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

DOI:10.22144/ctu.jvn.2016.505

# ĐỀ XUẤT MÔ HÌNH QUẢN LÝ VÀ TRỰC QUAN HÓA KẾT QUẢ THỐNG KÊ VĂN BẢN TRỰC TUYẾN – ỨNG DỤNG TRONG PHÂN TÍCH XU HƯỚNG NGHIÊN CỨU KHOA HỌC TẠI TRƯỜNG ĐẠI HỌC CẦN THƠ

Nguyễn Hùng Dũng[1], Trương Xuân Việt[1], Trương Quốc Định[2], Lương Huy Nhật[2], Huỳnh Gia Khương[2] và Nguyễn Hoàng Việt[1]

[1]*Trung tâm Công nghệ Phần mềm, Trường Đại học Cần Thơ*
[2]*Khoa Công nghệ Thông tin & Truyền Thông, Trường Đại học Cần Thơ*

***Thông tin chung:***
*Ngày nhận: 05/04/2016*
*Ngày chấp nhận: 29/08/2016*

***Title:***
*Recommending model management and visualize statistical results online text - Applying the analysis of trends in scientific research at Can Tho University*

***Từ khóa:***
*Big Data, Distributed File System, Inverted Index, Full-text Search, Solr, Lucene*

***Keywords:***
*Big Data, Distributed File System, Inverted Index, Full-text Search, Solr, Lucene*

**ABSTRACT**

*The objective of the article is to propose a suitable management model which could be used to exploit rich and diversified data in different formats (i.e. text and spreadsheet). Besides, we also propose specific solutions based on a common Big Data platform, including: (1) HDFS (Hadoop Distributed File System) of Hadoop, which could be used in file management, (2) Lucene, which could be used to establish reversed indexing for text and (3) Apache Solr, which could be used to support reversed indexing management mechanism, full text searching and advanced searching functions. This article also presents experimental results, aggregates statistical results and displays statistical chart of applying the model into the analysis of trends in scientific research at Can Tho University.*

**TÓM TẮT**

*Mục tiêu của bài viết là đề xuất mô hình quản lý và khai thác hữu hiệu các dữ liệu phong phú, đa dạng đang tồn tại dưới dạng các văn bản, bảng tính của một tổ chức. Bên cạnh đó, chúng tôi cũng đề xuất giải pháp công nghệ cụ thể dựa trên các nền tảng Big Data phổ biến, bao gồm: (1) HDFS (Hadoop Distributed File System) của Hadoop dùng trong quản lý tập tin, (2) Lucene để lập chỉ mục nghịch đảo (Inverted Index) cho văn bản và (3) Apache Solr hỗ trợ cơ chế quản lý chỉ mục nghịch đảo, tìm kiếm toàn văn và một số chức năng tìm kiếm nâng cao. Bài viết cũng trình bày kết quả thực nghiệm, tổng hợp kết quả và trình bày biểu đồ thống kê của việc áp dụng mô hình trong phân tích xu hướng nghiên cứu khoa học tại Trường Đại học Cần Thơ.*

Trích dẫn: Nguyễn Hùng Dũng, Trương Xuân Việt, Trương Quốc Định, Lương Huy Nhật, Huỳnh Gia Khương và Nguyễn Hoàng Việt, 2016. Đề xuất mô hình quản lý và trực quan hóa kết quả thống kê văn bản trực tuyến - ứng dụng trong phân tích xu hướng nghiên cứu khoa học tại Trường Đại học Cần Thơ. Tạp chí Khoa học Trường Đại học Cần Thơ. 45a: 1-11.

## 1 GIỚI THIỆU

Trong những năm qua, việc triển khai các ứng dụng CNTT trong quá trình điều hành các hoạt động của tổ chức đang được chú trọng. Tuy nhiên, các tổ chức nói chung cũng như Trường Đại học Cần Thơ nói riêng chủ yếu tiếp cận cách phát triển các hệ thống thông tin với dữ liệu đã chuẩn hóa và có cấu trúc. Điều đó có nghĩa là chúng ta đã và đang lãng phí một nguồn dữ liệu khổng lồ dạng bán cấu trúc (semi-structured data) và phi cấu trúc (unstructured data). Với những ưu điểm và tác động mạnh mẽ của *Dữ liệu lớn* (Big Data) vào các ứng dụng liên quan, Big Data đang được xem như một yếu tố quyết định đến việc phát triển cũng như mang lại lợi thế cạnh tranh của các tổ chức.

Các nghiên cứu tích hợp giữa Hadoop và Solr (hoặc Elastic Search) đã được quan tâm và triển khai tại các khung tích hợp Cloudera,

Hortonworks. Alhabashneh và công sự cũng đề xuất khung tích hợp của bộ ba Hadoop, Solr và Tiki, hỗ trợ lập chỉ mục ngữ nghĩa cho văn bản (O.Alhabashneh *et al.*, 2011). Trên thực tế, các khung tích hợp này chủ yếu được cấu thành từ các thành phần nguồn mở và miễn phí, sau đó đóng gói và thương mại hóa. Chúng tôi nhận thấy đây là cách tiếp cận hợp lý và hữu hiệu cho mục tiêu xây dựng một bộ quản lý và hỗ trợ tìm kiếm tài liệu cục bộ của một tổ chức, tuy nhiên việc tìm kiếm văn bản tiếng Việt chưa được hỗ trợ. Trong Cloudera, bộ trực quan hóa dựa trên ZoomData, trong khi đó Hortonworks sử dụng Kibana cho khung tích hợp của họ. Sau khi đánh giá và lựa chọn bộ trực quan, chúng tôi nhận thấy Banana − một phiên bản mở rộng của Kibana − là lựa chọn phù hợp với bộ tìm kiếm Solr.

Trong bài viết này, chúng tôi đề xuất mô hình tích hợp mới và thêm những tính năng phù hợp với các tập dữ liệu tiếng Việt nhưng vẫn chưa tìm thấy trong các nghiên cứu liên quan, trong đó chúng tôi sẽ bắt đầu nghiên cứu xử lý dữ liệu để chạy các dịch vụ phân tích, xử lý và trả lời các yêu cầu truy vấn của người dùng. Chúng tôi sử dụng phần mềm nguồn mở Hadoop (Phần mềm nguồn mở của Apache) và các dịch vụ liên quan như giải pháp chính cho mục tiêu nghiên cứu: HDFS (quản lý các tập tin), Lucene/Solr (cung cấp các hàm cơ bản hỗ trợ cho việc đánh chỉ mục và tìm kiếm). Chúng tôi tích hợp thêm VnAnalyzer (Cao Mạnh Đạt, 2013) để hỗ trợ tìm kiếm văn bản tiếng Việt và Banana dùng cho việc trực quan hóa kết quả thống kê. Thêm vào đó, chúng tôi đã cài đặt, tích hợp thành công và ứng dụng mô hình trong phân tích xu hướng nghiên cứu khoa học tại Trường Đại học Cần Thơ dựa trên các bài báo khoa học được công bố bởi tạp chí khoa học của Trường, với kết xuất đầu ra là các kết quả tìm kiếm và các biểu đồ *đánh giá sự tương quan giữa các nghiên cứu trên tạp chí này với định hướng nghiên cứu khoa học ưu tiên* tại Trường Đại học Cần Thơ (theo biên bản họp số: 1919/BB-ĐHCT-HĐKHĐT ngày 30 tháng 09 năm 2015 của Trường Đại học Cần Thơ − được nêu chi tiết trong phần thực nghiệm).

Bài báo được cấu trúc như sau: chúng tôi sẽ điểm qua cơ sở lý thuyết liên quan ở Phần 2. Trong Phần 3, chúng tôi giới thiệu mô hình quản lý đề xuất tìm kiếm tài liệu và trực quan hóa kết quả thống kê trên nền Hadoop và Lucene/Solr. Phần 4 chúng tôi sẽ trình bày một số kết quả đạt được dựa trên mô hình đã đề xuất trong Phần 3, ứng dụng mô hình đề xuất trên tập dữ liệu Tạp chí khoa học Đại học Cần Thơ. Cuối cùng, chúng tôi đưa ra kết luận về kết quả nghiên cứu của mô hình đã đề xuất.

## 2 CƠ SỞ LÝ THUYẾT

### 2.1 Dữ liệu lớn (Big data)

Dữ liệu lớn là thuật ngữ dùng để mô tả các bộ dữ liệu có kích thước rất lớn, khả năng phát triển nhanh nhưng rất khó thu thập, lưu trữ, quản lý và phân tích với các công cụ thống kê hay ứng dụng cơ sở dữ liệu truyền thống. Các đặc trưng cơ bản của Big Data được thể hiện qua thuật ngữ 5V (Volume, Velocity, Variety, Veracity, Value) (Bernard Marr, 2015).

### 2.2 Hệ sinh thái Hadoop

Hadoop là một khung ứng dụng nguồn mở của Apache cho phép triển khai hàng loạt các kỹ thuật quản lý dữ liệu, tìm kiếm, khai phá dữ liệu lớn, cho phép các hệ thống có cấu trúc và không có cấu trúc trao đổi và làm việc với nhau một cách hiệu quả. Hadoop được biết đến với khái niệm một hệ sinh thái do các khả năng tích hợp với đa dạng các dịch vụ và có được các tính năng mạnh mẽ như:

− Khả năng mở rộng: Cho phép thay đổi số lượng phần cứng mà không cần thay đổi định dạng dữ liệu hay khởi động lại hệ thống.

− Hiệu quả chi phí: Hỗ trợ lưu trữ và xử lý song song trên những máy chủ bình thường.

− Linh hoạt: Hỗ trợ bất kỳ loại dữ liệu từ bất kỳ nguồn nào.

− Chịu lỗi: Thiếu dữ liệu và phân tích thất bại là hiện tượng thường gặp trong phân tích Big Data. Hadoop có thể phục hồi và phát hiện nguyên nhân thất bại do tắc nghẽn mạng.

### 2.3 Lập chỉ mục văn bản với Lucene

Lucene là một thư viện mã nguồn mở, được phát triển bởi Dough Cutting. Thư viện này cung cấp các hàm cơ bản hỗ trợ cho việc đánh chỉ mục và tìm kiếm thông qua các hàm API. Lucene có thể lập chỉ mục và hỗ trợ các thư viện tìm kiếm các loại dữ liệu văn bản đa dạng: .doc, .pdf, .html, v.v... Lucene ban đầu được viết hoàn toàn bằng Java, sau đó được phát triển trên nhiều ngôn ngữ khác như C/C++ (CLucene), .NET (Lucene.NET), Perl (Plucene), Ruby( Ferret) và đặc biệt là PHP (Zend Framework).

Để tiến hành đánh chỉ mục được trong Lucene, trước hết phải chuyển dữ liệu thành dạng văn bản thuần túy (plain text) như tập tin .txt chẳng hạn. Lucene sẽ phân chia dữ liệu thành các chuỗi hoặc là các ký tự thông qua việc lựa chọn các toán tử thực thi trên chúng. Sau khi dữ liệu được phân tích, nó sẽ sẵn sàng cho việc lập chỉ mục. Lucene sẽ chứa dữ liệu này theo cấu trúc chỉ mục nghịch đảo (Inverted Index). Nguyên tắc của nó là thay vì phải tìm kiếm các từ nào chứa trong tài liệu đó thì với

cấu trúc này sẽ tối ưu hóa việc tìm ra câu trả lời *"tài liệu nào chứa từ khóa này"*.

Lucene vẫn chưa xây dựng một bộ phân tích từ vựng dành riêng cho tiếng Việt, điều này có thể làm giảm tính hiệu quả của việc tìm kiếm. Nhận thấy vấn đề này, tác giả Cao Mạnh Đạt đã xây dựng một bộ phân tích từ vựng, gọi là VNAnalyzer dành cho Lucene. Bộ phân tích này dựa trên module VnTokenizer của tác giả Lê Hồng Phương (Le-Hong *et al*., 2008), cùng những cài đặt phù hợp để có thể sử dụng trên Lucene. VNAnalyzer hiện tại đã giải quyết được hai vấn đề cơ bản trong quá trình phân tích đó là tách từ và loại bỏ từ dừng.

### 2.4 Bộ tìm kiếm văn bản Apache Solr

Apache Solr là một nền tảng tìm kiếm toàn văn (full-text) mã nguồn mở dựa trên Apache Lucene, chức năng chính là tìm kiếm, đánh chỉ số, cung cấp API để làm việc. Solr nhập dữ liệu chủ yếu dưới dạng XML/HTML và JSON. Solr cũng có thể sử dụng thư mục để nhập khối dữ liệu lớn. Người dùng có thể truy vấn dữ liệu lớn này thông qua HTTP GET và nhận về kết quả dưới dạng XML hoặc JSON. Solr sử dụng Apache Lucene làm thư viện cho việc đánh chỉ mục và tìm kiếm.

Các chức năng cơ bản của Solr:

– Khả năng tìm kiếm văn bản toàn văn (Full-Text Search giống cách thức Google).

– Chỉnh sửa để hiệu năng tốt hơn.

– Dựa trên các chuẩn mở trong giao tiếp với các hệ thống khác như XML, JSON và HTTP.

– Quản trị dưới dạng giao diện HTML đơn giản.

– Thống kê dưới dạng JMX.

– Khả năng mở rộng ra nhiều máy chủ Solr.

– Cấu hình đơn giản dễ dàng với định dạng XML.

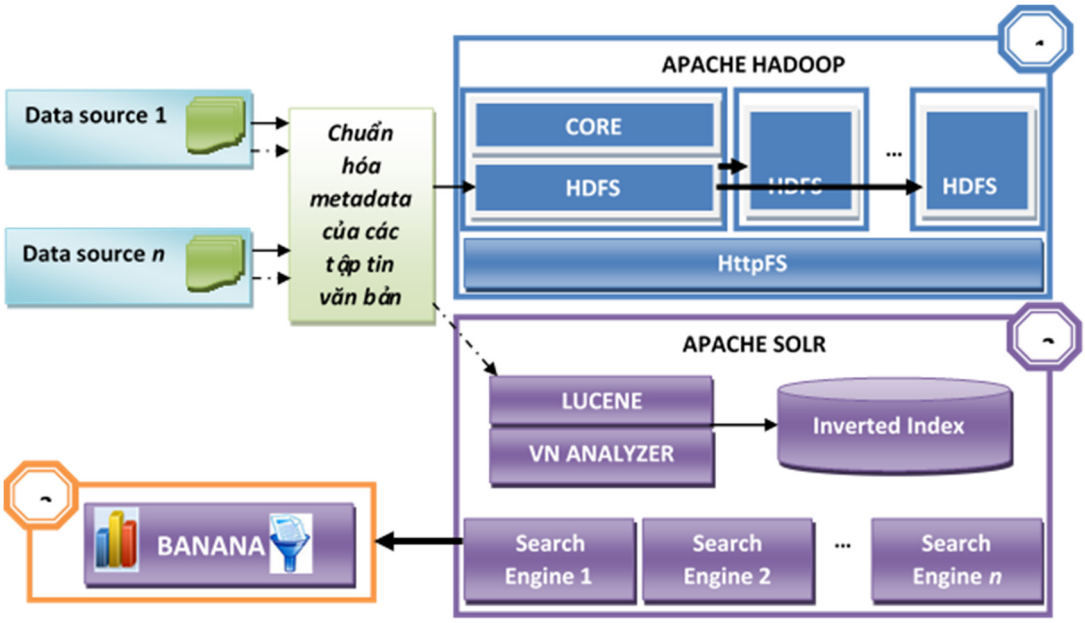
– Có khả năng bổ sung các phần mở rộng (plugin) mới. Ví dụ như phân tích mở rộng tiếng Việt: bắt lỗi chính tả, bỏ dấu…

### 2.5 Bộ công cụ trực quan hóa dữ liệu của Banana

Dự án Banana là một phân nhánh mã nguồn mở từ Kibana. Banana được xem như một công cụ có thể tạo ra các thống kê dữ liệu được lưu trữ trên Solr theo các dạng thống kê khác nhau. Việc kết hợp công cụ thống kê Banana vào Solr có thể giúp hiển thị dữ liệu một cách trực quan và đa dạng. Vì vậy, có giải quyết được nhiều vấn đề mà người dùng quan tâm về tập dữ liệu nhiều hơn và hơn hết là có thể khai thác được tập dữ liệu theo nhiều khía cạnh nhất có thể.

## 3 ĐỀ XUẤT MÔ HÌNH QUẢN LÝ, TÌM KIẾM TÀI LIỆU VÀ TRỰC QUAN HÓA KẾT QUẢ THỐNG KÊ

Trong bài viết này, chúng tôi đề xuất mô hình mới để quản lý và tìm kiếm văn bản với ba thành phần: (1) Hệ lưu trữ và phân phối tập tin dựa trên HDFS, (2) Hệ chỉ mục và tìm kiếm văn bản tiếng Việt dựa trên Lucene/Solr và (3) Bộ trực quan hóa dữ liệu. Dưới đây là mô hình và diễn giải từng thành phần trong mô hình mà chúng tôi đề xuất như sau:




**Hình 1: Mô hình quản lý và trực quan hóa kết quả thống kê văn bản**

Trong mô hình trên, dữ liệu đầu vào (Data source 1, 2…) của mô hình là các tập tin văn bản dạng .doc, .docx, .pdf, .xsl… và dữ liệu đầu ra là kết quả tìm kiếm theo từ khóa của người dùng, thống kê và trực quan hóa kết quả.

**Chuẩn hóa metadata**: trước khi nạp tài liệu vào HDFS, chúng tôi tiến hành chuẩn hóa metadata của tất cả các tập tin theo các trường (fields) như sau:

– tacgia: các tác giả tham gia NCKH.

– tuade: tên bài báo NCKH.

– ngaychapnhan: ngày bài báo NCKH được chấp nhận.

– donvi: tên khoa/đơn vị tác giả chính công tác.

– duongdan: thể hiện nơi lưu trữ tập tin.

Năm trường này được sử dụng cho việc thống kê và trực quan hóa dữ liệu bằng bộ công cụ của Banana.

Vai trò và chức năng cụ thể của từng thành phần trong mô hình là:

Hệ thống lưu trữ và phân phối tập tin dựa trên HDFS:

– Hệ thống dựa trên dịch vụ HDFS của Apache Hadoop.

– HDFS đóng vai trò tạo bản sao của dữ liệu nguồn và lưu trữ trên nhiều nút độc lập, đảm bảo an toàn dữ liệu và khả năng đáp ứng nhanh, mỗi văn bản nguồn cần quản lý đều có ít nhất một bản sao lưu tại một trong các nút của Hadoop.

Hệ thống chỉ mục, tìm kiếm văn bản và trình bày biểu đồ thống kê dựa trên Lucene/Solr:

– Hệ thống này cung cấp cơ chế lập chỉ mục nghịch đảo (Inverted Indexing) và máy tìm kiếm (Search Engine) cho văn bản nguồn.

– Kết quả tìm kiếm sẽ trả về văn bản gốc phù hợp đã được lưu trữ tại hệ thống lưu trữ (1). Do thư viện lập chỉ mục Lucene đã được tích hợp sẵn trong Apache Solr nên trên thực tế việc lập chỉ mục được tiến hành trực tiếp trên Solr mà không cần bổ sung bất cứ hỗ trợ nào khác.

– Việc thay thế các bộ phân tích ngôn ngữ cũng được dễ dàng cấu hình nên người dùng sẽ có thêm nhiều tùy chọn khi lập chỉ mục văn bản, cụ thể có thể thay thế ngôn ngữ mặc định tiếng Anh bằng các bộ phân tích ngôn ngữ tiếng Việt.

– Các chức năng tìm kiếm của Solr khá đa dạng và đáp ứng nhiều cách thức truy vấn khác nhau, trong đó chúng tôi tận dụng chủ yếu các tính năng nâng cao của tìm kiếm văn bản: tìm kiếm toàn văn (full-text search), tìm kiếm đa diện (faceted search), tìm kiếm theo điểm nhấn (hit highlighting). Bên cạnh đó, Solr cũng cung cấp cơ chế vận hành hiệu quả trên nhiều nút nhằm giúp tăng cường hiệu năng tìm kiếm của hệ thống.

– Trong Apache Solr, chúng tôi cũng tích hợp thêm bộ phân tích tiếng Việt đó là VnAnalyzer, giúp việc tìm kiếm thêm tài liệu với ngôn ngữ tiếng Việt được dễ dàng.

Bộ trực quan hóa dữ liệu:

– Đây là thành phần đóng vai trò lọc dữ liệu và trực quan hóa thống kê kết quả tìm kiếm được cung cấp bởi thành phần (2).

## 4 KẾT QUẢ THỰC NGHIỆM

Xây dựng hệ thống quản lý, tìm kiếm văn bản và trực quan hóa thống kê kết quả tìm kiếm để kiểm tra tính khả thi của các công nghệ đã được nghiên cứu, đồng thời ứng dụng hệ thống để *đánh giá sự tương quan giữa các nghiên cứu trên tạp chí này với định hướng nghiên cứu khoa học ưu tiên.* Ở đây, chúng tôi căn cứ theo các định hướng nghiên cứu của Đại học Cần Thơ tại Biên bản họp số 1919/BB-ĐHCT-HĐKHĐT ngày 30 tháng 09 năm 2015 của Trường Đại học Cần Thơ, theo đó các lĩnh vực ưu tiên trong nghiên cứu bao gồm: (a) Ứng dụng công nghệ cao trong nông nghiệp, thủy sản và môi trường; (b) Quản lý và sử dụng bền vững tài nguyên thiên nhiên; (c) Kỹ thuật công nghệ và công nghệ thông tin – truyền thông; (d) Khoa học giáo dục, luật và xã hội nhân văn; (e) Phát triển kinh tế, thị trường. Các lĩnh vực nghiên cứu này được sử dụng như các *từ khóa hoặc cụm từ khóa chính* để tìm kiếm và trực quan hóa. Chúng tôi tiến hành thực nghiệm trên tất cả 1.584 tập tin văn bản tạp chí Trường Đại học Cần Thơ từ năm 2011 đến 2015 (Nguồn: http://sj.ctu.edu.vn/ql/docgia/). Người dùng nhập từ khóa tìm kiếm thông tin, hệ thống xử lý và trả về kết quả tìm thấy. Đồng thời hệ thống sẽ kết xuất biểu đồ theo kết quả tìm kiếm tương ứng.

Để dễ dàng triển khai mô hình đề xuất trong Phần 3, chúng tôi đã xây dựng hệ thống thử nghiệm bao gồm 4 máy ảo. Chi tiết:

– Ba máy ảo chạy hệ thống HDFS của Hadoop để lưu trữ dữ liệu văn bản và 1 máy ảo Lucene/Solr cụ thể được liệt kê trong Bảng 1.

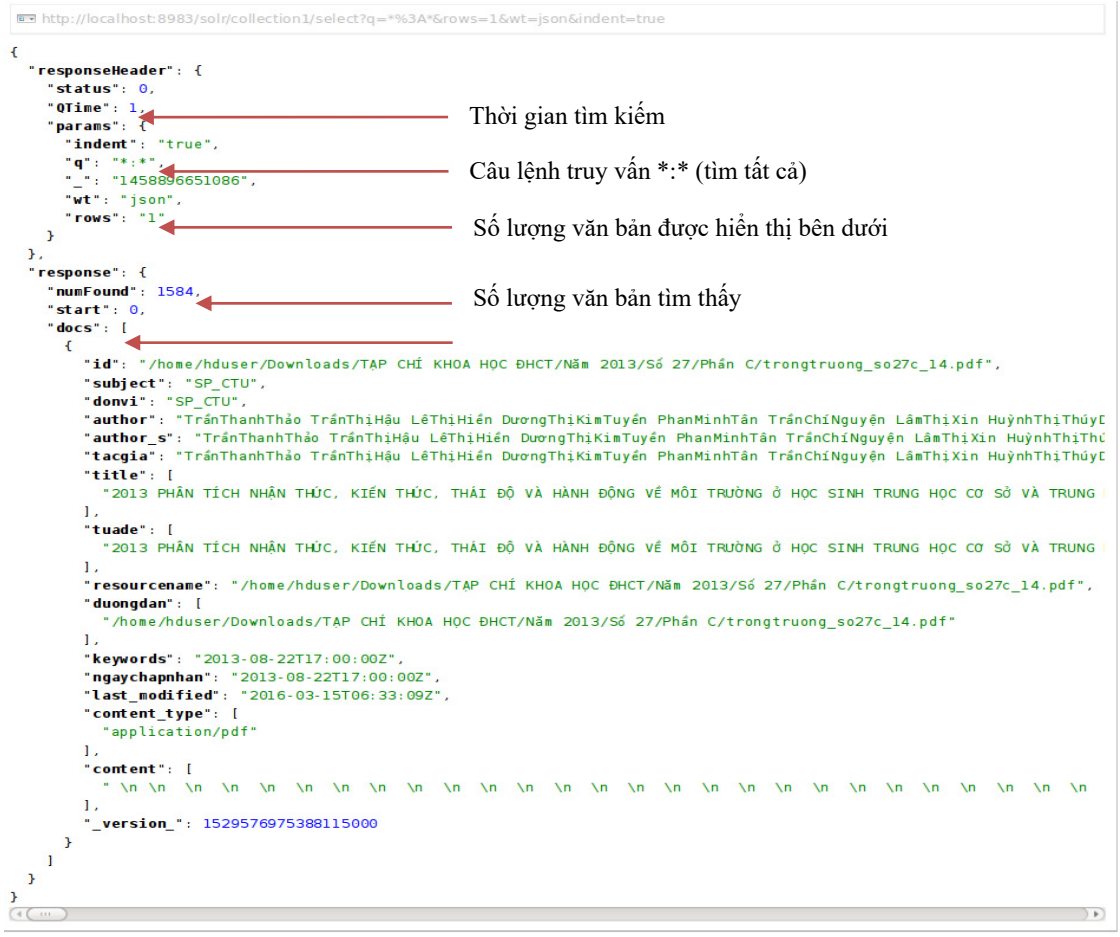
Khi tải lên các dữ liệu trên *master-node* (nút chính) dữ liệu sẽ được nhân rộng ra các slave-node (nút thứ cấp) còn lại. Chúng ta có thể truy cập vào địa chỉ của bất kỳ nút nào đang hoạt động để xem thông tin và lấy dữ liệu.

**Bảng 1: Liệt kê các máy chủ và chức năng của chúng trong hệ thống**

| Máy chủ | Tên máy chủ | Địa chỉ IP | Công việc |
|---|---|---|---|
| Master Node | Ctu1 | 192.168.0.1 | Máy chủ chính chạy Hadoop. |
| Slave node 1 | Ctu2 | 192.168.0.2 | Máy chủ thứ cấp 1 chạy Hadoop. |
| Slave node 2 | Ctu3 | 192.168.0.3 | Máy chủ thứ cấp 2 chạy Hadoop. |
| Lucene/Solr Server | Ctu4 | 192.168.0.4 | Một máy ảo chạy Solr để đánh chỉ mục và tìm kiếm văn bản, tích hợp thêm VnAnalyzer để hỗ trợ tìm kiếm văn bản tiếng Việt và Banana để phân tích và hiển thị biểu đồ thống kê kết quả. |

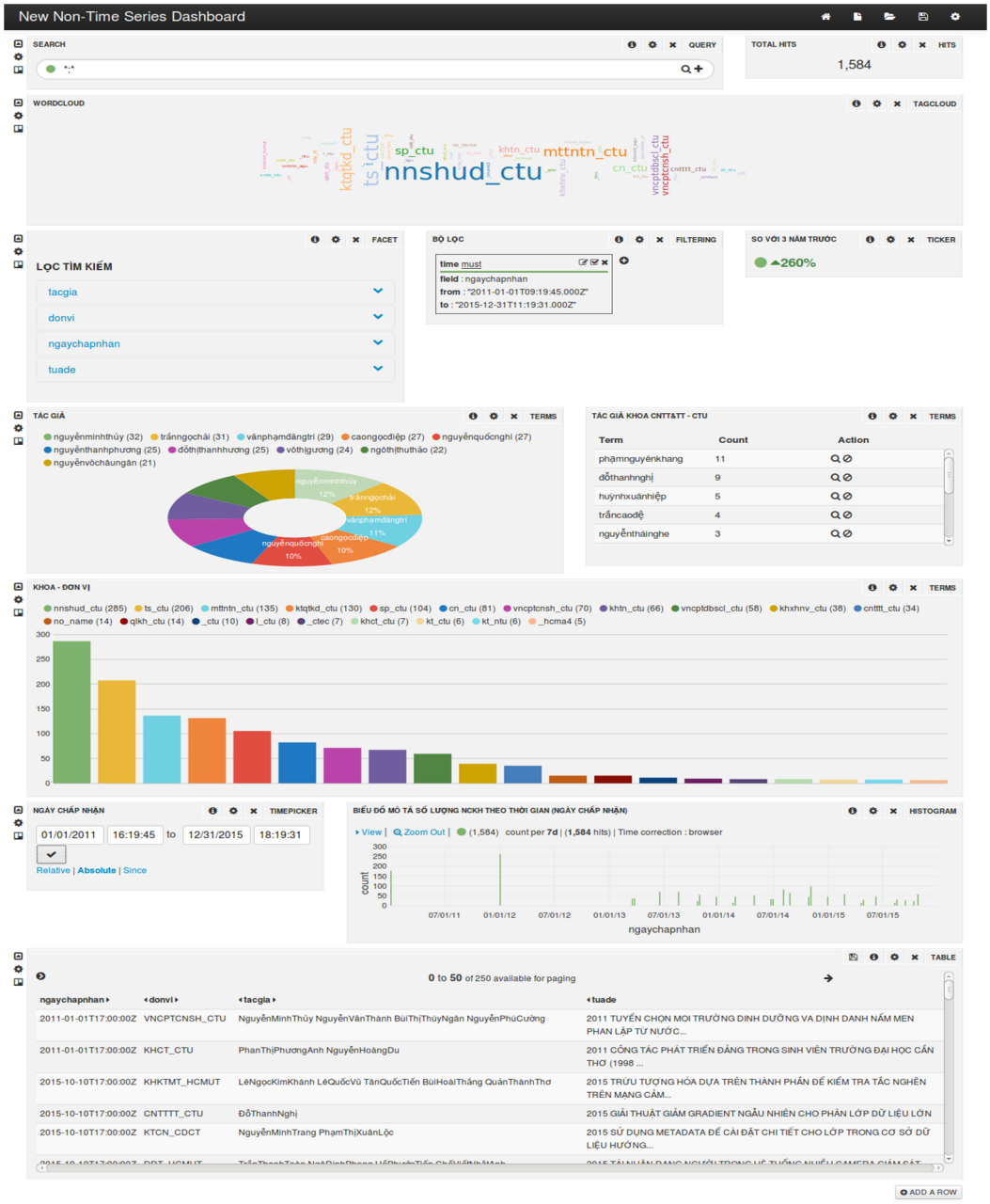
Sau khi tài liệu được đưa lên Solr, tất cả được đánh chỉ mục. Khi người dùng tìm kiếm theo tiêu chí nào đó, kết quả sẽ được kết xuất như Hình 2.

Tập dữ liệu "TẠP CHÍ KHOA HỌC ĐHCT" mà chúng tôi tạo ra chứa tổng cộng 1.584 bài báo nghiên cứu khoa học (NCKH) tại Trường Đại học Cần Thơ từ năm 2011 đến năm 2015, tất cả đã được đánh chỉ mục và định nghĩa các trường (fields) trong file schema.xml.




**Hình 2: Giao diện kết xuất dữ liệu khi tìm kiếm trực tiếp trên Solr**

Để thể hiện kết quả tìm kiếm một cách trực quan hơn, chúng tôi đã tích hợp vào hệ thống một giao diện người dùng thân thiện. Cách hiển thị kết quả thông qua giao diện này giúp người sử dụng có cái nhìn tổng thể và có thể so sánh về kết quả mà họ tìm kiếm. Giao diện được trình bày như sau:

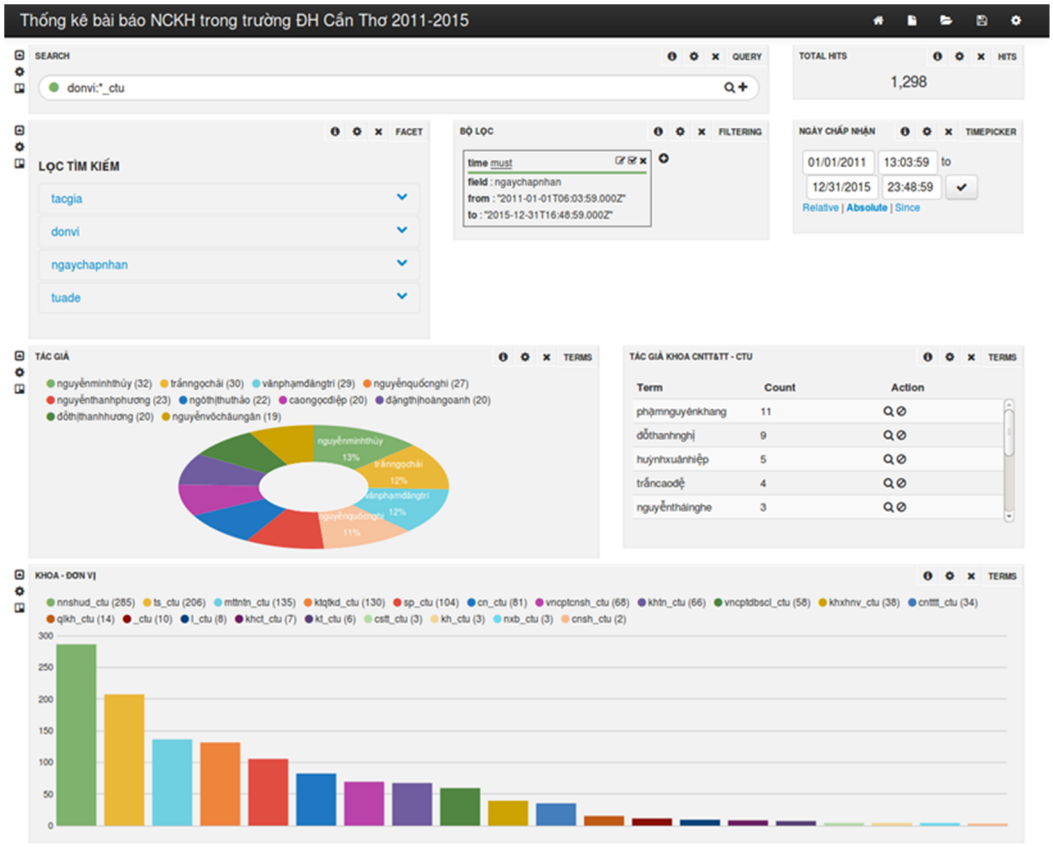



**Hình 3: Giao diện trực quan hóa kết quả tìm kiếm**

Dưới đây chúng tôi trình bày một số kết quả thực nghiệm điển hình về việc tìm kiếm, thống kê theo và trực quan hóa kết quả theo các từ khóa trên mô hình đã đề xuất như sau:

**(1) Tìm kiếm và thống kê bài báo NCKH tại Trường Đại học Cần Thơ trong 5 năm (2011-2015):**

Trường ***'donvi'*** được định nghĩa là khoa/đơn vị mà tác giả chính của bài báo NCKH công tác, để tìm kiếm những bài báo NCKH theo đơn vị thuộc Trường Đại học Cần Thơ, sử dụng truy vấn: ***donvi:*_ctu***.




**Hình 4: Thống kê bài báo NCKH tại Trường Đại học Cần Thơ theo khoa/đơn vị**

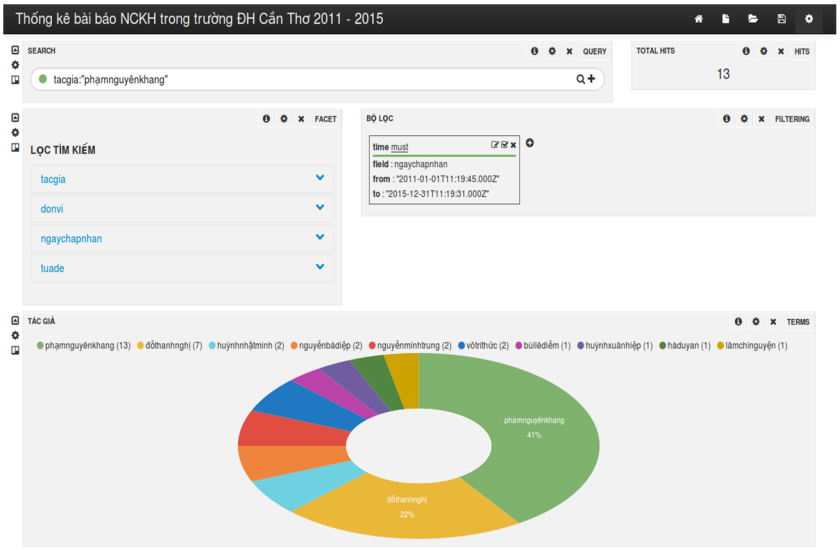
Kết quả trên tìm thấy có 1.298 bài báo NCKH được chấp nhận từ ngày 01/01/2011 đến ngày 31/12/2015. Khung **'Tác giả'** cho thấy biểu đồ thống kê theo số lượng đóng góp của các tác giả cho tạp chí. Chúng ta có thể thay đổi cách hiển thị danh sách tác giả (tăng dần hay giảm dần số lượng bài báo, số lượng tác giả, màu sắc biểu đồ,...) bằng cách nhấn chuột trái vào biểu tượng ⚙. Khung ***"Khoa – Đơn vị"*** cho thấy khoa Nông nghiệp – Sinh học ứng dụng (nnshud_ctu) có nhiều bài báo NCKH nhất (285 bài), khoa Thủy sản (ts_ctu) 206 bài, khoa Môi trường – Tài nguyên thiên nhiên (mttntn_ctu) có 135 bài, ...

Biểu đồ đường mô tả số lượng bài báo NCKH theo thời gian ở năm 2011 và 2012 không phân bố đều, tập trung vào 2 ngày là 01/01/2011 và 01/01/2012, lý do vì bài báo NCKH trong 2 năm này không có ngày chấp nhận mà chỉ có năm chấp nhận.

Qua kết quả thống kê, chúng ta dễ dàng nhận ra sự chênh lệch về số lượng bài báo NCKH giữa các khoa là khá lớn. Ngoài ra, năm 2014 là năm có số lượng bài báo nhiều nhất (397 bài).

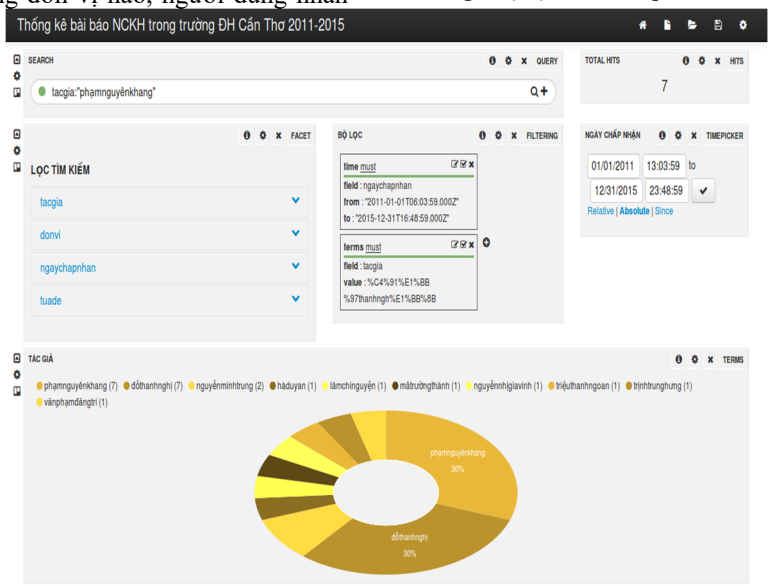
**(2) Tìm kiếm và thống kê bài báo NCKH theo tên tác giả**

Tên tác giả có thể truy vấn theo cấu trúc tacgia: "*<ten-tac-gia>*" hoặc tìm kiếm toàn văn với từ khóa "*<ten-tac-gia>*". Dưới đây là một ví dụ minh họa:




**Hình 5: Thống kê NCKH theo tên tác giả**

Để xem thống kê rõ hơn về mối tương quan giữa các tác giả, ví dụ hai tác giả khác nhau cùng nghiên cứu ở những đơn vị nào, người dùng nhấn chuột vào tên tác giả tương ứng ở biểu đồ hình tròn trong khung **'Tác giả'** để tạo thêm một bộ lọc ở khung **'Bộ lọc'** và kết quả được thống kê như sau:

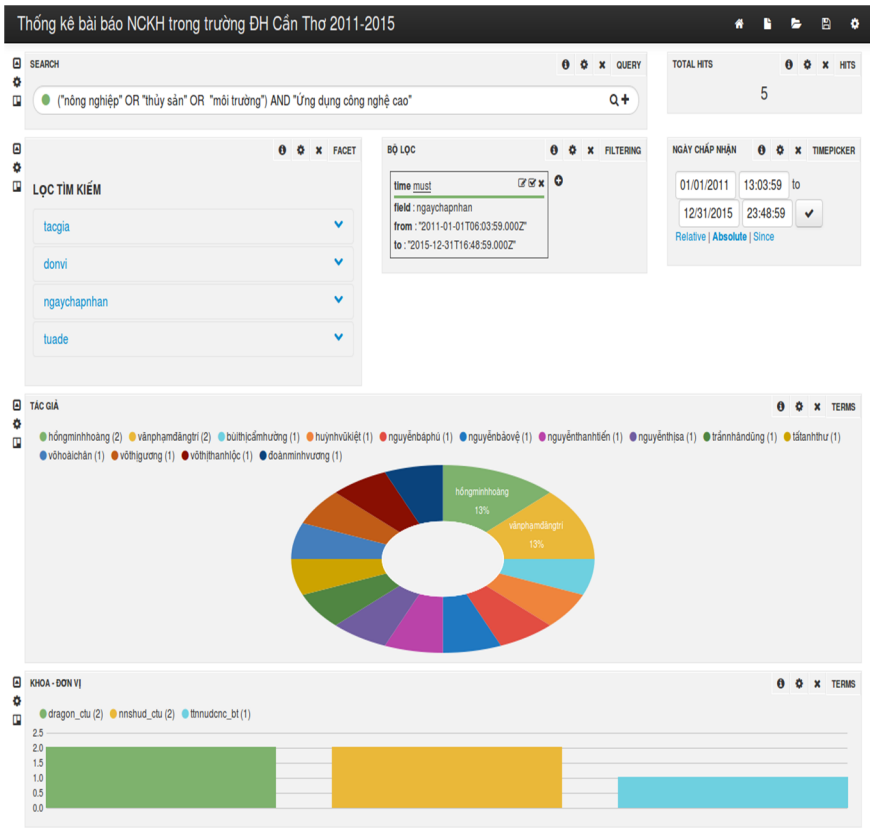



**Hình 6: Tương quan giữa hai tác giả**

Từ kết quả thống kê ở Hình 6 có thể thấy hai tác giả "Phạm Nguyên Khang" và "Đỗ Thanh Nghị" cùng tham gia nghiên cứu với tác giả Nguyễn Minh Trung (2 bài báo NCKH), Hà Duy An (1 bài), Lâm Chí Nguyện (1 bài),...Ngoài ra, số đơn vị mà tác giả Phạm Nguyên Khang tham gia để viết bài báo NCKH giảm từ 3 đơn vị xuống còn 2 đơn vị gồm Khoa CNTT-TT Đại học Cần Thơ (cntttt_ctu) là 6 bài và Khoa Kỹ thuật – Công nghệ Cao đẳng Cộng đồng Sóc Trăng (ktcn_stcc) là 1 bài, lý do vì tác giả Đỗ Thanh Nghị chỉ tham gia nghiên cứu chung ở 2 đơn vị này. Đồng thời có tổng cộng 7 bài báo NCKH do 2 tác giả trên viết chung ở hai năm 2013 và 2014.

**(3) Tìm kiếm và thống kê kết quả theo cụm từ**

Việc dùng các cụm từ tìm kiếm như "Ứng dụng công nghệ cao trong nông nghiệp, thủy sản và môi trường", "Quản lý và sử dụng bền vững tài nguyên thiên nhiên", "Kỹ thuật công nghệ và công nghệ thông tin – truyền thông",... và quan sát kết quả thống kê là điều có thể thực hiện được:




**Hình 7: Ứng dụng công nghệ cao trong nông nghiệp, thủy sản và môi trường**

Trong 3 đơn vị quan tâm đến "Ứng dụng công nghệ cao" có 2 đơn vị thuộc Đại học Cần Thơ: Viện nghiên cứu biến đổi khí hậu (dragon_ctu) với 2 bài, Khoa Nông nghiệp – Sinh học ứng dụng(nnshud_ctu) là 2 bài. Ngoài ra còn có trung tâm Nông nghiệp Ứng dụng công nghệ cao Bến Tre (ttnnudcnc_bt) là 1 bài. Kết quả tìm kiếm có 5 bài báo NCKH và có đến 4 bài được chấp nhận thời gian gần đây (từ cuối năm 2014 đến năm 2015). Từ đây có thể dự đoán được việc Ứng dụng công nghệ cao vào các lĩnh vực nông nghiệp, thủy sản và môi trường đang rất được quan tâm. Có thể loại bớt những kết quả thống kê của những năm trước (ví dụ không thống kê năm 2011) bằng cách sử dụng câu truy vấn: ***"nông nghiệp" OR "thủy sản" OR "môi trường") AND "ứng dụng công nghệ cao" -tuade:"2011*"***.

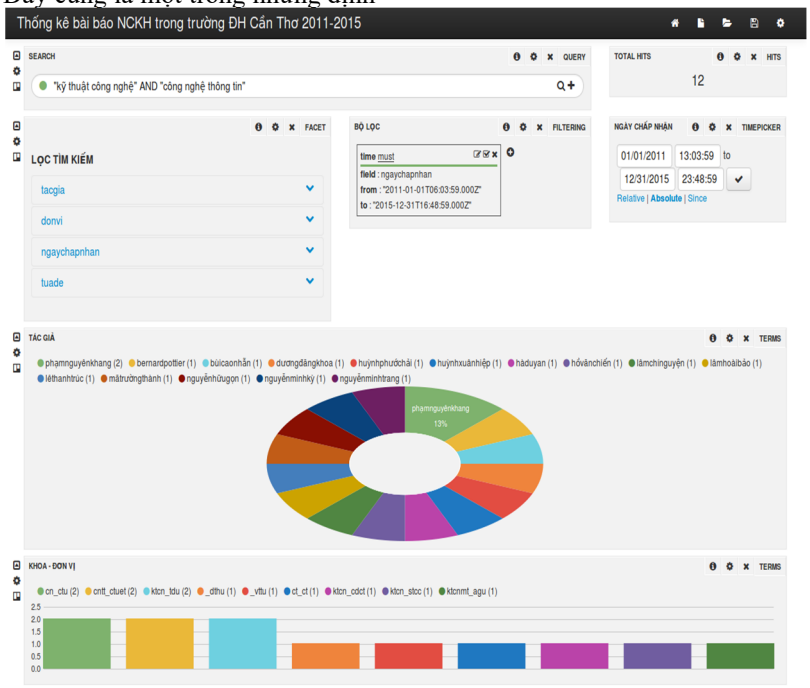
Các kết quả dưới đây, chúng tôi cho thấy được việc tìm kiếm đa dạng và phong phú hơn với việc kết hợp thêm các từ khóa để tìm kiếm:

***Hình 8*** với việc sử dụng từ khóa tìm kiếm: ***"tài nguyên thiên nhiên" AND "quản lý" AND "sử dụng" AND "bền vững"*** để cho thấy vấn đề quan tâm đến việc quản lý và sử dụng bền vững tài nguyên thiên nhiên như thế nào?

Có tổng cộng 55 bài báo NCKH liên quan đến vấn đề quản lý và sử dụng bền vững tài nguyên thiên nhiên. Khoa Môi trường – Tài nguyên thiên nhiên Đại học Cần Thơ (mttntn_ctu) đóng góp 27 bài, Khoa kinh tế - Quản trị kinh doanh (ktqtkd_ctu) với 3 bài, Viện nghiên cứu phát triển đồng bằng sông Cửu Long (vncptdbscl_ctu) là 3 bài... Khá nhiều khoa/đơn vị khác cũng tham gia NCKH về vấn đề này, cộng với việc tăng mạnh số lượng bài báo NCKH các năm gần đây (2013, 2014, 2015) nên có thể tạm kết luận, quản lý và sử dụng bền vững tài nguyên thiên nhiên đang được chú trọng phát triển, phù hợp với mục tiêu năm 2050 Việt Nam là quốc gia khai thác, sử dụng tài nguyên hợp lý, hiệu quả và bền vững (Nguồn http://www.vietnamplus.vn/quan-ly-su-dung-nguon-tai-nguyen-hop-ly-ben-vung/200977.vnp).


**Hình 8: Quản lý và sử dụng bền vững tài nguyên thiên nhiên**

Cuối cùng, chúng tôi trình bày kết quả tìm kiếm theo cụm từ khóa về "kỹ thuật công nghệ" và "công nghệ thông tin". Đây cũng là một trong những định hướng nghiên cứu khoa học được ưu tiên tại Trường Đại học Cần Thơ:




**Hình 9: Kỹ thuật công nghệ và công nghệ thông tin – truyền thông**

Có 12 bài báo NCKH liên quan đến Kỹ thuật công nghệ và Công nghệ thông tin – truyền thông. Những bài báo NCKH này được nghiên cứu ở các đơn vị về Kỹ thuật công nghệ như Khoa Kỹ thuật Công nghệ Cao đẳng Cần Thơ (cntt_cdct), Khoa Công nghệ (cn_ctu),... có cả trường Chính trị Thành phố Cần Thơ (ct_ct) cũng tham gia nghiên cứu.

## 5 KẾT LUẬN VÀ ĐỀ XUẤT

### 5.1 Kết luận

Trong bài viết này, chúng tôi đã đề xuất mô hình quản lý, tìm kiếm tài liệu và trực quan hóa kết quả thống kê dựa trên hai nền tảng Hadoop và Solr kết hợp một số thư viện của Lucene, bộ phân tích tiếng Việt và bộ công cụ trực quan hóa dữ liệu Banana. Mô hình đề xuất bao gồm 3 thành phần: (1) Hệ lưu trữ và phân phối tập tin dựa trên HDFS, (2) Hệ chỉ mục và tìm kiếm văn bản dựa trên Lucene/Solr, đối với văn bản tiếng Việt thì chúng tôi thay thế bộ phân tích của nó bằng VnAnalyzer và (3) Bộ trực quan hóa dữ liệu để thống kê và hiển thị biểu đồ bằng công cụ trực quan Banana. Mô hình này vừa đáp ứng nhu cầu tổng hợp và quản lý tập trung các nguồn dữ liệu phân tán của một tổ chức, vừa hỗ trợ hiệu quả cho việc lập chỉ mục, tìm kiếm và chỉ hướng nguồn dữ liệu. Các yếu tố liên quan đến cân bằng tải, tốc độ xử lý nhanh được chú trọng trong mô hình và được thể hiện trong hai thành phần (1) và (2) của mô hình, dựa trên cơ chế đa nút của Hadoop và Solr.

Cuối cùng, chúng tôi đã cài đặt, tích hợp thành công và ứng dụng mô hình trong phân tích xu hướng nghiên cứu khoa học tại Trường Đại học Cần Thơ với kết xuất đầu ra là các kết quả tìm kiếm và các biểu đồ cho thấy xu hướng nghiên cứu khoa học liên quan đến định hướng nghiên cứu khoa học ưu tiên tại Trường Đại học Cần Thơ. Đây cũng là công việc chưa được đề cập trong các nghiên cứu liên quan. Kết quả này có ý nghĩa thiết thực trong việc tìm kiếm, thống kê, kết xuất dữ liệu của một tổ chức khi các dữ liệu không phải ở dạng có cấu trúc như trước đây.

### 5.2 Đề xuất

Trong thực nghiệm, chúng tôi đã sử dụng 1.584 tập tin văn bản tạp chí của Trường Đại học Cần Thơ (http://sj.ctu.edu.vn/ql/docgia/). Tất cả các tập tin này, metadata chưa được chuẩn hóa nên việc tìm kiếm và kết xuất dữ liệu gặp rất nhiều khó khăn. Vì vậy, chúng tôi đề xuất các tập tin của bài báo trước khi được công bố cần được chuẩn hóa metadata theo chuẩn chung để có thể tìm kiếm, thống kê và kết xuất kết quả được dễ dàng. Ngoài ra, chúng tôi đề xuất ứng dụng mô hình này vào việc phân tích dữ liệu về NCKH cho Trường ĐHCT, điều này sẽ giúp cho các nhà quản lý có thêm thông tin để định hướng trong việc qui hoạch và xét duyệt các đề tài NCKH theo định hướng chung của Trường.

## TÀI LIỆU THAM KHẢO

Banana for Solr, 2015. [Online]. Available from: https://github.com/lucidworks/banana.

Bernard Marr, 2015. Why only one of the 5 Vs of big data really matters. [Online]. Available from: http://www.ibmbigdatahub.com/blog/why-only-one-5-vs-big-data-really-matters.

Cao Mạnh Đạt, 2013. Bộ phân tích từ vựng tiếng Việt cho Lucene. [Online]. Địa chỉ: https://caomanhdat.wordpress.com/2013/06/26/bo-phan-tich-tu-vung-tieng-viet-cho-lucene/.

Doug Cutting, 2013. Apache Lucene: Then and Now By Doug Cutting. [Online]. Available from: http://www.meetup.com/fr-FR/Hadoop-DC/events/140608632.

Khung tích hợp Cloudera, 2015. [Online]. Địa chỉ: http://www.cloudera.com.

Khung tích hợp Hortonworks, 2014. [Online]. Địa chỉ: http://hortonworks.com.

Le-Hong, P., T M H. Nguyen, A. Roussanaly, and T V. Ho, 2008. A hybrid approach to word segmentation of Vietnamese texts. Proceedings of the 2nd International Conference on Language and Automata Theory and Applications, Tarragona, Spain, Springer, LNCS 5196, pp. 240-249, 2008.

Lucene, 2015. [Online]. Available from: http://lucene.apache.org/solr/index.html.

O.Alhabashneh, R. Iqbal, N. Shah, S. Amin, A. James, 2011. Towards the Development of an Integrated Framework for Enhancing Enterprise Search Using Latent Semantic Indexing. In ICCS 2011, LNAI 6828, pp. 346–352, 2011, Springer-Verlag Berlin Heidelberg 2011. DOI: 10.1007/978-3-642-22688-5_29. ISBN: 978-3-642-22687-8.

Trương Quốc Định, Nguyễn Quang Dũng, 2012. Một giải pháp tóm tắt văn bản tiếng Việt tự động. Hội thảo quốc gia lần thứ XV: Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông- Hà Nội, 03-04/12/2012.