BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**NIÊN LUẬN**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**THIẾT KẾ WEBSITE KINH DOANH ĐỒ ĐIỆN TỬ GIA DỤNG**

**Sinh viên: Nguyễn Tuấn Khanh**

**Mã số: B1706481**

**Khóa: K43**

**Cần Thơ, 5/2021**

BỘ GIÁO DỤC VÀ ĐÀO TẠO

**TRƯỜNG ĐẠI HỌC CẦN THƠ**

**KHOA CÔNG NGHỆ THÔNG TIN & TRUYỀN THÔNG**

**BỘ MÔN CÔNG NGHỆ THÔNG TIN**

**NIÊN LUẬN**

**NGÀNH CÔNG NGHỆ THÔNG TIN**

**Đề tài**

**THIẾT KẾ WEBSITE KINH DOANH ĐỒ ĐIỆN TỬ GIA DỤNG**

Người hướng dẫnSinh viên thực hiện

**TS. Thái Minh Tuấn Nguyễn Tuấn Khanh**

**Mã số: B1706481**

**Khóa: K43**

***Cần Thơ, 5/2021***

# LỜI CẢM ƠN

Lời đầu tiên, tôi xin bày tỏ sự cảm ơn chân thành đối với TS. Thái Minh Tuấn – giáo viên hướng dẫn trực tiếp của tôi. Thầy Tuấn đã cho tôi những gợi ý và chỉ dẫn quý báu trong quá trình nghiên cứu và hoàn thiện niên luận.

Tôi cũng xin được cảm ơn gia đình, những người thân, và bạn bè tôi đã quan tâm, giới thiệu các tài liệu hữu ích trong thời gian học tập và nghiên cứu niên luận.

Mặc dù đã cố gắng hoàn thành niên luận nhưng chắc chắn sẽ không tránh khỏi những sai sót, tôi kính mong nhận được sự thông cảm và chỉ bảo của thầy và các bạn.

Tôi xin chân thành cảm ơn!

**MỤC LỤC**

[LỜI CẢM ƠN i](#_Toc72528404)

[TÓM LƯỢT iv](#_Toc72528405)

[ABSTRACT v](#_Toc72528406)

[PHẦN GIỚI THIỆU 1](#_Toc72528407)

[CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ ĐỀ TÀI 3](#_Toc72528408)

[**1.1** **MÔ TẢ ĐỀ TÀI** 3](#_Toc72528409)

[**1.2** **CÁC CHỨC NĂNG CỦA HỆ THỐNG** 3](#_Toc72528410)

[CHƯƠNG 2: THIẾT KẾ CÀI ĐẶT VÀ GIẢI PHÁP 4](#_Toc72528411)

[**2.1** **CƠ SỞ LÝ THUYẾT** 4](#_Toc72528412)

[**2.1.1** **Giới thiệu mô hình MVC** 4](#_Toc72528413)

[**2.1.2** **Các công nghệ sử dụng khác** 4](#_Toc72528414)

[**2.2** **SƠ ĐỒ USECASE** 5](#_Toc72528415)

[**2.2.1** **Khách vãng lai** 5](#_Toc72528416)

[**2.2.2** **Thành viên** 5](#_Toc72528417)

[**2.2.3** **Quản trị viên** 5](#_Toc72528418)

[CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ 6](#_Toc72528419)

[**3.1** **Tập dữ liệu** 6](#_Toc72528420)

[**3.2** **Tiền xử lý văn bản** 6](#_Toc72528421)

[**3.3** **Phương pháp đánh giá kết quả tóm tắt** 6](#_Toc72528422)

[**3.4** **Lựa chọn các phương pháp tóm tắt văn bản đầu vào** 6](#_Toc72528423)

[*Bảng 3. Bảng thông kê đặc trưng của 3 phương pháp đầu vào.* 7](#_Toc72528424)

[*Bảng 4: kết quả tóm tắt của 3 phương pháp* 7](#_Toc72528425)

[*Bảng 5: Kết quả tóm tắt của mô hình sử dụng kỹ thuật Voting* 8](#_Toc72528426)

[*Bảng 6: Bảng tổng kết kết quả thử nghiệm trên tập dữ liệu.* 8](#_Toc72528427)

[CHƯƠNG 4: KẾT LUẬN 9](#_Toc72528428)

[TÀI LIỆU THAM KHẢO 10](#_Toc72528429)

[PHỤC LỤC 12](#_Toc72528430)

**DANH MỤC BẢNG VÀ HÌNH VẼ**

[Hình 1-1 Các giai đoạn của hệ thống tóm tắt 3](#_Toc58432794)

[Bảng 1. Minh họa phương pháp Voting. 7](#_Toc58432805)

[Hình 2-1. Mô hình tóm tắt đơn văn bản tiếng Việt dựa vào kỹ thuật Voting. 8](#_Toc58432807)

[Bảng 2. Minh hoạ tóm tắt bằng kỹ thuật Voting 9](#_Toc58432808)

[Bảng 3. Bảng thông kê đặc trưng của 3 phương pháp đầu vào. 13](#_Toc58432818)

[Bảng 4: kết quả tóm tắt của 3 phương pháp 13](#_Toc58432819)

[Bảng 5: Kết quả tóm tắt của mô hình sử dụng kỹ thuật Voting 14](#_Toc58432820)

[Bảng 6: Bảng tổng kết kết quả thử nghiệm trên tập dữ liệu. 14](#_Toc58432821)

# TÓM LƯỢT

Trong xu thế phát triển mạnh mẽ của thương mại điện tử trên thế giới, các doanh nghiệp Việt Nam đặc biệt là các doanh nghiệp vừa và nhỏ, bộ phận chiếm tỷ lệ lớn trong tổng số các doanh nghiệp Việt Nam, cũng đã bước đầu nhận thức được ích lợi và tầm quan trọng của việc ứng dụng thương mại điện tử. Tuy nhiên, do còn nhiều hạn chế trong nhận thức của bản thân các doanh nghiệp cũng như các điều kiện cơ sở hạ tầng, việc ứng dụng thương mại điện tử trong các doanh nghiệp vừa và nhỏ ở Việt Nam có thể nói mới ở mức độ sơ khởi. Vì thế, các doanh nghiệp vừa và nhỏ cần xây dựng cho mình một chiến lược kinh doanh phù hợp với điều kiện thực tế hiện nay nhằm tiếp cận nhiều hơn nữa với thương mại điện tử để có thể khai thác tối đa lợi ích mà phương thức kinh doanh này đem lại.

Hiện tại các phương tiện kỹ thuật và các công cụ lập trình ứng dụng vào TMĐT chưa được ứng dụng rộng rãi trong các doanh nghiệp, để có đủ kiến thức cơ bản về TMĐT xây dựng và vận hành một website có hiệu quả, thực hiện marketing qua mạng, chọn lựa mô hình TMĐT phù hợp, xúc tiến thanh toán qua mạng, áp dụng các biện pháp an toàn mạng tối thiểu cần thiết v.v...

Từ những lý do trên, Tôi quyết định chọn đề tài “Thiết kế website kinh doanh đồ điện tử gia dụng” - xây dựng thử nghiệm cửa hàng kinh doanh điện tử gia dụng theo mô hình thương mại điện tử. Thông qua đó để tôi hiểu biết, nắm rõ hơn về lĩnh vực TMĐT và các kỹ năng ứng dụng lập trình cho website thương mại điện tử một cách hiệu quả.

# ABSTRACT

In the strong development trend of e-commerce in the world, Vietnamese enterprises, especially small and medium enterprises, which account for a large proportion of the total number of Vietnamese enterprises, have also begun to receive see the benefits and importance of the application of e-commerce. However, due to the limited awareness of infrastructure and business facilities, it can be said that e-commerce in small and medium enterprises in Vietnam is in its infancy. Therefore, small and medium enterprises need to build for themselves a business strategy in line with today's reality to have more access to e-commerce in order to make the most of the benefits that this business method offers bring.

At present, the technical means and programming tools applied to e-commerce have not been widely applied in enterprises, in order to have enough basic knowledge about e-commerce to build and operate an e-commerce website e-commerce website, do network marketing, choose the right e-commerce model, promote online payment, apply minimum necessary cybersecurity measures, etc.

From the above reasons, I decided to choose the topic "Designing a website for home electronics business" - building a trial of a home electronics business store according to the e-commerce model. Through that, I have a better understanding and understanding of the field of e-commerce and the skills to apply programming for e-commerce websites effectively.

# PHẦN GIỚI THIỆU

1. **Đặt vấn đề**

Thế giới ngày một phát triển, đời sống con người ngày càng hiện đại hóa,nhu cầu về đời sống vật chất và tinh thần của con người mỗi ngày một cao và có nhiều đòi hỏi khắt khe hơn.

Hằng ngày mỗi người trong chúng ta đều phải có những nhu cầu thiết yếu như : ăn uống , mua sắm, xem phim, nghe nhạc ...và vì vậy chúng ta phải đi tới các trung tâm giải trí hay mua sắm.... để mua những sản phẩm mình muốn hay đơn giản là thưởng thức những bản nhạc mới nhất.

Nắm bắt được nhu cầu đó và xét thấy sự phát triển vượt bậc của mạng Internet ở khắp nơi trên thế giới nói chung và tại nước ta trong những năm gần đây nói riêng nên việc ứng dụng công nghệ thông tin vào lĩnh vực tìm kiếm thông tin, truy cập dữ liệu từ xa không còn khó khăn. Người ta tận dụng tối đa những lợi ích mà lĩnh vực này mang lại.

Mục tiêu của việc xây dựng trang web này nhằm giúp cho khách hàng có thể mua hàng trực tiếp từ xa thông qua mạng internet phổ biến. Khách hàng ở nhà hay tại công ty vẫn có thể dễ dàng tham khảo thông tin sản phẩm mình tìm, so sánh giá cả các mặt hàng và lựa chọn cho mình loại sản phẩm phù hợp nhu cầu của mình, giúp công việc mua sắm một cách nhanh chóng, tiện lợi, tiết kiệm thời gian, đáp ứng được nhu cầu thực tế. Hệ thống tìm kiếm dễ dàng, giao diện thân thiện. Chỉ cần đăng nhập vào hệ thống với tài khoản đã có (nếu khách hàng đã là thành viên của cửa hàng) hay chỉ cần vài thao tác đăng kí đơn giản là khách hàng có thể tự do chọn mua và tạo đơn đặt hàng tại hệ thống.

1. **Mục tiêu đề tài**

Hiểu biết hình thức kinh doanh thương mại điện tử. Hiểu biết về các giải pháp xây dựng một website thương mại điện tử. Xây dựng một e mang lại thuận tiện cho người dùng, tiết kiệm thời gian khi có nhu cầu mua một sản phẩm. Xây dựng Website giúp cho nhân viên quản lý thông tin một cách trực quan, thuận tiện.

Củng cố và nâng cao kỹ năng lập trình

1. **Đối tượng và phạm vi nghiên cứu**

Website kinh doanh đồ điện tử gia dụng được xây dựng với tiêu chí phục vụ cho tất cả mọi người với mong muốn đáp ứng đầy đủ các mặt hàng, các chức năng của một website bán hàng trực tuyến.

Phạm vi nghiên cứu là: quản lý sản phẩm, quản lý đơn hang, đặt hang, tìm kiếm sản phẩm, lọc sản phẩm, lịch sử mua hàng, đăng ký thành viên, thống kê doanh số.

1. **Nội dung nghiên cứu**

Phương pháp thực hiện:

* Phân tích yêu cầu: thực hiện khảo sát thực trạng, phân tích đánh giá thực trạng, đặc tả yêu cầu xây dựng hệ thống.
* Thiết kê: yêu cầu chức năng, phân tích và thiết kế mô hình hệ thống: CDM, PDM.
* Cài đặt: ngôn ngữ lập trình PHP, hệ quản trị CSDL: MySQL.

1. **Bố cục niên luận cơ sở**

Niên luận gồm 4 chương chính.

Chương 1: Giới thiệu tổng quan đề tài

Chương 2: Thiết kế cài đặt và giải pháp

Chương 3: Kiểm thử và Đánh giá

Chương 4: Kết luận

# CHƯƠNG 1. GIỚI THIỆU TỔNG QUAN VỀ ĐỀ TÀI

1. **MÔ TẢ ĐỀ TÀI**

“Website kinh doanh đồ điện tử gia dụng” là trang web cung cấp các chức năng phục vụ cho việc mua hàng trực tuyến. Hệ thống bao các chức năng chính cho người dùng: thêm sản phảm vào giỏ hàng, đặt hang, đăng ký thành viên, thay đổi mật khẩu, tìm kiếm sản phẩm, lọc sản phẩm, hủy đơn hang. Hệ thống được xây dựng hướng đến các nhóm người dùng “Khách”, “Thành viên” và “Quản trị viên”.

Khi trang chủ được hiển thị, khách hàng có thể xem sản phẩm của các hãng đồ điện tử gia dụng có trong danh mục, các sản phẩm này có thể trùng nhau về tên, vềloại sản phẩm nhưng chỉ có người quản trị phân biệt được sản phẩm nào thuộc hãng đồ điện tử gia dụng nào. Sau khi đã chọn được sản phẩm, khách hàng có thể thêm sản phẩm vào giỏ hàng hoặc bấm vào nút mua ngay để thanh toán đơn hàng với điều khiện khách hàng phải đăng nhập vào website, nếu khách hàng chưa phải là thành viên có thể đăng ký trở thành thành viên của website. Trong giỏ hàng của mình, khách hàng có thể cập nhật số lượng, thêm hoặc bớt đi sản phẩm. Khách hàng sẽ lập đơn đặt hàng sau khi đã quyết định chọn những sản phẩm có trong giỏ hàng. Phiên giao dịch sẽ kết thúc khi khách hàng đặt xong đơn đặt hàng.

1. **CÁC CHỨC NĂNG CỦA HỆ THỐNG**

Hệ thống bao gồm các chức năng chính sau:

* Khác vãng lai: đăng ký, tìm kiếm sản phẩm, thêm sản phẩm vào giỏ hàng, xem danh mục sản phẩm, lọc sản phẩm theo danh mục, xem thông số kỹ thuật.
* Thành viên là người đã có tài khoản ngoài thực hiện các chức năng như người dùng Khách vãng lai còn có thể: đăng nhập, đặt hàng, thay đổi mật khẩu, hủy đơn, xem đơn hàng, xem thông tin tài khoản.
* Quản trị viên: đăng nhập, quản lý sản phẩm, quản lý đơn hàng, tìm kiếm sản phẩm, quản lý khách hàng, thống kê doanh số, quản lý mã giảm giá, quản lý loại sản phẩm, quản lý nhà sản xuất.

# CHƯƠNG 2: THIẾT KẾ CÀI ĐẶT VÀ GIẢI PHÁP

1. **CƠ SỞ LÝ THUYẾT**
2. **Giới thiệu mô hình MVC**

Mô hình MVC (model, controller, view) là mô hình chuẩn cho ứng dụng web được sử dụng nhiều nhất ngày nay. Mô hình MVC được sử dụng lần đầu tiên trong Smalltalk, sau đó được sử dụng phổ biến trong ngôn ngữ lập trình Java. Hiện nay, đã có hơn hàng chục PHP framework dựa trên mô hình này.

* **Model**: có nhiệm vụ thao tác với cơ sở dữ liệu, nghĩa là nó sẽ chứa tất cả các hàm, các phương thức truy vấn trực tiếp với dữ liệu và controller sẽ thông qua các hàm, phương thức đó để lấy dữ liệu rồi gửi qua View.
* **View**: Chịu trách nhiệm hiển thị dữ liệu đã được truy xuất từ model theo một format **nào** đó theo ý đồ của lập trình viên. Cách sử dụng của View tương tự như các module templates thường thấy trong các ứng dụng web phổ biến như WordPress, Joomla,…
* **Controller**: trung gian, làm nhiệm vụ xử lý cho model và view tương tác với nhau. Controller nhận request từ client, sau đó gọi các model để thực hiện các hoạt động được yêu cầu và gửi ra ngoài View. View sẽ chịu trách nhiệm format lại data từ controller gửi ra và trình bày dữ liệu theo 1 định dạng đầu ra (html).

1. **Các công nghệ sử dụng khác**

**PHP (Hypertext Preprocessor):** ngôn ngữ lập trình kịch bản mã nguồn mở, chủ yếu được dùng để phát triển các ứng dụng viết cho máy chủ. PHP rất thích hợp với web và có thể dễ dàng nhúng vào trang HTML. Do được tối ưu hóa cho các ứng dụng web, tốc độ nhanh, nhỏ gọn, cú pháp giống C và Java, dễ học và thời gian xây dựng sản phẩm tương đối ngắn hơn so với các ngôn ngữ khác nên PHP đã nhanh chóng trở thành một ngôn ngữ lập trình web phổ biến nhất thế giới.

**CSS (Cascading Style Sheets):** thường được sử dụng với HTML để tùy chỉnh giao diện người dùng.

**HTML (Hyper Text Markup Language – Ngôn ngữ đánh dấu siêu văn bản):** được coi là xương sống của một trang web, sử dụng để định dạng và hiển thị văn bản trên trình duyệt. Cha đẻ của HTML là Tim Berners-Lee, cũng là người khai sinh ra World Wide Web và chủ tịch của World Wide Web Consortium (W3C - tổ chức thiết lập ra các chuẩn trên môi trường Internet).

**JavaScript:** ngôn ngữ lập trình kịch bản, giúp tạo ra tính tương tác giữa người dùng và website.

**jQuery:** bộ thư viện được viết dựa trên ngôn ngữ JavaScript, giúp đơn giản hóa việc viết mã lệnh JavaScript. Là thư viện mã nguồn mở được phát triển bởi John Resig và phiên bản đầu tiên được ra đời vào tháng 8 năm 2006.

**Bootstrap:** là một CSS framework được Twitter phát triển, gồm tập hợp các bộ chọn, thuộc tính và giá trị có sẵn, giúp nhà phát triển tiết kiệm thời gian xây dựng giao diện cho dự án.

**Ajax (Asynchronous JavaScript And XML):** là phương thức trao đổi dữ liệu với máy chủ để cập nhật một hay nhiều phần của trang web - mà không cần tải lại toàn bộ trang.

**MySQL:** hệ quản trị cơ sở dữ liệu mã nguồn mở được các nhà phát triển rất ưa chuộng trong quá trình phát triển ứng dụng. Với tốc độ và tính bảo mật cao, MySQL rất thích hợp cho các ứng dụng có nối kết cơ sở dữ liệu trên Internet.

**PowerDesigner:** phần mềm hỗ trợ đắc lực trong việc thiết kế các mô hình quan hệ: CDM (Conceptual Data Model - Mô hình dữ liệu mức quan niệm), LDM (Logical Data Model - Mô hình dữ liệu mức luận lý), PDM (Physical Data Model - Mô hình dữ liệu mức vật lý)...

1. **SƠ ĐỒ USECASE**
2. **Khách vãng lai**
3. **Thành viên**
4. **Quản trị viên**

**CHƯƠNG 3: THỬ NGHIỆM VÀ ĐÁNH GIÁ**

1. **Tập dữ liệu**

Tập dữ liệu sử dụng trong niên luận cơ sở này được xây dựng theo quan điểm thu thập từ những trang báo mạng chính thống. Trong cấu trúc một bài báo mạng thường được chia làm 3 phần: Tiêu đề, tóm tắt, nội dung. Tôi xem phần tóm tắt chính là phần tóm tắt của con người thực hiện. Do vậy, tôi thu thập 1945 văn bản

Độ dài bản tóm tắt hệ thống là 3 câu gần tương đương với bản tóm tắt con người

1. **Tiền xử lý văn bản**

Tập dữ liệu mẫu được lấy từ trang thông tin điện tử nên sẽ còn gặp nhiều vấn đề khó khăn trong quá trình xây dựng và huấn luyện mô hình nên cần phải tiến hành tiền xử lý văn bản.

Ở bước tiền xử lý tôi sử dụng các bộ công cụ sau:

* Sử dụng thực viện Natural Language Toolkit(NLTK) để tiến hang tách câu trong tiếng Việt
* Sử dụng thư viện ViTokenizer để tách từ tiếng Việt
* Sử dụng thư viện ViPosTagger để gán nhãn từ loại

1. **Phương pháp đánh giá kết quả tóm tắt**

Phương pháp đánh giá ROUGE-N đã được đề cập trong mục 1.4.3.

Trong niên luận này tôi sử dụng tool [ROUGE](https://github.com/pltrdy/rouge) để đánh giá.

1. **Lựa chọn các phương pháp tóm tắt văn bản đầu vào**

Trong phần này, để có cơ sở đánh giá hiệu quả của phương pháp Voting. Tôi lựa chọn 03 phương pháp tóm tắt văn bản đầu vào cho phương pháp Voting. Trong đó mỗi phương pháp lựa chọn một số đặc trưng khác nhau để đại diện cho phương pháp, cụ thể: phương pháp 1 chọn 2 đặc trưng trong đó đặc trưng trọng số TF.ISF có vai trò lớn nhất đại diện cho phương pháp 1; phương pháp 2 lại chọn đặc trưng xác suất từ; phương pháp 3 chọn đặc trưng thực thể tên.

Để hiểu rõ hơn, xem bảng mô tả các đặc trưng được lựa chọn của từng phương pháp tóm tắt được thể hiện trong bảng 3:

*Bảng 3. Bảng thông kê đặc trưng của 3 phương pháp đầu vào.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Đặc trưng** | **Phương pháp (1)** | **Phương pháp (2)** | **Phương pháp (3)** |
| F1- Trọng số TF.ISF | x | x |  |
| F2- Xác suất từ |  | x | x |
| F3- Thực thể tên | x |  | x |

**Phương pháp 1:** dựa vào trọng số TF.ISF và thực thể tên.

**Phương pháp 2:** dựa vào trọng số TF.ISF và xác suất từ.

**Phương pháp 3:** dựa vào xác suất từ và thực thể tên.

Với 3 phương pháp nêu trên, tiến hành thực hiện tóm tắt văn bản. Kết quả của 3 phương pháp được mô tả trong bảng 4:

# *Bảng 4: kết quả tóm tắt của 3 phương pháp*

|  |  |  |
| --- | --- | --- |
| **Phương pháp** | **Kết quả tóm tắt (ROUGE-N)** | |
| N=1 | N=2 |
| Phương pháp (1) | 0.264 | 0.123 |
| Phương pháp (2) | 0.263 | 0.120 |
| Phương pháp (3) | 0.268 | 0.127 |

Tiếp theo, tôi tiến hình Voting trên 3 phương pháp là phương pháp (1)(2)(3), như ta quan sát trong bảng 4, mỗi phương pháp đều sử dụng đặc trưng có vái trò cao riêng làm chủ đạo cho phương pháp. Cụ thể: phương pháp(1) sử dụng đặc trưng trọng số TF.ISF, phương pháp (2) sử dụng đặc trưng xác suất từ, phương pháp (3) sử dụng đặc trưng thực thể tên.

Kết quả của mô hình tóm tắt văn bản sử dụng kỹ thuật Voting được mô tả dưới bảng 5:

# *Bảng 5: Kết quả tóm tắt của mô hình sử dụng kỹ thuật Voting*

|  |  |  |
| --- | --- | --- |
| **Phương pháp** | **Kết quả tóm tắt (ROUGE-N)** | |
| N=1 | N=2 |
| **Voting 3 phương pháp (1)(2)(3)** | 0.270 | 0.127 |

Với các kết quả, ta có bảng 6 tổng hợp kết quả tóm tắt:

# *Bảng 6: Bảng tổng kết kết quả thử nghiệm trên tập dữ liệu.*

|  |  |  |
| --- | --- | --- |
| **Phương pháp** | **Kết quả tóm tắt (ROUGE-N)** | |
| N=1 | N=2 |
| Phương pháp (1) | 0.264 | 0.123 |
| Phương pháp (2) | 0.263 | 0.120 |
| Phương pháp (3) | 0.268 | 0.127 |
| **Voting 3 phương pháp (1)(2)(3)** | 0.270 | 0.127 |

Quảng bảng tổng kết trên, cho thấy:

Trên tập dữ liệu thử nghiệm với kết quả của 3 phương pháp khá cạnh tranh nhau thì kết quả Voting cao hơn các phương pháp đơn lẻ. Vậy có thể kết luận rằng, phương pháp tóm tắt văn bản tiếng Việt theo hướng trích rút sử dụng kỹ thuật Voting cho kết quả tốt hơn các phương pháp tóm tắt đơn lẻ với tỉ lệ là 0.27.

# CHƯƠNG 4: KẾT LUẬN

Bài toán tóm tắt tiếng Việt có ý nghĩa quan trọng trong nâng cao hiệu quả khai thác thông tin từ các kho ngữ liệu văn bản tiếng Việt. Các công cụ tóm tắt tiếng Việt được ứng dụng nhiều trong các hệ thống tìm kiếm thông minh, đa ngôn ngữ, tổng hợp thông tin... Đối với lĩnh vực an ninh quốc phòng, tóm tắt tin tức có thể giúp cho cán bộ nghiệp vụ thu thập đủ các thông tin cần thiết và kịp thời theo dõi, đánh giá, xử lý nguồn thông tin một cách nhanh chóng. Nâng cao hiệu quả và độ chính xác của tóm tắt tiếng Việt là hướng nghiên cứu có ý nghĩa khoa học và thực tiễn luôn được các nhà khoa học quan tâm nghiên cứu. Chính vì vậy, mục tiêu nghiên cứu của luận án này là đề xuất các phương pháp tóm tắt văn bản mới phù hợp với văn bản tiếng Việt, có thể áp dụng xây dựng các phần mềm tóm tắt văn bản tiếng Việt chất lượng cao phục vụ trong nhiều lĩnh vực, nhất là lĩnh vực an ninh quốc phòng.

Trong niên luận cơ sở này tôi đã khái quát trình bày các bước cơ bản để xây dựng bài toán tóm tắt văn bản trong nâng cao hiệu quả khai thác thông tin từ các kho ngữ liệu văn bản tiếng Việt. Xây dựng bài toán tóm tắt văn bản theo hướng rút trích ES(Extraction Summarization) là thông qua tính toán các đặc trưng tần suất từ, thực thể tên, trọng số TF.ISF,… để chọn ra các câu quan trọng nhất theo tỉ lệ trích rút.

Phương pháp tóm tắt văn bản mới dựa theo kỹ thuật Voting với ý tưởng xem kết quả của mỗi phương pháp tóm tắt văn bản khác nhau là một lá phiếu đã được sắp xếp thứ tự ưu tiên các câu.Kết quả tóm tắt văn bản theo kỹ thuật Voting đạt kết quả khá khiếm tốn với 27%. Tuy nhiên, với phương pháp Voting còn gặp nhiều hạn chế. Những mặc hạn chế đó là nếu ta xem mỗi phương pháp là một lá phiếu thì có khả năng xảy ra là số phương pháp yếu nhiều hơn sẽ thắng số phương pháp tốt (theo quan điểm đa số) và ngược lại. Điều đó làm cho kết quả Voting có thể không tốt,… Ngoài ra, đề tài còn hạn chế về tập dữ liệu thu thập từ các trang báo điện tử và việc lấy phần tóm tắt con người để đánh giá còn nhiều khó khăn. Chính vì vậy, thách thức đặt ra để mở rộng đề tài này theo hướng tiếp cận tiếp theo là làm sao để giải quyết vấn đề các phương pháp yếu sẽ thắng các phương pháp tốt,cũng như cần mở rộng tập dữ liệu. Có thể kết hợp phương pháp Voting với giải thuật di truyền hay giải thuật tối ưu đàn kiến.

# TÀI LIỆU THAM KHẢO

**Tiếng Việt**

|  |  |
| --- | --- |
| [1] | D. Q. Ban, Ngữ Pháp Tiếng Việt, NXB giáo dục, 2006. |
| [2] | T. N. A. Nguyễn Nhật An, “Tóm tắt văn bản tiếng Việt dựa vào kỹ thuật Voting” trong *Chuyên san Công nghệ thông tin và Truyền thông (JICT) thuộc Tạp chí Khoa học và Kỹ thuật*, Học viện Kỹ thuật quân sự, tháng 4/2014, 2014, pp. pp.57-67. |
| [3] | N. V. Tu, Từ điển đồng nghĩa Tiếng Việt,, NXB giáo dục., 2001. |
| [4] | H. K. Đỗ Phúc, Rút trích ý chính từ văn bản tiếng Việt, Tạp chí Công nghệ Thông tin và Truyền thông, 2006. |
| [5] | L. T. Hương, Nghiên cứu một số phương pháp tóm tắt văn bản, Báo cáo tổng kết đề tài cấp KH và CN cấp bộ, Đại học Bách khoa Hà Nội, 2014. |
| [6] | N. T. T. Hà, “Phát triển một số thuật toán tóm tắt văn bản Tiếng Việt sử dụng phương pháp học bán giám sát,” trong *Luận án Tiến sỹ*, Học viện, 2012. |
| [7] | H. Phê, Từ điển tiếng Việt, NXB giáo dục, 1998. |
| [8] | N. H. T. A. Luận án Tiến sỹ, Tiếp cận đồ thị biểu diễn, khai thác văn bản và ứng dụng, Trường Đại học Khoa Học Tự Nhiên, ĐHQG-HCM, 2011. |
| [9] | T. N. A. P. T. N. H. Nguyễn Nhật An, “Kỹ thuật Voting trong bài toán tách từ tiếng Việt,” trong *Tạp chí NCKH&CNQS*, Đặc san CNTT 04/2014, pp. tr.54-61. |
| [10] | N. Q. D. Trương Quốc Định, “Một giải pháp tóm tắt văn bản tiếng Việt tự động,” trong *Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của Công nghệ*, 03-04/12/2012. |
| [11] | N. T. T. H. Luận án Tiến sỹ, “Phát triển một số thuật toán tóm tắt văn bản Tiếng Việt sử dụng phương pháp học bán giám sát,” trong *HV KTQS*, 2012, p. 175 trang. |
| [12] | T. M. Vũ, "Tóm tắt đa văn bản dựa vào trích xuất câu," in *Luận văn Thạc sỹ*, Trường ĐHCN, Đại học Quốc gia Hà Nội, 2010.  **Tiếng Anh** |
| [13] | S. A. X. H. P. T. B. H. H. S. M.L. Nguyen, "Sentence Extraction with Support Vector Machine Ensemble," in *Proceedings of the First World Congress of the International Federation for Systems Research : The New Roles of Systems Sciences For a Knowledge-based Society*, 2005. |
| [14] | P. T. A. T. C. D. N. X. H. Nguyen Quang Uy, "A Study on the Use of Genetic Programming for Automatic Text Summarization," in *Conference on Knowledge and Systems Engineering*, KSE, 2012 4th Inter, 2012, pp. pp.93-98. |
| [15] | V. project, "Vietnamese Language Processing," [Online]. Available: http://vlsp.vietlp.org. |
| [16] | E. Pacuit, "Voting Methods," The Stanford Encyclopedia of Philosophy , 2012. [Online]. Available: https://plato.stanford.edu/entries/voting-methods/. |
| [17] | R. a. B. R.-N. Baeza-Yates, Modern, Addison Wesley, 1999. |
| [18] | D. R. e. a. Radev, "Evaluation Challenges in Large-scale Document Summarization," in *ACL '03 Proceedings of the 41st Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics*, 2003, pp. pp.375-382. |
| [19] | S. G. a. B. C., "Term-weighting approaches in automatic text retrieval," in *Information Processing and Management 24, 1988.513-523.*, Reprinted in: Sparck-Jones, K.; Willet, P. (eds.) Readings in, pp. 323-328. |

# PHỤC LỤC

**GIAI ĐOẠN TIỀN XỬ LÝ VĂN BẢN**

import nltk

from pyvi import ViTokenizer, ViPosTagger

import numpy as np

import os

from string import punctuation

from tqdm import tqdm

class load\_data:

def read\_data\_train(path):

# xử lý văn bản

with open(path, 'r', encoding='utf-8') as f:

data = f.read()

find\_content = data.find("Content:")

doc = data [find\_content + 8: ]

sentences = nltk.sent\_tokenize(doc)

X = []

for sentence in sentences:

sent=[]

sentence\_tokenized = ViTokenizer.tokenize(sentence)

for word in sentence\_tokenized.split(" "):

if (word not in list(punctuation)):

sent.append(word)

X.append(" ".join(sent))

return X

def read\_data\_test(path):

# xử lý văn bản

f = open(path, 'r', encoding='utf-8')

data = f.read()

find\_summary = data.find("Summary:")

find\_content = data.find("Content:")

doc = data[find\_summary + 8: find\_content]

sentences = nltk.sent\_tokenize(doc)

X = []

for sentence in sentences:

sent=[]

sentence\_tokenized = ViTokenizer.tokenize(sentence)

for word in sentence\_tokenized.split(" "):

if (word not in list(punctuation)):

sent.append(word)

X.append(" ".join(sent))

return X

def tao\_danh\_sach(document):

# list các từ có trong văn bản

list\_word = []

# mang lưu số từ của 1 câu

word\_in\_sentence = []

# danh sách các từ trong câu

list\_sentence = []

for cau in document:

#cau=cau.lower()

cau = cau.split(" ")

word\_in\_sentence.append(len(cau))

list\_sentence.append(cau)

for word in cau:

list\_word.append(word)

return list\_word, word\_in\_sentence, list\_sentence

**TÍNH CÁC ĐẶC TRƯNG**

import numpy as np

**Trọng số TF.ISF**

def tinh\_TF\_ISF(sentence):

# tạo danh sách số từ trong văn bản,số từ của 1 câu,số câu trong văn bản

list\_word, word\_in\_sentence, list\_sentence = load\_data.tao\_danh\_sach(sentence)

# Tính TF

TF = np.array(word\_in\_sentence)

# -------------Tính ISF------------------------

SF\_dict = dict.fromkeys(list\_word, 0)

# tổng số câu (SF) có chứa từ X

for key in SF\_dict.keys():

count = 0

for n\_cau in list\_sentence:

dem\_in\_cau = n\_cau.count(key)

if(dem\_in\_cau > 0):

count += 1

SF\_dict[key] = count

# tinh ISF

import math

ISF\_word\_dict = {}

for k, v in SF\_dict.items():

s\_f = math.log10(len(list\_sentence)/v)

ISF\_word\_dict[k] = s\_f

#tinh ISF của câu

ISF\_sentence = []

for n\_cau in list\_sentence:

tong = 0

for k, v in ISF\_word\_dict.items():

if(k in n\_cau):

if(n\_cau.count(k) > 1):

v = v\*2

tong += v

ISF\_sentence.append(tong)

# tính trọng số

trong\_so\_TFISF=(np.array(TF)\*np.array(ISF\_sentence))/np.array(word\_in\_sentence)

# trọng số của 1 câu tính theo TF-ISF

if (max(trong\_so\_TFISF)):

trong\_so\_cau = np.array(trong\_so\_TFISF)/max(trong\_so\_TFISF)

else: trong\_so\_cau = np.array(trong\_so\_TFISF)

return trong\_so\_cau

**Xác suất từ**

#build từ điển số lần suất hiện của từ trong văn bản

def word\_dict(sentences,list\_word):

#tạo từ điển từ

wordict=dict.fromkeys(list\_word,0)

#số lần xuất hiện của từ trong toàn bộ văn bản

for n\_cau in sentences:

#n\_cau=n\_cau.lower()

for tu in n\_cau.split(" "):

wordict[tu]+=1

return wordict

#tống số lần xuất hiện

def sum\_lan\_xuat\_hien\_tu\_in\_cau(wordict,list\_sentence):

solanxuathien=[]

for sentence in list\_sentence:

dem=0

for words, count in wordict.items():

if (words in sentence):

if(sentence.count(words) > 1):

count=count\*2

dem+=count

solanxuathien.append(dem)

return solanxuathien

def xac\_suat\_tu(sentences):

# tạo danh sách số từ trong văn bản,số từ của 1 câu,số câu trong văn bản

list\_word, word\_in\_sentence, list\_sentence = load\_data.tao\_danh\_sach(sentences)

#tạo từ điển

wordict=xac\_suat.word\_dict(sentences,list\_word)

#tính tổng lần xuất hiện của 1 từ

solanxuathien=xac\_suat.sum\_lan\_xuat\_hien\_tu\_in\_cau(wordict,list\_sentence)

#tính xác xuất thực thực từ

trongso=np.array(solanxuathien)/len(list\_word)/np.array(word\_in\_sentence)

return trongso

**Thực thể tên**

def tinh\_thuc\_the\_ten(sentences):

X = []

for sent in sentences:

sent = ViPosTagger.postagging(sent)

X.append(sent)

list\_cau = []

list\_nhan = []

for x in X:

for y in range(len(x)):

if(y % 2 == 0):

list\_cau.append(x[y])

else:

list\_nhan.append(x[y])

# tính số thực thể tên có trong câu s

N\_name = []

Nw = [] # số thực từ có trong câu

for loai in list\_nhan:

lent = len(loai)

if("Np" in loai) or ("Ny" in loai):

dem\_in= loai.count('Np')+loai.count('Ny')

else: dem\_in = 0

N\_name.append(dem\_in)

Nw.append(lent)

# tính trọng số câu dựa vào thực thể tên

trong\_so\_cau = np.array(N\_name)/np.array(Nw)

return trong\_so\_cau

**MODEL VOTING**

import numpy as np

from tienxulyvannban import load\_data

from thuc\_the\_ten import danh\_tu\_rieng

from tf\_isf import TF\_ISF

from xac\_suat\_tu import xac\_suat

import os

#---------TÍNH CÁC PHƯƠNG PHÁP ------------------------#

def phuong\_phap(dac\_trung\_1, dac\_trung\_2,data):

PP = np.array(dac\_trung\_1)\*np.array(dac\_trung\_2)

# sắp xếp

sentence\_dict = dict.fromkeys(data, 0)

n = 0

for cau in sentence\_dict.keys():

if(n < len(PP)):

sentence\_dict[cau] = PP[n]

n += 1

PP\_dict = {}

for ts in sorted(PP, reverse=True):

for cau\_n, gt in sentence\_dict.items():

if (gt == ts):

PP\_dict[cau\_n] = gt

sort\_PP = []

for key in PP\_dict.keys():

sort\_PP.append(key)

## tóm tắt văn bản bằng phương pháp :

cau\_pp = []

so\_cau\_tom\_tat = 3

for key in PP\_dict.keys():

if(len(cau\_pp) < so\_cau\_tom\_tat):

cau\_pp.append(key)

tom\_tat = []

for cau\_n in data:

if(cau\_n in cau\_pp):

cau\_n=cau\_n.replace("\_"," ")

tom\_tat.append(cau\_n)

return sort\_PP, tom\_tat

#------------------------------TÍNH THỨ TỰ ƯU TIÊN---------------------

def uu\_tien(phuong\_phap\_n,sentence):

thu\_tu = []

n = len(phuong\_phap\_n)

for sent in sentence:

if sent in phuong\_phap\_n:

thu\_tu.append(n-phuong\_phap\_n.index(sent))

return thu\_tu

# -------------------MODEL VOTING(BORDA)-------------------------------------------

def model\_voting(pp1,pp2,pp3,data\_train):

thu\_tu\_1 = Voting.uu\_tien(pp1,data\_train)

thu\_tu\_2 = Voting.uu\_tien(pp2,data\_train)

thu\_tu\_3 = Voting.uu\_tien(pp3,data\_train)

voting = np.array(thu\_tu\_1) + np.array(thu\_tu\_2) + np.array(thu\_tu\_3)

voting\_dict = dict.fromkeys(data\_train, 0)

n = 0

for cau in data\_train:

if n < len(voting):

voting\_dict[cau] = voting[n]

n += 1

cau\_dict = {}

for ts in sorted(voting, reverse=True):

for cau, gt in voting\_dict.items():

if (gt == ts):

cau\_dict[cau] = gt

sort\_cau\_voting=[]

so\_cau\_tom\_tat = 3 #số câu cần tóm tắt

for key in cau\_dict.keys():

if(len(sort\_cau\_voting) < so\_cau\_tom\_tat):

sort\_cau\_voting.append(key)

tom\_tat = []

for cau\_n in data\_train:

if(cau\_n in sort\_cau\_voting):

cau\_n=cau\_n.replace("\_"," ")

tom\_tat.append(cau\_n)

return tom\_tat

#hàm chạy model

def run\_model( train\_paths,dir\_path):

count = 0

for f in os.listdir(train\_paths):

path\_dir =os.path.join(train\_paths,f)

if os.path.exists(dir\_path +"\\voting summary") ==0 :

os.mkdir(dir\_path +"\\voting summary")

os.mkdir(dir\_path +"\\voting summary\\"+ f)

if os.path.exists(dir\_path +"\\pp1 summary") ==0:

os.mkdir(dir\_path +"\\pp1 summary")

os.mkdir(dir\_path +"\\pp1 summary\\"+f)

if os.path.exists(dir\_path +"\\pp2 summary") ==0:

os.mkdir(dir\_path +"\\pp2 summary")

os.mkdir(dir\_path +"\\pp2 summary\\"+f)

if os.path.exists(dir\_path +"\\pp3 summary") ==0:

os.mkdir(dir\_path +"\\pp3 summary")

os.mkdir(dir\_path +"\\pp3 summary\\"+f)

for path in os.listdir(path\_dir):

if path != ".DS\_Store":

file\_path = os.path.join(path\_dir, path)

for data in os.listdir(file\_path):

sentences = []

txt = os.path.join(file\_path, data)

data = load\_data.read\_data\_train(txt)

for sent in data:

sentences.append(sent)

tf\_isf = TF\_ISF.tinh\_TF\_ISF(sentences)

xac\_suat\_tu = xac\_suat.xac\_suat\_tu(sentences)

thuc\_the\_ten = danh\_tu\_rieng.tinh\_thuc\_the\_ten(sentences)

sort\_pp1,tom\_tat\_pp1 = Voting.phuong\_phap(tf\_isf, thuc\_the\_ten,sentences)

sort\_pp2,tom\_tat\_pp2 = Voting.phuong\_phap(tf\_isf, xac\_suat\_tu,sentences)

sort\_pp3,tom\_tat\_pp3 = Voting.phuong\_phap(thuc\_the\_ten, xac\_suat\_tu,sentences)

tom\_tat\_voting = Voting.model\_voting(sort\_pp1,sort\_pp2,sort\_pp3,sentences)

count +=1

with open("./data/voting summary/"+ f + "/"+str(count) + ".txt",'w', encoding='utf-8') as f1:

f1.write(".\n".join(tom\_tat\_voting))

with open("./data/pp1 summary/"+ f + "/"+str(count) + ".txt",'w', encoding='utf-8') as f2:

f2.write(".\n".join(tom\_tat\_pp1))

with open("./data/pp2 summary/"+ f + "/"+str(count) + ".txt",'w', encoding='utf-8') as f3:

f3.write(".\n".join(tom\_tat\_pp2))

with open("./data/pp3 summary/"+ f + "/"+str(count) + ".txt",'w', encoding='utf-8') as f4:

f4.write(".\n".join(tom\_tat\_pp3))

**ĐÁNH GIÁ ROUGE**

import os

from rouge import Rouge

from model\_Voting import Voting

**#** hàm đánh giá

def danh\_gia(summary\_path,test\_paths):

tamp1 = dict.fromkeys(['f','p','r'],0)

tamp2 = dict.fromkeys(['f','p','r'],0)

tamp3 = dict.fromkeys(['f','p','r'],0)

avg\_scores ={}

count = 0

for paths in os.listdir(summary\_path):

path = os.path.join(summary\_path,paths)

for file\_summary in os.listdir(path):

file\_summarys = os.path.join(path,file\_summary)

summary = open(file\_summarys , encoding='utf-8').read()

for file\_tests in os.listdir(test\_paths):

if file\_tests == paths:

file\_test = os.path.join(test\_paths,file\_tests)

for test\_summary in os.listdir(file\_test):

if test\_summary == file\_summary:

test\_summary = os.path.join(file\_test,test\_summary)

test = open(test\_summary , encoding='utf-8').read()

scores = rouge.get\_scores(summary,test)

for key in scores:

for indx , val in key.items():

if indx =="rouge-1":

for var,gt in val.items():

tamp1[var] +=gt

avg\_scores[indx]=tamp1

if indx=="rouge-2":

for var,gt in val.items():

tamp2[var] +=gt

avg\_scores[indx] =tamp2

if indx=="rouge-l":

for var,gt in val.items():

tamp3[var] +=gt

avg\_scores[indx] =tamp3

count +=1

for index ,val in avg\_scores.items():

for ind1,val1 in val.items():

avg\_scores[index][ind1] = val1 / float(count)

return avg\_scores

#------ hàm main ----------------------

dir\_path = os.path.dirname(os.path.realpath(os.getcwd()))

dir\_path = os.path.join(dir\_path, 'module voting\\data')

test\_paths = os.path.join(dir\_path,'test')

train\_paths = os.path.join(dir\_path, 'train')

summary\_voing\_path = os.path.join(dir\_path, 'voting summary')

summary\_pp1\_path = os.path.join(dir\_path, 'pp1 summary')

summary\_pp2\_path = os.path.join(dir\_path, 'pp2 summary')

summary\_pp3\_path = os.path.join(dir\_path, 'pp3 summary')

# chạy model tạo folder summary của từng phương pháp

Voting.run\_model(train\_paths,dir\_path)

#---------------------thực hiện đánh giá ROUGE ------------------------

rouge = Rouge()

# điểm đánh giá của từng phương pháp

pp1 = danh\_gia(summary\_pp1\_path,test\_paths)

pp2 = danh\_gia(summary\_pp2\_path,test\_paths)

pp3 = danh\_gia(summary\_pp3\_path,test\_paths)

voting = danh\_gia(summary\_voing\_path,test\_paths)

print(pp1)

print("\n")

print(pp2)

print("\n")

print(pp3)

print("\n")

print(voting)