

Máy chủ tìm kiếm Solr và ứng dụng xây dựng hệ tìm kiếm theo yêu cầu người dùng

Nguyễn Văn Đông Anh – CNPM K51

GVHD: PGS.TS Huỳnh Quyết Thắng

ThS. Lê Quốc

Nội dung

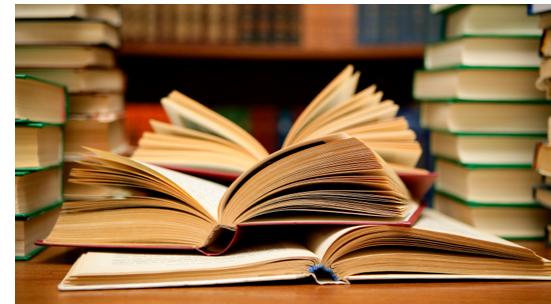
1. Đặt vấn đề và nhiệm vụ của đồ án
2. Kiến thức liên quan
3. Xây dựng hệ thống tìm kiếm
4. Cài đặt và đánh giá
5. Kết quả và định hướng phát triển

1. ĐẶT VẤN ĐỀ VÀ NHIỆM VỤ ĐỒ ÁN



Đặt vấn đề

- Tri thức
- Mạng cộng đồng chia sẻ tri thức BKProfile
 - Tìm kiếm
 - Chia sẻ
- Xây dựng hệ thống tìm kiếm
 - **Tìm kiếm tri thức**
 - **Tìm kiếm chuyên gia**



Nhiệm vụ của đồ án

- **Lý thuyết**

- Nghiên cứu chung hệ thống tìm kiếm
- Nghiên cứu máy chủ tìm kiếm Solr

- **Xây dựng**

- Hệ thống tìm kiếm trong BKProfile

2. KIẾN THỨC LIÊN QUAN



Tìm kiếm nói chung

- Xây dựng bộ dữ liệu
- Truy vấn tìm kiếm
- Xử lý truy vấn
- Tìm trong bộ dữ liệu
- Đánh điểm, sắp xếp
kết quả phù hợp [10]



Máy chủ tìm kiếm Solr

- Máy chủ tìm kiếm văn bản mã nguồn mở
- Hiệu năng cao
- Thuật toán đánh điểm tốt
- Nhiều chức năng [6]



Tại sao chọn Solr

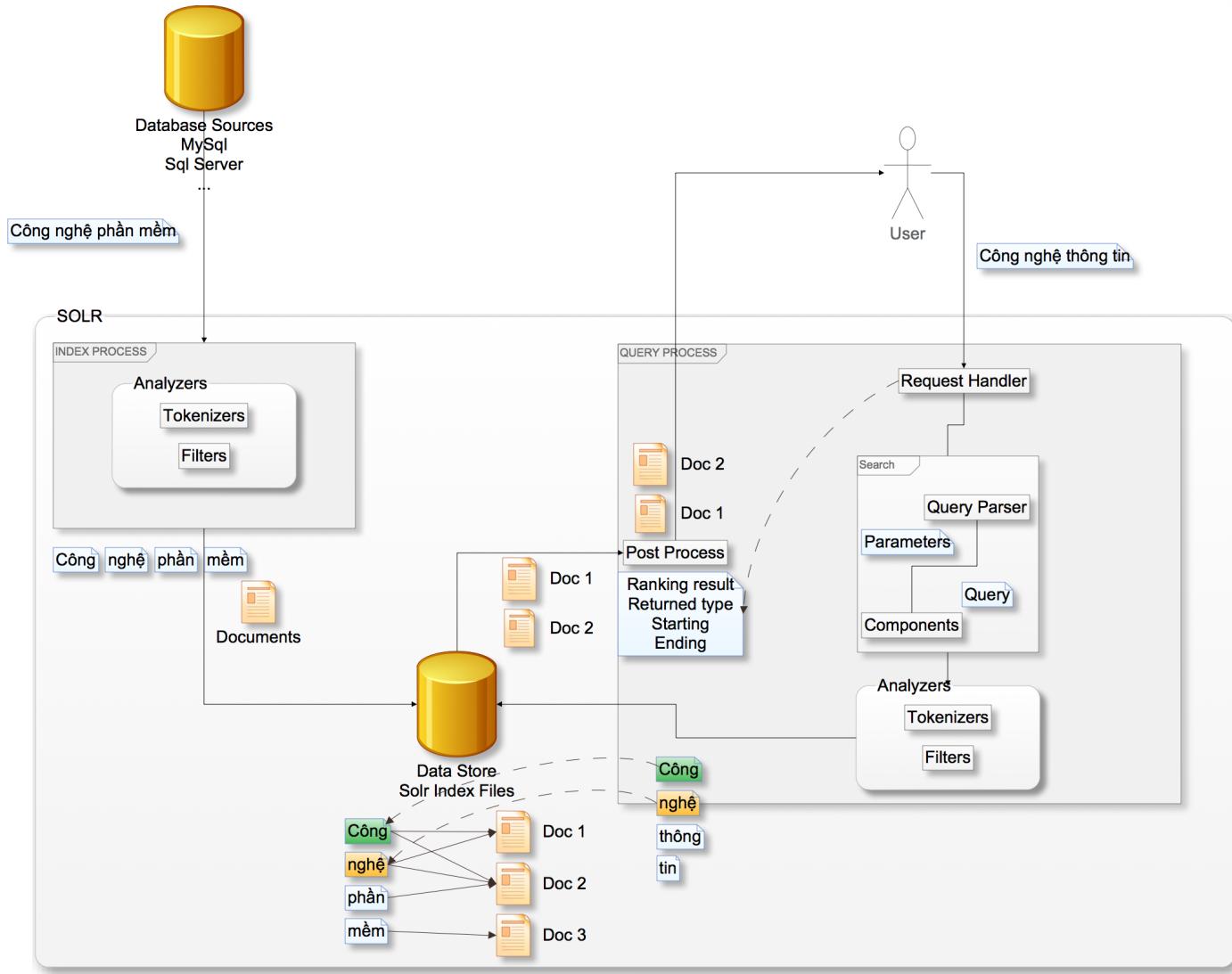
Plain	Insert/Index Time	HDD Size	Queries
MySQL Fulltext	16 min 10 sec	260 MB (.MYI file)	skipped
Sphinx	30 sec	137 MB (.spd file)	138 / sec*
Solr	5 min 23 sec	106 MB	465 / sec

Simple	Insert/Index Time	HDD Size	Queries
MySQL Fulltext	24 min 2 sec	344 MB (.MYI file)	skipped
Sphinx	42 sec	235 MB (.spd file)	90 / sec*
Solr	5 min 47 sec	106 MB	478 / sec

Structured	Insert/Index Time	HDD Size	Queries
MySQL Fulltext	26 min 18 sec	389 MB (.MYI file)	skipped
Sphinx	49 sec	272 MB (.spd file)	84 / sec*
Solr	6 min 3 sec	103 MB	466 / sec

* PHP script used to issue each query and format to HTML

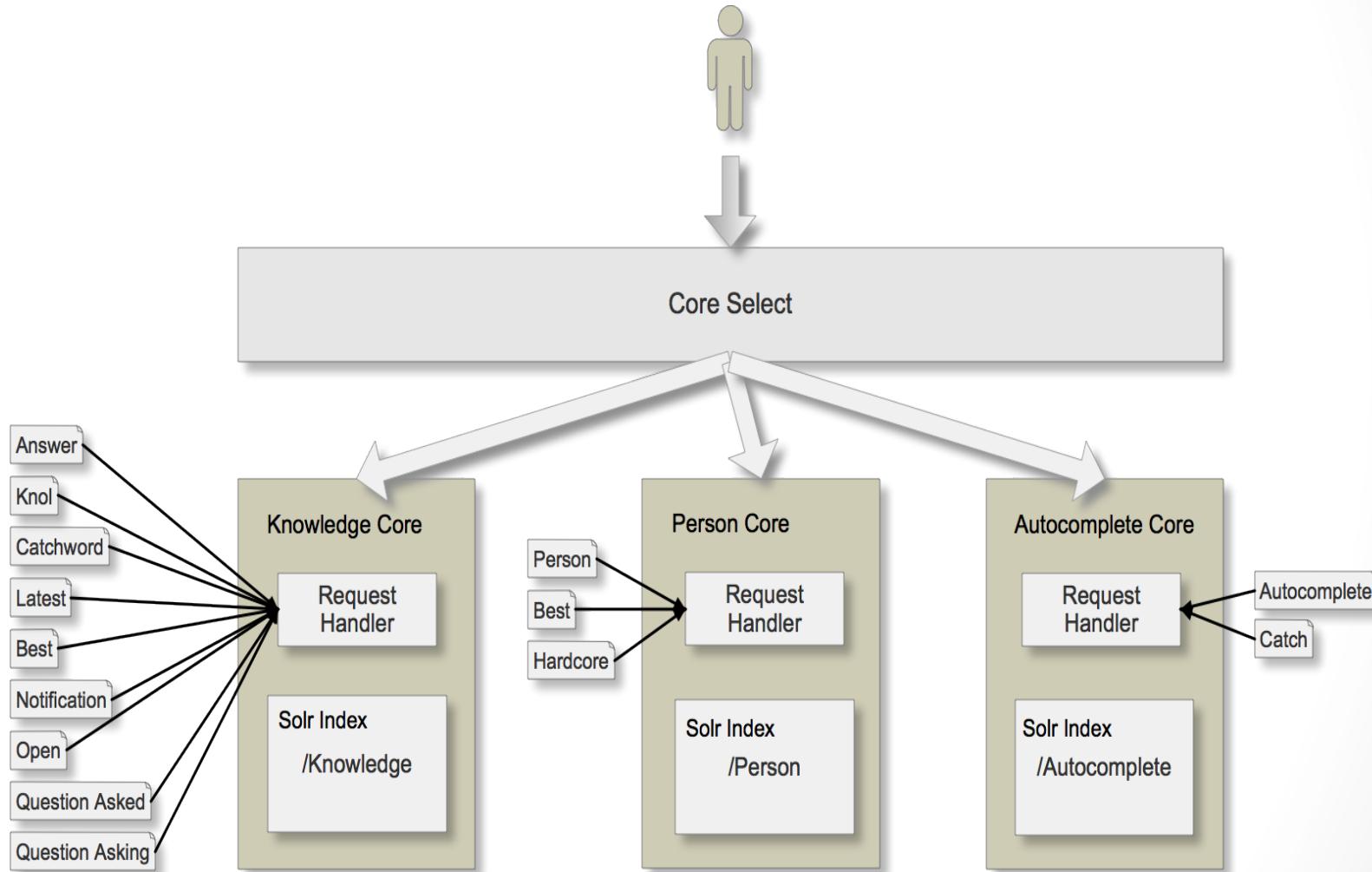
Tổng quan Solr



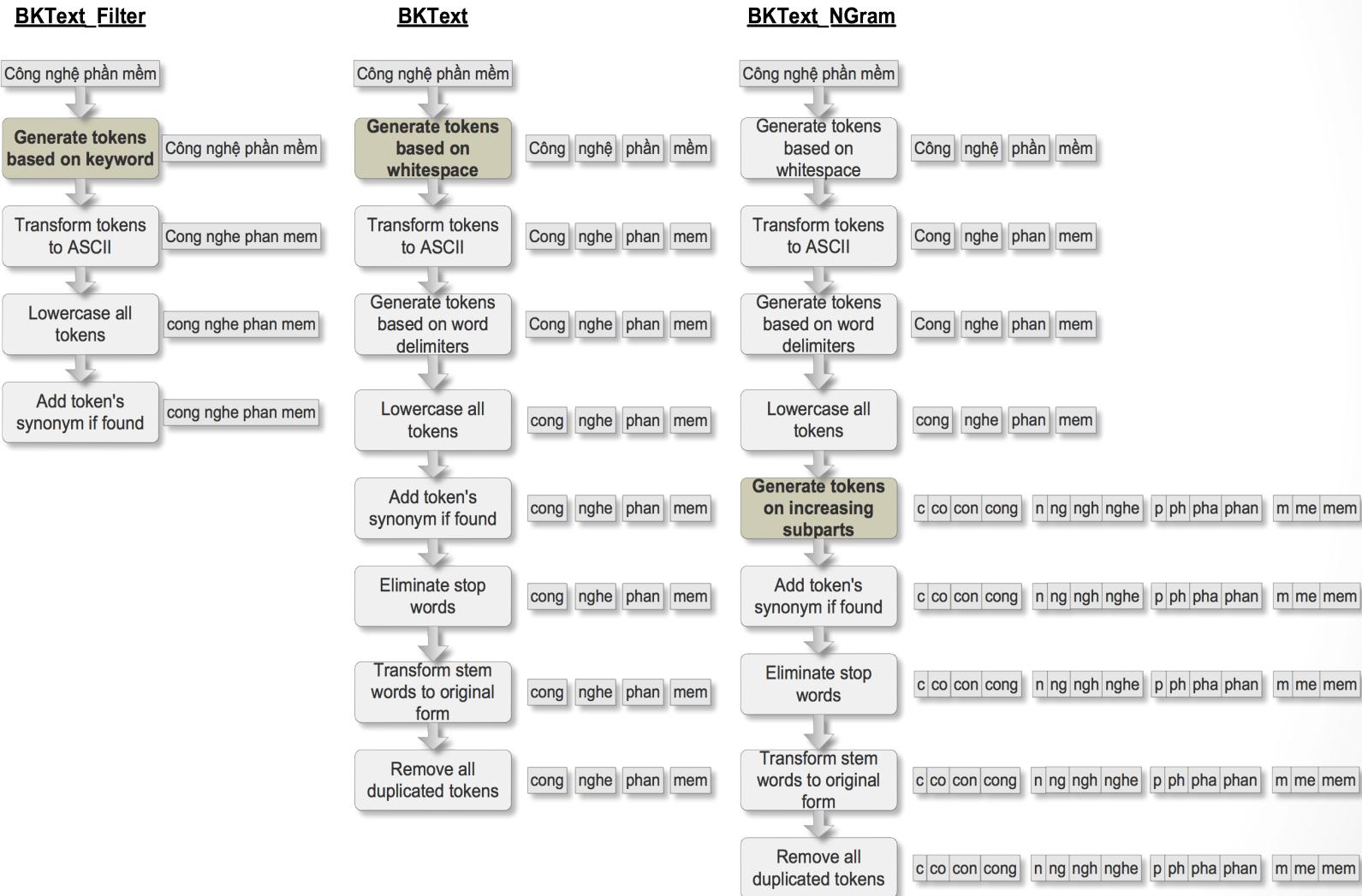
3. XÂY DỰNG HỆ THỐNG TÌM KIẾM



Kiến trúc tổng quan



Phân tích từ khóa



Bộ xử lý truy vấn

Knowledge Core

Answer

Knol

Catchword

Latest

Best

Notification

Open

Question Asked

Question Asking

Person Core

Person

Best

Hardcore

Autocomplete Core

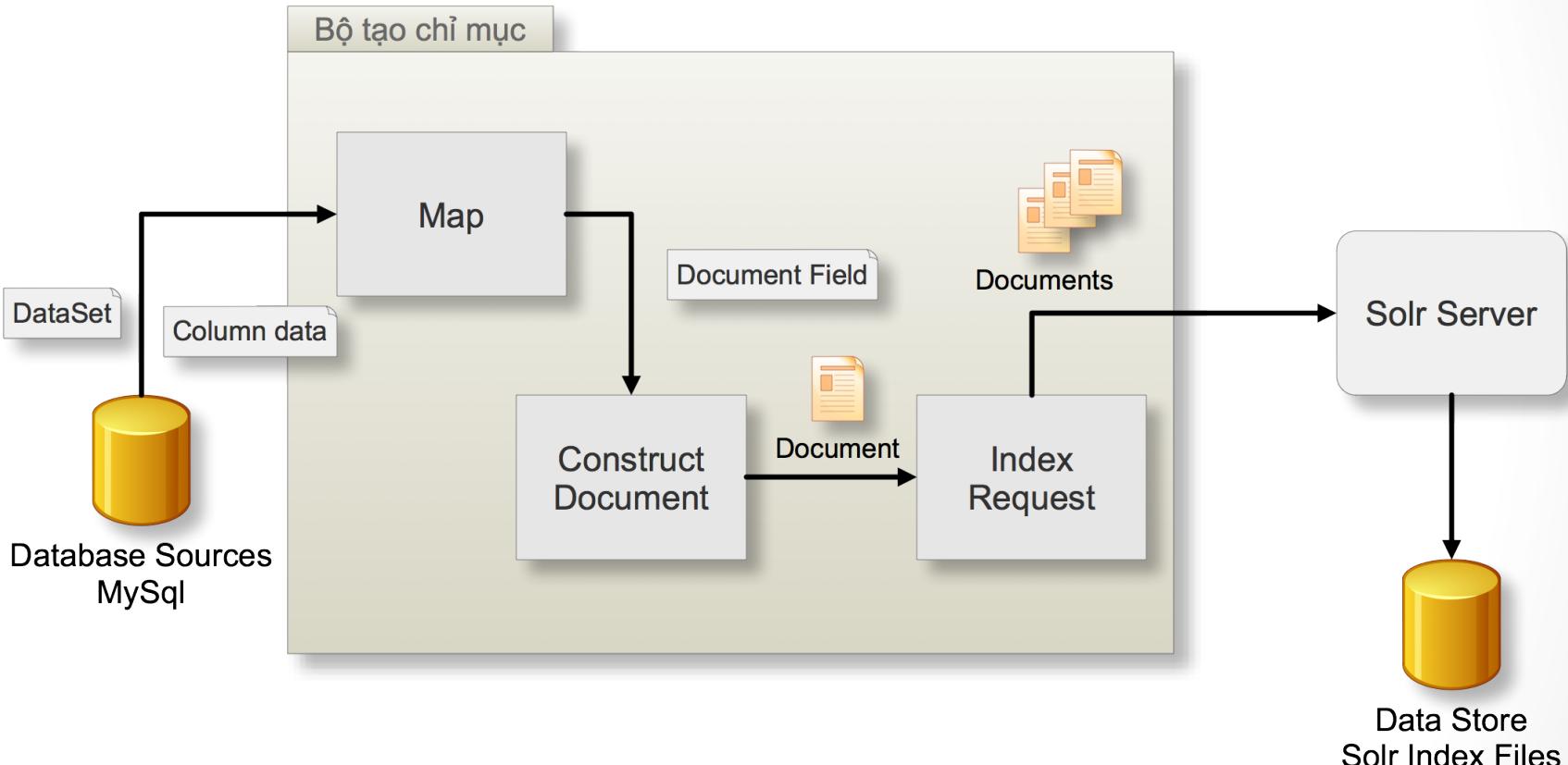
Autocomplete

Catch

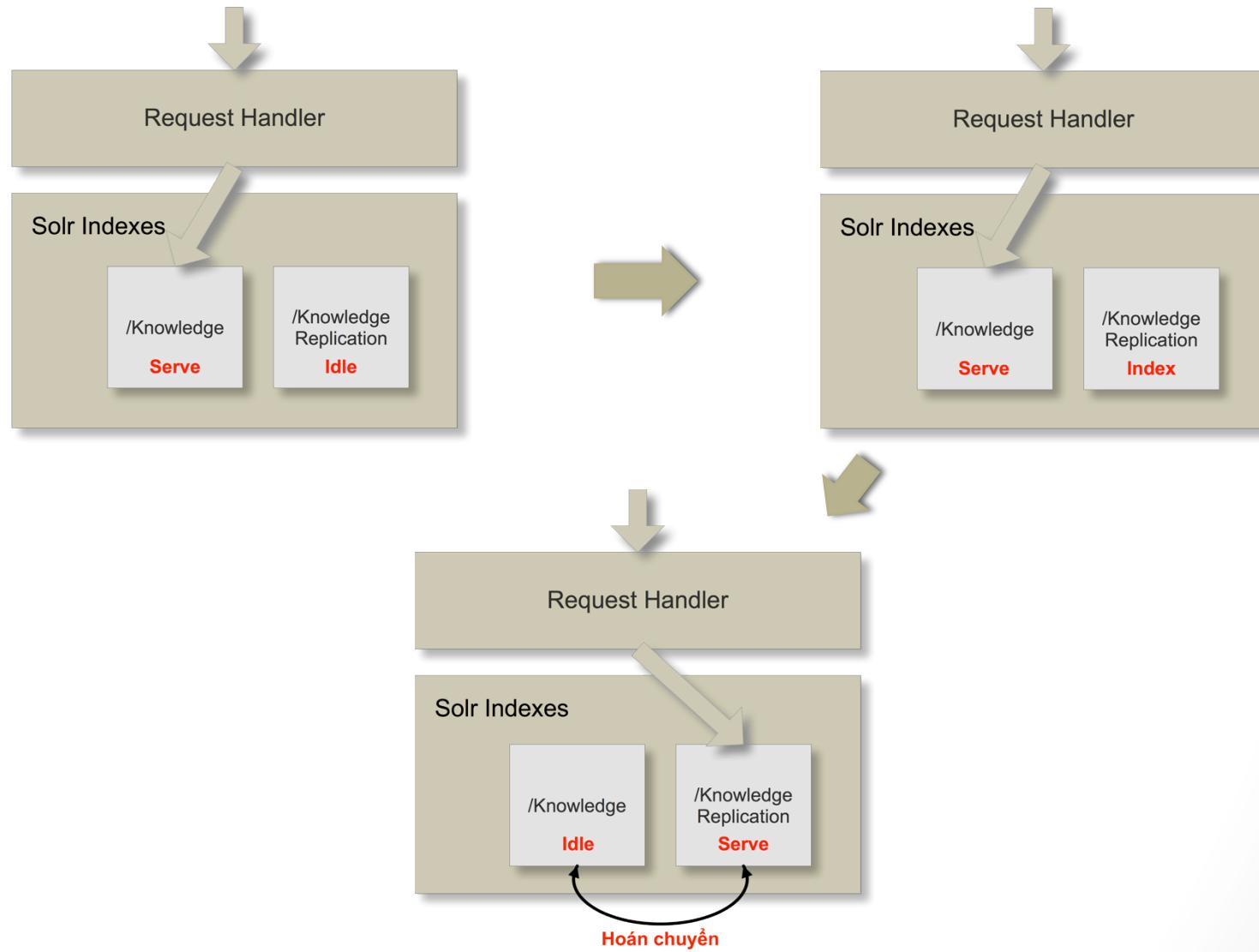
Trường truy vấn

Logic tính điểm

Bộ tạo dữ liệu chỉ mục



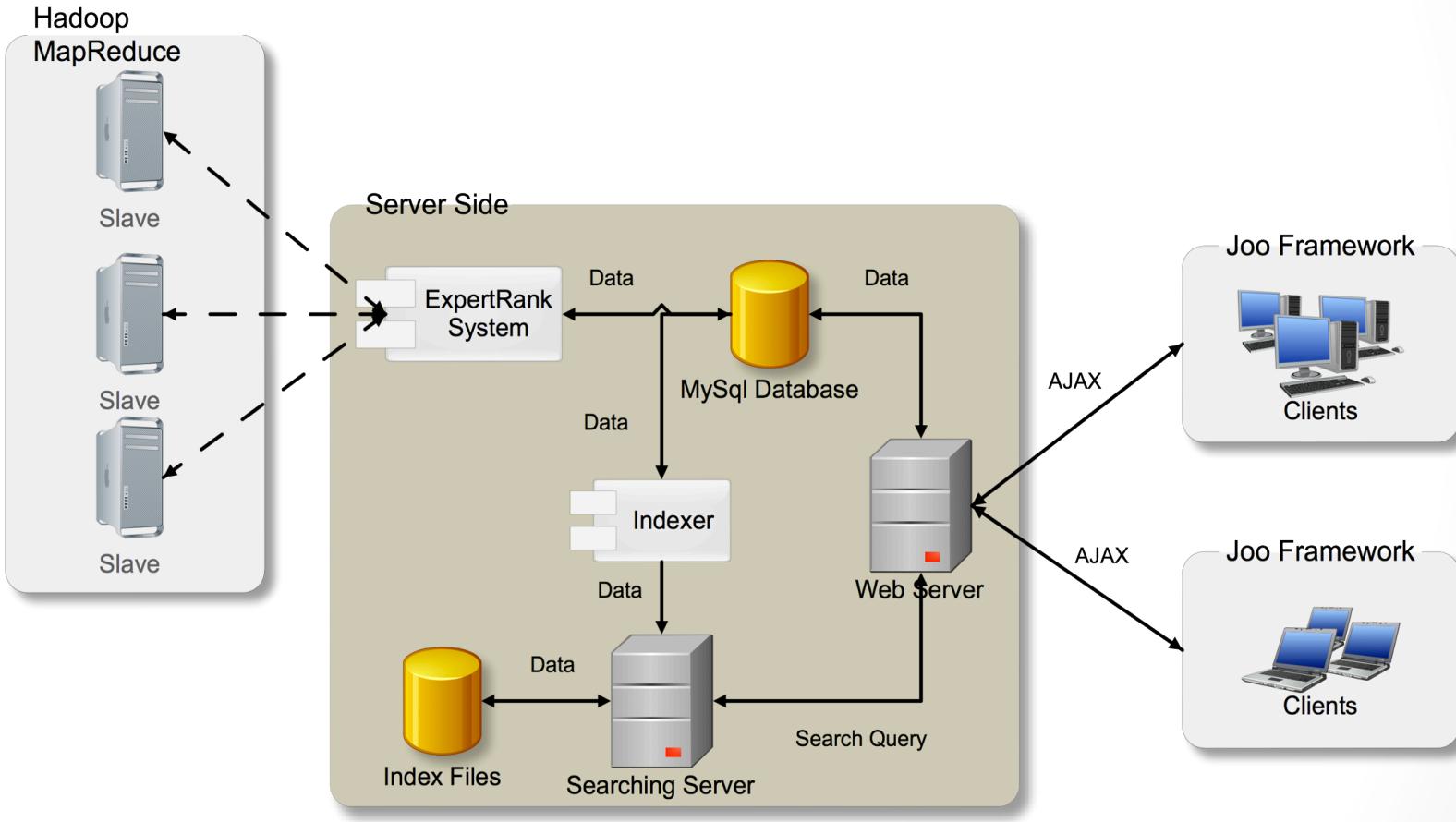
Nhân tìm kiếm



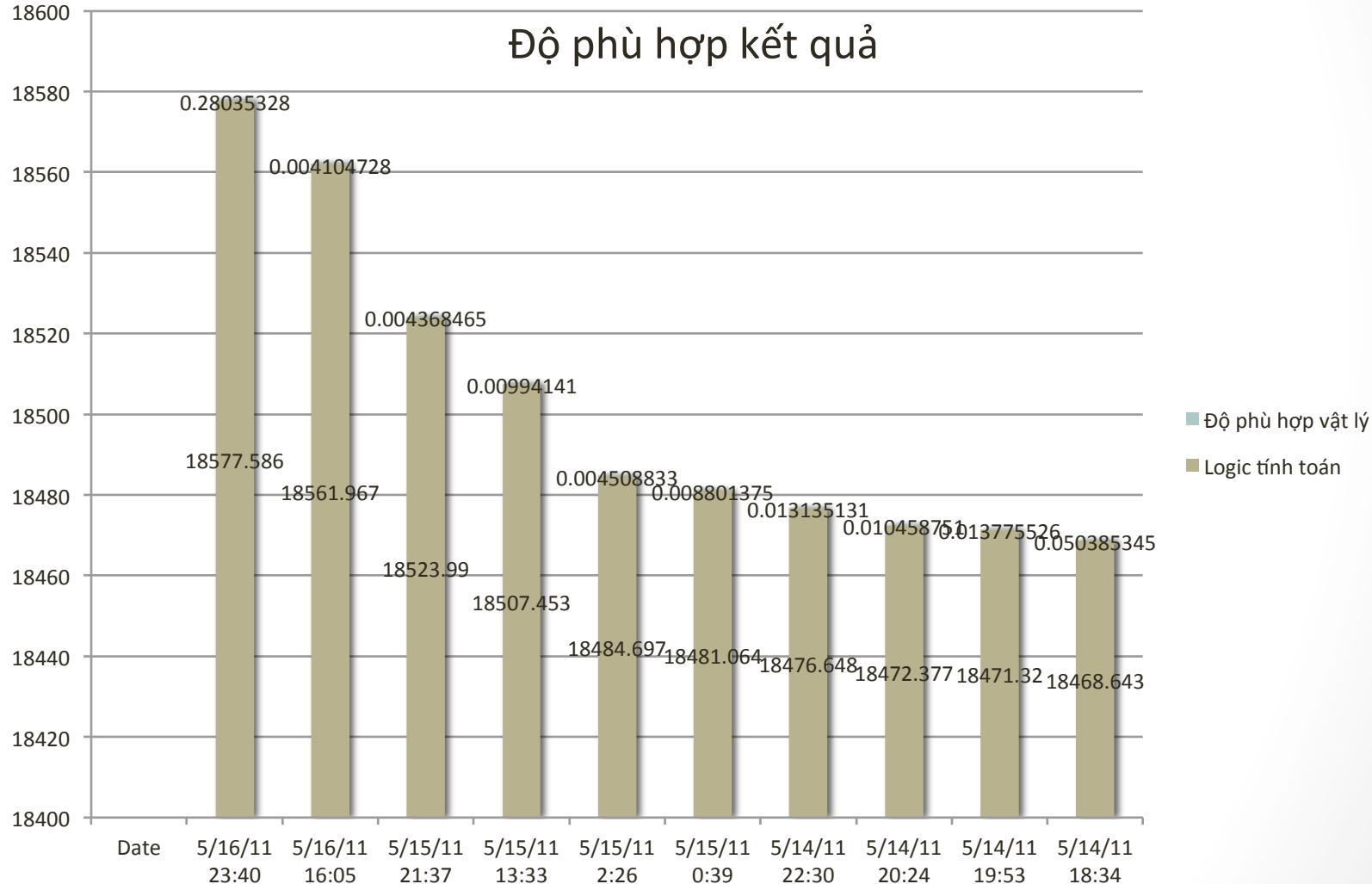
4. CÀI ĐẶT VÀ ĐÁNH GIÁ



Cài đặt



Đánh giá



Đánh giá

Trang chủ | Giới thiệu | Blog | Hướng dẫn | Góp ý Nguyễn Văn Đông Anh ▾

Tri thức | 🔍 Đặt câu hỏi ?

Tiêu đề	Mô tả	Số người quan tâm
C Chuyên môn	Chuyên môn	31 người quan tâm
C# Chuyên môn	Chuyên môn	30 người quan tâm
C++ Chuyên môn	Chuyên môn	23 người quan tâm
Cơ khí Chuyên môn	Chuyên môn	14 người quan tâm
Câu lạc bộ Chuyên môn	Chuyên môn	9 người quan tâm
Công việc Chuyên môn	Chuyên môn	3 người quan tâm
Cơ điện tử Chuyên môn	Chuyên môn	2 người quan tâm
Cơ khí động lực Chuyên môn	Chuyên môn	5 người quan tâm
Cloud Computing Chuyên môn	Chuyên môn	5 người quan tâm
Công nghệ phần mềm Chuyên môn	Chuyên môn	37 người quan tâm
Công nghệ thông tin Chuyên môn	Chuyên môn	53 người quan tâm
Cách để học và thi tốt môn C cơ bản là gì? C		
Ý nghĩa và cách dùng của con trỏ trong lập trình C là gì? C		
Đặt câu hỏi mới: "c"		

tiêu biểu

Tiến Cường Bùi
ws, Công nghệ
nông tin
người này

Câu hỏi tiêu biểu

o có hội thảo cu

ig Thương 11 giờ trước

Phạm Tuấn Long 23

(20)

5. KẾT QUẢ VÀ ĐỊNH HƯỚNG PHÁT TRIỂN



Kết quả

- Xây dựng được **hệ thống tìm kiếm** trên mạng cộng đồng chia sẻ tri thức BKProfile
 - Tìm kiếm chuyên gia theo **3** tiêu chí
 - Tìm kiếm tri thức theo **8** tiêu chí
 - Hỗ trợ gợi ý tri thức theo **2** tiêu chí
- **Bộ tạo dữ liệu** hoạt động tốt, ổn định
- Độ sẵn sàng phục vụ cao

Định hướng phát triển

- Hỗ trợ tìm kiếm theo ngữ nghĩa
- Phân tán hóa với MapReduce

Tài liệu tham khảo

1. David Smiley and Eric Pugh, Solr 1.4 Enterprise Search Server, 1st ed., Packt Publishing Ltd, 2009.
2. Lucid Imagination, Lucidwords for Solr Certified Distribution Reference Guide version 1.4, Lucid Imagination, 2009
3. Jimmy Lin and Chris Dyer, Data-Incentive Text Processing with MapReduce, 1st ed., Morgan and Claypool Publishers, 2010
4. Anthony Arnone and Neal Richther, Search Engine Rodeo, Summer 2007, www.cs.montana.edu/~richter/Search_Engine_Rodeo.pdf, last visited Jan 2011
5. Apache Lucene, <http://lucene.apache.org/>, last visited March 2011
6. Apache Solr, <http://lucene.apache.org/solr/>, last visited March 2011
7. Class Similarity, Apache Lucene, http://lucene.apache.org/java/2_4_0/api/org/apache/lucene/search/Similarity.html, last visited March 2011
8. Yonik, Apache Solr, <http://people.apache.org/~yonik/presentations/Solr.pdf>, last visited Jan 2011
9. Enterprise Search Engines – Critical Success Factors, <http://www.searchtools.com/slides/intranets2006/enterprise-search-critical-success.html>, last visited April 2011
10. Search Engine Process Diagram, SearchTools, <http://www.searchtools.com/slides/bestsearch/bls-03.html>, last visited April 2011

Q&A



PHỤ LỤC



Kết quả tìm kiếm tri thức tiêu biểu



TOEFL

Tiếng anh



Đâu là cách tốt nhất để đạt được TOEFL iBT 90 sau 6 tháng?



Bởi Nguyễn Thị Mai ngày 24/3/2011 lúc 8h:52

5 trả lời, trả lời cuối cùng ngày 11/5/2011 lúc 21h:59 bởi Nguyễn Tuấn

Trả lời được bình chọn nhiều nhất:



Chuẩn bịCó nhiều thứ để em phải học. Đầu tiên theo anh phần quan trọng nhất là phần từ vựng và ngữ pháp. Đây là 2 phần cơ bản nhấ...

Vào ngày 25/3/2011 lúc 19h:55 bởi Nguyễn Văn Đông Anh Xem chi tiết>>



Java



IDE nào là tốt nhất để bắt đầu lập trình Java?



Em mới tự học lập trình Java, có quá nhiều IDE để chọn nên em không biết sử dụng cái nào

Bởi Bùi Tiến Cường ngày 25/3/2011 lúc 19h:38

13 trả lời, trả lời cuối cùng ngày 5/5/2011 lúc 20h:29 bởi Tùng Cheng

Trả lời được bình chọn nhiều nhất:



Anh nghĩ lúc mới bắt đầu lập trình không nên dùng IDE. Sử dụng console và text editor sẽ giúp em hiểu ra được nhiều vấn đề về ngôn ngữ và cách sử dụng Java. Một số text editor anh hay dùng (và nghĩ là hay) bao gồm Notepad++ và vim. Chúc năn...

Vào ngày 26/3/2011 lúc 18h:42 bởi Lê Quốc

Xem chi tiết>>

(27)

Kết quả tìm kiếm tri thức mới



Công nghệ thông tin



Cách xây dựng hàm random?

Trong tin học có hàm Random. Hàm này để chọn một con số ngẫu nhiên trong một tập hợp số nào đó. Em có thắc mắc là người ta xây dựng hàm Random này như thế nào? Tính ngẫu nhiên của nó có đảm bảo khách quan không?

Bởi **Tú Tini** 2 ngày trước lúc 14h:19



1 trả lời, trả lời cuối cùng 19 giờ trước bởi **Phạm Hoàng Hà**

Trả lời được bình chọn nhiều nhất: 0



Vì có hàm sinh ra nó nên không thể nào gọi là ngẫu nhiên được mà chỉ là giả ngẫu nhiên thôi. Có khách quan hay không thì "tùy thuộc vào văn cảnh". Một hàm rất hay dùng là sử dụng phương pháp đồng dư tuyến tính. Với bộ 4 số: modul m, nhân tử a...

Vào 19 giờ trước bởi **Phạm Hoàng Hà**

Xem chi tiết>>



Algorithm

Công nghệ thông tin

Bách Khoa



Tại sao thuật toán dijstra lại không sử dụng được với đồ thị trọng số âm?



Mình đang ôn thi toán rời rạc thấy có vấn đề thắc mắc như trên mong mọi người giải đáp giúp? Và nếu trong trường hợp đồ thị trọng số âm mình cộng thêm k vào để thành đồ thị trọng số dương thì lúc đấy dùng dijstra liệu có chính xác.

Bởi **Bùi Tiến Cường** ngày 31/5/2011 lúc 10h:27

2 trả lời, trả lời cuối cùng 2 ngày trước lúc 2h:5 bởi **Phạm Tuấn Long**

Trả lời được bình chọn nhiều nhất: 3



Đồ thị có cạnh âm hay không thì thuật toán Dijkstra vẫn chạy và vẫn ra được kết quả nào đó. Vấn đề là kết quả đó không đúng. Bởi vì: Sau mỗi bước lặp của Dijkstra, thuật toán sẽ kết nạp một nút nào đó vào danh sách "visited", là danh sách cá...

Vào 2 ngày trước lúc 11h:47 bởi **Phạm Tuấn Long**

Xem chi tiết>>

Kết quả tìm kiếm chuyên gia

Có **297** kết quả phù hợp với truy vấn của bạn



Nguyễn Thanh Vi

Java PHP Công nghệ phần mềm Tiếng anh chuyên ngành công nghệ thông tin Câu lạc bộ
Đoàn Tiếp sức mùa thi Bách Khoa BKProfile TOEIC Âm nhạc

Số câu hỏi: 6 • Số câu trả lời: 23 • Số bình chọn: 77



Nguyễn Mai

Số câu hỏi: 26 • Số câu trả lời: 2 • Số bình chọn: 12



Nguyễn Văn Đông Anh

Java C# TOEFL Javascript MySQL MS SQL Server Mac Công nghệ thông tin Tiếng anh Bách
Khoa BKProfile

Số câu hỏi: 5 • Số câu trả lời: 11 • Số bình chọn: 44



Nguyễn Hằng

Java PHP MS SQL Server OOP Công nghệ phần mềm Hội sinh viên Phim ảnh

Số câu hỏi: 4 • Số câu trả lời: 13 • Số bình chọn: 13



Nguyễn Thị Mai

Số câu hỏi: 1 • Số câu trả lời: 0 • Số bình chọn: 8

Cấu trúc bản ghi

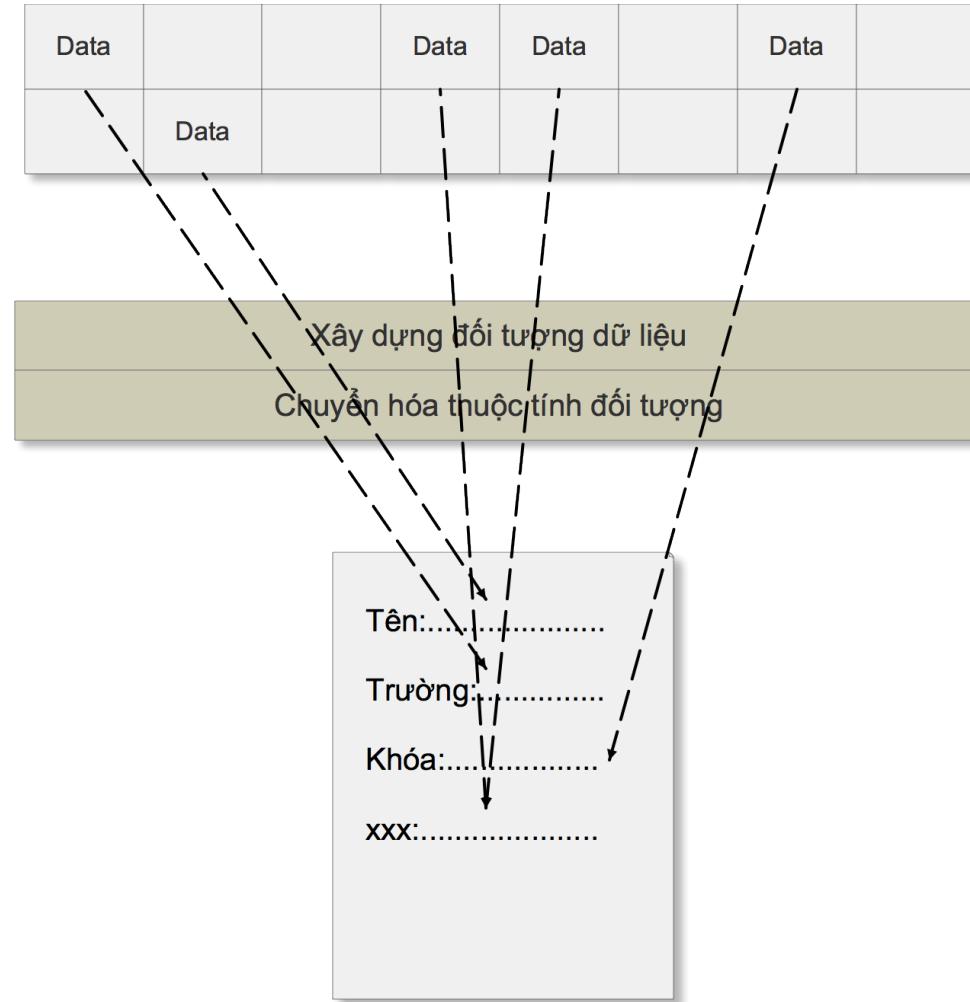
Tên (String):...Nguyễn Văn Đông Anh..

Trường (String):.....ĐHBKHN.....

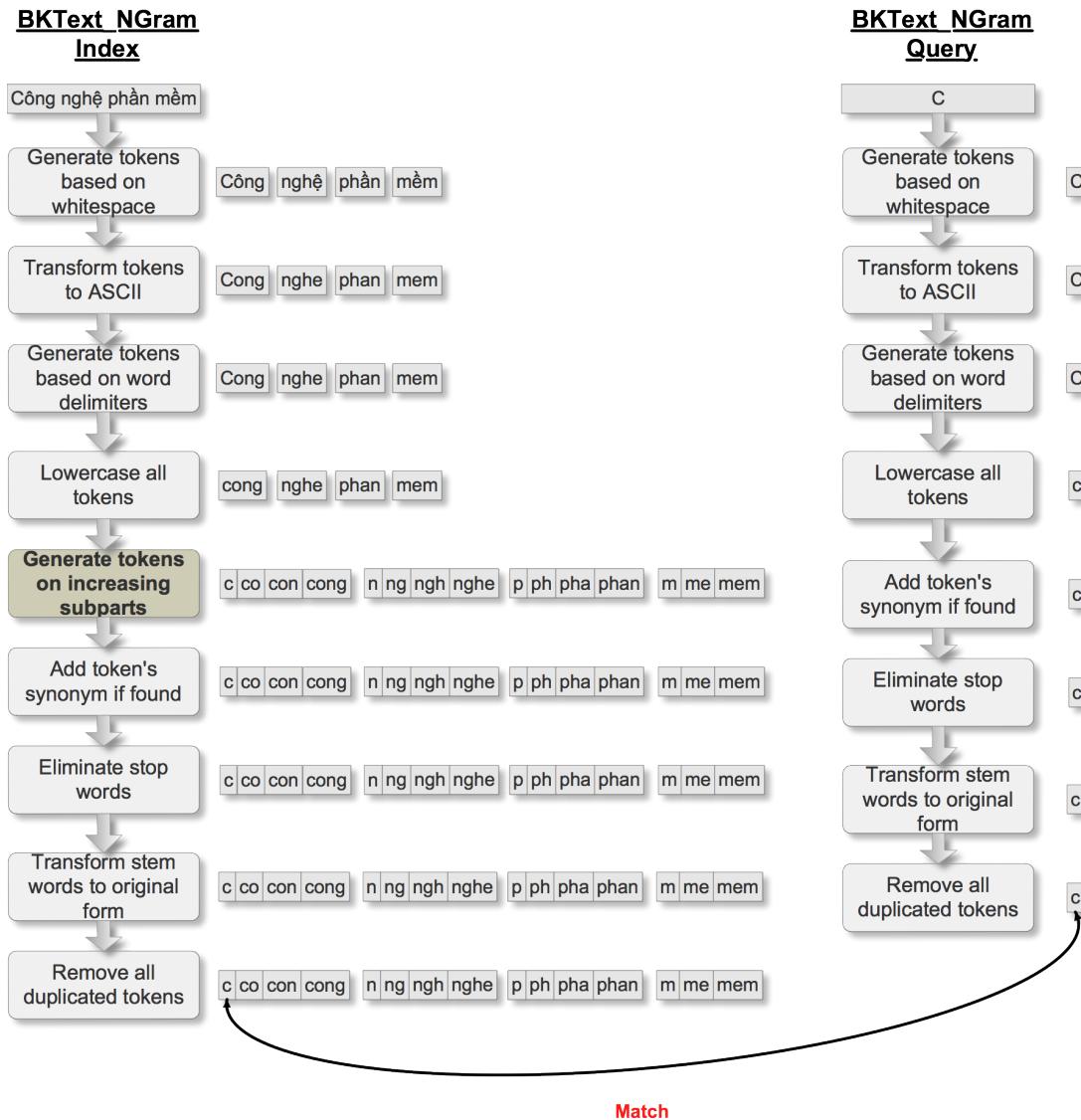
Khóa (Integer):.....51.....

xxx (Customed Type):.....

Mô hình chuyển dữ liệu

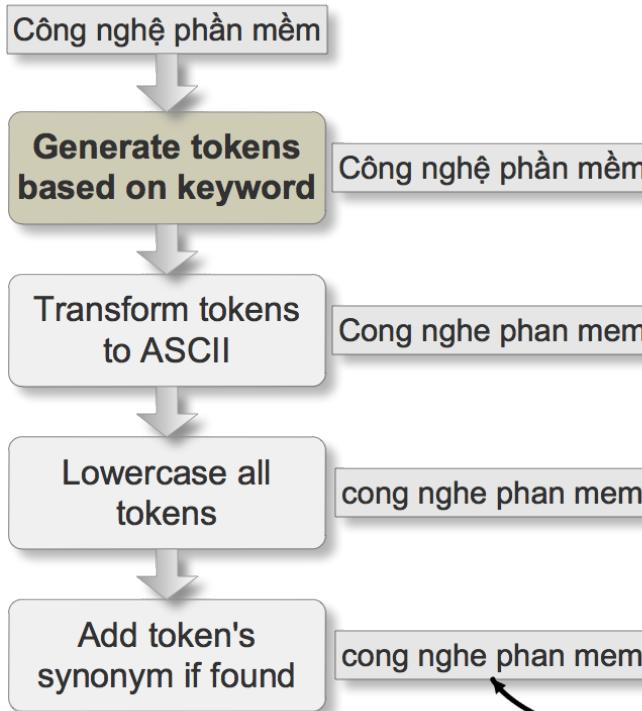


Chi tiết phân tích từ khóa

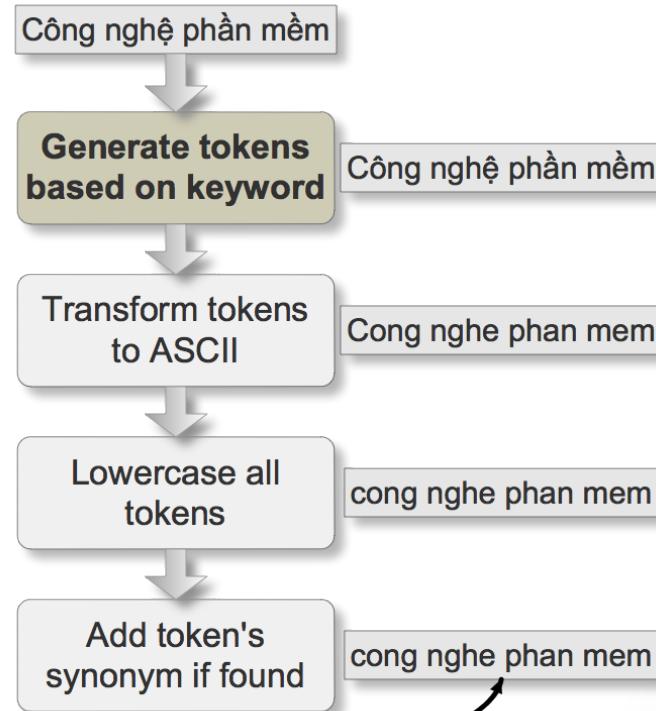


Chi tiết phân tích từ khóa

BKText Filter Index



BKText Filter Query



Match

Chi tiết phân tích từ khóa

