Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

XÂY DỰNG HỆ THỐNG HỖ TRỢ KHUYẾN NÔNG TRÊN CÂY LÚA QUA MẠNG THÔNG TIN DI ĐỘNG

Lương Thế Anh[1], Nguyễn Thái Nghe[2] và Nguyễn Chí Ngôn[3]

[1] *Trung Tâm NN-TH, Trường Đại học Xây dựng Miền Tây*
[2] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*
[3] *Khoa Công nghệ, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 08/04/2014
Ngày chấp nhận: 28/08/2014

Title:
Development of an mobile communication network-based agricultural extension support system

Từ khóa:
Hệ thống hỗ trợ khuyến nông, tách từ tiếng Việt, phân loại văn bản

Keywords:
Agricultural extension support system, Vietnamese word segmentation, text classification

ABSTRACT

The objective of this research is to build a link system between farmers and agricultural experts to support the agricultural extension via mobile communication network and to collect real data used for developing automatic agricultural extension systems in the future. This system can be considered as "24/7 farmers link program". To build this system, at first, we need to build modules for sending and receiving SMS and MMS messages. These modules are important for farmers to send data of rice status to agricultural experts for receiving consultations. Next, a message classification module is built based on a combination of machine learning methods (e.g. SVM) with image and text processing technologies. To make it more convenient for experts and system users, a website is build to integrate these modules into the whole system. Preliminary results show that the construction of this system is feasible. This is also the foundation for building an online automatic agricultural extension support system through mobile communication network.

TÓM TẮT

Nghiên cứu này được thực hiện với mục tiêu xây dựng một hệ thống nhịp cầu giữa nhà nông và các chuyên gia nông nghiệp nhằm hỗ trợ công tác khuyến nông (trước mắt là trên cây lúa) qua mạng thông tin di động, đồng thời thu thập dữ liệu thực tế dùng để phát triển các hệ thống khuyến nông tự động sau này. Hệ thống này có thể được xem là "nhịp cầu nhà nông 24/7". Để xây dựng được hệ thống, trước hết ta cần xây dựng mô-đun gửi và nhận tin nhắn SMS/MMS. Các mô-đun quan trọng này hỗ trợ cho nông dân gửi dữ liệu về tình trạng của cây lúa để được tư vấn bởi các chuyên gia nông nghiệp. Tiếp đến, một mô-đun phân loại tin nhắn được thiết lập dựa trên sự kết hợp các phương pháp máy học với công nghệ xử lý ảnh và xử lý văn bản. Để thuận lợi cho các chuyên gia và người dùng hệ thống, một website được xây dựng nhằm tích hợp các mô-đun trên lại với nhau. Kết quả nghiên cứu bước đầu cho thấy việc xây dựng hệ thống này là rất khả thi. Đó cũng là nền tảng để xây dựng hệ thống hỗ trợ khuyến nông tự động trực tuyến qua mạng thông tin di động.

1 GIỚI THIỆU

Việt Nam hiện là một nước nông nghiệp, phần lớn người dân sống chủ yếu dựa vào trồng trọt và chăn nuôi (Tổng cục thống kê, 2014a). Trong đó, cây lúa đóng vai trò quan trọng, là nguồn an ninh lượng thực chủ yếu (Tổng cục thống kê, 2014b,c). Ngành trồng lúa đã đạt được những thành tựu đáng kể đưa Việt Nam trở thành nước có sản lượng gạo xuất khẩu lớn hàng đầu thế giới.

Ngày nay, việc trồng lúa trở nên khó khăn, phức tạp hơn do thường xuyên phát sinh nhiều loại sâu bệnh lạ và môi trường, khí hậu bị ô nhiễm. Vì vậy, việc trồng lúa ngày nay đòi hỏi phải có sự tích lũy những kinh nghiệm, tích hợp các tri thức và thông tin từ nhiều nguồn khác nhau. Để duy trì khả năng cạnh tranh, nâng cao năng suất, chất lượng hạt gạo, người nông dân hiện đại thường dựa vào các chuyên gia và các cố vấn nông nghiệp để cung cấp kiến thức, thông tin cho việc ra quyết định. Khó khăn ở chỗ là các chuyên gia không phải lúc nào cũng luôn có sẵn khi nhà nông cần đến và chi phí mà nông dân bỏ ra để được hỗ trợ là khá cao.

Với sự phát triển của mạng thông tin và các thiết bị di động, khả năng tiếp cận với tri thức thông qua mạng thông tin di động ngày càng trở nên đơn giản và phổ biến với mọi thành phần xã hội, chẳng hạn như dịch vụ tra cứu điểm thi tuyển sinh qua SMS, dịch vụ tin nhắn SMS phục vụ bạn đọc sử dụng thư viện, dịch vụ tra cứu thông tin chứng khoán, tỷ giá, giá vàng, SMS Marketing, SMS Banking,… Bên cạnh đó, công nghệ thông tin và các giải pháp máy học đã phát triển khá mạnh mẽ trong những năm gần đây, trong khi đó nguồn dữ liệu nông nghiệp dành cho khai phá hiện còn rất khan hiếm. Hiện tại, ở Đồng bằng sông Cửu Long *chưa có hệ thống tin học nào được xây dựng để hỗ trợ công tác khuyến nông* và thực hiện thu thập dữ liệu qua mạng thông tin di động.

Từ thực tiễn đó, việc xây dựng một hệ thống nhằm hỗ trợ về mặt thông tin, kỹ thuật cho nhà nông đồng thời để thu thập dữ liệu thực tế qua mạng thông tin di động là rất cần thiết, cấp bách và hợp lý.

Trong bài viết này, *chúng tôi đề xuất một giải pháp mới nhằm hỗ trợ cho công tác khuyến nông, cụ thể là khuyến nông qua mạng thông tin di động bằng tin nhắn SMS/MMS*. Bài viết tập trung nghiên cứu các công nghệ cũng như các phương pháp để xây dựng những mô-đun thiết yếu cho hệ thống như mô-đun gửi và nhận tin nhắn SMS/MMS, mô-đun phân loại tin nhắn tự động bằng kỹ thuật phân loại văn bản dùng giải thuật SVM và sau cùng là xây dựng một website hoàn chỉnh để tích hợp các module trên, quản lý và cấu hình hệ thống.

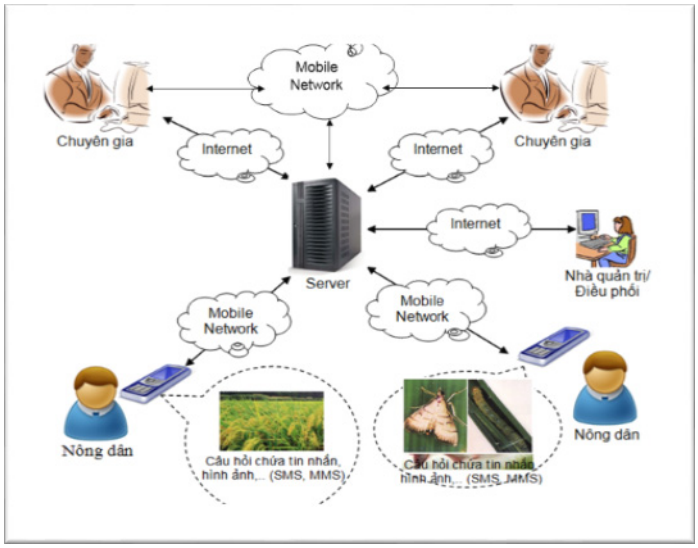
Với sự ra đời của hệ thống này sẽ khắc phục được một phần những khó khăn của người nông dân trong quá trình sản xuất lúa gạo, với khả năng ứng dụng rộng rãi trong lĩnh vực nông nghiệp nó được xem là một công cụ hữu ích với tiềm năng rộng lớn để hỗ trợ kỹ thuật cho nông dân một cách kịp thời, tiết kiệm chi phí và là một công cụ hiệu quả để thu thập dữ liệu thực tiễn làm cơ sở cho việc phát triển các hệ thống hỗ trợ hoàn toàn tự động sau này.

2 XÂY DỰNG HỆ THỐNG

2.1 Mô hình hoạt động tổng thể của hệ thống

Mô hình tổng thể của hệ thống được biểu diễn như trong Hình 1. Ở đó, khi nhà nông có vấn đề/câu hỏi (chẳng hạn như liên quan đến bệnh hại trên cây lúa, cần tư vấn cách điều trị,…) thì họ có thể đặt câu hỏi bằng tin nhắn SMS hoặc chụp lại hình ảnh hiện trạng (MMS) (có thể kèm theo câu hỏi) để gửi đến hệ thống bằng điện thoại di động. Yêu cầu này sẽ được hệ thống chuyển đến các chuyên gia thích hợp trong từng lĩnh vực để được giải đáp. Ngay sau khi nhận được phản hồi từ phía chuyên gia, hệ thống sẽ phản hồi lại kết quả cho nhà nông. Như vậy, với ý tưởng này, hệ thống cũng có thể xem như ***"Nhịp cầu nhà nông trực tuyến 24/7"***.

Đối với tin nhắn hình ảnh, hiện tại hệ thống chưa hỗ trợ phân loại tự động, khi hệ thống nhận được hình ảnh mà nhà nông gửi lên, điều phối viên sẽ xem xét và phân loại hình ảnh đó rồi gửi cho chuyên gia giải đáp, sau khi nhận được câu trả lời của chuyên gia thì hệ thống sẽ tự động gửi nội dung trả lời cho nhà nông.



Hình 1: Mô hình hoạt động của hệ thống

Đối với tin nhắn văn bản, khi nhận được câu hỏi của nhà nông qua mô-đun gửi nhận tin nhắn SMS, mô-đun phân loại tin nhắn SMS tự động sẽ tự động thực hiện một số bước tiền xử lý cơ bản như tách từ, chọn từ khóa hay loại bỏ từ dừng. Do văn bản là tin nhắn, nên số lượng từ khóa không nhiều và ít khi lặp lại, nghiên cứu này đưa ra hai phương án chọn từ khóa là phương án thủ công và phương án tự động.

– Với phương án thủ công thì hệ thống sẽ giữ lại những từ có trong danh sách từ khóa (tập đặc trưng văn bản) đã được xây dựng thủ công bởi các chuyên gia từ trước, sau đó véc-tơ hóa các từ được giữ lại đó và đưa vào mô hình của phương án này để phân loại.

– Với phương án tự động thì hệ thống sẽ chọn từ khóa bằng cách loại bỏ các từ dừng (stopwords) là những từ thường xuất hiện trong văn bản nhưng không có giá trị phân loại chẳng hạn như từ "và", "nhưng", "có", "không", sau đó véc-tơ hóa tất cả các từ còn lại và đưa vào mô hình của phương án này để phân loại.

Trong giai đoạn ban đầu này, bộ từ khóa và tập dữ liệu do nhóm tác giả thu thập và xây dựng chưa đủ lớn và đa dạng nên việc phân loại tự động chỉ nhằm mục đích minh họa cho mô-đun phân loại tự động và kiểm tra việc vận hành của hệ thống, còn việc phân loại vẫn là một hệ thống bán tự động, có sự kiểm tra, giám sát của điều phối viên. Do hệ thống hiện là bán tự động nên điều phối viên cần kiểm tra kết quả phân loại của mô-đun tự động và thực hiện phân loại lại để làm cơ sở cho việc xây dựng và huấn luyện lại mô hình, sau khi điều phối viên phân loại, câu hỏi mới được chuyển đến chuyên gia thích hợp để trả lời. Khi nhận được câu trả lời từ chuyên gia, hệ thống tự động gửi nội dung trả lời cho nhà nông.

Khi hệ thống nhận đủ số lượng tin nhắn SMS mới đến (theo cấu hình của hệ thống), hệ thống sẽ tự động xây dựng lại bộ từ khóa và huấn luyện lại mô hình với bộ từ khóa và dữ liệu mới, sau khi huấn luyện xong hệ thống sẽ sử dụng mô hình mới huấn luyện vào phân loại tin nhắn mới đến hệ thống. Hệ thống sẽ lặp đi lặp đi lặp lại việc xây dựng lại bộ từ khóa và huấn luyện lại mô hình cho đến khi lượng dữ liệu thu thập đủ lớn và độ chính xác phân loại là chấp nhận được (do quản trị viên của hệ thống quyết định).

Toàn bộ quy trình xây dựng hệ thống được tóm tắt qua các bước sau:

– Xây dựng mô-đun gửi/nhận tin nhắn SMS/MMS

– Xây dựng mô-đun phân loại nội dung tin nhắn văn bản bằng kỹ thuật SVM.

– Tách từ bằng công cụ VnTokenizer.

– Chọn từ khóa hoặc loại bỏ từ dừng (tùy theo cấu hình hệ thống).

– Xây dựng bộ dữ liệu huấn luyện cho mô hình SVM.

– Huấn luyện mô hình.

– Đánh giá kết quả.

– Xây dựng website hoàn chỉnh để tích hợp toàn bộ các mô-đun, quản lý và cấu hình hệ thống.

Sau đây chúng tôi sẽ trình bày chi tiết các bước để xây dựng các mô-đun này.

2.2 Xây dựng mô-đun gửi và nhận tin nhắn

2.2.1 Tin nhắn văn bản (SMS)

Gửi tin nhắn: Về tổng thể, có 2 cách để gửi tin nhắn SMS từ điện thoại di động đến máy tính (Developer's Home, 2014).

Cách 1: Kết nối điện thoại di động hoặc modem GSM/GPRS/3G vào máy tính. Sau đó dùng tập lệnh AT để chỉ thị cho điện thoại hoặc modem gửi tin nhắn SMS. Để gửi tin nhắn, trước hết cần lắp SIM được nhà mạng cung cấp vào điện thoại hoặc modem, sau đó kết nối modem vào máy vi tính bằng dây cáp, hồng ngoại hay bluetooth. Sau khi kết nối thành công, ta có thể điều khiển modem bằng cách gửi chỉ thị đến nó. Chỉ thị được sử dụng để điều khiển modem được gọi là tập lệnh AT (viết tắt của ATtention). Tập lệnh AT là những chỉ thị được sử dụng để điều khiển modem hay điện thoại di động, danh sách các lệnh thông dụng được mô tả trong Bảng 1.

Bảng 1: Một số lệnh AT dùng để gửi tin nhắn

Lệnh AT	Công dụng
AT + CMGS	Gửi tin nhắn
AT + CMSS	Gửi tin nhắn từ bộ lưu trữ
AT + CMGW	Ghi tin nhắn vào bộ nhớ
AT + CMGD	Xóa tin nhắn
AT + CMMS	Gửi thêm tin nhắn

Cách 2: Kết nối máy tính với Trung tâm SMS (SMSC) hoặc SMS Gateway của mạng không dây hoặc nhà cung cấp dịch vụ SMS. Sau đó gửi tin nhắn SMS bằng cách sử dụng các giao thức/giao diện được hỗ trợ bởi SMSC hoặc SMS Gateway.

Cách gửi tin nhắn thông qua modem hay điện thoại di động kết nối trực tiếp với máy tính có hạn chế là tốc độ gửi tin nhắn SMS rất thấp. Nếu cần tốc độ gửi cao hơn thì cần thiết phải thiết lập kết nối trực tiếp đến Trung tâm SMS hoặc SMS Gateway của mạng không dây. Kết nối này có thể được thực hiện qua mạng Internet hoặc kết nối quay số. Nếu không thể kết nối trực tiếp đến Trung tâm SMS hoặc SMS Gateway của mạng không dây thì ta có thể kết nối đến SMS Gateway của một nhà cung cấp dịch vụ SMS nào đó, lúc đó SMS Gateway này sẽ chuyển tiếp tin nhắn SMS đến một Trung tâm SMS thích hợp. Sau khi đăng ký và thiết lập tài khoản với nhà mạng không dây hoặc nhà cung cấp dịch vụ SMS, ta đã có thể bắt đầu gửi tin nhắn SMS bằng cách sử dụng các giao thức/giao diện được hỗ trợ bởi SMSC hoặc SMS Gateway.

Nhận tin nhắn***:*** Tương tự như việc gửi tin SMS, cũng có 2 cách để nhận tin nhắn SMS trên máy tính.

Cách 1: Kết nối điện thoại di động hoặc modem GSM/GPRS/3G vào máy tính. Sau đó dùng máy tính và tập lệnh AT để đọc tin nhắn nhận được từ điện thoại đi động hoặc modem. Việc nhận tin nhắn SMS thông qua một modem có một lợi thế là nhà mạng không dây thường không tính phí nhận tin nhắn khi dùng với một *Mô-đun Nhận diện Thuê bao* (thẻ SIM). Bất lợi của việc nhận tin nhắn theo cách này là modem không thể xử lý một số lượng lớn lưu lượng tin nhắn SMS truy cập. Có một cách để giải quyết vấn đề này đó là sử dụng nhiều modem để cân bằng tải lưu lượng SMS truy cập. Mỗi một modem sẽ có một thẻ SIM và số thuê bao riêng.

Bảng 2: Một số lệnh AT dùng để nhận tin nhắn SMS

Lệnh AT	Công dụng
AT + CNMI	Để xác định tin nhắn mới
AT + CMGL	Liệt kê tất cả tin nhắn
AT + CMGR	Đọc tin nhắn

Cách 2: Truy cập đến Trung tâm tin nhắn (SMSC) hoặc SMS Gateway của mạng không dây. Mọi tin nhắn SMS nhận được sẽ được chuyển tiếp đến máy tính thông qua giao thức/giao diện được hỗ trợ bởi SMSC hoặc SMS Gateway.

Cũng giống như việc gửi tin nhắn, việc nhận tin nhắn thông qua điện thoại hoặc modem GSM/GPRS có một số hạn chế, đó là tốc độ truyền tải SMS quá thấp. Nếu cần tốc độ cao hơn thì cần thiết phải thiết lập kết nối trực tiếp đến Trung tâm SMS hoặc SMS Gateway của mạng không dây. Sau khi thiết lập một tài khoản với nhà mạng không dây hoặc nhà cung cấp dịch vụ SMS, SMSC hoặc SMS Gateway sẽ bắt đầu chuyển tiếp các tin nhắn đến ứng dụng SMS bằng cách sử dụng một số các giao thức/giao diện. Cũng giống như việc gửi tin, để kết nối đến SMSC, bắt buộc phải có các giao thức SMSC. Việc nhận tin nhắn theo cách này cũng dễ như việc gửi.

2.2.2 Tin nhắn đa phương tiện (MMS)

Các giao ***thức của mô hình MMS:*** Các thiết bị di động (MMS Clients) và các Trung tâm tin nhắn đa phương tiện muốn liên lạc được với nhau phải thông qua các giao thức. Có hai tiêu chuẩn quan trọng để định nghĩa công nghệ MMS, một được

xuất bản bởi 3GPP, một được xuất bản bởi Open Mobile Alliance (OMA). Hai cơ quan tiêu chuẩn này hợp tác để định nghĩa các giao thức MMS. Khi nói đến MMS là nói đến các giao thức liên quan. Có tất cả mười một loại giao thức trong mô hình kiến trúc của MMS (NowSMS, 2014).

a. MM1 là giao thức được sử dụng giữa thiết bị di động với Trung tâm tin nhắn MMS (MMSC). Nó định nghĩa cách thức mà một điện thoại di động gửi và nhận tin nhắn thông qua MMSC.

b. MM2 là giao thức nằm giữa MMS server và MMS relay.

c. MM3 là giao thức được sử dụng giữa trung tâm MMS và các hệ thống tin nhắn khác. Giao thức này thông qua môi trường Internet để kết nối với server bên ngoài. Trên thực tế, giao thức này chủ yếu được thực hiện thông qua giao thức email SMTP.

d. MM4 là giao thức được sử dụng để kết nối hai trung tâm MMS lại với nhau. Giao thức này cần thiết cho việc trao đổi tin nhắn đa phương tiện giữa các môi trường MMS riêng biệt (như giữa hai mạng di động khác nhau).

e. MM5 là giao thức cho phép tác động qua lại giữa trung tâm MMS và các thành phần mạng khác như bộ ghi định vị thường trú HLR hoặc một DNS.

f. MM6 là giao thức cho phép tương tác giữa trung tâm MMS và cơ sở dữ liệu người dùng.

g. MM7 là giao thức được sử dụng để cho phép các ứng dụng của nhà cung cấp dịch vụ giá trị gia tăng (VASP) gửi và nhận tin nhắn MMS thông qua một MMSC. Giao thức MM7 được định nghĩa hoàn chỉnh bởi 3GPP, và là một giao thức dựa trên SOAP.

h. MM8 là giao thức được sử dụng giữa trung tâm MMS và một hệ thống thanh toán trả sau.

i. MM9 là giao thức được sử dụng giữa trung tâm MMS và hệ thống trả trước trực tuyến.

j. MM10 là giao thức cho phép tương tác giữa trung tâm MMS và một cơ quan kiểm soát dịch vụ tin nhắn (MSCF).

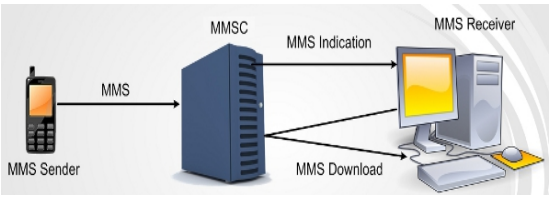
k. EAIF là giao thức độc quyền được định nghĩa bởi NOKIA, là giao thức mở rộng của giao thức MM1 vì thế nó có thể được sử dụng cho các nhà cung cấp dịch vụ giá trị gia tăng.

Trong khuôn khổ bài viết, chúng tôi không đi sâu nghiên cứu các giao thức MMS cũng như các kiến trúc chi tiết bên trong hệ thống MMS, chi tiết có thể tham khảo tại (NowSMS, 2014).

Cách nhận tin nhắn MMS: Về cơ bản, việc nhận tin nhắn từ máy tính có thể được thực hiện bằng hai phương thức kết nối khác nhau.

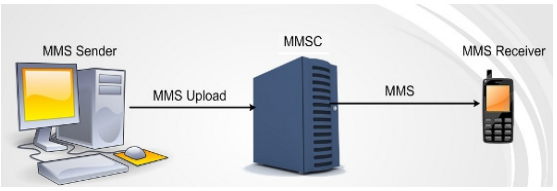
Cách 1: Tin nhắn MMS được nhận qua kết nối trực tiếp tới Trung tâm tin nhắn MMS của nhà mạng bằng cách sử dụng một trong những giao thức được hỗ trợ, bao gồm MM4, MM7, hoặc EAIF. Khi sử dụng bất kì giao thức nào trong các giao thức này, trung tâm tin nhắn của nhà mạng sẽ tự động kết nối đến MMS Gateway để lấy tin nhắn.

Cách 2: Tin nhắn MMS có thể được lấy về bằng cách sử dụng công nghệ SMS kết hợp với công nghệ WAP. Để nhận được một tin nhắn MMS cần trải qua hai giai đoạn. Giai đoạn một, modem nhận tin nhắn SMS, còn gọi là tin nhắn thông báo MMS. Tin nhắn này chứa URL của tin nhắn MMS trên Trung tâm tin nhắn đa phương tiện (MMSC). Giai đoạn hai, khi modem đã nhận được tin nhắn thông báo MMS, modem mở kết nối GPRS đến Wap Gateway để tải về nội dung tin nhắn MMS về từ trung tâm tin nhắn đa phương tiện.



Hình 2: Mô hình nhận tin nhắn MMS từ ứng dụng

Cách gửi tin nhắn MMS: Để gửi tin nhắn MMS thì ứng dụng khởi tạo một kết nối GPRS đến Wap Gateway của nhà mạng và thực hiện gửi tin nhắn MMS đến Trung tâm tin nhắn MMS (MMSC) thông qua kết nối WAP và GPRS.



Hình 3: Mô hình gửi tin nhắn MMS từ ứng dụng

Để thuận lợi và không mất nhiều thời gian cho việc xây dựng và phát triển hệ thống, nghiên cứu sử dụng thư viện SMSLIB (SMSLib, 2014) để gửi và nhận tin nhắn SMS và tin nhắn thông báo MMS, thư viện jWAP (jWAP, 2014) để kết nối đến Wap Gateway nhà mạng và JMMSLIB (jMmsLib, 2014) để giải mã tin nhắn MMS. Việc gửi và nhận tin

nhắn được thực hiện thông qua một modem 3G được kết nối trực tiếp với máy tính.

2.3 Xây dựng mô-đun phân loại tin nhắn SMS

2.3.1 Bài toán phân lớp (classification)

Là quá trình phân lớp một đối tượng dữ liệu vào một hay nhiều lớp cho trước nhờ một mô hình phân lớp mà mô hình này được xây dựng dựa trên một tập hợp các đối tượng dữ liệu đã được gán nhãn từ trước gọi là tập dữ liệu học (training data). Quá trình phân lớp còn được gọi là quá trình gán nhãn cho các đối tượng dữ liệu. Như vậy, nhiệm vụ của bài toán phân lớp dữ liệu là cần xây dựng mô hình (bộ) phân lớp để khi có một dữ liệu mới vào thì mô hình phân lớp sẽ cho biết dữ liệu đó thuộc lớp nào. Có nhiều bài toán phân lớp dữ liệu như phân lớp nhị phân, phân lớp đa lớp, phân lớp đa trị,…

Quá trình phân lớp dữ liệu thường gồm hai bước: bước "***xây dựng mô hình***" và bước "***sử dụng mô hình***".

Bước 1: Một mô hình sẽ được xây dựng dựa trên việc phân tích các đối tượng dữ liệu đã được gán nhãn từ trước. Tập các mẫu dữ liệu này còn được gọi là *tập dữ liệu huấn luyện* (*training data set*). Các nhãn lớp của tập dữ liệu huấn luyện được xác định bởi con người trước khi xây dựng mô hình. Trong bước này, chúng ta còn phải tính độ chính xác của mô hình bằng cách sử dụng một tập dữ liệu khác để kiểm tra, tập dữ liệu này gọi là *tập dữ liệu kiểm tra* (*test data set*) hoặc dùng phương pháp kiểm tra chéo trên tập dữ liệu huấn luyện. Nếu độ chính xác là chấp nhận được, mô hình sẽ được sử dụng để xác định nhãn lớp cho các dữ liệu khác mới trong tương lai.

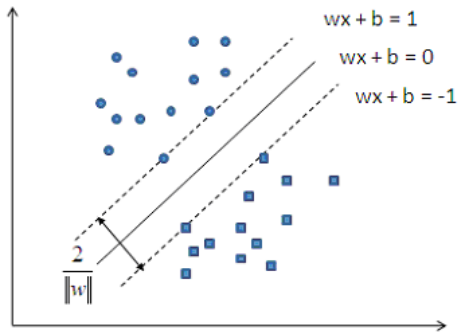
Bước 2: sử dụng mô hình đã được xây dựng ở bước 1 để phân lớp dữ liệu mới.

Như vậy, thuật toán phân lớp là một ánh xạ từ miền dữ liệu đã có sang một miền giá trị cụ thể của thuộc tính lớp, dựa vào giá trị các thuộc tính của dữ liệu.

2.3.2 Máy học Véc tơ hỗ trợ (SVM)

Phương pháp SVM ra đời từ lý thuyết học thống kê do Vapnik (1995) xây dựng, SVM có nhiều tiềm năng phát triển về mặt lý thuyết cũng như ứng dụng trong thực tiễn. SVM được đánh giá là một trong mười giải thuật quan trọng của khai mỏ dữ liệu. Các ứng dụng thực tế cho thấy phương pháp SVM có khả năng phân loại khá tốt đối với bài toán phân loại văn bản cũng như trong nhiều ứng dụng khác (như nhận dạng chữ viết tay, phát hiện mặt người trong các ảnh, ước lượng hồi quy,...).

Bài toán cơ bản của SVM là bài toán phân loại hai lớp: Cho trước n điểm trong không gian d chiều (mỗi điểm thuộc vào một lớp kí hiệu là +1 hoặc –1, mục đích của giải thuật SVM là tìm một siêu phẳng (hyperplane) phân hoạch tối ưu cho phép chia n điểm này thành hai phần sao cho các điểm cùng một lớp nằm về một phía với siêu phẳng này.



Hình 4: Phân lớp tuyến tính với SVM

Xét tập dữ liệu mẫu có thể tách rời tuyến tính $\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$ với $x_i \in R^d$ và $y_i \in \{\pm 1\}$. Siêu phẳng tối ưu phân tập dữ liệu này thành hai lớp là siêu phẳng có thể tách rời dữ liệu thành hai lớp riêng biệt với lề (margin) lớn nhất. Tức là, cần tìm siêu phẳng H: y = w.x + b = 0 và hai siêu phẳng H_1, H_2 hỗ trợ song song với H và có cùng khoảng cách đến H. Với điều kiện không có phần tử nào của tập mẫu nằm giữa H_1 và H_2, khi đó:

w.x + b >= +1 với y = +1

w.x + b >= –1 với y = –1

Kết hợp hai điều kiện trên ta có:

y(w.x + b) >= 1

Khoảng cách của siêu phẳng H_1 và H_2 đến H là ||w||. Ta cần tìm siêu phẳng H với lề lớn nhất, tức là giải bài toán tối ưu tìm $\min_{w,b}\|w\|$ với ràng buộc y(w.x + b) >= 1. Người ta có thể chuyển bài toán sang bài toán tương đương nhưng dễ giải hơn là $\min_{w,b}\frac{1}{2}\|w\|^2$ với ràng buộc y(w.x + b) >= 1. Lời giải cho bài toán tối ưu này là cực tiểu hóa hàm Lagrange:

$$L(w,b,a)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i(<w,x_i>+b)-1) \quad (1)$$

Trong đó α là các hệ số Lagrange, α≥0. Sau đó người ta chuyển thành bài toán đối ngẫu là cực đại hóa hàm W(α):

$$\max_{\alpha} W(\alpha) = \max_{\alpha} (\min_{w,b} L(w,b,a)) \quad (2)$$

Từ đó giải để tìm được các giá trị tối ưu cho w,b và α. Về sau, việc phân loại một mẫu mới chỉ là việc kiểm tra hàm dấu sign(wx +b).

Giải thuật SVM cơ bản giải quyết được bài toán phân lớp tuyến tính, tuy nhiên nếu ta kết hợp SVM với phương pháp hàm nhân (kernel-based method), sẽ cho phép giải quyết một số bài toán phi tuyến bằng cách ánh xạ dữ liệu vào một không gian có số chiều lớn hơn. Không có bất kỳ thay đổi cần thiết nào về mặt giải thuật, việc duy nhất cần làm là thay thế các tích vô hướng của hai véc-tơ u.v bởi hàm nhân K(u,v).

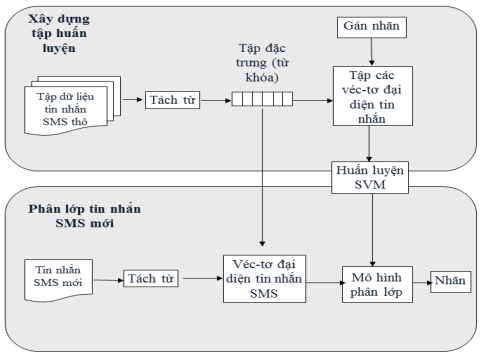
Bảng 3: Một số hàm nhân thường được dùng

Kiểu hàm	Công thức
Tuyến tính	$K(u,v) = u.v$
Đa thức bậc d	$K(u,v) = (u.v +c)^d$
Radial Basis Function	$K(u,v) = \exp(-\gamma\|u-v\|^2)$

Trong nghiên cứu này, chúng tôi thực hiện xây dựng và huấn luyện mô hình phân loại tin nhắn SMS nhờ vào sự hỗ trợ của công cụ LibSVM (Chang, C.C., Lin, C.J., 2011).

2.3.3 Phân loại tin nhắn văn bản (SMS)

Phân lớp văn bản được định nghĩa là việc gán tên các chủ đề (tên lớp/nhãn lớp) cho trước vào các văn bản dựa trên nội dung của nó. Phân lớp văn bản là công việc được sử dụng để hỗ trợ cho quá trình tìm kiếm thông tin (Information Retrieval), chiết lọc thông tin (Information Extraction), lọc văn bản hoặc tự động dẫn đường cho các văn bản tới những chủ đề xác định trước đó. Để xây dựng bộ phân lớp văn bản tự động, người ta sử dụng các thuật toán máy học (machine learning) có giám sát.



Hình 5: Mô hình phân lớp tin nhắn văn bản với SVM

Xây dựng mô hình phân loại tin nhắn

Chọn lớp (target class): Trong nghiên cứu này chúng tôi chia chuyên ngành lúa thành sáu chuyên ngành nhỏ hơn (Nguyễn Ngọc Đệ, 2008) được mô tả như trong Bảng 4.

Bảng 4: Các chuyên ngành nhỏ trên cây lúa

STT	Tên
1	Bệnh hại lúa
2	Sâu hại lúa
3	Cỏ hại lúa
4	Giống lúa
5	Kỹ thuật canh tác
6	Sau thu hoạch

Dựa vào đó hệ thống phân làm sáu lớp (nhãn) tương ứng với sáu chuyên ngành nhỏ này. Các chuyên gia tham gia vào hệ thống thuộc một hoặc nhiều chuyên ngành trong sáu chuyên ngành này. Như vậy, ta sẽ xây dựng tập dữ liệu huấn luyện với sáu nhãn tương ứng. Ngoài ra, ta cũng thêm một lớp là lớp tin nhắn rác nếu nó không thuộc một trong sáu lớp trên, đây là một dạng bài toán phân lớp đa lớp.

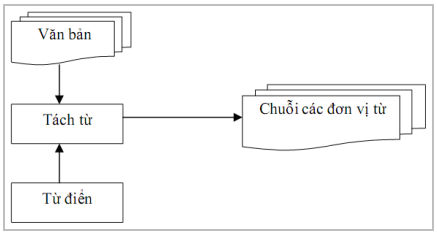
Tách từ tiếng việt

Để xây dựng hệ thống phân loại văn bản, việc tách văn bản thành từ độc lập có nghĩa là việc bắt buộc. Đối với văn bản Tiếng Anh, việc tách từ được thực hiện khá đơn giản vì mỗi từ Tiếng Anh phân biệt nhau bởi một khoảng trắng. Ngược lại, việc tách từ Tiếng Việt lại khá phức tạp vì một từ tiếng việt có thể có hoặc không có khoảng trắng. Có nhiều cách tiếp cận xử lý tách từ Tiếng Việt chẳng hạn như phương pháp dùng từ điển, phương pháp thống kê… Kỹ thuật tách từ Tiếng Việt cũng đã được nhiều nhóm tác giả nghiên cứu và xây dựng cho kết quả với độ chính xác khá cao.

Nghiên cứu sử dụng phần mềm tách từ VnTokenizer[1] để thực hiện việc tách tin nhắn thành các từ độc lập, công cụ này được phát triển dựa trên phương pháp so khớp tối đa (Maximum matching) với tập dữ liệu sử dụng là bảng âm tiết tiếng Việt và từ điển từ vựng tiếng Việt. Công cụ được xây dựng trên ngôn ngữ Java, mã nguồn mở. Có thể dễ dàng tích hợp vào các hệ thống phân tích tiếng Việt khác.

Quy trình thực hiện tách từ theo phương pháp so khớp tối đa được minh họa như trong Hình 6.

[1] http://mim.hus.vnu.edu.vn/phuonglh/node/33



Hình 6: Tách từ theo phương pháp so khớp tối đa

Các đơn vị từ được sinh ra từ công cụ này bao gồm các từ trong từ điển, các chuỗi số, chuỗi ký tự nước ngoài, dấu câu, các ký tự hỗn tạp khác trong văn bản, các từ mới hoặc các từ được sinh tự do theo một quy tắc nào đó (như phương thức thêm phụ tố hay phương thức láy) hoặc các chuỗi kí hiệu không được liệt kê trong từ điển. Công cụ này tách từ cho độ chính xác là 96% - 98% theo (Phương *et al.*, 2010). Ví dụ sau minh họa một đoạn văn bản được tách từ bởi VnTokenizer.

Ví dụ: văn bản gốc

Bệnh gây hại chủ yếu giai đoạn mạ – đẻ nhánh. Lúc đầu vết bệnh chỉ là những chấm nhỏ, màu xanh xám, sau lớn lên có dạng hình thoi (mắt én) đặc trưng.
Nhiệm vụ của bào tử này là hút các chất dinh dưỡng có trong cây lúa và ngoài ra còn tiết ra độc tố Pyricularin gây độc cho cây.
Bào tử nấm Pyricularia oryzae hay P. grisea phát triển tốt trong điều kiện nhiệt độ mát từ 24 – 28 độ C , ẩm độ cao trên 80%.

Ví dụ: Văn bản sau khi tách từ

Bệnh gây hại chủ_yếu giai_đoạn mạ – đẻ_nhánh . Lúc_đầu vết bệnh chỉ_là những chấm nhỏ , màu_xanh xám , sau lớn_lên có dạng hình_thoi (mắt én) đặc_trưng .
Nhiệm_vụ của bào_tử này là hút các chất dinh_dưỡng có trong cây lúa và ngoài_ra còn tiết ra độc_tố Pyricularin gây độc cho cây .
Bào_tử nấm Pyricularia_oryzae hay P. grisea phát_triển tốt trong điều_kiện nhiệt_độ mát từ 24 – 28 độ C , ẩm_độ cao trên 80% .

Như ta đã biết một từ tiếng Việt trong văn bản thông thường có thể có hoặc không có khoảng trắng (một hoặc nhiều âm tiết), do vậy với văn bản gốc ban đầu thì không thể phân biệt từ nào là từ đầy đủ trong từ điển, từ nào chỉ là một âm tiết (một phần của từ). Sau khi tách từ, các từ bao gồm nhiều âm tiết sẽ được nối lại với nhau bằng cách sử dụng ký tự "_" thay thế cho ký tự khoảng trắng. Với cách làm như vậy thì sau khi thực hiện tách từ, các từ (có nghĩa trong từ điển) sẽ được phân biệt nhau nhờ khoảng trắng giống như văn bản tiếng Anh. Từ đó, ta sẽ dễ dàng xây dựng bộ từ khóa cũng như xây dựng mô hình phân loại với văn bản đã tách.

Xây dựng bộ từ khóa (đặc trưng)

Bộ từ khóa đặc trưng là một danh sách các từ khóa đặc trưng cho sáu chuyên ngành lúa như trong Bảng 4 trên. Bước chọn từ khóa là một bước quan trọng quyết định nhiều đến kết quả phân loại của hệ thống. Trong nghiên cứu này, chúng tôi đề xuất hai phương án xây dựng bộ từ khóa:

- *Xây dựng bộ từ khóa thủ công:* Với phương án này thì cần có sự hỗ trợ của các chuyên gia về lúa, độ tin cậy của các từ khóa được chọn lọc cao hơn, số lượng từ khóa ít nhưng chất lượng hơn, điểm yếu của phương án này là mất nhiều thời gian và công sức để xây dựng và chọn lọc từ khóa.
- *Xây dựng bộ từ khóa tự động*: Từ tập dữ liệu thu thập được hệ thống sẽ thực hiện việc tách từ và loại bỏ từ dừng, do văn bản là tin nhắn nên số lượng từ khóa không nhiều và ít khi lặp lại nên hệ thống không thực hiện việc giảm số chiều (số đặc trưng) mà dùng tất cả các từ còn lại sau khi loại bỏ từ dừng để xây dựng bộ từ khóa, với phương án tự động thì việc xây dựng bộ từ khóa được thực hiện nhanh hơn và không mất nhiều công sức cũng như không cần sự trợ giúp của các chuyên gia nhưng chất lượng các từ khóa thì không cao vì không phải từ nào được giữ lại cũng có ý nghĩa phân loại, số lượng từ khóa sẽ nhiều hơn rất nhiều so với phương án thủ công.

Danh sách từ khóa được lưu vào một tập tin, mỗi dòng của tập tin một là từ khóa như ví dụ sau:

Đất
Giống
Nước
Sâu
... ...
Đạm
... ...

Biểu diễn tin nhắn bằng vec-tơ đặc trưng

Trong nghiên cứu này, chúng tôi sử dụng SVM để phân loại tin nhắn văn bản do SVM có nhiều ưu

điểm khi sử dụng cho phân loại văn bản (Đỗ Thanh Nghị, 2011; Trần Cao Đệ và Phạm Nguyên Khang , 2012). Định dạng từng dòng của tập tin huấn luyện như sau:

<label><index1>:<value1><index2>:<value2>...

Với ***<label>*** là nhãn phân lớp của văn bản (6 chuyên ngành về lúa như trong Bảng 4), ***<index>*** là chỉ số của từ khóa, chỉ số này tương ứng với số thứ tự của từ khóa trong tập tin từ khóa, ***<value>*** là giá trị trọng số của từ khóa. Mặc dù, có nhiều cách xác định giá trị trọng số của từ khóa, trong nghiên cứu này, do văn bản là tin nhắn SMS nên số lượng từ khóa không nhiều và ít khi lặp lại nên khi véc-tơ hóa ta không quan tâm từ khóa đó xuất hiện bao nhiêu lần mà chỉ cần quan tâm nó có xuất hiện hay không, nếu có xuất hiện thì phần giá trị trọng số được gán là 1, nếu không xuất hiện thì không cần phải lưu - định dạng này còn được gọi là định dạng thưa. Ví dụ về định dạng tập tin huấn luyện với dữ liệu là các tin nhắn.

Tin nhắn sau khi tách từ và loại bỏ từ dừng:

cho_biết phòng trị bệnh vàng lùn lúa cỏ

bệnh bạc lá lúa nguyên_nhân phòng_tránh

Định dạng tập tin huấn luyện của các tin nhắn trên như sau:

1 63:1 95:1 167:1 419:1 420:1 629:1 858:1 948:1
1 56:1 63:1 414:1 420:1 524:1 630:1

Như ta đã thấy, tin nhắn thứ nhất và thứ hai thuộc lớp sâu bệnh hại lúa, như vậy nhãn (lớp) của các tin nhắn này trong tập tin định dạng là 1. Với từ khóa "*cho_biết*" là từ khóa nằm ở vị trí thứ 63 trong tập tin từ khóa, như vậy để thể hiện một từ có trong bộ từ khóa và vị trí của nó là thứ 63 trong bộ từ khóa ta viết 63:1, thực hiện tương tự với các từ còn lại trong tin nhắn và với tất cả tin nhắn nhận được. Chú ý rằng thứ tự các từ trong tin nhắn không quan trọng, khi xác định trọng số các từ, ta viết theo thứ tự từ nhỏ đến lớn.

Như đã trình bày, việc xây dựng bộ từ khóa đặc trưng trong nghiên cứu này sử dụng hai phương án, phương án thủ công và phương án tự động. Như vậy, để xây dựng tập huấn luyện ta cũng cần phải áp dụng hai phương án này. Nếu như hệ thống được người sử dụng cấu hình là sử dụng phương án thủ công thì khi xây dựng tập huấn luyện, hệ thống sẽ chọn từ khóa trong tin nhắn sau khi tách từ bằng cách chỉ giữ lại những từ có trong danh sách từ khóa đã được chọn thủ công bởi các chuyên gia và thực hiện véc-tơ các từ đó và lưu vào tập tin huấn luyện theo phương án này. Nếu như hệ thống được người sử dụng cấu hình là sử dụng phương án tự động thì khi xây dựng tập huấn luyện, hệ thống sẽ làm việc loại bỏ các từ dừng (ngược lại so với phương án thủ công là chọn từ khóa trong bộ từ khóa) sau khi thực hiện tách từ, sau đó véc-tơ hóa toàn bộ các từ còn lại và lưu vào tập tin huấn luyện theo phương án này.

Một vấn đề quan trọng cần quan tâm khi xây dựng tập dữ liệu là thói quen nhắn tin tiếng Việt không bỏ dấu của người dùng, do vậy trong quá trình xây dựng tập dữ liệu và bộ từ khóa nếu ta chỉ sử dụng tiếng Việt có dấu thì sẽ làm cho kết quả phân loại trở nên không chính xác mặc dù nội dung tin nhắn có chứa từ khóa cần thiết cho phân loại, chỉ có khác là từ khóa đó không có dấu tiếng Việt.

Để giảm sai sót trong phân loại tin nhắn, chúng tôi đề xuất một giải pháp để xây dựng bộ từ khóa và tập tin huấn luyện trong mô hình phân loại tin nhắn. Do bộ từ khóa được xây dựng dựa trên hai cách như trên đã đề cập, với phương án xây dựng bộ từ khóa thủ công thì khi cập nhật thông tin cho từng chuyên ngành trong sáu chuyên ngành lúa trên ta cũng đồng thời cập nhật các từ khóa đặc trưng cho từng chuyên ngành, các từ khóa được sử dụng bao gồm cả các từ có dấu tiếng Việt, không có dấu tiếng Việt, các từ gần đúng với từ khóa nhưng không có nghĩa khác (phòng trường hợp người dùng viết sai chính tả,…), các từ khác nhau nhưng cùng nghĩa, các từ địa phương… Việc xây dựng tập dữ liệu được thực hiện bằng cách với một tin nhắn tiếng Việt có dấu đã được phân loại, hệ thống sẽ tự động tạo ra thêm một tin nhắn tiếng Việt không có dấu và lưu vào cơ sở dữ liệu để sử dụng cho việc xây dựng lại mô hình. Với cách chọn từ khóa theo phương án tự động dựa vào tập dữ liệu nên chỉ cần xây dựng tập dữ liệu như ở trên đã đề xuất là ta sẽ có một bộ từ khóa gồm cả từ có dấu lẫn từ không dấu tiếng Việt.

Tương tự việc xây dựng tập tin huấn luyện, ta xây dựng tập tin kiểm tra để kiểm tra độ chính xác của mô hình.

Sử dụng mô hình

Sau khi kiểm tra độ chính xác phân lớp của mô hình, nếu độ chính xác chấp nhận được, ta đưa mô hình vào sử dụng để phân loại các tin nhắn mới. Tương tự như quá trình xây dựng tập huấn luyện, tin nhắn mới đến hệ thống sẽ được tiền xử lý, tách từ và véc-tơ hóa theo định dạng giống như định dạng của tập tin huấn luyện với phương án mà

người sử dụng cấu hình hệ thống đã chọn. Chỉ có một điểm khác là nhãn phân lớp của tin nhắn mới này là nhãn tạm cho tin nhắn mới, sau khi đưa vào mô hình phân loại, nhãn tạm này sẽ được tự động thay thế bằng nhãn chính thức, định dạng của tập tin cần phân loại như sau:

<templabel> <index1>:<value1> <index2>:<value2> ...

2.4 Xây dựng website tích hợp và quản lý cấu hình hệ thống

Website hệ thống được xây dựng theo mô hình MVC trên nền ngôn ngữ Java (JSP và Servlet) với hệ quản trị cơ sở dữ liệu MySQL. Hệ thống quản lý hai nhóm đối tượng người dùng là chuyên gia và quản trị/điều phối viên.

Chức năng của nhóm chuyên gia:

– Đăng ký, xem, chỉnh sửa thông tin cá nhân.

– Trả lời các câu hỏi của nhà nông liên quan đến chuyên môn.

– Phân loại lại tin nhắn nếu có sai sót.

Chức năng của nhóm quản trị/ điều phối viên:

– Cập nhật các lớp (chuyên ngành) cần phân loại

– Cập nhật danh sách chuyên gia

– Cập nhật, phân loại tin nhắn

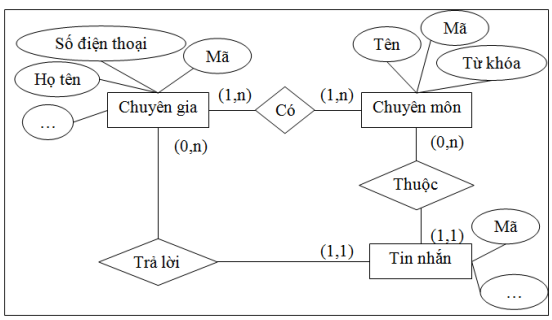
– Cập nhật thông tin Gateway

– Cấu hình hệ thống

Thống kê/báo cáo

Mô hình thực thể kết hợp (ERD)

Các thực thể chính của hệ thống bao gồm: Chuyên gia, chuyên môn và tin nhắn. Một chuyên gia có thể có nhiều chuyên môn, một chuyên môn có thể có nhiều chuyên gia đăng ký. Một chuyên gia có thể trả lời nhiều tin nhắn, một tin nhắn chỉ được trả lời bởi một chuyên gia. Hình 7 minh họa các thực thể của hệ thống được kết hợp lại theo mô hình thực thể kết hợp (ERD).



Hình 7: Mô hình thực thể kết hợp (ERD)

3 KẾT QUẢ THỰC NGHIỆM

Số lượng từ khóa bước đầu chúng tôi xây dựng để minh họa cho phương án chọn từ khóa được xây dựng thủ công là 243 từ, các từ khóa thuộc 06 chuyên ngành như đã đề cập, là các từ đặc trưng nhất, có ý nghĩa trong phân loại được chúng tôi cùng với sự trợ giúp của một số chuyên gia chọn lựa và xây dựng một cách thủ công. Tập dữ liệu ban đầu chúng tôi tự xây dựng gồm 200 câu hỏi (có dấu tiếng Việt) và 200 câu hỏi không có dấu tiếng Việt được hệ thống tự động sinh ra từ 200 câu hỏi có dấu. Để xác định nhãn (phân lớp chuyên ngành) cho các tin nhắn này thì chúng tôi dùng mô-đun phân loại bán tự động để gán thủ công. Với phương án chọn từ khóa tự động thì sau khi thực hiện các bước cần thiết để xây dựng bộ từ khóa trên tập dữ liệu gồm 400 tin nhắn câu hỏi như trên, chúng tôi thu được bộ từ khóa gồm 1044 từ, là những từ được giữ lại sau khi loại bỏ từ dừng, các ký tự đặc biệt và ký tự số không có ý nghĩa trong phân loại.

Để phân loại tin nhắn bằng SVM, chúng tôi sử dụng bộ thư viện LibSVM (Chang *et al.*, 2011). Bằng nghi thức kiểm tra chéo (10-folds) trên tập học, mô-đun phân loại tin nhắn cho độ chính xác đạt 69,94%. Với phương án tự động cho kết quả chính xác đạt 71,9%.

Do đang trong giai đoạn nghiên cứu và thu thập dữ liệu, nhóm tác giả chỉ mới thực hiện kiểm tra độ chính xác trên bộ dữ liệu thu thập và xây dựng được, nhóm tác giả chưa thực hiện kiểm tra độ chính xác với tin nhắn ngoài thực tế. Sau khi tin nhắn chứa câu hỏi được phân loại, nó được chuyển đến chuyên gia thích hợp để trả lời. Khi nhận được câu trả lời từ chuyên gia, hệ thống tự động gửi nội dung trả lời cho nhà nông.

3.1.1 Trang chủ hệ thống

Chuyên gia và quản trị đăng nhập vào hệ thống thông qua trang chủ. Sau khi đăng nhập thì tùy loại người dùng sẽ có những danh mục chức năng riêng như trên đã đề cập.

Hình 8: Trang chủ website hệ thống

3.1.2 Trang phân loại nội dung tin nhắn (bán tự động)

Có chức năng hiển thị các tin nhắn đã được phân loại bởi mô-đun phân loại tự động nhưng chưa được phân loại bởi quản trị/điều phối viên. Điều phối viên đăng nhập vào trang này để thực hiện phân loại (gán nhãn) tin nhắn.

Hình 9: Trang phân loại nội dung tin nhắn

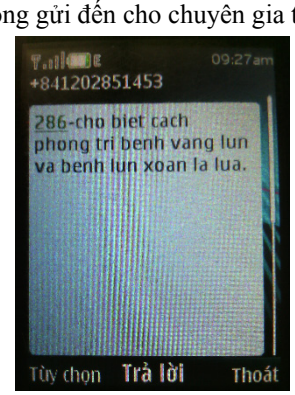
3.1.3 Trang cấu hình hệ thống

Cho phép thay đổi các thông số cấu hình hệ thống như thời gian hệ thống lặp lại việc truy vấn và huấn luyện lại mô hình, số lượng tin nhắn mới để thực hiện huấn luyện lại,…

Hình 10: Trang cấu hình hệ thống

3.1.4 Màn hình điện thoại chuyên gia

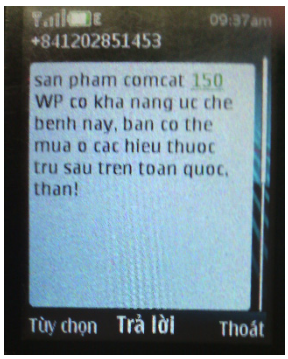
Sau khi hệ thống nhận được câu hỏi của nhà nông, hệ thống tiến hành phân loại câu hỏi và gửi câu hỏi đến cho chuyên gia trả lời. Câu hỏi đi kèm một mã số để xác định tin nhắn trong cơ sở dữ liệu tin nhắn. Hình 11 minh họa nội dung tin nhắn SMS được hệ thống gửi đến cho chuyên gia trả lời.



Hình 11: Tin nhắn được hệ thống gửi đến cho chuyên gia trả lời

3.1.5 Màn hình điện thoại nhà nông

Sau khi hệ thống nhận được câu trả lời của chuyên gia, hệ thống sẽ tự động gửi câu trả lời đến điện thoại nhà nông. Hình 12 minh họa nội dung tin nhắn SMS được hệ thống gửi cho nhà nông để giải đáp thắc mắc của nhà nông.



Hình 12: Tin nhắn được hệ thống gửi cho nhà nông để giải đáp thắc mắc của nhà nông

4 KẾT LUẬN VÀ ĐỀ XUẤT

Trong nghiên cứu này chúng tôi đã tìm hiểu và trình bày khái quát một số kiến thức về hệ thống thông tin di động, cách thức để cấu hình hệ thống, gửi và nhận tin nhắn từ máy vi tính đến điện thoại di động, các kỹ thuật phân loại văn bản và tin nhắn văn bản bằng SVM.

Chúng tôi đã tham khảo và tìm hiểu cách sử dụng một số thư viện, phần mềm ứng dụng trong hệ thống như SMSLib, jWAP, jMmsLib, LibSVM. Chúng tôi cũng xây dựng thủ công được bộ từ khóa gồm 243 từ thuộc các chuyên ngành lúa để minh họa cho việc phân loại văn bản tự động bằng SVM theo phương án thủ công, và một tập dữ liệu huấn luyện bước đầu của hệ thống gồm 200 tin nhắn SMS tham khảo tại (Thư viện KHCN Quảng Trị, 2014; Cty CP Phân bón Bình Điền, 2014; Bạn nhà nông, 2014; Sở KHCN Vĩnh Phúc, 2014). Ngoài ra, chúng tôi đã đề xuất các phương án để xây dựng bộ từ khóa và tập dữ liệu để nâng cao độ chính xác cho mô hình phân loại tin nhắn tiếng Việt không dấu do nhắn tin tiếng Việt không dấu thì dễ dàng và thuận lợi hơn cho nhà nông, nhắn được nhiều ký tự hơn (tiếng Việt không dấu là 160 ký tự/1 tin nhắn, tiếng Việt có dấu là 70 ký tự/1 tin nhắn) và nhà nông không cần phải có điện thoại cấu hình cao mới có thể đặt câu hỏi bằng SMS.

Trên nền tảng đó, chúng tôi đã phát triển được một ***hệ thống hỗ trợ khuyến nông trên cây lúa qua mạng thông tin di động*** với các chức năng cơ bản nhằm hỗ trợ về thông tin kỹ thuật cho nhà

nông một cách kịp thời, với chi phí hợp lý nhất, giải quyết được tương đối đầy đủ những thắc mắc, nhu cầu của nhà nông. Sau một thời gian vận hành và khai thác một cách bán tự động trên thực tế, hệ thống sẽ thu thập được một lượng dữ liệu thực cần thiết, kết hợp với công nghệ khai phá dữ liệu, xử lý văn bản và xử lý ảnh để xây dựng một hệ thống hỗ trợ khuyến nông trên cây lúa qua mạng thông tin di động một cách hoàn toàn tự động. Với kết quả bước đầu như đã trình bày trong phần kết quả cho thấy việc xây dựng và triển khai hệ thống này là hoàn toàn khả thi và có tiềm năng phát triển rộng rãi.

Hiện tại, chúng tôi xây dựng mô-đun gửi nhận tin nhắn đa phương tiện giới hạn chỉ với hai nhà mạng di động là Mobifone và Vinaphone. Việc phát triển hệ thống để hỗ trợ nhiều nhà mạng hơn sẽ được phát triển trong các giai đoạn sau. Giới hạn về số lượng ký tự trong một tin nhắn và việc người dùng có thói quen nhắn tin tiếng Việt không có dấu làm cho việc phân loại tin nhắn văn bản trở nên khó khăn hơn.

Nghiên cứu này làm tiền đề cho việc hoàn thiện hệ thống tự động hỗ trợ khuyến nông trên cây lúa qua mạng thông tin di động. Những công việc cần tiếp tục thực hiện gồm:

– Phát triển hệ thống để hỗ trợ gửi và nhận tin nhắn đa phương tiện (MMS) của tất cả các nhà mạng hiện hành ở Việt Nam.

– Phát triển mô-đun phân loại tin nhắn hình ảnh tự động bằng các kỹ thuật xử lý ảnh và máy học.

– Xây dựng một hệ thống hoàn toàn tự động dựa vào sự kết hợp hai kỹ thuật phân loại, phân loại tin nhắn văn bản và phân loại tin nhắn hình ảnh.

– Hoàn thiện thêm bộ từ khóa cho phân loại văn bản, xây dựng mô hình phân loại với độ chính xác cao hơn nhờ vào nguồn dữ liệu thực tế mà hệ thống thu thập được.

– Hoàn thiện thêm tập dữ liệu để có thể phân loại tin nhắn SMS không gõ dấu tiếng Việt được hiệu quả hơn.

– Phát triển mô hình sang nhiều lĩnh vực, nhiều chuyên ngành, chuyên môn khác ngoài cây lúa.

TÀI LIỆU THAM KHẢO

1. Vapnik, V. 1995. The Nature of Statistical Learning Theory. Springer, NewYork.
2. Chang, C.C., Lin, C.J., 2011. LIBSVM – a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm (Truy cập: 3/2014).
3. Nguyễn Ngọc Đệ, 2008. Giáo trình cây lúa, Viện Nghiên cứu Phát triển Đồng bằng Sông Cửu Long, Đại học Cần Thơ.
4. L. H. Phương, N. T. M Huyền và TV. Ho, 2010. A hybrid approach to word segmentation of Vietnamese texts, http://mim.hus.vnu.edu.vn/phuonglh/node/33 (Truy cập: 01/2014).
5. Đỗ Thanh Nghị, 2011. Khai mỏ dữ liệu – minh họa bằng ngôn ngữ R, NXB Đại học Cần Thơ.
6. Trần Cao Đệ và Phạm Nguyên Khang, 2012. Phân loại văn bản với Máy học vector hỗ trợ và Cây quyết định", Tạp chí khoa học (21a), tr. 52 – 63.
7. Tổng cục thống kê, 2014a. Thống kê lao động từ 15 tuổi trở lên đang làm việc tại thời điểm 1/7 hàng năm phân theo ngành kinh tế, http://www.gso.gov.vn/default.aspx?tabid=387&idmid=3&ItemID=14227 (Truy cập: 04/2014).
8. Tổng cục thống kê, 2014b. Thống kê diện tích gieo trồng một số cây hàng năm, http://www.gso.gov.vn/default.aspx?tabid=390&idmid=3&ItemID=14110 (Truy cập: 04/2014).
9. Tổng cục thống kê, 2014c. Thống kê sản lượng một số cây hàng năm, http://www.gso.gov.vn/default.aspx?tabid=390&idmid=3&ItemID=14108 (Truy cập: 04/2014).
10. Developer's Home, 2014. How to send SMS Messages from a Computer/PC?, http://www.developershome.com/sms/howToSendSMSFromPC.asp (Truy cập: 01/2014).
11. NowSMS, 2014. Documentation, http://www.nowsms.com/doc (Truy cập: 01/2014).
12. SMSLib, 2014. A universal API for sms messaging, http://smslib.org/ (Truy cập: 01/2014)
13. jWAP, 2014. http://jwap.sourceforge.net/ (Truy cập: 02/2014).
14. jMmsLib, 2014. http://jmmslib.sourceforge.net (Truy cập: 02/2014).

15. Thư viện KHCN Quảng Trị, 2014. Các câu hỏi thường gặp, http://elib.dostquangtri.gov.vn/thuvien/Include/TVDT.asp?option=4&noidungcantim=l%C3%BAa&CSDL=6 (Truy cập: 03/2014).
16. Cty CP Phân bón Bình Điền, 2014. Các câu hỏi về lúa, http://www.binhdien.com/articlebd.php?id=158&cid=1 (Truy cập: 03/2014).
17. Bạn nhà nông, 2014. Hỏi và đáp, http://www.bannhanong.vietnetnam.net/home.php?kh=&cat_id=29 (Truy cập: 01/2014).
18. Sở KHCN Vĩnh Phúc, 2014. hỏi đáp khoa học kỹ thuật, http://123.25.71.107:82/hoidap/index.php?language=vi&nv=news&op=search&q=c%C3%A2y+l%C3%BAa&mod=all (Truy cập: 01/2014).