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

Lời cảm ơn **i**

Đề cương chi tiết **ii**

Mục lục **ii**

Tóm tắt **iv**

1 Giới thiệu **1**

2 Các công trình liên quan **5**

3 Phương pháp đề xuất **6**

4 Thực nghiệm và kết quả **7**

5 Kết luận và hướng phát triển **8**

Danh mục công trình của tác giả **9**

Tài liệu tham khảo **10**

A Ngữ pháp tiếng Việt **11**

B Ngữ pháp tiếng Nôm **12**

Danh sách hình

Danh sách bảng

Chương 1

Giới thiệu

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.

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.

Chương 2

Các công trình liên quan

Chương 3

Phương pháp đề xuất

Chương 4

Thực nghiệm và 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.