Topics

What is a decentralized search engine? and why would you use that

Architecture
details about the YaCy technology

Integration
of YaCy in your web pages and services
We do not want centralization (of search engines)

we want:
freedom of information
anonymity when doing web search
YaCy has an integrated NoSQL Database. The database stores a Reverse Word Index, Metadata and the source documents.
Architecture of Large-Scale Search

horizontal scaling: more documents

vertical scaling: more queries per second
Construction of a Large-Scale Search Engine

Large Search Engine in a Data Center
The Large-Scale Search Engine in your Home!
YaCy connects search peers with a peer-to-peer protocol
YaCy peers store index fragments according to a ‘folded’ ordering on word-hashes and url-hashes in a distributed hash table (DHT). The index is distributed redundantly to save the index when some peers are not available. The redundancy also helps to increase search performance.
Community & Personal Use of Search Engines

Decentralized Search

non-Cloud Search
(keep your secrets)
Productivity #1/5: Project Search Engine

Your Project

Discussions  Wiki  Code  Bugtracker

Search Engine

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info
Michael Christen
http://yacy.net

Usage granted by
Productivity #2/5: Keep Secrets!

Enterprise Environment

Search Engine

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Michael Christen
http://yacy.net

Usage granted by
### Productivity #3/5: Personal Relevance

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appearance In Reference/Anchor Name</td>
<td>13</td>
</tr>
<tr>
<td>Position Of Phrase</td>
<td>8</td>
</tr>
<tr>
<td>Appearance In URL</td>
<td>14</td>
</tr>
<tr>
<td>Date</td>
<td>4</td>
</tr>
<tr>
<td>Hit Count</td>
<td>9</td>
</tr>
<tr>
<td>Preferred Language</td>
<td>13</td>
</tr>
<tr>
<td>Phrases In Text</td>
<td>3</td>
</tr>
<tr>
<td>Term Frequency</td>
<td>14</td>
</tr>
<tr>
<td>Position In Text</td>
<td>10</td>
</tr>
<tr>
<td>Category Image Appearance</td>
<td>0</td>
</tr>
<tr>
<td>Word Distance</td>
<td>15</td>
</tr>
<tr>
<td>Appearance In Emphasized Text</td>
<td>10</td>
</tr>
<tr>
<td>Category Index Page</td>
<td>0</td>
</tr>
<tr>
<td>Appearance In Tags</td>
<td>13</td>
</tr>
<tr>
<td>Domain Length</td>
<td>11</td>
</tr>
<tr>
<td>Category Video Appearance</td>
<td>0</td>
</tr>
<tr>
<td>Category App, Appearance</td>
<td>0</td>
</tr>
<tr>
<td>Appearance In Title</td>
<td>14</td>
</tr>
<tr>
<td>URL Length</td>
<td>13</td>
</tr>
<tr>
<td>YaCy Block Rank</td>
<td>9</td>
</tr>
<tr>
<td>Links To Local Domain</td>
<td>2</td>
</tr>
<tr>
<td>Position In Phrase</td>
<td>1</td>
</tr>
<tr>
<td>Category Audio Appearance</td>
<td>0</td>
</tr>
<tr>
<td>Words In Title</td>
<td>4</td>
</tr>
<tr>
<td>Appearance In Author</td>
<td>12</td>
</tr>
<tr>
<td>Authority of Domain</td>
<td>11</td>
</tr>
<tr>
<td>Links To Other Domain</td>
<td>3</td>
</tr>
<tr>
<td>URL Components</td>
<td>12</td>
</tr>
<tr>
<td>Words In Text</td>
<td>2</td>
</tr>
</tbody>
</table>

*that's what Lucene has*

*similar to Google PR*

*in YaCy, you can combine many weighted attributes*
Productivity #4/5: Download Helper

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info
Michael Christen
http://yacy.net
Usage granted by
Productivity #5/5: Become Independent

Data → Search → User

Free Software
Data under Creative Commons License
Open Access Repositories

as it is today:
PROPERTY & CENTRALISED:
it traces you & data can be censored, blocked, removed, spammed

User needs proprietary & centralised software to discover free content

is this what we want?

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info
Michael Christen http://yacy.net
Usage granted by
Impact of running your own search engine:

1. connect tools and people in projects
   free software projects need free search

2. keep secrets
   search tracks can reveal industrial research targets

3. your personal relevance
   create a ranking method for your personal needs

4. do more with search
   for example file sharing and downloading

5. support freedom
   free information cannot be free without free search
Examples #1/2: Search Cluster in a Data Center

'Sciencenet': Search Engine for scientific content in the Karlsruhe Institute of Technology:
34 computers running YaCy in its own network

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Michael Christen
http://yacy.net

Usage granted by
http://sciencenet.fzk.de
300 million documents
Examples #2/2: Decentralised Search for Everyone

People run their own YaCy search peer at home and create independent search for everyone.

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Michael Christen
http://yacy.net

Usage granted by
Architecture #1/4: The Search Engine Network

The 'freeworld' YaCy Search Engine Network

Peer Types:
- **Junior**: behind firewall or router
- **Senior**: has open server port
- **Principal**: publishes seed-lists

**DHT-Store**
**DHT-Read**

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Michael Christen
http://yacy.net

Usage granted by

T]
Architecture #2/4: Snippets & Link Verification

SRU
Facets
File Types, Protocols, Domains, Authors

every link is verified
before it is displayed: the content is loaded, parsed and used for a search snippet generation

Standards
APIs Opensearch (search results with RSS), JSON, AJAX tools
Tools search widget, ready-to-use code snippets to embed search everywhere

Michael Christen
http://yacy.net

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Usage granted by
Architecture #3/4: Data Acquisition

Crawler
with target host balancing

- target hosts (domain name)
- round-robin access
- robots.txt, latency and minimum access time 0.5s
- loader

Scan Sources
in a specific network

- Scan IP Range
- Discover Services
- Availability Mngt.
- RSS Feeds

OAI-PMH Loader
load opac records from libraries

Import Files
- Dublin Core Files
- Wikimedia Dump

Scan Sources
- FTP
- SMB

OAI-PMH Loader

Loader

Architecture #3/4: Data Acquisition

OAI-PMH Loader

Load opac records from libraries

Import Files

- Dublin Core Files
- Wikimedia Dump

Scanners Sources

- FTP
- SMB

Parser

- HTML, XHTML, RSS, RDF, XHTML+RDFa, FOAF, vCard, Flash, PDF, PS, Word, Excel, Visio, Powerpoint, OpenOffice, RTF, csv, gzip, zip, tar, rar, bzip2, 7zip, images(EXIF), Dublin Core XML, torrent files

Indexer

YaCy, Web Search by the people, for the people

@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Michael Christen
http://yacy.net

Usage granted by

Crawler

with target host balancing

- target hosts (domain name)
- round-robin access
- robots.txt, latency and minimum access time 0.5s
- loader
Architecture #4/4: Production / Monitoring

Network Animation

Scheduler

Data Visualization

Connections, Queues, Database

YaCy, Web Search by the people, for the people
@ RMLL - Rencontres Mondiales du Logiciel Libre - http://2011.rmll.info

Michael Christen
Usage granted by
http://yacy.net
How to integrate a YaCy Search Portal:
Just copy-paste the code snippet to your web page source code.

Code Snippet #2 looks like:

MySearch

Search

The YaCy administration interface offers more code snippets. An example from /ConfigSearchBox.html looks like:

MySearch

Search

Integration #1/3: Search Interface Integration

Code Snippet Example #1: a search window in an iframe

```
<iframe name="target2"
    src="http://141.52.175.43:8080/yacysearch.html?
    display=2&resource=local"
    width="100%" height="180"
    frameborder="0" scrolling="auto" id="target2"
    frameborder="0" scrolling="auto" id="target2">
</iframe>
```

Code Snippet Example #2: a search box (points to new page)

```
<form method="get" accept-charset="UTF-8"
    action="http://141.52.175.43:8080/yacysearch.html">
    <div>
        <div>MySearch</div>
        <input type="text" name="query" value="" maxlength="80" />
        <input type="hidden" name="verify" value="true" />
        <input type="hidden" name="maximumRecords" value="10" />
        <input type="hidden" name="meanCount" value="5" />
        <input type="hidden" name="resource" value="local" />
        <input type="hidden" name="urlmaskfilter" value=".*" />
        <input type="hidden" name="prefermaskfilter" value="" />
        <input type="hidden" name="display" value="2" />
        <input type="hidden" name="nav" value="all" />
        <input type="submit" name="Enter" value="Search" />
    </div>
</form>
```

Your YaCy peer provides help pages with code snippets for an easy integration!
Integration #2/3: External Index Retrieval

```bash
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type='text/xsl' href='/yacysearch.xsl' version='1.0'?>
<rss version="2.0"
    xmlns:yacy="http://www.yacy.net/
    xmlns:opensearch="http://a9.com/-/spec/opensearch/1.1/"
<!--[CDATA[very short example -->
<item>
    <title>Friend of a Friend (FOAF) project</title>
    <link>http://www.foaf-project.org</link>
    <pubDate>Fri, 23 May 2008 02:00:00 +0200</pubDate>
</item>
<item>
    <title>FOAF - Wikipedia</title>
    <link>http://de.wikipedia.org/wiki/FOAF</link>
    <pubDate>Tue, 08 Jan 2008 01:00:00 +0100</pubDate>
</item>
<item>
    <link>http://microformats.org/wiki/xfn-to-foaf</link>
    <pubDate>Fri, 09 May 2008 02:00:00 +0200</pubDate>
</item>
</rss>
```

**How to get Opensearch/JSON Search Results:**
- do a normal web search in YaCy
- replace the 'html' extension of the result page URL with 'rss'
- for json, replace the 'html' extension with 'json'

**Standards:**
The YaCy-internal Dublin Core Metadata Format fits into the RSS format for search result data in Opensearch standard very well.
If wanted, also JSON can be used as export format.

**Opensearch Standard:** [http://www.opensearch.org](http://www.opensearch.org)
**SRU Standard for Queries:** [http://www.loc.gov/standards/sru/specs/search-retrieve.html](http://www.loc.gov/standards/sru/specs/search-retrieve.html)
Integration #3/3: External Index Feeding

<xml version="1.0" encoding="utf-8">  
<!-- YaCy surrogate using dublin core notion -->  
<surrogates xmlns:dc="http://purl.org/dc/elements/1.1/">  

<record>  
  <dc:title><![CDATA[Alan Smithee]]></dc:title>  
  <dc:description><![CDATA['''Alan Smithee'' ist ein Anagramm von „The Alias Men“.]]></dc:description>  
  <dc:language>de</dc:language>  
  <dc:date>2009-04-14T00:00:00Z</dc:date>  
  <!-- date is in ISO 8601 -->  
</record>  
</surrogates>

Standards:  
YaCy can import standard Dublin Core Metadata XML files as input for indexing

How to import Dublin Core Files:  
just place the xml files into a hand-over directory at DATA/SURROGATES/in/

The Dublin Core XML File Standard:  
http://dublincore.org/documents/dc-xml-guidelines/
Where is a (demo) Search Portal?

There is no one-for-all demo portal for YaCy!
YaCy is about decentralized search and offering a central point for everyone would ruin the idea!

Decentralized Search in your browser:

http://peer-search.net
- JavaScript Code is loaded into your browser
- your browser loads a list of YaCy peers
- when you search, your browser contacts some of the YaCy peers and combines the search results from these peers; like a meta-search.

Peer Roulette, search on a random peer:

http://www.yacyweb.de/peers.htm
- yacyweb generates a list of YaCy peers
- when you click on a link you get the web interface of the peer directly
- when you search on that peer the content may be restricted to the rules of the peer owner

The best demo: run your own peer!
Thank You!

Download
- download YaCy from http://yacy.net

Please Help!
- the french interface translation and wiki pages
- run a peer
- become a developer

French Support Forum
- we don’t have that (yet). Please start a french forum!