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

Họ và tên HVCH

TÊN ĐỀ TÀI LUẬN VĂN

Chuyên ngành: Khoa học máy tính

Mã số chuyên ngành: 12345

LUẬN VĂN THẠC SĨ: KHOA HỌC MÁY TÍNH

NGƯỜI HƯỚNG DẪN KHOA HỌC:

1. Tên người hướng dẫn

Tp. Hồ Chí Minh, Năm 2015

Lời cảm ơn

Tôi xin chân thành cảm ơn ...

Mục lục

Danh sách hình

Danh sách bảng

Chương 1

Giới thiệu

1.1 Giới thiệu về gom nhóm

Gom nhóm là công việc tìm những nhóm đối tượng gần giống nhau trong dữ liệu. Vấn đề gom nhóm được nghiên cứu rộng rãi trong cơ sở dữ liệu và thống kê trong khai thác dữ liệu. Gom nhóm được thực hiện bằng cách dựa vào độ tương đồng của các đối tượng. Độ tương đồng giữa các đối tượng được xác định bằng hàm tương đồng. Dựa vào độ tương đồng này, gom nhóm có thể gom những đối tượng gần giống nhau vào thành những nhóm khác biệt. Qua đó, gom nhóm giúp cho chúng ta dễ dàng phân loại được dữ liệu. Đặc biệt, gom nhóm rất hữu dụng trong việc áp dụng cho văn bản để giúp phân loại và truy xuất văn bản được cải thiện.

Ngày nay, internet đang được phủ rộng khắp mọi nơi nên giúp cho việc cập nhật thông tin trở nên dễ dàng hơn. Tuy nhiên, mặt trái của internet chính là việc đưa thông tin ồ ạt mà không có chọn lọc. Điều đó khiến cho người dùng gặp nhiều khó khăn trước muôn vàn lựa chọn. Do đó, người dùng cần một giải pháp để giúp cho họ có thể tiếp cận thông tin một cách dễ dàng và có chọn lọc hơn. Gom nhóm văn bản có thể gom những bài báo, chủ để gần nhau giống nhau để giúp người dùng có

thể dễ dàng chọn lựa.

Gom nhóm không những giúp cho người dùng lựa chọn dễ dàng và còn chắt lọc thông tin cho người dùng. Việc chắt lọc này là do trong quá trình gom nhóm các loại thông tin gần giống nhau thành các nhóm. Từ đó, người dùng thay vì phải tìm kiếm thông tin dàn trải thì có thể chỉ vào những chủ đề mà mình yêu thích để tiết kiệm thời gian.

Chúng ta thường hay đọc tin tức trực tuyến thông qua các trang báo tin tức trên mạng. Nhưng việc phải vào từng trang báo chỉ để đọc một vài tin tức chọn lọc thì rất là mất thời gian. Chính vì vậy, các đại gia công nghệ đã bắt đầu tiến vào lĩnh vực truyền thông bằng các công cụ như Google news, Apple news, Facebook news. Đây là những công cụ tập hợp thông tin từ các trang báo trên mạng rồi từ đó gom nhóm thành chuyên mục tương tự nhau. Có thể thấy, gom nhóm văn bản đang đóng vai trò quan trọng trong việc kết nối người dùng trên mạng với nhau.

Việc nhiều công ty công nghệ lớn cùng tham gia vào lĩnh vực truyền thông cụ thể là mảng tin tức trực tuyến cho thấy được tầm quan trọng của gom nhóm văn bản. Tương tự như gom nhóm, nhiệm vụ của gom nhóm văn bản cũng là tìm kiếm những văn bản có độ tương đồng gần giống nhau thành các nhóm. Tuy nhiên, do văn bản là tập hợp các chữ, từ, câu nên không thể gom nhóm dưới dạng dữ liệu thô vì gây ra quá nhiều khó khăn. Vì vậy, văn bản cần được biểu diễn lại dưới dạng dữ liệu khác để giúp cho việc gom nhóm các văn bản với nhau trở nên dễ dàng hơn. Nhưng mà cách thực hiện, cũng như là ý tưởng của gom nhóm văn bản cũng không có gì khác so với gom nhóm.

1.2 Gom nhóm văn bản

Để thực hiện gom nhóm văn bản, chúng ta không thể sử dụng văn bản dưới dạng dữ liệu thô. Văn bản cần được biểu diễn thành kiểu dữ

liệu khác để giúp cho việc khai thác các thuật toán trên văn bản trở nên dễ dàng hơn. Từ đó, giúp cho việc gom nhóm văn bản thêm hiệu quả và chính xác.

Gom nhóm văn bản dùng để gom những văn bản gần giống nhau thành các nhóm. Do vậy, gom nhóm văn bản rất hữu ích trong việc phân chia văn bản. Việc gom các văn bản gần tương đồng cũng giúp ích cho việc tìm kiếm văn bản. Sự sắp xếp của các nhóm văn bản cũng hỗ trợ cho việc truy xuất văn bản được cải thiện. Như vậy, gom nhóm văn bản đem lại nhiều lợi ích cho việc phân loại, tìm kiếm cũng như là tăng tốc độ truy xuất văn bản.

Nghiên cứu của vấn đề gom nhóm đứng trước tính khả dụng khi ứng dụng vào văn bản để thỏa mãn một 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 thống 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.

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. Tuy nhiên, văn bản có định dạng không phải là số nên khi gom nhóm văn bản ta cần chuyển đổi văn bản bằng một dạng thể hiện khác mà ta có thể thao tác được. Thông thường, văn bản sẽ được chuyển đổi và biểu diễn thành một vector để giúp cho chúng ta số hóa văn bản. Việc số hóa văn bản giúp cho chúng ta có thể gom nhóm dễ dàng hơn so với việc phải thao tác trên các từ, chữ, câu của văn bản thông thường.

Sự thay đổi định dạng của văn bản từ dạng chữ sang số với thể hiện là vector đã giúp cho mang lại nhiều lợi ích trong việc gom nhóm. Do thể hiện của văn bản là dạng vector nên các thuật toán gom nhóm vẫn có thể được áp dụng được. Không những thế, việc để định dạng là số giúp cho việc tính toán cũng như là thay đổi trở nên dễ dàng hơn. Khi ta số hóa các định dạng chữ của văn bản cũng là cách để tiết kiệm bộ nhớ, qua đó cải thiện thuật toán, tăng tốc độ thực thi. Có thể thấy, việc số hóa văn bản đã đem lại những lợi ích to lớn trong quá trình gom nhóm văn bản.

1.3 Các phương pháp gom nhóm văn bản

1.4 Các công thức tính khoảng cách

Trong gom nhóm văn bản, các công thức tính khoảng cách được sử dụng để đo lường giữa hai văn bản. Sau đây, một số công thức tính khoảng cách thường được sử dụng:

- 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 : $\| \ a \ - \ b \ \|_1 = \sum\limits_i \mid a_i \ - \ b_i \mid$
- Khoảng cách cực đại : $\| \ a \ - \ b \ \|_\infty = \max\limits_i \mid a_i \ - \ b_i \mid$
- Khoảng cách Mahalanobis : $\sqrt{(a \ - \ b)^\top \, S^{-1} \, (a \ - \ b)}$ với S là ma trận covariance

1.5 Lựa chọn công thức tính khoảng cách

Chương 2

Các công trình liên quan

2.1 Gom nhóm phân cấp

2.1.1 Giới thiệu

Gom nhóm phân cấp là phương pháp gom nhóm dùng để xây dựng các nhóm thành các cấp bậc khác nhau. Các phương pháp gom nhóm được liệt kê ở chương 1 cho ta thấy được sự hiệu quả trong việc gom nhóm và tương đối đơn giản. Tuy nhiên, các phương pháp này chỉ cho ta thấy được kết quả gom nhóm cuối cùng, không cho chúng ta thấy được cấu trúc của dữ liệu. Hơn thế nữa, các phương pháp này đa phần cần phải xác định số lượng phân nhóm ban đầu. Vì vậy, một phương pháp khác được giới thiệu là gom nhóm phân cấp để khắc phục các nhược điểm trên.

2.1.2 Các hướng tiếp cận

Để xây dựng nên cấu trúc cây của các phân nhóm trong gom nhóm phân cấp, ta có hai hướng tiếp cận khác nhau. Đó là hướng tiếp cận tích tụ và hướng tiếp cận phân chia.

2.1.2.1 Hướng tiếp cận tích tụ

Hướng tiếp cận tích tụ là một trong những cách tiếp cận của gom nhóm phân cấp. Đây là phương pháp tiếp cận gom nhóm theo cách đi từ dưới lên. Phương pháp này bắt đầu từ việc quan sát mỗi phân nhóm, các phân nhóm sẽ gom lại với nhau và chuyển thành cấp cao hơn trong cây. Trong trường hợp thông thường, độ phức tạp của thuật toán theo hướng tiếp cận tích tụ là $O(n^2 \log(n))$. Tuy nhiên, thuật toán có thể tối ưu hóa để làm giảm độ phức tạp bằng cách sử dụng liên kết đơn hoặc là liên kết toàn phần.

Cho dữ liệu thô sau, ta sẽ sử dụng hướng tiếp cận tích tụ để gom nhóm phân cấp với công thức tính khoảng cách là Euclid.

2.1.2.2 Hướng tiếp cận phân chia

Hướng tiếp cận phân chia là một trong những cách tiếp cận của gom nhóm phân cấp. Đây là phương pháp tiếp cận gom nhóm theo cách đi từ trên xuống. Phương pháp này bắt đầu từ việc quan sát tất cả chỉ trong một phân nhóm, và khi di chuyển xuống sẽ tách dần dần thành các phân nhóm con. Quá trình gộp và tách thường sẽ tốn nhiều chi phí cho thuật toán. Khác với hướng tiếp cận tích tụ, hướng tiếp cận phân chia có độ phức tạp lớn hơn, $O(2^n)$.

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 hiệu quả 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.

Chương 3

Phương pháp đề xuất

Chương 4

Thực nghiệm và kết quả

4.1 Thực nghiệm

4.1.1 Môi trường

4.1.2 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.

4.1.2.1 Dữ liệu tiếng Anh

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.

4.1.2.2 Dữ liệu tiếng Việt

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.

4.2 Các phương pháp đánh giá

4.2.1 Giới thiệu 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.

4.2.2 Cách đánh giá NMI

4.2.2.1 Giới thiệu

NMI có nguồn gốc từ MI(mutual information), được sử dụng nhiều trong lý thuyết xác suất và lý thuyết thông tin. NMI 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.

4.2.2.2 Công thức

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

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.

4.2.2.3 Ví dụ

4.2.3 Cách đánh giá ARI

4.2.3.1 Giới thiệu

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.

4.2.3.2 Công thức

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

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:

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

4.2.3.3 Ví dụ

4.3 Kết quả

Chương 5

Kết luận và hướng phát triển

Danh mục công trình của tác giả

1. Tạp chí ABC
2. Tạp chí XYZ

Tài liệu tham khảo

Phụ lục A

Ngữ pháp tiếng Việt

Đây là phụ lục.

Phụ lục B

Ngữ pháp tiếng Nôm

Đây là phụ lục 2.