tự, gán các điểm vào các nhóm sao cho khoảng cách từ điểm đến trung tâm là thấp nhất.
* Bước 3: Lựa chọn trung vị của mỗi nhóm để làm trung tâm mới.
* Bước 4: Khi các trung tâm nhóm không có gì thay đổi so với lần lặp trước thì dừng thuật toán. Ngược lại quay trở về bước 2.

1. Ví dụ:

Cho các điểm dữ liệu A1(2, 10), A2(2, 5), A3(8, 4), A4(5, 8), A5(7, 5), A6(6, 4). Phân chia các điểm dữ liệu này vào 3 nhóm.

* Khởi tạo hạt giống. Giả sử ta khởi tạo được 3 trung tâm là C1(2, 5), C2(5, 8), C3(6, 4).
* Tính khoảng cách euclidean từ 3 trung tâm đến mỗi điểm và gán vào các nhóm, ta được kết quả: 1: {A2}, 2: {A1, A4}, 3: {A3, A5, A6}.
* Cập nhật lại các trung tâm bằng các giá trị trung vị. Đối với nhóm có 2 phần tử thì chọn ngẫu nhiên 1 trong 2 để làm giá trị trung vị: C1(2, 5), C2(5, 8), C3(7, 5).
* Thực hiện gắn nhãn với các trung tâm mới, ta được kết quả: 1: {A2}, 2: {A1, A4}, 3: {A3, A5, A6}.
* Do không có sự thay đổi nhóm ở lần lặp thứ 2, thuật toán kết thúc.

## OPTICS:

# Kết quả thực nghiệm

# Các ứng dụng

# Kết luận