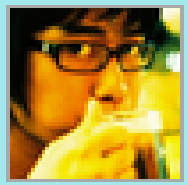
Full Text Search with Apache Solr

Pittaya Sroilong
pittaya@gmail.com

Who am I?

Envious and overconfident blogger is me

03:48 PM March 24, 2008 from im



pittaya

Solr?

Not her!

But a search server

based on Lucene

Lucene?

Full-text search library

100% java
:-(

Solr is based on Lucene

XML/HTTP, JSON interface

Open Source

Shield us from using Java

:-)

Who use Solr/Lucene?

Who use Solr/Lucene?

What is our problem?

How do we implement this?

```
SELECT * FROM post WHERE
topic LIKE '%aoi%' OR author
LIKE '%aoi%' ORDER BY id DESC
```

```
SELECT * FROM post WHERE
(topic LIKE '%aoi%' OR author
LIKE '%aoi%')
OR
(topic LIKE '%miyabi%' OR
author LIKE '%miyabi%')
ORDER BY id DESC
```

Full table scan

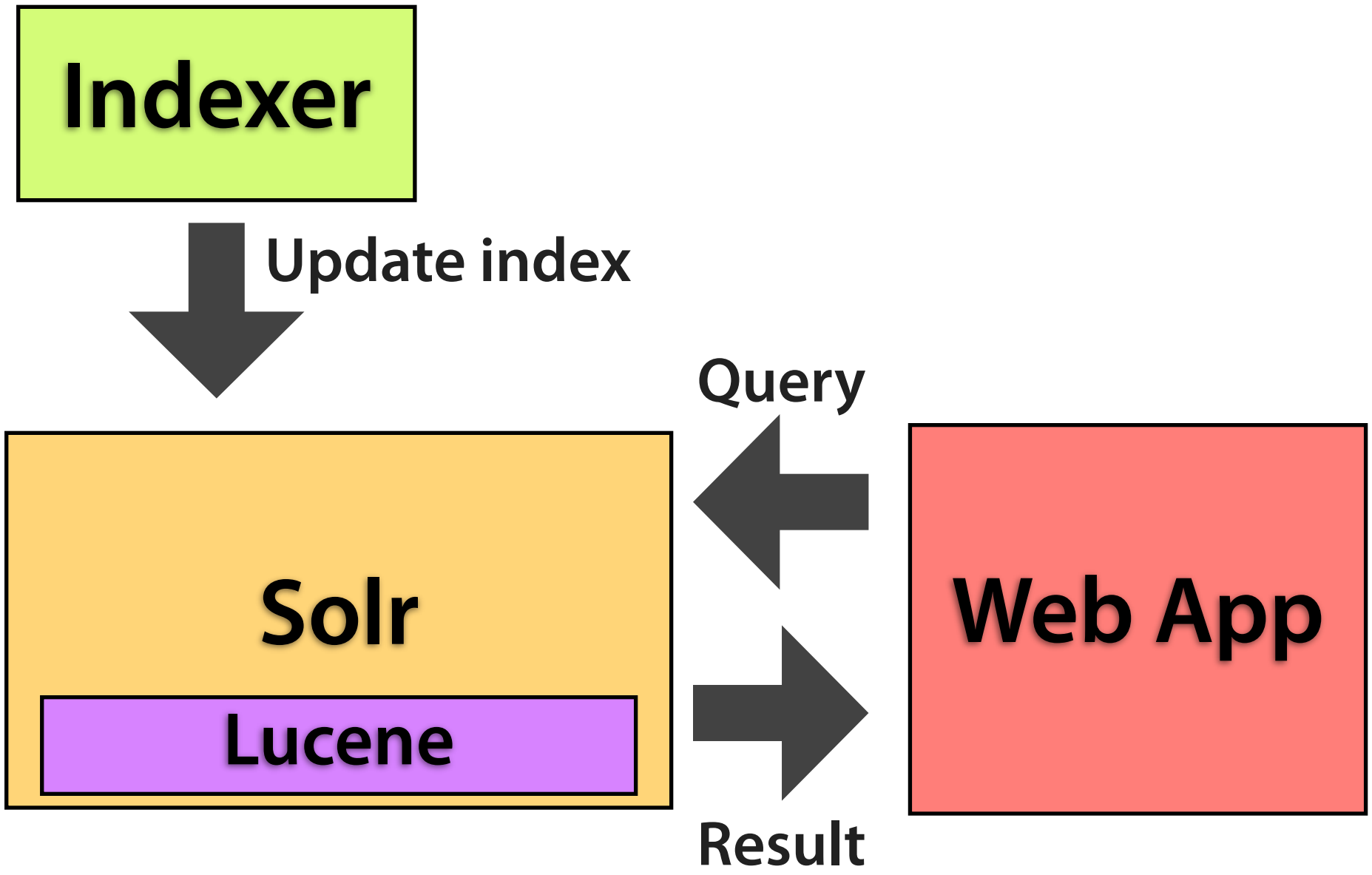
=

Performance killer

No search scoring

RDBMS isn't designed to do this

Use the right tool!


Indexer
Update index
Query
Solr
Lucene
Web App
Result

1

Define schema.xml

```
<field name="id" type="string"
indexed="true" stored="true" />
<field name="fullname" type="string"
indexed="true" stored="true" />
<field name="position" type="string"
indexed="true" stored="true" />
<field name="tag" type="stringi"
indexed="true" stored="true"
multiValued="true" />
```

2

Deploy on any J2EE container

Tomcat, Jetty, etc.

3

Index documents

Document format

```
<add><doc>
 <field name=”id”>555</field>
 <field name=”fullname”>Kaka</field>
 <field name=”position”>Midfielder</field>
 <field name=”tag”>AC Milan</field>
 <field name=”tag”>Brazil</field>
</doc></add>
```

Post to Solr

http://<host>/solr/update

Any language that can do HTTP POST

PHP, Perl, Python

cURL

Commit

<commit />

4

Search

Query from http://<host>/solr/select

Use Solr query syntax

http://<host>/solr/select?
q=tag:madrid&start=0&rows
=2&fl=fullname,position,tag

Response in XML or JSON (configurable)

```
<response>
 <result numFound=”46” start=”0”>
  <doc>
    <str name=”fullname”>Sergio Ramos</str>
    <str name=”position”>Defender</str>
    <str name=”tag”>Real Madrid</str>
    <str name=”tag”>Spain</str>
  </doc>
  <doc>
    <str name=”fullname”>Diego Forlan</str>
    <str name=”position”>Striker</str>
    <str name=”tag”>Atletico Madrid</str>
    <str name=”tag”>Uruguay</str>
  </doc>
 </result>
</response>
```

&wt=json

```
{
 “result”: { “numFound”: 46, “start”: 0,
  “docs” : [
   { “fullname”: “Sergio Ramos”,
     “position”: “Defender”,
     “tag”: [“Real Madrid”, “Spain”] },
   { “fullname”: “Diego Forlan”,
     “position”: “Striker”,
     “tag”: [“Atletico Madrid”, “Uruguay”] }
  ]
 }
}
```

Query examples

- David Pizzarro
  - Equiv: David OR Pizzarro
  - Default operator is "OR" (configurable)
  - Result: David Villa, David Pizzarro, Claudio Pizzarro, David Seaman

- +David +tag:Roma
  - Equiv: David AND tag:Roma
  - Result: David Pizzarro

- +David +position:(Striker OR Midfielder)
  - Result: David Villa, David Pizzarro

Updating

Post new document to

http://<host>/solr/update

Deleting

<delete>
<id>345</id>
</delete>

```
<delete>
<query>tag:Brazil</query>
</delete>
```

```
<delete>
<query>*:*</query>
</delete>
```

Thai support

fwdder.com

Sharing forward mails

ผลการค้นหา: ชุดว่ายน้ำ

เจ๊ียบกับชุดว่ายน้ำ

โดย: ข้าวตุ๋น

1 เดือน 20 วันที่ผ่านมา

4 ความเห็น / 610 คนอ่าน

Tag : sexy น่ารัก ชุดว่ายน้ำ เจ๊ียบ

มิสอเมริกา2008 ชุดว่ายน้ำ

โดย: ข้าวตุ๋น

3 เดือน 17 วันที่ผ่านมา

1 ความเห็น / 425 คนอ่าน

Tag : สวย ชุดว่ายน้ำ มิสอเมริกา

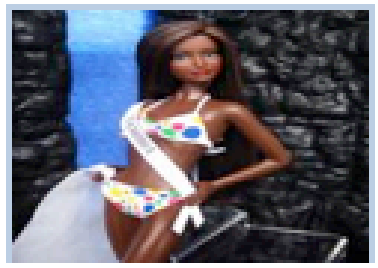
นก-สินจัย ถ่ายชุดว่ายน้ำ รับร้อน

โดย: PaNe

1 เดือน 11 วันที่ผ่านมา

7 ความเห็น / 855 คนอ่าน

Tag : ดารา สวย ชุดว่ายน้ำ star สินจัย



Miss Beauty Doll 2008 ในชุดว่ายน้ำ

โดย: PaNe

5 วัน 14 ชั่วโมงที่ผ่านมา

0 ความเห็น / 65 คนอ่าน

ผลการค้นหา: ชุดว่ายน้ำ แฟชั่น

~ แฟชั่นชุดว่ายน้ำ ~

โดย: kumiko

3 เดือน 28 วันที่ผ่านมา

1 ความเห็น / 421 คนอ่าน

Tag : แฟชั่น ชุดว่ายน้ำ fashion swim wear บิกินนี่

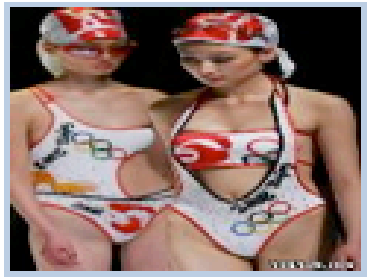
~ แฟชั่นชุดว่ายน้ำ ~

โดย: kumiko

3 เดือน 10 วันที่ผ่านมา

0 ความเห็น / 664 คนอ่าน

Tag : sexy แฟชั่น เซ็กซี่ ชุดว่ายน้ำ ชุดชั้นใน >>



แฟชั่นชุดว่ายน้ำ โอลิมปิก 2008 เจ้าภาพจีน

โดย: ด-ะ-ว้-น-ม-า-แว้-ว

1 เดือน 20 วันที่ผ่านมา

1 ความเห็น / 266 คนอ่าน

Tag : ว่ายน้ำ ก๊ฟ้า sports โอลิมปิก

บิกินีชั้นล่างหลุด ระหว่างเดินแฟชั่น

โดย: Superboy

1 เดือน 28 วันที่ผ่านมา

19 ความเห็น / 4893 คนอ่าน

Use customized field in schema.xml

```
    <fieldType name="html_th" class="solr.TextField"
positionIncrementGap="100">
      <analyzer type="index">
        <tokenizer
class="solr.HTMLStripStandardTokenizerFactory"/>
        <filter class="solr.ThaiWordFilterFactory" />
        <filter class="solr.StopFilterFactory"
ignoreCase="true" words="stopwords.txt"/>
        <filter class="solr.LowerCaseFilterFactory"/>
        <filter class="solr.EnglishPorterFilterFactory"
protected="protwords.txt"/>
        <filter
class="solr.RemoveDuplicatesTokenFilterFactory"/>
      </analyzer>
    </fieldType>
```

```
<field name="id" type="string"
indexed="true" stored="true" />
<field name="title" type="html_th"
indexed="true" stored="true" />
<field name="detail" type="html_th"
indexed="true" stored="true" />
<field name="tag" type="stringi"
indexed="true" stored="true"
multiValued="true" />
<field name="userid" type="integer"
indexed="false" stored="true" />
```

Index analyzer

Field name	title
Field value (Index) verbose output ☑ highlight matches ☑	การตื่นตัวของปรากฏการณ์โลกร้อน
Field value (Query) verbose output ☐	โลก
	Analyze

Index Analyzer

org.apache.solr.analysis.HTMLStripStandardTokenizerFactory {}

term position	1
term text	การตื่นตัวของปรากฏการณ์โลกร้อน
term type	<ALPHANUM>
source start,end	0,30

org.apache.solr.analysis.ThaiWordFilterFactory {}

term position	1	2	3	4	5	6
term text	การ	ตื่นตัว	ของ	ปรากฏการณ์	โลก	ร้อน
term type	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>	<ALPHANUM>
source start,end	0,3	3,10	10,13	13,23	23,26	26,30

Debugging

&debugQuery=on

Further readings

- http://lucene.apache.org/solr/
- http://wiki.apache.org/solr
- http://www.xml.com/pub/a/2006/08/09/solr-indexing-xml-with-lucene-andrest.html
- http://lucene.apache.org/java/docs/scoring.html

Q & A