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

HỆ THỐNG HỖ TRỢ TƯ VẤN TUYỂN SINH ĐẠI HỌC

Nguyễn Thái Nghe[1] và Trương Quốc Định[1]
[1] *Khoa Công nghệ Thông tin & Truyền thông, Trường Đại học Cần Thơ*

Thông tin chung:
Ngày nhận: 19/09/2015
Ngày chấp nhận: 10/10/2015

Title:
A consultancy support system for university entrance test

Từ khóa:
Phân loại văn bản, phân loại tin nhắn SMS, tư vấn tự động, tìm kiếm thông tin, hệ gợi ý

Keywords:
Text classification, SMS classification, automatic question-answer, text recommendation

ABSTRACT

In this study, we propose a solution to build a semi-automatic consultancy system (a semi-automatic question-answering system) using mobile/Internet networks and machine learning approaches. To build the system, at first, we need to build modules for sending and receiving SMS/email messages. These modules are important for pupils to send their questions that need to be consulted. Next, a message classification module is built using a combination of machine learning method (e.g., SVM) and text processing technologies. Finally, a whole web-based system is conducted to integrate these modules. The initial results show that the system can classify the questions at 82.33% of accuracy, thus, the proposed approach is feasible.

TÓM TẮT

Trong bài viết này, chúng tôi đề xuất một giải pháp xây dựng Hệ thống hỗ trợ tư vấn tuyển sinh bán tự động sử dụng kết hợp các kỹ thuật trong xử lý văn bản, máy học SVM và xử lý tin nhắn SMS trong hệ thống thông tin di động. Hệ thống tư vấn này có khả năng tiếp nhận câu hỏi của thí sinh từ trang Web/email hoặc qua tin nhắn SMS, sau đó, câu hỏi sẽ được phân loại tự động bằng máy học SVM để chuyển đến chuyên gia thích hợp trong từng lĩnh vực. Sau khi có câu trả lời từ chuyên gia, hệ thống sẽ phản hồi tức thì cho thí sinh. Bên cạnh đó, ngay sau khi thí sinh đặt câu hỏi, hệ thống sẽ xử lý và tìm độ tương đồng của câu hỏi hiện tại so với các câu đã được trả lời trước đây, nhằm gợi ý cho thí sinh có thêm thông tin. Thử nghiệm trên tập dữ liệu thu thập được từ 447 câu hỏi thuộc 8 lĩnh vực thường được nhiều thí sinh quan tâm cho thấy hệ thống đạt độ chính xác 82.33%. Độ chính xác này sẽ còn được cải thiện theo thời gian khi mà lượng câu hỏi đủ lớn cho mô hình máy học, vì thế, giải pháp đề xuất này sẽ mở ra một hướng mới trong hỗ trợ tư vấn tuyển sinh.

1 GIỚI THIỆU

Gần đây do công tác tuyển sinh có nhiều thay đổi cả về nội dung lẫn hình thức nên rất nhiều thí sinh và cả gia đình khá bối rối, việc tư vấn tuyển sinh và chọn ngành học phù hợp là nhu cầu mà xã hội đang rất quan tâm. Mặc dù hàng năm, phần lớn các trường (đơn vị) đều tổ chức các đợt tư vấn cho thí sinh, tuy nhiên việc này còn phụ thuộc vào nhiều yếu tố như địa điểm, thời gian, nhân sự,.. do vậy chỉ hỗ trợ được một bộ phận thí sinh ở thành phố hoặc những thí sinh có điều kiện tham dự. Phần lớn các thí sinh ở vùng sâu vùng xa không có điều kiện tham gia. Bên cạnh đó, một số tổ chức cũng đã thiết lập các trang web để nhận và trả lời các câu hỏi của thí sinh, như: tuvantuyensinh.vn, huongnghiep.tuvantuyensinh.vn,... Tuy nhiên, các trang này đa phần là nhận câu hỏi của thí sinh sau

đó việc giải đáp cũng được tổ chức theo định kỳ chứ không trực tuyến.

Hiện tại, điện thoại di động không còn xem là một mặt hàng xa xỉ mà nó đang là phương tiện truyền/nhận thông tin tức thời và hiệu quả, đặc biệt là ở những nơi vùng xâu, vùng xa. Với giới trẻ, việc sử dụng Internet hay điện thoại để gửi tin nhắn là việc hết sức đơn giản. Chính vì thế, việc tư vấn tuyển sinh qua hệ thống tin nhắn sẽ đảm bảo tính tức thời và hiệu quả, nhằm giúp các em cập nhật thông tin, được giải đáp các câu hỏi một cách nhanh nhất trong tuyển sinh và những vấn đề liên quan. Từ những thực trạng trên, nhu cầu cần một hệ thống tư vấn tuyển sinh có thể hoạt động một cách tự động 24/7, để có thể hỗ trợ cả thí sinh lẫn gia đình là rất cần thiết. Tuy nhiên, vẫn chưa thấy có hệ thống nào có khả năng đáp ứng được các yêu cầu trên.

Trong bài viết này, chúng tôi đề xuất một giải pháp xây dựng Hệ thống hỗ trợ tư vấn tuyển sinh (bán) tự động sử dụng kết hợp các kỹ thuật trong xử lý văn bản (xử lý ngôn ngữ tự nhiên), máy học SVM, và xử lý tin nhắn SMS trong hệ thống thông tin di động. Thử nghiệm trên tập dữ liệu thu thập được từ 447 câu hỏi thuộc 8 lĩnh vực khác nhau cho thấy hệ thống đạt độ chính xác khá tốt, vì thế, giải pháp đề xuất này sẽ mở ra một hướng mới trong hỗ trợ tư vấn tuyển sinh một cách tự động.

2 KIẾN TRÚC CỦA HỆ THỐNG

Kiến trúc của hệ thống được trình bày trong Hình 1. Ở đó, khi thí sinh cần được tư vấn, họ sẽ đặt câu hỏi thông qua email, website, hoặc tin nhắn SMS. Câu hỏi này sẽ được hệ thống xử lý (như tách từ, loại bỏ từ dừng, chọn từ khóa) và đưa vào bộ phân loại SVM. Câu hỏi sau khi được phân loại sẽ được gửi tới các chuyên gia (cán bộ chuyên trách) thuộc lĩnh vực tương ứng như Giáo vụ, Tài vụ, các ngành công nghệ thông tin,... Ngay sau khi nhận được câu trả lời từ các chuyên gia hệ thống sẽ phản hồi tức thì cho thí sinh (qua email hoặc qua tin nhắn SMS tùy công cụ mà người hỏi sử dụng). Bên cạnh đó, ngay sau khi thí sinh đặt câu hỏi và trong thời gian chờ câu trả lời từ chuyên gia, hệ thống sẽ tự động hiển thị các câu hỏi tương tự mà đã có câu trả lời trước đó thông qua chức năng tìm các câu hỏi tương đồng.

Hệ thống này có khả năng làm việc song song để tiếp nhận và phản hồi các câu trả lời thông qua website, email và tin nhắn SMS, hoạt động liên tục 24/7 trong năm.

Để đáp ứng được yêu cầu của hệ thống như đã mô tả, chúng tôi tiến hành xây dựng các modules và thực hiện các công việc như sau:

– Xây dựng module quản lý, tiếp nhận và trả lời câu hỏi qua giao diện web (gồm cả việc quản lý, gửi và nhận email)

– Xây dựng module quản lý, tiếp nhận và trả lời câu hỏi qua SMS

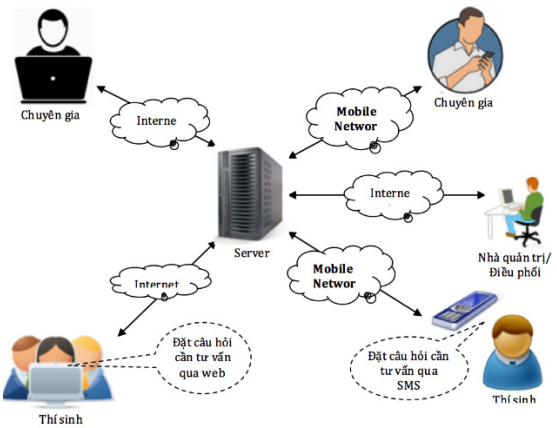
– Xây dựng module xử lý câu hỏi (tách từ, loại bỏ từ dừng, chọn từ khóa)

– Xây dựng module phân loại (tự động) câu hỏi theo từng lĩnh vực bằng kỹ thuật phân loại máy học véc-tơ hỗ trợ (SVM)

– Xây dựng module quản lý chuyên gia và nội dung phản hồi từ chuyên gia

– Xây dựng module gợi ý các câu hỏi liên quan (dùng tf-idf và độ tương đồng Cosine)

– Phân tích, thiết kế và xây dựng hệ thống (nền web) hoàn chỉnh để tích hợp các module trên.



Hình 1: Kiến trúc của hệ thống tư vấn tuyển sinh

Do tin nhắn SMS rất ngắn và cô đọng, nên số lượng từ khóa không nhiều và ít khi lặp lại, chúng tôi đề xuất ba phương án chọn từ khóa là phương án thủ công, phương án tự động và kết hợp cả 2.

– Phương án chọn từ khóa thủ công: Hệ thống sẽ sử dụng những từ có trong danh sách từ khóa (tập đặc trưng văn bản) đã được xây dựng thủ công bởi các chuyên gia/admin. Ví dụ, liên quan đến lĩnh vực CNTT thì có những từ như Hệ thống thông tin, Khoa học máy tính, trí tuệ nhân tạo,… Phương án này đòi hỏi tốn chi phí về thời gian và công sức của chuyên gia.

– Phương án chọn từ khóa tự động: Hệ thống sẽ tự động chọn từ khóa bằng cách tách từ, 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 "và", "nhưng", "có", "không",..) 234.

– Phương án kết hợp: Do trong giai đoạn ban đầu, bộ từ khóa và tập dữ liệu do nhóm tác giả thu thập và xây dựng chưa nhiều, chưa phong phú nên chúng tôi kết hợp cả 2 phương án trên để xây dựng bộ từ khóa.

Sau khi có bộ từ khóa, hệ thống sẽ véc-tơ hóa chúng để làm đầu vào cho bộ phân lớp SVM. Hiện tại, trong giai đoạn thử nghiệm nên hệ thống vận hành theo cơ chế bán tự động, nghĩa là sau khi hệ thống phân loại câu hỏi, người quản trị sẽ kiểm tra kết quả và thực hiện phân loại lại (nếu có sai sót) để làm cơ sở (gán nhãn) cho việc xây dựng và huấn luyện lại mô hình sau này. Trong phần tiếp theo, chúng tôi sẽ mô tả chi tiết cách xây dựng các modules như đã trình bày.

3 XÂY DỰNG CÁC MODULES HỖ TRỢ PHÂN LOẠI CÂU HỎI TỰ ĐỘNG

3.1 Xây dựng module tiếp nhận câu hỏi

3.1.1 Moulde tiếp nhận câu hỏi qua SMS

Gửi tin nhắn SMS: Về tổng thể, có 2 cách để gửi tin nhắn SMS từ máy tính đến điện thoại di động:

Cách 1: Kết nối điện thoại di động hoặc modem GSM/GPRS/3G vào máy tính. Sau đó dùng tập lệnh AT (AT là từ viết tắt của ATtention) để chỉ thị cho điện thoại hoặc modem gửi tin nhắn SMS.

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.

Nhận tin nhắn SMS: Tương tự như việc gửi tin SMS, ta 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 tập lệnh AT để đọc tin nhắn nhận được từ điện thoại đi động hoặc modem. 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. Sau đó việc gửi và nhận tin nhắn SMS thông qua tập lệnh AT

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.

Trong nghiên cứu này, chúng tôi dùng thư viện SMSLIB 14 để hỗ trợ việc gửi và đọc tin nhắn SMS từ modem 3G và lưu vào cơ sở dữ liệu hệ thống.

3.1.2 Module tiếp nhận câu hỏi qua Web/Email

Tương tự như những trang web truyền thống, người dùng sẽ thông qua một form để điền gửi các thông tin cần được tư vấn. Để gửi và nhận email, hệ thống sử dụng giao thức smtp để gửi và pop3 để nhận.

3.2 Xây dựng module rút trích tập đặc trưng văn bản tiếng Việt

Các câu hỏi sau khi được tiếp nhận sẽ được xử lý bằng các phương pháp như trong xử lý ngôn ngữ tự nhiên 234. Việc xử lý này được thực hiện qua hai bước: Tách từ và lựa chọn đặc trưng (từ khóa).

3.2.1 Tách từ

Tách từ là một trong những bước tiền xử lý cơ bản trong việc phân loại văn bản 234. Việc tách từ tiếng Anh khá đơn giản do trong tiếng Anh mỗi từ là một nhóm ký tự có nghĩa, được phân cách bởi ký tự khoảng trắng trong câu. Trong khi đó tiếng Việt phải đối mặt với vấn đề ngược lại do thực tế một từ tiếng Việt có thể có nhiều hơn một âm tiết được tách ra do đó khoảng trắng không phải luôn luôn là ký tự để phân tách một từ tiếng Việt. Chính vì thế ta không thể áp dụng các thuật toán tách từ tiếng Anh cho tiếng Việt. Việc tách từ tiếng Việt đã được nhiều tổ chức và cá nhân quan tâm nghiên cứu với nhiều cách tiếp cận khác nhau, trong đó 2 đã cho thấy phương pháp so khớp tối đa (Maximum Matching) cho kết quả tách từ đạt độ chính xác 96%-98%.

Bên cạnh đó, trong văn bản, một số từ gọi là từ dừng (stopwords) như từ nối, dấu chấm câu, ký hiệu đặc biệt, từ chỉ số lượng ("và", "các", "những, "mỗi",…) không có giá trị trong phân loại. Do vậy, để giảm bớt số lượng đặc trưng, nâng cao tốc độ tính toán, các từ dừng này cần được loại bỏ. Một vài phương pháp thường được sử dụng để loại bỏ các đặc trưng không quan trọng: Tần suất xuất hiện của từ (chỉ số TF), độ lợi thông tin, thông tin tương hỗ, độ mạnh của từ và một số phương pháp khác 234.

Trong nghiên cứu này, chúng tôi sử dụng công cụ VnTokenizer 2 để tách từ và loại bỏ từ dừng. 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. Công cụ này tách từ cho độ chính xác là 96% - 98%.

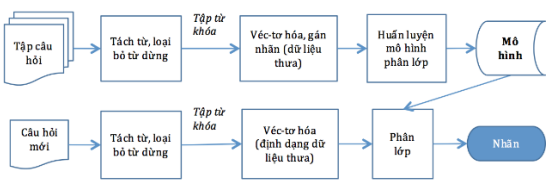
3.2.2 Xây dựng bộ từ khóa đặc trưng

Bộ từ khóa đặc trưng là một danh sách các từ khóa đặc trưng cho các nhóm lĩnh vực cần phân loại (sẽ được trình bày chi tiết sau, trong Bảng 1). Trong nghiên cứu này, chúng tôi đề xuất ba phương án chọn từ khóa là phương án thủ công, phương án tự động và kết hợp cả 2 phương án trên như đã trình bày trong phần II.

Một vấn đề quan trọng cần quan tâm khi xây dựng tập dữ liệu là thói quen nhắn tin tiếng Việt không có 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. Do đó, để giảm sai sót trong phân loại tin nhắn, chúng tôi đề xuất một giải pháp để xây dựng bộ từ khóa là tương ứng với 1 câu hỏi, sẽ tạo ra hai mẫu tin: một cho tiếng Việt và một cho tiếng Anh.

3.3 Phân loại câu hỏi bằng SVM

Trong nghiên cứu này, chúng tôi cài đặt, huấn luyện và sử dụng SVM cho phân loại tin nhắn thông qua công cụ LibLinear 13 .



Hình 2: Quy trình phân loại câu hỏi

3.3.1 Phân loại câu hỏi bằng SVM

Có khá nhiều lĩnh vực liên quan trong tư vấn tuyển sinh, để minh họa, trong nghiên cứu này chúng tôi chọn tám nhóm lĩnh vực như trình bày trong Bảng 1. Tuy nhiên, hệ thống hoàn toàn có thể được mở rộng bằng cách thêm vào các nhóm lĩnh vực khác sau này.

Bảng 1: Các lớp (lĩnh vực) cần phân loại

Mã loại	Lĩnh vực liên quan
1	CNTT và truyền thông
2	Khoa học xã hội và nhân văn
3	Kinh tế
4	Kĩ thuật
5	Nông nghiệp
6	Sư phạm
7	Quy chế - hồ sơ
8	Điểm chuẩn - nguyện vọng

Quy trình phân loại câu hỏi bằng SVM được thực hiện như mô tả như trong Hình 2. Tập câu hỏi thu thập được sẽ được tách từ, loại bỏ từ dừng và lựa chọn từ khóa. Sau đó chúng được véc-tơ hóa để làm đầu vào cho bộ phân lớp SVM. Việc tách từ và chọn từ khóa đã được trình bày ở phần B.

Để vec-tơ hóa các từ khóa (đặc trưng), do văn bản là tin nhắn SMS/email 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 ghi 1, nếu không xuất hiện thì không cần phải lưu, định dạng này còn được gọi là định dạng thưa 5612. Định dạng từng dòng của tập tin huấn luyện như sau:

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

Với ***<label>*** là nhãn (lớp - class) của câu hỏi, ***<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. Với value = 0, ta có thể không cần lưu như ví dụ sau:

1 1:1 5:1 7:1
1 1:1 2:1
...
3 3:1 67:1 90:1 130:1
...
6 6:1 9:1 123:1 149:1

3.4 Xây dựng module gợi ý câu hỏi liên quan

Ngay sau khi thí sinh đặt câu hỏi, trong thời gian chờ đợi trả lời, chúng tôi đề xuất xây dựng một module gợi ý các câu hỏi có liên quan đã được trả lời trước đây để thí sinh có thêm thông tin hỗ trợ quyết định. Ở đây chúng tôi sử dụng phương pháp tính độ tương đồng Cosine.

Để tính độ tương đồng Consine, trước hết tất cả các câu trong văn bản sẽ được vector hóa thành các vector có độ dài bằng nhau thông qua việc tính TF-IDF.

TF (*term frequency*):

$$\text{tf}(t, d) = \frac{f(t,d)}{\max\{f(w,d) : w \in d\}}$$

Trong đó:

f(t,d) là số lần xuất hiện từ **t** trong văn bản **d**.

max{f(w,d) : w ∈ d} là số lần xuất hiện nhiều nhất của một từ bất kỳ **w** trong văn bản.

IDF (*inverse document frequency*):

$$\text{idf}(t, D) = log\frac{|D|}{|\{d \in D : t \in d\}|}$$

Trong đó:

|D|: Tổng số văn bản trong tập **D**

$|\{d \in D : t \in d\}|$: Số văn bản chứa từ nhất định, với điều kiện *t* xuất hiện. Nếu từ đó không xuất hiện ở bất cứ 1 văn bản nào trong tập thì mẫu số sẽ bằng 0 => phép chia cho không không hợp lệ, vì thế người ta thường thay bằng mẫu thức 1+ $|\{d \in D: t \in d\}|$.

TF-IDF:

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D)$$

Sau khi xác định tf-idf, ta tạo ra tập các vector chứa chỉ số TF*IDF cho từng câu hỏi. Sau cùng là tính độ tương đồng Cosine của vec-tơ câu hỏi hiện tại (a) và các vec-tơ của các câu hỏi trước đây (b1, b2, …bn) , theo công thức:

$$\cos\theta = \frac{\vec{a}.\vec{b}}{\overrightarrow{|a|} + \overrightarrow{|b|}}$$

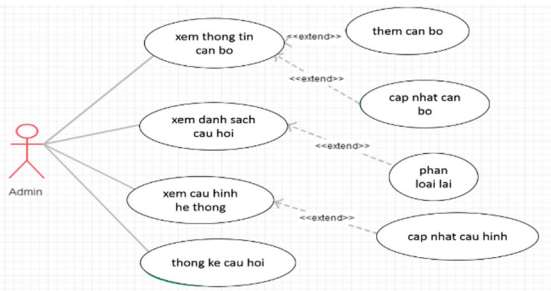
Từ kết quả độ tương đồng Cosine, ta có thể chọn ra top-N câu hỏi tương đồng để gợi ý.

4 XÂY DỰNG HỆ THỐNG THÔNG TIN VÀ TÍCH HỢP CÁC MÔ MODULES

Tương tự như việc xây dựng các hệ thống thông tin quản lý khác, hệ thống này cũng được phân tích, thiết kế, xây dựng và cài đặt và sau đó là tích hợp với các modules quản lý, phân loại câu hỏi. Tuy nhiên, do giới hạn số trang của bài viết, chúng tôi chỉ mô tả một số mô hình/sơ đồ cơ bản.

Ngoài người dùng là thí sinh (người đặt câu hỏi), hệ thống quản lý hai đối tượng người dùng khác là chuyên gia (cán bộ) và quản trị/điều phối viên.

Một phần của sơ đồ use case được biểu diễn như trong Hình 3.

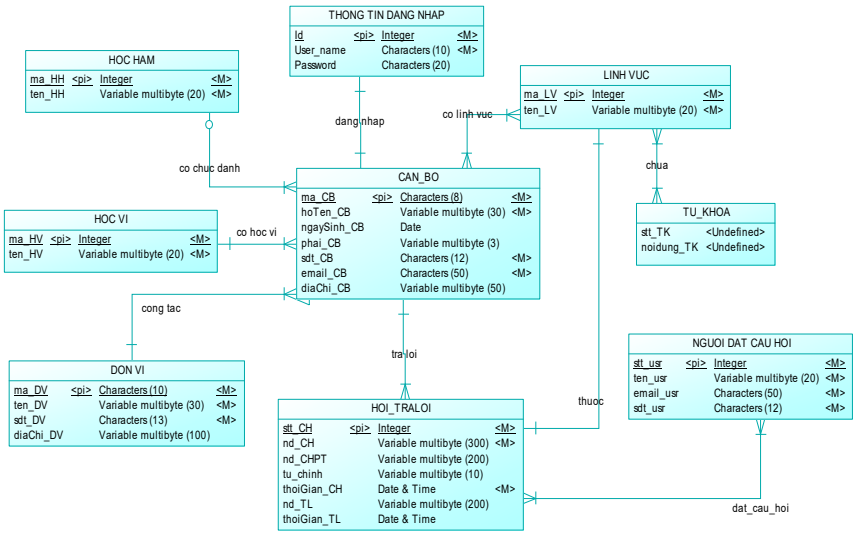


Hình 3: Một phần của Sơ đồ use cases

Các chức năng chính của người dùng chuyên gia: Xem, chỉnh sửa thông tin cá nhân; Trả lời các câu hỏi của thí sinh liên quan đến chuyên môn; Phân loại lại tin nhắn nếu có sai sót. Các chức năng chính của quản trị/ điều phối viên: Cập nhật các lĩnh vực (các lớp) của phân loại; Thêm/cập nhật cán bộ mới; Cập nhật, phân loại tin nhắn; Cấu hình hệ thống; Thống kê.

Chi tiết sơ đồ thực thể kết hợp được trình bày trong Hình 4. Trong đó, các thực thể chính của hệ thống bao gồm: Thí sinh, Cán bộ, lĩnh vực và câu hỏi. Một cán bộ có thể phụ trách nhiều lĩnh vực. Một câu hỏi có thể thuộc một hay nhiều lĩnh vực…

Sau bước phân tích, thiết kế ta tiến hành cài đặt và tích hợp các modules vào hệ thống. Hệ thống tổng thể được xây dựng theo mô hình MVC trên nền ngôn ngữ Java (Spring MVC framewwork) với hệ quản trị cơ sở dữ liệu MySQL.



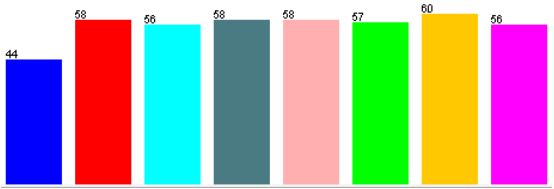
Hình 4: Sơ đồ thực thể kết hợp (ERD)

5 KẾT QUẢ MINH HỌA

5.1 Độ chính xác của mô hình phân loại

5.1.1 Dữ liệu thử nghiệm

Để thử nghiệm độ tin cậy của mô hình dự đoán, chúng tôi thu thập tập dữ liệu gồm 447 câu hỏi, trong đó có 235 câu hỏi (có dấu tiếng Việt) và 212 câu hỏi không có dấu tiếng Việt được hệ thống tự động sinh ra. Sau khi tách từ và loại bỏ từ dừng, còn lại 431 từ khóa. Các câu hỏi trong tập dữ liệu này thuộc 8 lĩnh vực như đã trình bày trong Bảng 1, phân bố khá đồng đều như trong Hình 5, điều này sẽ giúp tránh tình trạng mất cân bằng dữ liệu (imbalanced data) sẽ làm ảnh hưởng đến kết quả phân lớp.

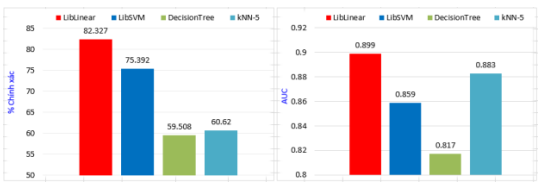


Hình 5: Phân phối dữ liệu của 8 lĩnh vực tương ứng từ trái sang (8 lớp – target class)

5.1.2 Độ chính xác

Bằng nghi thức kiểm tra chéo (10-folds cross validation), mô hình phân loại câu hỏi bằng SVM (dùng công cụ LibLinear) cho kết quả cao nhất đạt 82,33% trong khi Decision Tree (C4.5/J48) chỉ đạt độ chính xác 59.51% như minh họa trong Hình 6 (trái). Từ kết quả này ta nhận thấy rằng do câu hỏi qua tin nhắn SMS (email) rất ngắn nên tập dữ liệu biểu diễn cho các câu hỏi này rất thưa (sparse) vì vậy chúng tôi đã chọn phương pháp biểu diễn dữ liệu thưa và dùng SVM để phân lớp là phù hợp.

Kiểm tra trên độ đo Area Under the ROC Curve (AUC – đây là độ đo thường được dùng trong xếp hạng (rank) các phương pháp), kết quả này cũng gần tương tự cho các phương pháp (SVM dùng Liblinear đạt 0.899) như minh họa trong Hình 6 (phải).



Hình 6: Độ chính xác (trái) và độ đo AUC (phải)

5.2 Các giao diện minh họa

Hình 7 minh hoạ giao diện "Đặt câu hỏi" để gửi yêu cầu thông qua giao diện web. Trong thời gian chờ câu trả lời, hệ thống sẽ tự động gợi ý các câu hỏi liên quan (phần dưới của Hình 7) đến câu vừa hỏi thông qua module tính độ tương đồng của câu hỏi đang được truy vấn và các câu hỏi đã được trả lời trước đây, nhằm hỗ trợ thông tin tốt nhất cho thí sinh.

Các chức năng chính của người dùng là chuyên gia:

Trả lời các câu hỏi liên quan đến chuyên môn: Sau khi đăng nhập thành công, chuyên gia có thể trả lời các câu hỏi liên quan đến lĩnh vực mà họ đã đăng ký (nếu dùng giao diện web). Các câu hỏi này được chuyển cho từng cán bộ nhờ vào hệ thống phân loại câu hỏi tự động hoặc bán tự động (điều

phối viên sẽ chuyển) tùy thuộc vào việc cấu hình hệ thống.

Phân loại lại câu hỏi: Được cài đặt cùng trang với phần trả lời câu hỏi, nếu chuyên gia thấy câu hỏi không đúng chuyên môn của mình thì họ sẽ phân loại lại câu hỏi đó để chuyển đến đúng cán bộ phụ trách, như minh họa trong Hình 8.

Hình 7: Hệ thống tự động gợi ý các câu hỏi liên quan

Các chức năng chính của người dùng là quản trị/điều phối viên:

– Cập nhật, phân loại lại tin nhắn như của chuyên gia

– *Cấu hình hệ thống:* Cho phép thay đổi một số thông số hệ thống như thời gian hệ thống lặp lại việc truy vấn và huấn luyện lại mô hình, số lượng tin nhắn để thực hiện huấn luyện lại,…

– *Thống kê tin nhắn:* Cho phép thống kê tổng số lượng câu hỏi hệ thống nhận được, số lượng tin nhắn đã trả lời, số lượng tin nhắn theo từng chuyên ngành, số lượng tin nhắn đã trả lời của từng cán bộ, theo từng tháng, từng năm…

Hiện tại, hệ thống là một hệ bán tự động, mục đích chủ yếu là thu thập dữ liệu để xây dựng các mô hình phân loại nên trong quá trình vận hành hệ thống ngoài thực tế, hệ thống cần thường xuyên kiểm tra và huấn luyện lại các mô hình để nâng cao độ chính xác cho phân loại tự động. Sau một khoảng thời gian xác định, hệ thống sẽ tiến hành kiểm tra số lượng tin nhắn mới thu thập được, nếu số lượng tin nhắn đủ số lượng quy định để huấn luyện lại mô hình thì hệ thống sẽ thực hiện huấn luyện lại mô hình và sử dụng mô hình mới vào phân loại tin nhắn mới đến hệ thống.

xây dựng tập dữ liệu tốt hơn sau này

Hệ thống sẽ 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 thì hệ thống sẽ được chuyển sang giai đoạn hai của đề tài là xây dựng hệ thống hỗ trợ tư vấn tuyển sinh một cách tự động hoàn toàn.

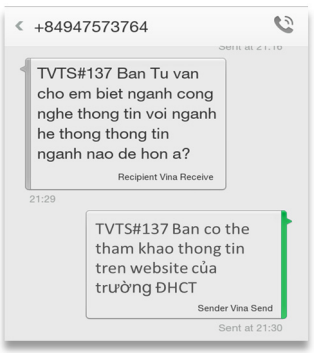
Hình 8: Các câu hỏi đã được phân loại tự động và cũng cho phép chuyên gia phân loại lại

Ở module tư vấn qua tin nhắn SMS, để tránh sai sót và thuận lợi cho hệ thống nhận diện được đâu là tin nhắn câu hỏi của thí sinh, đâu là câu trả lời của cán bộ hoặc tin nhắn rác (spam SMS) thì câu trả lời theo quy tắc mà hệ thống đưa ra, chẳng hạn như:

Tin nhắn SMS từ hệ thống gửi đến cho cán bộ có cấu trúc: TVTS# + mã câu hỏi + khoảng trắng + nội dung câu hỏi (-Tu: HeThongTuVanTuyenSinh)

Khi chuyên gia trả lời câu hỏi, dùng cú pháp: TVTS# + mã câu hỏi + khoảng trắng + nội dung câu trả lời

Ví dụ: Thí sinh đặt câu hỏi qua SMS: "Xin cho biết ngành công nghệ thông tin ra trường có thể làm việc ở đâu?". Câu hỏi được hệ thống xử lý và chuyển đến chuyên gia: "*TVTS#526* Xin cho biết ngành công nghệ thông tin ra trường có thể làm việc ở đâu?", trong đó 526 là mã câu hỏi. Chuyên gia sẽ trả lời theo quy tắc: "*TVTS#526* câu trả lời…". Một ví dụ khác được minh hoạ trong Hình 9.



Hình 9: Giao diện minh họa tư vấn qua tin nhắn SMS

6 KẾT LUẬN

Bài viết này đã đề xuất một giải pháp xây dựng Hệ thống tư vấn tuyển sinh bán tự động sử dụng kết hợp các kỹ thuật trong xử lý văn bản, máy học SVM và xử lý tin nhắn SMS trong hệ thống thông tin di động. Hệ thống tư vấn này có khả năng tiếp nhận câu hỏi của thí sinh từ trang Web hoặc qua tin nhắn SMS, sau đó, câu hỏi sẽ được phân loại tự động bằng máy học SVM để chuyển đến chuyên gia thích hợp trong từng lĩnh vực. Sau khi có câu trả lời từ chuyên gia, hệ thống sẽ phản hồi tức thì cho thí sinh. Bên cạnh đó, ngay sau khi thí sinh đặt câu hỏi, hệ thống sẽ xử lý và tìm độ tương đồng của câu hỏi hiện tại so với các câu đã được trả lời trước đây, nhằm gợi ý cho thí sinh có thêm thông tin.

Để hoàn thiện hơn, hệ thống cần được triển khai ngoài thực tiễn để thu thập thêm dữ liệu thực, cập nhật thêm bộ từ khóa, từ đó huấn luyện lại mô hình phân lớp nhằm đạt độ chính xác cao hơn.

LỜI CẢM TẠ

Chân thành cảm ơn các em Đỗ Lê Nhật Thanh, Nguyễn Nam Nhi và Lương Thế Anh đã hỗ trợ cài đặt demo. Nghiên cứu này là một phần trong đề tài NCKH cấp Trường Đại học Cần Thơ, mã số đề tài T2015-32.

TÀI LIỆU THAM KHẢO

1. V.Vapnik. The Nature of Statistical Learning Theory. Springer, NewYork, 1995.
2. Phuong, L. H., Thi Minh Huyên, N., Roussanaly, A., & Vinh, H. T. (2008, June). A Hybrid Approach to Word Segmentation of Vietnamese Texts. In Language and Automata Theory and Applications (pp. 240-249). Springer-Verlag.
3. Huang, X., Peng, F., Schuurmans, D., Cercone, N., & Robertson, S. E. (2003). Applying machine learning to text segmentation for information retrieval. Information Retrieval, 6(3-4), 333-362.
4. Chang, P. C., Galley, M., & Manning, C. D. (2008, June). Optimizing Chinese word segmentation for machine translation performance. In Proceedings of the Third Workshop on Statistical Machine Translation (pp. 224-232). Association for Computational Linguistics.
5. Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features (pp. 137-142). Springer Berlin Heidelberg.
6. Ikonomakis, M., Kotsiantis, S., & Tampakas, V. (2005). Text classification using machine learning techniques. WSEAS Transactions on Computers, 4(8), 966-974.
7. Dalal, Mita K., and Mukesh A. Zaveri. "Automatic text classification: a technical review." International Journal of Computer Applications 28.2 (2011): 37-40.
8. Song, G., Ye, Y., Du, X., Huang, X., & Bie, S. (2014). Short text classification: A survey. Journal of Multimedia, 9(5), 635-643.
9. Arnaud Henry-Labordere and Vincent Jonack. 2004. SMS and MMS Interworking in Mobile Networks. Artech House, Inc., Norwood, MA, USA.
10. Trần Cao Đệ, 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.
11. Lương Thế Anh, Nguyễn Thái Nghe, và Nguyễn Chí Ngôn. 2014. Xây dựng hệ thống hỗ trợ khuyến nông trên cây lúa qua mạng thông tin di động. Trang 9-21, số 33a, Tạp chí Khoa học Trường Đại học Cần Thơ, ISSN: 1859-2333.
12. Chang, C.C., Lin, C.J (2011), LIBSVM – a library for support vector machines, http://www.csie.ntu.edu.tw/~cjlin/libsvm
13. R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification Journal of Machine Learning Research 9(2008), 1871-1874.
14. SMSLib, a universal API for sms messaging, http://smslib.org/, retrieved 01/2015
15. jwap, http://jwap.sourceforge.net/, retrieved 01/2015
16. jMmsLib, http://jmmslib.sourceforge.net, retrieved 01/2015