BỘ GIÁO DỤC VÀ ĐÀO TẠO

ĐẠI HỌC ĐÀ NẴNG

NGUYỄN THỊ THUÝ KIỀU

PHÂN LOẠI VĂN BẢN TỰ ĐỘNG TRONG HỆ THỐNG ĐIỀU HÀNH TÁC NGHIỆP TẠI SỞ THÔNG TIN VÀ TRUYỀN THÔNG QUẢNG NAM

Chuyên ngành: KHOA HỌC MÁY TÍNH

Mã số: 60.48.01

TÓM TẮT LUẬN VĂN THẠC SĨ KỸ THUẬT

Đà Nẵng - Năm 2011

Công trình được hoàn thành tại

ĐẠI HỌC ĐÀ NẴNG

Người hướng dẫn khoa học: **PGS.TS. Phan Huy Khánh**

Phản biện 1: **TS. Nguyễn Thanh Bình**

Phản biện 2: **PGS.TS. Lê Mạnh Thạnh**

Luận văn được bảo vệ trước Hội đồng chấm Luận văn tốt nghiệp thạc sĩ kỹ thuật họp tại Đại học Đà Nẵng vào ngày 15 tháng 10 năm 2011

Có thể tìm hiểu luận văn tại:

- Trung tâm Thông tin - Học liệu, Đại học Đà Nẵng

- Trung tâm Học liệu, Đại học Đà Nẵng

MỞ ĐẦU

1. Lý do chọn đề tài

Hiện nay, Tin học hóa các hoạt động QLNN đang ngày càng đặc biệt chú trọng. Tại Việt Nam, phần lớn các cơ quan hành chính nhà nước đang sử dụng các hệ thống phần mềm hỗ trợ điều hành, xử lý hồ sơ văn bản. Về lâu dài, cơ sở dữ liệu về hồ sơ văn bản của các hệ thống này ngày càng phình to, khối lượng dữ liệu lớn gây khó khăn trong việc tìm kiếm văn bản phục vụ công việc.

Để nâng cao hiệu quả trong việc sử dụng các hệ thống phần mềm này, chúng ta cần phải xây dựng chức năng có khả năng phân loại tự động nhằm sắp xếp, phân loại các văn bản trong CSDL văn bản để quá trình tìm kiếm, truy xuất của người dùng được nhanh nhạy và dễ dàng hơn.

Tại Sở Thông tin & Truyền thông Quảng Nam hiện đang thực hiện quy trình nghiệp vụ văn bản thông qua hệ thống phần mềm Điều hành tác nghiệp Q- Office hoạt động trong môi trường internet. Hệ thống lưu trữ CSDL văn bản khá lớn và mỗi ngày càng phình to, do đó cần thiết phải xây dựng chức năng phân loại tự động nhằm hỗ trợ người xử dụng tìm kiếm văn bản một cách dễ dàng nhằm nâng cao hiệu quả công việc, tiết kiệm thời gian…

Xuất phát từ nhu cầu đó, tôi đã chọn đề tài **"Phân loại văn bản tự động trong hệ thống điều hành tác nghiệp tại Sở Thông tin và Truyền thông Quảng Nam"** làm đề tài cho luận văn thạc sỹ của mình.

2. Mục tiêu và nhiệm vụ

Mục đích của đề tài là:

Nâng cao tính hiệu quả trong việc tìm kiếm, tra cứu văn bản trong CSDL văn bản của hệ thống điều hành tác nghiệp Q-Office tại Sở Thông tin & Truyền thông Quảng Nam…

Mục tiêu của tôi trong đề tài này là tập trung vào việc nghiên cứu các vấn đề:

Nghiên cứu các phương pháp tách từ đã được áp dụng thành công trong một số ngôn ngữ như: tiếng Anh, tiếng Trung… Có đánh giá về các phương pháp này khi áp dụng cho tiếng Việt

3. Đối tượng và phạm vi nghiên cứu

Trong khuôn khổ của luận văn này tôi tập trung nghiên cứu một số các phương pháp phân loại văn bản cũng như một số các phương pháp tách từ áp dụng cho văn bản tiếng Việt. Trên cơ sở đó tôi tiến hành so sánh đánh giá các phương pháp :

Các phương pháp phân loại văn bản trong luận văn gồm :

- Phương pháp xác xuất Naive Bayes
- Phương pháp K người láng giềng gần nhất (K – Nearest Neighbours)
- Phương pháp sử dụng mạng Neral
- Phương pháp phân loại văn bản bằng cây quyết định
- Phân loại văn bản bằng phương pháp hồi qui
- Phương pháp máy học sử dụng vector hỗ trợ (SVM)
- Tìm hiểu các phương pháp tách từ trong văn bản
- Tách từ tiếng Việt dùng mô hình WFST
- Tách từ tiếng Việt dùng mô hình Maximum Matching
- Tách từ tiếng Việt dùng mô hình MMSeg
- Tách từ tiếng Việt dùng mô hình Maximum Entropy

4. Phương pháp nghiên cứu

Để có thể *"Phân loại được văn bản tiếng Việt tự động trong Hệ thống Điều hành tác nghiệp tại Sở Thông tin và Truyền thông Quảng Nam"* thì điều đầu tiên là cần phải tách văn bản thành các từ và cụm từ có nghĩa trong tiếng Việt. Vì thế trong đề tài này, tôi tiến hành nghiên cứu một số phương pháp tách từ áp dụng cho tiếng Việt và xây dựng công cụ tách từ hiệu quả trên văn bản tiếng Việt. Từ đó, áp dụng vào bài toán phân loại văn bản để xây dựng công cụ phân loại tự động văn bản tiếng Việt theo chủ đề.

5. Ý nghĩa khoa học và thực tiễn của đề tài

Việc xây dựng thành công công cụ *Phân loại văn bản tự động trong Hệ thống Điều hành tác nghiệp tại Sở Thông tin và Truyền thông Quảng Nam* sẽ có thể được áp dụng vào nhiều ứng dụng cụ thể trong đời sống, góp phần giảm thiểu sự tiêu tốn về thời gian và công sức con người. Đồng thời, việc xây dựng thành công công cụ tách từ hiệu quả trên văn bản tiếng Việt mở ra điều kiện thuận lợi cho các bài toán xử lý ngôn ngữ tự nhiên khác trên tiếng Việt.

Đề tài cũng đóng góp một hướng tiếp cận mới giúp giải quyết các bài toán xử lý ngôn ngữ tự nhiên trong điều kiện chúng ta chưa có một công cụ tách từ cho độ chính xác.

6. Bố cục luận văn

Luận văn được chia làm 3 chương có nội dung như sau :

Chương 1. Tìm hiểu phân loại văn bản tiếng Việt

Chương 2. Phân tích và thiết kế hệ thống

Chương 3. Xây dựng ứng dụng và đánh giá kết quả thử nghiệm.

CHƯƠNG 1
TÌM HIỂU PHÂN LOẠI VĂN BẢN

1.1 LÝ THUYẾT VỀ VĂN BẢN VÀ PHÂN LOẠI VĂN BẢN

1.1.1 Khái niệm văn bản

Theo Wikipedia (http://en.wikipedia.org/wiki/Text) thì văn bản (text, document) được giới thiệu như sau:

Trong ngôn ngữ học (linguistics), văn bản là một hoạt động giao tiếp, thi hành 7 nguyên tắc cấu thành cơ bản và 3 nguyên tắc điều khiển của văn bản học. Cả tiếng nói, ngôn ngữ viết hay ngôn ngữ thông thường đều có thể xem như văn bản trong ngôn ngữ học.

1.1.2 Phân lớp văn bản

TheoWikipedia (http://en.wikipedia.org/wiki/Categorization) thì khái niệm về phân lớp (classification, categorization) là một tiến trình trong đó các đối tượng và sự việc được nhận ra, được phân biệt và hiểu được. Sự phân lớp hàm ý rằng các đối tượng được nhóm thành các bộ phân loại, thường thì phục vụ cho một vài mục đích đặc biệt.

1.1.3 Phân loại văn bản

Phân loại văn bản được định nghĩa như quá trình gán các văn bản vào một hay nhiều nhóm phù hợp được xác định trước dựa trên nội dung của văn bản đó.

Có nhiều cách để phân loại văn bản để dễ dàng trong việc tìm kiếm:

Phân loại bằng cách lưu trữ các loại công văn giấy tờ vào các hệ thống tủ đựng hồ sơ để khi tìm kiếm sẽ dễ dàng hơn. Tuy nhiên việc này cũng tốn khá nhiều thời gian công sức khi một ngày các cơ quan này tiếp nhận không biết bao nhiêu công văn giấy tờ gửi đến.

Chính vì bất lợi đó mà bài toán phân loại công văn giấy tờ trên máy đã được người ta nghĩ đến. Hệ thống sẽ tìm ra đặc thù của văn bản một cách tương đối máy móc, thuật toán này tương đối hiệu quả song lại chỉ phù hợp cho các nhóm dữ liệu tương đối đặc thù.

Phương pháp này cũng mất rất nhiều thời gian và công sức. Ngoài phương pháp phân loại thủ công như trên, để xây dựng công cụ phận loại văn bản tự động người ta thường dùng các thuật toán học máy (Machine Learning).

1.1.4 Các ứng dụng của phân loại văn bản

Phân loại văn bản là bài toán nền tảng trong lĩnh vực ***truy hồi thông tin*** (Information Retrieval: IR) có liên quan một phần đến xử lý ngôn ngữ tự nhiên (Natural Language Processing: NLP). Phân loại văn bản là bài toán ứng dụng rất nhiều trong lĩnh vực xử lý ngôn ngữ hiện nay :

Search engines, hệ thống lọc Spam mail, hệ thống phân loại để phục vụ cho việc lưu trữ và tìm kiếm…

1.2 BÀI TOÁN PHÂN LOẠI VĂN BẢN

1.2.1 Một số khái niệm cơ bản

1.2.1.1 Hạng (Term)

Hạng trong một văn bản có thể là từ đơn (Single words), hoặc ngữ (phrases).

1.2.1.2 Từ khóa

Từ khóa là các từ xuất hiện trong một văn bản hay bài báo ở dạng nguyên thể.

1.2.1.3 Từ dừng (Stopword)

Trước hết có thể quan sát thấy rằng trong các ngôn ngữ tự nhiên, rất nhiều từ được dùng để biểu diễn cấu trúc câu nhưng hầu như không mang ý nghĩa về mặt nội dung, chẳng hạn các loại từ: giới từ, liên từ,… Những từ đó được gọi là từ dừng (stopword).

1.2.1.4 Thuật ngữ

Thuật ngữ là các từ khóa có nghĩa liên quan đến một lĩnh vực nào đó.

1.2.1.5 Khái niệm

Khái niệm là các thuật ngữ nhưng nó là sự khái quát hóa, tổng quát hóa của nhiều thuật ngữ khác.

1.2.1.6 Lớp (Category)

Lớp của các tài liệu là sự gom nhóm các tài liệu có nội dung tương tự nhau.

1.2.1.7 Trọng số (Weight)

Là một giá trị đặc trưng cho hạng, giá trị này thường là số thực. Công thức người ta thường dùng là TFIDF (Term Frequency and Inverse Document Frequency) và một số mở rộng của nó như logTF_IDF, TF_IWF,…

1.2.1.8 Đặc trưng (Feature)

Đặc trưng của văn bản là những hạng (term) trong văn bản. Cơ bản thì có 2 loại thuật toán (algorithm) để biểu diễn không gian đặc trưng (feature space) trong quá trình phân lớp.

1.2.1.9 Chọn lựa đặc trưng (Feature Selection)

Chọn lựa đặc trưng là chọn lựa 1 tập con (subset) các đặc trưng biểu diễn từ không gian đặc trưng gốc.

1.2.1.10 Rút trích đặc trưng (Feature Extraction)

Rút trích đặc trưng là biến đổi (transform) không gian đặc trưng gốc (đầu vào) thành một không gian đặc trưng nhỏ hơn để giảm chiều đặc trưng.

1.2.2 Tổng quan bài toán phân loại văn bản

Để phân loại, rất nhiều cách tiếp cận đã được áp dụng như :

- Dựa vào từ khóa.
- Dựa vào thống kê tần số xuất hiện của các từ trong văn bản...

1.2.3 Lịch sử nghiên cứu đối với bài toán phân loại văn bản

So với bài toán phân loại văn bản áp dụng trên tiếng Anh, phân loại văn bản tiếng Việt mới có trong thời gian gần đây. Nhiều áp dụng thử nghiệm các phương pháp phân loại đã kiểm chứng cho kết quả tốt trên tiếng Anh được áp dụng cho văn bản tiếng Việt. Tuy nhiên, so với bài toán phân loại văn bản tiếng Anh, bài toán phân loại văn bản tiếng Việt hoàn toàn chưa có một kết quả nào được công bố chính thức.

1.2.4 Các phương pháp tiếp cận bài toán

Tiếp cận theo hướng dãy các từ (Bag of Words – BOW)
Tiếp cận theo hướng mô hình ngôn ngữ thống kê N-Gram
Kết hợp 2 phương pháp trên

1.2.5 Phân loại văn bản tiếp cận theo hướng dãy từ

Phân loại văn bản tiếp cận theo hướng dãy từ có các phương pháp sau:

1.2.5.1 Xác suất Naive Bayes

Naive Bayes là phương pháp phân loại dựa trên xác suất được sử dụng rộng rãi trong lĩnh vực máy học, được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron năm 1961 và ngày càng trở nên phổ biến.

1.2.5.2 K-láng giềng gần nhất

kNN là phương pháp truyền thống khá nổi tiếng về hướng tiếp cận dựa trên thống kê đã được nghiên cứu trên nhận dạng mẫu hơn suốt bốn thập kỷ qua. kNN là phương pháp đơn giản và không cần huấn luyện để nhận dạng mẫu trong tập huấn luyện như các phương pháp khác.

1.2.5.3 Sử dụng mạng neural

Mạng neural nhân tạo là phương pháp máy học cung cấp phương pháp hiệu quả để tạo ra các giá trị xấp xỉ của những hàm có giá trị thực, giá trị rời rạc, vector. NN mô phỏng theo hệ thống sinh học thực tế, với các tế bào thần kinh gọi là neural liên kết với nhau tạo thành một mạng gọi là mạng neural. Mỗi neural nhận một hoặc nhiều giá trị đầu vào và tạo ra một giá trị thực duy nhất ở đầu ra, giá trị ở đầu ra này có thể trở thành đầu vào cho một neural khác.

1.2.5.4 Phân loại văn bản bằng cây quyết định

Có một lớp các thuật toán không sử dụng xác suất hay còn gọi là không sử dụng số học mà thay vào đó là sử dụng các mô hình thể hiện. Trong những phương pháp này có thể kể đến hai phương pháp điển hình là ***phương pháp học luật quy nạp*** và ***cây quyết định***.

1.2.5.5 Hồi quy

Hồi quy được định nghĩa là hàm xấp xỉ giá trị thực f thay cho giá trị nhị phân trong bài toán phân lớp. Hàm f sẽ có nhiệm vụ học từ kho ngữ liệu huấn luyện. Ở đây, tôi chỉ đề cập đến một mô hình, LLSF (Linear Least-Squares Fit) ứng dụng trong bài toán phân loại văn bản.

1.2.5.6 Phân loại văn bản sử dụng Support Vector Machines(SVM)

SVM là phương pháp nhận dạng do Vladimir N.Vapnik đề xuất năm 1995. SVM là phương pháp nhận dạng dựa trên lý thuyết học thống kê ngày càng được sử dụng phổ biến trong nhiều lĩnh vực, đặc biệt là lĩnh vực phân loại mẫu và nhận dạng mẫu.

1.2.6 Phân loại văn bản tiếp cận theo hướng mô hình ngôn ngữ thống kê N-Gram

Một công trình khá nổi tiếng là Tree Augmented Naive Bayes (TAN) classifier của [Friedman, 1997]. Bộ phân lớp này cho phép sự phụ thuộc có cấu trúc cây giữa các biến quan sát ngoài phụ thuộc truyền thống trên biến gốc (root) của cây. Tuy nhiên, quá trình học mạng Bayes theo cấu trúc cây có chi phí khá lớn, và vì thế mô hình này hiếm khi được sử dụng bài toán phân loại văn bản. [Fuchun Peng et al] trong công trình đã đề nghị một phương pháp rất hay để giải quyết vấn đề giả sử độc lập, công trình này có tên là mô hình Chain Augmented Naive Bayes (CAN). Mô hình này đơn giản hơn mô hình trước bằng cách là nó giới hạn tính phụ thuộc giữa các biến đến 1 dãy Markov (Markov chain) thay vì dùng cây (tree).

Nhờ đó, mô hình này lại gián tiếp liên quan đến mô hình ngôn ngữ n-gram. Cũng trong công trình này, các tác giả đã đề xuất mô hình ngôn ngữ thống kê n-gram kết hợp phương pháp Naive Bayes ứng dụng cho bài toán phân loại văn bản.

1.2.7 . Tiếp cận theo hướng kết hợp 2 loại trên

1.2.7.1 Giới thiệu về cụm từ hay ngữ trong văn bản

Về ý nghĩa ngôn ngữ, các cụm từ này thường mang ý nghĩa cao hơn từ nhưng lại không đầy đủ như câu. Những cụm từ như vậy được gọi là "cụm từ ngữ pháp", tức là được cấu thành theo đúng cấu trúc ngữ pháp của ngôn ngữ.

1.2.7.2 Vector hóa văn bản bằng phương pháp ngữ thống kê

Trong cách tiếp cận này, tác giả đã dùng ngữ thống kê, hay còn được gọi là n-gram để xây dựng vector đặc trưng cho văn bản cần phân loại.

1.3 BÀI TOÁN TÁCH TỪ TRONG TIẾNG VIỆT

1.3.1 Khái niệm về từ

Theo một số nhà ngôn ngữ học có một số khái niệm tiêu biểu sau đây về từ :

B.Golovin quan niệm: "từ là đơn vị nhỏ nhất có nghĩa của ngôn ngữ, được vận dụng độc lập, tái hiện tự do trong lời nói để xây dựng nên câu".

Còn Solncev thì lại quan niệm: "Từ là đơn vị ngôn ngữ có tính hai mặt: âm và nghĩa. Từ có khả năng độc lập về cú pháp khi sử dụng trong lời"

Trong tiếng Việt, cũng có nhiều định nghĩa về từ như:

Theo Trương Văn Trình và Nguyễn Hiến Lê thì: “Từ là âm có nghĩa, dùng trong ngôn ngữ để diễn tả một ý đơn giản nhất, nghĩa là ý không thể phân tích ra được”.

Nguyễn Kim Thản thì định nghĩa: “Từ là đơn vị cơ bản của ngôn ngữ, có thể tách khỏi các đơn vị khác của lời nói để vận dụng một cách độc lập và là một khối hoàn chỉnh về ý nghĩa (từ vựng hay ngữ pháp) và cấu tạo”.

1.3.2 Các vấn đề trong bài toán tách từ Tiếng Việt

1.3.2.1 Xử lý nhập nhằng

Nhập nhằng trong tách từ được phân thành hai loại :

- Nhập nhằng chồng (Overlapping Ambiguity)
- Nhập nhằng hợp (Combination Ambiguity)

1.3.3 Lịch sử nghiên cứu đối với bài toán tách từ

- Công trình của nhóm **VCL**(http://vcllab.com
- Công trình của tác giả Lê Hà An [Lê An Hà, 2003]
- Công trình của [H. Nguyen, 2005]
- Công trình “**Hệ phân tách từ Việt**” nằm trong nhóm sản phẩm của đề tài KC01.01/06-10 về:"**Nghiên cứu và phát triển một số sản phẩm thiết yếu về xử lý tiếng nói và văn bản tiếng Việt**" viết tắt là VLSP.

1.3.4 Các hướng tiếp cận chính cho bài toán tách từ

- Tiếp cận dựa vào từ điển cố định.
- Tiếp cận dựa vào thống kê thuần túy.
- Tiếp cận dựa trên cả hai phương pháp trên.

1.3.5 Chuyển dịch trạng thái hữu hạn có trọng số

Đây có thể được xem là mô hình tách từ đầu tiên dành cho tiếng Việt. Mô hình này là một cải tiến của mô hình WFST (Weighted Finite State Transducer) của [Richard, 1996] áp dụng cho tiếng Trung Quốc để phù hợp hơn với tiếng Việt.

1.3.6 So khớp tối đa (MM: Maximum Matching)

Maximum Matching (MM) được xem như là phương pháp tách từ dựa trên từ điển đơn giản nhất. MM cố gắng so khớp với từ dài nhất có thể có trong từ điển. Đó là một thuật toán ăn tham (Greedy Algorithms) nhưng bằng thực nghiệm đã chứng minh được rằng thuật toán này đạt được độ chính xác> 90% nếu từ điển đủ lớn [Chih-Hao, Tsai, 2000].

1.3.7 MMSeg (Maximum Maching Segment)

Thực chất đây là mô hình tách từ của tiếng Trung Quốc và mô hình này cho kết quả tương đối khả quan (94,5% trên tiếng Việt). Nói chính xác là đây là mô hình bổ sung cho mô hình MM sử dụng thêm 1 số luật heuristic trên ngôn ngữ để đánh giá dựa trên 2 mô hình MM

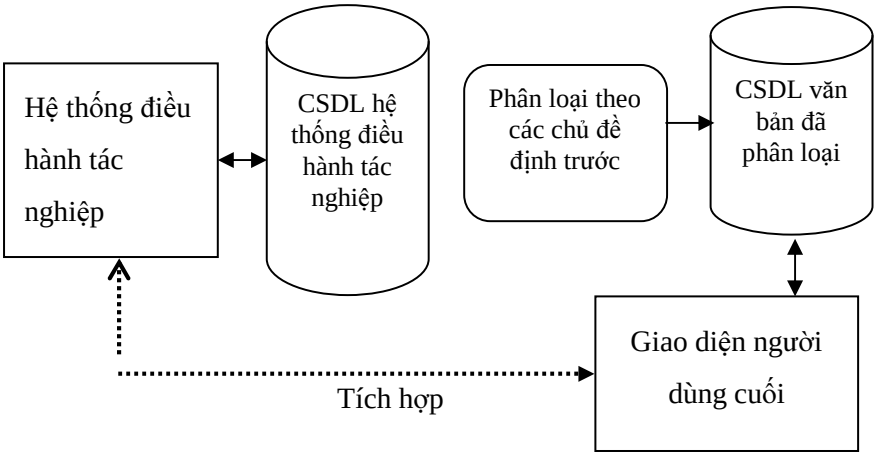
1.3.8 Maximum Entropy

Mô hình tách từ bằng phương pháp Maximum Entropy dựa trên ý tưởng của mô hình gán nhãn từ loại (POS Tagger) dùng phương pháp Maximum Entropy cho tiếng Anh của Adwait Ratnaparkhi. Các tác giả của công trình đã cài đặt thành công mô hình này cho tiếng Việt.

CHƯƠNG 2
PHÂN TÍCH VÀ THIẾT KẾ HỆ THỐNG

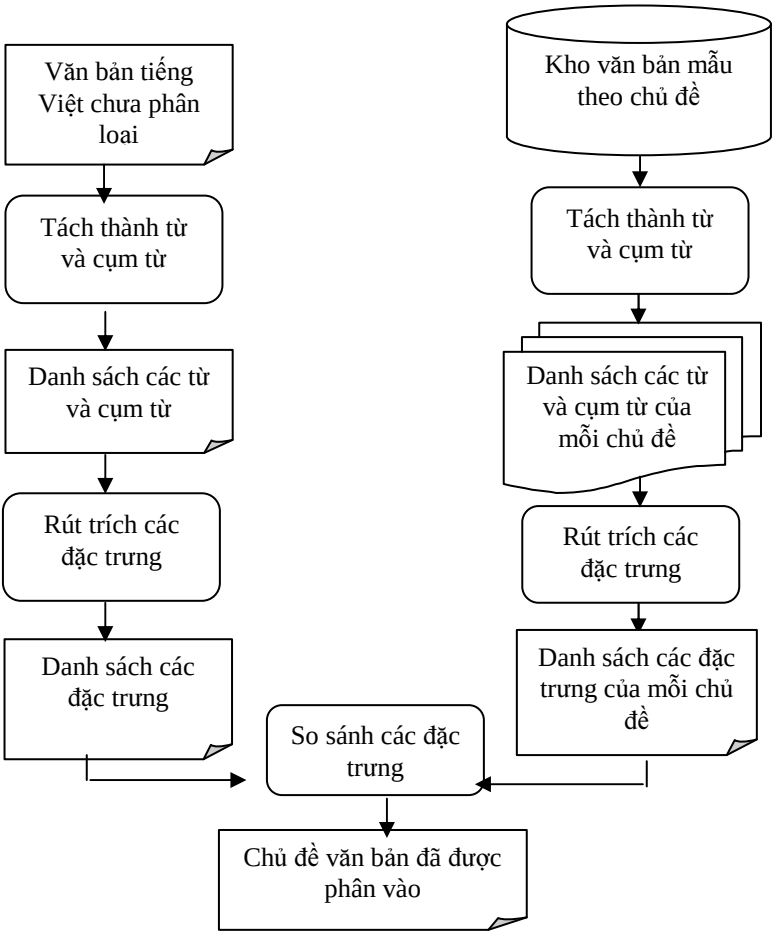
2.1 PHÂN TÍCH HỆ THỐNG

2.1.1 Kiến trúc tổng quát của hệ thống



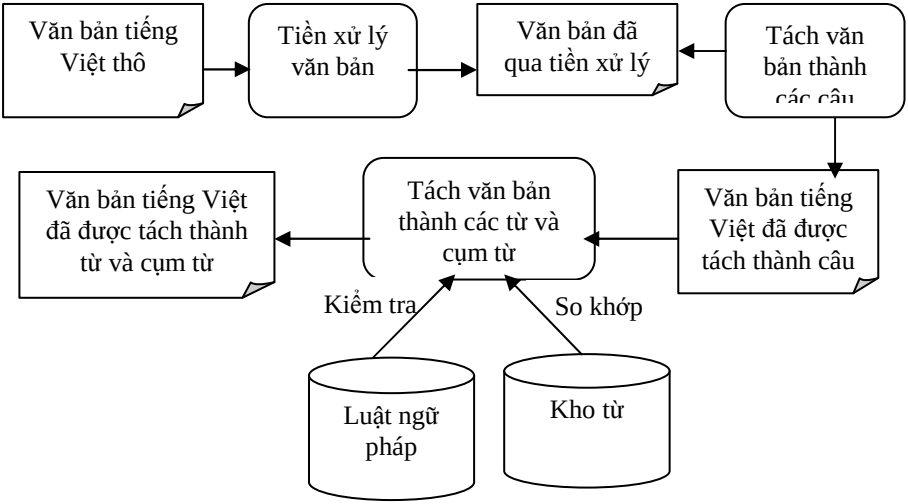
Hình 2.1 Mô hình tng quan của hệ thống

2.1.1.1 .Phân loại văn bản



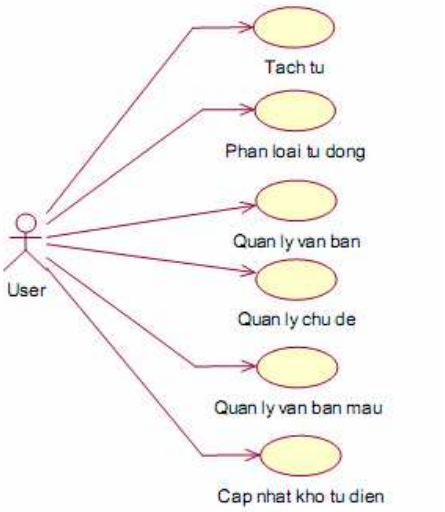
Hình 2.2 Quá trình phân loại tài liệu

2.1.1.2 Tách từ



Hình 2.3 Quá trình tách từ tiếng Việt thành các từ và cụm từ

2.1.2 Các chức năng chính của chương trình



Hình 2.4 Các chức năng chính của chương trình

2.2 THIẾT KẾ HỆ THỐNG

2.2.1 Thiết kế cơ sở dữ liệu

Loại văn bản

Bảng 2.2. Loại văn bản

Tên cột	Kiểu dữ liệu	Độ dài	Ràng buộc	Chú thích
LoaiVanBanID	Int		Khóa chính	Tự động tăng và Lưu trữ thông tin mã loại văn bản
TenLoaiVanBan	nvarchar	255		Lưu trữ thông tin về tên loại văn bản

Từ khóa

Bảng 2.3. Từ Khóa

Tên cột	Kiểu dữ liệu	Độ dài	Ràng buộc	Chú thích
TuKhoaID	int		Khóa chính	Tự động tăng và Lưu trữ thông tin mã từ khóa
LoaiVanBanID	int	255	Khóa ngoại	Lưu trữ thông tin về mã loại văn bản
TuKhoa	nvarchar	255		Lưu trữ Thông tin về từ khóa
HeSo	float			Hệ số mỗi từ khóa, bình thường là 1, hệ số càng cao sẽ là đặc trưng cho mỗi loại văn bản Công thức chung để tính toán: $\sum_1^n SoTuKhoa * HeSo$

Phân loại

Bảng 2.4. Phân loại

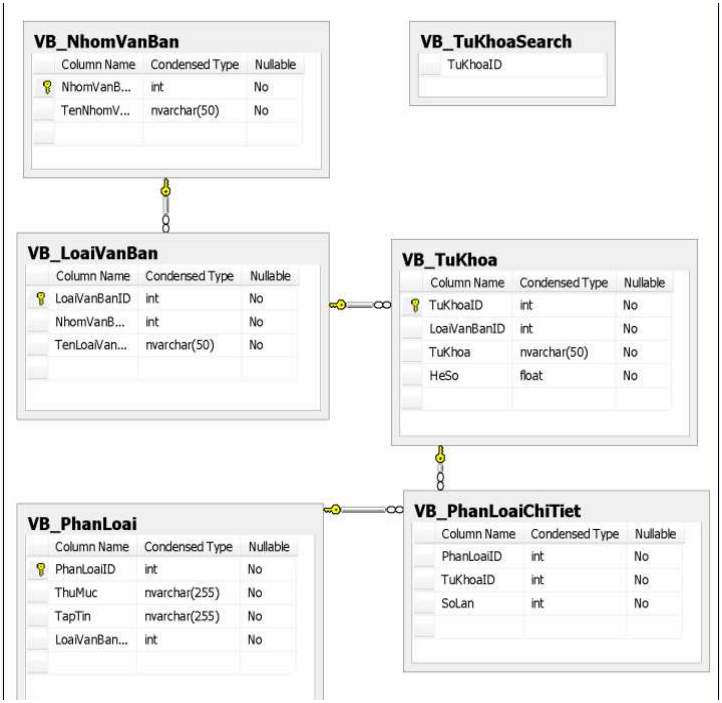
Tên cột	Kiểu dữ liệu	Độ dài	Ràng buộc	Chú thích
PhanLoaiID	int		Khóa chính	Tự động tăng và Lưu trữ thông tin mã phân loại văn bản
ThuMuc	nvarchar	255		Lưu trữ thông tin về Địa chỉ thư mục lưu trữ văn bản
TapTin	nvarchar	255		Lưu trữ Tên tập tin văn bản lưu trữ
LoaiVanBanID	int		Khóa ngoại	Văn bản thuộc loại văn bản nào

Phân loại chi tiết

Bảng 2.5 Phân loại chi tiết

Tên cột	Kiểu dữ liệu	Độ dài	Ràng buộc	Chú thích
PhanLoaiID	int		Khóa ngoại	ID phân loại để xác định tập tin văn bản
TuKhoaID	int		Khóa ngoại	ID từ khóa được tìm thấy trong văn bản
SoLan	int			Số lần xuất hiện từ khóa trong văn bản

2.2.2 Lược đồ cơ sở dữ liệu quan hệ



Hình 2.11 Lược đồ cơ sở dữ liệu quan hệ

2.2.3 Thiết kế giao diện người sử dụng

*** Phân loại:**

- Chọn văn bản: mở tập tin Microsoft Word để phân loại văn bản

- Lưu trữ: lưu kết quả phân loại văn bản vào cơ sở dữ liệu. Cho phép chọn văn bản thuộc loại văn bản nào (mặc định chọn loại văn bản có trọng số lớn nhất). Cho phép chọn thư mục lưu trữ văn bản.

- Xem chi tiết: hiển thị thông tin chi tiết từng từ khóa đã tìm thấy từ tập tin văn bản, các thong tin về từ khóa, tần số xuất hiện, trọng số của từ khóa.

- Kết quả văn bản mới: xem kết quả phân tích từ văn bản mới đã được chọn.

*** Xuất dữ liệu:**

- MS Excel: xuất kết quả phân loại bên dưới sang định dạng Microsoft Excel

- MS Word: xuất kết quả phân loại bên dưới sang định dạng Microsoft Word

- PDF: xuất kết quả phân loại bên dưới sang định dạng PDF

2.2.4 Môi trường và công cụ phát triển hệ thống

Phần mềm thực thi trên các phiên bản hệ điều hành Windows.

Công cụ phát triển:

- Microsoft Visual Studio.NET 2008
- Microsoft .NET Framework 2.0
- Microsoft SQL Server 2005

Công cụ hỗ trợ:

- *DevExpress.NET 8.1.2*

CHƯƠNG 3
XÂY DỰNG ỨNG DỤNG VÀ ĐÁNH GIÁ KẾT QUẢ THỬ NGHIỆM

3.1 XÁC ĐỊNH NGUỒN VĂN BẢN (NGỮ LIỆU)

3.1.1 Chọn kho ngữ liệu

Tôi tiến hành xây dựng ngữ liệu chuẩn cho tiếng Việt dựa trên nguồn tài nguyên chính là "Hệ thống điều hành tác nghiệp tại sở thông tin và truyền thông Quảng Nam."

3.1.2 Văn bản cần phân loại

Văn bản tiếng Việt ở đây là "Hệ thống điều hành tác nghiệp tại sở thông tin và truyền thông Quảng Nam". Vì vậy, dựa vào những đặt điểm của nguồn tôi qui ước một số ràng buộc chung cho dữ liệu đầu vào như sau:

Định dạng file:.txt

- .dpf
- .doc

Chuẩn chính tả: các văn bản được đưa vào phân loại phải đúng chính tả.

Các văn bản không nên quá dài, mà cụ thể ở đây <1000 từ. Điều này cho phép thời gian thực hiện của chương trình có thể ổn định trong khoảng thời gian chấp nhận được. Mỗi lần thực hiện, chương trình phân loại cho 1 văn bản.

3.2 XỬ LÝ NGUỒN NGỮ LIỆU

3.2.1 Kho ngữ liệu

Kho ngữ liệu mà tôi thu thập được dựa trên các văn bản trong hệ thống Q-Office.

Kết quả là tôi có được bộ ngữ liệu tương đầy đủ, số lượng khoảng ~400 tài liệu được chia thành 2 dạng:

Dạng thô (raw form): Gồm 4 chủ đề lớn chia thành: Công nghệ thông tin, Bưu chính viễn thông, Báo chí xuất bản, thanh tra.

Dạng mịn (smooth form): Gồm những chủ đề nhỏ chia thành: Phần cứng, phần mềm, bưu chính, viễn thông, phát thanh-truyền hình,

3.2.2 Văn bản đầu vào

Văn bản tiếng Việt sau khi được lấy về sẽ được xử lí để đưa vào bước tách từ. Lưu ý: văn bản tiếng Việt mà chúng ta cần phân loại có thể chứa cả hình ảnh, bản biểu, biểu đồ…Trong quá trình phân loại, các thành phần này sẽ được loại bỏ, ta chỉ quan tâm đến văn bản thuần dạng DOC

Văn bản dạng này có thể bao gồm các thẻ định dạng như HTML. Tuy nhiên, nó có thể bao gồm các bản biểu, bểu đồ, hình ảnh.v.v.

3.2.3 Tổ chức dữ liệu

Cơ sở dữ liệu được lưu trữ tại bất kỳ nơi nào trên ổ đĩa cứng, Attack cơ sở dữ liệu vào hệ quản trị dữ liệu Microsoft SQL Server 2005. Thư viện sử dụng trong chương trình được lưu trữ tại thư mục chứa phần mềm thực thi (phanloaivanban.exe).

Các văn bản sau khi phân loại được lưu trữ vào các thư mục tại thư mục chứa phần mềm thực thi (phanloaivanban.exe).

KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

1. Kết luận

Phân loại văn bản là một bài toán mà đến bây giờ vẫn còn nhiều đáp án khác nhau. Tuy nhiên, bên cạnh vấn đề khó khăn trong xử lý ngôn ngữ, chúng ta vẫn có thể tìm thấy nhiều bất cập. Chính những bất cập này là động lực thôi thúc tôi tham gia nghiên cứu, góp phần giải quyết những vấn đề cần phải làm sáng tỏ.

Ứng dụng tôi xây dựng ở đây đã có thể đạt được một số kết quả:

Việc xây dựng thành công công cụ *Phân loại văn bản tự động trong Hệ thống Điều hành tác nghiệp tại Sở Thông tin và Truyền thông Quảng Nam* sẽ có thể được áp dụng vào nhiều ứng dụng cụ thể trong đời sống, góp phần giảm thiểu sự tiêu tốn về thời gian và công sức con người. Đồng thời, việc xây dựng thành công công cụ tách từ hiệu quả trên văn bản tiếng Việt mở ra điều kiện thuận lợi cho các bài toán xử lý ngôn ngữ tự nhiên khác trên tiếng Việt.

Đề tài cũng đóng góp một hướng tiếp cận mới giúp giải quyết các bài toán xử lý ngôn ngữ tự nhiên trong điều kiện chúng ta chưa có một công cụ tách từ cho độ chính xác.

Tuy nhiên, trong giới hạn một ứng dụng của đề tài này, thì chương trình vẫn còn những điểm chưa đạt:

Chưa tích hợp trên hệ thống điều hành tác nghiệp, xem như cơ sở dữ liệu đã được lấy về phục vụ cho việc phân loại tự động.

Một số trường hợp tên riêng vẫn chưa xử lý được. Ứng dụng vẫn còn phụ thuộc quá nhiều vào từ điển ở phần tách từ nên nếu từ điển không đầy đủ thì kết quả tách từ sẽ không chính xác, dẫn đến việc phân loại văn bản không có kết quả cao.

2. Hướng phát triển

Trong công cụ *Phân loại văn bản tự động trong Hệ thống Điều hành tác nghiệp tại Sở Thông tin và Truyền thông Quảng Nam*, vấn đề tách tên riêng vẫn là một thách thức. Chính vì sự không thống nhất trong cách viết của người Việt nên vấn đề tên riêng càng trở thành một vấn đề thật sự nan giải. Do vậy, để có thể giải quyết tốt bài toán tách từ tiếng Việt, chúng ta phải bắt tay vào giải quyết vấn đề tách tên riêng. Trong bài toán phân loại văn bản, vấn đề phát triển tiếp theo là xây dựng khả năng học tăng cường cho bộ phân loại. Trong thực tế, số chủ đề phân loại rất phong phú, vì vậy khả năng học tăng cường sẽ giải quyết tốt các nhu cầu của mọi người khi áp dụng vào thực tế. Ngoài ra, tôi cũng sẽ tiến hành nghiên cứu cho hệ thống của mình khả năng học tăng cường (online learning) thích nghi với mọi điều kiện thay đổi của thực tế.

Mục tiêu phát triển tiếp theo của tôi là hoàn thiện hệ thống của mình để có thể truy tìm thông tin online hay nói đúng hơn là một động cơ tìm kiếm cho tiếng Việt với hai tiêu chí: tốc độ nhanh hơn, độ chính xác cao hơn, tùy khuôn khổ ứng dụng và nhu cầu của người sử dụng.

Tích hợp module phân loại văn bản trên hệ thống điều hành tác nghiệp để có thể rút trích thông tin một cách chính xác. Đó cũng là một hướng phát triển mà tôi quan tâm. Nguyên cứu thêm các luật ngữ pháp trong tiếng Việt để khử nhập nhằng hiệu quả trong bài toán tách từ