Kỷ yếu Hội nghị Khoa học Quốc gia lần thứ IX "Nghiên cứu cơ bản và ứng dụng Công nghệ thông tin (FAIR'9)"; Cần Thơ, ngày 4-5/8/2016
DOI: 10.15625/vap.2016.00041

MÔ HÌNH QUẢN LÝ TẬP DỮ LIỆU VĂN BẢN LỚN CHO PHÉP TÌM KIẾM TOÀN VĂN VÀ PHÂN TÍCH THỐNG KÊ TRỰC QUAN

Nguyễn Hùng Dũng[1], Trương Xuân Việt[1], Trương Quốc Định[2], Nguyễn Hoàng Việt[1]
[1]Trung tâm Công nghệ Phần mềm − Đại học Cần Thơ
[2] Khoa Công nghệ Thông tin & Truyền thông − Đại học Cần Thơ
hungdung@ctu.edu.vn, txviet@ctu.edu.vn, tqdinh@cit.ctu.edu.vn, nhviet@ctu.edu.vn

TÓM TẮT— *Mục tiêu của bài viết là đề xuất mô hình mới cho phép quản lý tập dữ liệu lớn phi cấu trúc, tồn tại dưới dạng các tập tin văn bản, bảng tính. Mô hình đề xuất dựa trên một tập các công nghệ nguồn mở của Big Data theo hướng tích hợp dịch vụ và chuẩn hóa dữ liệu nối kết, 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 tiếng Việt, 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 và (3) Bộ trực quan hóa dữ liệu dựa trên Banana. Kết quả thực nghiệm được thực hiện trên tập dữ liệu tất cả các bài báo khoa học đăng trên Tạp chí Khoa học trường Đại học Cần Thơ từ năm 2011 đến 2015.*

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

I. GIỚI THIỆU

Tìm kiếm và trực quan hoá dữ liệu phi cấu trúc là một trong các nhu cầu thiết thực và được đánh giá như một trong các chìa khóa chính hỗ trợ các tổ chức ra quyết định. Nhu cầu trả lời các câu hỏi phức tạp ngoài phạm vi của các truy vấn SQL phổ biến chủ yếu được thực hiện thủ công hoặc dựa trên các phỏng đoán của các nhà hoạch định – các phỏng đoán này thường không dựa trên một nền tảng của một tập dữ liệu đầy đủ. Để ra một quyết định ngắn, nhân viên tổng hợp cần phải đọc lại hàng ngàn, thậm chí vài chục ngàn văn bản để truy vấn các thông tin cần thiết. Đây là một công việc thường không được thực hiện một cách thấu đáo trong môi trường làm việc văn phòng. Truy vấn dữ liệu phi cấu trúc dựa trên các kỹ thuật Big Data, do vậy, trở thành một phương tiện hữu hiệu giải quyết các khó khăn đã nêu. Tuy nhiên, không nhiều nghiên cứu về việc phát triển các nền tảng này tại Việt Nam, đặc biệt cho ngôn ngữ tiếng Việt.

Trong nghiên cứu này chúng tôi mong muốn tìm kiếm một mô hình mới và hoàn chỉnh nhằm các mục tiêu sau: (1) quản lý được một tập văn bản lớn, cho phép dễ dàng truy lục khi cần thiết, (2) hỗ trợ cơ chế tìm kiếm trực tuyến và toàn văn trên tập dữ liệu tiếng Việt với các toán tử Bool và (3) trực quan hóa kết quả tìm kiếm và tương thích với đa dạng thiết bị hiển thị. Giải pháp này, nếu được xây dựng sẽ cho phép dễ dàng quản lý và truy vấn rất nhiều câu hỏi liên quan đến điều hành mà các nhà quản trị cần biết, vượt qua được các hạn chế về sự thiếu hụt thông tin. Dữ liệu lớn (Big data) là công nghệ ưu tiên được lựa chọn do sự phù hợp về tiêu chí và đặc biệt, một hệ sinh thái phong phú nguồn mở có sẵn. Về lý thuyết, 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) [3].

Các thành phần chính yếu của mô hình đề xuất dựa trên 3 nhóm công nghệ sau: (1) HDFS – dịch vụ quản lý tập tin của Hadoop, (2) Lucene/Solr – dịch vụ cung cấp chỉ mục hóa và tìm kiếm toàn văn trực tuyến và (3) Banana [2] – bộ công cụ trực quan hóa dựa trên nền tảng Kibana [15]. Một trong các thuận lợi cơ bản là các nghiên cứu về ngôn ngữ tiếng Việt đã được thực hiện rất thấu đáo và các thư viện phân tích từ vựng như VNTokenizer đã được phát triển bởi Lê Hồng Phương [10]. Đặc biệt, Cao Mạnh Đạt [4] đã phát triển một bộ phân tích từ vựng VnAnalyzer dựa trên VNTokenizer và tương thích với Apache Lucene. Thư viện này được chúng tôi ưu tiên lựa chọn do sự phù hợp với giải pháp Apache Solr [11]. Như vậy, có thể nói, các thành tố quan trọng nhất để phát triển mô hình theo đề xuất đã tồn tại, vấn đề tiếp theo là đề xuất cách thức bắt tay giữa chúng để đạt mục tiêu nghiên cứu.

Tập dữ liệu tất cả các bài báo khoa học đăng trên Tạp chí Khoa học Đại học Cần Thơ được sử dụng để kiểm tra vận hành của mô hình đề xuất. Trong đó, các nội dung cụ thể liên quan đến hoạt động nghiên cứu khoa học tại trường Đại học Cần Thơ được tìm kiếm dựa trên các câu hỏi truy vấn khác nhau. Trên thực tế, tập dữ liệu này có thể mở rộng không giới hạn để trả lời các câu hỏi rộng hơn về hoạt động nghiên cứu của trường Đại học Cần Thơ.

Bài báo được cấu trúc như sau: chúng tôi sẽ điểm qua các công trình nghiên cứu 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.

II. NGHIÊN CỨU LIÊN QUAN

Trên thực tế, mô hình chúng tôi đề xuất không quá mới trên thế giới. Các nghiên cứu tích hợp giữa Hadoop và Solr đã được thực hiện trong khung tích hợp Cloudera [8]; tương tự giữa Hadoop và Elastic Search trong khung tích hợp Hortonworks [9]. 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 [13]. Vấn đề đặt ra là, đây là các giải pháp thương mại hóa, nhưng các thành tố cấu thành chúng lại chủ yếu dựa trên mã nguồn mở. Hơn nữa, các khung tích hợp này chưa hỗ trợ phân tích từ vựng trên ngôn ngữ tiếng Việt.

Dù không tái sử dụng lại các khung tích hợp 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 [8], bộ trực quan hóa dựa trên ZoomData, trong khi đó Hortonworks [9] 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 [15] − là lựa chọn phù hợp với bộ tìm kiếm Solr.

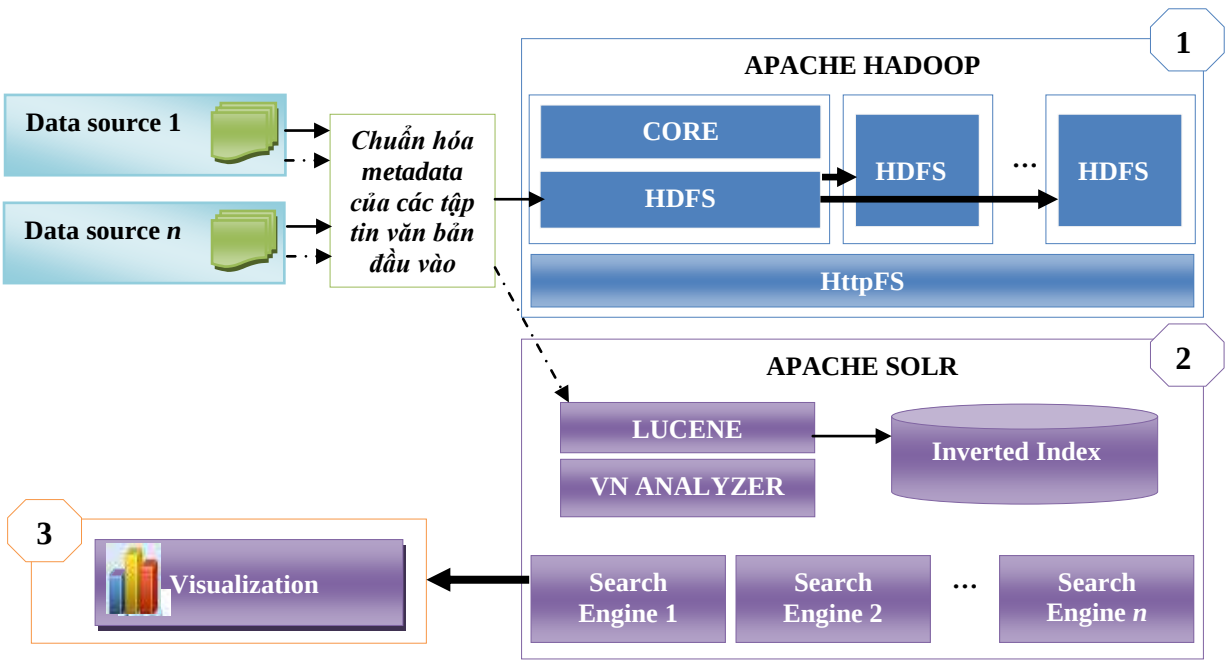
Trong nghiên cứu này, bộ lập chỉ mục Lucene đóng vai trò chủ đạo. Lucene là một thư viện mã nguồn mở, được phát triển bởi Dough Cutting. Thư viện này cung cấp các hàm cơ bản hỗ trợ cho việc đánh chỉ mục và tìm kiếm thông qua các hàm API. Lucene có thể lập chỉ mục và hỗ trợ các thư viện tìm kiếm các loại 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".

Trong Hortonworks [9], quá trình xây dựng chỉ mục nghịch đảo (Inverted index) cho văn bản được thực hiện dựa trên cơ chế ánh xạ/rút gọn (map/reduce) ngay bên trong Hadoop. Cơ chế này có thể vẫn dựa trên bộ thư viện Lucene nhưng được thực hiện bên ngoài Solr, khi đó Solr chỉ đóng vai trò hỗ trợ tìm kiếm toàn văn dựa trên tập chỉ mục xây dựng sẵn. Chúng tôi nhận thấy rằng bộ chỉ mục Lucene đã được tích hợp sẵn trong Solr, do vậy đã dịch chuyển quá trình chỉ mục hóa này vào Solr, thay vì sử dụng map/reduce bên trong Hadoop. Cách làm này đơn giản hóa mô hình và giúp tinh giảm phần nào kích thước lưu trữ, khi đó chúng tôi chỉ sử dụng duy nhất dịch vụ HDFS của Hadoop để quản lý hệ thống văn bản.

III. ĐỀ XUẤT MÔ HÌNH QUẢN LÝ, TÌM KIẾM TOÀN VĂN VÀ TRỰC QUAN HÓA KẾT QUẢ

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, tìm kiếm toàn văn và trực quan hóa kết quả

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 mà chúng tôi sử dụng để thực nghiệm theo các trường (fields) như sau:

- tacgia: Các tác giả tham gia nghiên cứu khoa học (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:

(1) 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.

(2) Hệ thống chỉ mục, tìm kiếm văn bản dựa trên Lucene/Solr:

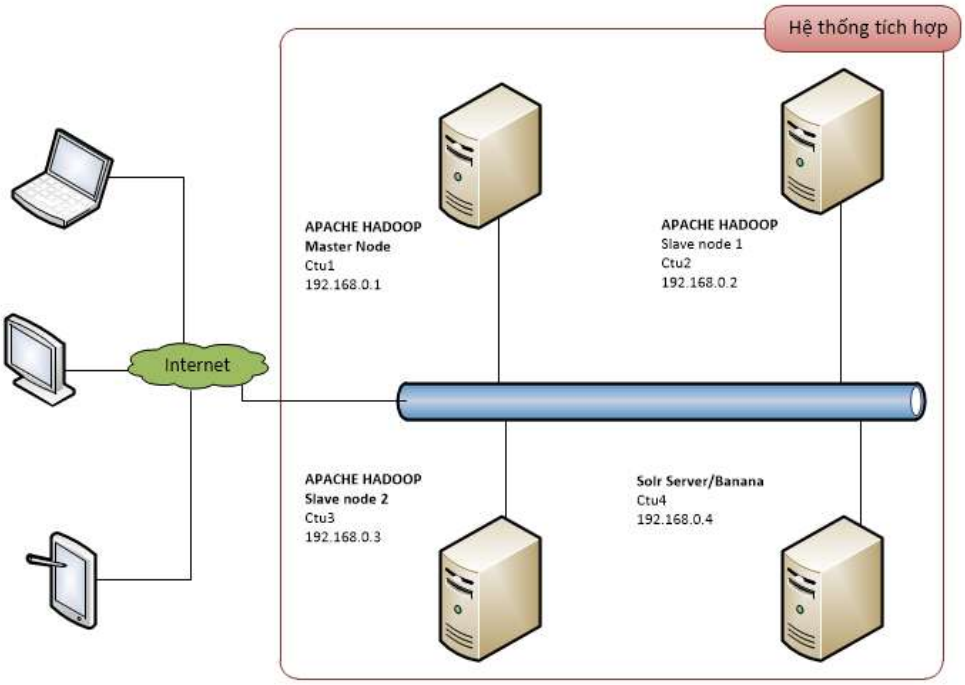
- Hệ thống này cung cấp cơ chế lập chỉ mục nghịch đảo (Inverted Indexing) và máy tìm kiếm (Search Engine) cho văn bản nguồn.
- Kết quả tìm kiếm sẽ trả về văn bản gốc phù hợp đã được lưu trữ tại hệ thống lưu trữ (1). Do thư viện lập chỉ mục Lucene đã được tích hợp sẵn trong Apache Solr nên trên thực tế 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 [4], giúp việc phân tích và tìm kiếm thêm tài liệu với ngôn ngữ tiếng Việt được dễ dàng.

(3) 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).

IV. 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 9 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. Hệ thống thử nghiệm được chúng tôi cài đặt theo kiến trúc như sau:



Hình 2. Kiến trúc hệ thống của mô hình thử nghiệm

Sau khi tài liệu được đưa lên Solr, tất cả được đánh chỉ mục. Khi người dùng tìm kiếm theo tiêu chí nào đó, dữ liệu sẽ được mô tả trên Solr với các tham số được liệt kê theo bảng sau:

Bảng 1. Mô tả các tham số lưu trữ thông tin trên Solr

Tham số	Mô tả
QTime	Thể hiện thời gian tìm kiếm.
q	Trình bày câu truy vấn.
rows	Số lượng văn bản được hiển thị.
numFound	Số lượng văn bản được tìm thấy.
docs	Liệt kê trường: các giá trị được định nghĩa trong lược đồ (schema.xml).

Tập dữ liệu "TẠP CHÍ KHOA HỌC ĐHCT" mà chúng tôi tạo ra chứa tổng cộng 1.584 bài báo 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.

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. Dưới đây chúng tôi trình bày một số kết quả thực nghiệm điển hình về việc tìm kiếm, thống kê 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)

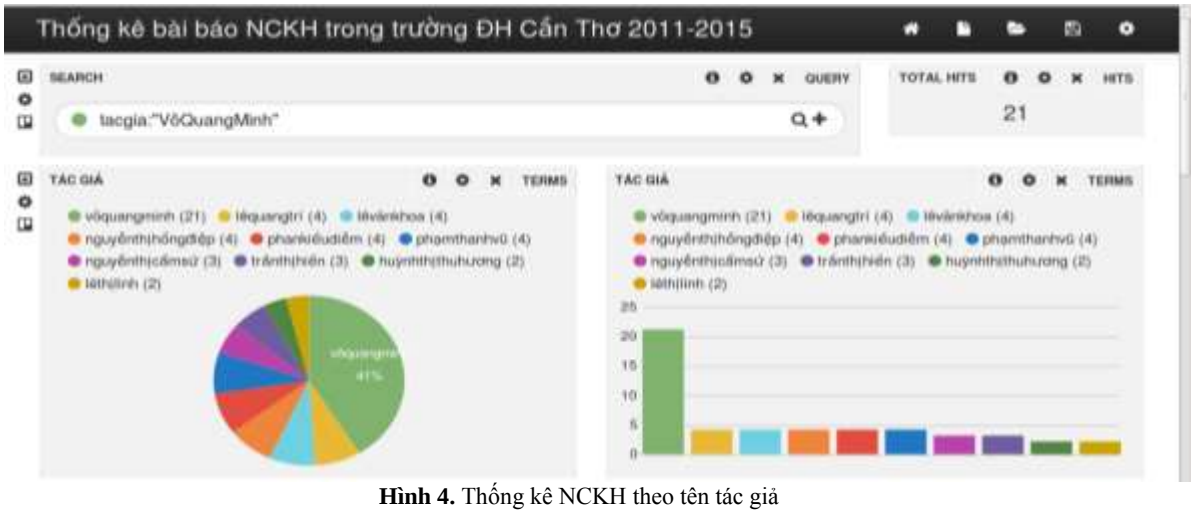
Trường ***'donvi'*** được định nghĩa là khoa/đơn vị mà tác giả chính của bài báo NCKH công tác, để tìm kiếm những bài báo NCKH theo đơn vị thuộc Trường Đại học Cần Thơ, sử dụng truy vấn: ***donvi:*_ctu***.

Kết quả của câu truy vấn trên được hiển thị như ***Hình 3,*** ứng dụng 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, ...
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.

Hình 3. Thống kê bài báo NCKH tại Trường Đại học Cần Thơ theo tác giả và theo khoa/đơn vị

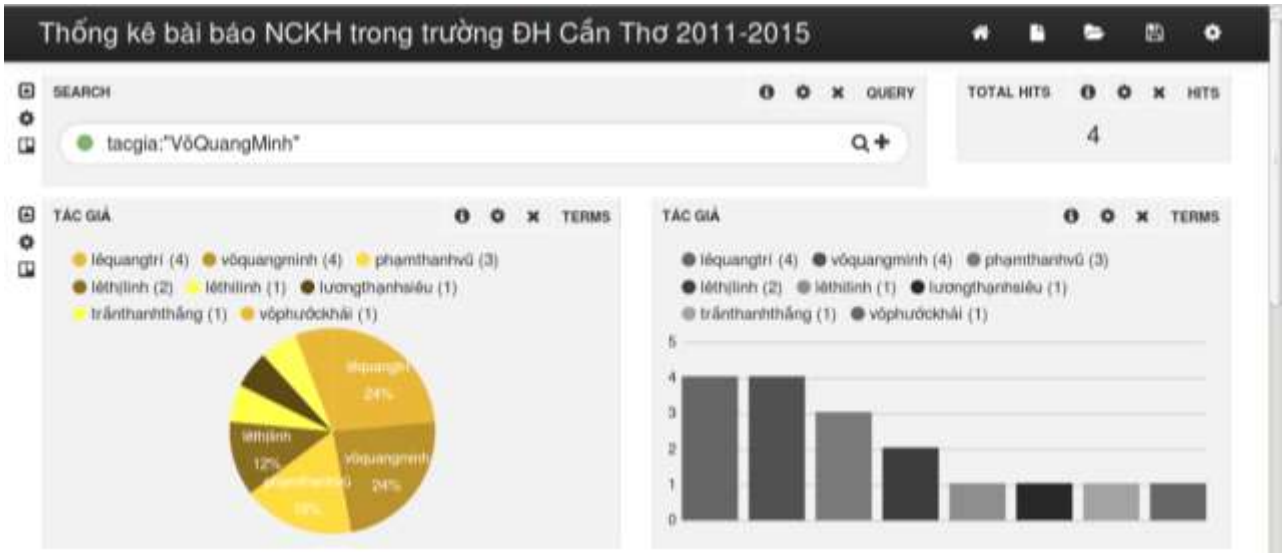
(2) Tìm kiếm và thống kê bài báo NCKH theo tên tác giả

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 hiển thị kết quả tìm thấy tất cả các bài báo NCKH của tác giả và trực quan hóa kết quả theo hai dạng biểu đồ hình tròn và cột:



Hình 4. 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 và kết quả được thống kê như sau:

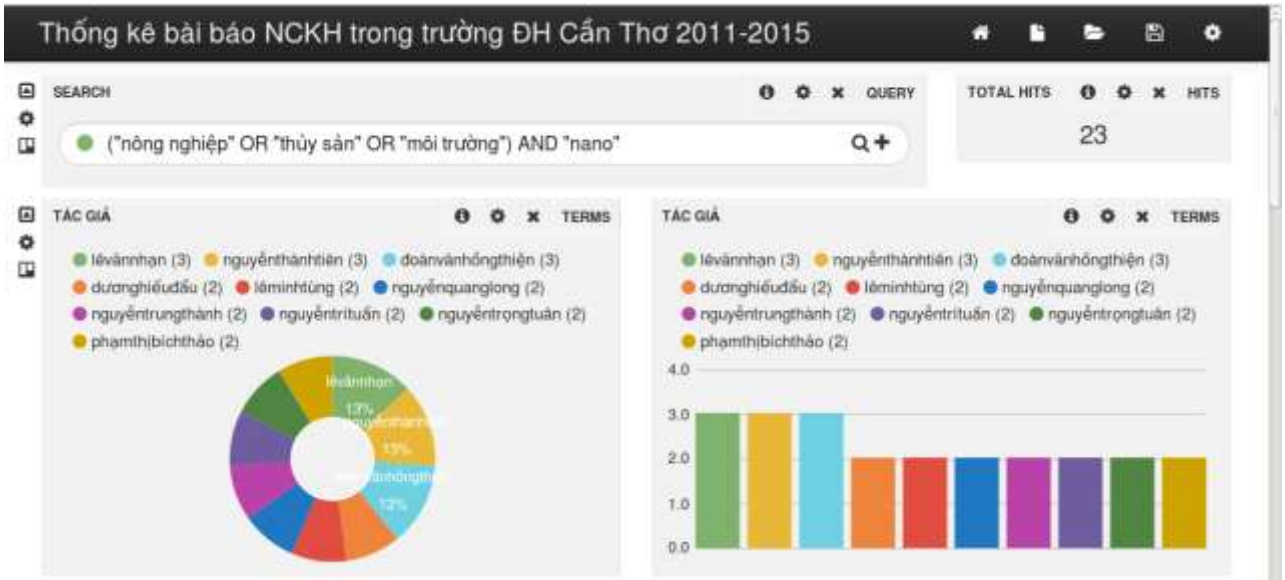


Hình 5. Tương quan giữa hai tác giả

Từ kết quả thống kê ở ***Hình 5*** có thể thấy hai tác giả "Võ Quang Minh" và "Lê Quang Trí" cùng tham gia nghiên cứu với tác giả Phạm Thanh Vũ (3 bài báo NCKH), Lê Thị Linh (2 bài), Võ Phước Khải (1 bài).

(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.

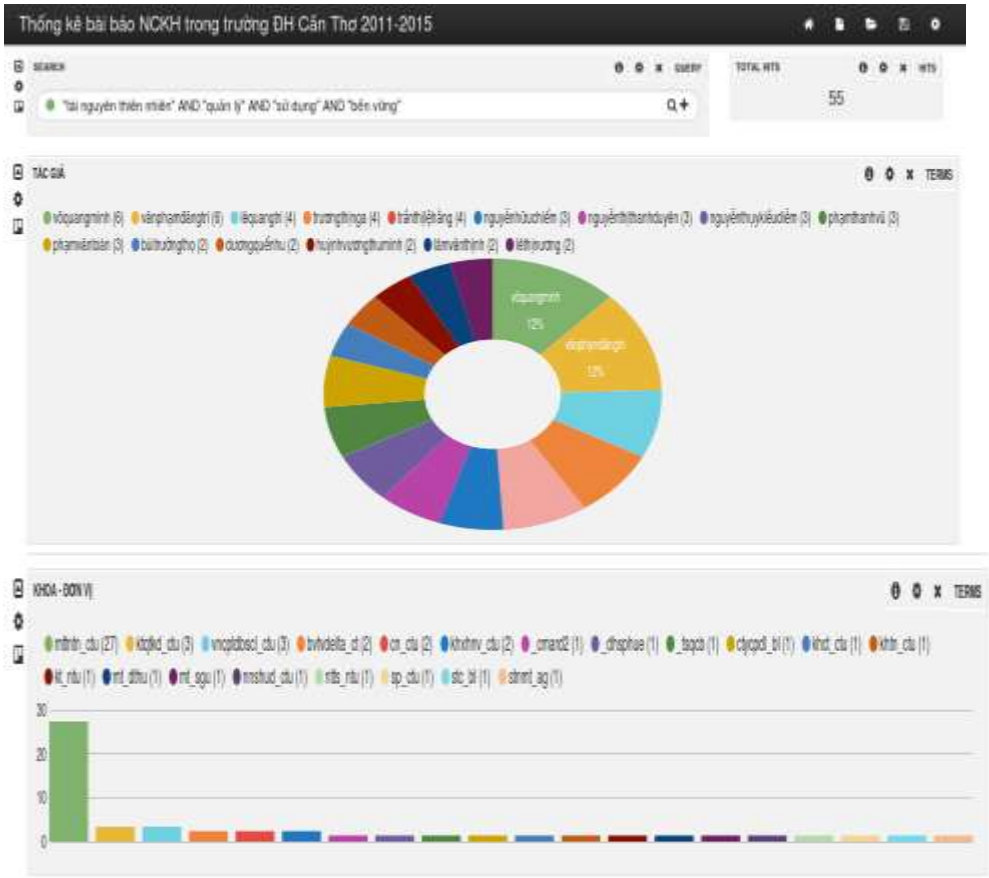


Hình 6. Ứng dụng công nghệ cao trong nông nghiệp, thủy sản và môi trường

Tổng cộng tìm kiếm được 23 bài liên quan đến vấn đề ứng dụng công nghệ cao trong nông nghiệp, thủy sản và môi trường. 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 "nano" -tuade:"2011*"***.

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

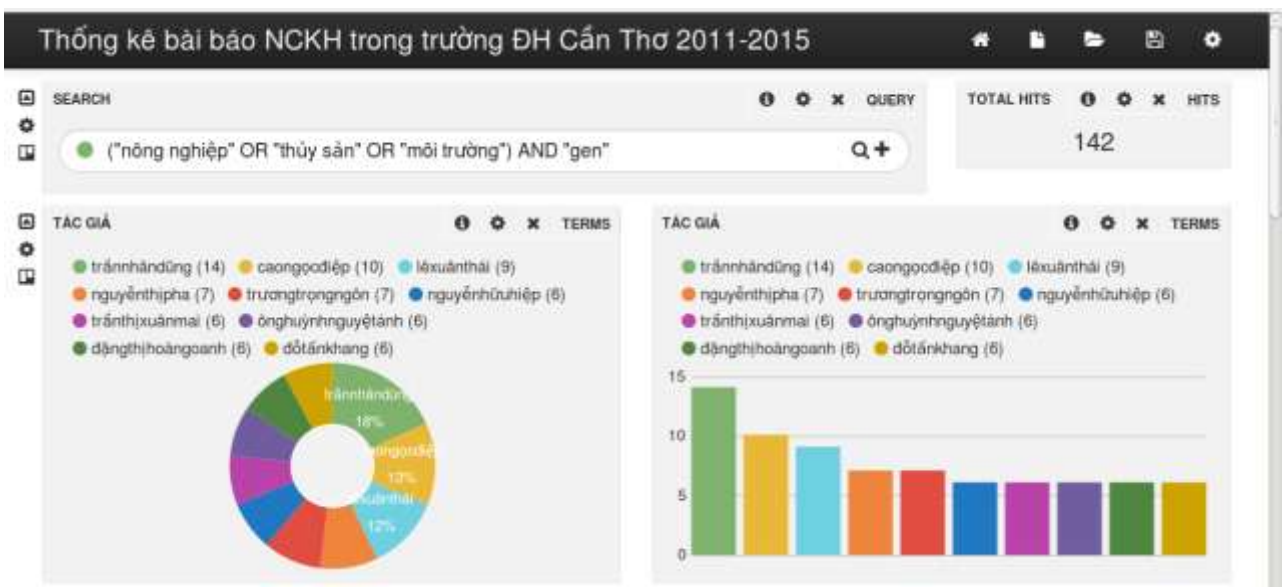
Hình 7, với việc sử dụng từ khóa tìm kiếm: ***"tài nguyên thiên nhiên" AND "quản lý" AND "sử dụng" AND "bền vững"*** cho thấy vấn đề quan tâm đến việc quản lý và sử dụng bền vững tài nguyên thiên nhiên như thế nào?



Hình 7. Quản lý và sử dụng bền vững tài nguyên thiên nhiên

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.

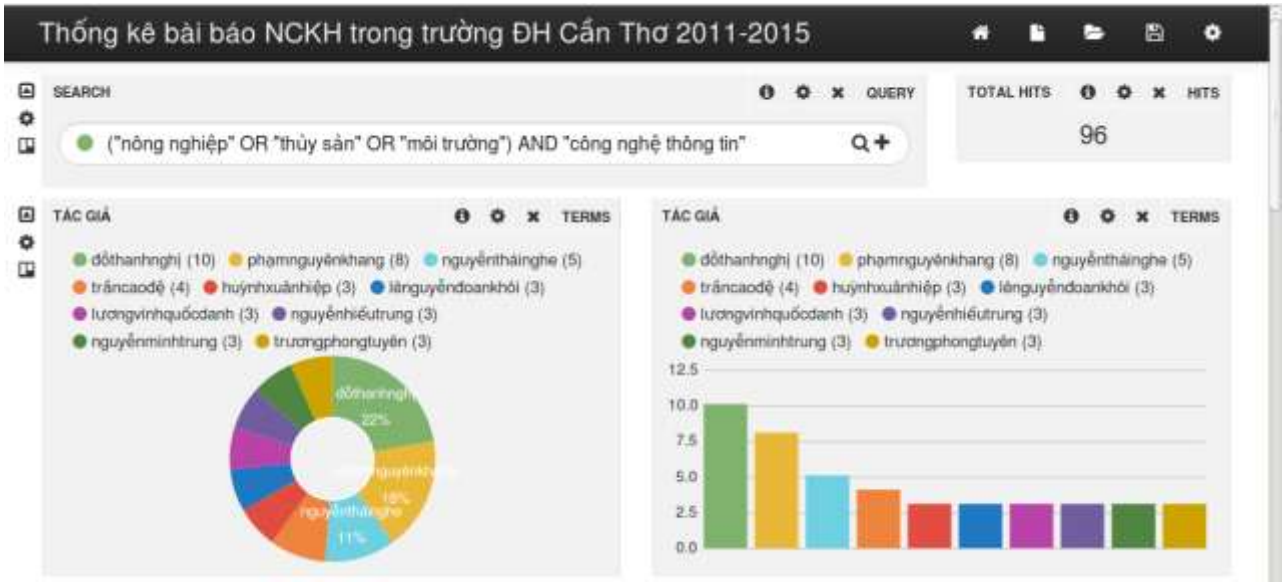
Để thấy được tầm quan trọng của ***'gen'*** trong lĩnh vực nông nghiệp, thủy sản và môi trường, chúng tôi tiến thành tìm kiếm các bài báo NCKH liên quan đến vấn đề này. ***Hình 8***, minh họa kết quả và trực quan hóa kết quả tìm kiếm:



Hình 8. Thống kê NCKH về lĩnh vực nông nghiệp, thủy sản và môi trường liên quan đến gen

Kết quả có 142 bài báo NCKH liên quan được tìm thấy, chứng tỏ vấn đề này có rất nhiều tác giả quan tâm. Đây cũng là một trong những định hướng nghiên cứu chủ đạo của Trường.

Ngoài ra những nghiên cứu về nông nghiệp, thủy sản và môi trường cũng có sự góp phần không nhỏ của công nghệ thông tin. ***Hình 9***, trình bày kết quả tìm kiếm và trực quan hóa việc ứng dụng công nghệ thông tin vào nghiên cứu trong lĩnh vực này.



Hình 9. Sử dụng công nghệ thông tin trong lĩnh vực nông nghiệp, thủy sản và môi trường

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 10. 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.

V. KẾT LUẬN VÀ ĐỀ XUẤT

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.

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 Đại học Cần Thơ, điều này sẽ giúp cho các nhà quản lý có thêm thông tin để định hướng trong việc quy hoạch và xét duyệt các đề tài NCKH theo định hướng chung của Trường.

VI.TÀI LIỆU THAM KHẢO

[1] A. Hemanth, Dr. R. V. Krishnaiah, 2013. The Hadoop Distributed Filesystem: Balancing Portability. International Journal of Computer Engineering & Applications, Vol. III, Issue III. ISSN: 2321-3469.

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

[3] 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.

[4] 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/.

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

[6] Hao Wu, Guoliang Li, and Lizhu Zhou, 2013. Ginix: Generalized Inverted Index for Keyword Search. Tsinghua Science and Technology, Volume 18, Number 1, February 2013. ISSN 1007-0214 10/12 pp77-87.

[7] Jeffrey Dean and Sanjay Ghemawat, 2008. MapReduce: Simplified Data Processingon Large Clusters. Magazine: Communications of the ACM - 50th anniversary issue: 1958 - 2008, Volume 51 Issue 1, January 2008, Pages 107-113.

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

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

[10] 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.

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

[12] Marcus Fontoura, Maxim Gurevich, Vanja Josifovski, Sergei Vassilvitskii, 2011. Efficiently Encoding Term Co-occurrences in Inverted Indexes. CIKM '11 Proceedings of the 20th ACM international conference on Information and knowledge management. ISBN: 978-1-4503-0717-8, Pages 307-316

[13] 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.

[14] 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.

[15] Kibana analytics and search dashboard for Elasticsearch, 2016. [Online]. https://www.elastic.co/products/kibana.

MÔ HÌNH QUẢN LÝ TẬP DỮ LIỆU VĂN BẢN LỚN CHO PHÉP TÌM KIẾM TOÀN VĂN VÀ PHÂN TÍCH THỐNG KÊ TRỰC QUAN

Nguyen Hung Dung, Truong Xuan Viet, Truong Quoc Dinh, Nguyen Hoang Viet

ABSTRACT— *The article objective is proposing a new model for managing large unstructured data set existed in the text files, spreadsheets form. The proposed model is based on the open source set of Big Data with service integration and link data normalization, including: (1) HDFS (Hadoop Distributed File System) used in the file management, (2) Lucene set up inverted index for the Vietnamese text, Apache Solr supported inverted indexes management mechanisms, full-text search and some of advanced search functions and (3) the data visualization based on Banana. The experimental results are performed on the data set of the scientific journals published on Can Tho University journals of science from 2011 to 2015.*

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