ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

DƯƠNG XUÂN LONG

GOM NHÓM TIN TỨC TIẾNG VIỆT CÓ CÙNG CHỦ ĐỀ

ĐỒ ÁN THẠC SĨ: CÔNG NGHỆ THÔNG TIN

Tp. Hồ Chí Minh, Năm 2015

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

DƯƠNG XUÂN LONG

GOM NHÓM TIN TỨC TIẾNG VIỆT CÓ CÙNG CHỦ ĐỀ

Chuyên ngành: KHOA HỌC MÁY TÍNH
Mã số chuyên ngành: 60.48.01.01

ĐỒ ÁN THẠC SĨ: CÔNG NGHỆ THÔNG TIN

NGƯỜI HƯỚNG DẪN KHOA HỌC:
TS. Nghiêm Quốc Minh

Tp. Hồ Chí Minh, Năm 2015

Lời cảm ơn

Tôi chân thành cảm ơn khoa công nghệ thông tin, trường đại học khoa học tự nhiên, đại học quốc gia Tp. Hồ Chí Minh đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập và thực hiện đồ án tốt nghiệp.

Tôi nói lên lòng biết ơn sâu sắc đối với thầy Nghiêm Quốc Minh đã luôn quan tâm và tận tình hướng dẫn em trong quá trình học tập, nghiên cứu và thực hiện đề tài.

Chân thành cảm ơn quý thầy cô trong khoa công ghệ thông tin đã tận tình giảng dạy, trang bị cho em những kiến thức quý báu trong những năm học vừa qua.

Gửi lòng biết ơn đến thầy cô và bạn bè trong lớp đã giúp đỡ, động viên tinh thần em rất nhiều trong suốt quá trình thực hiện đồ án này.

Mặc dù đã cố gắng hoàn thành đồ án trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót, kính mong nhận được sự góp ý và tận tình chỉ bảo của quý thầy cô và các bạn.

Một lần nữa, tôi chân thành cảm ơn và mong luôn nhận được những tình cảm chân thành của tất cả mọi người.

Mục lục

Lời cảm ơn **ii**

Mục lục **iii**

Danh sách hình **iv**

Danh sách bảng **v**

Danh sách thuật ngữ **vi**

Tổng quan đề tài **1**

1 Gom nhóm văn bản **4**

1.1 Giới thiệu về gom nhóm 4

1.2 Gom nhóm văn bản 5

1.3 Các phương pháp gom nhóm 7

1.4 Mục tiêu đồ án 9

2 Phương pháp gom nhóm phân cấp **11**

2.1 Hướng tiếp cận agglomerative 11

2.2 Hướng tiếp cận divisive 11

2.3 Đồng gom nhóm phân cấp 11

2.3.1 Giới thiệu 11

2.3.2 Độ tương đồng Goodman-Kruskal τ 14

3 Phương pháp đề xuất **17**

3.1 Goodman-Krusal τ 17

3.2 Goodman-krusal trong gom nhóm phân cấp 17

4 Thực nghiệm và kết quả **18**

4.1 Dữ liệu 18

4.2 Phương pháp đánh giá 19

4.2.1 Normalized Mutual Information - NMI 20

4.2.2 Adjusted Rand Index - ARI 21

4.3 Kết Quả 21

5 Kết luận và hướng phát triển **22**

Tài liệu tham khảo **23**

Danh sách hình

Danh sách bảng

2.1 Bảng dữ liệu giữa *Salary* và *Job* 14

4.1 Bảng dữ liệu Reuters 19

Danh sách thuật ngữ

1. Clustering - Cách phân nhóm
2. High-dimensional - Dữ liệu có số chiều cao
3. Curse of dimensionality - Lời nguyền về số chiều
4. Cosine similarity - Độ tương đồng cosine
5. Feature selection - Lựa chọn đặc trưng
6. Co-clustering - Đồng gom nhóm
7. Keyword - Từ khóa chính trong một văn bản
8. Stopword - Từ phổ biến trong một ngôn ngữ nhưng mang ít thông tin

Tổng quan

Ngày nay, internet đang phát triển ở một tốc độ chóng mặt dẫn đến thông tin được phổ cập đến người dùng diễn ra một cách nhanh chóng, khác hoàn toàn với thời trước khi mà chúng ta cần đến báo chí, đài phát thanh để biết tin tức. Ngày nay, mọi thông tin đều hiện diện trên internet. Có thể nói, internet chứa nguồn tài nguyên vô hạn không khác gì một kho tàng bất tận cho những ai biết tận dụng nó.

Trong nguồn tài nguyên vô hạn đó thì tin tức là một tài nguyên mà được nhiều người tìm đến nhất. Có lẽ ngày nay việc chúng ta hay đọc báo trên mạng không không còn là chuyện mới mẻ gì nữa. Và cũng ngày càng có nhiều người lựa chọn hình thức đọc báo trên mạng. Điều đó góp phần thúc đẩy báo mạng phát triển và ngày càng có nhiều trang báo mạng được ra đời như nấm mọc sau mưa.

Chính tốc độ phát triển như vũ bão của báo mạng đang đe dọa đến sự tồn tại của ngành công nghiệp báo chí truyền thống. Chính vì thế mà chúng ta thấy được một cuộc cách mạng khi mà các tờ báo giấy đang chuyển mình thành các trang báo mạng. Và không chỉ có sự tham gia của các tờ báo truyền thống, các đại gia trong ngành công nghệ thông tin cũng gia nhập trào lưu này như Facebook, Google và Apple khiến cho cuộc đua tranh ở lĩnh vực này đang ngày một khốc liệt. Facebook thì tích hợp hệ thống đọc tin tức Feed ngay trên trang mạng xã hội của mình. Google thì cho ra mắt hệ thống đọc tin tức Google news. Còn Apple thì mới tích hợp Apple news ngay trong iOS_9.

Việc các đại gia trong ngành công nghệ thông tin đua nhau vào mảng nội dung, tin tức trên mạng cho thấy sắp tới đây sẽ là lĩnh vực đầy tìm năng và hứa hẹn. Tuy nhiên, việc chỉ có tin tức hay nội dung không thì vẫn là chưa đủ, người dùng ngoài việc đọc báo ra còn muốn theo dõi các tin tức cũng như là các nội dung liên quan có cùng chủ đề. Tin tức trên mạng thì tràn lan đại khải, một số người thì chỉ quan tâm đến chủ đề này, còn một số khác thì quan tâm đến chủ đề khác. Cho nên việc cho người dùng lựa chọn các bài báo mà có chủ đề mà mình yêu thích là nhu cầu thiết thực và hợp lí. Đề làm được như vậy, chúng ta cần phân loại văn bản hay bài báo để giúp cho người dùng dễ theo dõi các tin tức mà mình mong muốn.

Vì thế mà em đã quyết định chọn bài toán phân loại văn bản để thực hiện đồ án tốt nghiệp của mình.

Phân loại là một tiến trình tổ chức dữ liệu thành những nhóm mà mỗi nhóm có các đối tượng có độ tương đồng cao. Đây là công việc chính trong khai thác dữ liệu, và là kỹ thuật chung cho trong phân tích dữ liệu thống kê được sử dụng trong nhiều lĩnh vực khác nhau như là máy học, nhận dạng mẫu, phân tích ảnh và truy vấn thông tin ...

Phân loại cũng đồng thời giúp các nhà nghiên cứu thị trường phát hiện ra được những nhóm khách hàng riêng biệt thông qua quá trình mua hàng và từ đó giúp cho các nhà nghiên cứu thị trường có thể đặc trưng hóa nhóm khách hàng này. Ngoài ra, trong lĩnh vực sinh học, phân loại cũng có thể được sử dụng để dẫn xuất gen của cây cối, động vật thành những nhóm có cùng chức năng để có thể hiểu được hoạt động tổng quan của chúng. Mục đích của việc phân loại tổ chức dữ liệu thành nhóm để thể hiện cấu trúc bên trong của dữ liệu và đôi khi việc chia cắt dữ liệu cũng là mục đích chính. Ngoài ra thì phân loại cũng là bước chuẩn bị cho kỹ thuật AI(tóm tắt văn bản).

Vấn đề của phân loại là tìm được những đối tượng tương đồng, nhưng làm cách nào để có thể tìm được những đối tượng đó? Khi nói đến đây chúng ta sẽ

liên tưởng đến việc tính khoảng cách của các đối tượng. Việc tính toán đó sẽ giúp chúng ta tìm được những đối tượng gần nhau để từ đó có thể dễ dàng xếp thành những nhóm có đặc trưng gần giống nhau. Chúng ta thấy là phân loại được áp dụng cho nhiều lĩnh vực khác nhau nên khái niệm khoảng cách cũng được hiểu theo tùy ngữ cảnh. Và chính vì có quá nhiều lĩnh vực áp dụng phân loại nên em sẽ không thể nào nghiên cứu hết được. Ở đây em chọn lĩnh vực là phân loại văn bản cho đồ án tốt nghiệp của mình.

Sau đây, em giới thiệu sơ bộ chác chương trong đồ án của mình:

Chương 1 - Giới thiệu

Chương 2 - Các phương pháp phân loại

Chương 3 - Phương pháp đề xuất

Chương 4 - Thực nghiệm và kết quả

Chương 5 - Kết luận và hướng phát triển

Chương 1

Gom nhóm văn bản

1.1 Giới thiệu về gom nhóm

Vấn đề gom nhóm đã được nghiên cứu rộng rãi trong cơ sở dữ liệu và thống kê trong khai thác dữ liệu. Công việc của gom nhóm là tìm những nhóm gồm các đối tượng giống nhau trong dữ liệu. Độ tương đồng giữa các đối tượng được đo bằng cách sử dụng hàm tương đồng. Gom nhóm dữ liệu rất hữu dụng trong văn bản, khi được gom nhóm từ các đối tượng khác nhau như : văn bản, đoạn văn bản, câu hoặc là từ. Đặt biệt, gom nhóm rất hữu dụng trong tổ chức văn bản đề cải thiện truy xuất văn bản.

Nghiên cứu của vấn đề gom nhóm đứng trước tính khả dụng khi ứng dụng vào văn bản. Các phương pháp truyền thống cho gom nhóm thường tập trung vào dữ liệu lớn, khi thuộc tính của dữ liệu là số. Vấn đề này cũng được nghiên cứu trong phân loại dữ liệu, khi mà thuộc tính có giá trị nặc danh. Vấn đề của gom nhóm là tìm được tính khả dụng trong các nhiệm vụ sau :

- Duyệt và tổ chức văn bản : tổ chức phân cấp của văn bản vào trong các hạng mục mạch lạc. Điều này có thể giúp ích cho việc duyệt hệ thốn của tập hợp văn bản. Ví dụ kinh kiển cho phương pháp này là Scatter/Gather. Phương pháp này cung cấp kỹ thuật duyệt hệ thống với sử dụng gom nhóm tổ chức của tập hợp văn bản.

- Tóm tắt corpus : kỹ thuật gom nhóm cung cấp tóm tắt mạch lạc của tập hợp trong dạng nhóm tài tiệu hoặc nhóm từ. Thứ này được sử dụng đề cung cấp tóm tắt trong phần nội dung tổng kết của corpus căn bản. Lính vực này có nhiều phương pháp, đặc biệt là gom nhóm câu dùng để tóm tắt văn bản. Vấn đề của gom nhóm liên quan đến việc giảm số chiều và mô hình hóa chủ đề.

- Phân loại văn bản : Gom nhóm là phương pháp học không giám sát. Nó thể được đòn bẩy hóa để cải thiện chất lượng kết quả trong giám sát. Cụ thể, các nhóm từ và phương thức đồng huấn luyện có thể được sử dụng để cải thiện độ chính xác phân loại của ứng dụng giám sát với tác dụng của kỹ thuật gom nhóm.

1.2 Gom nhóm văn bản

Một tài liệu văn bản có thể được biểu diễn dưới dạng nhị phân. Khi đó, chúng ta sử dụng sự hiện diện theo thứ tự của từ để tạo thành vector nhị phân. Các phương pháp chung của các thuật toán như : $K - means$, phân cấp được sẻ dụng cho bất kì loại dữ liệu nào, bao gồm cả dữ liệu văn bản. Trong nhiều trường hợp, chúng ta sử dụng nhiều thuật toán gom nhóm phân loại dữ liệu trong thể hiện nhị phân. Trọng số của từ là dựa vào tần số xuất hiện trong toàn bộ tập hợp. Thuật toán gom nhóm dữ liệu có thể kết nối tần số của từ để tạo ra nhóm liên quan.

Tuy nhiên, các kỹ thuật ngây thơ thường không hiệu quả cho gom nhóm dữ liệu văn bản. Điều này là vì dữ liệu văn bản có một số thuộc tính độc nhất. Cho nên, các thuật toán đòi hỏi cần thiết kế đặc biệt cho nhiệm vụ này. Những đặc trưng riêng biệt của biểu diễn văn bản như sau :

- Số chiều của biểu diễn văn bản là rất lớn, nhưng cơ bản là dữ liệu thì thưa thớt. Nói cách khác, vốn từ trong dữ liệu có thể là 10^5 nhưng một tài liệu

thì chỉ có khoảng vài trăm từ. Vấn đề này càng trở nên nghiêm trọng hơn khi các tài liệu được gom nhóm lại quá ngắn.

- Trong khi vốn từ cho sẵn của các tài liệu có thể rất lớn, những từ kinh điển liên kết với từ khác. Điều này có nghĩa là số lượng của niệm(hoặc là thành phần cơ bản) trong dữ liệu nhỏ hơn không gian đặc trưng. Điều đó đòi hỏi thuật toán cần thiết kế cẩn thận đề quan tâm đến mối liên kết từ trong tiến trình go nhóm.

- Số lượng từ (hoặc là khác không) trong nhiều tài liệu khác nhau là khác biệt lớn. Vì thế, điều quan trọng là phải chuẩn hóa các thể hiện của văn bản thích hợp trong suốt nhiệm vụ gom nhóm.

Sự thưa thớt và số chiều cao trong thể hiện của nhiều tài liệu văn bản là vấn đề cần được quan tâm. Vì vậy, thuật toán được đòi hỏi thiết kế đặc thù cho thể hiện của văn bản. Nghiên cứu về chủ đề tối ưu tính thể hiện của văn bản đưa ra nhiều kỹ thuật để cải thiện truy vấn văn bản. Hầu hết những kỹ thuật này cũng có thể cải thiện thể hiện cua văn bản cho vấn đề gom nhóm.

Để tăng hiệu quả gom nhóm, tần số của từ cần được chuẩn hóa trong toàn dữ liệu. Nhìn chung, biểu diễn văn bản bằng TF-IDF là cách phổ biến. Trong TF-IDF, tần số của từ được chuẩn hóa bằng tần số văn bản đảo ngược(IDF). Sự chuẩn hóa tần số văn bản đảo ngược giảm trọng cho từ mà hay xuất hiện trong tập hợp. Điều này giảm tầm quan trọng của từ thường, tăng tác động của từ tách biệt.

Ngoài ra, hàm chuyển đổi tuyến tính phụ được áp dụng cho tần số từ. Điều này giúp tránh việc ảnh hưởng của một từ quá phổ biến trong văn bản. Công việc chuẩn hóa văn bản chính bản thân nó đã là nhánh nghiên cứu rât lớn. Vì vậy, nhiều kỹ thuật khác nhau dành cho việc chuẩn hóa đã ra đời.

1.3 Các phương pháp gom nhóm

Phương pháp phân chia miền

Phương pháp phân chia miền được thực hiện bằng cách tái phân bổ các đối tượng. Bắt đầu từ miền khởi tạo, phương pháp này dịch chuyển đối tượng từ phân nhóm này sang phân nhóm khác. Phương pháp này thường đòi hỏi số lượng phân nhóm phải được thiết lập bởi người dùng. Do số lượng phân nhóm được thiết lập bởi ban đầu nên kết quả thường không được tối ưu. Vì vậy, phương pháp này cần một quá trình liệt kê đầy đủ tất cả các phân vùng có thể.

Nhưng việc liệt kê đầy đủ tất cả các phân vùng có thể không phải là điều khả thi. Thay vào đó, các thuật toán heuristics được sử dụng để có hiệu quả tương đương. Các thuật toán heuristics thường được áp dụng trong các hình thức tối ưu hóa vòng lặp. Nói cách khác, việc tái phân bổ các đối tượng theo vòng lặp để phân phối các các đối tượng trong k phân nhóm.

Giả sử chúng ta có một cơ sở dữ liệu có n đối tượng. Mục tiêu là phân chia n đối tượng vào k miền cho trước. Và mỗi một miền được chia thể hiện một phân nhóm và $k \leq n$. Điều đó có nghĩa là sẽ phân loại dữ liệu vào k nhóm dể thỏa mãn các điều sau:

- Mỗi nhóm có ít nhất một đối tượng.
- Mỗi đối tượng phải nằm trong một nhóm duy nhất.

Phương pháp dựa vào mật độ

Phương pháp này phân loại dựa vào mật độ của các điểm. Ý tưởng cơ bản là cứ tiếp tục phát triển phân nhóm cho đến mật độ của hàng xóm vượt ngưỡng. Nghĩa là mỗi điểm thuộc về một phân nhóm, bán kính của phân nhóm phải chứa ít nhất một số điểm. Các phân nhóm có mật độ dày đặc được chia cắt bởi các phân nhóm có mật độ thấp. Mật độ các điểm của phân nhóm tạo thành nên các hình thù ngẫu nghiên của phân nhóm đó.

Phương pháp dựa vào lưới tọa độ

Đây là thuật toán sử dụng mạng lưới cấu trúc dữ liệu đa phân giải. Thông thường, bài toán phân loại trên dữ liệu lớn có độ tính toán phức tạp cao. Tuy nhiên, thuật toán này có ưu điểm lớn là giảm được độ phức tạp khi tính toán, đặc biệt là dữ liệu lớn. Hướng tiếp cận của thuật toán này cũng cũng khác so với các thuật toán phân loại thường gặp. Thuật toán này không tập trung vào điểm dữ liệu mà vào giá trị xung quanh điểm dữ liệu.

Phương pháp dựa vào mô hình

Đây là phương pháp dựa vào giả thiết dữ liệu được tạo ra bởi sự pha trộn của các phân phối xác suất. Trong phương pháp này, một mô hình được giả thiết cho mỗi phân nhóm đề tìm sự thích hợp tốt nhất của dữ liệu cho mô hình đã cho. Hay nói cách khác, thuật toán cố gắng tối ưu độ tương thích giữa dữ liệu và mô hình toán học. Phương pháp này định vị phân nhóm bằng hàm mật độ phân loại. Đồng thời, nó phản ánh phân phối của không gian của những điểm dữ liệu.

Phương pháp phân cấp

Gom nhóm phân cấp là phương pháp xây dựng nên cấu trúc phân cấp của các nhóm. Phương pháp gom nhóm này có hai cách tiếp cận là agglomerative

và divisive. Kết quả sau khi gom nhóm thường là đồ thị hình cây thể hiện cấu trúc của các nhóm. Quá trình phân chia hay gộp lại cần có phương pháp đo lường(Chương 2). Trong hầu hết các phương pháp, khoảng cách đo gữa các điểm trong không gian thường được sử dụng.

Việc lựa chọn cách tính khoảng cách sẽ ảnh hưởng đến việc gom nhóm. Công thức khoảng cách có thể khiến cho một đối tượng gần với đối tượng này hoặc xa rời đối tượng khác. Ví dụ về tính khoảng cách cho 2 điểm $(1, 0)$ và $(0, 0)$ trong không gian hai chiều. Công thức định mức thường thì cho ra giá trị 1, khi sử dụng khoảng cách Manhattan thì cho ra giá trị 2. Còn công thức tính khoảng cách Euclidean thì cho ra giá trị là $\sqrt{2}$.

Nếu dữ liệu là văn bản hay không phải là số thì khoảng cách Hamming hay khoảng cách Levenshtein hay sử dụng. Thống kê cho thấy khoảng cách đo lường hay sử dụng là khoảng cách Euclidean. Sau đây, tôi xin giới thiệu các công thức tính khoảng cách :

- Khoảng cách Euclidean : $\| a - b \|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
- Khoảng cách Euclidean vuông : $\| a - b \|_2^2 = \sum_i (a_i - b_i)^2$
- Khoảng cách Manhatthan : $\| a - b \|_1 = \sum_i | a_i - b_i |$
- Khoảng cách cực đại : $\| a - b \|_\infty = \max_i | a_i - b_i |$
- Khoảng cách Mahalanobis : $\sqrt{(a - b)^\top S^{-1} (a - b)}$ với S là ma trận covariance

1.4 Mục tiêu đồ án

Mục tiêu của đồ án là gom nhóm thành các nhóm con chính xác hơn. Để thực hiện được điều này, tôi sử dụng gom nhóm phân cấp kết hợp với đồng gom

nhóm. Ở mỗi cấp thì thuật toán tái sử dụng lại kết quả của cấp trước để gom thành nhóm nhỏ. Văn bản được gom nhóm dựa vào từ còn từ được gom nhóm thì dựa vào văn bản. Điều này lặp lại cho đến cấp cuối cùng và sự kết hợp này mang đến sự hiệu quả.

Chương 2

Phương pháp gom nhóm phân cấp

2.1 Hướng tiếp cận agglomerative

Giới thiệu sơ theo hướng tiếp cận agglomerative.

Ví dụ về agglomerative.

Các đoạn sau liệt kê các paper theo hướng tiếp cận agglomerative.

2.2 Hướng tiếp cận divisive

Giới thiệu so theo hướng tiếp cận divisive.

2.3 Đồng gom nhóm phân cấp

2.3.1 Giới thiệu

Như mọi người đã biết, gom nhóm là phương pháp để chúng ta gom những văn bản gần giống nhau thành một nhóm riêng biệt. Tuy nhiên, các phương pháp truyền thống gặp hạn chế để tìm được kết quả tối ưu khi mà chiều của dữ liệu tăng cao. Để giải quyết vấn đề này, nhiều phương pháp tính khoảng cách

được đề xuất như là độ tương đồng cosine(sử dụng khi số chiều của dữ liệu cao) và lựa chọn đặc trưng(dùng để giảm số chiều của dữ liệu). Tuy nhiên, những cách này không thực sự giải quyết được triệt để vấn đề. Vì nếu cố gắng làm giảm số chiều của dữ liệu, ta gặp phải một vấn đề khác thường hay gọi là lời nguyền về số chiều của dữ liệu. Chính vì vậy, một phương pháp mới được đề xuất là đồng gom nhóm để giúp chúng ta giải quyết vấn đề trên.

Đồng gom nhóm là phương pháp gom nhóm cả hai chiều của dữ liệu(bao gồm gom nhóm văn bản và gom nhóm đặc trưng). Đây là phương pháp mạnh mẽ vì khai thác được độ tương đồng của các phân nhóm trong chiều này của dữ liệu để gom nhóm trong chiều khác. Điều này có nghĩa các phân nhóm của văn bản được đánh giá bằng các phân nhóm của đặc trưng và ngược lại. Bằng cách này, các phân nhóm của văn bản có thể được tiến hành dựa trên các phân nhóm đặc trưng để giúp làm giảm số chiều của dữ liệu. Như vậy, đồng gom nhóm là phương pháp hữu hiệu để giúp ta gom nhóm văn bản đồng thời làm giảm số chiều của dữ liệu.

Khi ta kết hợp đồng gom nhóm với gom nhóm phân cấp thì ta có được đồng gom nhóm phân cấp. Phương pháp này không những cho ta có được những điểm mạnh của đồng gom nhóm mà đồng thời còn có được điểm mạnh của gom nhóm phân cấp. Qua đó, quá trình gom nhóm có thể cho ta thấy được dendrogram, nơi lưu trữ quá trình hợp nhất hoặc phân tách giữa các phân nhóm. Việc tạo thành cây phân cấp giúp cho chúng ta thấy được vị trí liên quan giữa các phân nhóm. Qua đó, ta có thể hiểu được độ tương đồng liên quan giữa các phân nhóm. Từ đó, chúng ta có thể có hình dung được bức tranh tổng thể cũng như là cấu trúc của dữ liệu.

Ngoài ra, đồng gom nhóm phân cấp còn tạo ra công cụ ý niệm hữu ích để

giúp cho ta thấy được mối quan hệ giữa các phân nhóm trong văn bản. Phương pháp này cung cấp thể hiện hữu hình của kết quả phân nhóm đồng thời giải thích được quá trình phân nhóm. Xa hơn nữa, ta có thể sử dụng kết quả phân nhóm để duyệt và tìm kiếm văn bản đồng thời tìm kiếm đặc trưng chung hoặc riêng. Vì vậy, khi kết hợp hai phương pháp lại với nhau, ta sẽ gom nhóm phân cấp văn bản dựa vào gom nhóm phân cấp đặc trưng và ngược lại. Sự kết hợp này giúp cho chúng ta có nhiều kết quả và ý nghĩa hơn những cách thông thường.

Việc tương tác qua lại giữa hai chiều dữ liệu là ý chính trong phương pháp gom nhóm phân cấp. Khi ta xây dựng gom nhóm phân cấp cho văn bản, phương pháp này cũng đồng thời đang nghiên cứu vấn đề của chiều còn lại(đặc trưng). Vì vậy, khi ta xây dựng gom nhóm phân cấp cho đặc trưng, ta mượn lại thông tin đã có từ quá trình gom nhóm phân cấp văn bản để xây dựng. Quá trình này lặp đi lặp lại cho đến khi kết thúc. Như vậy, nhờ có sự tương tác qua lại giữa hai quá trình gom nhóm phân cấp văn bản và gom nhóm phân cấp đặc trưng mà ta mới có được nhiều ích lợi trong quá trình gom nhóm.

Với những ích lợi như vậy, đồng gom nhóm phân cấp cho ta thấy được đây là một phương pháp hiệu quả để gom nhóm văn bản. Phương pháp này không những giúp chúng ta giảm được số chiều của dữ liệu mà kết quả của nó có thể giúp ích cho việc tìm kiếm dữ liệu sau này. Vì vậy, em quyết định sử dụng đồng gom nhóm phân cấp để gom nhóm văn bản. Qua đó, em sử dụng thể hiện văn bản là tần số của đặc trưng, tức là số lần từ đó xuất hiện trong văn bản. Kết quả của phương pháp này sẽ là hai gom nhóm phân cấp, một cho văn bản và một cho đặc trưng.

2.3.2 Độ tương đồng Goodman-Kruskal τ

Mỗi phương pháp gom nhóm đều sử dụng một cách đo độ tương đồng khác nhau. Đối với đồng gom nhóm phân cấp, độ tương đồng Goodman-Kruskal, được ký hiệu là τ, được sử dụng để đo độ tương đồng giữa các văn bản. τ được chọn vì trong nhiều đánh giá về đồng gom nhóm phân cấp thì đây là lựa chọn tốt.

τ được đề xuất như là phương pháp đo lường kết hợp giữa hai biến rời rạc. Phương pháp này đo giảm tỷ lệ trong lỗi dự đoán của biến phụ thuộc được cho bởi thông tin của biến độc lập. τ bị giới hạn bởi 0(không thể kết hợp) và 1(kết hợp hoàn hảo). Chúng ta sử dụng contingency table để lưu trữ phân phối của giá trị cho hai biến. Khi đó, phương pháp này dự đoán tần số thực tế trong bảng cho mỗi giá trị của biến phụ thuộc.

Nguyên tắc cơ bản để τ xác định giảm tỷ lệ trong lỗi bằng hai luật. Luật thứ nhất là quyết định số lượng lỗi của dự đoán mà không sử dụng thông tin biến độc lập. Luật thứ nhất cho ra kết quả E_1, số lượng lỗi của dự đoán cho mỗi giá trị của biến phụ thuộc(dành cho hàng). Luật thứ hai cho thông tin về biến độc lập(dành cho cột) thường dùng để dự đoán tần só của mỗi giá trị của biến phụ thuộc. Luật thứ hai cho ra kết quả về lỗi dự đoán, E_2.

Ví dụ, cho hai biến rời rạc *Job* và *Salary* trong đó *Salary* là biến phụ thuộc, còn *Job* là biến độc lập. Bảng Bảng 2.1 cho dữ liệu như sau:

	$Job = Clerk$	$Job = Teacher$	$Job = Manager$	$Job = Journalist$	Tot
$Salary = Low$	d_{11}	d_{12}	d_{13}	d_{14}	d_{r1}
$Salary = Medium$	d_{21}	d_{22}	d_{23}	d_{24}	d_{r2}
$Salary = High$	d_{31}	d_{12}	d_{13}	d_{14}	d_{r1}
	c_1	c_2	c_3	c_4	T

Bảng 2.1: Bảng dữ liệu thống kê giữa *Salary* và *Job*

Với: d_{ri} là tổng số dữ liệu của dòng thứ i, c_j là tổng số dữ liệu của cột thứ j và T là tổng số dữ liệu của bảng.

Công thức tính τ:

$$\tau = \frac{E_1 - E_2}{E_1} \tag{2.1}$$

Công thức dành cho luật thứ nhất:

$$E_1 = \sum_i (\frac{n - R_i}{n} \times Ri) \tag{2.2}$$

Công thức tính luật thứ hai:

$$\begin{aligned} E_{2j} &= \textstyle\sum_j (\frac{C_j - O_{ij}}{C_j} \times O_{ij}) \\ E_2 &= \textstyle\sum_j E_{2j} \end{aligned} \tag{2.3}$$

Dựa vào công thức tính τ, ta có thể thấy nó thỏa mãn tính chất để thể hiện mối liên quan giữa hai biến. Công thức tính τ thể hiện phép toán hoán vị giữa các trị dòng và cột. Ngoài ra, công thức còn cho ta thấy mối liên hệ về việc giảm lỗi trong dự đoán của biến phụ thuộc dựa vào biến không phụ thuộc. Công thức τ để đo tính hợp lệ cho giải pháp đồng phân nhóm từ hai phân nhóm, một cho giá trị của biến phụ thuộc, còn lại cho biến không phụ thuộc. Ta có thể cải thiện thêm τ bằng cách cải tiến contingency table từ Bảng 2.1

	CC_1	CC_2	
RC_1	t_{11}	t_{12}	T_{RC_1}
RC_2	t_{21}	t_{22}	T_{RC_2}
	T_{CC_1}	T_{CC_2}	T

Với RC_i biểu diễn phân nhóm thứ i của dòng, còn CC_j biểu diễn phân nhóm thứ j trên cột. Trong ví dụ trên, ta giả thiết rằng RC_1 đã gọp lại những dòng từ bảng gốc của *Salary* thành Low, Medium. Trong khi đó, RC_2 còn lại vẫn là $Salary = High$. Tương tự cho cột, CC_1 đã gộp những cột của *Job* thành Clerk, Teacher, còn CC_2 chứa cột Manager, Journalist. t_{ij} biểu diễn giá trị chứa trong

ô giao nhau của dòng thứ i và cột thứ j.

$$t_{ij} = \sum_{Salary=x_value \in RC_i} \; (\sum_{Job=y_value \in CC_j} d_{xy}) \tag{2.4}$$

Với x có miền từ tất cả các giá trị của biến *Salary* và y là miến giá trị của tất cả biến *Job*. Trong cách thể hiện mới này, ta có thể cải thiện cách đánh giá theo từng bước một của thuật toán theo sự kết hợp của hai phân nhóm(dòng và cột).

Chương 3

Phương pháp đề xuất

3.1 Goodman-Krusal τ

phần này giới thiệu phương pháp tính toán goodman-krusal.

Ví dụ về Goodman-krusal τ

3.2 Goodman-krusal trong gom nhóm phân cấp

phần này áp dụng Goodman-krusal τ vào gom nhóm phân cấp.

Chương 4

Thực nghiệm và kết quả

4.1 Dữ liệu

Dữ liệu sử dụng trong chương trình bao gồm dữ liệu tiếng Anh và dữ liệu tiếng Việt. Trong đó, bộ dữ liệu tiếng Việt được tổng hợp từ các trang tin tức nổi tiếng của Việt Nam như vnexpress, dân trí, tuổi trẻ, ... Còn bộ dữ liệu tiếng Anh là các bài báo được lấy từ trang tin tức Reuters được tổng hợp lại thành Reuters-21578. Ngoài ra, cả hai bộ dữ liệu đều là các bài báo đã được phân lớp sơ bộ.

Như đã đề cập, bộ dữ liệu tiếng Việt gồm các trang : vnexpress, dân trí, tuổi trẻ, ... được lấy từ trang tin tức tổng hợp của Google.

Reuters-21578 bao gồm 21,578 bài báo và các bài báo thuộc nhiều danh mục khác nhau. Bộ dữ liệu này được tổng hợp bởi David D. Lewis vào năm 1987. Sau đó thì bộ dữ liệu tiếp tục được phân lớp và chỉnh sửa lại bởi nhiều người khác nhau thuộc Reuters và Carnegie. Bộ dữ liệu này xuất bản và phân phối miễn phí cho mục đích nghiên cứu.

Trong bộ dữ liệu Reuters-21578, nhiều bài chỉ có nội dung là một dòng, thậm chí có bài còn là rỗng. Vì vậy, trước khi sử dụng, ta sử dụng bộ lọc để loại bỏ những bài như vậy. Sau đó, ta chọn ra hai mẫu ngẫu nhiên trong bộ dữ liệu Reuters-21578. Mẫu thứ nhất được đặt tên là re1 và có 1,504 gồm nhiều bài

thuộc các mục khác nhau. Tương tự, ta đặt tên cho mẫu thứ hai là re2 và có 1,657 bài cũng thuộc nhiều mục khác nhau. Việc chọn ra hai mẫu ngẫu nhiên giúp cho việc kiểm nghiệm kết quả của thuật toán được khách quan.

Mục	Bài
Exchanges	7
Orgs	1
People Orgs	1
Places	7013
Plcaes Exchanges	323
Places Orgs	226
Places People	442
Places People Exchanges	7
Places People Orgs	60
Topics	38
Topics Exchanges	1
Topics Orgs	9
Topics People Orgs	2
Topics Places	9298
Topics Places Exchanges	61
Topics Places Orgs	466
Topics Places Orgs Exchanges	2
Topics Places People	412
Topics Places People Exchanges	1
Topics Places People Orgs	87
Others	586

Bảng 4.1: Bảng dữ liệu Reuters-21578 sau khi lọc

4.2 Phương pháp đánh giá

Để đánh giá kết quả gom nhóm văn bản, ta có hai loại chỉ số để sử dụng: chỉ số ngoại vi và chỉ số nội tại. Chỉ số nội tại dùng để đo độ tốt của cấu trúc gom nhóm không cần thông tin ngoài. Chỉ số ngoại vi dùng dùng để đo độ tương đồng giữa hai phân nhóm. Trong đó, phân nhóm thứ nhất là cấu trúc gom nhóm

gốc đã được biết. Còn phân nhóm thứ hai là kết quả từ quá trình gom nhóm. Trong bài toán, ta sử dụng hai chỉ số đánh giá ngoại vi là : NMI(normalized mutual information) và ARI (adjusted rand index).

Như đã đề cập ở phần trên, ta sẽ sử dụng hai chỉ số ngoại vi để đánh giá. Ta có tập $\mathbf{C} = C_1 \dots C_j$ là tập phân nhóm của đối tượng được xây dựng ở một cấp độ nhất định. Tập $\mathbf{P} = P_1 \dots P_j$ là tập hợp được chia bởi phân lớp ban đầu. J và I là tương đương với số phân nhóm của ($|\mathbf{C}|$) và số phân lớp của ($|\mathbf{P}|$). Ta biểu diễn n là tổng số đối tượng trong thuật toán.

4.2.1 Normalized Mutual Information - NMI

NMI có nguồn gốc từ MI(mutual information), được sử dụng nhiều trong lý thuyết xác suất và lý thuyết thông tin. MI là phương pháp đo độ phụ thuộc lẫn nhau giữa hai biến. Trong đây, NMI được nâng cấp để đo phụ thuộc lẫn nhau giữa hai nhóm. Từ đó, NMI cung cấp thông tin cân bằng liên quan đến số lượng phân nhóm. Ngoài ra, NMI còn cho ra kết quả chia sẻ thông tin với lớp thực sự được gán và thông tin hỗn hợp trung bình giữa những cặp của phân nhóm và phân lớp:

$$\mathbf{NMI} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J} x_{ij} \log \frac{n x_{ij}}{x_i x_j}}{\sqrt{\sum_{i=1}^{I} x_i \log \frac{x_i}{n} \sum_{j=1}^{J} x_j \log \frac{x_j}{n}}} \tag{4.1}$$

Với x_{ij} là số lượng phần tử của các đối tượng mà xuất hiện trong cả tập C_j và P_i. x_j là số lượng phần tử chỉ xuất hiện trong tập C_j. x_i là số lượng phần tử chỉ xuất hiện trong tập P_i. Giá trị của chỉ số này nằm trong khoảng từ 0 đến 1.

4.2.2 Adjusted Rand Index - ARI

ARI có nguồn gốc từ RI (rand index), được sử dụng thống kê và gom nhóm dữ liệu. RI dùng để đo độ tương đồng giữa các nhóm dữ liệu. Vấn đề của RI là giá trị mong muốn của hai phân nhóm ngẫu nhiên nằm trong khoảng từ 0 và 1. Vì vậy, ARI ra đời là phiên bản chỉnh sửa có thể định nghĩa cho việc điều chỉnh cho cơ hội gom nhóm các thành phần. Giá trị của ARI có thể nằm trong phạm vi từ -1 đến 1.

$$E[\alpha] = \frac{\pi(C) \cdot \pi(P)}{n(n-1)/2} \tag{4.2}$$

Với $\pi(C)$ và $\pi(P)$ biểu thị tương ứng với số lượng các cặp đối tượng của cùng phân nhóm trong $\mathbf{C}$ và cùng phân lớp trong $\mathbf{P}$. Giá trị lớn nhất cho α có thể đạt được là:

$$\max(\alpha) = \frac{1}{2}(\pi(C) + \pi(P)) \tag{4.3}$$

Độ tương đồng giữa $\mathbf{C}$ và $\mathbf{P}$ có thể được ước lượng bởi adjusted rand index như sau:

$$AR(\mathbf{C}, \mathbf{P}) = \frac{\alpha - E[\alpha]}{\max(\alpha) - E[\alpha]} \tag{4.4}$$

4.3 Kết Quả

Chương 5

Kết luận và hướng phát triển

Các vấn đề đã giải quyết và khiếm khuyết của thuật toán.

Hướng phát triển.

Tài liệu tham khảo