ĐẠI HỌC THÁI NGUYÊN
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

LÈNG HOÀNG LÂM

PHÂN LOẠI VĂN BẢN HÀNH CHÍNH TIẾNG VIỆT VÀ ỨNG DỤNG VÀO CÁC CƠ QUAN NHÀ NƯỚC TỈNH BẮC KẠN

Chuyên ngành: Khoa học máy tính
Mã số: 60 48 0101

LUẬN VĂN THẠC SỸ KHOA HỌC MÁY TÍNH

Người hướng dẫn khoa học: PGS.TS. ĐOÀN VĂN BAN

Thái Nguyên - 2017

LỜI CAM ĐOAN

Tôi xin cam đoan đây là sản phẩm nghiên cứu, tìm hiểu của cá nhân tôi. Các số liệu, kết quả trình bày trong luận văn là trung thực. Những nội dung trình bày trong luận văn hoặc là của bản thân, hoặc là được tổng hợp từ những nguồn tài liệu có nguồn gốc rõ ràng và được trích dẫn hợp pháp, đầy đủ.

Tôi xin hoàn toàn chịu trách nhiệm cho lời cam đoan của mình.

Thái Nguyên, tháng 4 năm 2017

HỌC VIÊN

Lèng Hoàng Lâm

LỜI CẢM ƠN

Trân trọng cảm ơn các thầy giáo, cô giáo trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên; các giảng viên đến từ Viện Hàn lâm Khoa học và Công nghệ Việt Nam, Trường Đại học Quốc gia Hà Nội... đã tạo điều kiện tốt nhất cho học viên trong quá trình học tập và làm luận văn. Đặc biệt, xin được bày tỏ lòng biết ơn chân thành và sâu sắc nhất tới thầy giáo, PGS.TS. Đoàn Văn Ban, người đã định hướng và luôn tận tình chỉ bảo, hướng dẫn em trong việc nghiên cứu, thực hiện luận văn này.

Trong suốt quá trình học tập và thực hiện đề tài, học viên luôn nhận được sự ủng hộ, động viên của gia đình, đồng nghiệp, đặc biệt là sự quan tâm tạo điều kiện của Ban lãnh đạo Trung tâm Công nghệ thông tin và Truyền thông tỉnh Bắc Kạn - nơi học viên đang công tác. Xin trân trọng cảm ơn!

Thái Nguyên, tháng 4 năm 2017

HỌC VIÊN

Lèng Hoàng Lâm

MỤC LỤC

DANH MỤC CÁC TỪ VIẾT TẮT

Từ viết tắt	Giải thích
CSDL	Cơ sở dữ liệu
KDD	Knowledge Discovery from Data
IDF	Inverse Document Frequency
kNN	K-Nearest Neighbor
NB	Naïve Bayers
SVM	Support Vector Machine
S^3VM	Semi-Supervised Support Vector Machine
TBL	Transformation - based Learning
TF	Term Frequency
WFST	Weighted Finite - State Transducer

DANH MỤC CÁC HÌNH

DANH MỤC CÁC BẢNG

MỞ ĐẦU

1. Đặt vấn đề

Trong thời đại bùng nổ Công nghệ thông tin hiện nay, phương thức sử dụng văn bản giấy truyền thống đã dần được số hóa, chuyển sang dạng các văn bản điện tử lưu trữ trên máy tính và được chia sẻ, truyền tải trên mạng. Với rất nhiều tính năng ưu việt của tài liệu số như: Lưu trữ gọn nhẹ, linh hoạt; thời gian lưu trữ lâu dài; dễ hiệu chỉnh và đặc biệt tiện dụng trong trao đổi, chia sẻ nên ngày nay, số lượng văn bản điện tử được sử dụng trong các cơ quan nhà nước tăng lên rất nhanh chóng. Do đó, một vấn đề đặt ra là làm thế nào để có thể tìm kiếm và khai thác thông tin từ nguồn dữ liệu phong phú này. Các kỹ thuật để giải quyết vấn đề này được gọi là "Text Mining" hay Khai phá dữ liệu văn bản.

Khai phá dữ liệu văn bản đề cập đến tiến trình trích lọc các mẫu hình thông tin hay tri thức đáng quan tâm hoặc có giá trị từ các tài liệu văn bản. Trong đó, phân loại văn bản là một bài toán cơ bản nhất của lĩnh vực khai phá dữ liệu văn bản. Phân loại văn bản là công việc phân tích nội dung của văn bản và sau đó ra quyết định (hay dự đoán) văn bản thuộc nhóm nào trong các nhóm văn bản đã cho trước. Văn bản được phân loại có thể thuộc một nhóm, nhiều nhóm, hoặc không thuộc nhóm văn bản mà ta đã định nghĩa trước. Phân loại văn bản có thể thực hiện bằng nhiều cách như sử dụng tiếp cận lý thuyết tập thô, cách tiếp cận theo luật kết hợp hoặc dựa trên cách tiếp cận máy học. Đây là một lĩnh vực mang tính khoa học cao, ứng dụng được rất nhiều trong các bài toán thực tế hiện nay như tìm kiếm thông tin, lọc văn bản, tổng hợp tin tức tự động, thư viện điện tử,… Do vậy, học viên quyết định chọn đề tài "*Phân loại văn bản hành chính tiếng Việt và ứng dụng vào các cơ quan nhà nước tỉnh Bắc Kạn*" để nghiên cứu, thực hiện luận văn tốt nghiệp của mình.

Mục tiêu của đề tài luận văn là khảo sát, tìm hiểu một số phương pháp

phân loại văn bản thường được sử dụng hiện nay, trên cơ sở đó đề xuất lựa chọn một phương án phân loại văn bản tiếng Việt tự động và ứng dụng thử nghiệm phân loại cho một đối tượng cụ thể là văn bản hành chính tiếng Việt.

2. Đối tượng và phạm vi nghiên cứu

Đối tượng nghiên cứu bao gồm: Các thuật toán phân loại văn bản và các vấn đề liên quan đến bài toán phân loại văn bản tiếng Việt.

Phạm vi nghiên cứu của luận văn tập trung vào một số thuật toán phân loại văn bản thông dụng; các đặc trưng của văn bản tiếng Việt; các kỹ thuật liên quan trong xử lý phân loại văn bản và ứng dụng thuật toán học bán giám sát trong phân loại văn bản tiếng Việt.

3. Hướng nghiên cứu của đề tài

Nghiên cứu lý thuyết cơ bản về khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản với một số thuật toán phân loại văn bản thông dụng như Naïve Bayers, K-Nearest Neighbor, Support Vector Machine.

Nghiên cứu về các đặc trưng của văn bản tiếng Việt và các kỹ thuật liên quan trong xử lý phân loại văn bản tiếng Việt như tách từ, biểu diễn văn bản, đánh trọng số của từ, tính độ tương đồng văn bản.

Từ kết quả thu được tiến hành cài đặt ứng dụng trong bài toán phân loại văn bản hành chính tiếng Việt.

4. Những nội dung chính

Nội dung chính của luận văn được trình bày trong 3 chương với tổ chức cấu trúc như sau:

Chương 1. Tổng quan về phân loại văn bản tiếng Việt.

Chương này trình bày khái quát về khai phá dữ liệu, khai phá dữ liệu văn bản và bài toán phân loại văn bản tiếng Việt; đồng thời làm rõ các đặc trưng của văn bản tiếng Việt và giới thiệu sơ bộ về công tác quản lý văn bản tại các cơ quan thuộc tỉnh Bắc Kạn.