ỨNG DỤNG KỸ THUẬT TÌM KIẾM THÔNG TIN VÀO HỆ THỐNG TRA CỨU TÀI LIỆU THƯ VIỆN TẠI TRƯỜNG ĐẠI HỌC TRÀ VINH

Nguyễn Ngọc Đan Thanh *

Tóm tắt

Lucene là một thư viện mã nguồn mở hỗ trợ các chức năng cần thiết của một hệ thống tìm kiếm thông tin. Thư viện Lucene được phát triển dựa trên nền tảng Java, sau đó được mở rộng ở nhiều ngôn ngữ lập trình khác nhau như Perl, Python, Ruby, C/C++, PHP, C#,... Trong bài báo này, tác giả sẽ trình bày tổng quan các vấn đề nghiên cứu về thư viện Lucene và triển khai ứng dụng tìm kiếm trên tài liệu thư viện tại Trường Đại học Trà Vinh. Kết quả của bài báo đề ra hướng tìm kiếm mới nhằm nâng cao chất lượng tìm kiếm thông tin.

Từ khóa: Tìm kiếm thông tin, mã nguồn mở Lucene, lập chỉ mục, mô hình không gian vector, truy tìm.

Abstract

Lucene is an open source library that supports some important features of an information retrieval system. It is developed based on Java programing language and is expanded to many other platforms such as Perl, Python, Ruby, C/C++, PHP, C#.

This paper will give an overview of Lucene and carry out the application in searching document in the Library of Tra Vinh University. The paper opens a new method in order to improve quality for searching information.

Keywords: information retrieval, Lucene open source, indexing, Vector Space Model, retrieval.

1. Giới thiệu về tìm kiếm thông tin

Khái niệm

Tìm kiếm thông tin (Information Retrieval - IR) là tìm kiếm tài nguyên (thường là các tài liệu - documents) trên một tập các dữ liệu phi cấu trúc (thường là văn bản dạng text) được lưu trữ trên máy tính nhằm thỏa mãn nhu cầu về thông tin (Hồ Bảo Quốc), (Huỳnh Đức Việt, Võ Duy Thanh, Võ Trung Hùng).

Nguyên tắc hoạt động

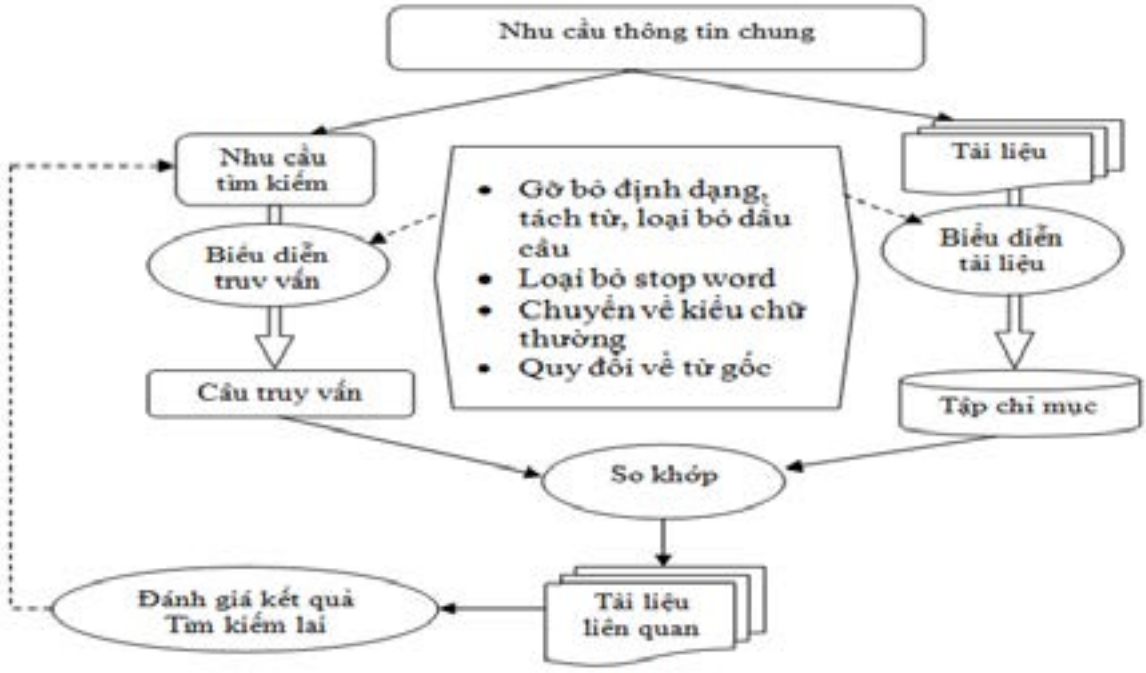
Nguyên tắc hoạt động cơ bản của hệ thống tìm kiếm thông tin là so khớp nhu cầu thông tin của người sử dụng với các tài liệu được lưu trữ trong cơ sở dữ liệu của hệ thống. Đối tượng tài liệu thường là văn bản hoặc những loại dữ liệu khác như hình ảnh, đồ thị,…

Bắt đầu với nhu cầu chung về quản lý và truy tìm thông tin, các tài liệu thô ban đầu như sách, tạp chí,… sẽ được lưu trữ dưới dạng tài liệu điện tử. Qua quá trình xử lý, các tài liệu này sẽ được chuyển sang biểu diễn dưới dạng cấu trúc đặc biệt nhằm giúp hệ thống có thể truy tìm thông tin một cách tốt nhất. Quá trình này được gọi là lập chỉ mục (indexing). Sau quá trình lập chỉ mục sẽ thu được một tập chỉ mục có lưu trữ các tài liệu dưới dạng biểu diễn mới. Từ đó, mọi thông tin truy vấn sẽ thao tác trực tiếp trên tập chỉ mục này.

Cũng như các tài liệu, nhu cầu truy vấn của người sử dụng sẽ được phân tích và biểu diễn ở một dạng cấu trúc đặc biệt. Để xác định được các tài liệu liên quan, hệ thống sẽ tiến hành so khớp câu truy vấn và các tài liệu trong tập chỉ mục. Sau cùng sẽ tiến hành đánh giá kết quả, có thể dựa trên cách xếp hạng của tài liệu hay mức độ thoả mãn nhu cầu của người sử dụng.

Mô hình hoạt động cơ bản của một hệ thống IR (Hồ Bảo Quốc), (Tanveer Siddiqui) được minh họa cụ thể trong Hình 1.

* BM Công nghệ TT, Khoa Kỹ thuật và Công nghệ, Trường Đại học Trà Vinh



Hình 1: Mô hình hoạt động cơ bản của hệ thống tìm kiếm thông tin

Đánh giá kết quả tìm kiếm

Các phương pháp đánh giá dựa trên cơ sở nào cũng đều đòi hỏi một tập tài liệu và một câu truy vấn trên tập tài liệu đó. Giả sử rằng mỗi tài liệu có thể liên quan hoặc không liên quan đến câu truy vấn.

Độ chính xác (Precision): được định nghĩa là tỷ lệ của các tài liệu liên quan trong tập kết quả trả về, đo lường tính chính xác của hệ thống, hay rõ hơn là ước tính có bao nhiêu tài liệu thật sự liên quan được tìm thấy (Huỳnh Đức Việt, Võ Duy Thanh, Võ Trung Hùng), (Tanveer Siddiqui):

$$\text{Độ chính xác} = \frac{|\{\text{Tập tài liệu liên quan}\} \cap \{\text{Tập kết quả}\}|}{|\{\text{Tập kết quả}\}|}$$

Độ bao phủ (Recall): được định nghĩa là tỷ lệ của các tài liệu liên quan trong cơ sở dữ liệu tài liệu, đo lường tính toàn diện của hệ thống (Huỳnh Đức Việt, Võ Duy Thanh, Võ Trung Hùng), (Tanveer Siddiqui):

$$\text{Độ bao phủ} = \frac{|\{\text{Tập tài liệu liên quan}\} \cap \{\text{Tập kết quả}\}|}{|\{\text{Tập tài liệu liên quan}\}|}$$

Kết quả sai (Fall-out): được đo bởi tỉ lệ các tài liệu không có liên quan trả về trên tổng các tài liệu không liên quan (Huỳnh Đức Việt, Võ Duy Thanh, Võ Trung Hùng), (Tanveer Siddiqui):

$$\text{Kết quả sai} = \frac{|\{\text{Tập tài liệu không liên quan}\} \cap \{\text{Tập kết quả}\}|}{|\{\text{Tập tài liệu không liên quan}\}|}$$

Các mô hình tìm kiếm thông tin

Mô hình tìm kiếm thông tin định nghĩa nhiều mặt khác nhau của thủ tục truy tìm thông tin như cách biểu diễn các tài liệu và các câu truy vấn, cách hệ thống tìm kiếm các tài liệu liên quan đến câu truy vấn của người sử dụng hay cách xếp hạng các tài liệu tìm kiếm được. Hệ thống tìm kiếm thông tin gồm có mô hình biểu diễn cho các tài liệu, mô hình cho các câu truy vấn của người sử dụng và hàm so khớp các câu truy vấn với các tài liệu. Mục tiêu chính của mô hình là truy tìm tất cả các tài liệu liên quan đến câu truy vấn (Ayse Goker, John Davies), (Ricardo Baeza-Yates, Berthier Ribeiro-Neto), (Tanveer Siddiqui).

Có ba nhóm mô hình phổ biến là:

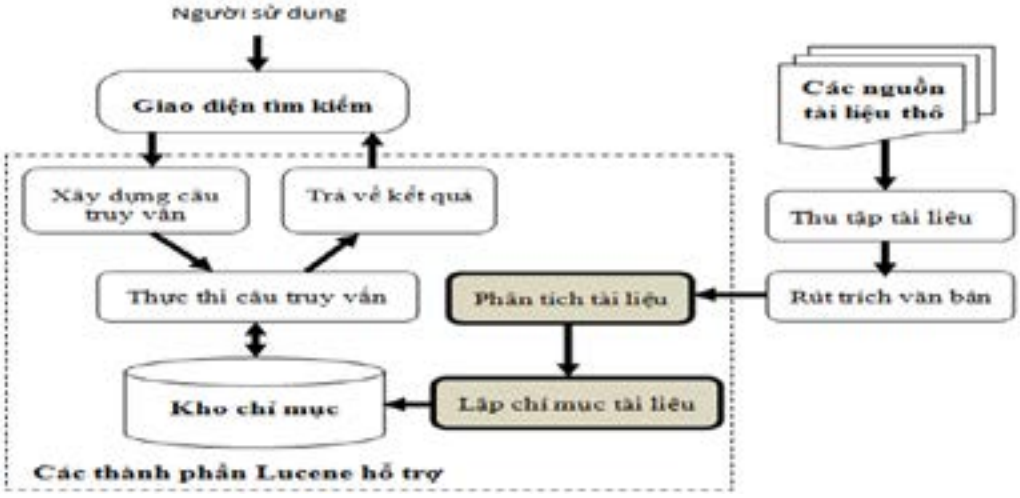
Mô hình cổ điển (Classical model) được xây dựng dựa trên kiến thức toán học. Mô hình này đơn giản, hiệu quả và dễ triển khai. Phần lớn các hệ thống thương mại hiện nay đều dựa trên các mô hình cổ điển.

Mô hình phi cổ điển (Non-Classical model) thực hiện truy tìm thông tin dựa trên các kỹ thuật logic riêng biệt (specical logic techique), lý thuyết tình huống (situation theory) hoặc các khái niệm về sự tương tác (concept of interaction).

Mô hình lựa chọn (Alternative model), đây là một trường hợp nâng cao của mô hình cổ điển. Nó sử dụng các kỹ thuật đặc biệt trong nhiều lĩnh vực khác nhau, gồm có một số mô hình như mô hình phân cụm (cluster model), mô hình mờ (fuzzy model).

Thư viện tìm kiếm toàn văn Lucene

Lucene không phải là một ứng dụng tìm kiếm hoàn chỉnh, nó chỉ là một thư viện mã nguồn mở, cung cấp các thành phần cần thiết của một ứng dụng tìm kiếm. Lập trình viên có thể tích hợp thư viện Lucene vào ứng dụng để sử dụng các tính năng sẵn có của nó hoặc mở rộng thêm một số thành phần khác phù hợp với ứng dụng của mình. Lucene hỗ trợ hai thành phần chính: lập chỉ mục và tìm kiếm (Michael McCandless, Erik Hatcher, Otis Gospodnetić).



Hình 2: Các thành phần cơ bản của một ứng dụng tìm kiếm

Các lớp đối tượng lập chỉ mục

IndexWriter: Lớp đối tượng trung tâm của tiến trình lập chỉ mục.

Directory: Lớp đối tượng xác định vị trí của tập chỉ mục.

Analyzer: Được sử dụng để phân tích văn bản trước khi được lập chỉ mục.

Document: Một lớp đối tượng biểu diễn tập hợp các trường, mỗi trường sẽ chứa nội dung văn bản cần lập chỉ mục.

Field: Trường thông tin của tài liệu. Mỗi Field sẽ có tên và giá trị phù hợp để lưu trữ một trường thông tin nhất định.

Các lớp đối tượng tìm kiếm

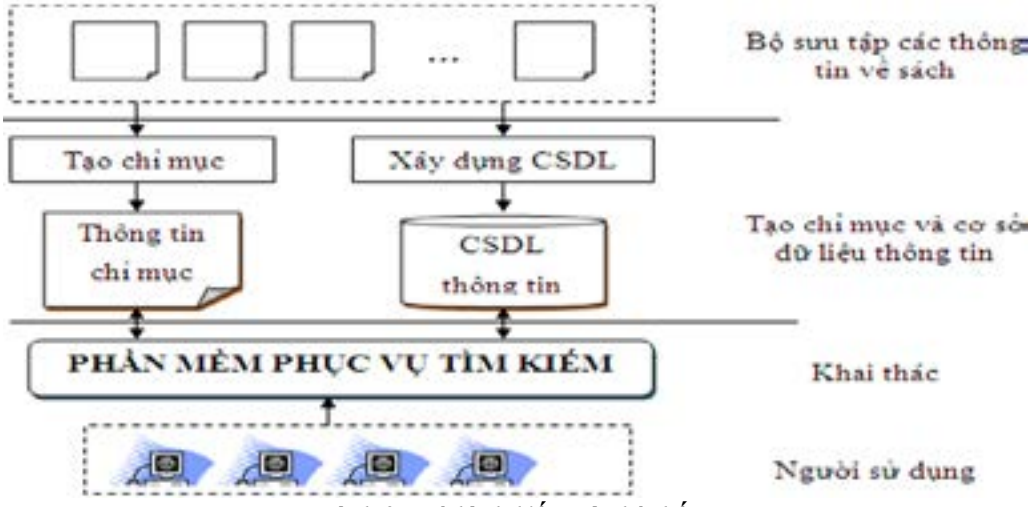
IndexSearcher: Mở tập chỉ mục đã được tạo trước bởi đối tượng IndexWriter và tiến hành tìm kiếm trên nó.

Term: Đơn vị cơ bản nhất để tìm kiếm.

Query: Lớp đối tượng truy vấn thông tin. Thư viện Lucene hỗ trợ một số loại truy vấn như: TermQuery, BooleanQuery, PhraseQuery, PrefixQuery,…

TopDocs: Lớp đối tượng đơn giản chứa liên kết đến N tài liệu có liên quan nhiều nhất đến câu truy vấn. Mỗi tài liệu trong danh sách sẽ có mã xác định docID để truy xuất đến tài liệu kết quả.

3. Xây dựng hệ thống tra cứu tài liệu



Hình 3: Mô hình kiến trúc hệ thống

Bộ sưu tập các thông tin về sách

Nguồn dữ liệu sách sử dụng được cập nhật từ kho sách của thư viện. Mỗi quyển sách trong thư viện khi được nhập về sẽ được biên mục nội dung và lưu trữ trên máy tính dưới dạng tập tin văn bản. Thông tin chung của mỗi quyển sách như: tên sách, tác giả, năm xuất bản, nhà xuất bản,… sẽ được ghi lần lượt trên từng dòng riêng biệt trong tập tin lưu giữ nội dung của quyển sách.

Hình 4: Cấu trúc mẫu lưu trữ nội dung một quyển sách

Tạo chỉ mục và cơ sở dữ liệu

Giai đoạn tạo chỉ mục sẽ sử dụng các lớp đối tượng được cung cấp bởi thư viện mã nguồn mở Lucene phiên bản 4.0: IndexWriter, Document, Analyzer, …

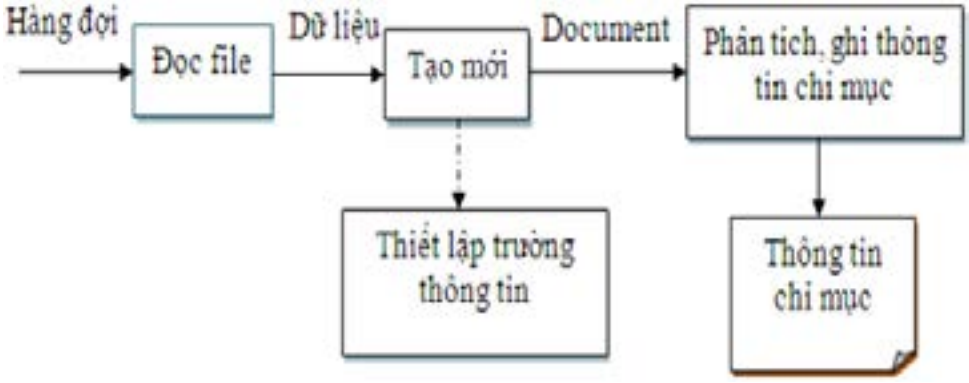
```
StandardAnalyzer analyzer = new StandardAnalyzer(Version.LUCENE_40);
FSDirectory dir = FSDirectory.open(new File(indexDir));
IndexWriterConfig config = new
IndexWriterConfig(Version.LUCENE_40, analyzer);
IndexWriter writer = new IndexWriter(dir, config)
```

Quá trình tạo chỉ mục sẽ thực hiện các công việc chính:

Đọc các tập tin văn bản đầu vào đưa vào hàng đợi.

Duyệt qua hàng đợi, tạo các đối tượng Document tương ứng với từng tập tin văn bản và thiết lập một trường thông tin cần thiết.

Tiến hành phân tích và ghi thông tin chỉ mục.



Hình 5: Tạo chỉ mục các tập tin văn bản đầu vào

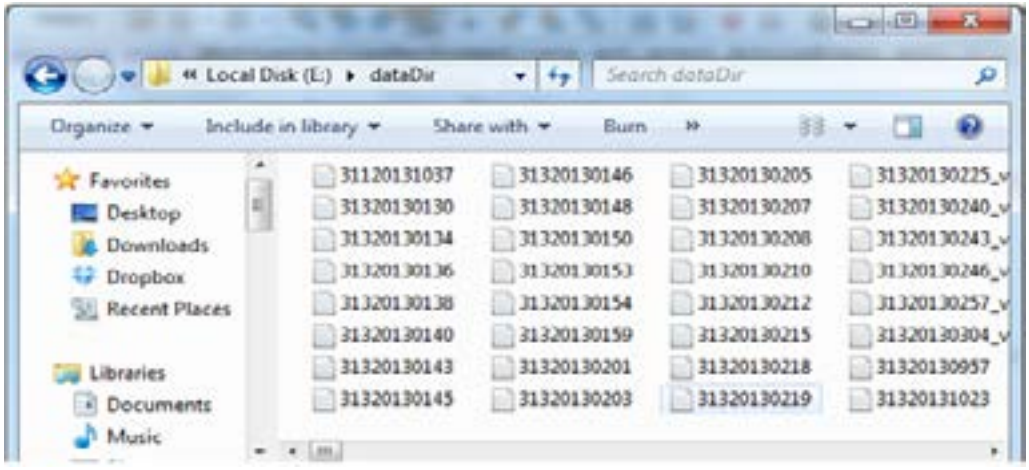
Ban đầu các tập tin văn bản chuẩn bị lập chỉ mục sẽ được lưu trữ trên máy tính, hệ thống sử dụng phương thức addFile (File f, Queue q) để lưu trữ các tập tin vào hàng đợi. Sau đó, thực hiện tạo ra các đối tượng Document, thiết lập các trường thông tin và ghi thông tin chỉ mục.

```
//Tạo mới đối tượng Document
Document doc = new Document();
//Đọc tập tin văn bản
FileReader fr = new FileReader(f);

/*Thiết lập các trường thông tin cho tài liệu*/
//Nội dung tài liệu
doc.add(new TextField("contents", fr));

//Đường dẫn tập tin văn bản
doc.add(new StringField("path", f.getPath(),
Field.Store.YES));

//Tên tập tin văn bản
doc.add(new StringField("filename",f.getName(),Field.Store.
YES));

//Tiêu đề quyển sách
doc.add(new TextField("ten_sach",ten_sach,Field.Store.YES));

//Tác giả quyển sách
doc.add(new TextField("tac_gia",tac_gia,Field.Store.YES));
…
//Ghi thông tin chỉ mục
writer.addDocument(doc);
```

4. Cài đặt thử nghiệm

Chuẩn bị dữ liệu

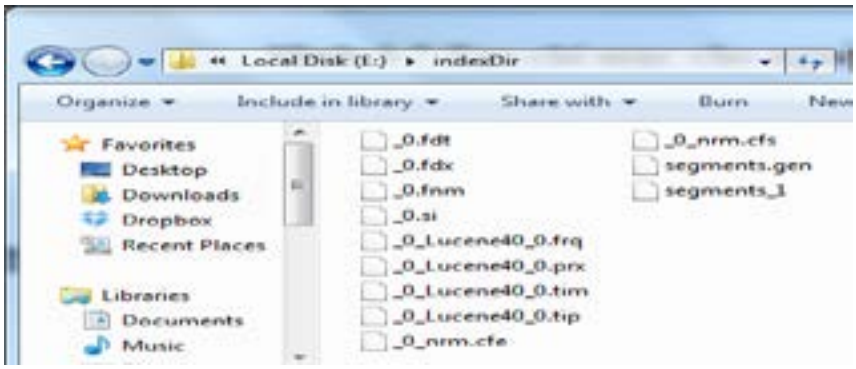
Dữ liệu đầu vào của hệ thống là bộ sưu tập các thông tin về sách. Đây là tập hợp các tập tin văn bản đã được biên mục sẵn. Các tập tin này có thể có phần mở rộng là .doc hoặc pdf. Tuy nhiên, trong phạm vi thử nghiệm của bài báo này, tác giả sử dụng một số tập tin văn bản thuần (.txt).



Hình 6: Bộ sưu tập các thông tin về sách

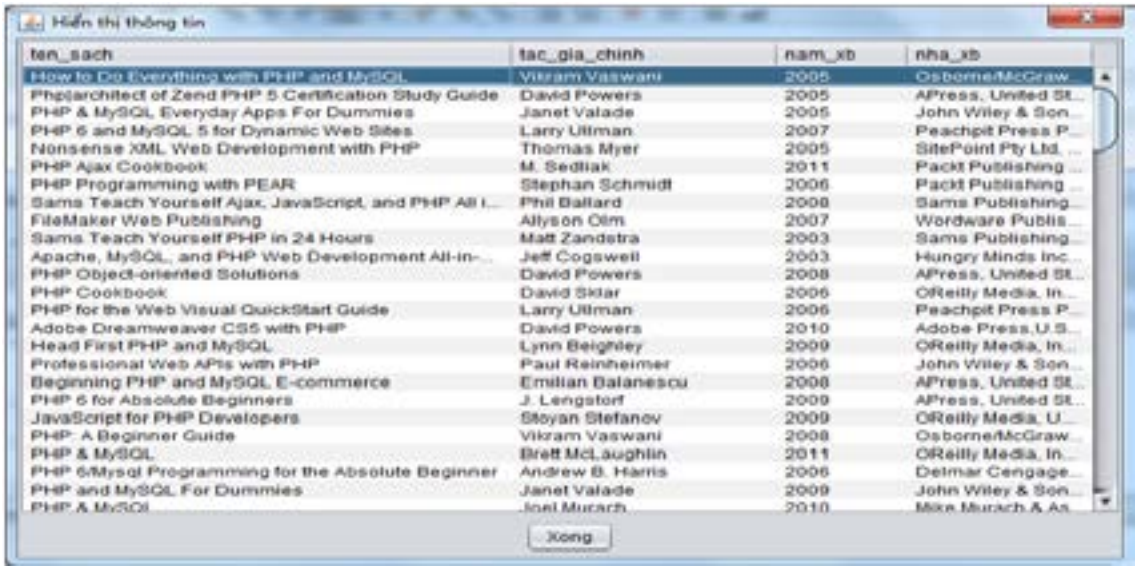
Tạo chỉ mục và cơ sở dữ liệu

Kết quả của tiến trình tạo chỉ mục sẽ phát sinh ra các tập tin lưu trữ thông tin chỉ mục theo cấu trúc quy định của thư viện Lucene.



Hình 7: Kết quả của quá trình tạo chỉ mục

Song song với tiến trình lập chỉ mục thì thông tin chung của các tài liệu tương ứng với nguồn dữ liệu đầu vào cũng được tự động lưu xuống cơ sở dữ liệu thông tin.



Hình 8: Kết quả thông tin sách lưu trữ trong cơ sở dữ liệu

Khai thác

Ứng dụng tìm kiếm phục vụ quá trình khai thác từ phía người sử dụng - Đây là gói ứng dụng web cho phép người sử dụng có thể tra cứu thông tin từ xa và nhận kết quả trả về trước khi họ có nhu cầu mượn tài liệu. Tùy theo nhu cầu của người sử dụng mà có thể thực hiện tra cứu thông tin theo hai hướng chính:

Hình 9: Các hướng khai thác thông tin

- Người sử dụng sẽ tra cứu dựa trên các thông tin chung của tài liệu. Khi đó, quá trình khai thác thông tin sẽ được truy xuất trực tiếp từ cơ sở dữ liệu.

- Người sử dụng sẽ tra cứu theo nội dung bên trong của tài liệu. Khi đó quá trình khai thác sẽ thực hiện dựa trên thông tin đã được lập chỉ mục trước đó - Đây cũng là hướng tra cứu mở rộng có ứng dụng kỹ thuật tìm kiếm thông tin mà hệ thống đang áp dụng.

Giả sử người sử dụng tìm kiếm thông tin dựa trên nội dung của các tài liệu ứng với nội dung tìm kiếm là "data structure" và thiết lập tùy chọn như Hình 10 thì kết quả trả về là danh sách các tài liệu có chứa các từ chỉ mục tương ứng các từ chỉ mục được phân tích trong nội dung tìm kiếm "data structure".

Hình 10: Ứng dụng tìm kiếm với giao diện web

Kết quả tìm thấy sẽ là các thông tin cơ bản cần thiết hỗ trợ bạn đọc mượn sách như: tiêu đề sách, tác giả, kho lưu trữ, …

Hình 11: Kết quả tìm kiếm với giao diện web

5. Kết luận

Trong bài báo, chúng đã trình bày các lý thuyết liên quan đến tìm kiếm thông tin cũng như các tiến trình phân tích, lập chỉ mục văn bản, … được hỗ trợ bởi thư viện tìm kiếm toàn văn Lucene. Từ đó, chúng tôi triển khai ứng dụng minh họa với hướng tìm kiếm mới kết hợp giữa tìm kiếm thông tin trên tập chỉ mục và trên cơ sở dữ liệu quan hệ. Mặc dù hệ thống còn đơn giản, chưa hỗ trợ phân tích các từ ngữ theo cấu trúc văn phạm tiếng Việt nhưng bước đầu nó cũng đạt được kết quả nhất định. Dựa trên hướng nghiên cứu này, hệ thống có thể mở rộng phát triển theo hướng kết hợp tìm kiếm ngữ nghĩa và khả năng tìm kiếm từ xa với máy chủ đa chỉ mục.

Tài liệu tham khảo

Ayse Goker, John Davies. 2009. *Information Retrieval: Searching in the 21st Century*, John Wiley and Sons.

Hồ Bảo Quốc. 2012. "*Giới thiệu về tìm kiếm thông tin*", Bài báo khoa học. Trường Đại học Khoa học tự nhiên TP. Hồ Chí Minh.

Huỳnh Đức Việt, Võ Duy Thanh, Võ Trung Hùng. 2010. "*Nghiên cứu ứng dụng mã nguồn mở Lucene để xây dựng phần mềm tìm kiếm thông tin trên văn bản*", Tạp chí Khoa học và Công nghệ - Đại học Đà Nẵng. Số 4 (39). tr307-316.

Michael McCandless, Erik Hatcher, Otis Gospodnetić. 2010. *Lucene In Action second Edition*. Manning Publications Co. The United States of America.

Ricardo Baeza-Yates, Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. Addison Wesley. 1st edition.

Tanveer Siddiqui. 2008. *Natural Language Processing and Information Retrieval*. Oxford University Press. India.