

**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**



**TRẦN THỊ TUYẾT HOA**

**XÂY DỰNG HỆ THỐNG TRUY HỒI HỌC LIỆU  
CHO SINH VIÊN NGÀNH ĐIỆN - ĐIỆN TỬ**

**Chuyên ngành: Hệ thống thông tin**

**Mã số: 8.48.01.04**

**TÓM TẮT LUẬN VĂN THẠC SĨ**

**TPHCM - NĂM 2022**



Luận văn được hoàn thành tại:  
**HỌC VIỆN CÔNG NGHỆ BƯU CHÍNH VIỄN THÔNG**

Người hướng dẫn khoa học: **TS. TÂN HẠNH**

Phản biện 1: .....

Phản biện 2: .....

Luận văn sẽ được bảo vệ trước Hội đồng chấm luận văn thạc sĩ tại  
Học viện Công nghệ Bưu chính Viễn thông

Vào lúc: ..... giờ ..... ngày ..... tháng ..... .. năm .....

Có thể tìm hiểu luận văn tại:

- Thư viện của Học viện Công nghệ Bưu chính Viễn thông.



# MỞ ĐẦU

## 1. Lý do chọn đề tài

Hiện nay, với sự phát triển không ngừng của ngành công nghệ thông tin nên nhu cầu tìm kiếm dữ liệu, tài liệu môn học của sinh viên cũng được phát triển theo. Một sinh viên không cần phải mất nhiều thời gian và công sức đi đến thư viện của trường học để tìm kiếm tài liệu các môn học như trước đây nữa, mà cách tìm kiếm đã được thay đổi hoàn toàn nhanh chóng hơn chỉ trong vài cái click chuột trên bàn phím máy vi tính thông qua Google hay các địa chỉ Web thông dụng.

Bên cạnh đó trong bối cảnh Thế giới hiện tại, thì tình hình dịch Covid đang là vấn đề đáng quan tâm và lo ngại nên môi trường học tập của sinh viên – học sinh dần chuyển sang hình thức học trực tuyến đã được phổ biến rộng khắp các trường học ở Việt Nam. Với hình thức học trực tuyến – online thì nhu cầu tìm kiếm thông tin, tài liệu chính xác về môn học cũng được nâng cao theo. Khi một học sinh tìm lại liệu trên Web mạng thì sẽ có hàng loạt các nội dung liên quan có khi đúng, gần đúng, khi không

chính xác hoặc có nhiều địa chỉ truy cập xấu hiện ra. Lý do là vì phải đối mặt với nhiều nguồn tài liệu, hệ thống truy hỏi thông tin dựa vào từ khóa sẽ không mang đến kết quả chính xác chuẩn như mong muốn.

Trước vấn đề đó, bản thân là một giáo viên trong ngành điện – điện tử tại trường Trung cấp Kinh tế - kỹ thuật Tây Ninh mong muốn xây dựng một hệ thống truy hỏi cho sinh viên trong ngành truy nhập tìm kiếm học liệu một cách hiệu quả chính xác nhất thông qua nguồn thư viện tài liệu tin cậy tại trường.

Từ mong muốn và ý tưởng trên tôi cũng sự ủng hộ và đồng ý hướng dẫn của TS. Tân Hạnh tôi chọn đề tài luận văn: “Xây dựng hệ thống truy hỏi học liệu cho sinh viên ngành điện – điện tử”, luận văn hoàn thành sẽ góp phần vào việc giải quyết các vấn đề cần thiết cấp bách trong thực tế.

## **2. Tổng quan về vấn đề nghiên cứu**

Đề tài hướng đến xây dựng và áp dụng có hiệu quả hệ thống tìm kiếm học liệu cho sinh viên ngành điện – điện tử nhằm hỗ trợ kịp thời cho sinh viên làm tài liệu tham khảo học trực tuyến trong tình hình chung căn cứ

theo chỉ đạo của thủ tướng chính phủ về việc giãn cách xã hội và kế hoạch đào tạo cần phải hoàn thành của trường.

Để thực hiện được mục tiêu ý tưởng đề ra, đề tài cần phải nghiên cứu và tiến hành các nội dung sau:

- Tìm hiểu và phân tích nhu cầu các nội dung môn học liên quan đến ngành điện – điện tử.

- Nghiên cứu từ cơ sở lý thuyết về hệ thống tìm kiếm, truy hồi thông tin, phân loại thể loại văn bản tìm kiếm. Từ đó áp dụng làm nền tảng để xây dựng và triển khai ứng dụng hệ thống tìm kiếm thông tin.

### **3. Mục tiêu nghiên cứu**

- Mục tiêu chính: xây dựng hệ thống truy hồi học liệu cho sinh viên ngành điện – điện tử
- Mục tiêu cụ thể của hệ thống gồm có các chức năng:
  - Truy hồi thông tin theo từ khóa
  - Truy hồi thông tin theo từ khóa và loại văn bản
  - Chức năng học và phân loại văn bản theo thể loại học liệu

- Chức năng học và phân loại theo chủ đề thuộc lĩnh vực điện - điện tử.

## **4. Đối tượng và phạm vi nghiên cứu**

### ***4.1. Đối tượng:***

- Các lý thuyết về truy hồi thông tin (Information Retrieval - IR)
- Hệ thống tìm kiếm thông tin (Information Retrieval Systems- IRS)
- Nghiên cứu về các quá trình truy hồi thông tin, các hướng tiếp cận giải quyết bài toán về truy hồi thông tin.
- Nghiên cứu về phân loại ngữ nghĩa văn bản tự động dựa trên kỹ thuật máy học (machine learning techniques)
- Phân tích, khảo sát và xây dựng hệ thống truy hồi học liệu cho sinh viên ngành điện – điện tử tại trường trung cấp kinh tế kỹ thuật Tây Ninh.



#### **4.2 Phạm vi:**

- Học liệu thuộc ngành điện – điện tử
- Ngôn ngữ tiếng Anh, Việt
- Thể loại học liệu: giáo trình, sách tham khảo

#### **5. Phương pháp nghiên cứu**

- Phương pháp phân tích và tổng hợp lý thuyết về truy hồi thông tin, phân loại văn bản dựa trên học máy.
- Phương pháp thực nghiệm khoa học: Xây dựng mô hình ứng dụng nhằm đánh giá hiệu quả của giải pháp và đánh giá kết quả thực nghiệm.

# Chương 1: TỔNG QUAN VỀ TRUY HỒI THÔNG TIN

Hiện nay, truy hồi thông tin là vấn đề khá phổ biến trên toàn Thế giới. Nó đã, đang và sẽ được ứng dụng rộng rãi trong tất cả các lĩnh vực khi có nhu cầu tìm kiếm và truy hồi thông tin. Qua quá trình tìm hiểu từ những cơ sở lý thuyết đến thực tiễn về truy hồi thông tin, tôi đã xây dựng một giải pháp cơ bản nhằm giải quyết yêu cầu bài toán đặt ra với trình tự các phương pháp sau.

Tổng quan về quá trình truy hồi thông tin, chia làm 2 giai đoạn:

## 1. Giai đoạn tiền xử lý.

- + Xử lý ngôn ngữ tự nhiên.
- + Chỉ mục và đánh trọng số thuật ngữ liên quan.

## 2. Giai đoạn thu thập.

- + Xử lý truy vấn ứng dụng mô hình Boolean
- + Tìm kiếm thông qua thuật ngữ liên quan có

trong chỉ mục.

+ Xếp hạng thứ tự liên quan của tài liệu trả về.

+ Phản hồi độ liên quan: dùng công thức tính độ chính xác (Precision) và độ bao phủ (Recall) đánh giá mức độ liên quan của tài liệu trả về.

Bên cạnh các giải pháp trên, tôi đã áp dụng phần mềm mã nguồn mở Lucene để thực hiện giải quyết bài toán đưa ra một cách hiệu quả nhất.

Nội dung chương 1 là giới thiệu sơ lược về truy hồi thông tin. Các nội dung được đề cập đến như sau:

- Khái niệm về truy hồi thông tin
- Các giai đoạn trong quá trình truy hồi thông tin
- Giới thiệu phần mềm mã nguồn mở Lucene
- Các phương pháp giải quyết truy hồi thông tin.
- Đánh giá hiệu quả của việc truy hồi thông tin

## **1.1. Các khái niệm truy hồi thông tin**

IR là lĩnh vực khoa học máy tính chuyên về lý thuyết và thực hành tìm kiếm thông tin. Vì văn bản là

phương tiện phổ biến nhất được sử dụng để biến đổi và phân phối thông tin một cách hiệu quả, nên hầu hết các nghiên cứu về IR đều tập trung vào việc tìm kiếm thông qua các bộ sưu tập văn bản của tài liệu [25].

## **1.2. Quá trình truy hồi thông tin**

Truy hồi thông tin là hoạt động thu thập, hiển thị thông tin liên quan dựa trên các dữ liệu có sẵn hoặc từ thao tác cập nhật dữ liệu trên máy chủ [2].

Quá trình truy hồi thông tin là quá trình truy vấn dữ liệu từ nhu cầu của người dùng đến sự phản hồi của nguồn dữ liệu có sẵn được tạo ra từ trước.

Quá trình truy hồi chia làm 2 pha.

- **Pha 1 (Chỉ mục indexing)**
- **Pha 2 (Truy hồi)**

### **1.2.1. Giai đoạn tiền xử lý**

- Tiền xử lý tài liệu, tiến hành các bước sau:

+ Phân tích từ vựng [25].

+ Loại bỏ từ dừng (stopword)

+ Lấy gốc từ

- Đánh chỉ mục: cho phép tích hợp ngữ nghĩa thu được từ kho dữ liệu riêng.

Cấu trúc chỉ mục gồm tập hợp các thuật ngữ đã xử lý. Tần suất xuất hiện thuật ngữ càng nhiều thì tầm quan trọng của chúng càng lớn.

### ***1.2.2. Giai đoạn thu thập:***

- *Xử lý truy vấn*
- *Tìm kiếm*
- *Xếp hạng*
- *Phản hồi về độ liên quan*

## **1.3. Giới thiệu phần mềm Lucene**

Hiện nay trên thế giới có một số thư viện mã nguồn mở chuyên hỗ trợ xây dựng hệ thống tìm kiếm thông tin như: Lucene, Egothor, Xapian, MG4J,... và Lucene chính là thư viện mã nguồn mở được nhiều cá nhân, tổ chức lựa chọn và sử dụng nhiều nhất [24].

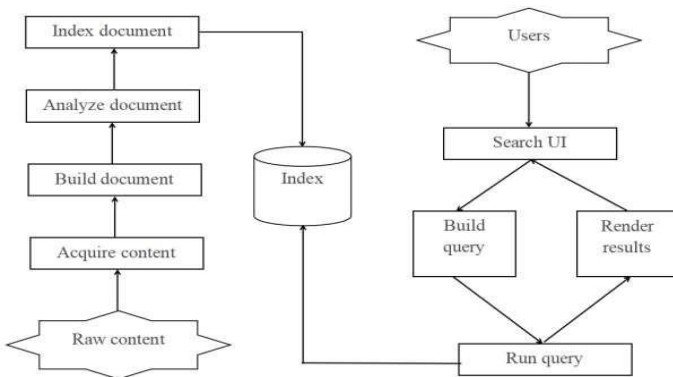
Luận văn kế thừa thư viện mã nguồn mở Lucene

để xây dựng truy hồi với hai thành phần chính là lập chỉ mục và tìm kiếm văn bản. Tìm hiểu tính năng, hoạt động của mã nguồn mở Lucene và sử dụng Lucene.Net để xây dựng thử nghiệm hệ thống tìm kiếm truy hồi thông tin.

### 1.3.1 Tổng quát

Lucene là thư viện hoạt động khá hiệu quả trong hệ thống truy hồi thông tin.

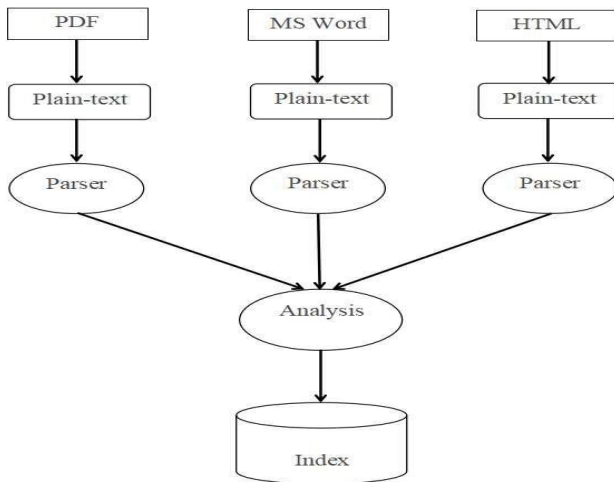
Lucene cho phép xử lý dữ liệu đầu vào dưới dạng văn bản nhằm tạo ra tập chỉ mục và cung cấp phương thức tìm kiếm trên tập chỉ mục đó. Đồng thời, nó cũng cho phép người dùng kế thừa, phát triển và thao tác với nhiều ngôn ngữ khác nhau.



**Hình 1.1: Sơ đồ Lucene trong hệ thống tìm kiếm truy hồi thông tin [25]**

### 1.3.2 Quy trình lập chỉ mục

Đầu tiên, chúng ta chuyển đổi tất cả nội dung trong các file dữ liệu như: PDF, MS Word, PDF, ... sang nội dung chứa dữ liệu dạng văn bản (text). Sau đó Lucene tiến hành phân tích và xử lý dữ liệu, loại bỏ từ vô nghĩa, tách các từ, cụm từ,... sau khi dữ liệu được phân tích xong sẽ chuyển sang đánh chỉ mục (Index) [19].



Hình 1.2: Quy trình đánh chỉ mục [18]

### 1.3.3 Các toán tử đánh chỉ mục cơ bản

Lucene hỗ trợ các toán tử đánh chỉ mục cơ bản như sau:

- Thêm tài liệu mới
- Xóa tài liệu (Remove Document)
- Cập nhật tài liệu

### ***1.3.4 Tối ưu hóa đánh chỉ mục***

### ***1.3.5 Bộ phân tích Analyzer***

## **1.4 Các phương pháp giải quyết vấn đề truy hồi thông tin**

- Các phương pháp tiếp cận dựa trên thống kê, các tài liệu thu thập được xếp hạng cao vì những tài liệu được xác định là phù hợp nhất cho truy vấn.

- Các loại hướng tiếp cận là Mô hình truy hồi Boolean (Boolean Retrieval Model) và mô hình không gian Vec-tơ (Vector Space Model).

+ Truy hồi Boolean dựa trên mệnh đề logic.

+ Mô hình Vec-tơ không gian, là các tài liệu các truy vấn được biểu diễn dưới dạng Vec-tơ.

## **1.5 Đánh giá hiệu quả của việc truy hồi thông tin**

Việc đánh giá mức độ chính xác của kết quả gọi là đánh giá truy hồi thông tin. Cùng với thước đo hiệu



suất phần mềm, hiệu suất truy hồi là vấn đề then chốt của hệ thống IR.



**Hình 1.3: Các tài liệu thu nhận được so với tài liệu liên quan.**

Nhận xét: Phần giao nhau giữa hai hình tròn nhỏ bên trong chính là phần mà dữ liệu tối ưu hóa được chọn.

## **Chương 2: CHỈ MỤC VĂN BẢN TỰ ĐỘNG**

Trong chương này nhằm mục đích phân loại tự động các văn bản thành các danh mục xác định trước và sắp xếp chúng để cho việc truy hồi linh hoạt và hiệu quả.

### **2.1 Học máy**

Học máy là một ứng dụng của Trí tuệ nhân tạo, là một lĩnh vực giúp hệ thống tự động hiểu được dữ liệu từ dữ liệu được đào tạo mà chúng ta không cần lập trình cụ thể. Học máy chia làm 3 phần: học có giám sát, học bán giám sát và học không giám sát.

Trong nghiên cứu này, tôi chọn phương pháp học máy có giám sát để áp dụng phân tích nội dung văn bản và trả về kết quả có nội dung liên quan đến truy vấn.

### **2.2 Phân loại văn bản**

Phân loại văn bản được áp dụng trong một số miền như: lập chỉ mục tài liệu dựa trên vốn từ vựng được kiểm soát, lọc tài liệu, phân loại cảm giác tài liệu...

Cách tiếp cận chủ đạo để phân loại văn bản dựa vào kỹ thuật máy học: là quy trình quy nạp chung tự tạo bộ phân loại bằng cách học từ một tập hợp các tài liệu đã

được phân loại trước dựa vào các đặc điểm của danh mục.

### ***2.2.1 Xử lý ngôn ngữ tự nhiên – thuật toán tách từ (tokenizer)***

### ***2.2.2 Loại bỏ từ dừng***

## **2.3 Chỉ mục văn bản**

### ***2.3.1 Tổng quan***

Các bước cơ bản trong xây dựng chỉ mục:

- Thu thập tài liệu cần chỉ mục
- Mã hóa văn bản
- Thực hiện tiền xử lý ngôn ngữ của mã thông báo
- Lập chỉ mục các tài liệu thuật ngữ xuất hiện.

### ***2.3.2 Xác định từ, cụm từ quan trọng để lập chỉ mục***

Từ, cụm từ quan trọng là từ có khả năng đại diện cho nội dung của tài liệu hay còn gọi mục từ. Mục từ là đơn vị cơ sở cho quá trình lập chỉ mục.

### ***2.3.3 Lập chỉ mục với Lucene***

Lucene sẽ quản lý chỉ mục trên thư viện tài liệu động, nó sẽ cập nhật rất nhanh khi thêm hoặc xóa bỏ tài liệu ra khỏi thư viện.

## 2.4 Đánh trọng số

Tiêu chí tầm quan trọng của thuật ngữ được sử dụng để xếp hạng mức độ liên quan của tài liệu với truy vấn có chứa thuật ngữ đó.

## 2.5 Các mô hình xếp hạng truyền thống

Mô hình xếp hạng là một hệ thống giải quyết hoặc xây dựng các vấn đề IR khác nhau.

Thông thường, mô hình xếp hạng được viết vắn tắt bằng bốn chữ cái D, Q, F, R.

### 2.5.1. Mô hình Boolean

### 2.5.2 Mô hình không gian Vec-tơ

## 2.6 Đánh giá hệ thống thông qua các độ đo

**Độ đo Precision** (Mức chính xác)

**Độ đo Recall** (Độ bao phủ, độ nhạy hoặc độ triệu hồi)

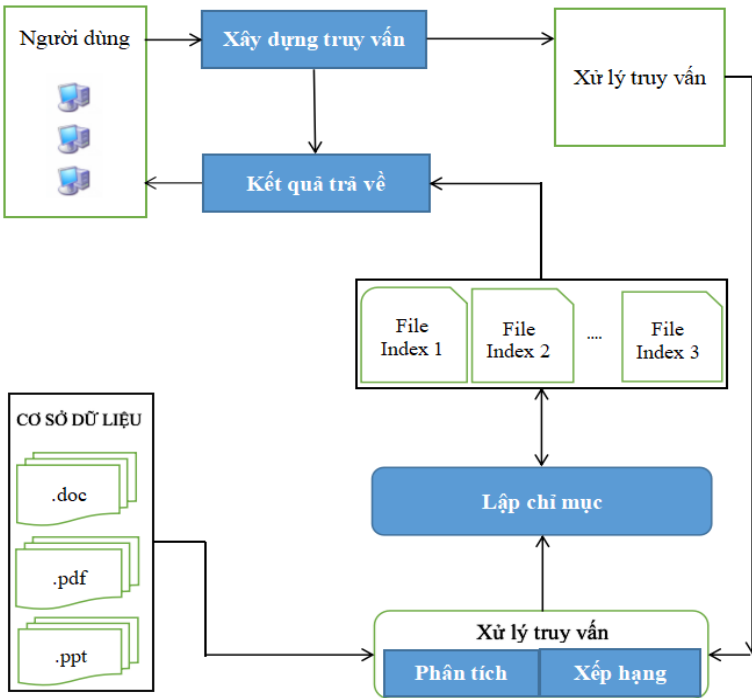
**Độ đo Accuracy** (Độ chính xác)

**Độ đo Specificity** (Độ đặc hiệu)

## Chương 3: XÂY DỰNG THỰC NGHIỆM HỆ THỐNG TRUY HỒI THÔNG TIN

Nội dung chương: mô tả cơ chế làm việc của hệ thống truy hồi thông tin; giới thiệu về thư viện mã nguồn mở Lucene và phần chạy demo thực nghiệm hệ thống.

### 3.1 Mô tả hệ thống



Hình 3.1: Mô tả hệ thống truy hồi thông tin

Hệ thống truy hồi thông tin được mô tả qua các thành phần sau:

- Xây dựng cơ sở dữ liệu
- Bước xử lý dữ liệu
- Thành phần lập chỉ mục

## **3.2 Dữ liệu**

### ***3.2.1 Loại tài liệu***

Nguồn dữ liệu được xây dựng dùng trong việc thử nghiệm là các file tài liệu chuyên ngành điện - điện tử được thể hiện trong tập tin word, ppt, pdf.

### ***3.2.2 Khối lượng tài liệu***

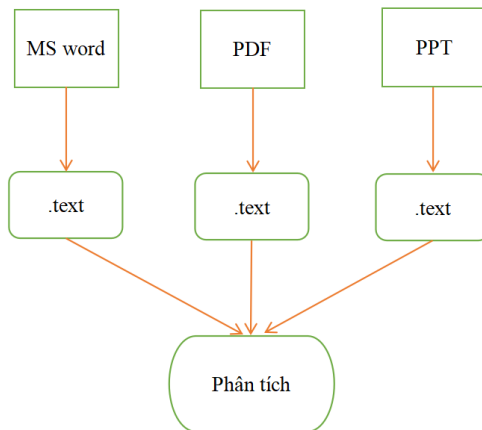
Thư viện lưu trữ 511 file dữ liệu, bao gồm 220 file tài liệu tiếng việt, 291 file tiếng anh. Được xây dựng từ thực tế giảng dạy của cá nhân và các giáo viên tại trường, còn lại là nguồn tài liệu sưu tầm từ Internet.

## **3.3. Tiền xử lý dữ liệu**

- Giai đoạn xử lý dữ liệu là quá trình xử lý các dữ liệu gốc nhằm nâng cao chất lượng dữ liệu và nâng cao

hiệu quả của việc tìm kiếm dữ liệu.

- Để Lucene dễ dàng phân tích và đánh chỉ mục, trước hết ta phải chuyển tài liệu về dạng văn bản thuần túy (.txt) từ những tài liệu đầu vào ở nhiều định dạng khác nhau như word, pdf...

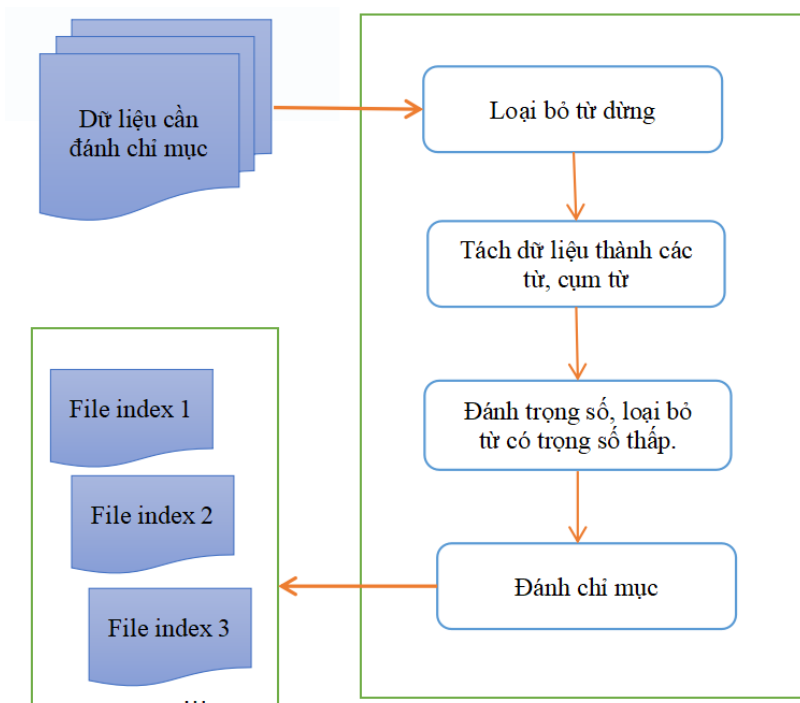


**Hình 3.2: Mô hình chuyển file văn bản**

### 3.4. Chỉ mục Lucene

Sau khi tiền xử lý dữ liệu ta tiến hành lập chỉ mục. Để chuẩn bị cho việc lập chỉ mục, Lucene sẽ phân tích dữ liệu, phân chia dữ liệu thành các chuỗi hoặc các ký tự thông qua lựa chọn các toán tử đã thực thi hoặc loại bỏ các từ không có nghĩa. Kế đến, Lucene sẽ lưu dữ liệu này theo cấu trúc chỉ mục. Cấu trúc này cho phép thực hiện

tìm kiếm nhanh hơn các từ khóa trong quá trình tìm kiếm.



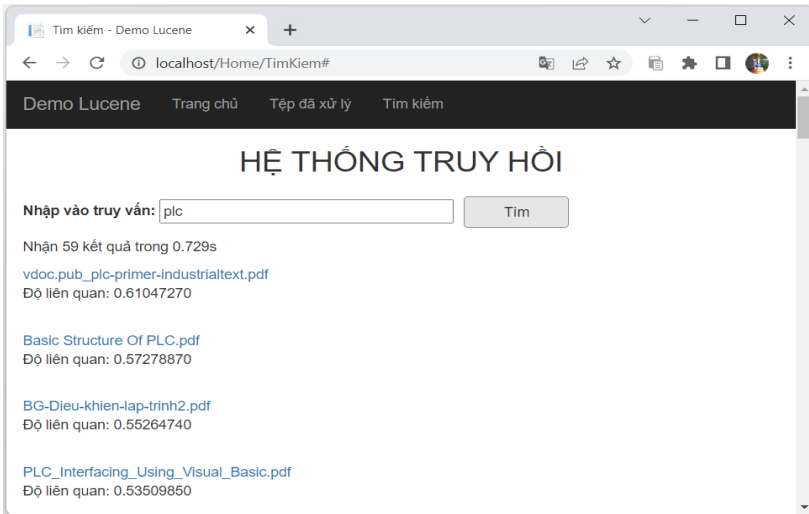
**Hình 3.3: Quy trình lập chỉ mục Lucene**

### 3.5. Thử nghiệm

Phần trình duyệt tìm kiếm được xây dựng ở dạng giao diện Web. Cho phép người dùng nhập các từ khóa tìm kiếm, hệ thống sẽ thực hiện tìm kiếm từ khóa trong file chỉ mục, sắp xếp kết quả và trả về danh sách các kết quả theo mức độ phù hợp giữa truy vấn và tài liệu trong



ơ sở dữ liệu chỉ mục.



**Hình 3.4:** Giao diện hệ thống sau khi truy hỏi thông tin

### 3.6. Đánh giá

Trong truy hỏi thông tin, độ chính xác (Precision) và độ bao phủ (Recall) được xác định theo nghĩa của một tập hợp các tài liệu được truy hỏi. [15].

#### 3.6.1 Độ chính xác ( $P$ )

#### 3.6.2 Độ bao phủ ( $R$ )

#### 3.6.3 Đánh giá kết quả thực nghiệm

Để minh họa thực nghiệm sử dụng 10 câu truy vấn

gồm tiếng anh và tiếng việt nhằm trải nghiệm độ tin cậy của hệ thống truy hồi.

**Bảng 3.1: Thống kê độ chính xác và độ bao phủ của hệ thống**

<b>ST T</b>	<b>Truy vấn</b>	<b>Tài liệu tìm được</b>	<b>Tài liệu liên quan</b>	<b>P (%)</b>	<b>R (%)</b>
1	Diode	47	35	41	74
2	Điện tử	64	22	57	34
3	Scada	19	15	82	79
4	PLC	59	51	62	86
5	Mạch điện	57	5	48.5	61
6	Electrical circuits	43	14	50	56
7	Vẽ điện	13	10	65	77
8	Circuit	27	6	60	70

	theory				
9	89C51	8	5	38.8	63
10	Sensors	59	5	35.8	76
<b>Giá trị trung bình</b>				<b>54%</b>	<b>71.2%</b>

Trong truy hồi văn bản trên một tập hợp tài liệu thì độ chính xác là số kết quả đúng chia cho số tất cả các kết quả được trả về. Độ chính xác tính đến tất cả các tài liệu đã truy hồi tuy nhiên nó cũng được đánh giá ở một thứ hạng nhất định, vì vậy chỉ nên xem xét các kết quả cao nhất được hệ thống trả về.

# KẾT LUẬN

## 1. Kết quả đạt được

Bài toán giải quyết được vấn đề tìm kiếm và truy hồi thông tin mang lại hiệu quả trong việc tập trung vào dữ liệu từng lĩnh vực nhằm tránh xử lý nguồn dữ liệu lớn không liên quan.

Luận văn tiếp cận đến nghiên cứu các vấn đề về truy hồi thông tin, các đánh giá về hệ truy hồi thông tin giúp xác định khả năng tự tìm kiếm của truy hồi thông tin; nghiên cứu và các kỹ thuật lập chỉ mục mô hình truy hồi thông tin; và phân loại văn bản dựa vào kỹ thuật máy học (Machine learning techniques). Từ đó thực hiện bài toán “Xây dựng hệ thống truy hồi học liệu cho sinh viên ngành điện – điện tử”.

Luận văn cũng tìm hiểu một cách hệ thống các tính năng và hoạt động của mã nguồn mở Lucene như: Lucene cung cấp khả năng phân tích dữ liệu, tiền xử lý, tạo chỉ mục cho các tài liệu để xây dựng nên hệ thống chỉ mục, cung cấp khả năng tiếp nhận các câu truy vấn của người dùng, thực hiện tìm kiếm dựa trên hệ thống

chỉ mục đã có và truy hồi kết quả tìm kiếm.

## **2. Hạn chế**

Bên cạnh những kết quả đạt được thì luận văn cũng có những mặt hạn chế như sau:

- Phần trình bày các nội dung của luận văn tương đối hạn chế dẫn đến tính thuyết phục của bài toán chưa cao.

- Khả năng áp dụng phương pháp IF và IDF để đánh trọng số và xếp hạng liên quan của tài liệu với truy vấn chưa phân tích hết mức độ liên quan của từng thuật ngữ trong tài liệu mà chỉ dựa trên số lần của từ xuất hiện trên văn bản, dẫn đến kết quả độ chính xác và độ bao phủ chưa cao.

## **3. Hướng phát triển**

- Tìm hiểu các cơ sở lý thuyết liên quan và các kỹ thuật học máy ứng dụng trong giải pháp giải quyết bài toán mang tính thuyết phục cao.

- Áp dụng kết hợp kỹ thuật phân tích ngữ nghĩa tiềm ẩn (LSA) trong tự nhiên và lập chỉ mục ngữ nghĩa

tiềm ẩn (LSI) với phương pháp TF và IDF trong việc đánh trọng số và lập chỉ mục để mang lại kết quả tìm kiếm truy hồi dữ liệu có độ chính xác hơn.

Thông qua cơ sở lý thuyết và bài toán thực nghiệm, tôi sẽ đề xuất áp dụng đề tài vào thực tế tại trường trung cấp kinh tế kỹ thuật Tây Ninh nơi tôi đang công tác, và có thể phát triển, thay đổi hệ thống để đưa ra khả năng tìm kiếm thông tin tốt nhất. Khi đó đề tài không những chỉ áp dụng cho sinh viên ngành điện – điện tử mà có thể áp dụng cho tất cả các ngành nghề đào tạo tại trường hay ở những trường học khác nhằm đáp ứng nhu cầu học tập của sinh viên học sinh.