TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN HỆ THỐNG THÔNG TIN

SINH VIÊN THỰC HIỆN

NGUYỄN TRẦN THIÊN THANH - TRẦN KHẢI HOÀNG

TÌM HIỂU CÁC HƯỚNG TIẾP CẬN BÀI TOÁN PHÂN LOẠI VĂN BẢN VÀ XÂY DỰNG PHẦN MỀM PHÂN LOẠI TIN TỨC BÁO ĐIỆN TỬ

KHÓA LUẬN CỬ NHÂN TIN HỌC

Tp.HCM, 2005

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN
BỘ MÔN HỆ THỐNG THÔNG TIN

SINH VIÊN THỰC HIỆN

- **NGUYỄN TRẦN THIÊN THANH - 0112243**
- **TRẦN KHẢI HOÀNG - 0112305**

TÌM HIỂU CÁC HƯỚNG TIẾP CẬN BÀI TOÁN PHÂN LOẠI VĂN BẢN VÀ XÂY DỰNG PHẦN MỀM PHÂN LOẠI TIN TỨC BÁO ĐIỆN TỬ

KHÓA LUẬN CỬ NHÂN TIN HỌC

GIÁO VIÊN HƯỚNG DẪN

Cử nhân : NGUYỄN VIỆT THÀNH

Thạc sĩ : NGUYỄN THANH HÙNG

Niên khóa 2001-2005

LỜI CẢM ƠN

Chúng em xin gửi lời cảm ơn chân thành và sâu sắc nhất đến thầy Nguyễn Việt Thành và thầy Nguyễn Thanh Hùng đã tận tụy hướng dẫn, động viên, giúp đỡ chúng em trong suốt thời gian thực hiện đề tài.

Chúng em xin chân thành cảm ơn quý Thầy Cô trong Khoa Công Nghệ Thông Tin truyền đạt kiến thức quý báu cho chúng em trong những năm học vừa qua.

Chúng con xin nói lên lòng biết ơn đối với Ông Bà, Cha Mẹ luôn là nguồn chăm sóc, động viên trên mỗi bước đường học vấn của chúng con.

Xin chân thành cám ơn các anh chị và bạn bè đã ủng hộ, giúp đỡ và động viên chúng em trong thời gian học tập và nghiên cứu.

Mặc dù chúng em đã cố gắng hoàn thành luận văn trong phạm vi và khả năng cho phép nhưng chắc chắn sẽ không tránh khỏi những thiếu sót. Chúng em kính mong nhận được sự cảm thông và tận tình chỉ bảo của quý Thầy Cô và các bạn.

Sinh viên thực hiện,
Nguyễn Trần Thiên Thanh & Trần Khải Hoàng
07/2005

LỜI NÓI ĐẦU

Trong những năm gần đây, sự phát triển vượt bậc của công nghệ thông tin đã làm tăng số lượng giao dịch thông tin trên mạng Internet một cách đáng kể đặc biệt là thư viện điện tử, tin tức điện tử.... Do đó mà số lượng văn bản xuất hiện trên mạng Internet cũng tăng theo với một tốc độ chóng mặt. Theo số lượng thống kê từ Broder et al (2003), lượng thông tin đó lại tăng gấp đôi sau từ 9 đến 12 tháng, và tốc độ thay đổi thông tin là cực kỳ nhanh chóng.

Với lượng thông tin đồ sộ như vậy, một yêu cầu lớn đặt ra đối với chúng ta là làm sao tổ chức và tìm kiếm thông tin có hiệu quả nhất. Phân loại thông tin là một trong những giải pháp hợp lý cho yêu cầu trên. Nhưng một thực tế là khối lượng thông tin quá lớn, việc phân loại dữ liệu thủ công là điều không tưởng. Hướng giải quyết là một chương trình máy tính tự động phân loại các thông tin trên.

Chúng em đã tập trung thực hiện đề tài "***Tìm hiểu các hướng tiếp cận cho bài toán phân loại văn bản và xây dựng ứng dụng phân loại tin tức báo điện tử***" nhằm tìm hiểu và thử nghiệm các phương pháp phân loại văn bản áp dụng trên tiếng Việt. Để thực hiện việc phân loại, điều bắt buộc đối với tiếng Việt đó là việc tách từ. Trong luận văn này, chúng em cũng tìm hiểu một số cách tách từ tiếng Việt và thử nghiệm một phương pháp tách từ mới thích hợp cho việc phân loại mà không dùng bất kỳ từ điển hoặc tập ngữ liệu nào. Cuối cùng, chúng em xây dựng phần mềm phân loại văn bản tích hợp vào trang web "Toà soạn báo điện tử" (Luận văn khoá 2000 - Hoàng Minh Ngọc Hải (0012545), Nguyễn Duy Hiệp (0012038)) nhằm phục vụ cho việc phân loại tin tức báo điện tử.

Hiện nay, trang web của khoa chúng ta vẫn chưa thực hiện được việc phân loại tự động các tin tức lấy về, do đó gây ra rất nhiều lãng phí về thời gian và công sức của nhà quản trị cũng như làm giới hạn việc thu thập tin tức từ nhiều nguồn khác nhau. Ứng dụng phân loại tin tức báo điện tử tích hợp với việc lấy tin tức tự động của chúng em hy vọng sẽ đem đến một cách quản trị mới, nhanh chóng và hiệu quả hơn cách lấy tin truyền thống. Ngoài ra, trong điều kiện cần cập nhật thông tin một

cách nhanh chóng như hiện nay, phần mềm phân loại văn bản tự động của chúng em còn có khả năng ứng dụng cho nhiều loại trang báo điện tử tiếng Việt khác.

Nội dung của luận văn được trình bày bao gồm 8 chương; trong đó, 3 chương đầu trình bày các hướng tiếp cận cho phân loại văn bản và tách từ tiếng Việt hiện nay; 2 chương tiếp theo trình bày hướng tiếp cận của luận văn đối với phân loại văn bản và tách từ tiếng Việt; 3 chương cuối trình bày hệ thống thử nghiệm văn bản, ứng dụng vào phân loại tin tức bán tự động, và cuối cùng là đánh giá, kết luận quá trình nghiên cứu của luận văn.

- **Chương 1. Tổng quan:** giới thiệu sơ lược về các phương pháp phân loại văn bản và các hướng tiếp cận cho việc tách từ tiếng Việt; đồng thời xác định mục tiêu của đề tài.
- **Chương 2. Một số phương pháp phân loại văn bản:** giới thiệu tóm tắt một số phương pháp phân loại văn bản dành cho tiếng Anh.
- **Chương 3. Phương pháp tách từ tiếng Việt hiện nay:** trình bày tóm tắt một số phương pháp tách từ tiếng Việt hiện nay, ưu điểm và hạn chế của các phương pháp đó.
- **Chương 4. Phương Tách từ Tiếng Việt không dựa trên tập ngữ liệu đánh dấu (annotated corpus) hay từ điển (lexicon) – Một thách thức:** trình bày phương pháp tách từ tiếng Việt mới chỉ dựa vào việc thống kê từ Internet thông qua Google mà không cần bất kỳ từ điển hay tập ngữ liệu nào.
- **Chương 5. Bài toán phân loại tin tức báo điện tử:** trình bày hướng tiếp cận cho bài toán phân loại tin tức báo điện tử.
- **Chương 6. Hệ thống thử nghiệm phân loại văn bản:** giới thiệu về hệ thống thử nghiệm các phương pháp tách từ và phân loại văn bản do chúng em xây dựng. Ngoài ra, trong chương 6, chúng em trình bày về dữ liệu dùng để thử nghiệm và các kết quả thử nghiệm thu được.
- **Chương 7. Ứng dụng phân loại tin tức báo điện tử bán tự động:** giới thiệu ứng dụng phân loại tin tức báo điện tử do chúng em xây dựng tích hợp

trên trang web do luận văn "Tòa soạn báo điện tử" khóa 2000 xây dựng của sinh viên Hoàng Minh Ngọc Hải (0012545), Nguyễn Duy Hiệp (0012038)

- **Chương 8. Tổng kết:** là chương cuối cùng của đề tài, tóm lại các vấn đề đã giải quyết và nêu một số hướng phát triển trong tương lai.

MỤC LỤC

DANH SÁCH HÌNH

DANH SÁCH BẢNG

Chương 1
TỔNG QUAN

Đặt vấn đề

Các phương pháp phân loại văn bản

Tách từ tiếng Việt – Một thách thức thú vị

Mục tiêu của luận văn

- Phần tìm hiểu các thuật toán phân loại văn bản
- Phần tách từ tiếng Việt
- Phần mềm phân loại tin tức báo điện tử bán tự động

Chương 1. TỔNG QUAN

1.1. Đặt vấn đề

Trong thời đại bùng nổ công nghệ thông tin hiện nay, phương thức sử dụng giấy tờ trong giao dịch đã dần được số hoá chuyển sang các dạng văn bản lưu trữ trên máy tính hoặc truyền tải trên mạng. Bởi nhiều tính năng ưu việt của tài liệu số như cách lưu trữ gọn nhẹ, thời gian lưu trữ lâu dài, tiện dụng trong trao đổi đặc biệt là qua Internet, dễ dàng sửa đổi… nên ngày nay, số lượng văn bản số tăng lên một cách chóng mặt đặc biệt là trên world-wide-web. Cùng với sự gia tăng về số lượng văn bản, nhu cầu tìm kiếm văn bản cũng tăng theo. Với số lượng văn bản đồ sộ thì việc phân loại văn bản tự động là một nhu cầu bức thiết.

Tại sao phải phân loại văn bản tự động? Việc phân loại văn bản sẽ giúp chúng ta tìm kiếm thông tin dễ dàng và nhanh chóng hơn rất nhiều so với việc phải bới tung mọi thứ trong ổ đĩa lưu trữ để tìm kiếm thông tin. Mặt khác, lượng thông tin ngày một tăng lên đáng kể, việc phân loại văn bản tự động sẽ giúp con người tiết kiệm được rất nhiều thời gian và công sức.

Do vậy, các phương pháp phân loại văn bản tự động đã ra đời để phục vụ cho nhu cầu chính đáng đó.

1.2. Các phương pháp phân loại văn bản

Theo Yang & Xiu (1999), "*việc phân loại văn bản tự động là việc gán các nhãn phân loại lên một văn bản mới dựa trên mức độ tương tự của văn bản đó so với các văn bản đã được gán nhãn trong tập huấn luyện*".

Từ trước đến nay, phân loại văn bản tự động trong tiếng Anh đã có rất nhiều công trình nghiên cứu và đạt được kết quả đáng khích lệ. Dựa trên các thống kê của Yang & Xiu (1999) và nghiên cứu của chúng em, một số phương pháp phân loại thông dụng hiện nay là: *Support Vector Machine* [Joachims, 1998], *k-Nearest Neighbor* [Yang, 1994], *Linear Least Squares Fit* [Yang and Chute, 1994] *Neural Network* [Wiener et al, 1995], *Naïve Bayes* [Baker and Mccallum, 2000], *Centroid-based* [Shankar and Karypis, 1998]. Các phương pháp trên đều dựa vào xác suất

thống kê hoặc thông tin về trọng số của từ trong văn bản. Chi tiết về ý tưởng và công thức tính toán của mỗi phương pháp sẽ được chúng em trình bày ở chương 3, mục 3.3.

Mỗi phương pháp phân loại văn bản đều có cách tính toán khác nhau, tuy nhiên, nhìn một cách tổng quan thì các phương pháp đó đều phải thực hiện một số bước chung như sau: đầu tiên, mỗi phương pháp sẽ dựa trên các thông tin về sự xuất hiện của *từ* trong văn bản (ví dụ tần số, số văn bản chứa từ…) để biểu diễn văn bản thành dạng vector; sau đó, tuỳ từng phương pháp mà ta sẽ áp dụng công thức và phương thức tính toán khác nhau để thực hiện việc phân loại.

Đối với tiếng Anh, các kết quả trong lĩnh vực này rất khả quan, còn đối với tiếng Việt, các công trình nghiên cứu về phân loại văn bản gần đây đã có một số kết quả ban đầu nhưng vẫn còn nhiều hạn chế. Nguyên nhân là ngay ở bước đầu tiên, chúng ta đã gặp khó khăn trong việc xử lý văn bản để rút ra tần số xuất hiện của từ. Trong khi đó, để phân loại văn bản thì có thể nói bước đầu tiên là quan trọng nhất bởi vì nếu ở bước tách từ đã sai thì việc phân loại hầu như không thể thành công được. Phần trình bày tiếp theo sẽ cho chúng ta biết những *thách thức* đặt ra trong việc tách từ tiếng Việt, cũng như những ứng dụng thú vị của nó.

1.3. Tách từ Tiếng Việt – Một thách thức thú vị

Đối với tiếng Anh, "*từ là một nhóm các ký tự có nghĩa được tách biệt với nhau bởi khoảng trắng trong câu*" (Webster Dictionary), do vậy việc tách từ trở nên rất đơn giản. Trong khi đối với tiếng Việt, ranh giới từ không được xác định mặc định là khoảng trắng mà tùy thuộc vào ngữ cảnh dùng câu tiếng Việt. Ví dụ các từ trong tiếng Anh là *"book", "cat", "stadium"* thì trong tiếng Việt là *"quyển sách", "con mèo", "sân vận động"* … Vấn đề trên thực sự đưa ra một ***thách thức*** đối với chúng ta - những người làm tin học.

Tuy nhiên, thách thức nào cũng có cái ***thú vị*** của nó. Nếu chúng ta giải quyết được việc tách từ một cách thoả đáng, thì thành quả mà chúng ta đạt được là một nền tảng để phát triển cho các hướng nghiên cứu khác có liên quan đến việc xử lý ngôn ngữ tự nhiên như: phân loại văn bản, dịch tự động, kiểm tra lỗi chính tả, kiểm

tra ngữ pháp… Đó là các ứng dụng rất thiết thực với đời sống con người và là mục tiêu của con người đang chinh phục.

Một số nước châu Á như Trung Quốc, Nhật Bản, Hàn Quốc, Việt Nam sử dụng loại hình ngôn ngữ gần như tương tự nhau về mặt hình thái và cú pháp. Do đó ta có thể áp dụng, cải tiến một số phương pháp tách từ của các nước bạn đặc biệt là Trung Quốc vào việc tách từ tiếng Việt.

Theo Đinh Điền (2004), các phương pháp tách từ sau có nguồn gốc từ tiếng Hoa đã được thử nghiệm trên tiếng Việt : *Maximum Matching: forward/backward* hay còn gọi LRMM (Left Right Maximum Matching); giải thuật học cải biến *TBL*; mạng chuyển dịch trạng thái hữu hạn có trọng số WFST (Weighted finite-state Transducer); giải thuật dựa trên nén (compression);….Theo các cách tiếp cận trên, điều kiện quan trọng cần có là một hệ thống từ điển (LRMM) và ngữ liệu đánh dấu (TBL, WFST) đầy đủ, chuẩn xác. Một từ điển hay một tập ngữ liệu không hoàn chỉnh sẽ làm giảm hiệu suất của thuật toán.

Tuy nhiên, khó có thể tạo ra được một từ điển hoàn chỉnh nhất là trong thời đại ngày nay, ngôn ngữ còn tiếp tục phát triển và thay đổi từng ngày. Xét về mặt phổ biến, tiếng Anh là ngôn ngữ được dùng rộng rãi trong giao dịch trên thế giới. Do đó để tạo ra một tập ngữ liệu tiếng Anh thỏa các tiêu chí chọn mẫu ngữ liệu là không quá phức tạp. Trong khi đó, Việt Nam chỉ mới cho phép truy cập Internet trong vòng chục năm trở lại đây, do đó số lượng trang web tiếng Việt là không nhiều. Cho đến nay, vẫn chưa có một tập ngữ liệu huấn luyện chuẩn nào dành cho việc tách từ và phân loại trang web tiếng Việt được công bố.

Gần đây, một phương pháp tách từ mới được giới thiệu có ưu điểm là không cần dùng tập ngữ liệu hay từ điển để lấy thông tin thống kê hay trọng số của từ, đó là phương pháp Internet and Genetics Algorithm-based Text Categorization (IGATEC) của H. Nguyen et al (2005). Điểm sáng tạo của thuật toán là kết hợp thuật toán di truyền với việc trích xuất thông tin thống kê từ Internet thông qua một công cụ tìm kiếm (như Google chẳng hạn) thay vì lấy từ tập ngữ liệu như các phương pháp trước.

Chúng em thực hiện bước tách từ trong luận văn này dựa trên ý tưởng của thuật toán IGATEC nhưng có bổ sung nhiều cải tiến đáng kể để tăng độ chính xác đồng thời thực hiện các thí nghiệm chi tiết nhằm so sánh các cách áp dụng thuật toán để tìm ra cách tối ưu nhất.

1.4. Mục tiêu của luận văn

1.4.1. Phần tìm hiểu các thuật toán phân loại văn bản

Trong khuôn khổ luận văn này, chúng em tìm hiểu ở mức cơ bản một số phương pháp phân loại văn bản hiện có đang áp dụng cho tiếng Anh và đưa ra một số so sánh nhất định giữa các phương pháp: *Support Vector Machine* (Joachims, 1998), *k-Nearest Neighbor* (Yang, 1994), *Linear Least Squares Fit* (Yang and Chute, 1994) *Neural Network* (Wiener et al, 1995), *Naïve Bayes* (Baker and Mccallum, 2000), *Centroid-based* (Shankar and Karypis, 1998).

Sau đó, chúng em sẽ chọn và áp dụng một phương pháp cho bài toán phân loại tin tức báo điện tử tiếng Việt chấp nhận được, phù hợp với mức độ và thời gian cho phép của một luận văn đại học.

1.4.2. Phần tách từ tiếng Việt

Hiện nay các phương pháp tách từ tiếng Việt được công bố vẫn chưa nhiều và hướng tiếp cận chủ yếu dựa vào tập huấn luyện và từ điển. Như chúng ta đã biết, việc tạo ra hệ thống dữ liệu đó không phải là một sớm một chiều, mà yêu cầu đầu tư khá nhiều công sức, thời gian và tiền bạc.

Trong luận văn này, chúng em cố gắng tìm hiểu, cải tiến, cài đặt, thử nghiệm một phương pháp tách từ tiếng Việt theo hướng tiếp cận IGATEC, có độ chính xác chấp nhận được, và điều quan trọng là không cần dùng tập ngữ liệu (corpus) để phân định ranh giới từ.

Sau đó, chúng em sẽ cài đặt, thử nghiệm độ chính xác của phương pháp tách từ này trong khía cạnh phân loại văn bản

1.4.3. Phần mềm phân loại tin tức báo điện tử bán tự động

Để thử nghiệm hướng nghiên cứu tách từ tiếng Việt và phân loại văn bản của luận văn, chúng em tích hợp phần mềm phân loại tin tức vào trang web báo điện tử có sẵn được xây dựng trên nền DotNetNuke Portal của luận văn khoá 2000 (Hoàng Minh Ngọc Hải (0012545), Nguyễn Duy Hiệp (0012038))

Như chúng ta đều biết, điều kiện mạng cung cấp cho các trường đại học ở nước ta hiện nay là khá hạn chế, khó đáp ứng được hoàn toàn việc cho phép các sinh viên lên mạng Internet để xem các tin tức mới hằng ngày. Để giải quyết phần nào vấn đề trên, chúng ta có thể chọn lọc một số tin tức từ các nguồn khác, đăng tải trên trang web nội bộ của trường. Trên cơ sở đó, chúng em tích hợp phần mềm phân loại tin tức báo điện tử tự động vào toà soạn báo điện tử cho phép lấy tin tự động từ các trang web khác. Nhờ vậy, công việc lấy tin và phân loại tin tức giờ đây đã trở nên rất dễ dàng và nhanh chóng, tiết kiệm nhiều công sức và thời gian cho nhà quản trị.

Không chỉ ứng dụng cho các trường đại học, phần mềm phân loại tin tức của chúng em còn có thể ứng dụng, hỗ trợ cho nhiều công việc khác như : lưu trữ (clipping) báo chí, xây dựng bộ ngữ liệu cho các bài toán cần dữ liệu được phân loại, tiền đề cho các bài toán khác như phân loại website.

1.4.4. Đóng góp của luận văn

Luận văn đã thực hiện việc được nhiều cải tiến của hướng tiếp cận tách từ tiếng Việt dùng trong phân loại văn bản theo phương pháp dựa trên thống kê Internet.

Đối với tách từ tiếng Việt, chúng em đề nghị thêm một công thức tính toán độ tương hỗ mới, từ đó thực hiện thử nghiệm tính hiệu quả của cách tính này so với cách công thức ở những công trình khác.

Trong quá trình xây dựng thuật toán di truyền dùng trong tách từ, chúng em đã cải tiến hình thức đột biến mới phù hợp với hình thức cấu tạo từ trong câu.

Đối với việc phân loại văn bản, chúng em cải tiến công thức tính trong hướng tiếp cận Naïve Bayes phù hợp với phương pháp tính dựa trên thống kê từ Google.

Chương 2

CÁC PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN TIẾNG ANH

Bối cảnh các phương pháp phân loại văn bản hiện nay

Các phương pháp phân loại văn bản tiếng Anh hiện hành

- Biểu diễn văn bản
- Support vector Machine (SVM)
- K–Nearest Neighbor (kNN)
- Naïve Bayes (NB)
- Neural Network (NNet)
- Linear Least Square Fit (LLSF)
- Centroid- based vector

Kết luận

Chương 2. CÁC PHƯƠNG PHÁP PHÂN LOẠI VĂN BẢN TIẾNG ANH

2.1. Bối cảnh các phương pháp phân loại văn bản hiện nay

Phân loại văn bản tự động là một lĩnh vực được chú ý nhất trong những năm gần đây. Để phân loại người ta sử dụng nhiều cách tiếp cận khác nhau như dựa trên từ khóa, dựa trên ngữ nghĩa các từ có tần số xuất hiện cao, mô hình Maximum Entropy, tập thô … Tiếng Anh là một trong những ngôn ngữ được nghiên cứu sớm và rộng rãi nhất với kết quả đạt được rất khả quan. Một số lượng lớn các phương pháp phân loại đã được áp dụng thành công trên ngôn ngữ này : mô hình hồi quy [Fuhr et al,1991], phân loại dựa trên *láng giềng gần nhất* (k-nearest neighbors) [Dasarathy, 1991], phương pháp dựa trên *xác suất Naïve Bay*es [Joachims, 1997], *cây quyết định* [Fuhr et al,1991], *học luật quy nạp* [William & Yoram, 1996], *mạng nơron* (neural network)[Wiener et al, 1995], *học trực tuyến*[William & Yoram, 1996], và *máy vector hỗ trợ* (SVM-support vector machine) [Vapnik, 1995]. Hiệu quả của các phương pháp này rất khác nhau ngay cả khi áp dụng cho tiếng Anh. Việc đánh giá gặp nhiều khó khăn do việc thiếu các tập ngữ liệu huấn luyện chuẩn. Thậm chí đối với tập dữ liệu được sử dụng rộng rãi nhất, Reuter cũng có nhiều phiên bản khác nhau. Hơn nữa, có rất nhiều độ đo được sử dụng như recall, precision, accuracy hoặc error, break-even point, F-measure …Chương này giới thiệu các thuật toán phân loại được sử dụng phổ biến nhất đồng thời so sánh giữa các phương pháp sử dụng kết quả của [Yang, 1997].

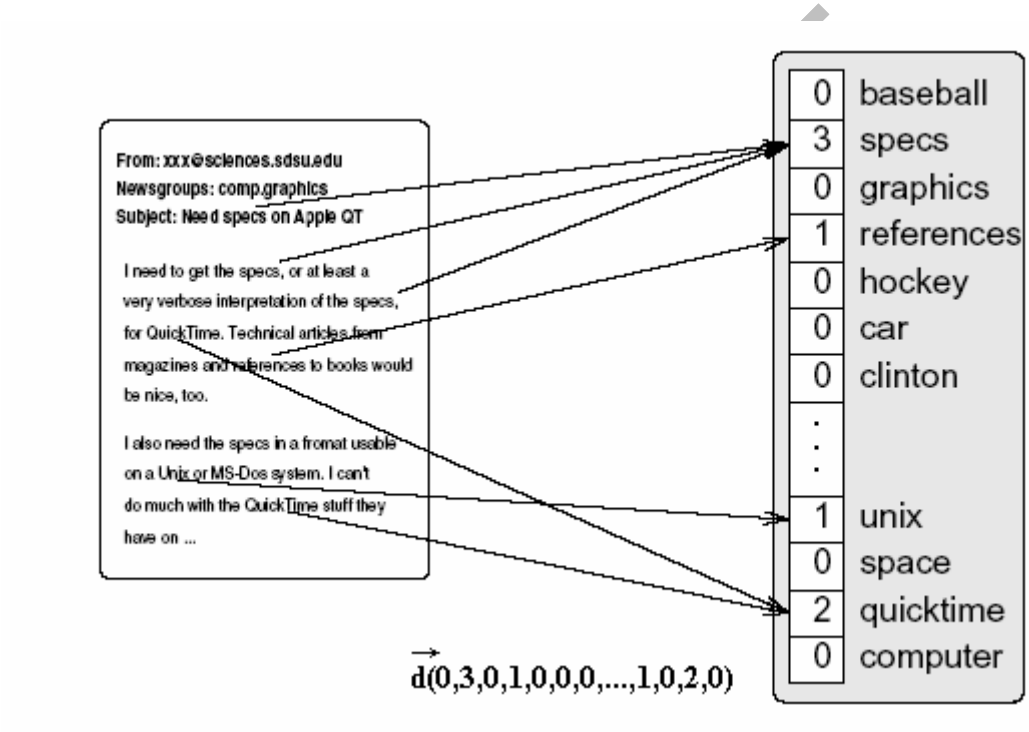
2.2. Các phương pháp phân loại văn bản tiếng Anh hiện hành

2.2.1. Biểu diễn văn bản

Bước đầu tiên của mọi phương pháp phân loại là chuyển việc mô tả văn bản dùng chuỗi ký tự thành một dạng mô tả khác, phù hợp với các thuật toán học theo mẫu và phân lớp. Hầu hết các thuật toán đều sử dụng cách biểu diễn văn bản sử dụng vector đặc trưng, sự khác nhau có chăng là việc chọn không gian đặc trưng khác nhau. Vì vậy ở phần này chúng em sẽ trình bày sơ lược về vector đặc trưng.

Ý tưởng chính là xem mỗi văn bản d_i tương ứng là một vector đặc trưng $\vec{d_i}\left(TF(w_1),TF(w_2),...,TF(w_n)\right)$ trong không gian các từ W^n (w_i là một từ, một đặc trưng, tương ứng một chiều của không gian). Gía trị của $TF(w_i)$ chính là số lần xuất hiện của từ w_i trong văn bản d_i . Từ được chọn là một đặc trưng khi nó xuất hiện trong ít nhất 3 văn bản [Joachims, 1997]. Để không bị phụ thuộc vào chiều dài văn bản vector đặc trưng sẽ được chuẩn hóa về chiều dài đơn vị :

$$\vec{di}(\frac{TF(w_1)}{\sum TF^2(w_i)},\frac{TF(w_2)}{\sum TF^2(w_i)},...,\frac{TF(w_n)}{\sum TF^2(w_i)})$$



Hình 2. 1. Biểu diễn văn bản

Trong thực tế để cải thiện tốc độ và kết quả người ta thường sử dụng $IDF(w_i)$ hoặc $TFIDF(\mathrm{w_i})$ thay cho $TF(w_i)$:

$$IDF(w_i)=\log(\frac{m}{DF(w_i)})$$

$$TFIDF(w_i)=TF(w_i).IDF(w_i)$$

Với

- m chính là số văn bản huấn luyện

- $DF(w_i)$ là số văn bản có chứa từ w_i.

Một vấn đề nảy sinh khi biểu diễn văn bản theo hướng vector đặc trưng chính là việc chọn đặc trưng và số chiều cho không gian. Cần phải chọn bao nhiêu từ và chọn những từ nào ? theo những cách nào ? Có nhiều hướng tiếp cận trong vấn đề này mà tiêu biểu là sử dụng *Information Gain* [Yang & Petersen, 1997] ngoài ra còn có các phương pháp như *DF-Thresolding* [Yang & Petersen, 1997], $\chi^2 - Test$ [Schütze et al,1995] hoặc *Term Strength* [Yang & Wilbur,1997]. Phương pháp *Information Gain* sử dụng độ đo *Mutual Information(MI)* [Yang & Petersen, 1997] để chọn ra tập đặc trưng con f gồm những từ có giá trị MI cao nhất.

Các đặc trưng của văn bản khi biểu diễn dưới dạng vector :

- Số chiều không gian đặc trưng thường rất lớn (trên 10000)
- Có các đặc trưng độc lập nhau, sự kết hợp các đặc trưng này thường không có ý nghĩa trong phân loại
- Đặc trưng rời rạc : vector $\vec{d_i}$ có rất nhiều giá trị 0 do có nhiều đặc trưng không xuất hiện trong văn bản d_i.
- Hầu hết các văn bản có thể được phân chia một cách tuyến tính bằng các hàm tuyến tính.

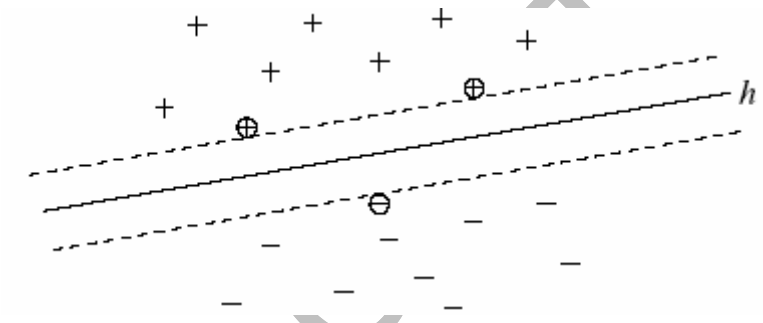
Việc phân loại sẽ tốt hơn nếu các thuật toán tận dụng được những đặc trưng này. Phần tiếp theo sẽ nói rõ hơn về các thuật toán phân loại.

2.2.2. Support vector Machine(SVM)

SVM là phương pháp tiếp cận phân loại rất hiệu quả được Vapnik giới thiệu năm 1995 [Vapnik, 1995] để giải quyết vấn đề nhận dạng mẫu 2 lớp sử dụng nguyên lý Cực tiểu hóa Rủi ro có Cấu trúc (Structural Risk Minimization) [Vapnik, Cortes, 1995].

2.2.2.1. Ý tưởng

Cho trước một tập huấn luyện được biểu diễn trong không gian vector trong đó mỗi tài liệu là một điểm, phương pháp này tìm ra một siêu mặt phẳng h quyết định tốt nhất có thể chia các điểm trên không gian này thành hai lớp riêng biệt tương ứng lớp + và lớp –. Chất lượng của siêu mặt phẳng này được quyết định bởi khoảng cách (gọi là biên) của điểm dữ liệu gần nhất của mỗi lớp đến mặt phẳng này. Khoảng cách biên càng lớn thì mặt phẳng quyết định càng tốt đồng thời việc phân loại càng chính xác. Mục đích thuật toán SVM tìm được khoảng cách biên lớn nhất. Hình sau minh họa cho thuật toán này :



Hình 2. 2. Siêu mặt phẳng h phân chia dữ liệu huấn huyện thành 2 lớp + và – với khoảng cách biên lớn nhất. Các điểm gần h nhất là các vector hỗ trợ ,Support Vector (được khoanh tròn)

2.2.2.2. Công thức chính

SVM thực chất là một bài toán tối ưu, mục tiêu của thuật toán này là tìm được một không gian H và siêu mặt phẳng quyết định h trên H sao cho sai số phân loại là thấp nhất

Phương trình siêu mặt phẳng chứa vector $\vec{d_i}$ trong không gian như sau :

$$\vec{d_i} \cdot \vec{w} + b = 0$$

Đặt $h(\vec{d_i}) = sign(\vec{d_i} \cdot \vec{w} + b) = \begin{cases} +1, \vec{d_i} \cdot \vec{w} + b > 0 \\ -1, \vec{d_i} \cdot \vec{w} + b < 0 \end{cases}$

Như thế $h(\vec{d_i})$ biểu diễn sự phân lớp của $\vec{d_i}$ vào hai lớp như đã nói. Gọi $y_i = \{\pm 1\}$, y_i = + 1, văn bản $\vec{d_i}$ ∈ lớp +; y_i = - 1, văn bản $\vec{d_i}$ ∈ lớp - Khi này để có siêu mặt phẳng h ta sẽ phải giải bài toán sau :

Tìm Min $\left\|\vec{w}\right\|$ với $\vec{w}$ và b thõa điều kiên sau :

$$\forall i \in \overline{1,n} : y_i\left(sign(\vec{d_i} \cdot \vec{w} + b)\right) \geq 1$$

Bài toán SVM có thể giải bằng kỹ thuật sử dụng toán tử Lagrange để biến đổi thành dạng đẳng thức.

Điểm thú vị ở SVM là mặt phẳng quyết định chỉ phụ thuộc vào các vector hỗ trợ (Support Vector) có khoảng cách đến mặt phẳng quyết định là $\frac{1}{\left\|\vec{w}\right\|}$. Khi các điểm khác bị xóa đi thì thuật toán vẫn cho kết quả giống như ban đầu. Chính đặc điểm này làm cho SVM khác với các thuật toán khác như kNN,LLSF, NNet và NB vì tất cả dữ liệu trong tập huấn luyện đều được dùng để tối ưu hóa kết quả. Các phiên bản SVM tốt có thể kể đến là SVM^{Light} [Joachims, 1998] và Sequential Minimal Optimization (SMO) [Platt, 1998]

2.2.3. K–Nearest Neighbor (kNN)

kNN là phương pháp truyền thống khá nổi tiếng về hướng tiếp cận dựa trên thống kê đã được nghiên cứu trong nhận dạng mẫu hơn bốn thập kỷ qua [Dasarathy, 1991]. kNN được đánh giá là một trong những phương pháp tốt nhất (áp dụng trên tập dữ liệu Reuters phiên bản 21450), được sử dụng từ những thời kỳ đầu của việc phân loại văn bản [Marsand et al, 1992] [Yang, 1994] [Iwayama, Tokunaga, 1995].

2.2.3.1. Ý tưởng

Khi cần phân loại một văn bản mới, thuật toán sẽ tính khoảng cách (khoảng cách Euclide, Cosine ...) của tất cả các văn bản trong tập huấn luyện đến văn bản này để tìm ra k văn bản gần nhất (gọi là k "láng giềng"), sau đó dùng các khoảng cách này đánh trọng số cho tất cả chủ đề. Trọng số của một chủ đề chính là tổng tất cả khoảng cách ở trên của các văn bản trong k láng giềng có cùng chủ đề, chủ đề nào

không xuất hiện trong k láng giềng sẽ có trọng số bằng 0. Sau đó các chủ đề sẽ được sắp xếp theo mức độ trọng số giảm dần và các chủ đề có trọng số cao sẽ được chọn là chủ đề của văn bản cần phân loại.

2.2.3.2. Công thức chính

Trọng số của chủ đề c_j đối với văn bản $\vec{x}$:

$$W(\vec{x}, c_j) = \sum_{\vec{d_i} \in \{kNN\}} sim(\vec{x}, \vec{d_i}).y(\vec{d_i}, c_j) - b_j$$

Trong đó

- $y\left(\vec{d_i}, c_j\right) \in \{0,1\}$, với
  - ✓ y = 0 : văn bản $\vec{d_i}$ không thuộc về chủ đề c_j
  - ✓ y = 1 : văn bản $\vec{d_i}$ thuộc về chủ đề c_j
- $sim\left(\vec{x}, \vec{d_i}\right)$: độ giống nhau giữa văn bản cần phân loại $\vec{x}$ và văn bản $\vec{d}_i$. Có thể sử dụng độ đo cosine để tính $sim\left(\vec{x}, \vec{d_i}\right)$

$$sim\left(\vec{x}, \vec{d_i}\right) = \cos(\vec{x}, \vec{d_i}) = \frac{\vec{x}.\vec{d_i}}{\|\vec{x}\|.\|\vec{di}\|}$$

- b_j là ngưỡng phân loại của chủ đề c_j được tự động học sử dụng một tập văn bản hợp lệ được chọn ra từ tập huấn luyện

Để chọn được tham số k tốt nhất cho việc phân loại, thuật toán phải được chạy thử nghiệm trên nhiều giá trị k khác nhau, giá trị k càng lớn thì thuật toán càng ổn định và sai sót càng thấp [Yang, 1997]. Giá trị tốt nhất được sử dụng tương ứng trên hai bộ dữ liệu Reuter và Oshumed là k = 45 [Joachims, 1997].

2.2.4. Naïve Bayes (NB)

NB là phương pháp phân loại dựa vào xác suất được sử dụng rộng rãi trong lĩnh vực máy học [Mitchell, 1996] [Joachims, 1997] [Jason, 2001] được sử dụng lần đầu tiên trong lĩnh vực phân loại bởi Maron vào năm 1961 [Maron, 1961] sau đó trở nên phổ biến dùng trong nhiều lĩnh vực như trong các công cụ tìm kiếm [Rijsbergen et al, 1970], các bộ lọc mail [Sahami et al, 1998]...

2.2.4.1. Ý tưởng

Ý tưởng cơ bản của cách tiếp cận Naïve Bayes là sử dụng xác suất có điều kiện giữa từ và chủ đề để dự đoán xác suất chủ đề của một văn bản cần phân loại. Điểm quan trọng của phương pháp này chính là ở chỗ giả định rằng sự xuất hiện của tất cả các từ trong văn bản đều độc lập với nhau. Như thế NB không tận dụng được sự phụ thuộc của nhiều từ vào một chủ đề cụ thể

Giả định đó làm cho việc tính toán NB hiệu quả và nhanh chóng hơn các phương pháp khác với độ phức tạp theo số mũ vì nó không sử dụng việc kếp hợp các từ để đưa ra phán đoán chủ đề.

2.2.4.2. Công thức chính

Mục đích chính là tính được xác suất $\Pr(Cj, d')$, xác suất để văn bản d' nằm trong lớp Cj. Theo luật Bayes, văn bản d' sẽ được gán vào lớp Cj nào có xác suất $\Pr(Cj, d')$ cao nhất. Công thức sau dùng để tính $\Pr(Cj, d')$ [Joachims, 1997]

$$H_{BAYES}(d') = \arg\max_{Cj \in C} \left(\frac{\Pr(C_j).\prod_{i=1}^{|d'|} \Pr(w_i \mid C_j)}{\sum_{C' \in C} \Pr(C').\prod_{i=1}^{|d'|} \Pr(w_i \mid C')} \right)$$

$$= \arg\max_{Cj \in C} \left(\frac{\Pr(Cj).\prod_{w \in F} \Pr(w \mid C_j)^{TF(w,d')}}{\sum_{C' \in C} \Pr(C').\prod_{w \in F} \Pr(w \mid C')^{TF(w,d')}} \right)$$

Với

- $TF(w_i, d')$ là số lần xuất hiện của từ w_i trong văn bản d'
- $|d'|$ là số lượng các từ trong văn bản d'
- w_i là một từ trong không gian đặc trưng F với số chiều là $|F|$
- $\Pr(C_j)$ được tính dựa trên tỷ lệ phần trăm của số văn bản mỗi lớp tương ứng trong tập dữ liệu luyện : $\Pr(C_j) = \frac{\|C_j\|}{\|C\|} = \frac{\|C_j\|}{\sum_{C' \in C} \|C'\|}$

- $\Pr(w_i \mid C_j)$ được tính sử dụng phép ước lượng Laplace [Napnik, 1982] :
- $$\Pr(w_i \mid C_j) = \frac{1 + TF(w_i, C_j)}{|F| + \sum_{w' \in |F|} TF(w', C_j)}$$

Ngoài ra còn có các phương pháp NB khác có thể kể ra như sau ML Naive Bayes, MAP Naive Bayes, Expected Naive Bayes, Bayesian Naive Bayes [Jason, 2001]. Naive Bayes là một công cụ rất hiệu quả trong một số trường hợp. Kết quả có thể rất tồi nếu dữ liệu huấn luyện nghèo nàn và các tham số dự đoán (như không gian đặc trưng) có chất lượng kém. Nhìn chung đây là một thuật toán phân loại tuyến tính thích hợp trong phân loại văn bản nhiều chủ đề. NB có ưu điểm là cài đặt đơn giản, tốc độ nhanh, dễ dàng cập nhật dữ liệu huấn luyện mới và có tính độc lập cao với tập huấn luyện, có thể sử dụng kết hợp nhiều tập huấn luyện khác nhau. Tuy nhiên NB ngoài giả định tính độc lập giữa các từ còn phải cần đến một ngưỡng tối ưu để cho kết quả khả quan. Nhằm mục đích cải thiện hiệu năng của NB, các phương pháp như multiclass-boosting, ECOC [Berger, 1999] [Ghani, 2000] có thể được dùng kết hợp.

2.2.5. Neural Network (NNet)

Nnet được nghiên cứu mạnh trong hướng trí tuệ nhân tạo. Wiener là người đã sử dụng Nnet để phân loại văn bản, sử dụng 2 hướng tiếp cận : kiến trúc phẳng (không sử dụng lớp ẩn) và mạng nơron 3 lớp (bao gồm một lớp ẩn)[Wiener et al, 1995]

Cả hai hệ thống trên đều sử dụng một mạng nơron riêng rẽ cho từng chủ đề, NNet học cách ánh xạ phi tuyến tính những yếu tố đầu vào như từ, hay mô hình vector của một văn bản vào một chủ đề cụ thể.

Khuyết điểm của phương pháp NNet là tiêu tốn nhiều thời gian dành cho việc huấn luyện mạng nơron.

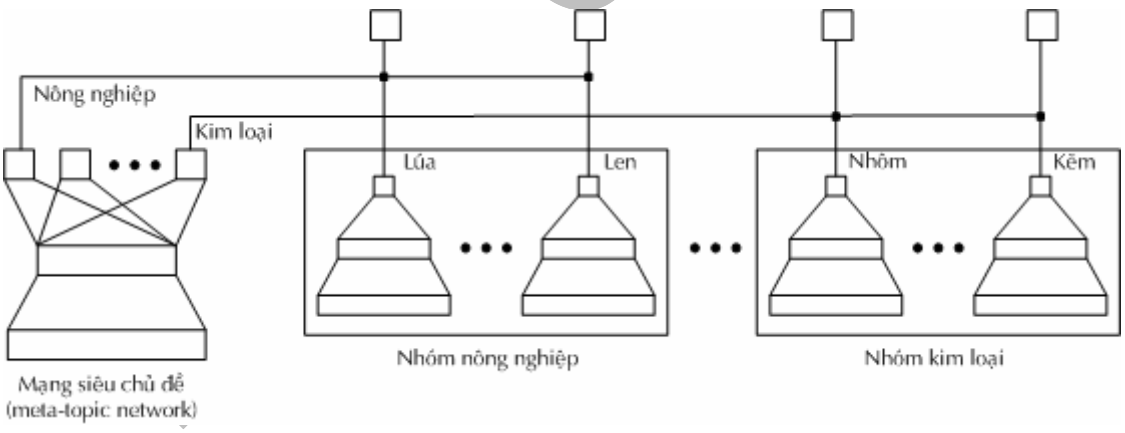
2.2.5.1. Ý tưởng

Mô hình mạng neural gồm có ba thành phần chính như sau: *kiến trúc* (architecture), *hàm chi phí* (cost function), và *thuật toán tìm kiếm* (search

algorithm). *Kiến trúc* định nghĩa dạng chức năng (functional form) liên quan giá trị nhập (inputs) đến giá trị xuất (outputs).

Kiến trúc phẳng (flat architecture) : Mạng phân loại đơn giản nhất (còn gọi là mạng logic) có một đơn vị xuất là kích hoạt kết quả (logistic activation) và không có lớp ẩn, kết quả trả về ở dạng hàm (functional form) tương đương với mô hình hồi quy logic. Thuật toán tìm kiếm chia nhỏ mô hình mạng để thích hợp với việc điều chỉnh mô hình ứng với tập huấn luyện. Ví dụ, chúng ta có thể học trọng số trong mạng kết quả (logistic network) bằng cách sử dụng không gian trọng số giảm dần (gradient descent in weight space) hoặc sử dụng thuật toán interated-reweighted least squares là thuật toán truyền thống trong hồi quy (logistic regression).

Kiến trúc mô dun (modular architecture): Việc sử dụng một hay nhiều lớp ẩn của những hàm kích hoạt phi tuyến tính cho phép mạng thiết lập các mối quan hệ giữa những biến nhập và biến xuất. Mỗi lớp ẩn học để biểu diễn lại dữ liệu đầu vào bằng cách khám phá ra những đặc trưng ở mức cao hơn từ sự kết hợp đặc trưng ở mức trước.



Hình 2. 3. Hình Kiến trúc mô đun (Modular Architecture) . Các kết quả của từng mạng con sẽ là giá trị đầu vào cho mạng siêu chủ đề và được nhân lại với nhau để dự đoán chủ đề cuối cùng .

2.2.5.2. Công thức chính

Trong công trình của Wiener et al (1995) dựa theo khung của mô hình hồi quy, liên quan từ đặc trưng đầu vào cho đến kết quả gán chủ đề tương ứng được học từ

tập dữ liệu. Do vậy, để phân tích một cách tuyến tính, tác giả dùng hàm sigmoid sau làm hàm truyền trong mạng neural:

$$p = \frac{1}{1+e^{-\eta}}$$

Trong đó, $\eta = \beta^T x$ là sự kết hợp của những đặc trưng đầu vào và p phải thỏa điều kiện $p \in (0,1)$

2.2.6. Linear Least Square Fit (LLSF)

LLSF là một cách tiếp cận ánh xạ được phát triển bởi Yang và Chute vào năm 1992 [Yang & Chute, 1992] Đầu tiên, LLSF được Yang và Chute thử nghiệm trong lĩnh vực xác định từ đồng nghĩa sau đó sử dụng trong phân loại vào năm 1994 [Yang & Chute, 1994]. Các thử nghiệm của Yang cho thấy hiệu suất phân loại của LLSF có thể ngang bằng với phương pháp kNN kinh điển.

2.2.6.1. Ý tưởng

LLSF sử dụng phương pháp hồi quy để học từ tập huấn luyện và các chủ đề có sẵn [Yang & Chute, 1994]. Tập huấn luyện được biểu diễn dưới dạng một cặp vector đầu vào và đầu ra như sau :

Vector đầu vào một văn bản bao gồm các từ và trọng số

Vector đầu ra gồm các chủ đề cùng với trọng số nhị phân của văn bản ứng với vector đầu vào

Giải phương trình các cặp vector đầu vào/ đầu ra, ta sẽ được ma trận đồng hiện của hệ số hồi quy của từ và chủ đề(matrix of word-category regression coefficients)

2.2.6.2. Công thức chính

$$F_{LS} = \arg\min_F \|FA - B\|^2$$

Trong đó

- A, B là ma trận đại diện tập dữ liệu huấn luyện (các cột trong ma trận tương ứng là các vector đầu vào và đầu ra)
- F_{LS} là ma trận kết quả chỉ ra một ánh xạ từ một văn bản bất kỳ vào vector của chủ đề đã gán trọng số

Nhờ vào việc sắp xếp trọng số của các chủ đề, ta được một danh sách chủ đề có thể gán cho văn bản cần phân loại. Nhờ đặt ngưỡng lên trọng số của các chủ đề mà ta tìm được chủ đề thích hợp cho văn bản đầu vào. Hệ thống tự động học các ngưỡng tối ưu cho từng chủ đề, giống với kNN. Mặc dù LLSF và kNN khác nhau về mặt thống kê, nhưng ta vẫn tìm thấy điểm chung ở hoạt động của hai phương pháp là việc học ngưỡng tối ưu.

2.2.7. Centroid- based vector

Là một phương pháp phân loại đơn giản, dễ cài đặt và tốc độ nhanh do có độ phức tạp tuyến tính O(n) [Han, Karypis 2000]

2.2.7.1. Ý tưởng

Mỗi lớp trong dữ liệu luyện sẽ được biểu diễn bởi một vector trọng tâm. Việc xác định lớp của một văn bản thử bất kì sẽ thông qua viêc tìm vector trọng tâm nào gần với vector biểu diễn văn bản thử nhất. Lớp của văn bản thử chính là lớp mà vector trọng tâm đại diện. Khoảng cách được tính theo độ đo cosine.

2.2.7.2. Công thức chính

Công thức tính vector trọng tâm của lớp i

$$\overrightarrow{C_i} = \frac{1}{\|\{i\}\|} \sum_{d_j \in \{i\}} \overrightarrow{d_j}$$

Độ đo khoảng cách giữa vector $\vec{x}$ và $\overrightarrow{C_i}$

$$\cos\left(\vec{x}, \overrightarrow{C_i}\right) = \frac{\vec{x} \cdot \overrightarrow{C_i}}{\left\|\vec{x}\right\| * \left\|\overrightarrow{C_i}\right\|}$$

Trong đó :

- $\vec{x}$ là vector văn bản cần phân loại
- $\{i\}$ là tập hợp các văn bản thuộc chủ đề C_i

Chủ đề của $\vec{x}$ là C_x thõa $\cos(\vec{x}, \overrightarrow{C_x}) = \arg\max(\cos(\vec{x}, \overrightarrow{C_i}))$

2.3. Kết luận

Các thuật toán phân loại trên từ thuật toán phân loại 2 lớp (SVM) đến các thuật toán phân loại đa lớp (kNN) đều có điểm chung là yêu cầu văn bản phải được biểu diễn dưới dạng vector đặc trưng. Ngoài ra các thuật toán như kNN,NB,LLSF đều phải sử dụng các ước lượng tham số và ngưỡng tối ưu trong khi đó thuật toán SVM có thể tự tìm ra các tham số tối ưu này. Trong các phương pháp SVM là phương pháp sử dụng không gian vector đặc trưng lớn nhất (hơn 10000 chiều) trong khi đó chỉ là 2000 đối với NB, 2415 cho kNN và LLSF, 1000 cho Nnet [Yang, 1997]. Thời gian huấn luyện cũng khác nhau đối với từng phương pháp, Nnet (sử dụng mỗi mạng tương ứng một chủ đề) và SVM là hai phương pháp có thời gian huấn luyện lâu nhất trong khi đó kNN,NB,LLSF và Centroid là các phương pháp có tốc độ (thời gian huấn luyện, phân loại) nhanh và cài đặt dễ dàng.

Về hiệu suất, dựa vào thử nghiệm của Yang [Yang, Liu, 1997] trên tập dữ liệu Reuter-21578 với hơn 90 chủ đề và trên 7769 văn bản, ta có thể sắp xếp các phương pháp phân loại văn bản theo thứ tự như sau SVM > kNN >> {LLSF,NB,Nnet}. Tuy nhiên kết quả trên có thể không còn đúng khi áp dụng thử nghiệm phân loại trên Tiếng Việt. Các lý do chính như sau :

Thứ nhất: không có một tập dữ liệu chuẩn dành riêng cho việc phân loại.

Thứ hai: hiện tại chưa có chuẩn thống nhất nào cho vấn đề font và dấu câu cho Tiếng Việt.

Thứ ba: viêc biểu diễn văn bản Tiếng Việt bằng vector đặc trưng gặp nhiều trở ngại do bị phụ thuộc nhiều vào các phương pháp tách từ. Trong khi đó các phương pháp này không đạt được hiệu quả cao như trong tiếng Anh.

Để có thể áp dụng các phương pháp phân loại văn bản đã được sử dụng thành công trên nhiều ngôn ngữ (Anh, Pháp,…) như đã liệt kê trên, điều kiện tiên quyết là phải tìm ra một phương pháp tách từ tốt để thông qua đó cải thiện hiệu quả của các thuật toán phân loại. Trong tiếng Anh, đơn vị nhỏ nhất là "từ" nên việc tách từ trở nên khá đơn giản, trong khi đối với một số ngôn ngữ như tiếng Hoa, Nhật, Hàn Quốc... và Tiếng Việt của chúng ta phải xử lý hoàn toàn khác do đơn vị nhỏ nhất lại

là “tiếng”. Do đó, trước khi thực hiện phân loại, chúng ta phải tìm hiểu về các hướng tiếp cận cho việc tách từ tiếng Việt, một vấn đề khá thú vị không kém các phương pháp phân loại.

Chương 3

CÁC PHƯƠNG PHÁP TÁCH TỪ TIẾNG VIỆT HIỆN NAY

Tại sao tách từ tiếng Việt là một thách thức?

So sánh giữa tiếng Việt và tiếng Anh

Nhận xét

Bối cảnh các phương pháp tách từ hiện nay

Bối cảnh chung

Các hướng tiếp cận dựa trên từ

Các hướng tiếp cận dựa trên ký tự

Một số phương pháp tách từ tiếng Việt hiện nay

Phương pháp Maximum Matching: forward/backward

Phương pháp giải thuật học cải tiến

Mô hình tách từ bằng WFST và mạng Neural

Phương pháp quy hoạch động

Phương pháp tách từ tiếng Việt dựa trên thống kê từ Internet và thuật toán di truyền

Kết luận

Chương 3. CÁC PHƯƠNG PHÁP TÁCH TỪ TIẾNG VIỆT HIỆN NAY

3.1. Tại sao tách từ tiếng Việt là một thách thức?

3.1.1. So sánh giữa tiếng Việt và tiếng Anh

Dựa vào các đặc điểm của tiếng Anh và tiếng Việt được trình bày trong [Đinh Điền, 2004], chúng em lập bảng so sánh các đặc điểm chủ yếu giữa tiếng Anh và tiếng Việt như sau

Đặc điểm của Tiếng Việt	Đặc điểm của Tiếng Anh
➢ Được xếp là *loại hình đơn lập* (isolate) hay còn gọi là loại hình phi hình thái, không biến hình, đơn tiết	➢ Là *loại hình biến cách* (flexion) hay còn gọi là loại hình khuất chiết
➢ Từ không biến đổi hình thái, ý nghĩa ngữ pháp nằm ở ngoài từ Ví dụ : Chị ngã em nâng và Em ngã chị nâng	➢ Từ có biến đổi hình thái, ý nghĩa ngữ pháp nằm ở trong từ. Ví dụ: I see him và He sees me.
➢ Phương thức ngữ pháp chủ yếu: trật tự từ và hư từ. Ví dụ: Gạo xay và Xay gạo; đang học và học rồi ; "nó bảo sao không tới", "sao không bảo nó tới", "sao không tới bảo nó"..	➢ Phương thức ngữ pháp chủ yếu là : phụ tố. Ví dụ: studying và studied
➢ Ranh giới từ không được xác định mặc nhiên bằng khoảng trắng	➢ Kết hợp giữa các hình vị là chặt chẽ, khó xác định, được nhận diện bằng khoảng trắng hoặc dấu câu.
➢ Tồn tại loại từ đặc biệt " từ chỉ loại" (classifier) hay còn gọi là	➢ Hiện tượng cấu tạo bằng từ ghép thêm phụ tố (affix) vào gốc từ là

phó danh từ chỉ loại kèm theo với danh từ, như: *cái bàn, cuốn sách, bức thư, con chó, con sông, vì sao…* ➢ Có hiện tượng láy và nói lái trong tiếng Việt Ví dụ: lấp lánh, lung linh Hiện đại -> hại điện, thầy giáo-> tháo giầy…	rất phổ biến. Ví dụ: *anticomputerizational* (anti-compute-er-ize-ation-al)

Bảng 3. 1. So sánh giữa tiếng Việt và tiếng Anh

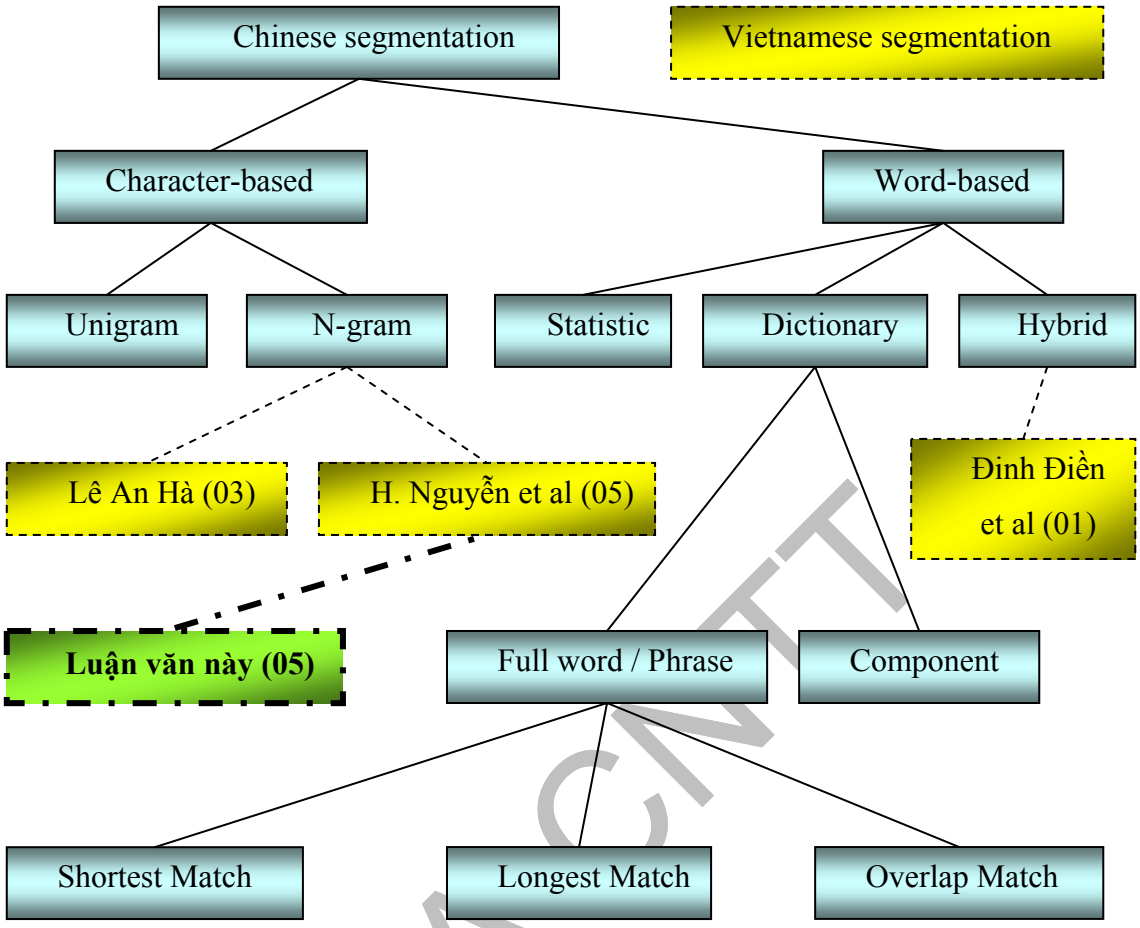
3.1.2. Nhận xét

- Tiếng Việt là loại hình phi hình thái nên việc phân biệt loại từ (danh từ, động từ, tính từ …) và ý nghĩa từ là rất khó, cho dù có sử dụng từ điển.
- Việc tiền xử lý văn bản (tách từ, tách đoạn, tách câu…) sẽ thêm phức tạp với phần xử lý các hư từ, phụ từ, từ láy…
- Phương thức ngữ pháp chủ yếu là trật tự từ nên nếu áp dụng phương pháp tính xác suất xuất hiện của từ có thể không chính xác như mong đợi
- Ranh giới từ không được xác định mặc nhiên bằng khoảng trắng. Điều này khiến cho việc phân tích hình thái (tách từ) tiếng Việt trở nên khó khăn. Việc nhận diện ranh giới từ là quan trọng làm tiền đề cho các xử lý tiếp theo sau đó, như: kiểm lỗi chính tả, gán nhãn từ loại, thống kê tần suất từ,…
- Vì giữa tiếng Anh và tiếng Việt có nhiều điểm khác biệt nên chúng ta không thể áp dụng y nguyên các thuật toán tiếng Anh cho tiếng Việt

3.2. Bối cảnh các phương pháp tách từ hiện nay

3.2.1. Bối cảnh chung

Dựa trên cơ sở thống kê các phương pháp tách từ trên tiếng Hoa của [Foo and Li, 2004], chúng em xin trình bày bối cảnh các phương pháp tách từ hiện nay cho tiếng Việt như sau:



Hình 3.4. Các hướng tiếp cận cơ bản trong tách từ tiếng Hoa và các hướng tiếp cận hiện tại được công bố trong tách từ tiếng Việt

3.2.2. Các hướng tiếp cận dựa trên từ (Word-based approaches)

Hướng tiếp cận dựa trên từ với mục tiêu tách được các từ hoàn chỉnh trong câu. Hướng tiếp cận này có thể chia ra là ba hướng: *dựa trên thống kê* (statistics-based), *dựa trên từ điển* (dictionary-based) và *hydrid* (kết hợp nhiều phương pháp với hy vọng đạt được những ưu điểm của các phương pháp này)

3.2.2.1. Các công trình tách từ tiếng Hoa

Hướng tiếp cận dựa trên thống kê (statistics-based) dựa trên các thông tin như tần số xuất hiện của từ trong tập *dữ liệu huấn luyện đầu*. Hướng tiếp cận này đặc

biệt dựa trên tập ngữ liệu huấn luyện, nhờ vậy nên hướng tiếp cận này tỏ ra rất linh hoạt và hữu dụng trong nhiều lãnh vực riêng biệt [Nie et al.,1996].

Hướng tiếp cận dựa trên từ điển (dictionary-based) thường được sử dụng trong tách từ. Ý tưởng của hướng tiếp cận này là những cụm từ được tách ra từ văn bản phải khớp với các từ trong từ điển. Những hướng tiếp cận khác nhau sẽ sử dụng những loại từ điển khác nhau. Hướng tiếp cận "*full word / phrase*" cần sử dụng một từ điển hoàn chỉnh để có thể tách được đầy đủ các từ hoặc ngữ trong văn bản, trong khi đó, hướng tiếp cận *thành phần* (component) lại sử dụng *từ điển thành phần* (component dictionary)[Wu & Tseng, 1993] . Từ điển hoàn chỉnh chứa tất cả các từ và ngữ được dùng trong tiếng Hoa, trong khi từ điển thành phần (component dictionary) chỉ chứa các thành phần của từ và ngữ như hình vị và các từ đơn giản trong tiếng Hoa.

Tùy theo cách chọn để khớp từ (match), hướng tiếp cận "full word/ phrase" có thể được chia ra thành *khớp dài nhất* (longest match – bằng cách duyệt văn bản tuần tự để tìm ra từ dài nhất có trong từ điển) và *khớp ngắn nhất* (shortest match – bằng cách duyệt văn bản tuần tự và chọn từ đầu tiên có trong từ điển). Ngoài hai cách thông dụng nhất là *khớp dài nhất* và *khớp ngắn nhất,* He et. al. (1996)còn đề nghị một cách thứ ba là cách *kết hợp* (overlap). Trong cách kết hợp này, mỗi chuỗi được phát sinh từ văn bản có thể chồng lấp lên chuỗi khác nếu chuỗi đó có trong từ điển (ví dụ : học sinh học, ta sẽ có các token là "học sinh", "sinh học" chứ không phải chỉ có một cách như *khớp dài nhất* hoặc *khớp ngắn nhất*). Tại thời điểm hiện tại, hướng tiếp cận *khớp dài nhất* được xem là phương pháp quan trọng và hiệu quả nhất trong *hướng tiếp cận dựa trên từ điển* [Foo & Li, 2002].

Tuy nhiên, *hướng tiếp cận dựa trên từ điển* vẫn có một số hạn chế trong việc tách từ vì thực hiện hoàn toàn dựa trên một từ điển hoàn chỉnh. Trong thực tế, để xây dựng một bộ từ điển thật sự hoàn hảo chứa tất cả các từ tiếng Hoa là không thật sự cần thiết và khó thành hiện thực. Hướng tiếp cận dựa trên thành phần (component) phát triển cũng với mục đích làm nhẹ bớt mặt hạn chế này bằng cách nối các hình vị và từ thành những từ và ngữ hoàn chỉnh [Wu & Tseng,1993,1995].

Hướng tiếp cận Hybrid với mục đích kết hợp các hướng tiếp cận khác nhau để thừa hưởng được ưu điểm của nhiều kỹ thuật khác nhau. Hướng tiếp cận này thường kết hợp giữa hướng dựa trên thống kê và dựa trên từ điển nhằm lấy được ưu thế chung và các mặt vượt trội riêng của mỗi phương pháp. Một số thành công của phương pháp này được trình bày trong [Nie et al, 1996]. Mặc dù hướng tiếp cận hibrid có được những ưu điểm của phương pháp khác nhưng lại gặp phải các phức tạp khác như thời gian xử lý, không gian đĩa và đòi hỏi nhiều chi phí.

3.2.2.2. Các công trình tách từ tiếng Việt

Công trình của Đinh Điền et al (2001) đã cố gắng xây dựng tập ngữ liệu huấn luyện riêng (khoảng 10M) dựa trên các thông tin có nguồn gốc từ Internet như tin tức, e-book… Tuy nhiên tập ngữ liệu vẫn còn khá nhỏ để đảm bảo dung lượng và độ phong phú cho việc tách từ. Mặc khác, do tập ngữ liệu được xây dựng một cách thủ công, nên sẽ phần nào mang tính chủ quan. Và một hạn chế nữa là việc đánh giá lại được những thay đổi hằng ngày rất chậm, và có thể xảy ra hiện tượng flip-flop (hiện tượng khi khắc phục lỗi này lại dẫn đến lỗi khác không ngờ tới)

Ở hướng tiếp cận dựa trên từ điển, các từ được tách phải tương ứng với những từ có trong từ điển. Hiện tại, ta vẫn chưa xây dựng được một bộ từ điển Việt Nam chứa toàn bộ các từ và ngữ.

3.2.3. Các hướng tiếp cận dựa trên ký tự (Character-based approaches)

Cần phân biệt rằng hình vị nhỏ nhất của tiếng Việt là "tiếng", được cấu tạo bởi nhiều ký tự trong bảng chữ cái, trong khi hình vị nhỏ nhất của tiếng Hoa là một ký tự. Vì chữ viết tiếng Hoa là chữ tượng hình, không dựa trên bảng chữ cái Latin như tiếng Việt nên trong trường hợp tiếng Hoa, người ta xét hình vị là "ký tự". Tuy nhiên, mỗi *ký tự* (character) trong tiếng Hoa được phát âm thành một "tiếng", nên xét về mặt âm vị, ta có thể xem "tiếng" trong tiếng Hoa và tiếng Việt là tương tự nhau. Vì vậy, để tránh sự hiểu nhầm ý nghĩa giữa *ký tự* trong tiếng Hoa và *tiếng* trong tiếng Việt, chúng em xin phép dùng từ "tiếng" để chỉ cho *ký tự* tiếng Hoa và *tiếng* trong tiếng Việt ở một số trường hợp trình bày về cách tách từ.

Mặc dù có cách viết khác nhau, nhưng về cấu tạo từ và ngữ pháp của tiếng Hoa và tiếng Việt có nhiều điểm tương đồng nhau. Xét về nguồn gốc, tiếng Việt là hình thức phiên âm của chữ Nôm do nhân dân ta sáng tạo nên, vốn có nguồn gốc từ tiếng Trung Hoa thời xưa.

3.2.3.1. Các công trình tách từ tiếng Hoa

Hướng tiếp cận này đơn thuần rút trích một số lượng nhất định các tiếng trong văn bản như rút trích từ 1 ký tự (unigram) hay nhiều ký tự (n-gram). Mặc dù hướng tiếp cận này tương đối đơn giản hơn các hướng khác, nhưng nó cũng mang lại nhiều kết quả khả quan trong tiếng Hoa [Foo and Li, 2004].

Hướng tiếp cận dựa trên một ký tự (unigram) chia văn bản ra các ký tự đơn lẻ để thực hiện việc tách từ. Ngày nay, hầu như người ta không sử dụng phương pháp này như hướng tiếp cận chính trong việc tách từ nữa.

Hướng tiếp cận dựa trên nhiều ký tự (n-gram) chia văn bản ra thành nhiều chuỗi, mỗi chuỗi gồm hai, ba ký tự trở lên. So với hướng tiếp cận dựa trên một ký tự, hướng tiếp cận này cho nhiều kết quả ổn định hơn [Kwok, 1997a;1997b]. Do hơn 75% từ trong tiếng Hoa là từ gồm hai ký tự, nên các phương pháp phổ biến là dựa trên việc tách từ gồm hai ký tự sẽ cho kết quả nhiều từ đúng hơn [Wu & Tseng, 1993].Ví dụ, ta có một câu ABCDEF, hướng tiếp cận trên sẽ chia câu thành AB CD EF. Một biến thể của phương pháp tách từ hai ký tự là hướng tiếp cận cách chia chồng lên nhau, ví dụ ta có ABCDEFG, hướng tiếp cận này sẽ chia thành AB BC CD DE DF FG. Nhóm nghiên cứu của Swiss Federal Institute of Technology (ETH) áp dụng phương pháp biến thể và có thể cải tiến là sử dụng thêm danh sách stoplist (tương tự như các hư từ trong tiếng Việt như à, ơi..) để tách các ngữ của câu trước khi tách từ [Mateev et al, 1997]. Nhờ vậy, mà kích thước văn bản cần tách từ được giảm xuống nhưng có khuyết điểm là nó có thể làm mất ý nghĩa của câu gốc.

Ưu điểm nổi bật của hướng tiếp cận dựa trên nhiều ký tự là tính đơn giản và dễ ứng dụng, ngoài ra còn có thuận lợi là ít tốn chi phí cho việc tạo chỉ mục (index) và xử lý nhiều câu truy vấn (query processing). Qua nhiều công trình nghiên cứu,

hướng tiếp cận tách từ dựa trên nhiều ký tự, đặc biệt là cách tách từ hai ký tự được xem là sự lựa chọn thích hợp[Foo & Li, 2002].

3.2.3.2. Các công trình tách từ tiếng Việt

Trong trường hợp tiếng Việt, hướng tiếp cận này được xem là hướng tiếp cận dựa trên *tiếng*, khác với tiếng Hoa là dựa trên *ký tự*. Ở Việt Nam, hướng tiếp cận này cũng đã có một số công trình được phổ biến. [Lê An Hà, 2003] xây dựng tập ngữ liệu thô 10M, sử dụng phương pháp quy hoạch động để cực đại hóa tổng xác suất xuất hiện của các ngữ. Gần đây nhất có thể kể đến công trình của [H. Nguyen et al, 2005], thay vì sử dụng ngữ liệu thô, công trình của họ có sáng tạo là lấy thông tin thống kê từ Internet và sử dụng thuật toán di truyền (Genetic Algorithm) để tìm cách tách từ tối ưu nhất. Mặc dù công trình của họ còn mang tính sơ bộ, và việc thử nghiệm chưa hoàn chỉnh, nhưng chúng em tin rằng ý tưởng mới lạ này đem lại nhiều hứa hẹn khả quan.

Hướng tiếp cận cho việc tách từ của chúng em mở rộng trên ý tưởng này, ngoài ra, chúng em thực hiện một số thay đổi quan trọng nhằm nâng cao tính chính xác của việc tách từ. Thêm nữa, chúng em đã thực hiện một số thử nghiệm trên số lượng dữ liệu đáng kể nhằm đưa ra các đánh giá một cách bao quát hơn, chính xác hơn.

3.3. Một số phương pháp tách từ tiếng Việt hiện nay

3.3.1. Phương pháp Maximum Matching: forward/backward

3.3.1.1. Nội dung

Phương pháp khớp tối đa (Maximum Matching) còn gọi là Left Right Maximum Matching (LRMM). Theo phương pháp này, ta sẽ duyệt một ngữ hoặc câu từ trái sang phải và chọn từ có nhiều âm tiết nhất có mặt trong từ điển, rồi cứ thể tiếp tục cho từ kế tiếp cho đến hết câu. Thuật toán được trình bày trong [Chih-Hao Tsai, 2000]

Dạng đơn giản được dùng giải quyết nhập nhằng từ đơn. Giả sử có một chuỗi ký tự (tương đương với chuỗi tiếng trong tiếng Việt) $C_1, C_2, \ldots, C_2$. Ta bắt đầu từ đầu chuỗi. Đầu tiên kiểm tra xem C_1, có phải là từ hay không, sau đó kiểm tra xem C_1C_2

có phải là từ hay không. Tiếp tục tìm cho đến khi tìm được từ dài nhất. Từ có vẻ hợp lý nhất sẽ là từ dài nhất. Chọn từ đó, sau đó tìm tiếp như trên cho những từ còn lại cho đến khi xác định được toàn bộ chuỗi từ.

Dạng phức tạp: Quy tắc của dạng này là phân đoạn có vẻ hợp lý nhất là đoạn ba từ với chiều dài tối đa. Thuật toán bắt đầu như dạng đơn giản. Nếu phát hiện ra những cách tách từ gây nhập nhằng (ví dụ, C_1 là từ và C_1C_2 cũng là từ), ta xem các chữ kế tiếp để tìm tất cả các đoạn ba từ có thể có bắt đầu với C_1 hoặc C_1C_2. Ví dụ ta được những đoạn sau:

- C_1 C_2 C_3 C_4
- C_1C_2 C_3 C_4 C_5
- C_1C_2 C_3 C_4 C_5 C_6

Chuỗi dài nhất sẽ là chuỗi thứ ba. Vậy từ đầu tiên của chuỗi thứ ba (C_1C_2) sẽ được chọn. Thực hiện lại các bước cho đến khi được chuỗi từ hoàn chỉnh.

3.3.1.2. Ưu điểm

- Với cách này, ta dễ dàng tách được chính xác các ngữ/câu như " hợp tác xã || mua bán", "thành lập || nước || Việt Nam || dân chủ || cộng hòa"
- Cách tách từ đơn giản, nhanh, chỉ cần dựa vào từ điển
- Trong tiếng Hoa, cách này đạt được độ chính xác 98,41% [Chih-Hao Tsai, 2000].

3.3.1.3. Hạn chế

- Độ chính xác của phương pháp phụ thuộc hoàn toàn vào tính đủ và tính chính xác của từ điển
- Phương pháp này sẽ tách từ sai trong các trường hợp " học sinh || học sinh|| học", "một || ông || quan tài || giỏi", "trước || bàn là || một || ly || nước"…

3.3.2. Phương pháp giải thuật học cải biến (Transformation-based Learning, TBL)

3.3.2.1. Nội dung

Đây là cách tiếp cận dựa trên ngữ liệu đã đánh dấu. Theo cách tiếp cận này, để huấn luyện cho máy tính biết cách nhận diện ranh giới từ tiếng Việt, ta có thể cho máy "học" trên ngữ liệu hàng vạn câu tiếng Việt đã được đánh dấu ranh giới từ đúng.

Sau khi học xong, máy sẽ xác định được các tham số (các xác suất) cần thiết cho mô hình nhận diện từ.

3.3.2.2. Ưu điểm

- Đặc điểm của phương pháp này là khả năng tự rút ra quy luật của ngôn ngữ
- Nó có những ưu điểm của cách tiếp cận dựa trên luật vì cuối cùng nó cũng dựa trên luật được rút ra) nhưng nó khắc phục được khuyết điểm của việc xây dựng các luật một cách thủ công bởi các chuyên gia.
- Các luật được thử nghiệm tại chỗ để đánh giá độ chính xác và hiệu quả của luật (dựa trên ngữ liệu huấn luyện)
- Có khả năng khử được một số nhập nhằng như "The singer sang a lot of a??as", thì hệ có thể xác định được "a??as" là "arias" (dân ca) thay vì "areas" (khu vực) của các mô hình ngôn ngữ theo kiểu thống kê.

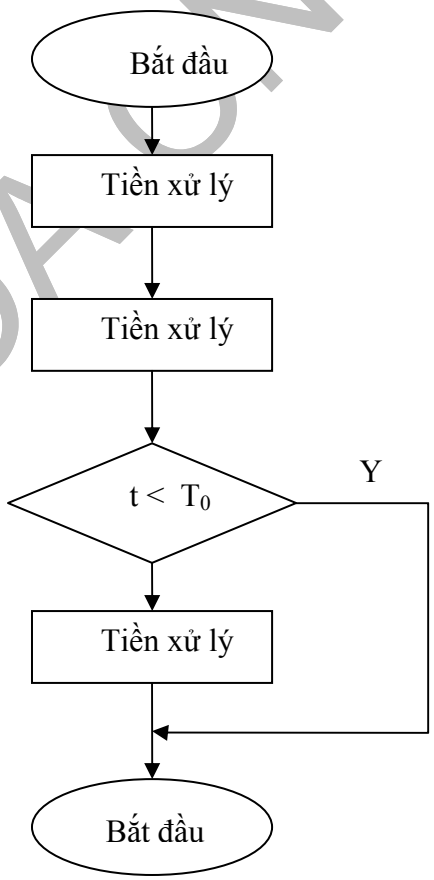
3.3.2.3. Hạn chế

- Phương pháp này "dùng ngữ liệu có gán nhãn ngôn ngữ để học tự động các qui luật đó"[Đinh Điền, 2004]. Như đã nói ở chương 1, việc xây dựng một tập ngữ liệu đạt được đầy đủ các tiêu chí của tập ngữ liệu trong tiếng Việt là một điều rất khó, tốn kém nhiều về mặt thời gian và công sức.
- Hệ phải trải qua một thời gian huấn luyện khá lâu để có thể rút ra các luật tương đối đầy đủ
- Cài đặt phức tạp

3.3.3. Mô hình tách từ bằng WFST và mạng Neural

3.3.3.1. Nội dung

Mô hình mạng chuyển dịch trạng thái hữu hạn có trọng số WFST (Weighted finit–state Transducer) đã được [Richard et al, 1996] áp dụng để tách từ tiếng Trung Quốc. Ý tưởng cơ bản là áp dụng WFST kết hợp với trọng số là xác suất xuất hiện của mỗi từ trong ngữ liệu. Dùng WFST để duyệt qua câu cần xét. Cách duyệt có trọng số lớn nhất sẽ là cách tách từ được chọn. Giải pháp này cũng đã đượng áp dụng trong [Đinh Điền et al, 2001] kèm với mạng neutral để khử nhập nhằng. Hệ thống tách từ tiếng Việt của [Đinh Điền, 2001] gồm hai tầng: tầng WFST ngoài việc tách từ còn xử lý thêm các vấn đề liên quan đến đặc thù của tiếng Việt như từ láy, tên riêng… và tầng mạng neural dùng để khử nhập nhằng nếu có.



Hình 3.5. Sơ đồ hệ thống WFST

- **Tầng WFST :**gồm có ba bước
  - ✓ *Xây dựng từ điển trọng số* : theo mô hình WFST, việc phân đoạn từ được xem như là một sự chuyển dịch trạng thái có xác suất (Stochastic Transduction). Chúng ta miêu tả từ điển D là một đồ thị biến đổi trạng thái hữu hạn có trọng số. Giả sử:
    - *H*: là tập các từ chính tả tiếng Việt (còn gọi là "tiếng")
    - *P*: là từ loại của từ (POS: Part – Of – Speech).

    Mỗi cung của D có thể là:
    - Từ một phần tử của H tới một phần tử của H, hoặc
    - Từ ε (ký hiệu kết thúc từ) tối một phần tử của *P*

    Các nhãn trong D biểu thị một chi phí ước lượng (estimated cost) bằng công thức :

    $$Cost = - \log(f/N)$$

    - Với f: tần số của từ, N: kích thước tập mẫu.

    Đối với các trường hợp từ mới chưa gặp, tác giả áp dụng xác suất có điều kiện Goog-Turning (Baayen) để tính toán trọng số.
  - ✓ *Xây dựng các khả năng phân đoạn từ* : Để giảm sự bùng nổ tổ hợp khi sinh ra các dãy các từ có thể từ một dãy các tiếng trong câu, tác giả đề xuất một phương pháp mới là kết hợp dùng từ điển để hạn chế sinh ra các bùng nổ tổ hợp. Khi phát hiện thấy một cách phân đoạn từ nào đó không phù hợp (không có trong từ điển, không phải là từ láy, không phải là danh từ riêng…) thì tác giả loại bỏ các nhánh xuất phát từ cách phân đoạn từ đó.
  - ✓ *Lựa chọn khả năng phân đoạn từ tối ưu* : Sau khi được một danh sách các cách phân đoạn từ có thể có của câu, tác giả chọn trường hợp phân đoạn từ có trọng số bé nhất như sau:
    - Ví dụ: input = "Tốc độ truyền thông tin sẽ tăng cao"
      - Dictionary "tốc độ" 8.68
        "truyền" 12.31

“truyền thông”	1231
“thông tin”	7.24
“tin”	7.33
“sẽ”	6.09
“tăng”	7.43
“cao”	6.95

Id(D)*D* = “Tốc độ # truyền thông # tin # sẽ # tăng # cao.” 48.79

(8.68 +12.31 + 7.33 + 6.09 + 7.43 +6.95 = 48.79)

Id(D)*D* = “Tốc độ # truyền # thông tin # sẽ # tăng # cao.” 48.70

(8.68 +12.31 + 7.24 + 6.09 + 7.43 +6.95 = 48.79)

Do đó, ta có được phân đoạn tối ưu là “Tốc độ # truyền # thông tin # sẽ # tăng # cao.”

- **Tầng mạng neural :** Mô hình mạng neural mà tác giả đề xuất được dùng để lượng giá 3 dãy từ loại: NNV,NVN, VNN (N: Noun, V: Verb). Mô hình này được học bằng chính các câu mà cách phân đoạn từ vẫn còn nhập nhằng sau khi qua mô hình thứ nhất.

3.3.3.2. Ưu điểm

- Độ chính xác trên 97% [Đinh Điền et al, 2001]
- Mô hình cho kết quả phân đoạn từ với độ tin cậy (xác suất) kèm theo.
- Nhờ có tầng mạng neural nên mô hình có thể khử nhập nhằng các trường hợp tầng WFST cho ra nhiều ứng viên có kết quả ngang nhau
- Phương pháp này cho kết quả với độ chính xác khá cao vì mục đích của tác giả muốn nhắm đến việc tách từ thật chính xác để là nền tảng cho việc dịch máy.

3.3.3.3. Hạn chế

- Cũng tương tự như phương pháp TBL, việc xây dựng tập ngữ liệu là rất công phu, nhưng thật sự rất cần thiết để phục vụ cho mục đích dịch máy sau này của tác giả.

3.3.4. Phương pháp quy hoạch động (dynamic programming)

3.3.4.1. Nội dung

Phương pháp quy hoạch động [Le An Ha, 2003] chỉ sử dụng tập ngữ liệu thô để lấy thông tin về tần số thống kê của từ , làm tăng độ tin cậy cho việc tính toán. Việc tính toán bắt đầu với những đơn vị chắc chắn như câu, các ngữ (chunk) được phân cách bởi dấu câu (như dấu phẩy, gạch nối, chấm phẩy…) vì những thành phần này không có tính nhập nhằng ngay cả trong văn viết cũng như nói. Sau đó, tác giả cố gắng tối đa hoá xác suất của ngữ bằng cách tìm ra nhiều cách tách ngữ đó. Cách tách cuối cùng là cách tách là cho ngữ đó có xác suất cao nhất. Ý tưởng của cách tách từ này cho một ngữ cần tách từ, ta phải tìm ra các tổ hợp từ tạo nên ngữ đó sao cho tổ hợp đó đạt được xác suất tối đa. Tuy nhiên trong phương pháp tính toán này, tác giả gặp phải vấn đề bùng nổ tổ hợp và phân tích ngữ liệu thô. Để giải quyết vấn đề trên, tác giả đã sử dụng phương pháp *quy hoạch động* (dynamic programming) vì lúc đó, xác suất cực đại của một ngữ nhỏ hơn chỉ phải tính toán một lần và sử dụng lại trong các lần sau.

3.3.4.2. Ưu điểm

- Không cần sử dụng tập ngữ liệu đã đánh dấu chính xác

3.3.4.3. Hạn chế

- Trong thí nghiệm, tác giả chỉ dừng lại ở việc tách các từ có ba tiếng bởi vì tập ngữ liệu đầu vào vẫn còn khá nhỏ.
- Xác suất từ đúng là 51%, xác suất từ chấp nhận được 65% [Le An Ha, 2003]. Xác suất này tương đối thấp so với các phương pháp tách từ khác đã đề cập ở trên.

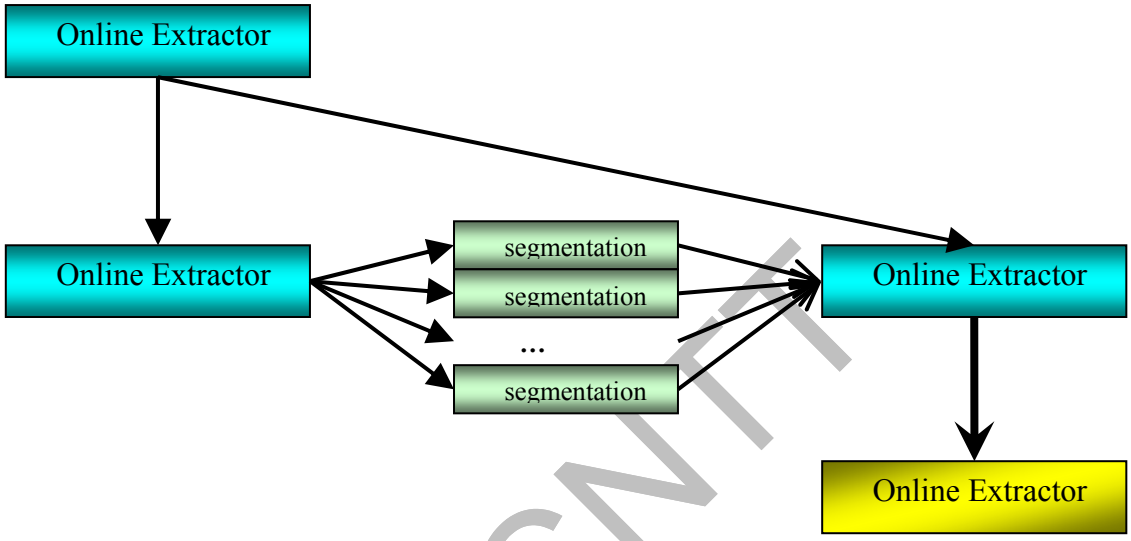
3.3.5. Phương pháp tách từ tiếng Việt dựa trên thống kê từ Internet và thuật toán di truyền (Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese - IGATEC)

3.3.5.1. Nội dung

Phương pháp IGATEC do H.Nguyễn et al (2005) giới thiệu là một hướng tiếp cận mới cho việc *tách từ* với mục đích *phân loại văn bản* mà không cần dùng đến

một từ điển hay tập huấn luyện nào. Trong hướng tiếp cận này, tác giả kết hợp giữa thuật toán di truyền (Genetics Algorithm - GA) với dữ liệu thống kê được trích xuất từ Internet tiến hoá một quần thể gồm các cá thể là các khả năng tách từ trong câu.

Hệ thống gồm ba phần



Hình 3.6. Toàn cảnh hệ thống IGATEC

- *Online Extractor* : Phần này có tác dụng lấy thông tin về tần số xuất hiện của các từ trong văn bản bằng cách sử dụng một search engine nổi tiếng như Google. Sau đó, tác giả sử dụng các công thức sau đây để tính toán mức độ phụ thuộc lẫn nhau (mutual information) để là cơ sở tính fitness cho GA engine.
  - ✓ Tính xác suất các từ xuất hiện trên Internet
    - $p(\mathrm{w}) = \frac{count(w)}{MAX}$
    - $p(w_1 \,\&\, w_2) = \frac{count(w_1 \,\&\, w_2)}{MAX}$

      Trong đó, $MAX = 4 * 10^9$;

      count(w) số lượng văn bản trên Internet được tìm thấy có chứa từ *w* hoặc cùng chứa w_1 và w_2 đối với $count(w_1 \,\&\, w_2)$
  - ✓ Tính xác suất độ phụ thuộc của một từ lên một từ khác

- $p(w_1 \mid w_2) = \frac{p(w_1 \& w_2)}{p(w_1)}$

✓ Thông tin phụ thuộc lẫn nhau (mutual information) của các từ ghép được cấu tạo bởi n tiếng (cw = $w_1 w_2 \ldots w_n$)

✓ $MI(cw) = \frac{p(w_1 \ \& \ w_2 \ \& \ \ldots \ \& \ w_n)}{\sum_{j=1}^{n} p(w_j) \ - \ p(w_1 \ \& \ w_2 \ \& \ \ldots \ \& \ w_n)}$

➢ *GA Engine for Text Segmentation* : mỗi cá thể trong quần thể được biểu diễn bởi chuỗi các bit 0,1, trong đó, mỗi bit đại diện cho một tiếng trong văn bản, mỗi nhóm bit cùng loại đại diện cho một segment.

✓ Các cá thể được khởi tạo ngẫu nhiên, trong đó, mỗi segment được giới hạn trong khoảng 5. GA engine sau đó thực hiện các bước đột biến và lai ghép nhằm mục đích làm tăng giá trị fitness của các cá thể, để đạt được cách tách từ tốt nhất có thể.

➢ *Text Categorization* : tác giả dùng độ hỗ trợ (support degree) của văn bản cần phân loại cho các từ khoá để phân loại văn bản.

3.3.5.2. Ưu điểm

➢ Không cần sử dụng bất cứ tập huấn luyện hoặc từ điển nào

➢ Phương pháp tương đối đơn giản.

➢ Không tốn thời gian huấn luyện

3.3.5.3. Hạn chế

➢ So với các phương pháp trước, IGATEC có độ chính xác thấp hơn LRMM và WFST nhưng vẫn chấp nhận được đối với mục đích tách từ dành cho phân loại văn bản.

➢ Thời gian chạy ban đầu khá chậm do phải lấy thông tin từ Internet mà đường truyền ở Việt Nam còn hạn chế.

➢ Chưa có các thử nghiệm trên tập dữ liệu đủ lớn.

3.4. So sánh các phương pháp tách từ Tiếng Việt hiện nay

Nhìn một cách tổng quan, phương pháp dựa trên từ (word-base) cho độ chính xác khá cao (trên 95%) nhờ vào tập ngữ liệu huấn luyện lớn, được đánh dấu chính xác, tuy nhiên hiệu suất của thuật toán phụ thuộc hoàn toàn vào ngữ liệu huấn luyên. Bởi vì mục đích của các tác giả [Đinh Điền et al, 2001] là thực hiện tách từ thật chính xác để phục vụ cho việc dịch máy nên tác giả đã chọn phương pháp WFST. Với các phương pháp cần phải sử dụng từ điển hoặc tập huấn luyện, ngoài việc tách từ thật chính xác, ta còn có thể nhờ vào các thông tin đánh dấu trong tập ngữ liệu để thực hiện các mục đích khác cần đến việc xác định từ loại như dịch máy, kiểm lỗi chính tả, từ điển đồng nghĩa... Do vậy, mặc dù thời gian huấn luyện khá lâu, cài đặt khá phức tạp, chi phí tạo tập ngữ liệu huấn luyện rất tốn kém, nhưng kết quả mà hướng tiếp cận dựa trên từ mang lại cho mục đích dịch máy là rất xứng đáng cho công sức bỏ ra.

Hướng tiếp cận dựa trên ký tự (character-based) có ưu điểm là dễ thực hiện, thời gian thực thi tương đối nhanh, tuy nhiên lại có độ chính xác không cao bằng phương pháp dựa trên từ. Hướng tiếp cận này thích hợp cho các mục đích nghiên cứu không cần đến độ chính xác tuyệt đối cũng như các thông tin về từ loại như phân loại văn bản, lọc spam, firewall... Nhìn trên bình diện chung, hướng tiếp cận dựa trên từ có nhiều ưu điểm đáng kể, và đem lại nhiều hứa hẹn lạc quan cho các hướng nghiên cứu tiếp theo để nâng cao độ chính xác của phương pháp tách từ này.

3.5. Kết luận

Dựa trên các phân tích về ưu khuyết điểm của các phương pháp, chúng em chọn hướng tiếp cận dựa trên "*tiếng*" (character-based) cho mục tiêu phân loại văn bản của mình.

Bởi vì, mục tiêu của luận văn là phân loại tin tức báo điện tử, một loại hình cực kỳ phong phú về nội dung và ngôn ngữ, nên việc tạo ra một từ điển hoàn chỉnh và có khả năng cập nhật các thay diễn ra liên tục của ngôn ngữ là khó thực hiện được. Hệ thống xử lý cần phải có khả năng linh hoạt, tự động cập nhật những thay đổi

hằng ngày, nên hướng tiếp cận không dựa trên từ điển hoặc tập ngữ liệu là cực kỳ thích hợp.

Hơn nữa, hệ thống phân loại tin tức cần có tốc độ xử lý chấp nhận được để có thể xử lý kịp thời các thông tin mới xuất bản hằng ngày. Do đó, với ưu điểm đơn giản, tốc độ thực thi chấp nhận đươc, hướng tiếp cận IGATEC là một lựa chọn hoàn toàn phù hợp.

Mặt khác, việc phân loại văn bản không yêu cầu việc tách từ phải có độ chính xác cao đến mức từng từ. Ta có hoàn toàn có thể thực hiện thêm việc loại bỏ các từ không cần thiết cho việc phân loại như các hư từ, thán từ... để tăng tốc độ và sự chính xác của bước tách từ, chuẩn bị cho việc phân loại văn bản.

Chương 4

TÁCH TỪ TIẾNG VIỆT KHÔNG DỰA TRÊN TẬP NGỮ LIỆU HAY TỪ ĐIỂN – MỘT THÁCH THỨC

Giới thiệu

Các nghiên cứu về thống kê dựa trên Internet

Các phương pháp tính độ liên quan giữa các từ dựa trên thống kê

Tiền xử lý

Hướng tiếp cận tách từ dựa trên thống kê từ Internet và thuật toán di truyền

Công cụ trích xuất thông tin từ Google

Công cụ tách từ dùng thuật toán di truyền

Kết quả thực nghiệm

Kết luận

Chương 4. TÁCH TỪ TIẾNG VIỆT KHÔNG DỰA TRÊN TẬP NGỮ LIỆU ĐÁNH DẤU (ANNOTATED CORPUS) HAY TỪ ĐIỂN (LEXICON) – MỘT THÁCH THỨC

4.1. Giới thiệu

Như chúng ta đã tìm hiểu ở những phần trên, việc khó xác định ranh giới từ đã làm cho việc xử lý tính nhập nhằng trong ngôn ngữ tiếng Việt càng thêm phức tạp.Ví dụ như: câu "ông lão già đi rất nhanh", ta có thể phân chia từ theo nhiều cách mà câu vẫn có nghĩa "ông ||già đi || rất || nhanh", "ông già || đi || rất || nhanh", "ông || già || đi || rất || nhanh" …

Nhìn chung, đối với tiếng Anh, về mặt lý thuyết tiếng Anh có nhiều thuận lợi vì là loại *ngôn ngữ hoà kết* hay *biến cách* (flexion) [Đinh Điền, 2004] , hệ thống ngữ pháp và từ loại đã được quy định rõ ràng, do đó việc phân định ranh giới từ cũng như xây dựng tập ngữ liệu đánh dấu là tương đối dễ dàng.

Còn đối với tiếng Việt, về mặt lý thuyết tiếng Việt là *loại hình đơn lập* [Đinh Điền, 2004], phương thức ngữ pháp chủ yếu là trật tự từ và hư từ, vì vậy chỉ xét về mặt phân định ranh giới từ đã có thể có nhiều cách phân định cho cùng một câu mà vẫn đúng ngữ pháp Việt Nam.

Ở phần này, chúng em xin trình bày hướng tiếp cận cho việc tách từ tiếng Việt theo một hướng mới mà không cần sử dụng tập ngữ liệu huấn luyện hay từ điển. Hướng tiếp cận của chúng em dựa trên ý tưởng của bài báo IGATEC, và có nhiều cải tiến đang kể hàm làm tăng chất lượng cho bước tách từ tiếng Việt phục vụ cho việc phân loại tin tức báo điện tử.

4.2. Các nghiên cứu về thống kê dựa trên Internet

4.2.1. Giới thiệu

Với sự phát triển nhanh chóng của Internet, world-wide-web đã trở thành nguồn dữ liệu lớn nhất trên thế giới, và là nguồn thông tin ngữ nghĩa tiềm tàng được hàng triệu người dùng trên thế giới tạo ra. Đối với con người, việc xem xét mức độ liên quan giữa hai từ là rất dễ dàng bởi vì con người có thể dựa vào kiến thức thông

thường của mình để suy ra ngữ cảnh thích hợp, ví dụ giữa từ "cái nón" và "màu đỏ", con người dễ dàng nhận ra sự liên quan là "cái nón có màu đỏ". Tuy nhiên, máy tính của chúng ta không có khả năng như con người, vì vậy, chúng ta phải tìm ra một cách biểu diễn ngữ nghĩa mà máy tính có thể "tiêu hoá" được. Có ý kiến cho rằng ta có thể tạo một mạng ngữ nghĩa đồ sộ như một hệ thống trí tuệ ban đầu, sau đó các kiến thức về cuộc sống thực sẽ tự động xuất hiện. Tuy nhiên hướng giải quyết này đòi hỏi lượng chi phí khổng lồ cho việc thiết kế cấu trúc có khả năng tính toán tri thức và việc nhập các dữ liệu chuẩn xác do các chuyên gia thực hiện. Trong khi nỗ lực này vẫn còn đang trong cuộc đua đường dài, chúng ta hãy sử dụng những thông tin hiện có trên world-wide-web để thực hiện việc biểu diễn ngữ nghĩa.

Chúng ta đều biết rằng Internet là kho dữ liệu vô tận, do vậy việc khai thác các thông tin trên đó không thể thực hiện thủ công mà chúng ta phải thông qua sự hỗ trợ của một công cụ tìm kiếm trên mạng. Nói đến công cụ tìm kiếm (search engine), có lẽ tên tuổi đầu tiên mà chúng ta nghĩ đến là Google, một công cụ tìm kiếm hàng đầu bởi tốc độ và chất lượng mà Google đem lại cho người dùng. Và điều đó càng được chứng minh cụ thể hơn khi có ngày càng nhiều các công trình nghiên cứu về thống kê trên Internet dựa vào công cụ tìm kiếm Google như trong phần trình bày tiếp theo sau đây.

4.2.2. Một số công trình nghiên cứu về thống kê dựa trên Internet

Theo Rudi Cilibrasi & Paul Vitanyi (2005), công cụ tìm kiếm Google có thể dùng để tự động khám phá ý nghĩa của từ. Ví dụ : Google tìm thấy từ "student" và "book" cùng xuất hiện với nhau trên Internet với tần số là 57.600.000, trong khi từ "student" và "apple" lại chỉ xuất hiện 8.110.000. Rõ ràng, chúng ta có thể nhận thấy "student" và "book" có liên quan với nhau mật thiết hơn là "student" và "apple".

Tác giả đã sử dụng kết quả tìm kiếm của Google để huấn luyện ngữ nghĩa của các từ (semantic meaning of words) cho phần mềm – một vấn đề trọng tâm trong ngành trí tuệ nhân tạo. Giả sử muốn tính toán mức độ liên quan giữa từ x với từ y, Rudi & Paul (2005) đã đưa ra công thức tính khoảng cách NGD (Normalise Google Distance) như sau:

$$NGD = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}} \quad (1)$$

Trong đó :

- *f(x)* :số trang web chứa từ *x* mà Goole trả về
- *f(x,y)* : số trang web chứa đồng thời từ *x* và từ *y*
- M = 8.058.044.651 là số trang web hiện tại mà Google đã đánh chỉ mục

Với công thức trên, giá trị của NGD càng *nhỏ* thì mức độ liên quan giữa hai từ càng *cao*.

Ví dụ: tần số xuất hiện của "student"= 401.000.000, "book" = 387.000.000, đồng thời là 57.600.000, còn "apple" là 144.000.000, "student" & "apple"= 8.110.000. Với M = 8.058.044.651, ta có

$$NGD(student, book) \approx \frac{\log 401.10^6 - \log 57,6.10^6}{\log 8058044651 - \log 387.10^6} \approx 0.64$$

$$NGD(student, apple) \approx \frac{\log 401.10^6 - \log 8,11.10^6}{\log 8058044651 - \log 144.10^6} \approx 0.97$$

Từ kết quả trên, ta có *NGD(student,book)* ≈0.64 < *NGD(student,apple)* ≈0.97, nên có thể kết luận là "student" liên quan với "book" nhiều hơn là "apple".

Nếu NGD của hai từ lớn hơn 1 thì tác giả nhận xét rằng hai từ đó thường xuất hiện cùng với nhau trong trang web mà không vì một mối liên quan nào cả.

Ví dụ: tần số xuất hiện của "by" là 2.770.000.000, "with" là 2.566.000.000, đồng thời "by" và "with" là 49.700.000. Với M = 8.058.044.651, ta có NGD(by,with) ≈ 3.51

Hơn nữa, NGD là *số tỉ lệ bất biến* (scale-invariant) nên có tính ổn định với sự tăng trưởng số lượng trang web trên Google. Đây là tính chất rất quan trọng bởi vì M số lượng trang web do Google đánh chỉ mục tăng thường xuyên, do đó, số trang web chứa các ngữ tìm kiếm cũng tăng lên ứng với tỉ lệ đó. Điều này có nghĩa là nếu M tăng gấp đôi thì tần số xuất hiện của các ngữ cũng tăng gấp đôi. Công trình của Rudi & Paul (2005) đã mở ra một hướng tiếp cận mới cho các công trình nghiên cứu khác nhờ tính chất không giới hạn bởi dữ liệu, dễ dàng thực thi và là nền móng cho các phương pháp nghiên cứu khác [Rudi & Paul, 2005].

Ngoài ra, theo James & Daniel (2005) còn có một số công trình nghiên cứu về phương pháp thống kê khác trên Internet như tính toán kết quả tìm kiếm bằng hàm luỹ thừa [Simkin & Roychowdhurry, 2003] [Bagrow et al, 2004] , hay phương pháp được đánh giá tốt hơn là dựa vào *giá trị tương tự cực đại* (Maximum Likelihood) [James & Daniel, 2005]…. Mục đích của việc sử dụng *giá trị tương tự cực đại* để tìm ra chỉ số gần giống nhau nhất giữa hai khái niệm. Tuy nhiên, theo kết luận của James & Daniel(2005), các phương pháp tính toán dựa trên hàm mũ cho kết quả chưa khả quan lắm và còn mang tính chủ quan.

4.2.3. Nhận xét

- Hướng thống kê dựa trên Internet hứa hẹn nhiều kết quả khả quan vì không cần phụ thuộc vào tập dữ liệu huấn luyện truyền thống mà chúng ta có thể tận dụng khả năng vô tận của Internet thông qua công cụ tìm kiếm.
- Dựa trên nhận xét của Rudi & Paul (2005), tỉ lệ xuất hiện của từ trên Internet là khá ổn định, điều này cho phép ta thực hiện các tính toán chính xác và ổn định vì ít phụ thuộc vào số lượng trang web trên Internet tăng lên theo thời gian.
- Hiện nay, các công trình nghiên cứu theo hướng tiếp cận mới này chủ yếu được thực hiện trên tiếng Anh, còn đối với tiếng Việt thì có thể nói IGATEC là công trình đầu tiên áp dụng phương pháp này nhưng đã đạt được kết quả rất đáng quan tâm. Chúng em hy vọng rằng rằng những nỗ lực nghiên cứu và cải tiến phương pháp IGATEC sẽ đạt được kết quả tốt hơn.

4.3. Các phương pháp tính độ liên quan giữa các từ dựa trên thống kê

Trong ngôn ngữ tự nhiên, nhất là loại ngôn ngữ phụ thuộc nhiều vào ngữ cảnh như tiếng Việt, đối với con người, chúng ta có thể dễ dàng xác định được ranh giới từ trong câu. Tuy nhiên, do chưa có một quy định cụ thể nào về ranh giới từ tiếng Việt, nên có thể nhiều người Việt có nhiều cách tách từ khác nhau. Đối với người chúng ta vẫn chưa thống nhất được, nên khi dùng máy tính để xử lý ngôn ngữ ta vẫn chưa có một chuẩn nào để xác định đâu là ranh giới từ. Vì vậy, đã có rất nhiều công

trình nghiên cứu cách tính toán độ liên quan giữa các từ để khắc phục các công việc phức tạp do cách phân tích cấu trúc ngữ pháp trong câu đem lại.

Trong phần này, chúng em sẽ trình bày hai nội dung chính:

- Hai thước đo chuẩn dùng để tính toán độ liên quan giữa hai từ trong tiếng Anh là *thông tin tương hỗ* (Mutual Information) và *t-score.*
- Một số ứng dụng và cải tiến của hai công cụ đo trên trong việc tách từ tiếng Hoa và tiếng Việt.

4.3.1. Thông tin tương hỗ (Mutual Information) và t-score dùng trong tiếng Anh

Thông tin tương hỗ (Mutual Information) và *t-score* là hai khái niệm rất quan trọng trong học thuyết về thông tin (Information Theory) và thống kê được trình bày trong [Church et al, 1991] cho mục đích tính toán mức độ liên quan của hai từ trong tiếng Anh.

4.3.1.1. Thông tin tương hỗ MI (Mutual Information) – thước đo đặc điểm tương tự (A Measure of Similarity)

Theo Church et al (1991), việc thống kê *thông tin tương hỗ* (Mutual Information) dùng để nhận biết các trường hợp ngôn ngữ thú vị, bao gồm từ *mối quan hệ ngữ nghĩa* (semantic relations) như bác sĩ/y tá (dạng content word/content word) cho đến mối quan hệ từ vựng-cú pháp (lexico-syntactic) như sự xuất hiện đồng thời giữa động từ và giới từ (dạng content word/ funtion word).

MI có nhiệm vụ so sánh xác suất xuất hiện đồng thời (joint probability) của từ x và từ y so với xác suất tìm thấy x và y xuất hiện độc lập. Công thức tính MI cho hai từ tiếng Anh trong [Church et al, 1991] như sau:

$$I(x;y) \equiv \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Trong đó:

- x và y là hai từ tiếng Anh cần kiểm tra mức độ kết hợp lẫn nhau.
- $I(x;y)$ là thông tin tương hỗ của hai từ.
- $P(x)$, $P(y)$ là xác suất xuất hiện độc lập của x và của y.
- $P(x,y)$ là xác suất xuất hiện đồng thời x và y.

Theo Church et al (1991), giá trị $I(x,y)$ càng lớn thì khả năng kết hợp của x và y càng cao.

4.3.1.2. t-score – thước đo sự khác biệt (A Measure of Dissimilarity)

Chúng ta dễ dàng nhận ra sự giống nhau giữa *strong* và *powerful*, tuy nhiên làm cách nào để phân biệt sự khác nhau giữa chúng. Ví dụ, chúng ta đều biết rằng người ta thường nói *strong tea*, *powerful car* hơn là nói *powerful tea* và *strong car*. Nhưng làm sao cho máy tính nhận ra được sự khác biệt này?

Giả sử , ta biết rằng *strong support* được dùng phổ biến hơn là *powerful support*, Church et al (1991) đã đưa ra công thức tính t-score để đo sự khác biệt trên:

$$t = -\frac{P(w \mid w_1) \; - \; P(w \mid w_2)}{\sqrt{\sigma^2(P(w \mid w_1) \; + \; \sigma^2(w \mid w_2))}}$$

Trong đó:

- w_1, w_2 là hai từ tương tự nhau cần phải phân biệt (ở ví dụ trên là *strong* và *powerful*) .
- w là từ dùng để phân biệt (ở ví dụ trên là *support*).
- $P(w|w_1)$, $P(w|w_2)$ là xác suất của từ w xuất hiện đi kèm với từ w_1, w_2

Lúc đó:

$$t = -\frac{P(powerful\ support) \; - \; P(strong\ support)}{\sqrt{\sigma^2(P(powerful\ support)) + \sigma^2(P(strong\ support))}}$$

$$\approx -\frac{\frac{f(powerful\ support)}{N} \; - \; \frac{f(strong\ support)}{N}}{\sqrt{\frac{f(powerful\ support)}{N^2} + \frac{f(strong\ support)}{N^2}}}$$

$$\approx -\frac{2-175}{\sqrt{2+175}} \approx -13$$

Ta nói rằng *powerful support* có độ lệch chuẩn (standard deviation) kém *strong support* 13 lần. Nhờ vậy, ta có thể phân biệt được sự khác nhau giữa *powerful* và *strong* trong việc sử dụng hai từ này.

4.3.2. Một số cải tiến trong cách tính độ liên quan ứng dụng trong tách từ tiếng Hoa và tiếng Việt

4.3.2.1. Thông tin tương hỗ (Mutual Information)

Khi áp dụng thông tin tương hỗ MI trong tách từ tiếng Hoa, Su et al (1993) cho rằng *thông tin tương hỗ* (Mutual Information) là thước đo mức độ kết hợp của một từ. Nó có nhiệm vụ so sánh xác suất một nhóm các *ký tự* (tương tự như "tiếng" trong tiếng Việt – xem giải thích ở mục 3.2.3.) xuất hiện đồng thời (joint probability) so với xác suất tìm thấy từng ký tự xuất hiện độc lập.

Theo Su et al (1993) cách tính MI cho từ có 2 *ký tự* có thể áp dụng công thức của Church et al (1991) với ý nghĩa của *x* và *y* lúc này không còn là "từ" (word) như trong tiếng Anh mà được hiểu là *tiếng* (xem giải thích ở mục 3.2.3.) trong tiếng Hoa.

$$I(x;y) \equiv \log_2 \frac{P(x,y)}{P(x)P(y)} \qquad (1a)$$

Trong đó:

- *x* và *y* là hai tiếng cần kiểm tra mức độ kết hợp lẫn nhau trong tiếng Hoa.
- *I(x;y)* là thông tin tương hỗ của hai tiếng.
- *P(x), P(y)* là xác suất xuất hiện độc lập của tiếng *x* và của tiếng *y*.
- *P(x,y)* là xác suất xuất hiện đồng thời tiếng *x* và tiếng *y*.

Cách tính MI dành cho từ ghép 3 tiếng như sau [Su et al, 1991]:

$$I(x;y;z) \equiv \log_2 \frac{P_D(x,y,z)}{P_I(x,y,z)} \qquad (1b)$$

Trong đó:

- $P_D(x,y,z) \equiv P(x,y,z)$ là xác suất xuất hiện đồng thời của x, y và x, (**D**ependently)

- $P_I(x,y,z)$ là xác suất xuất hiện độc lập của x,y, z (**I**ndependently) với $P_I(x,y,z) \equiv P(x)P(y)P(z) + P(x)P(y,z) + P(x,y)P(z)$.

Nhìn chung I(.) >>0 sẽ cho biết từ ghép đó có mức độ liên quan giữa các tiếng là rất chặt chẽ. Ngược lại, các tiếng có xu hướng xuất hiện một cách độc lập.

Một cách tính MI khác cũng được Ong & Chen (1999) đề nghị như sau:

$$MI(cw) = \frac{p(w_1 \,\&\, w_2 \,\&\, \ldots \,\&\, w_n)}{p(lw) + p(rw) - p(w_1 \,\&\, w_2 \,\&\, \ldots \,\&\, w_n)} \qquad (2)$$

Trong đó

- cw = p(w_1 & w_2 ...& w_{n-1})
- lw = p(w_1 & w_2 ...& w_{n-1})
- rw = p (w_2 & w_3 ...& w_n)

Theo nghiên cứu của chúng em, hiện nay công trình nghiên cứu về cách tách từ dựa trên độ tương hỗ MI trên tiếng Việt chưa nhiều. Ở đây, chúng em xin giới thiệu cách tính MI được đề nghị trong IGATEC trong [H. Nguyen et al, 2005]

$$MI(cw) = \frac{p(w_1 \;\&\; w_2 \;\&\; \ldots \;\&\; w_n)}{\sum_{j=1}^{n} p(w_j) \;-\; p(w_1 \;\&\; w_2 \;\&\; \ldots \;\&\; w_n)} \qquad (3)$$

Nhìn vào các công thức tính MI, ta có thể dự đoán được mỗi công thức ưu tiên cho một loại từ khác nhau. Phần tiếp theo sau đây sẽ trình bày một số nhận xét về các công thức trên để làm cơ sở đưa ra lựa chọn phù hợp nhất.

4.3.2.2. Cách tính tần số tương đối (Relative Frequency Count)

Cách tính tần số tương đối cho từ ghép có i tiếng được định nghĩa như sau [Su et al, 1993]:

$$r_i = \frac{f_i}{K}$$

Trong đó, f_i là số lần xuất hiện của từ ghép có i tiếng (i^{th} n-gram) trong tập ngữ liệu, và K là số lần xuất hiện trung bình của một từ. Nói một cách khác, f_i được bình thường hoá bằng cách chia cho K để lấy tỉ lệ liên quan. Một cách trực quan, ta sẽ

nhận ra, cách tính RFC sẽ ưu tiên cho những từ xuất hiện với tần số rất cao mà nó sẽ bỏ mất những xuất hiện trong từ điển với tần số thấp. Vì vậy, RFC được dùng như một thuộc tính hỗ trợ thêm cho việc tách từ.

4.3.2.3. Nhận xét về cách sử dụng MI và RFC

Nếu ta sử dụng đồng thời MI và RFC cho việc tách từ sẽ đem lại kết quả như mong đợi bởi vì nếu chỉ sử dụng một công cụ tính toán, kết quả chúng ta đạt được có thể chỉ ưu tiên cho một cách tách nào đó. Nếu chỉ sử dụng RFC, hệ thống của chúng ta có xu hướng chọn những từ xuất hiện nhiều lần nhưng lại có độ liên quan MI thấp. Ví dụ, nếu *P(x)* và *P(y)* rất lớn, nó có thể tạo ra *P(x,y)* cũng rất lớn mặc dù *x* và *y* không hề liên quan gì cả vì *P(x,y)/ P(x) x P(y)* rất nhỏ.

Mặc khác, nếu chỉ sử dụng MI thôi, thì ở trường hợp P(x) và P(y) quá nhỏ sẽ dẫn đến kết quả không đáng tin cậy. Một từ n-gram có thể có MI cao không bởi vì chúng kết hợp chặt chẽ với nhau mà bởi vì khi chia hai số cùng nhỏ như nhau, ta sẽ có số MI lớn.

Tóm lại, ta nên sử dụng cả hai thông tin MI và RFC vì thực tế, một nhóm các từ vừa có RFC và MI cao sẽ có xu hướng vừa kết hợp chặt chẽ với nhau, vừa được sử dụng rộng rãi.

4.3.3. Nhận xét về các cách tính độ liên quan khi áp dụng cho tiếng Việt

- Tiếng Hoa là loại ngôn ngữ đơn lập giống tiếng Việt, nên ta có thể áp dụng một số công tình nghiên cứu trên tiếng Hoa lên tiếng Việt.
- Về mặt lý thuyết, ta hoàn toàn có thể sử dụng các công thức MI trên để áp dụng cho tiếng Việt, và quan thực nghiệm, chúng ta sẽ đề xuất thêm một số cải tiến để công thức tính MI phù hợp với việc tách tiếng Việt hơn nữa.
- Đối với công thức RFC, ta cần phân biệt khái niệm f trong công thức là tần số xuất hiện của từ trong tập ngữ liệu, K là số lần xuất hiện trung bình của một từ (real word) trong tập ngữ liệu. Khi sử dụng tập ngữ liệu, các số f và K là hoàn toàn tính được. Tuy nhiên, phương pháp IGATEC mà chúng em sử dụng lại lấy kết quả số lượng trang web p chứa từ cần tìm nên chúng ta không thể tính được số K (vì không thể dựa vào số lượng trang web trả về

mà quyết định đó là từ hay không). Do vậy, hiện tại, chúng em vẫn chưa áp dụng cách tính RFC trên tiếng Việt.

- Bản chất của phương pháp tính t-score là tìm sự khác nhau trong việc sử dụng từ trong tiếng Anh, chúng em nhận thấy chưa thật sự cần thiết trong việc tách từ làm tăng tính phức tạp của việc tính toán. Do đó, chứng em chưa áp dụng t-score vào tách từ.

4.4. Tiền xử lý (Pre-processing)

Bởi vì các bài báo điện tử được trình bày dưới dạng html, nên trước khi thực hiện tách từ để phân loại, chúng em phải xử lý văn bản để lấy ra những nội dung quan tâm.

4.4.1. Xử lý văn bản đầu vào

Nội dung tóm tắt của bài báo là rất quan trọng vì nó thể hiện nội dung bài báo một cách cô đọng, súc tích, rõ ràng, giúp người xem dự đoán được đề tài của bài báo muốn đề cập đến. Chính vì lý do đó, chúng em quyết định thực hiện việc phân loại tin tức dựa trên phần tóm tắt của bài báo để tiết kiệm thời gian xử lý và đạt được kết quả chính xác cao.

Trong mỗi văn bản, khối tiền xử lý sẽ nhận diện tiêu đề, tóm tắt… của bài báo bằng cách dựa vào thông tin định dang của các thẻ trong trang html. Theo khảo sát của chúng em về cấu trúc hiển thị nội dung trang báo điện tử ở các trang web tin tức ở Việt Nam, tác giả luôn trình bày nội dung tóm tắt (abstract) của bài báo trước bài viết chi tiết, nên hướng phân loại dựa trên tóm tắt của bài báo là khả thi.

Hình 4. 1. Nội dung thông tin cần lấy

Sau khi rút trích được nội dung cần thiết, chúng em tiếp tục thực hiện tách ngữ, phục vụ cho công việc tách từ.

4.4.2. Tách ngữ & tách stopwords

Tách ngữ: Ứng với mỗi văn bản đã rút trích từ trang web, chúng em tiến hành loại bỏ các ký hiệu, các chữ số không cần thiết, sau đó, phân tích văn bản thành các ngữ phân cách bởi dấu câu.

Tách stopword: Nhằm làm tăng tốc độ tính toán của GA và lượt bớt các từ không có nghĩa phân loại trong câu, chúng em có thử nghiệm tách stopword trước khi tiến hành tách từ. Bước tách stopword tỏ ra khá hiệu quả trong việc làm tăng tốc độ GA nhờ chia nhỏ các ngữ ra thành những ngữ nhỏ hơn. Tuy nhiên, cách tách stopword không phải lúc nào cũng cho kết quả như mong đợi bởi vì tách stopword trước khi tách từ sẽ có nhiều khả năng làm sai lạc ý nghĩa của câu, ảnh hưởng đến việc phân loại sau đó. Do đó, chúng em đã thử nghiệm việc tách stopword sau khi

đã tách từ, kết quả phân loại sau khi đã loại bỏ stopword là khả quan hơn cách thực hiện ban đầu. (*Xin xem chương 6 để biết kết quả thực nghiệm.*)

4.5. Hướng tiếp cận tách từ dựa trên thống kê từ Internet và thuật toán di truyền (Internet and Genetic Algorithm-based)

Chúng em xây dựng hai công cụ hỗ trợ cho việc tách từ gồm: *công cụ trích xuất thông tin từ Google và công cụ tách từ dùng thuật toán di truyền*.

4.5.1. Công cụ trích xuất thông tin từ Google

4.5.1.1. Mục đích

Ngày nay, cùng với sự phát triển nhanh chóng của các công nghệ thông tin hiện đại, Internet đã trở thành một thư viện tuyệt vời với một khối lượng văn bản đồ sộ. Do đó, việc khai thác thông tin từ world-wide-web như một tập ngữ liệu khổng lồ cho các công trình nghiên cứu sẽ rút ngắn được thời gian và công sức tự xây dựng một tập ngữ liệu riêng. Với sự giúp sức của công cụ tìm kiếm miễn phí trên mạng, những thông tin cần thiết sẽ được lấy về một cách nhanh chóng và chính xác. Chúng em chọn Google là công cụ tìm kiếm chính bởi vì những ưu thế về tính nhanh chóng, chính xác, và phổ biến của nó so với các công cụ tìm kiếm khác.

Trong luận văn này, chúng em cần hai loại thông tin*:*

- *Tần số xuất hiện của các văn bản chứa các từ (document frequency)* trên các trang web để làm tính công thức MI, dự đoán khả năng tồn tại của một từ là đúng hay không
- *Tần số các văn bản chứa từ với từ khóa đại diện cho chủ đề* dùng để tính mức độ liên quan của từ với các chủ đề cần phân loại.

Do vây, nhiệm vụ của công cụ trích xuất thông tin từ Google sẽ lấy kết quả tìm kiếm của Google, trả về cho chương trình khi chúng ta đưa yêu cầu tìm kiếm.

4.5.1.2. Các công thức tính xác suất và độ tương hỗ

4.5.1.2.1. Các công thức tính xác suất

Khi nhận được kết quả trả về, dựa vào nền tảng của các công trình nghiên cứu về thống kê trên Internet của Rudi & Paul (2005), chúng em sẽ sử dụng các công thức sau đây để tính toán chỉ số MI.

Các công thức tính xác suất các từ xuất hiện trên Internet :

- Gọi *count(w)* là số lượng trang web chứa từ w
  count(w_1 & w_2) là số trang web chứa đồng thời w_1 và w_2
- $p(\text{w}) = \frac{count(w)}{MAX}$
- $p(w_1 \& w_2) = \frac{count(w_1 \& w_2)}{MAX}$
- Trong đó, $MAX = 4 * 10^9$;

4.5.1.2.2. Các công thức tính độ tương hỗ (Mutual Information – MI)

Đối với hướng tiếp cận N-Gram để tách từ, công thức MI để tính toán khả năng tồn tại một ngữ cần tách trong câu là rất quan trọng. Độ tương hỗ (Mutual Information) cho biết thông tin phụ thuộc lẫn nhau của các từ ghép được cấu tạo bởi *n* tiếng ($cw = w_1\ w_2\ \ldots\ w_n$) . Đối với từ một tiếng, ta quy ước $MI = p(w)$. Đối với từ ghép từ 2 tiếng trở lên, chúng em thử nghiệm 3 cách tính MI để tìm ra các tính hiệu quả nhất.

- MI theo cách tính của IGATEC [H. Nguyen et al, 2005]) (đã được trình bày ở mục 4.3.2.1.)
  - $MI(cw) = \frac{p(w_1\ \&\ w_2\ \&\ \ldots\ \&\ w_n)}{\sum_{j=1}^{n} p(w_j)\ -\ p(w_1\ \&\ w_2\ \&\ \ldots\ \&\ w_n)}$ (2)
- MI theo cách tính của [Ong & Chen, 1999] (đã được trình bày ở mục 4.3.2.1.)
  - Giả sử ta có
    - $cw = p(w_1 \& w_2 \ldots \& w_{n-1})$
    - $lw = p(w_1 \& w_2 \ldots \& w_{n-1})$

- $rw = p\ (\ w_2 \& \ w_3 ... \& \ w_n)$

✓ $MI(cw) = \frac{p(w_1 \& w_2 \& \ ... \ \& w_n)}{p(lw) + p(rw) - p(w_1 \& w_2 \& \ ... \ \& w_n)}$ (3)

➢ MI do chúng em đề nghị:

✓ Giả sử ta có

- $cw = p(\ w_1 \& \ w_2 ... \& \ w_{n-1})$
- Với n chẵn : $lw = p(\ w_1 \& \ w_2 ... \& \ \ w_{n/2})$, $rw = p\ (\ w_{n/2+1} \& \ w_{n/2+2} ... \& \ w_n)$
- Với n lẻ: $lw = p(\ w_1 \& \ w_2 ... \& \ \ w_{n-1})$, $rw = p\ (\ w_2 \& \ w_3 ... \& \ w_n)$

✓ $MI(cw) = \frac{p(w_1 \& w_2 \& \ ... \ \& w_n)}{p(lw) + p(rw) - p(w_1 \& w_2 \& \ ... \ \& w_n)}$ (4)

Chúng ta sẽ sử dụng các công thức trên để tính độ thích nghi của các cá thể trong thuật toán di truyền dưới đây. Kết quả của mỗi công thức tính MI sẽ ưu tiên cho những loại từ ghép khác nhau mà ta sẽ hiểu rõ hơn trong kết quả thực nghiệm ở chương 6.

4.5.2. Công cụ tách từ dùng thuật toán di truyền (Genetic Algorithm – GA)

Mục đích của chúng ta là tìm ra các cách tách từ hợp lý nhất cho văn bản, tuy nhiên, chúng ta gặp phải trở ngại là không gian tìm kiếm (search space) quá lớn do sự bùng nổ tổ hợp khi sinh ra dãy các từ. Như chúng ta đều biết, thuật toán di truyền (Genetic Algorithm – GA) được biết đến với khả năng duyệt tắt qua những không gian tìm kiếm lớn một cách hiệu quả và đưa ra những giải pháp toàn cục tối ưu nhất. GA thực hiện tiến hoá một số thế hệ để tạo ra một quần thể gồm những cá thể tối ưu nhờ vào các bước lai ghép (cross-over), đột biến (mutation), sinh sản (reproduction), và cách chọn lựa cá thể. Chất lượng của mỗi cá thể được tính toán dựa trên chỉ số fitness cho mỗi cá thể và quần thể. Trong quá trình thử nghiệm, chúng em chọn top N cá thể chất lượng nhất sau khi thực hiện các bước lai ghép, đột biến, sinh sản.

4.5.2.1. Khảo sát độ dài của "từ" trên từ điển

Như chúng ta đều biết, thuật toán di truyền đòi hỏi phải có rất nhiều tham số cho các bước thực hiện như số cá thể trong quần thể, số thế hệ tiến hoá, tỉ lệ lai ghép, tỉ lệ đột biến… Do vậy, chất lượng lựa chọn các tham số trên sẽ quyết định kết quả của thuật toán di truyền. Chính vì tính chất quan trọng của các tham số, chúng em thực hiện một khảo sát nhỏ về số lượng từ tương ứng với chiều dài từ trên từ điển thông dụng tại http://dict.vietfun.com để làm cơ sở cho các tham số sau này.

Độ dài từ (tiếng)	Tần số xuất hiện	Tỉ lệ
1	8933	12.2
2	48995	67.1
3	5727	7.9
4	7040	9.7
≥ 5	2301	3.1
Tổng cộng	72994	100

Bảng 4. 1. Thống kê độ dài từ trong từ điển

Có một điều cần lưu ý là tại thời điểm này, chúng ta vẫn chưa có một từ điển chuẩn nào được dùng cho việc xử lý ngôn ngữ, do đó, chúng em quyết định dùng loại từ điển phổ dụng để thống kê. Theo kết quả thống kê, trên 67% là từ ghép hai tiếng, còn lại khoảng 30% là các từ ghép một, ba, bốn tiếng. Các cụm từ dài hơn bốn tiếng chiếm khoảng 3%, tuy nhiên các cụm từ đó đa số là các câu thành ngữ của Việt Nam.

Kết quả thống kê trên có ý nghĩa rất quan trọng đối với công cụ tách từ bằng GA của chúng em. Dựa trên tỉ lệ của các loại từ, chúng em thực hiện việc khởi tạo cá thể ngẫu nhiên có thêm thông tin về xác suất xuất hiện của từ và đó là cơ sở để chúng em quyết định cách tách từ phù hợp với thực tế của tiếng Việt. Chi tiết về các ứng dụng của kết quả khảo sát sẽ được chúng em trình bày ở các phần sau.

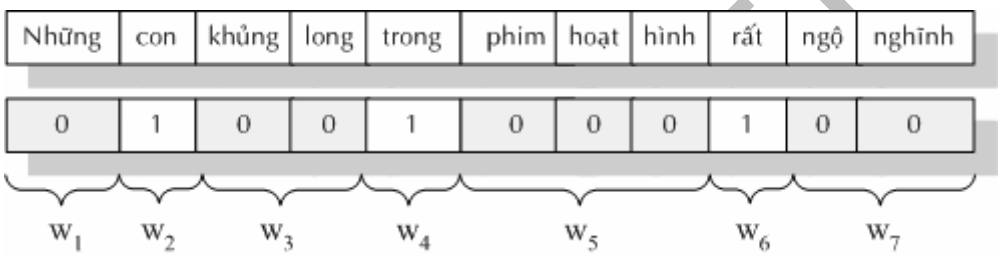
4.5.2.2. Khởi tạo quần thể

4.5.2.2.1. Biểu diễn cá thể

Giả sử văn bản đầu vào *t* được tạo thành bởi *n* tiếng (syllables) như sau: $t=s_1s_2...s_n$. Mục đích của quá trình chạy GA là tìm ra cách tách từ có độ chấp nhận cao nhất : $t=w_1w_2...w_m$, với $w_k= s_i... s_j$ $(1 \le k \le m, 1 \le i,j \le n)$

Tương tự như IGATEC, chúng em cũng biểu diễn mỗi cá thể (*id*) trong quần thể (*pop*) bởi chuỗi các bit 0,1, trong đó, mỗi bit đại diện cho một tiếng trong văn bản, mỗi nhóm bit cùng loại đại diện cho một *từ* (*word)*.

Ví dụ: Với câu "Những || con || khủng long || trong || phim hoạt hình || rất || ngộ nghĩnh", chúng em sẽ biểu diễn dưới dạng các bit 0, 1 như sau:



Hình 4. 2. Biểu diễn cá thể bằng các bit 0,1

4.5.2.2.2. Khởi tạo các tham số

Ở bước khởi tạo tham số, ta phải thiết lập một vài tham số cơ bản cho GA như số thế hệ tiến hoá (generations), kích thước quần thể (population size), tỉ lệ lai ghép (reproduction fraction)… Ngoài ra, vì mỗi cá thể của chúng ta là một thể hiện cách tách từ trong câu, nên ta sẽ lợi dụng tính chất liên kết của các từ để thực hiện khởi tạo cá thể ngẫu nhiên ban đầu. Tính chất liên kết của từ được thể hiện qua tỉ lệ của các từ trong từ điển, nên ta sẽ có thêm tham số về khả năng xuất hiện từ trong câu ở bảng tham số dưới đây.

Tham số	Giá trị
Số thế hệ tiến hoá	100
Kích thước quần thể	50
Tỉ lệ lai ghép	95%
Tỉ lệ đột biến	5%
Top N cá thể được chọn	100
Tỉ lệ từ 1 tiếng (mono-gram)	10%
Tỉ lệ từ 2 tiếng (bigram)	70%
Tỉ lệ từ 3 tiếng (trigram)	10%
Tỉ lệ từ 4 tiếng (quadgram)	10%

Bảng 4. 2. Tham số thực hiện GA

4.5.2.2.3. Khởi tạo cá thể

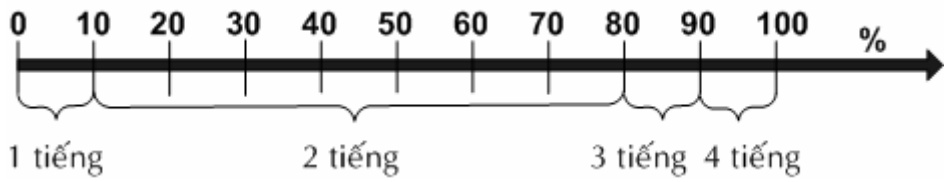
Như chúng ta đều biết, quy tắc của thuật toán di truyền là thực hiện tiến hoá các cá thể qua các thế hệ nhằm đạt đến *độ hội tụ* của *chỉ số thích nghi* (sẽ được nói rõ ở mục **4.5.2.3**). Nếu cá thể được khởi tạo ngẫu nhiên sẽ có độ thích nghi thấp, chúng ta phải tiến hoá qua rất nhiều thế hệ để đạt đến độ hội tụ cần thiết. Và hậu quả là số thế hệ tiến hoá càng nhiều thì thời gian tiêu tốn và chi phí tính toán càng cao. Giải pháp khắc phục nhược điểm trên là khởi tạo một số cá thể ban đầu gần với điểm hội tụ, nhờ vậy có thể rút ngắn được số thế hệ tiến hoá, tăng tốc độ. Ở bước khởi tạo quần thể, chúng em tạo ra cá các thể bằng hai cách: khởi tạo ngẫu nhiên và khởi tạo dựa trên phương pháp *MM:forward/backward* [Chih-Hao Tsai, 2000].

4.5.2.2.3.1. Khởi tạo cá thể ngẫu nhiên

Theo thống kê ở bảng 4.1, chúng em quyết định đặt ra một số giới hạn cho việc tạo cá thể ngẫu nhiên. Đầu tiên, tất cả các từ ghép w_k tạo ra có độ dài không quá 4.

Thứ hai, chúng em khởi tạo ngẫu nhiên các cá thể có số lượng từ tương ứng với tỉ lệ về độ dài từ ở trên, nhằm tạo ra điểm xuất phát tốt cho quá trình thực hiện GA.

Ví dụ: Giả sử ta có câu đầu vào "Những con khủng long trong phim hoạt hình rất đáng yêu" gồm 11 tiếng. Theo các tham số khởi tạo của bảng 4.2., chúng em thiết lập giới hạn tạo từ ngẫu nhiên trong câu:



Hình 4. 3. Thang tỉ lệ phát sinh loại từ

Một bộ phát sinh ngẫu nhiên sẽ phát sinh xác suất f $(0 \le f \le 1)$ để chọn loại từ:

- Nếu $0 \le f < 0.1$: phát sinh loại từ 1 tiếng
- Nếu $0.1 \le f < 0.8$: phát sinh loại từ 2 tiếng
- Nếu $0.8 \le f < 0.9$: phát sinh loại từ 3 tiếng
- Nếu $0.9 \le f \le 1$: phát sinh loại từ 4 tiếng

4.5.2.2.3.2. Khởi tạo cá thể bằng Maximum Matching : forward/backward

(Phương pháp *Maximum Matching : forward/backward* [Chih-Hao Tsai, 2000] đã được trình bày ở mục 3.2.1.)

Đây là bước khởi tạo rất quan trọng và điểm cải tiến đáng kể so với IGATEC. Chúng em chọn phương pháp *MM: forward/backward* để khởi tạo cá thể ban đầu vì độ chính xác khá cao của phương pháp này sẽ tạo ra cá các thể gần đúng nhất, giúp tăng tốc quá trình tiến hoá. Ngoài ra, việc áp dụng phương pháp MM theo dạng đơn giản chỉ cần duyệt tuyến tính, sẽ giảm thiểu được chi phí và thời gian tính toán so với các phương pháp khác.

Chúng em thực hiện tách từ theo hai hướng từ trái sang phải, và từ phải sang trái. Nếu hai cách tách từ trên trùng nhau, chúng em sẽ chọn một và gộp vào các cá thể đã được phát sinh ngẫu nhiên ở trên.

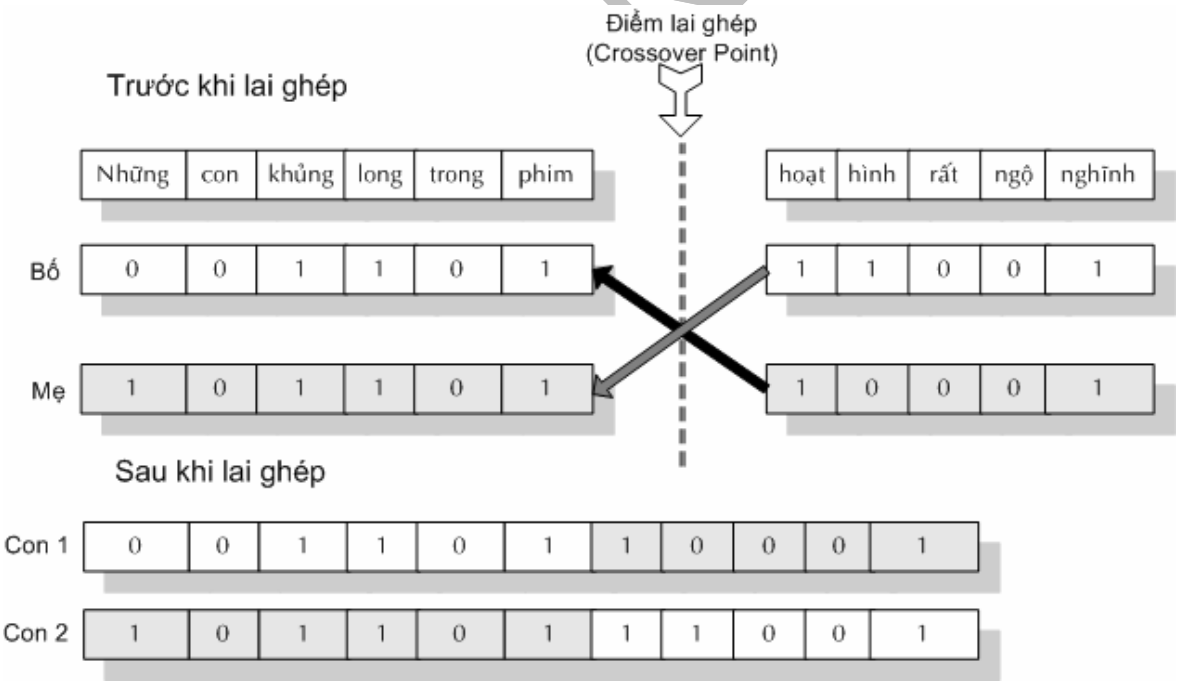
Sau khi khởi tạo xong, quần thể sẽ được tiến hóa qua các quá trình lai ghép, đột biến, sinh sản,

4.5.2.3. Thực hiện tiến hoá

4.5.2.3.1. Quá trình lai ghép (cross-over)

Chúng em áp dụng phương pháp chuẩn của lai ghép là dựa trên một điểm ngẫu nhiên trong chuỗi bit của cá thể. Khi có một cặp cá thể *bố mẹ* , thế hệ con được tạo ra dựa trên sự kết hợp từ phần đầu tiên của *bố* với phần cuối của *mẹ* và ngược lại. Tuy nhiên, trong quá trình lai ghép, chúng em nhận thấy giới hạn từ ghép tối đa 4 tiếng có thể bị phá vỡ, do đó, đối với những phân đoạn w_k nào có độ dài hơn chúng em sẽ thực hiện việc chuẩn hóa từ vị trí đó đến cuối sao cho không có một từ nào vượt quá 4 tiếng.

Ví dụ:



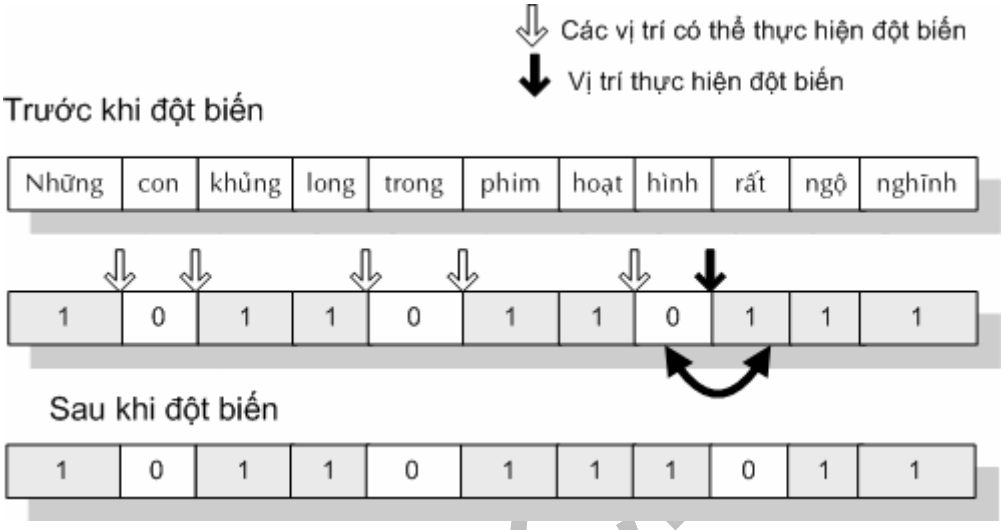
Hình 4. 4.Quá trình lai ghép

4.5.2.3.2. Quá trình đột biến (mutation)

Thay vì thực hiện phương pháp bật tắt bit (bit flip), chúng em thực hiện việc hoán chuyển vị trí của hai bit liền nhau tại một vị trí ngẫn nhiên. Ý tưởng thực hiện

đột biến như thế này bởi vì, trong việc phân định ranh giới từ, ta dễ dàng nhận ra rằng một tiếng nếu kết hợp với tiếng trước không phù hợp thì có thể kết hợp với từ đứng sau sẽ phù hợp hơn, hoặc là đứng một mình. Tương tự như phần lai ghép, chúng em thực hiện chuẩn hoá các cá thể sau khi đột biến.

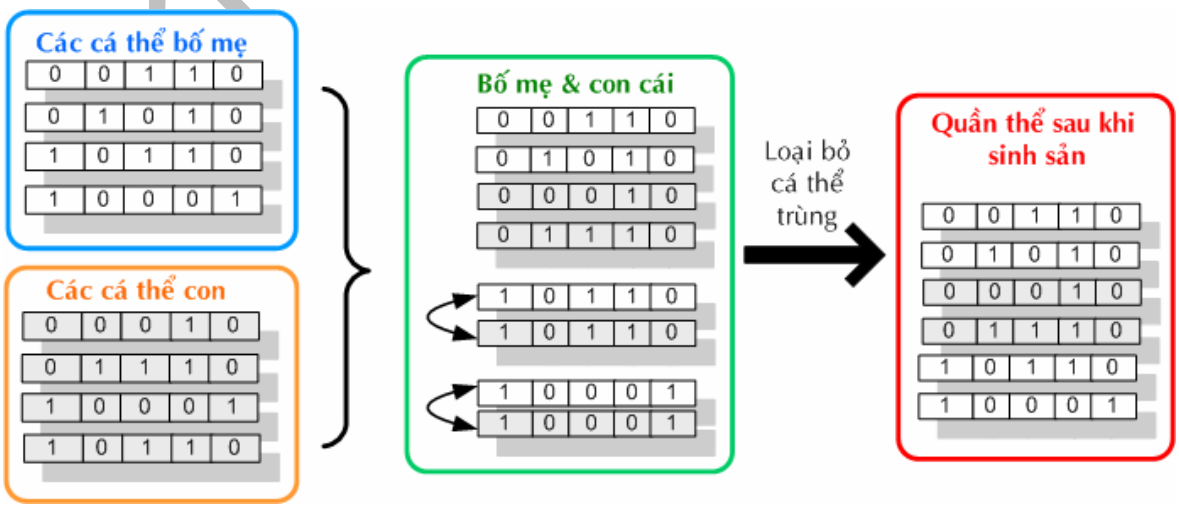
Ví dụ:



Hình 4. 5. Quá trình đột biến

4.5.2.3.3. Quá trình sinh sản (reproduction)

Sau khi lai ghép và đột biến, chúng em kết hợp các cá thể bố mẹ với cá thể con vừa được tạo ra để phục vụ cho bước chọn cá thể. Sau khi kết hợp, chúng em lọc bỏ các cá thể trong quần thể, để đạt được nhiều cách tách từ tốt. Ví dụ:



Hình 4. 6. Quá trình sinh sản

4.5.2.3.4. Quá trình chọn cá thể (selection)

Ở mỗi thế hệ, chúng em chỉ chọn top N cá thể từ quá trình sinh sản ở trên. Trước tiên, các cá thể sẽ được tính độ thích nghi (fitness) chính là tổng giá trị MI của các từ được tách trong câu. Sau đó, quần thể sẽ được sắp xếp theo giá trị của độ thích nghi giảm dần, quá trình chọn lọc cá thể sẽ chọn top N cá thể có độ thích nghi cao nhất để tạo nên quần thể tiếp tục tiến hoá ở các thế hệ sau.

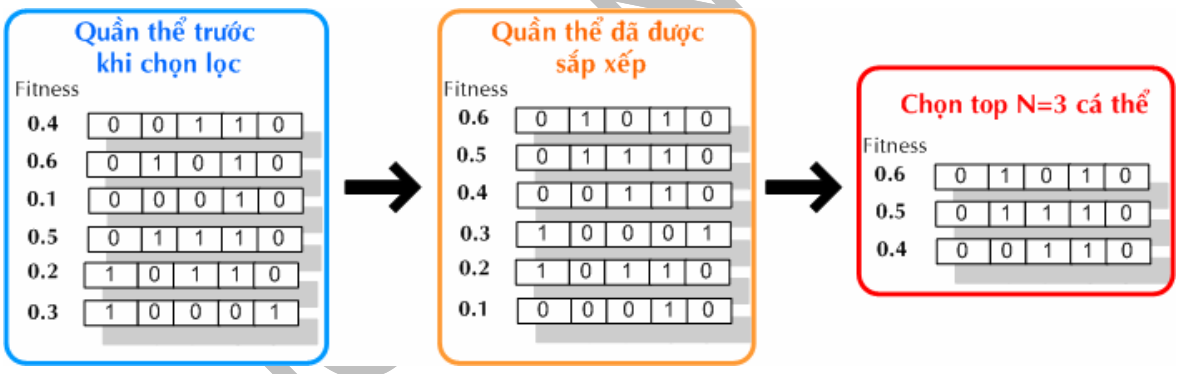
Cách thức lựa chọn cá thể như sau:

$$fit(id) = fit(w_1 w_2 ... w_m) = \sum_{k=1}^{m} MI(w_k)$$

$$fit(pop) = \sum_{i=1}^{N} fit(id_i)$$

Trong đó, $id = w_1 w_1 ... w_1$ là một cá thể trong quần thể $pop = \{id_1, id_2\}$

Ví dụ:



Hình 4. 7. Quá trình chọn cá thể

Có thể nói đây là quá trình quan trọng nhất trong cả tiến trình tiến hoá vì sự lựa chọn cá thể ở bước này sẽ quyết định cá thể tiến hoá có tốt hay không. Ở quá trình chọn lọc cá thể này, chúng em đã thử nghiệm một số công thức tính độ tương hỗ (Mutual Information) như đã trình bày ở trên, và thu được các kết quả khác nhau khi sử dụng các công thức khác nhau. Từ đó, chúng em rút ra một số kết luận và nhận xét quan trọng về ưu khuyết điểm của các công thức MI.

Kết quả thực nghiệm và nhận xét về các công thức MI sẽ được chúng em trình bày ở chương 6.

4.5.2.3.5. Độ hội tụ (convergence)

Quá trình thực hiện GA cố gắng làm tăng độ thích nghi (fitness) của mỗi cá thể cũng đồng nghĩa với việc tăng chất lượng của từ được tách. Ở mỗi thế hệ tiến hoá, chỉ số thích nghi của quần thể sẽ tăng dần đến một ngưỡng gọi là độ hội tụ α. Khi đó, độ chênh lệnh chỉ số thích nghi của quần thể giữa hai thế hệ sẽ nhỏ dần và tiến dần đến 0. Vì vậy, chúng em thực hiện việc ngừng GA một cách tự động khi giá trị fitness của các thế hệ đạt đến độ hội tụ có chỉ số $\alpha = 10^{-7}$ hoặc số thế hệ đạt đến số lượng mặc định đã trình bày ở trên.

Việc ngừng GA tự động sẽ giúp chúng ta giảm thiểu thời gian và chi phí tính toán không cần thiết, đồng thời là tăng tốc độ của việc tách từ.

4.6. Kết luận

Phương pháp tách từ dựa trên thống kê Internet và thuật toán di truyền tương đối đơn giản hơn các phương pháp khác và tỏ ra khá linh hoạt với sự biến động của ngôn ngữ trong tin tức điện tử. Ngoài ra, đây là hướng tiếp cận khá mới mẻ, hạn chế được khuyết điểm cơ bản của các phương pháp tách từ lâu nay là dựa trên tập ngữ liệu đã đánh dấu và từ điển chuyên biệt. Với ưu điểm là thuật toán đơn giản, dễ hiểu, dễ cài đặt, nhưng phương pháp IGATEC vẫn cho một kết quả tách từ chấp nhận được, có thể dùng trong phân loại văn bản.

Chương 5

BÀI TOÁN PHÂN LOẠI TIN TỨC ĐIỆN TỬ

Lý do chọn phương pháp Naïve Bayes

Thuật toán Naïve Bayes

- Công thức xác suất đầy đủ Bayes
- Tính độc lập có điều kiện (Conditional Independence)
- Nguồn gốc Naïve Bayes
- Naïve Bayes trong phân loại văn bản
- Hai mô hình sự kiện trong phân loại văn bản bằng Naïve Bayes

Bài toán phân loại tin tức điện tử tiếng Việt

Kết quả

Chương 5. BÀI TOÁN PHÂN LOẠI TIN TỨC ĐIỆN TỬ

Nhằm tận dụng phương pháp tách từ IGATEC đã được đề cập ở chương trên, trong chương này chúng em sẽ giới thiệu cách phân loại tin tức điện tử tự động sử dụng phương pháp Naïve Bayes và giải thích sự phù hợp của Naïve Bayes với phương pháp tách từ IGATEC.

5.1. Lý do chọn phương pháp Naïve Bayes

Như đã được giới thiệu trong chương 2, Naïve Bayes là một phương pháp rất phổ biến sử dụng xác suất có điều kiện giữa từ và chủ đề để xác định chủ đề của văn bản. Các xác suất này dựa trên việc thống kê sự xuất hiện của từ và chủ đề trong tập huấn luyện. Tập huấn luyện lớn có thể mang lại kết quả khả quan cho Naïve Bayes. Internet với hơn 10 tỷ trang web là một tập huấn luyện rất phong phú về mọi chủ đề trong cuộc sống. Hơn nữa, với số lượng chủ đề tin tức điện tử không nhiều (khoảng 20 chủ đề) thì việc sử dụng Internet như cơ sở dữ liệu huấn luyện rất phù hợp. Trong báo chí, với mỗi chủ đề luôn có các từ chuyên môn với tần số xuất hiện rất cao, việc tận dụng tần số phụ thuộc của các từ này vào chủ đề có thể đem lại kết quả khả quan cho phân loại.

Với dữ liệu được tạo ra nhờ công cụ tách từ GA và trích xuất thông tin từ Google, theo đánh giá của chúng em, thì phương pháp Naïve Bayes là khá phù hợp vì các dữ liệu đầu vào cho hướng phân loại này hoàn toàn phù hợp với dữ liệu hiện có. Điều này sẽ giúp chúng em tiết kiệm được rất nhiều thời gian và công sức tạo thêm nhiều tập dữ liệu nếu chọn phương pháp phân loại khác.

Mặt khác, phương pháp Naïve Bayes là phương pháp khá cổ điển được sử dụng đầu tiên bởi Maron vào năm 1961 [Maron, 1961], và sau đó rất phổ biến trong các lãnh vực tìm kiếm, lọc mail, các bộ lọc mail… nên chúng ta có thể tin tưởng về xác suất chính xác và các ưu khuyết điểm của phương pháp này để áp dụng phù hợp.

Một lý do nữa mà chúng em chọn Naïve Bayes bởi phương pháp đơn giản, tốc độ nhanh, cài đặt tương đối không quá phức tạp phù hợp với thời gian cho phép của luận văn. Chúng em không sử dụng kNN, do tập dữ liệu thử nghiệm hiện có là tập

các tin tức vắn tắt lấy ngẫu nhiên từ trang VnExpress.net còn khá nhỏ (dưới 1000). Trong khi đó để có thể sử dụng phương pháp kNN hiệu quả số lượng chủ đề và dữ liệu thử nghiệm phải lớn hơn nhiều. SVM tuy là một phương pháp được cho là có hiệu suất cao, nhưng thời gian huấn luyện lại rất lâu. Nnet lại cài đặt quá phức tạp.

Với những lý do trên, chúng em đề xuất chọn phương pháp Naïve Bayes để phân loại văn bản.

5.2. Thuật toán Naïve Bayes

Theo tác giả Mitchell (2005), thuật toán phân loại Naïve Bayes có đặc điểm nổi bật là có khả năng giảm độ phức tạp tính toán từ $2(2^n - 1)$ về còn $2n$. Thế đặc điểm nào giúp Naïve Bayes có khả năng đó?

5.2.1. Công thức xác suất đầy đủ Bayes

Giả sử ta muốn tính toán một hàm không biết giá trị đích $f : X \to Y$ tương đương với P(Y|X).

Đầu tiên, ta cho rằng Y là biến ngẫu nhiên có giá trị luận lý (boolean).

X là vector gồm n thuộc tính luận lý (boolean), $X = \langle X_1, X_2, ..., X_n \rangle$

Áp dụng luật Bayes, $P(Y=y_i|X)$ được tính như sau:

$$P(Y = y_i \mid X = x_k) = \frac{P(X = x_k \mid Y = c_i)P(Y = y_i)}{\sum_j P(X = x_k \mid Y = c_j)P(Y = y_i)} \quad (2.1)$$

Trong đó $P(X|Y)$ và $P(Y)$ được học từ tập huấn luyện. Tuy nhiên để tính toán chính xác $P(X|Y)$ thường đòi hỏi rất nhiều dữ liệu huấn luyện. Để hiểu tại sao, chúng ta sẽ tính toán số lượng tham số cần thiết khi Y là biến boolean, X là vector gồm n thuộc tính boolean :

$$\theta_{ij} = P(X = x_i \mid Y = y_j)$$

Trong i phải dựa trên 2^n giá trị có thể cho những giá trị của vector X và j cần 2 giá trị. Do đó, chúng ta cần tính toán khoảng 2^{n+1} tham số. Mặc khác, ta phải đảm bảo $1 = \sum_{i=1}^{n} \theta_{ij}$ cho bất kỳ j cố định nào. Vì vậy, ứng với một giá trị đặc biệt y_j, và 2^n

giá trị có thể của x_i, chúng ta chỉ cần tính toán $2^n - 1$ tham số độc lập. Dựa trên giá trị của Y (Y là biến boolean), chúng ta cần tính tổng cộng là $2(2^n - 1)$ tham số θ_{ij} .

5.2.2. Tính độc lập có điều kiện (Conditional Independence)

Định nghĩa: cho các biến ngẫu nhiên X, Y và Z, chúng ta nói rằng X là độc lập có điều kiện với Y gây ra Z, nếu và chỉ nếu xác suất phân phối chủ đạo X là độc lập với giá trị của Y gây ra Z. Lúc đó:

$$(\forall i, j, k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

Ví dụ: ta xem ba biến luận lý (boolean) ngẫu nhiên trên đại diện cho các trạng thái của thời tiết là : *Sấm , Mưa,* và *Sét*. Chúng ta đều biết rằng sự kiện *Sấm* xảy ra hoàn toàn độc lập với sự kiện *Mưa* gây ra *Sét*. Bởi vì khi có *Sét* sẽ gây ra tiếng *Sấm*, nên một khi chúng ta biết rằng có *Sét* hay không thì ta có thể biết được giá trị của *Sấm* mà không cần thêm thông tin nào từ *Mưa*. Trên thực tế, rõ ràng có sự phụ thuộc giữa *Mưa* và *Sấm*, tuy nhiên ta không cần thêm thông tin đó một khi ta đã có thông tin về *Sét*.

5.2.3. Nguồn gốc thuật toán Naïve Bayes

Thuật toán phân loại Naïve Bayes dựa trên luật Bayes, với giả định tất cả các thuộc tính $X_1 \ldots X_n$ đều độc lập có điều kiện với nhau do sự kiện Y gây ra. Chính giả thiết này đã đơn giản hóa cách tính của $P(X|Y)$, và vấn đề ước lượng $P(X|Y)$ từ tập ngữ liệu huấn luyện.

Chúng ta hãy xét ví dụ sau, giả sử ta có $X = \langle X_1, X_2 \rangle$, lúc đó

$$\begin{aligned} P(X \mid Y) &= P(X_1, X_2 \mid Y) \\ &= P(X_1 \mid X_2, Y) P(X_2 \mid Y) \\ &= P(X_1 \mid Y) P(X_2 \mid Y) \end{aligned}$$

Kết quả của dòng thứ nhất là theo cách tính thông thường của xác suất, và dòng thứ ba là phân tích trực tiếp theo định nghĩa về độc lập có điều kiện.

Từ đó, ta tổng quát hóa lên khi X chứa n thuộc tính đều độc lập với nhau do sự kiện Y gây ra được biểu diễn như sau:

$$P(X_1 \ldots X_n \mid Y) = \prod_{i=1}^{n} P(X_i \mid Y) \quad (2.2)$$

Chú ý, khi Y và X_i là biến luận lý (boolean), chúng ta chỉ cần $2n$ tham số để định nghĩa $P(X_i=x_{ik}|Y=y_j)$.

Bây giờ, chúng ta hãy xét đến nguồn gốc của thuật toán Naïve Bayes. Giả sử Y là một biến bất kỳ mang giá trị riêng biệt, và các thuộc tính $X_i...X_n$ là thuộc tính rời rạc hoặc liên tục. Mục đích của chúng ta là huấn luyện để thuật toán phân loại trả ra sự phân phối xác suất trên các giá trị của Y đối với mỗi thể hiện X mà ta cần phân loại. Biểu thức sau đây biểu diễn cho xác suất ứng với giá trị thứ k của Y:

$$P(Y = y_k \mid X_1...X_n) = \frac{P(Y = y_k)P(X_1...X_n \mid Y = y_k)}{\sum_j P(Y = y_j)P(X_1...X_n \mid Y = y_j)}$$

Trong đó, tổng giá trị ở mẫu của biểu thức là tổng cho bởi tất cả các giá trị y_j của Y. Lúc này, sử dụng công thức (2.2), ta có thể viết lại công thức trên như sau:

$$P(Y = y_k \mid X_1...X_n) = \frac{P(Y = y_k)\prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i \mid Y = y_j)} \quad (2.3)$$

Công thức (2.3) là công thức cơ bản của phương pháp phân loại Naïve Bayes. Khi cho một thể hiện $X^{new}=\langle X_1...X_n\rangle$, theo công thức trên, ta sẽ tính toán được các xác suất của Y gây ra bởi X^{new} bằng cách dựa vào P(Y) và $p(X_i|Y)$ được ước lượng từ tập ngữ liệu.Nếu chúng ta chỉ quan tâm đến giá trị lớn nhất của Y, thì sử dụng công thức sau:

$$Y \leftarrow \underset{y_k}{\operatorname{argmax}} \frac{P(Y = y_k)\prod_i P(X_i \mid Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i \mid Y = y_j)}$$

5.2.4. Phương pháp Naïve Bayes trong phân loại văn bản

5.2.4.1. Công thức xác suất đầy đủ Bayes

Phương pháp Naïve Bayes tìm chủ đề của văn bản d bằng các xác định chủ đề có xác suất $P(Y = c_i \mid X = d)$, xác suất để văn bản d nằm trong lớp c_i, lớn nhất thông qua việc sử dụng công thức xác suất đầy đủ Bayes :

$$P(Y = c_i \mid X = d) = \frac{P(X = d \mid Y = c_i)P(Y = c_i)}{\sum_j P(X = d \mid Y = c_j)P(Y = c_j)} \quad (2.7)$$

Trong đó

- c_j là chủ đề thứ j
- $d = (w_1, w_2, ..., w_n)$ là văn bản cần phân loại.
- $P(Y=c_i \mid X=d)$ gọi là xác suất xảy ra văn bản d thuộc về chủ đề c_i.
- $P(X=d \mid Y=c_i)$ gọi là xác suất chủ đề c_i có chứa văn bản d trong tập huấn luyện.

Một cách để xác định $P(Y \mid X)$ là sử dụng tập huấn luyện để ước lượng $P(X \mid Y)$ và $P(Y)$. Sau đó sử dụng công thức xác suất đầy đủ trên để xác định $P(Y = c_i \mid X = d)$ với d bất kỳ.

5.2.4.2. Ước lượng P(X\|Y)

Giả sử với mỗi chủ đề, ta có biến cố các từ phụ thuộc vào chủ đề là độc lập có điều kiện (conditional independence) với nhau. Ta có công thức của biểu diễn sự độc lập có điều kiện của 2 biến cố X,Z vào Y được trình bày ở 5.2.2 như sau :

$$P(X \mid Y, Z) = P(X \mid Z)$$

Sử dụng giả định trên ta tính được $P(X = d \mid Y = c_i)$:

$$\begin{aligned} P(X = d \mid Y = c_i) &= P(w_1, w_2, .., w_n \mid Y = c_i) \\ &= P(w_1 \mid Y = c_i)P(w_2 \mid Y = c_i)...P(w_n \mid Y = c_i) \\ &= \prod_{j=1}^{n} P(w_j \mid Y = ci) \end{aligned} \quad (2.8)$$

Từ (2.8), (2.7) được viết lại như sau :

$$P(Y = c_i \mid w_1, w_2, ..., w_n) = \frac{P(Y = c_i)\prod_k P(w_k \mid Y = c_i)}{\sum_j P(Y = c_j)\prod_k P(w_k \mid Y = c_j)} \quad (2.9)$$

Nhờ thống kê trên tập huấn luyện D, $P(X \mid Y)$ có thể được ước lượng theo :

$$P(X = w_j \mid Y = c_i) \simeq \frac{\#D\{X = w_j \wedge Y = c_i\}}{\#D\{Y = c_i\}} \quad (2.10)$$

Trong đó

- $\#D\{X = w_j \wedge Y = c_i\}$: số văn bản trong tập huấn luyện chứa đồng thời w_j và c_i
- $\#D\{Y = c_i\}$: số văn bản trong tập huấn luyện chứa c_i

Công thức ước lượng trên sẽ cho kết quả $P(X = w_j \mid Y = c_i) = 0$ khi không có văn bản chứa đồng thời cả hai (w_j và c_i). Nhằm tránh trường hợp này, ta nên sử dụng phép ước lượng đã được làm mịn sau :

$$P(X = w_j \mid Y = c_i) \simeq \frac{\#D\{X = w_j \wedge Y = c_i\} + l}{\#D\{Y = c_i\} + lR} \tag{2.11}$$

Với

- R : số lượng chủ đề
- l : quyết định độ mịn của phép ước lượng

5.2.4.3. Ước lượng P(Y)

Việc ước lượng $P(Y=c_i)$ đơn giản là tính phần trăm số văn bản trong tập huấn luyện có chủ đề c_i :

$$P(Y = c_i) = \frac{\#D\{Y = c_i\}}{\|D\|} \tag{2.12}$$

5.2.5. Hai mô hình sự kiện trong phân loại văn bản bằng phương pháp Naïve Bayes

5.2.5.1. Giới thiệu

Phân loại văn bản là một lĩnh vực có phạm vi thuộc tính (attribute) rất nhiều bởi vì thuộc tính của những văn bản cần phân loại là từ (word), mà số lượng từ khác nhau thì vô cùng lớn. Và thuật toán Naïve Bayes đã thành công trong việc ứng dụng vào lĩnh vực phân loại với khả năng làm giảm độ phức tạp trên. Mặc dù đây là thuật toán khá phổ biến, nhưng trong cộng đồng phân loại văn bản vẫn có một vài điều lẫn lộn về phương pháp phân loại Naïve Bayes bởi vì có hai mô hình phát sinh khác nhau vẫn thường được sử dụng. Cả hai mô hình đều sử dụng "naïve Bayes assumption" và cả hai đều được giới phân loại gọi là "naïve Bayes".

5.2.5.2. Mô hình đa biến trạng Bernoulli (Multi-variate Bernoulli Model)

Một mô hình biểu diễn một văn bản là một vector có thuộc tính nhị phân cho biết rằng từ nào có hay không xuất hiện trong văn bản. Số lần xuất hiện của một từ trong văn bản là không cần thiết. Ở đây chúng ta có thể hiểu rằng văn bản là sự kiện (event) và sự có mặt hay vắng mặt của các từ trở thành thuộc tính của sợ kiện. Đấy chính là mô hình sự kiện đa biến trạng Bernoulli (multi-variate Bernoulli event model), một mô hình khá truyền thống, đã được nhiều người sử dụng trong phân loại văn bản. Theo McCallum & Nigam (1998), một số công trình tiêu biểu về hướng tiếp cận này là Robertson & Sparck-Jones (1976), Lewis(1992), Kalt & Croft (1996), Larkey & Croft (1996), Koller & Sahami (1997), Sahami (1996).

5.2.5.3. Mô hình đa thức (Multinomial Model)

Mô hình thứ hai cho rằng một văn bản đại diện tập hợp tần số xuất hiện của từ trong văn bản. Do đó, thứ tự xuất hiện của từ được bỏ qua nhưng tần số xuất hiện được giữ lại. Ở đây, chúng ta có thể hiểu rằng những tần số xuất hiện của các từ là những sự kiện (events) và văn bản trở thành tập hợp các sự kiện của từ (word events). Chúng ta gọi đây là sự kiện mô hinh đa thức (Multinomial event model). Đây là hướng tiếp cận thông thường trong mô hình ngôn ngữ học thống kê. Hướng tiếp cận này cũng được rất nhiều người sử dụng mà theo McCallum & Nigam (1998) các công trình tiêu biểu như Lewis & Gale (1994), Kalt & Croft (1996), Joachims (1997), Mitchell (1997), McCallum et al (1998)…

5.2.5.4. Nhận xét

Đối với phương pháp multi-variate model, việc không nắm bắt thông tin tần số xuất hiện của từ có thể đưa đến khuyết điểm không phân biệt được văn bản ưu tiên cho chủ đề nào hơn nếu cả 2 văn bản đều xuất hiện cùng một từ nào đó nhưng tần số lại khác nhau rất nhiều. Ví dụ, nếu từ "thể thao" sẽ xuất hiện nhiều trong các tin tức về thể thao, và sẽ ít xuất hiện trong các tin tức có nội dung khác, nhưng do phương pháp multi-variate không sử dụng thông tin tần số nên không phân biệt được văn bản ưu tiên cho thể thao hơn. Trong khi đó, hướng tiếp cận multinomial model rõ ràng đã sử dụng thông tin về xác suất phân phối từ trong văn bản.

Đối với phương pháp mulnomial, do sử dụng tần số xuất hiện của từ nên sẽ phụ thuộc vào chiều dài văn bản, vì tài liệu càng dài, sự xuất hiện của các từ càng nhiều.

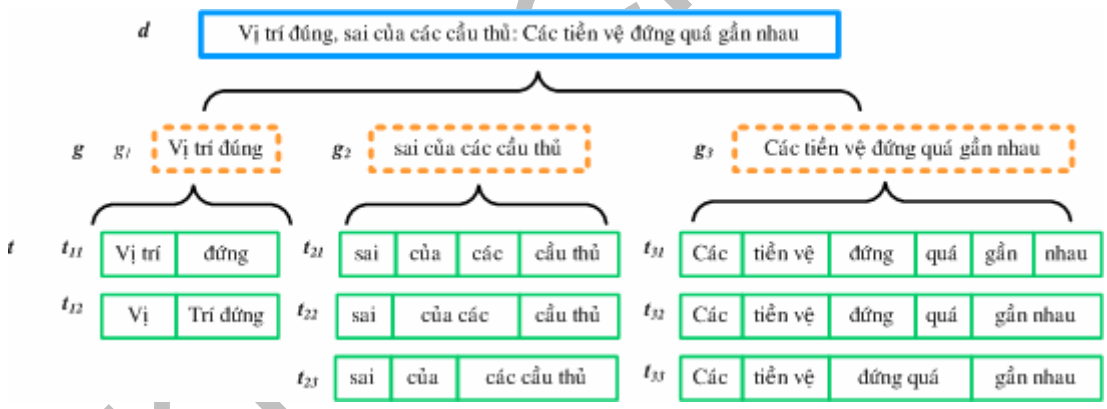
Theo kết quả đạt được của thí nghiệm so sánh giữa hai phương pháp Naïve Bayes trên, McCallum & Nigam (1998) đã đưa ra kết quả là hướng tiếp cận đa biến trạng thực hiện tốt với kích thước từ vựng nhỏ (<500 từ), còn phương pháp mô hình đa thức thường cho kết quả tốt hơn đối với kích thước từ vựng lớn (>500 từ).

5.3. Bài toán phân loại tin tức điện tử tiếng Việt

5.3.1. Quy ước

Với mỗi văn bản d , sau khi sử dụng GA để loại bỏ dấu câu và stopword, ta thu được d được tách thành nhiều ngữ g dưới dạng sau $d=\{g_1,g_2,...,g_m\}$, với g_i là tập hợp gồm n cách tách của một ngữ, $g_i = \{t_{i1},t_{i2},...,t_{in}\}$ trong đó t_{ij} là một cách tách ngữ., $t_{ij} = \{w_1,w_2,...,w_p\}$.

Ví dụ:



Hình 5. 1. Minh họa quy ước cho văn bản

Việc phân loại sẽ gán một chủ đề $c_h \in C=\{c_1,c_2,...,c_q\}$ cho văn bản, mỗi chủ đề lại bao gồm nhiều từ khóa (keyword) $K=\{k_1,...,k_r\}$. Cây phân cấp chủ đề và từ khóa thể hiện như sau :

Hình 5. 2.Minh họa chủ đề "Xã hội"

Trong phần này chúng em sẽ trình bày các phương pháp tính toán được sử dụng trong phân loại bao gồm: công thức được dùng trong IGATEC [H.Nguyen et al, 2005]và công thức Naïve Bayes [Mitchell, 2005].

5.3.2. Công thức phân loại văn bản trong IGATEC [H. Nguyen et al, 2005]

Công thức phân loại văn bản trong IGATEC [H.Nguyen et al, 2005] do chính tác giả đề nghị theo cách sử dụng độ phụ thuộc của văn bản vào chủ đề. Độ phụ thuộc này được tính dựa vào xác suất đồng xuất hiện của các từ trong văn bản với một từ khóa nhất định. Chi tiết cách tính này như sau :

Cho trước một từ khóa *k* , độ phụ thuộc của từ *w* vào *k* được tính như sau:

$$p(k \mid w) = \frac{p(k \,\&\, w)}{p(w)}$$

Trong đó

- *p(w)* là xác suất xuất hiện của từ *w* trên Google được tính theo công thức $p(w)=\frac{count(w)}{MAX}$ (đã trình bày ở mục 4.5.1.2)
- *p(k & w)* là xác suất xuất hiện đồng thời của chủ đề *k* và từ w_i trên Google với: $p(k \,\&\, w) = \frac{count(k \,\&\, w)}{MAX}$ (đã trình bày ở mục 4.5.1.2.)

Tiếp theo, độ liên quan (relative) của một cách tách ngữ *t* với từ khóa *k* bằng tổng xác suất của tất cả các từ *w* xuất hiện đồng thời với từ khóa *k* như sau:

$$rel(t,k) = \sum_{i=1}^{p} p(k \mid w_i)$$

Độ hỗ trợ (support) của cách tách ngữ *t* trên vào chủ đề $c=\{k_1,k_2,...,k_s\}$ là :

$$SP(t,c) = \frac{1}{s}\sum_{i=1}^{s} rel(t,k_i)$$

Theo công thức trên, tác giả cho rằng văn bản có độ hỗ trợ vào một chủ đề càng cao thì khả năng văn bản đó thuộc về chủ đề này càng lớn. Dựa vào các công thức, độ phụ thuộc của câu được xác định theo công thức:

$$SP(d,c) = \sum_{i=1}^{m} SP(g_i,c) = \sum_{i=1}^{m} \frac{1}{n}\sum_{j=1}^{n} SP(t_{ij},c)$$

Theo các công thức trên, văn bản *d* sẽ thuộc về chủ đề có *SP(d,c)* lớn nhất.

5.3.3. Công thức Naïve Bayes trong bài toán phân loại tin tức điện tử tiếng Việt sử dụng thống kê từ Google

Ở mục 5.2, chúng em đã trình bày các công thức Naïve Bayes cơ bản dùng thông tin xác suất học được từ tập dữ liệu huấn luyện. Tuy nhiên, hướng tiếp cận của chúng em không sử dụng tập ngữ liệu mà sử dụng thông tin thống kê từ Google nên các công thức trên được chúng em cải tiến cho phù hợp.

5.3.3.1. Ước lượng P(X|Y)

Với công thức (2.11) được trình bày ở mục 5.2. như sau:

$$P(X = w_j \mid Y = c_i) \simeq \frac{\#D\{X = w_j \wedge Y = c_i\}}{\#D\{Y = c_i\}}$$

nếu sử dụng cho tập ngữ liệu có sẵn, công thức có ý nghĩa là xác suất chủ đề c_i chứa văn văn bản có w_j bằng số văn bản có chứa w_j thuộc c_i trên tổng số văn bản thuộc chủ đề c_i. Tuy nhiên, trong hướng tiếp cận dựa trên Google, chúng ta không thể xác định được số lượng văn bản thực sự thuộc chủ đề c_i . Do đó, chúng em đề xuất cách tính xác suất khác phù hợp với hướng tiếp cận dựa trên thống kê Google:

$$P(X = w_j \mid Y = c_i) \simeq \frac{\#D\{X = w_j \wedge Y = c_i\}}{\#D\{Y = c_i\}} = \frac{p(w_j \,\&\, c_i) + 1}{\sum_k p(w_j \,\&\, c_k) + |Y|} \qquad (4.1)$$

Trong đó:

- $p(w_j \,\&\, c_i)$ là xác suất xuất hiện đồng thời w_j và c_i .
- k số thứ tự của các chủ đề, $k \in \{1,...,|Y|\}$

Công thức trên cho kết quả dựa trên xác suất xuất hiện đồng thời w_j và c_i trên tổng số lần xuất hiện số lần xuất hiện w_j trong tất cả các chủ đề.

5.3.3.2. Ước lượng P(Y)

Với công thức (2.12) được trình bày ở mục 5.2 là:

$$P(Y = c_i) = \frac{\#D\{Y = c_i\}}{\|D\|} \qquad (4.2)$$

Ở công thức này, ta giả sử các trang web chứa từ khóa c_i đều thuộc chủ đề c_i. Lúc đó, $P(Y=c_i)$ bằng xác suất xuất hiện c_i trên tổng số trang web chứa tất cả các chủ đề:

$$P(Y = c_i) = \frac{\#D\{Y = c_i\}}{\|D\|} = \frac{p(c_i)}{\sum_j p(c_j)}$$

Trong đó

- $p(c_i)$: tần số xuất hiện của chủ đề c_i trên Google
- j : là chỉ số của các chủ đề cần phân loại

5.3.3.3. Ước lượng P(Y\|X)

Khi đó công thức Naïve Bayes cho phân loại văn bản (2.9) sẽ có dạng :

$$P(Y = c_i \mid w_1, w_2, ..., w_n) = \frac{p(c_i)\prod_k p(w_k \,\&\, c_i)}{\sum_j p(c_j)\prod_k p(w_k \,\&\, c_j)} \tag{4.3}$$

Vì tần số xuất hiện $p(w)$ (mục 4.5.1) của từ trên Google rất nhỏ nên việc tính xác suất $P(Y = c_i \mid w_1, w_2, ..., w_n)$ theo công thức (4.3) có thể dẫn đến việc tràn số do nhân các số thực gần với 0. Chúng em khắc phục vấn đề này bằng cách chuyển công thức (4.3) sang sử dụng log :

$$\begin{aligned} P'(Y = c_i \mid w_1, w_2, ..., w_n) &= -\frac{\log\left(p(c_i)\prod_k p(w_k \,\&\, c_i)\right)}{\sum_j \log\left(p(c_j)\prod_k p(w_k \,\&\, c_j)\right)} \\ &= -\frac{\log\left(p(c_i)\right) + \sum_k \log\left(p(w_k \,\&\, c_i)\right)}{\sum_j \left(\log\left(p(c_j)\right) + \sum_k \log\left(p(w_k \,\&\, c_j)\right)\right)} \end{aligned}$$

Văn bản d sẽ được phân loại vào chủ đề c_i có giá trị $P'(Y = c_i \mid w_1, w_2, ..., w_n)$ cao nhất.

5.4. Kết luận

Các phương pháp phân loại văn bản dựa trên công thức của IGATEC và phương pháp Naïve đều tương đối đơn giản, không bị hạn chế về tập huấn luyện như khi sử dụng các phương pháp khác. Ngoài ra, các phương pháp trên cũng không gặp trường hợp sai lạc do có sự thay đổi trong tập huấn luyện bởi tính linh hoạt đối với sự thay đổi nhờ dùng thông tin thống kê từ Google.

Các kết quả trên thu nhận được thông qua việc chạy hệ thống thử nghiệm phân loại ViKass sẽ được mô tả chi tiết trong chương tiếp theo.

Chương 6

HỆ THỐNG THỬ NGHIỆM PHÂN LOẠI VĂN BẢN

Giới thiệu hệ thống thử nghiệm Vikass

Thử nghiệm các cách trích xuất thông tin

Dữ liệu thử nghiệm

Thử nghiệm các công thức tính độ tương hỗ MI

Thử nghiệm phân loại tin tức điện tử

Chương 6. HỆ THỐNG THỬ NGHIỆM PHÂN LOẠI VĂN BẢN

6.1. Giới thiệu hệ thống thử nghiệm Vikass

6.1.1. Chức năng hệ thống Vikass

Hệ thống thử nghiệm phân loại văn bản Vikass được xây dựng nhằm mục đích kiểm nghiệm phương pháp tách từ IGATEC và các phương pháp phân loại đề cập ở chương trước nhằm tìm ra được các tham số tối ưu trước khi tích hợp vào toà soạn báo điện tử. Các tham số này bao gồm các tham số chạy thuật toán di truyền như số lượng cá thể ban đầu, số thế hệ tối ưu, tỉ lệ lai ghép, tỉ lệ đột biến; cách tính MI hiệu quả và phương pháp phân loại nào cho kết quả tốt hơn. Ngoài tích hợp mô-đun trích tần số xuất hiện từ Google, hệ thống còn cung cấp các tính năng khác như trích tin tức, chỉnh sửa từ khóa. Chức năng của hệ thống sẽ được mô tả chi tiết trong các phần tiếp theo.

6.1.2. Tổ chức và xử lý dữ liệu

6.1.2.1. Giới thiệu chung

Hướng tiếp cận của luận văn dựa trên thống kê từ Google, điều đó có nghĩa là mỗi lần cần lấy tần số xuất hiện của một từ mới, hệ thống phải thực hiện lấy thông tin từ Internet. Điều này làm tiêu tốn rất nhiều thời gian chờ đợi, do vậy mỗi khi lấy được thông tin từ Google, chúng em lưu lại vào một file dữ liệu đệm để có thể sử dụng lại mỗi khi cần đến.

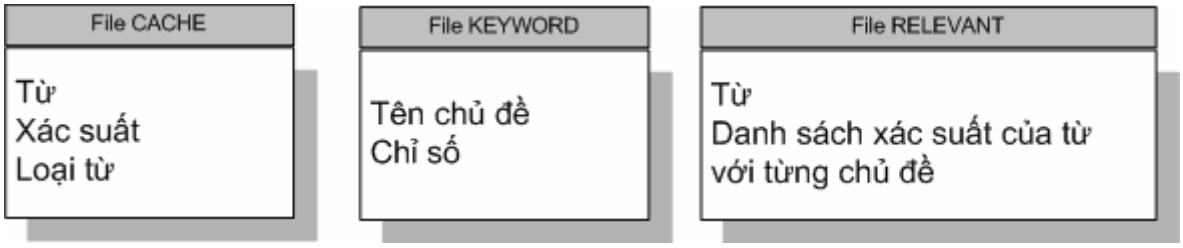
Với mục đích làm tăng tốc độ xử lý của chương trình thử nghiệm, việc quản lý dữ liệu hoàn toàn được thực hiện trên file văn bản thông thường trên kiểu phông phổ biến của tiếng Việt là phông Unicode UTF8.

Hệ thống thử nghiệm cần hai loại thông tin như sau:

- Đối với thử nghiệm tách từ tiếng Việt, hệ thống cần thông tin về xác suất xuất hiện của các từ trên Google.
- Đối với việc thử nghiệm phân loại văn bản, hệ thống cần thông tin về xác suất xuất hiện đồng thời của từ và từ khoá tương ứng với chủ đề.

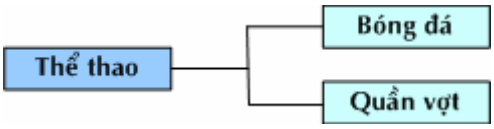
6.1.2.2. Tổ chức dữ liệu

Từ những yêu cầu trên, hệ thống dữ liệu được thiết kế thành ba file có nội dung như sau:



Hình 6. 1. Tổ chức file dữ liệu

- *File CACHE:* là dạng file văn bản thông thường, chứa thông tin:
  - ✓ Từ: từ đã tìm từ Google
  - ✓ Xác suất: xác suất của từ đó trên Google
  - ✓ Loại từ: mang một trong các giá trị W(là từ), NW (không là từ), WC (có thể là từ), NWC (không thể là từ), UD (chưa phân loại).
- *File KEYWORD:* File được viết dưới dạng xml bao gồm thông tin về tên chủ đề các cấp:
  - ✓ Tên chủ đề: tên của chủ đề các cấp (cấp 1 và cấp 2)
  - ✓ Chỉ số: chỉ số của mỗi chủ đề cho biết vị trí của chủ đề trong danh sách xác suất của từ với từng chủ đề trong file Relevant.
  - ✓ Chọn dạng xml để lưu tên chủ đề vì tính chất lồng nhau ở từng cấp của chủ đề rất thích hợp với cấu trúc dạng cây của tài liệu xml.
  - ✓ Ví dụ, ta có các chủ đề cấp 1 là "thể thao" và các chủ đề cấp 2 của nó là "Bóng đá", "Quần vợt" như hình vẽ dưới đây"



Hình 6. 2. Chủ đề Thể thao

Lúc đó, nội dung file chủ đề sẽ có nội dung như sau:

```
<?xml version="1.0" encoding="utf-8" ?>
<keyword>
  <topic name="thể thao" value="1">
      <topic name="bóng đá" value="2" />
      <topic name="quần vợt" value="3" />
  </topic>
</keyword>
```

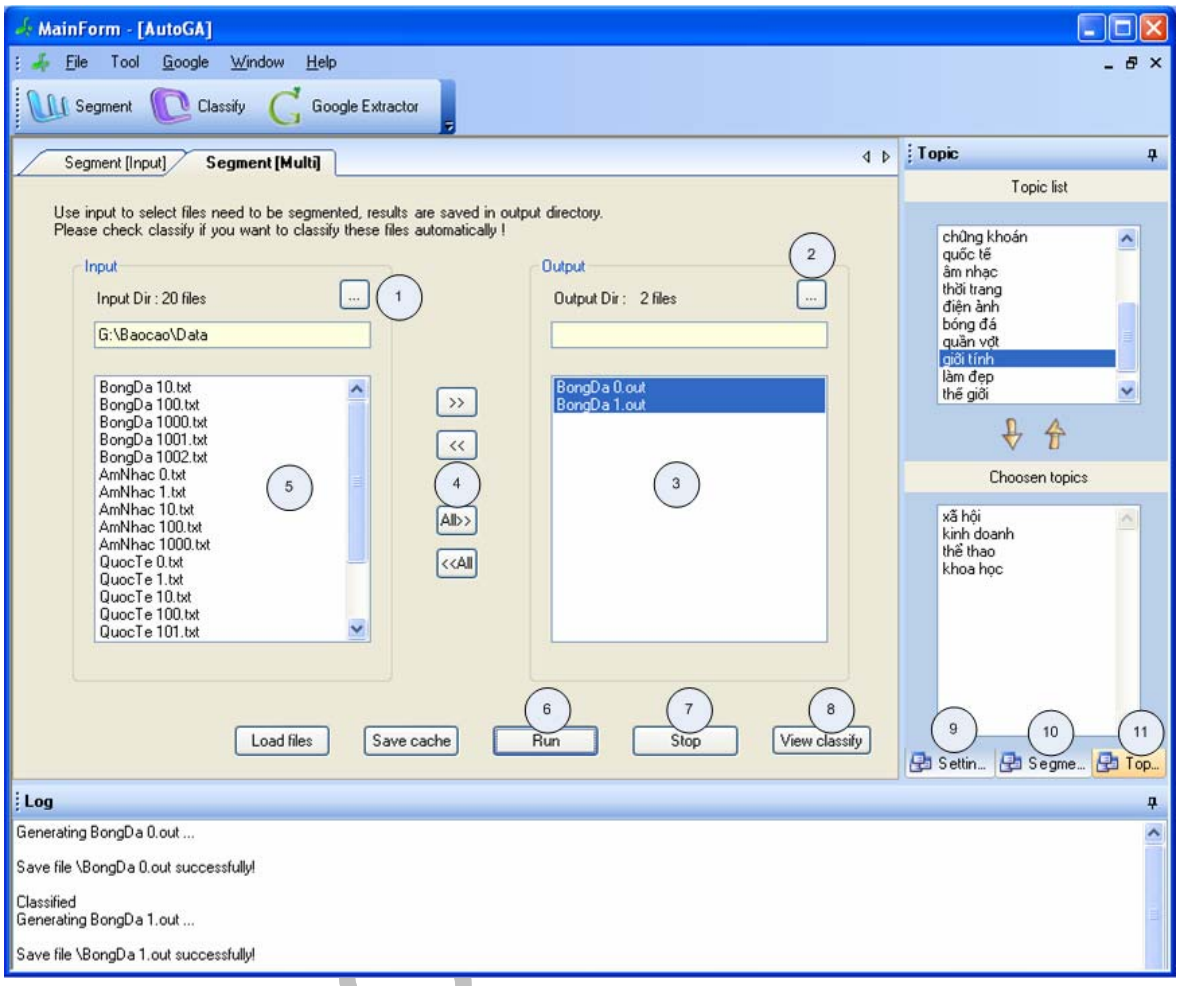
- *File RELEVANT:* chứa thông tin:
  - ✓ Từ: từ đã tìm
  - ✓ Danh sách xác suất của từ với từng chủ đề: xác suất xuất hiện đồng thời của từ ứng với từng chủ đề theo chỉ số được lưu trong file KEYWORD.

Sau khi thực hiện thử nghiệm, dung lượng file CACHE đã lên đến gần 10M và file RELEVANT xấp xỉ 50M. Với khối lượng dữ liệu lớn như vậy, việc sử dụng một hệ quản trị cơ sở dữ liệu là không cần thiết bởi vì việc xử lý thông tin trong hệ thống là đơn giản và yêu cầu tiên quyết của chương trình là tốc độ xử lý cao. Như vậy, chọn lựa lưu trữ thông tin dưới dạng văn bản bình thường là phù hợp với yêu cầu hệ thống.

6.1.2.3. Xử lý dữ liệu

Khi bắt đầu hoạt động, hệ thống tự động thực hiện đọc các file dữ liệu, phân tích chuỗi trong file để lấy thông tin và đưa vào bộ nhớ dưới dạng "bảng băm" (hashtable). Hệ thống thử nghiệm được phát triển nên ngôn ngữ C#, là một ngôn ngữ khá mạnh hỗ trợ nhiều cấu trúc lưu trữ thông tin trong đó có hỗ trợ bảng băm. Nhờ vậy mà việc tổ chức dữ liệu trở nên đơn giản hơn rất nhiều. Ngoài ra, cách xử lý như vậy sẽ làm tăng tốc độ tìm kiếm thông tin của từ nhờ các ưu điểm tổ chức dữ liệu của bảng băm.

6.1.3. Một số màn hình của hệ thống Vikass

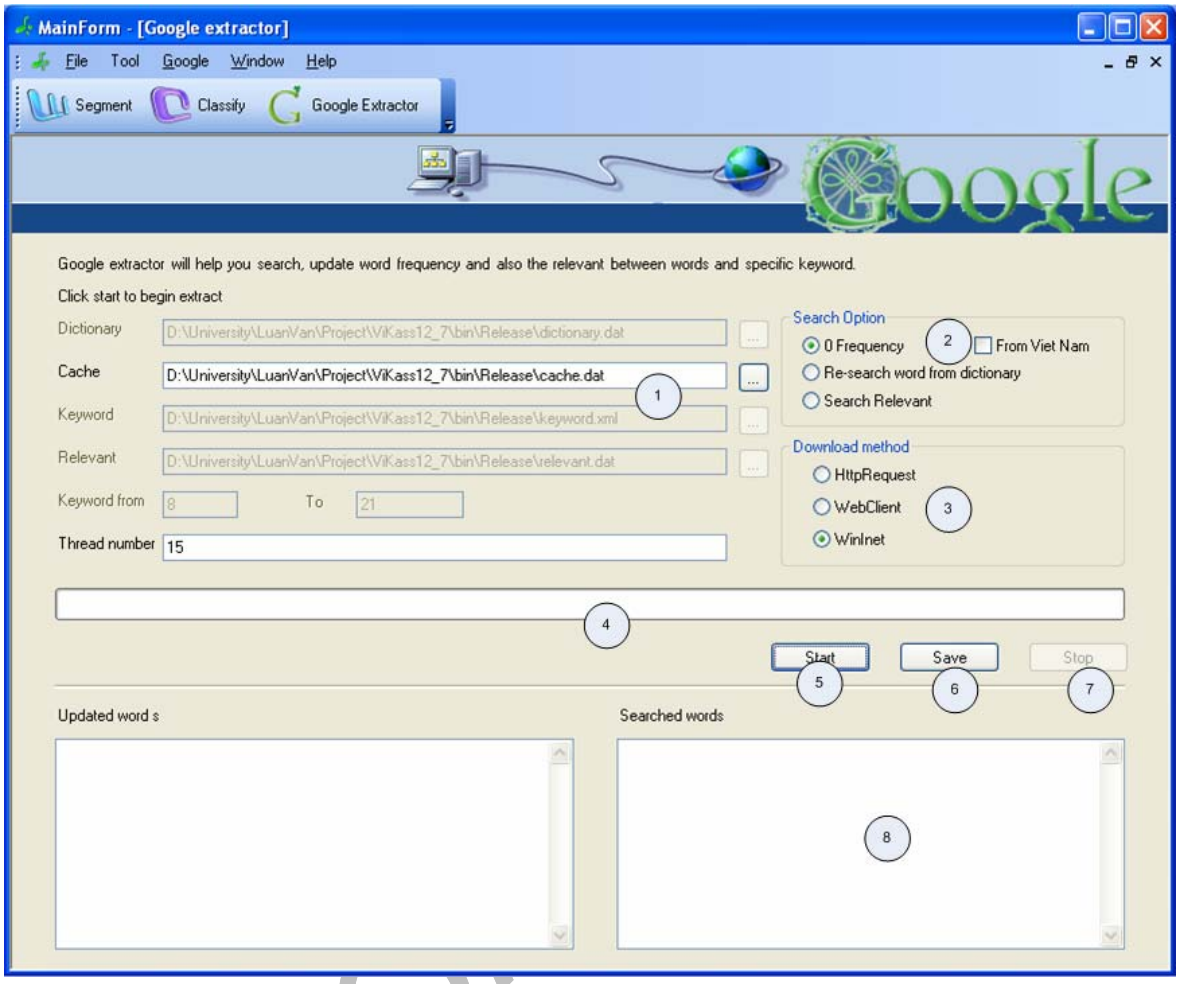


Hình 6. 3. Màn hình tách từ và phân loại

STT	Mô tả
1	Chọn thư mục chứa các tập tin cần tách từ và phân loại
2	Chọn thư mục lưu kết quả
3	Liệt kê tên các tập tin được chọn tách từ và phân loại
4	Di chuyển các tập tin qua lại để chọn các tập tin thực hiện tách từ
5	Liệt kê tên tất cả các tập tin có trong thư mục (1)
6	Thực hiện tách từ và phân loại
7	Dừng tách thực thi
8	Xem tập tin kết quả phân loại
9	Tab tùy chọn các thông số chạy GA
10	Tab tùy chọn các thông số như loại MI sử dụng, có sử dụng stopword hay không ?
11	Tab chọn các từ khóa sẽ sử dụng cho việc phân loại

Bảng 6. 1. Mô tả một số control của màn hình tách từ

Màn hình môđun trích xuất từ Google:

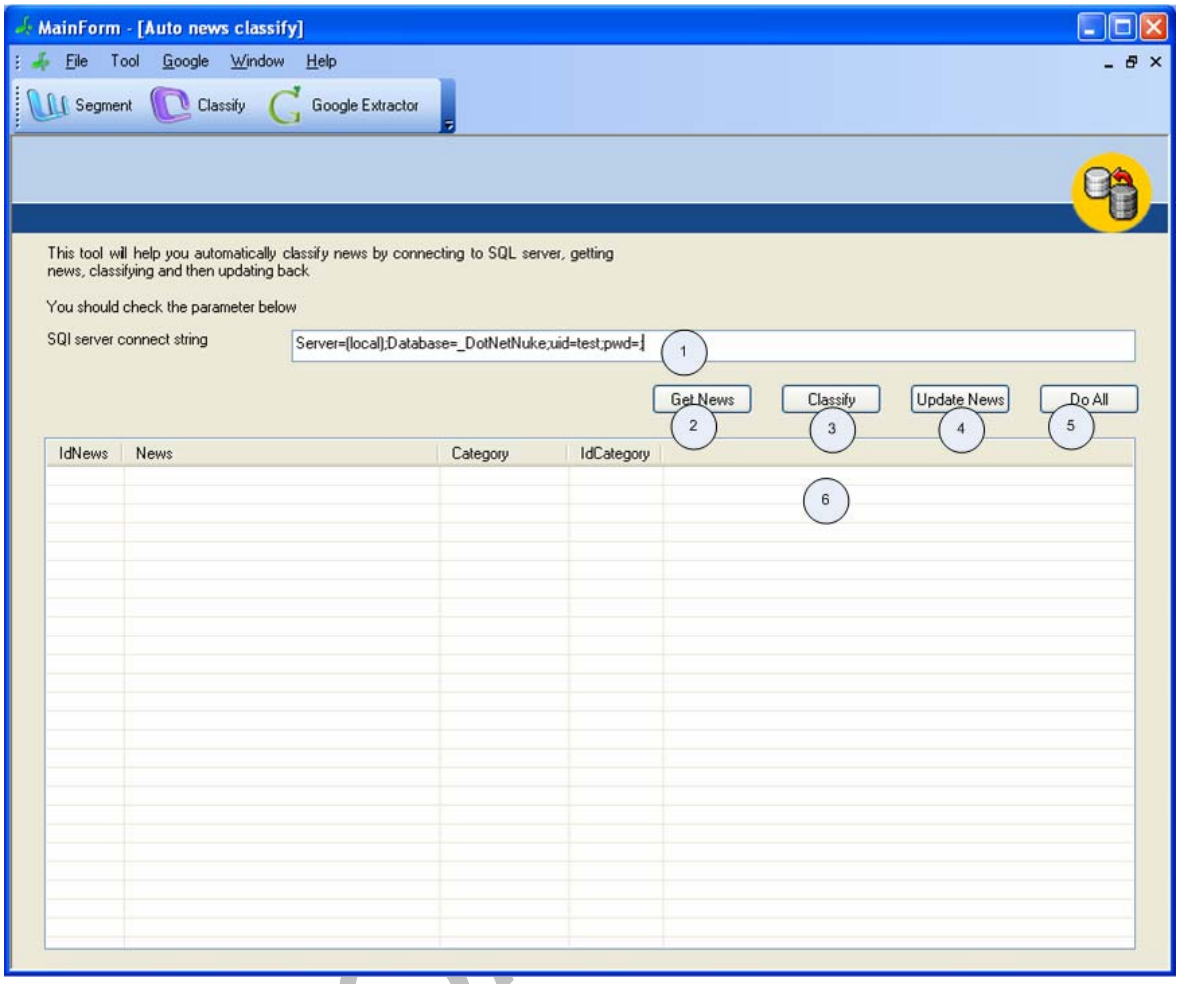


Hình 6. 4. Màn hình trích xuất từ Google

STT	Mô tả
1	Chọn thư mục chứa các tập tin như tập tin đệm, tập tin chứa độ liên quan của từ và từ khóa,…
2	Các tùy chọn như chỉ tìm kiếm các từ có tần số 0, chỉ tìm các trang .vn, tìm kiếm độ liên quan của từ và từ khóa…
3	Các phương pháp tải về sử dụng
4	Thanh biểu thị tiến trình tìm kiếm và trích từ
5	Thực hiện tìm kiếm và trích xuất
6	Lưu lại tập tin đệm và tập tin chứa độ liên quan
7	Dừng việc tìm kiếm
8	Danh sách các từ đã được tìm kiếm

Bảng 6.2. Mô tả một số control của màn hình trích từ Google

Màn hình phân loại tin tức điện tử hỗ trợ toà soạn báo điện tử :



Hình 6. 5. Màn hình phân loại tin tức điện tử

STT	Mô tả
1	Thiết lập các tham số kết nối đến SQL server
2	Lấy các tin tức được toà soạn báo điện tử tải về
3	Thực hiện phân loại
4	Cập nhật các tin tức đã được phân loại vào SQL server
5	Thực hiện tất cả các bước (2),(3),(4)
6	Hiển thị các thông tin như : nội dung tin, tên của chủ đề được phân loại,…

Bảng 6.3. Bảng mô tả một số control của màn hình phân loại tin tức điện tử

6.2. Thử nghiệm các cách trích xuất thông tin

Việc trích xuất thông tin về tần số xuất hiện của từ, độ liên quan giữa từ và chủ đề được thực hiện thông qua module Google Extractor. Nhằm mục đích tăng tốc trích thông tin từ Google, chúng em đã thử nghiệm trích thông tin bằng nhiều cách khác nhau và thực hiện kết nối đến Google sử dụng nhiều luồng (>=15). Bên cạnh đó, để tránh việc phải thực hiện tìm kiếm nhiều lần, các tập tin đệm được sử dụng với mục đích lưu lại hay cập nhất kết quả các lần tìm kiếm trước.

6.2.1. Các phương pháp thử nghiệm

Chúng em sử dụng 3 cách khác nhau để lấy kết quả tìm kiếm bao gồm sử dụng dịch vụ web do Google cung cấp, tải trang kết quả về máy cục bộ sau đó sử dụng XPath hay tìm kiếm chuỗi.

6.2.1.1. Google web service

Dịch vụ web là một ứng dụng cung cấp giao diện lập trình, hỗ trợ sự truyền thông từ ứng dụng này đến ứng dụng khác qua mạng dùng XML. Dịch vụ web của Google tại địa chỉ http://api.google.com/GoogleSearch.wsdl là một phương pháp tiện lợi để khai thác công cụ tìm kiếm này. Tuy nhiên, ta phải đăng kí tài khoản trước khi sử dụng. Với mỗi tài khoản Google giới hạn số lượng truy vấn là 1000 truy vấn/ngày. Các tham số cần biết khi sử dụng dịch vụ :

Tham số tìm kiếm	
q	Câu truy vấn
n	Số kết quả trả về trên từng trang
lr	Giới hạn phạm vi ngôn ngữ tìm kiếm
ie	Bảng mã câu truy vấn sử dụng
oe	Bảng mã của kết quả trả về

Bảng 6. 4. Tham số sử dụng dịch vụ Google

Một số câu truy vấn đặc biệt trên Google :

Truy vấn đặc biệt	Câu truy vấn	Ý nghĩa
Loại bỏ một từ	`bass –music`	**"-"** để loại bỏ 1 từ ra khỏi kết quả tìm kiếm
Từ khóa OR	`vacation london OR paris`	**OR**
Giới hạn site	`Admission site:www.stanford.edu`	**site:** chỉ tìm kiếm trong site được chỉ định
Giới hạn ngày	`Star Wars daterange:2452122-2452234`	**daterange:** chỉ trả về các file có nhãn thời gian thõa điều kiện
Lọc file	`Google filetype:doc OR filetype:pdf`	**filetype:** chỉ tìm kiếm các file có kiểu mở rộng được liệt kê
Loại trừ file	`Google doc -filetype: -filetype:pdf`	**-filetype**: ngược lại với **filetype:**
Tìm theo tiêu đề	`intitle:Google search`	**intitle:** chỉ tìm kiếm tiêu đề web

Bảng 6. 5. Một số câu truy vấn đặc biệt của Google

Trong quá trình thử nghiệm sử dụng dịch vụ web của Google, chúng em nhận thấy thời gian đáp ứng không được nhanh (khoảng >5s cho một truy vấn-sử dụng mạng Internet của trường) hơn nữa còn tồn tại nhiều lỗi. Lý do có thể kể đến như phiên bản dịch vụ đang trong quá trình thử nghiệm (bản β), hạn chế do dung lượng mạng, chi phí chứng thực. Giới hạn 1000truy vấn/ngày cũng ảnh hưởng đến chương trình khi phải thực hiện trích xuất trên lượng lớn các từ. Để khắc phục vấn đề này, chúng em sử dụng biện pháp tải trang kết quả về.

6.2.1.2. Xpath và tìm kiếm chuỗi

Trang kết quả trả về sẽ được chuyển sang định dạng xHTML dùng cho việc trích xuất dùng Xpath (http://www.w3.org/TR/XPath20) hay thực hiện tìm kiếm trên chuỗi. Cả hai phương pháp này đều cho hiệu suất tốt (khoảng 1-3s/truy vấn).

Xpath là định dạng được W3C đề nghị được sử dụng rộng rãi trong việc truy vấn tập tin XML. Sử dụng Xpath có thuận lợi hơn tìm kiếm chuỗi ở chỗ có thể sử dụng trích xuất trên nhiều ngôn ngữ trả về từ Google và nếu cấu trúc của trang web thay

đổi thì ta vẫn lấy được thông tin trả về của Google. Trong khi đó việc tìm kiếm chuỗi sẽ phụ thuộc vào các câu đặc biệt (như "các kết quả "...). Do đó, nếu các trang trả về của Google trình bày khác đi, cách tìm kiếm chuỗi sẽ không cho kết quả mong muốn. Tuy nhiên, sử dụng cách tìm kiếm chuỗi sẽ cho kết quả nhanh hơn dùng Xpath vì hệ thống không phải tốn một thời gian phân tích dữ liệu thành dạng tài liệu XML.

6.2.2. Nhận xét

Hiện tại, điều chúng ta quan tâm hàng đầu là tốc độ trích thông tin từ Google. Mặt khác, trang web Google có cấu trúc khả ổn định, hầu như không thay đổi. Vì vậy khi thực hiện thử nghiệm, chúng em sử dụng cách thức tìm kiếm chuỗi để đạt tối độ cao nhất. Tuy nhiên, chúng em vẫn xây dựng các lựa chọn rút trích để tạo tính linh hoạt trong thử nghiệm.

6.3. Dữ liệu thử nghiệm

6.3.1. Nguồn dữ liệu

Dữ liệu thử nghiệm được lấy từ trang tin tức VnExpress.net (www.vnexpress.net) tại thời điểm tháng 6/2005. Đây là một trong những trang tin tức điện tử đầu tiên tại Việt Nam ra đời vào ngày 26/2/2001, đến nay đã hơn bốn năm hoạt động với lượng độc giả đông đảo trong cả nước và quốc tế. Ngoài các trang mục do phóng viên của tờ báo viết, VnExpress.net còn mở rộng đón nhận các bài viết do độc giả gửi về từ khắp nơi để làm phong phú thêm cho nội dung của tờ báo và cập nhật tin tức thường xuyên nhanh chóng.

6.3.2. Số lượng dữ liệu thử nghiệm

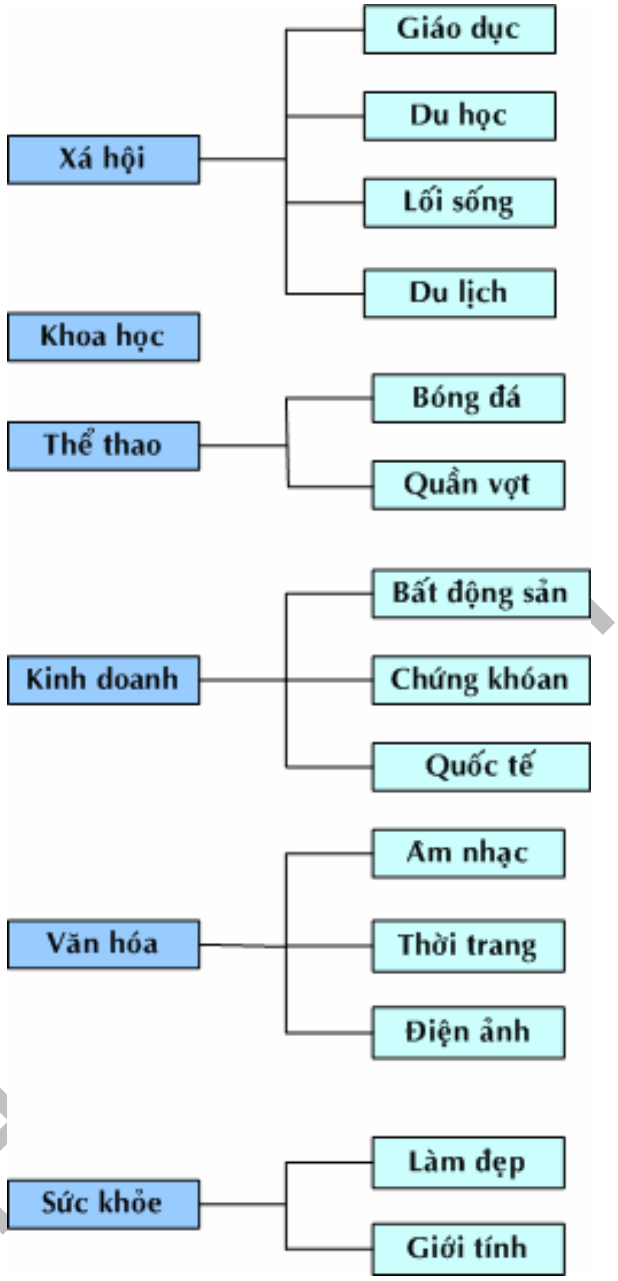
Từ các mục của VnExpress.net, đầu tiên chúng em chọn lọc ra một số mục chính để lấy dữ liệu thử nghiệm.

Vì chúng em quy định từ khóa cho chủ đề chính là tên chủ đề đó nên trong quá trình thử nghiệm, chúng em phát hiện ra một số trường hợp nhập nhằng.

Đầu tiên, từ khóa Thế giới, Xã hội có ý nghĩa bao quát có thể về Kinh tế thế giới, chính trị thế giới, văn hóa xã hội…, nên khả năng các tin tức được phân loại vào chủ đề này là rất cao do tần số xuất hiện của chủ đề này với các từ phổ biến lớn.

Thứ hai, một số mục có tên không đồng nhất giữa các tờ báo điện tử như trang VnExpress.net dùng Vi tính trong khi đó TuoiTre.com.vn lại dùng Nhịp sống số, Vnn.vn dùng Công nghệ thông tin và Viễn thông.... Việc này làm giảm kết quả khi sử dụng từ khóa khóa Vi tính cho chủ đề này vì từ khóa này không bao quát được cho các trang sử dụng tên chủ đề khác mặc dù cùng trình bày một nội dung.

Do vậy, chúng em chỉ sử dụng một số mục có từ khóa rõ ràng. Đối với mỗi tin tức, chúng em chỉ tách lấy phần tiêu đề, phần tóm lược và phần chú thích ảnh. Đây là các phần có ý nghĩa phân loại cao do được người viết bài tóm lược và chọn lọc. Ứng mỗi chủ đề, chúng em lấy ngẫu nhiên 100 tin. Còn cách giải quyết phần nhập nhằng trình bày ở trên sẽ là hướng mở rộng của luận văn. Tổng dữ liệu thử nghiệm là 1500 tập tin bao gồm 15 chủ đề cấp 2, mỗi chủ đề 100 tập tin.



Hình 6. 6. Cây chủ đề

6.3.3. Nhận xét

Mặc dù dữ liệu dùng thử nghiệm khá nhỏ do hạn chế về mặt thời gian, nhưng cách thức chọn dữ liệu và chủ đề thử nghiệm phân loại của chúng em đã mở rộng rất nhiều so với 35 văn bản thử nghiệm của [H. Nguyen et al, 2005] trên 5 chủ đề Chính trị, Giáo dục, Kinh doanh, Sức khỏe, Thể thao.

6.4. Thử nghiệm các công thức tính độ tương hỗ MI

6.4.1. Các phương pháp thử nghiệm

Nhằm xác định hiệu quả của các cách tính MI trong việc tách từ tiếng Việt, chúng em thực hiện thử nghiệm 3 công thức MI đã được trình bày ở mục 4.5: một công thức tính MI của [H.Nguyen et al, 2005] (gọi là MI1) , một của [Ong & Chen, 1999] (gọi là MI2), một do chúng em đề nghị (gọi là MI3) . Ứng với mỗi công thức tính MI trên, chúng em thử nghiệm thêm việc tách stopword và không tách stopword trước khi tách từ. Mục đích của việc tách stopword trước khi tách từ nhằm tạo ra nhiều ngữ nhỏ hơn khi đã bỏ các từ không có ý nghĩa, để làm tăng tốc độ tách từ của hệ thống.

Như vậy, tổng cộng có 6 thử nghiệm tách từ như sau:

- MI1 tách stop word (MI1_NonSW)
- MI1 không tách stop word (MI1_SW)
- MI2 tách stop word (MI2_NonSW)
- MI2 không tách stop word (MI2_NonSW)
- MI3 tách stop word (MI3_NonSW)
- MI3 không tách stop word (MI3_NonSW)

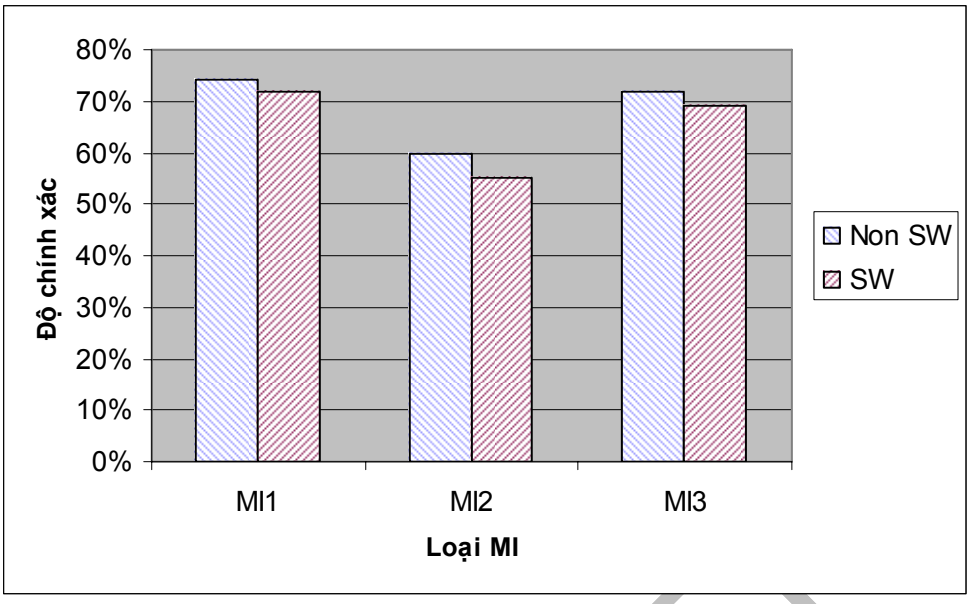
Chúng em thử nghiệm các công thức trên 1500 nội dung tóm tắt các tin tức của VnExpress.net

6.4.2. Kết quả

Độ chính xác của các công thức tính độ tương hỗ như sau:

Cách tính MI	Không tách stop word	Có tách stopword
MI 1 [H. Nguyen et al, 2005]	74%	72%
MI 2 [Ong & Chen, 1999]	60%	55%
MI 3 (chúng em đề nghị)	72%	69%

Bảng 6. 6. Kết quả thực nghiệm các công thức tính độ tương hỗ MI



Hình 6. 7. Biểu đồ so sánh kết quả các công thức tính độ tương hỗ MI

6.4.3. Nhận xét

Trong 6 cách thử nghiệm, cách tách từ dùng công thức MI1. có độ chính xác cao nhất.

Thời gian chạy tách từ lúc đầu khá lâu (trung bình khoảng 10 phút cho một mẫu tóm tắt dài khoảng 100 tiếng) đa phần là do thời gian lấy thông tin từ Google. Nhưng khi thông tin về tần số xuất hiện của các từ đã được lưu lại tương đối lớn (độ lớn file cache khoảng 10M), thì tốc độ tách từ giảm xuống đáng kể (trung bình <1giây đối với các văn bản không cần lấy thông tin từ Internet)

Cách tiếp cận của công thức MI1 là ưu tiên dựa trên từ ghép có hai tiếng, mà theo thống kê dựa trên từ điển của chúng em, số từ 2 tiếng chiếm đa số trong từ vựng tiếng Việt. Cách tính này cho kết quả khá tốt vì vừa thoả mãn được tính chất tự nhiên dựa trên ưu thế áp đảo của từ 2 tiếng, vừa được chứng minh bằng thực nghiệm.

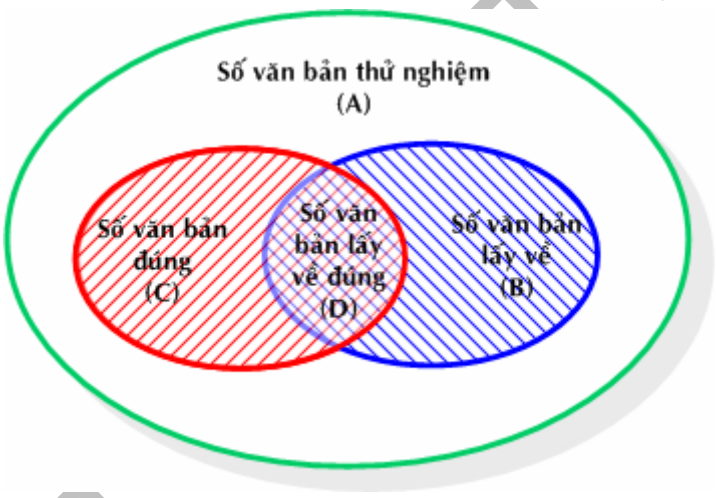
Trong các trường hợp thử nghiệm có tách stopword, thời gian tách từ giảm đi rất nhiều (trung bình 5 phút cho văn bản mới). Tuy nhiên, trong quá trình thử nghiệm, chúng em nhận thấy việc tách stopword có thể làm sai lạc ý nghĩa của văn bản ban

đầu do danh sách stopword đưa vào không hoàn chỉnh. Vì vậy kết quả tách từ có tách stopword không cao như cách tách thuần tuý.

6.5. Thử nghiệm phân loại tin tức điện tử

6.5.1. Thước đo kết quả phân loại văn bản

Để đánh giá hiệu quả phân loại văn bản, thông thường người ta dùng các chỉ số về độ thu về-recall và độ chính xác-precision [Yang, 2000]. Cho một phương pháp phân loại văn bản, đầu vào là một văn bản, và kết quả trả về là một danh sách các chủ đề được gán cho văn bản đó, chỉ số độ thu về, độ chính xác có thể được tính như sau:



Hình 6. 8. Các thông số dùng tính độ thu về, độ chính xác

Hình trên mô tả các thông số sau:

- (A) là tất cả văn bản thực hiện phân loại văn bản cho chủ đề T
- (B) là số văn bản được phân loại lấy về cho chủ đề T
- (C) là số văn bản thực sự thuộc về chủ đề T
- (D) là số văn bản lấy về chính xác.

Các tham số trên được dùng trong công thức tính độ thu về-recall, độ chính xác-precision dưới đây:

$$\text{recall} = \frac{\text{Số văn bản lấy về đúng (D)}}{\text{Số văn bản đúng (C)}}$$

$$\text{precision} = \frac{\text{Số văn bản lấy về đúng (D)}}{\text{Số văn bản lấy về (B)}}$$

Việc gán nhãn chủ đề của các phương pháp phân loại văn bản có thể được đánh giá bằng cách dùng bảng trường hợp hai chiều ứng với từng loại chủ đề:

	Chủ đề đang xét ĐÚNG với chủ đề văn bản	Chủ đề đang xét SAI với chủ đề văn bản
Phân loại ĐÚNG với chủ đề văn bản	**a**	**b**
Phân loại SAI với chủ đề văn bản	**c**	**d**

Bảng 6. 7. Bốn trường hợp của phân loại văn bản

Như vậy, với mỗi kết quả phân loại cho một văn bản, ta sẽ có được một trong 4 trường hợp a,b,c hoặc d. Từ đó, ta tính được các chỉ số sau:

- $recall = \frac{a}{a+c}$ nếu a + c >0, ngược lại là không xác định.
- $precision = \frac{a}{a+b}$ nếu a + b >0, ngược lại là không xác định.
- Tuy nhiên, cách tính với độ thu về, độ chính xác riêng rẽ sẽ cho kết quả không cân đối. Ví dụ nếu số văn bản lấy về đúng (D) gần bằng với số văn bản đúng thực sự (C) thì chỉ số độ thu về sẽ cao, tuy nhiên nếu số văn bản lấy về (B) khá nhiều so với (D) sẽ cho chỉ số độ chính xác nhỏ. Do vậy, thông thường người ta thêm một chỉ số F1 [Yang , 1997] để phản ánh sự cân đối giữa 2 độ đo trên:

$$F1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

Ngoài ra, để tính toán hiệu quả thực thi trên toàn bộ chủ đề, thông thường người ta còn sử dụng hai phương pháp *macro-averaging* và *micro-averaging*.

Macro-averaging tính trung bình các chỉ số recall, precision, fallout, Acc,Err của tất cả các chủ đề.

Micro-averaging tính toán các chỉ số dựa trên tổng giá trị a, b, c, d của từng chủ đề dựa theo các công thức áp dụng tính cho một chủ đề.

Sự khác nhau chủ yếu giữa hai cách tính *macro-averaging* và *micro-averaging* là : *micro-averaging* tính toán dựa trên trọng số của mỗi văn bản, nên cho kết quả trung bình trên mỗi văn bản (per-document average); trong khi đó, *macro-averaging* tính toán trọng số trên mỗi chủ đề, do đó, kết quả cho sẽ đại diện cho giá trị trung bình trên mỗi chủ đề (per-category average).

6.5.2. Các phương pháp thử nghiệm

Ở phần phân loại văn bản, chúng em thử nghiệm 2 công thức đã được trình bày ở 5.3. là công thức phân loại được sử dụng trong [H. Nguyen et al, 2005] (gọi tắt là công thức IClass) và công thức tính Naïve Bayes được cải tiến cho phù hợp với hướng tiếp cận dựa trên Google (gọi tắt là NBClass).

Ứng với công thức phân loại, chúng em thử nghiệm với 2 công thức tính MI: một của [H. Nguyen et al, 2005] (gọi tắt là MI1) và một công thức MI do chúng em đề xuất (gọi tắt là MI3) cho hai trường hợp tách và không tách stopword.Ở phần này chúng em không thử nghiệm với MI2 của [Ong & Chen, 1999] vì kết quả tách từ của công thức này thấp hơn các công thức khác khá nhiều sẽ cho kết quả không tốt.

Như vậy tổng cộng chúng em thực hiện 8 lần thử nghiệm phân loại như sau:

- Công thức IClass + MI1 + tách stop word
- Công thức IClass + MI1 + không tách stop word
- Công thức IClass + MI3 + tách stop word
- Công thức IClass + MI3 + không tách stop word
- Công thức NBClass + MI1 + tách stop word
- Công thức NBClass + MI1 + không tách stop word
- Công thức NBClass + MI3 + tách stop word
- Công thức NBClass + MI3 + không tách stop word

6.5.3. Kết quả

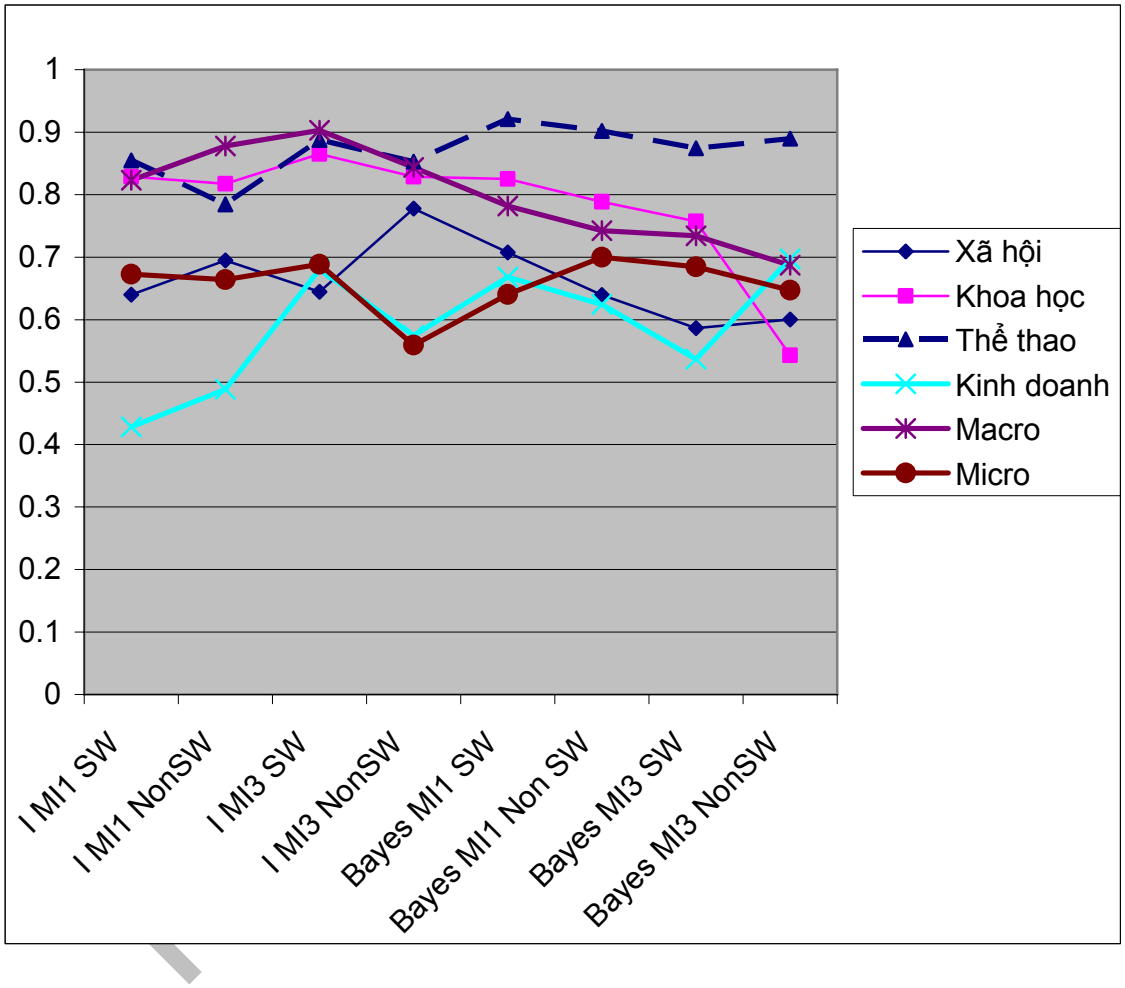
Sau khi thực hiện phân loại văn bản, chúng em sử dụng các độ đo đã được trình bày ở mục 6.5.1. để tính toán kết quả chính xác của các thử nghiệm phân loại. Kết quả tính toán được trình bày trong bảng thống kê sau:

Phương pháp	Tên chủ đề	R	P	F1
IClass + MI 1 +tách stopword	Xã hội	0.62625	0.654047	0.639847
	Khoa học	0.72	0.975434	0.828475
	Thể thao	0.765	0.968245	0.854706
	Kinh doanh	0.795	0.293358	0.428571
	Macro	0.763437	0.892427	0.822908
	Micro	0.663	0.682801	0.672755
IClass + MI 1 +không tách stopword	Xã hội	0.764	0.636667	0.694545
	Khoa học	0.7216	0.942131	0.81725
	Thể thao	0.65625	0.975	0.784483
	Kinh doanh	0.816	0.348718	0.488623
	Macro	0.814333	0.951923	0.877769
	Micro	0.656	0.672131	0.663968
IClass + MI 3 +tách stopword	Xã hội	0.630	0.660	0.645
	Khoa học	0.857	0.873	0.865
	Thể thao	0.861	0.915	0.887
	Kinh doanh	0.630	0.740	0.681
	Macro	0.913	0.892	0.903
	Micro	0.678	0.700	0.689
IClass + MI 3	Xã hội	0.772	0.784	0.778
	Khoa học	0.808	0.851	0.829

+không tách stopword	Thể thao	0.882	0.825	0.853
	Kinh doanh	0.637	0.523	0.575
	Macro	0.858	0.830	0.844
	Micro	0.553	0.566	0.559
NBClass + MI 1 +tách stopword	Xã hội	0.680	0.738	0.708
	Khoa học	0.810	0.841	0.825
	Thể thao	0.924	0.918	0.921
	Kinh doanh	0.725	0.620	0.668
	Macro	0.785	0.779	0.782
	Micro	0.648	0.633	0.640
NBClass + MI 1 +không tách stopword	Xã hội	0.591	0.697	0.640
	Khoa học	0.704	0.897	0.789
	Thể thao	0.886	0.918	0.902
	Kinh doanh	0.675	0.581	0.625
	Macro	0.714	0.773	0.742
	Micro	0.783	0.633	0.700
NBClass + MI 3 +tách stopword	Xã hội	0.544	0.636	0.586
	Khoa học	0.680	0.855	0.757
	Thể thao	0.708	1.142	0.874
	Kinh doanh	1.404	0.332	0.537
	Macro	0.748	0.721	0.734
	Micro	0.725	0.648	0.684
NBClass + MI 3	Xã hội	0.611	0.590	0.600
	Khoa học	0.485	0.616	0.543
	Thể thao	0.749	1.095	0.890

+không tách stopword	Kinh doanh	0.660	0.739	0.697
	Macro	0.626	0.760	0.687
	Micro	0.647	0.647	0.647

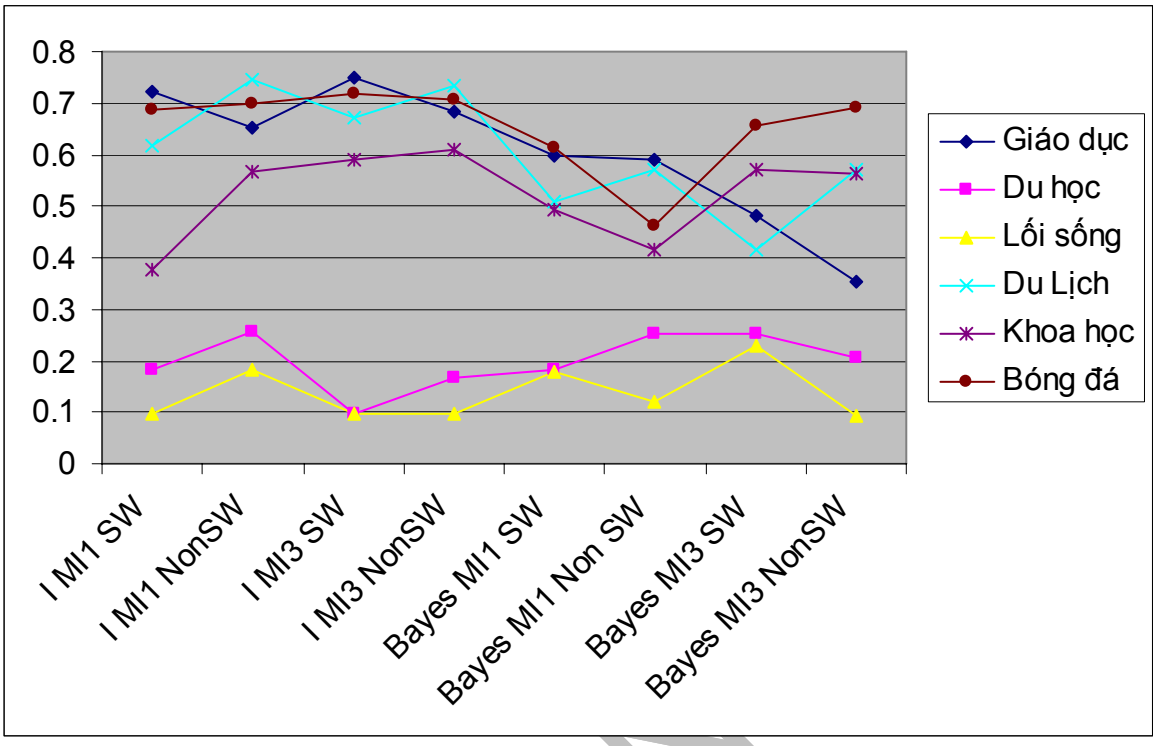
Bảng 6. 8. Kết quả phân loại văn bản cho từng chủ đề ở cấp 1

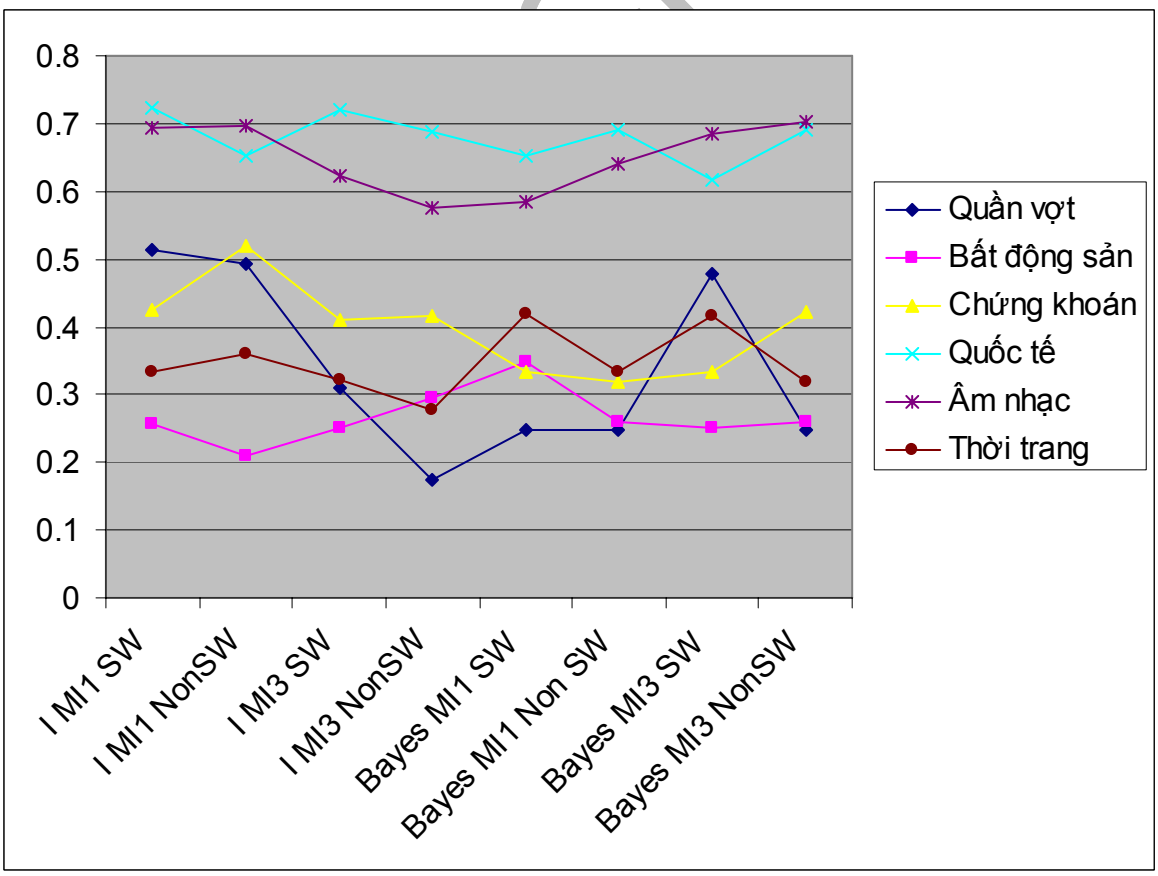


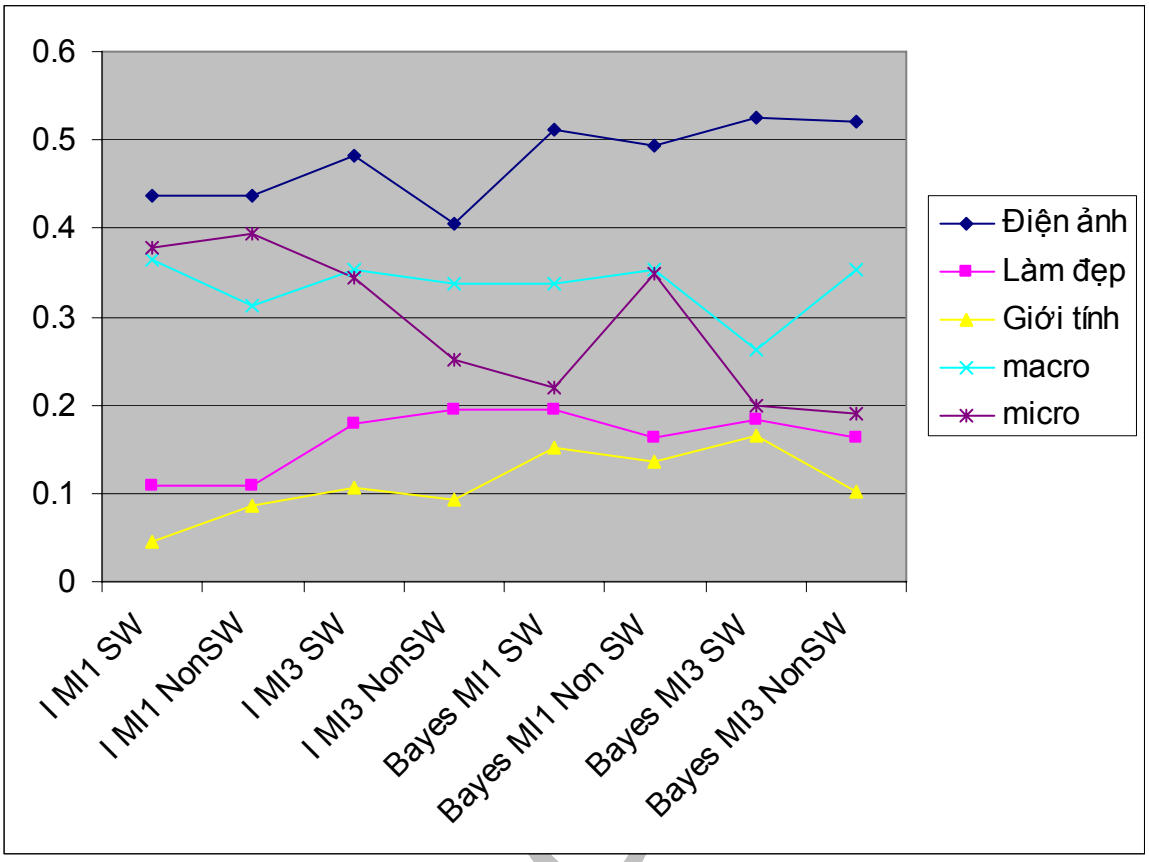
Hình 6. 9. Biểu đồ F1 cho cấp 1

Vì kết quả của phần thử nghiệm phân loại ở cấp hai rất dài, nên chúng em chỉ xin trình bày biểu đồ kết quả phân loại mà không trình bày chi tiết bảng kết quả cho từng chủ đề.

Sau đây là kết quả phân loại cho các chủ đề cấp 2.


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
I MI1 SW
I MI1 NonSW
I MI3 SW
I MI3 NonSW
Bayes MI1 SW
Bayes MI1 Non SW
Bayes MI3 SW
Bayes MI3 NonSW
Giáo dục
Du học
Lối sống
Du Lịch
Khoa học
Bóng đá


0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
I MI1 SW
I MI1 NonSW
I MI3 SW
I MI3 NonSW
Bayes MI1 SW
Bayes MI1 Non SW
Bayes MI3 SW
Bayes MI3 NonSW
Quần vợt
Bất động sản
Chứng khoán
Quốc tế
Âm nhạc
Thời trang



Hình 6. 10. Biểu đồ F1 cho cấp 2

6.5.4. Nhận xét

Trong hai mức phân loại chủ đề, ta nhận thấy kết quả phân loại ở mức 1 cho độ chính xác cao hơn mức 2. Lý do là vì số lượng chủ đề của cấp 2 nhiều hơn cấp 1 rất nhiều (15 so với 4 ở cấp 1) và một số chủ đề của cấp 2 chưa thực sự tốt như Bất động sản, Lối sống, Làm đẹp, Giới tính. Từ đó, ta thấy được việc xây dựng danh sách từ khoá cho mỗi chủ đề một yêu cầu cần thiết để nâng hiệu suất phân loại văn bản.

Dựa vào kết quả thử nghiệm ta nhận thấy rằng trong việc phân loại sử dụng Bayes tốt hơn công thức phân loại của H. Nguyen et al (2005) trong nhiều trường hợp. Trong các thử nghiệm công thức của H.Nguyen et al (2005), độ hỗ trợ của kết quả vào chủ đề đối có giá trị rất gần nhau, khi áp dụng cho các chủ đề hầu như không có sự khác biệt. Trong khi đó, với công thức Naïve Bayes, có một số chủ đề

nổi trội hơn hẳn các chủ đề khác và kết quả thống kê cũng cho thấy Naïve Bayes cho kết quả chính xác hơn.

Kết quả của thử nghiệm công thức trong [H.Nguyen et al, 2005] với độ chính xác chưa cao lắm bởi vì đấy là công thức do chính tác giả đề nghị chưa dựa trên cơ sở lý thuyết vững chắc. Trong khi đó, phương pháp Naïve Bayes đã xuất hiện khá lâu, được chứng minh trên lý thuyết và thực nghiệm nên độ tin cậy rất cao. Việc sử dụng hướng tiếp cận Naïve Bayes cho phân loại văn bản dựa trên Google có thể nói là bước cải tiến đáng khích lệ so với cách phân loại cũ.

Dựa vào biểu đồ, ta nhận thấy sự kết hợp giữa phương pháp phân loại Naïve Bayes và công thức tính độ tương hỗ (MI) của [H. Nguyen et al, 2005] cho kết quả phân loại tốt nhất. Trong đó, tỉ lệ trung bình của phương pháp cho các chủ đề ở cấp 1 là 75%, và cho các chủ đề ở cấp 2 là 67%. Kết quả này hợp lý vì thực nghiệm cho thấy công thức MI1 của H.Nguyen et al (2005) cho kết quả tách từ chính xác cao nhất nên đã góp phần làm cho kết quả phân loại tốt hơn.

Kết quả phân loại văn bản trung bình giữa 8 cặp là 75%, là kết quả chấp nhận được đối với phân loại văn bản tiếng Việt. Kết quả không cao so với kết quả phân loại bằng tiếng Anh bởi vì như chúng ta đã biết phần tách từ tiếng Việt gặp rất nhiều phức tạp.

Chương 7

ỨNG DỤNG PHÂN LOẠI TIN TỨC ĐIỆN TỬ TỰ ĐỘNG

Giới thiệu tòa soạn báo điện tử

Tính cần thiết của phân loại tin tức tự động

Phân tích hiện trạng

Mô hình DFD quan niệm cấp 2 hiện hành cho ô xử lý Nhận bài và Trả bài

Phê phán hiện trạng

Mô hình DFD quan niệm cấp 2 mới cho ô xử lý Nhận bài và Trả bài

Triển khai DLL

Chương trình cài đặt "Tòa soạn báo điện tử" đã tích hợp module phân loại tin tức

Kết quả

Chương 7. ỨNG DỤNG PHÂN LOẠI TIN TỨC ĐIỆN TỬ TỰ ĐỘNG

Nhằm đánh giá hiệu quả thực tế của việc phân loại sử dụng IGATEC và Naïve Bayes, chúng em đã xây dựng công cụ phân loại thành một module đồng thời tích hợp vào trong tòa soạn báo điện tử. Trong chương này, chúng em sẽ giới thiệu sơ lược về tòa soạn báo điện tử và mô tả cách thức tích hợp module phân loại.

7.1. Giới thiệu tòa soạn báo điện tử

Phần mềm tòa soạn báo điện tử (Luận văn khóa 2000-Hoàng Minh Ngọc và Nguyễn Duy Hiệp) xây dựng trên nền tảng DotNetNuke tuân thủ theo qui trình của một tòa soạn thực tế đi từ soạn bài, duyệt bài và đăng bài. Mỗi biên tập viên sẽ phụ trách một mảng chủ đề. Cộng tác viên hay người dùng sau khi viết bài phải được biên tập viên duyệt. Nếu nội dung và hình thức chấp nhận được thì bài được chuyển lên vị trí có chức năng đưa bài lên website chính thức. Người quản trị sẽ phân công chuyên mục và chủ đề cho các biên tập viên. Nếu đã qua các cấp kiểm duyệt, bài viết được phép đưa lên website. Nếu tại một cấp nào đó, người quản lý thấy bài viết cần được chỉnh sửa thì bài viết sẽ được trả về đúng cấp có thẩm quyền.

Ngoài ra, tòa soạn báo điên tử còn hỗ trợ việc thu thập tin tức điện tử từ nhiều nguồn khác nhau. Tin tức được tải về sau đó phải được các biên tập viên xác định chủ đề và chuyên mục mà bài báo thuộc về để tiến hành thủ tục đăng bài. Việc phân loại tin tức ở giai đoạn thực hiện luận văn này là hoàn toàn thủ công.

7.2. Tính cần thiết của phân loại tin tức tự động

Việc thực hiện phân loại thủ công trên số lương lớn các tin tức được tải về có thể ngốn rất nhiều thời gian và công sức. Nhằm làm tăng tính hiệu quả cũng như hỗ trợ tối đa cho các biên tập viên tập trung vào các công việc khác quan trọng hơn. Module phân loại tin tức tự động đã được xây dựng. Nhiệm vụ của module này là thực hiện phân loại tự động các tin tức tải về nhằm đề xuất sắp xếp tin tức này vào một chuyên mục hợp lý. Module được viết dưới dạng một thư viện dll thực hiện các

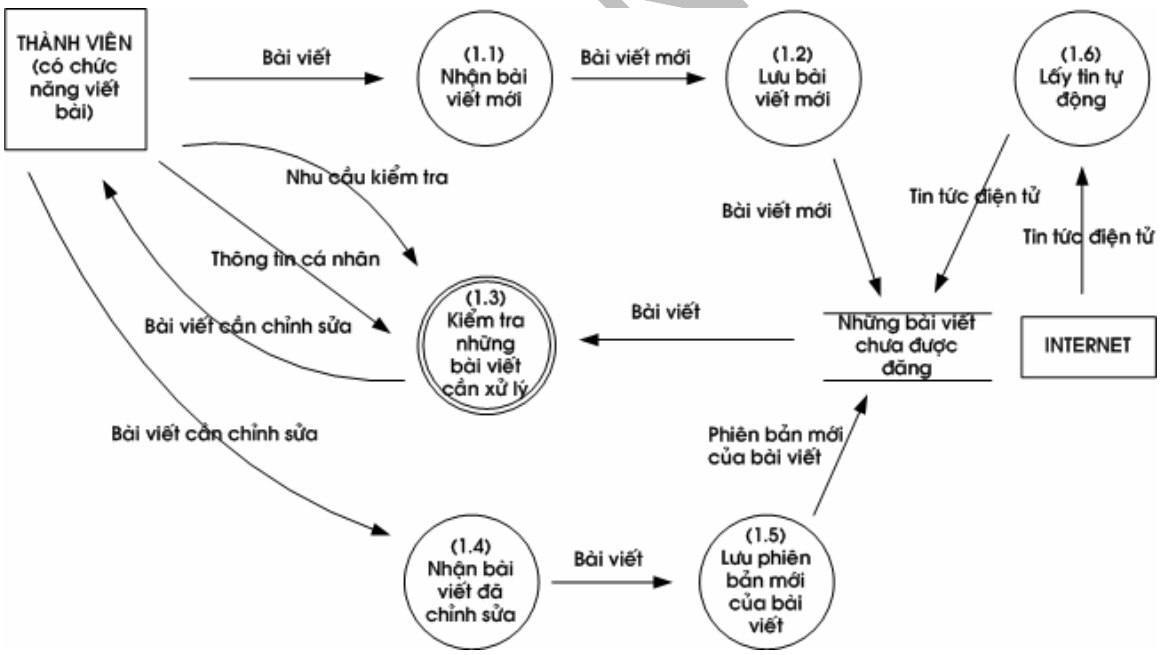
công việc như sau: lấy các tin tức được tải về, tiến hành phân loại và cập nhật vào cơ sở dữ liệu.

7.3. Phân tích hiện trạng

Mục đích của luận văn chúng em là tích hợp phần xử lý phân loại trang web tự động vào phần duyệt bài viết và sửa bài viết nên chúng em chỉ trình bày mô hình DFD cho ô xử lý "Nhận bài và Trả bài". Để tìm hiểu về toàn cảnh mô hình DFD của toà soạn báo điện tử, xin tham khảo luận văn "Toà soạn báo điện tử" của Hoàng Minh Ngọc Hải (0012545), Nguyễn Duy Hiệp (0012038))

7.3.1. Mô hình DFD quan niệm cấp 2 hiện hành cho ô xử lý Nhận bài và Trả bài

7.3.1.1. Mô hình



Hình 7. 1.Mô hình DFD hiện hành

7.3.1.2. Mô tả mô hình

Thành viên có chức năng viết bài nhận bài viết mới được giao, sau khi hoàn thành thì lưu xuống kho dữ liệu những bài viết chưa đăng để chờ duyệt. Sau khi bài viết được duyệt, thành viên kiểm tra xem bài viết có cần chỉnh sửa không, nếu có thì

thực hiện chỉnh sửa sau đó lưu phiên bản mới của bài viết chờ duyệt tiếp. Ngoài ra, các bài báo được lấy tự động từ Internet xuống cũng được lưu trong kho dữ liệu các bài viết chưa đăng để chờ duyệt.

7.3.1.2.1. Mô tả kho dữ liệu

Hệ thống thông tin: Xây dựng toà soạn báo điện tử	**Mô hình quan niệm xử lý** Hiện tại [] Tương lai[]	Trang :
Ứng dụng : Xây dựng toà soạn báo điện tử	***Mô tả kho dữ liệu : NHỮNG BÀI VIẾT CHƯA ĐƯỢC ĐĂNG*** Tờ :	Ngày lập : 28/6/2004 Người lập : 1. Hoàng Minh Ngọc Hải 2. Nguyễn Duy Hiệp

Dòng dữ liệu vào :

Bài viết đã chỉnh sửa

Bài viết mới

Dòng dữ liệu ra :

Bài viết cần chỉnh sửa

Diễn giải :

Kho này lưu trữ những bài viết đang nằm trong dây chuyền

Cấu trúc dữ liệu:

MA_BAI_VIET

MA_CHUYEN_MUC

MA_TAC_GIA

NGAY_VIET

TIEU_DE

NOI_DUNG

DUONG_DAN_ANH

KICH_THUOC_ANH

CHIEU_DAI

CHIEU_RONG

Khối lượng :

- Hiện tại : Không xác định
- Tương lai : Không xác định

Thông tin thường truy xuất :

MA_BAI_VIET

MA_CHUYEN_MUC

TIEU_DE

NOI_DUNG

Bảng 7. 1. Bảng kho dữ liệu những bài viết chưa được đăng

7.3.1.2.2. Mô tả ô xử lý

Ô xử lý	Tên	Dòng dữ liệu vào	Dòng dữ liệu ra	Diễn giải
(1.1)	Nhận bài viết mới	Bài viết	Bài viết mới	Phóng viên sau khi viết một bài mới sẽ gửi vào hệ thống. Những bài viết này được lưu dưới dạng những bài viết chưa được xử lý.
(1.2)	Lưu bài viết mới	Bài viết mới	Bài viết mới	Lưu bài viết dưới tình trạng "Chưa xử lý"

(1.3)	Kiểm tra những bài viết cần xử lý	Nhu cầu kiểm tra Thông tin cá nhân	Bài viết cần chỉnh sửa	Kiểm tra các bài viết đã được duyệt xem có cần chỉnh sửa không
(1.4)	Nhận bài viết đã chỉnh sửa	Bài viết đã chỉnh sửa	Bài viết đã chỉnh sửa	Bài viết sau khi thành viên (có chức năng chỉnh sửa) duyệt, chỉnh sửa và trả lại cho thành viên phụ trách bài viết đó.
(1.5)	Lưu phiên bản mới của bài viết	Bài viết đã chỉnh sửa	Bài viết đã chỉnh sửa	Bài viết đã chỉnh sửa được lưu vào CSDL dưới tình trạng "Đã xử lý" tại cấp vừa chỉnh sửa và dưới tình trạng "Chưa xử lý" tại cấp được chuyển bài về
(1.6)	Lấy tin tự động	Tin tức điện tử	Tin tức điện tử	Hệ thống tự động lấy tin tức từ các trang báo khác và lưu xuống kho dữ liệu

Bảng 7. 2. Bảng mô tả các ô xử lý của mô hình DFD hiện hành

7.3.2. Phê phán hiện trạng

Hiện tại, hệ thống tự động lấy tin tức từ các trang báo điện tử khác về và gán vào các mục đã được chỉ định sẵn. Tuy nhiên, việc chỉ định chủ đề cho các tin tức lấy về một cách cứng nhắc chỉ đúng trong trường hợp trang web lấy tin có cấu trúc chủ đề tương ứng với chủ đề trong toà soạn báo điện tử của mình. Đối với những trang báo có cấu trúc khác đi, việc gán nhãn mặc định cho các bài báo sẽ không còn đúng nữa.

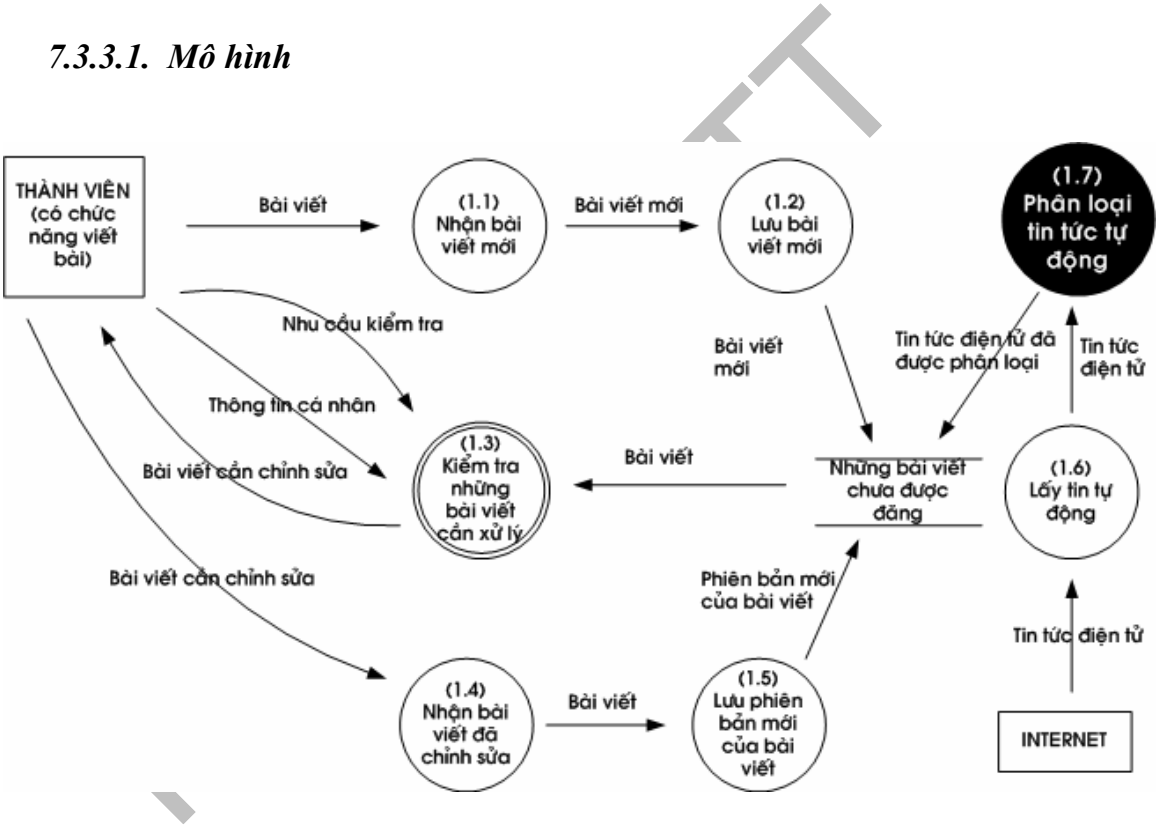
Ví dụ ở toà soạn báo điện tử của chúng ta có mục Kinh doanh\Quốc tế, còn ở báo www.vnexpress.net có mục Thế giới bao gồm nhiều nội dung, trong đó có một số tin tức về Kinh doanh quốc tế, một số tin tức về chính trị thế giới, một số bài về văn hoá chẳng hạn. Như vậy nếu ta chỉ định các bài báo lấy từ mục tin Thế giới ở www.vnexpress.net đều được xếp vào mục Kinh doanh\Quốc tế thì kết quả không còn đúng hoàn toàn nữa. Lúc đó, các thành viên duyệt bài lại phải đọc lần lượt các

bài báo được lấy về một cách thủ công để phân loại chủ đề của tin tức cho phù hợp với cấu trúc chủ đề của mình.

Để hạn chế trường hợp trên, chúng em đưa ra giải pháp là tích hợp module phân loại văn bản vào việc xử lý lấy tin tự động từ Internet. Các tin tức vừa được lấy về sẽ được module phân loại văn bản phân loại tự động vào các chủ đề có sẵn của toà soạn báo. Như vậy, chúng ta sẽ tiết kiệm được nhiều công sức và thời gian duyệt bài của các thành viên một cách đáng kể.

7.3.3. Mô hình DFD quan niệm cấp 2 mới cho ô xử lý Nhận bài và Trả bài

7.3.3.1. Mô hình



Hình 7. 2. Mô hình DFD cải tiến

7.3.3.2. Mô tả mô hình

Mô hình mới chỉ thêm một ô xử lý việc phân loại tin tức tự động sau khi hệ thống lấy tin tức từ trang web khác về.

7.3.3.2.1. Mô tả ô xử lý

Ô xử lý	Tên	Dòng dữ liệu vào	Dòng dữ liệu ra	Diễn giải
(1.7)	Phân loại tin tức tự động	Tin tức điện tử	Tin tức điện tử đã phân loại	Module phân loại văn bản mới tích hợp vào hệ thống thực hiện phân loại tự động các tin tức vừa lấy về.

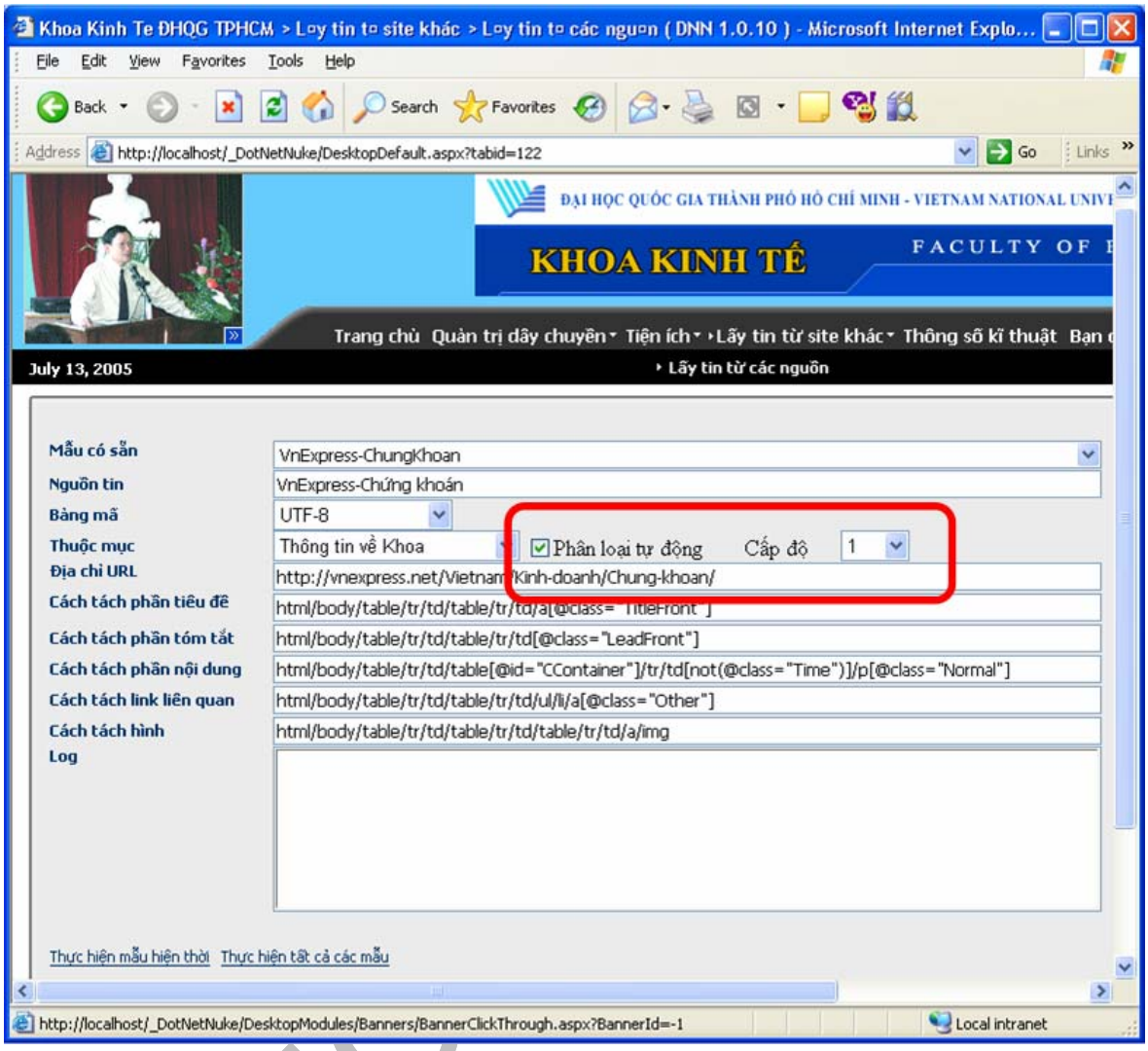
Bảng 7. 3. Bảng mô tả ô xử lý phân loại tin tức tự động

7.4. Triển khai DLL

Chương trình phân loại văn bản tự động được viết trên ngôn ngữ C#, trong khi "Tòa soạn báo điện tử" của luận văn khóa 2000 được viết mã trên nền VB.Net. Do đó, để tích hợp hai hệ thống lại, chúng em đã xây dựng các thành phần chính dùng trong phân loại văn bản thành DLL.

Có thể nói, việc đóng gói chương trình thành dạng DLL ngoài tính tiện lợi trong việc tích hợp giữa các hệ thống xây dựng trên các ngôn ngữ khác nhau, goíi DLL còn có ưu điểm là khả năng sử dụng đơn giản, dễ mang chuyển, là yếu tố quan trọng trong việc xây dựng chương trình.

"Tòa soạn báo điện tử" của luận văn khóa 2000 được xây dựng khá công phu về mặt hình thức lẫn nội dung, cho nên khi tích hợp DLL mới vào, chúng em nhận thấy không cần thiết phải thiết lập thêm giao diện nào nữa. Chúng em chỉ tạo thêm một số lựa chọn cho người dụng có thể bật tắt chức năng phân loại.

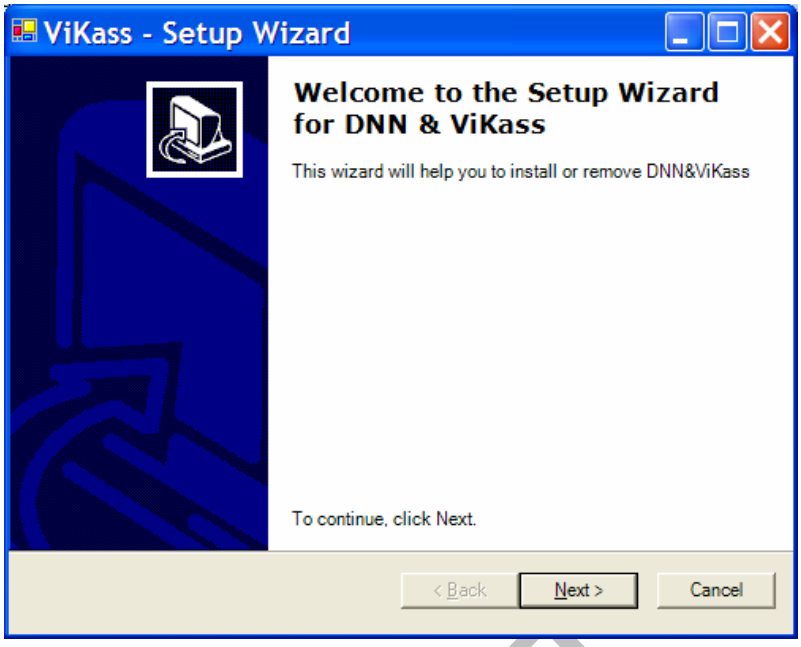


Hình 7. 3. Màn hình lấy tin tức cho phép phân loại tự động

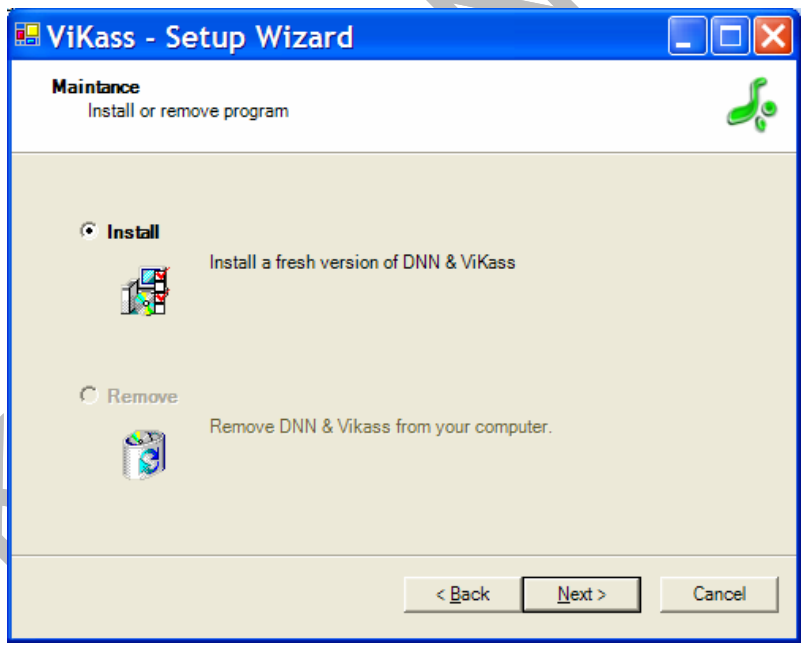
7.5. Chương trình cài đặt "Tòa soạn báo điện tử" đã tích hợp module phân loại tin tức

"Tòa soạn báo điện tử" của luận văn khóa 2000 hiện tại chưa xây dựng công cụ cài đặt vài gỡ chương trình tự động (Install và Uninstall), đòi hỏi người dùng phải có nhiều kiến thức về SQL Server để có thể cài đặt cơ sở dữ liệu một cách thủ công. Vì vậy, nhằm tăng thêm tính tiện dụng của "Tòa soạn báo điện tử", chúng em tự xây dựng công cụ cài đặt tự động "Tòa soạn báo điện tử" vào máy chỉ với thao tác click chuột. Công cụ cài đặt thực hiện việc thiết lập cơ sở dữ liệu vào hệ quản trị SQL Server, thư mục ảo chứa nội dung trang web trong IIS, và tạo shorcut trên desktop.

Một số giao diện của công cụ cài đặt:

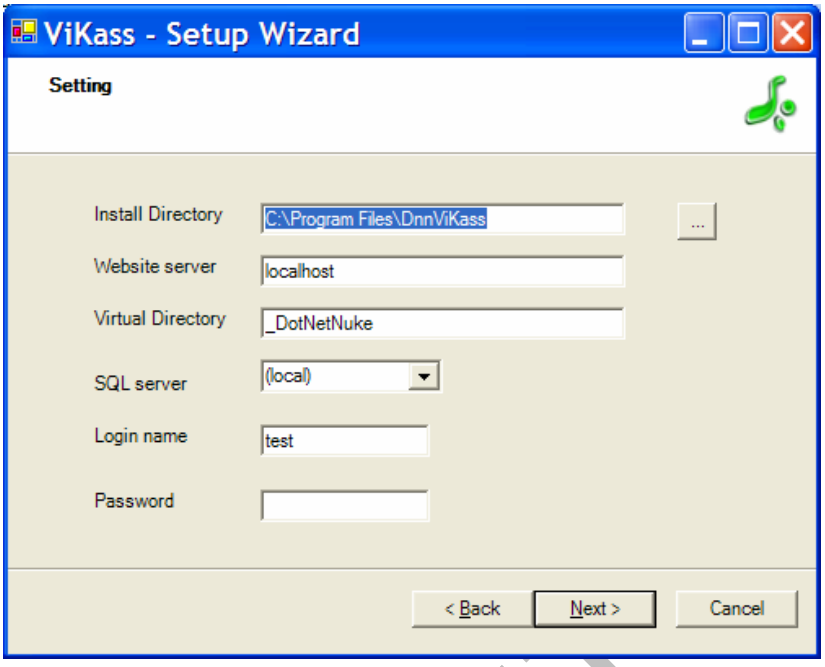


Hình 7. 4. Màn hình bắt đầu. Click Next để bắt đầu cài đặt



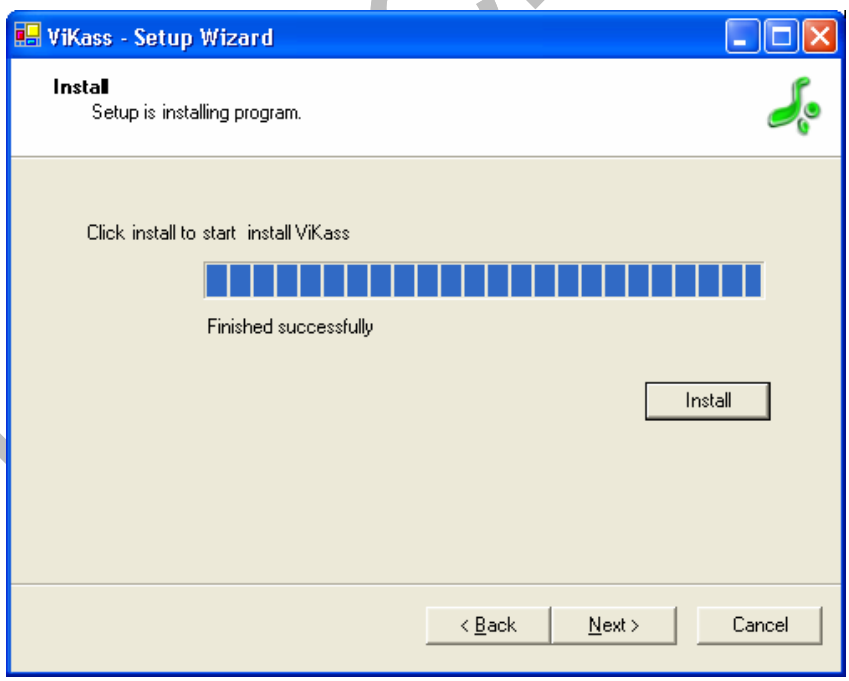
Hình 7. 5.Màn hình chọn chế độ cài đặt hoặc tháo gỡ chương trình.

Chọn Install và click Next để sang bước tiếp theo

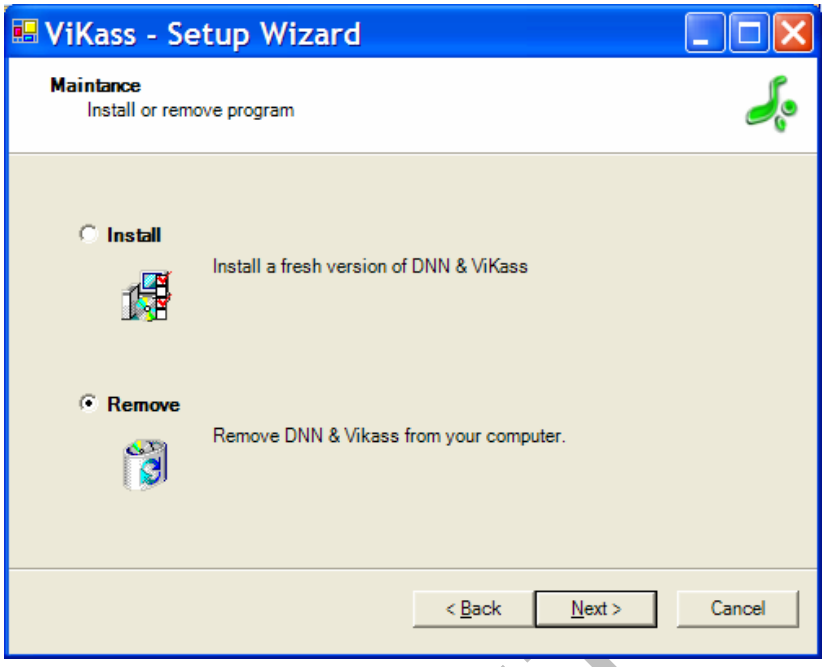


Hình 7. 6.Màn hình chọn đường dẫn để cài đặt chương trình.

Sau khi chọn xong các đường dẫn phù hợp, nhấp vào Next để thực hiện cài đặt.

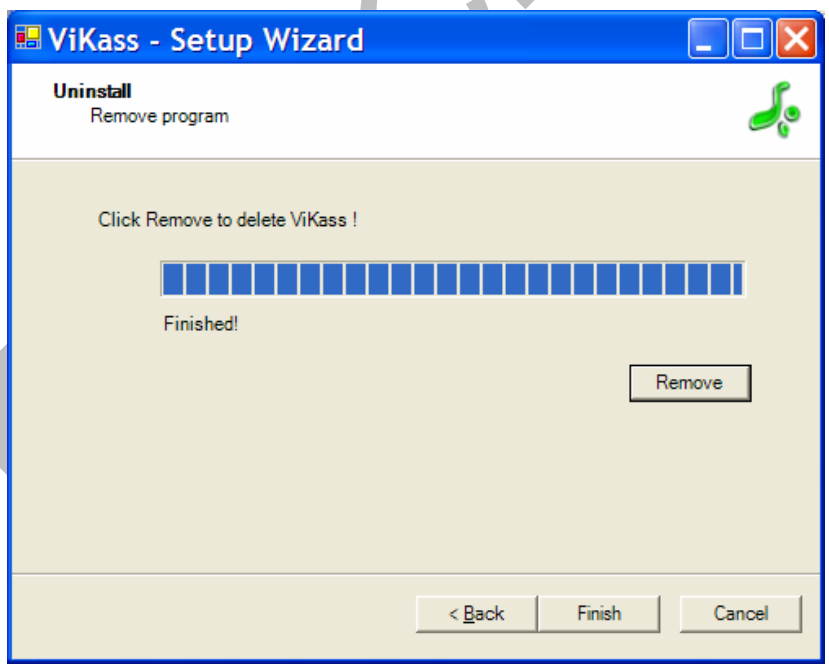


Hình 7. 7.Màn hình cài đặt chương trình



Hình 7. 8.Màn hình chọn chức năng gỡ chương trình.

Chọn Remove để gỡ chương trình đã cài đặt trên máy.



Hình 7. 9.Màn hình gỡ chương trình thành công

7.6. Kết quả

Nhờ việc tích hợp module phân loại văn bản vào trong web "Tòa soạn báo điện tử" mà giờ đây công việc phân loại tin tức điện tử đã trở nên nhanh chóng và tiện lợi hơn. Tuy xác suất phân loại đúng chưa đảm bảo cho hệ thống phân loại văn bản hoàn toàn tự động, mà cần có sự duyệt bài lại để đảm bào chính xác hoàn toàn, nhưng module phân loại văn bản bán tự động cũng đã cung cấp cho người dùng một tiện ích vô cùng hữu hiệu.

Chương 8
TỔNG KẾT

Kết quả đạt được

Về mặt lý thuyết

Về mặt thực hành

Hạn chế và hướng giải quyết

Kết luận

Chương 8. TỔNG KẾT

8.1. Kết quả đạt được

8.1.1. Về mặt lý thuyết

Phân loại văn bản là một bài toán khó và rất thú vị. Khó bởi vì vấn đề phân loại văn bản cần phải thực hiện xử lý ngôn ngữ, mà như chúng ta đều biết, ngôn ngữ tự nhiên là muôn hình vạn trạng, không chỉ phong phú về từ vựng, cú pháp mà còn phức tạp về ngữ nghĩa. Nhưng đây lại là bài toán rất thú vị vì với mỗi ngôn ngữ khác nhau, chúng ta phải thực hiện những cách xử lý khác nhau đối với ngôn ngữ.

Trong khuôn khổ luận văn này, những vấn đề liên quan đến đề tài như các phương pháp tách từ và phương pháp phân loại văn bản đã được chúng em tiến hành nghiên cứu khá công phu theo cả chiều rộng lẫn chiều sâu về. Trên cơ sở nghiên cứu đó, các hướng tiếp cận áp dụng cho tiếng Anh và tiếng Hoa phù hợp đã được lựa chọn và thử nghiệm lên tiếng Việt.

Đặc biệt, ở giai đoạn tách từ chuẩn bị cho phân loại, chúng em đã tìm hiểu một cách sâu sắc về hướng thống kê dựa trên Internet. Dựa trên nền tảng đó, chúng em mạnh dạn thực hiện cải tiến phương pháp tách từ dựa trên Internet và thuật toán di truyền thay vì sử dụng lại các công cụ tách từ tiếng Việt đã được công bố trước đây. Hướng tiếp cận mới này không những hạn chế được nhược điểm phụ thuộc vào tập ngữ liệu của các phương pháp khác mà còn đem lại khả năng khai thác vô tận nguồn dữ liệu khổng lồ của nhân loại : word-wide-web. Kết quả đạt được của phương pháp này là hoàn toàn khả quan và chấp nhận được đối với một hướng tiếp cận mới cho tách từ tiếng Việt dùng trong phân loại văn bản.

Phương pháp phân loại văn bản Naïve Bayes thường được dùng trong phân loại văn bản tiếng Anh, nay được áp dụng trong tiếng Việt với hướng tiếp cận dựa trên thống kê từ Google tỏ ra khá hiệu bởi. Nhờ tính đơn giản, các thông số tính toán không cần quá lớn như các phương pháp khác, khả năng linh hoạt đối với sự thay đổi về thông tin huấn luyện, thời gian phân loại phù hợp yêu cầu, Naïve Bayes đã tỏ ra rất phù hợp với các yêu cầu đề ra.

8.1.2. Về mặt thực nghiệm

Công trình nghiên cứu của luận văn đã thực hiện được nhiều thử nghiệm đối với từng hướng tiếp cận tách từ tiếng Việt dựa trên Google cũng như phân loại văn bản. Nhờ vậy, kết quả thực nghiệm đã chứng minh được tính hiệu quả cho các công thức trên lý thuyết.

Qua kết quả thực nghiệm, chúng em nhận thấy công thức tách từ của [H. Nguyen et al, 2005] và công thức MI do chúng em đề nghị cho hiệu quả gần tương đương nhau, tuy cách tính của [H. Nguyen et al, 2005] có vẻ chính xác hơn cho các từ có hai tiếng.

Kết quả thực nghiệm ở phần phân loại văn bản cho thấy công thức phân loại trong [H. Nguyen et al, 2005] là mang tính chủ quan của tác giả, và dữ liệu thực nghiệm không đủ lớn để có thể kết luận. Nhưng khi áp dụng thử nghiệm trên số lượng văn bản và chủ đề nhiều hơn thì cách tính này cho ra kết quả thấp hơn nhiều so với kết quả mà tác giả trình bày. Kết quả sử dụng công thức Naïve Bayes đã cho kết quả khả quan hơn nhờ dựa vào lý thuyết đã được chứng minh từ các công trình trước.

8.2. Hạn chế và hướng phát triển

Với những kết quả thử nghiệm ban đầu, hệ thống phân loại văn bản đã bước đầu hoạt động hiệu quả , góp phần thực hiện phân loại văn bản bán tự động, giúp tiết kiệm được thời gian và công sức đọc văn bản một cách thủ công. Mặc dù những kết quả của hệ thống là chấp nhận được, tuy nhiên hệ thống có thể được cải thiện về độ chính xác và tốc độ nếu ta khắc phục một số hạn chế của hệ thống và thực hiện thêm các hướng mở rộng khác được trình bày sau đây.

Phương pháp tách từ dựa trên Internet và thuật toán di truyền tỏ ra khá linh hoạt trong việc xử lý ngôn ngữ. Tuy nhiên với mặt bằng chất lượng Internet hiện nay ở Việt Nam, bước đầu thực hiện việc tách từ sẽ khá lâu vì phải mất thời gian lấy thông tin từ công cụ tìm kiếm trên mạng. Nhưng khi các thông tin trên được lưu lại tương đối lớn, tốc độ phân định ranh giới từ sẽ được cải thiện.

Trong phần thử nghiệm phân loại văn bản, hiện tại chúng em quy định một chủ đề chỉ có một từ khóa chính là tên của chủ đề đó. Chính đây là một điểm hạn chế dẫn đến kết quả phân loại văn bản chưa cao như trong các công trình phân loại văn bản tiếng Anh. Do vậy, nhu cầu xây dựng một công cụ chiết xuất từ khóa tự động từ tập dữ liệu tin tức thô là rất cần thiết. Khi đã có tập từ khóa, độ chính xác của việc phân loại văn bản sẽ tăng lên đáng kể.

Hiện tại, luận văn thực hiện phân loại theo hướng tiếp cận Naïve Bayes với các từ được tách trong câu mà không có sự chọn lựa những từ đặc trưng để thực hiện phân loại. Điều này dẫn đến một số từ không có ý nghĩa phân loại vẫn xem như có vai trò tương tự như những từ có ý nghĩa phân loại cao. Nếu chúng ta nghiên cứu thực hiện chọn lựa các đặc trưng của văn bản (feature selection) rồi mới phân loại, chứng ta sẽ đạt được tỉ lệ chính xác cao hơn và tăng tốc độ xử lý của hệ thống sẽ tăng lên đáng kể.

Trong luận văn này, chúng em chỉ mới chọn thực hiện thử nghiệm phân loại tiếng Việt với hướng tiếp cận Naïve Bayes mà chưa chọn các phương pháp khác. Điều này là do phần nhiều bởi tính chủ quan và một số giới hạn về sự nghiên cứu. Do đó, việc mở rộng thử nghiệm phân loại văn bản tiếng Việt trên các hướng tiếp cận khác như SVM, kNN… sẽ có thể đem lại nhiều kết quả cao hơn trong lĩnh vực này.

8.3. Kết luận

Hệ thống phân loại văn bản ứng dụng công cụ tách từ tiếng Việt dựa trên thống kê Internet và thuật toán di truyền là ứng dụng một hướng tiếp cận mới đầy hứa hẹn cho phương pháp tách từ tiếng Việt, vốn hiện nay vẫn còn nhiều hạn chế. Ngoài ra, phần mềm phân loại bán tự động tin tức của luận văn có nhiều ý nghĩa thực tiễn trong việc quản trị thông tin của các tờ báo điện tử nói riêng, và trong các lĩnh vực đòi hỏi đến việc xử lý ngôn ngữ nói chung. Với ý nghĩa to lớn đó, chúng em nguyện cố gắng nhiều hơn nữa tìm hiểu, nghiên cứu cải tiến hệ thống đạt hiệu quả ngày càng cao.

TÀI LIỆU THAM KHẢO

[Broder et al, 2003] Andrei Z. Broder (NY), Marc Najork(CA), Janet L. Wiener(CA). *Efficient URL Caching for World Wide Web Crawling*, 2003.

[Bagrow et al, 2004J.P. Bagrow, H.D. Rozenfeld, E.M. Bollt, and D. ben-Avraham, *"How Famous is a Scientist? – Famous to Those Who Know Us.",* arxiv.org/abs/cond-mat/0404515, Europhys. Lett., 67, (4) 511-516 (2004).

[Berger, 1999] Adam Berger, *Error-correcting output coding for text classification*. In proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering, Stockholm, Sweeden, 1999.

[Chien et al, 1997] Lee-Feng Chien, T. I. Huang, M. C. Chen. 1997. *PATTree-Based Keyword Extraction for Chinese Information Retrieval*, Proceedings of 1997 ACM SIGIR Conference, Philadelphia, USA, 50-58.

[Chih-Hao Tsai, 2000] Chih-Hao Tsai, 2000. MMSEG: *A Word Identification System for Mandarin Chinese Text Based on Two Variants of the Maximum Matching Algorithm*. Web publication at http://technology.chtsai.org/mmseg/

[Church et al, 1991] Kenneth Church, William Gale, Patrick Hanks, Donald Hindle, *Using Statistics in Lexical Analysis,* Bell Laboratories and Oxford University Press, 1991.

[Dasarathy, 1991] Belur V. Dasarathy. Nearest Neighbor (NN) Norms: *NN Pattern Classication Techniques.* McGraw-Hill Computer Science Series. IEEE Computer Society Press, Las Alamitos, California, 1991.

[Đinh Điền et al, 2001] Đinh Điền, Hoang Kiem, Nguyen Van Toan. 2001. *Vietnamese Word Segmentation*. pp. 749 -756. The sixth Natural Language Processing Pacific Rim Symposium, Tokyo, Japan.

[Đinh Điền,2004] Đinh Điền, *Giáo trình xử lý ngôn ngữ tự nhiên*, Đại học Khoa Học Tự Nhiên Tp.HCM, 12/2004

[Foo & Li, 2004] Foo S., Li H. 2004. *Chinese Word Segmentation and Its Effect on Information Retrieval*, Information Processing & Management: An International Journal, 40(1): 161-190.

[Fuhr et al, 1991] N. Fuhr, S. Hartmanna, G. Lustig, M. Schwantner, and K. Tzeras. *Air/x – a rule-based multistage indexing system for large subject fields*. In 606-623, editor, Proceedings of RIAO'91, 1991.

[Ghani, 2000] Rayid Ghani, *Using error-correcting codes for text classification*. In proceedings of Seventeenth International Conference on Machine Learning, 2000

[Goldberg et al, 1992] Goldberg, D.E., Deb, K., & Clark, J.H. (1992). *Genetic algorithms, noise, and the sizing of populations*. Complex Systems, 6. 333-362.

[H. Nguyen et al, 2005] H. Nguyen, H. Nguyen, T. Vu, N. Tran, K. Hoang ,2005. *Internet and Genetics Algorithm-based Text Categorization for Documents in Vietnamese*, Research, Innovation and Vision of the Future, the 3rd International Conference in Computer Science, (RIFT 2005), Can Tho, Vietnam.

[He et al, 1996] He, J., Xu, J., Chen, A., Meggs, J, & Gey, F. C. (1996). *Berkeley Chinese information retrieval at TREC–5: Technical report*. http://trec.nist.gov/pubs/trec5/t5_proceedings .html, Maryland.

[James & Daniel, 2005] James P.Pagrow & Daniel ben-Avraham. *On the Google – Fame of Scientist and other populations*, 2005.

[Jason, 2001] Jason D.M Rennie, *Improving Multi-class Text Classification with Naive Bayes,* 2001

[Joachims, 1998] Thorsten Joachims. *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. In European Conference on Machine Learning (ECML), 1998

[Kwok, 1997a] Kwok, K.L. (1997a) *Comparing representations in Chinese information retrieval*.http://ir.cs.qc.edu/#publi_.

[Kwok, 1997b] Kwok, K.L. (1997b) *Lexicon effects on Chinese information retrieval*. http://ir.cs.qc.edu/#publi_.

[Le An Ha, 2003] Le An Ha, 2003. A *method for word segmentation Vietnamese*. Proceddings of Corpus Linguistics 2003, Lancaster, UK.

[Maron, 1961] Maron, *Automated indexing*, JACM, 1961

[Mateev et al, 1997] Mateev, B., Munteanu, E., Sheridan, P., Wechsler, M., & Schuble, P. (1997). ETH TREC-6: *Routing, Chinese, cross-language and spoken document retrieval*. http://trec.nist.gov/pubs/ trec6/t6_ proceedings.html, Maryland.

[McCallum & Nigam, 1998] Andrew McCallum & Kamal Nigam. *A comparison of Event Models for Naïve Bayes Text Classification,* 1998.

[Mitchell, 2005] Tom M. Mitchell. *Generative and Discriminative Classifiers: Naïve Bayes and Logistic Regression,* textbook Machine Learning, DRAFT OF March 6, 2005.

[Nie et al, *1996] Nie, J.Y., Brisebois, M., & Ren, X.B. (1996).* On Chinese text retrieval. Proceedings of SIGIR '96, Zurich, Switzerland, 225-233.

[Ong & Chen, 1999] Thian-Huat Ong & Hsinchun Chen. Updateable PAT-Tree *Approach to Chinese Key Phrase Extraction using Mutual Information: A Linguistic Foundation for Knowledge Management*, Proceedings of the Second Asian Digital Library Conference, pp.63-84, 1999.

[Platt, 1998] J.Platt. *Sequential minimal optimization : A fast algorithm for training support vector machines*. In Technical Report MST-TR-98-14. Microsoft Research,1998

[Richard et al,1996] Richard W Sproat. Chilin Shih, William Gale, and Nancy Chang. *A stochastic finite-state word-segmentation algorithm for Chinese*. CL, 22(3):377-404. 1996

[Rijsbergen et al, 1970] Van Rijsbergen, Robertson, Sparck Jones, Croft, Harper (early 1970's) *–search engines*

[Rudi & Paul, 2005] Rudi Cilibrasi & Pau Vitanyi, *Automatic Meaning Discovery Using Google*, Neitherlands, 2005.

[Sahami et al, 1998] Sahami, Dumais, Heckerman, Horvitz (1998) –spam filtering

[Schütze et al,1995] Schütze, H. Hull, D. , and Pedersen, J. (1995). *A comparison of classifier and document representations for the routing problem*. In International ACM SIGIR Conference on Research and Development in Information Retrieval.

[Simkin & Roychowdhurry, 2003] M.V. Simkin and V.P. Roychowdhury, *"Theory of Aces:Fame by chance or merit?"* (preprint, arxiv.org/abs/condmat/0310049, 2003).

[Su et al, 1993] Keh-Yih Su, Ming-Wen Wu, Jing-Shin Chang. A Corpus-based *Approach to Automatic Compound Extraction*, 1993

[Vapnik & Cortes, 1995] C.Cortes and V.Vapnik, *Support Vector Network. Machine Learning*, 20:273-297,1995

[Vapnik, 1995] V.Vapnik, *The Nature of Statistical Learning Theory*. Springer, NewYork, 1995.

[Wiener et al, 1995] Erik Wiener, Jan O. Pedersen, and Andreas S. Weigend. *A Neural Network Approach to Topic Spotting*. In Proceedings of the Fourth Annual Sumposium on Document Analysis and Information Retrieval (SDAIR'95), 1995.

[William & Yoram, 1996] William W. Cohen and Yoram Singer. *Context-sensitive learning methods for text categorization*. In SIGIR '96: Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996. 307-315.

[Wu & Tseng, 1993] Wu, Z.M., & Tseng, G. (1993*). Chinese text segmentation for text retrieval: Achievements and problems*. Journal of the American Society for Information Science, 44 (9), 532-542.

[Wu & Tseng, 1995] Wu, Z.M., & Tseng, G. (1995). ACTS: *An automatic Chinese text segmentation system for full text retrieval*. Journal of the American Society for Information Science, 46(2), 83-96

[Yang & Chute, 1992] Y. Yang and G.Chute. *A Linear Least Squares Fit Mapping Method for Information Retrieval from Natural Language Texts*, 1992

[Yang & Chute, 1994] Y. Yang and G.Chute. *An example-based mapping method for text categorization and retrieval*. ACM Transaction on Information Systems(TOIS), 12(3):252-277,1994

[Yang & Petersen, 1997] Yang, Y. and Petersen, J. (1997). *A comparative study on feature selection in text categorization*. In International Conference on Machine Learning(ICML).

[Yang & Wilbur, 1996] Yang, Y. and Wilbur, J. (1996). *Using corpus statistics to remove redundant words in text categorization*. Journal of the American Society for Information Science, 47(5):357-369.

[Yang & Xiu, 1999] Yiming Yang and Xin Liu, *A re-examination of text categorization methods*. Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR' 99)

[Yang, 2000] Yiming Yang. *An Evaluation of Statistical Approaches to Text Categorization*, Kluwer Academic Publishers, 2000.