

Grow-your-own search engine



An introduction to Solr
Andrew Clegg

<http://lucene.apache.org/solr/>



Search engine toolkit

Indexing

Relevance ranking

Spellchecking

Hit highlighting

Text processing tools

etc. etc.

Originally in Java

Ported to C#.NET, Python, C (*Lucy*)

Perl & Ruby bindings for Lucy

Not an off-the-shelf application — requires coding

Solr is a search server built on Lucene



Web admin interface

Configured via XML — no coding required

Controlled via HTTP — REST-like web services

Database-like schema — typed fields, unique keys

Extended query syntax

Caching, replication, clustering, delta indexing, sharding

Used by: AOL, news.com, Digg, Apple, Disney, MTV, CiteSeerX, PubGet, NASA, the White House...

Indexing



'documents'



indexing



index

XML

CSV

HTTP request

Database query

Tokenization etc.

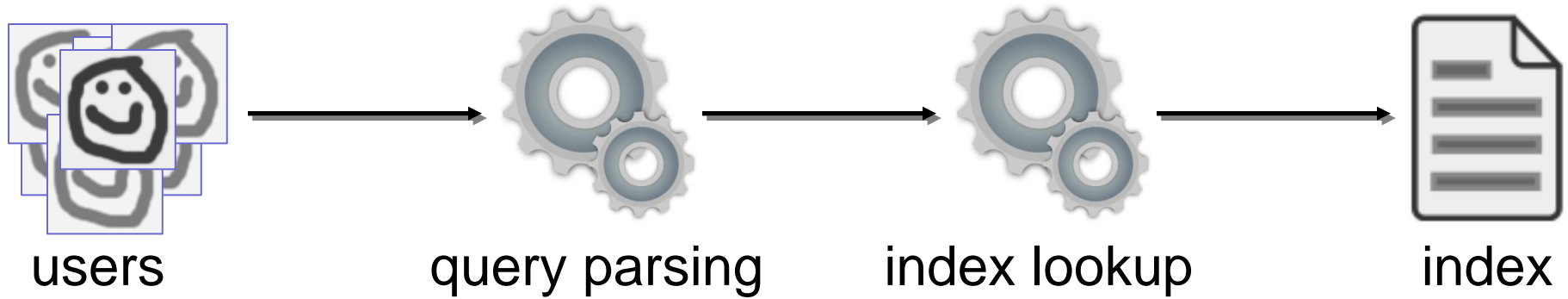
Term extraction

Frequency

calculations

Web crawling & PDF/Word/etc. extraction via other tools

Querying



ranked hits:
database/document IDs
... and/or ...
content **stored** in index

Query syntax example:

name:(kinase OR phosphotransferase)
AND species:Saccharomyces
AND description:phosphoryl*
AND createdate:[-1YEAR/DAY TO NOW]

“name contains kinase or phosphotransferase
and species contains Saccharomyces
and description contains word starting with phosphoryl
and createdate at any time between one year ago
(rounded to the nearest day) and now”

Ranking based on *tf.idf* model...

tf = term frequency

idf = inverse document frequency

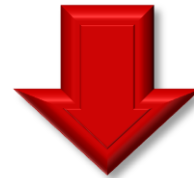


Query terms appearing frequently in *this* document are **up-weighted**



... but ...

Query terms appearing in *many* documents are **down-weighted**



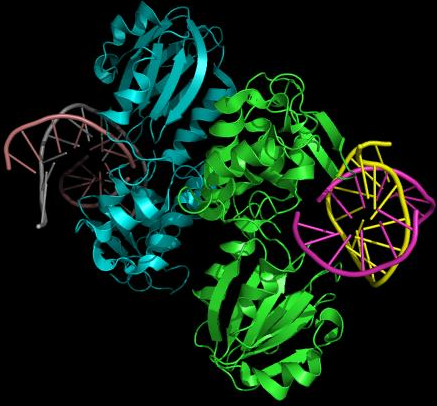
Highest ranked docs contain multiple instances of rare query terms

Case study: indexing the CATH protein domain database with Solr

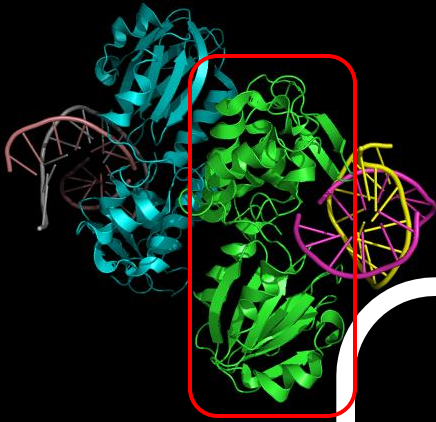


<http://cathdb.info/>

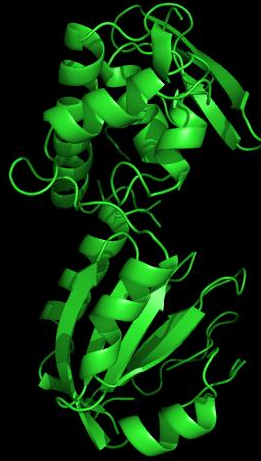
Whole protein



Whole protein



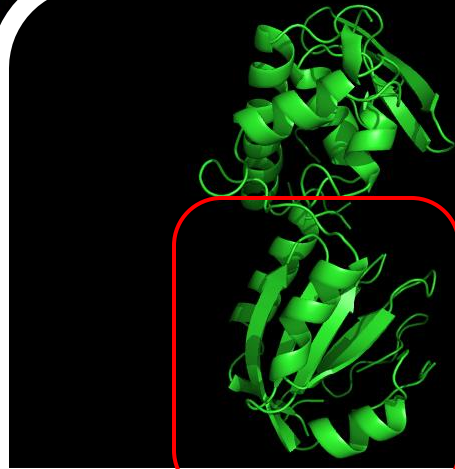
Chain



Whole protein



Chain



Domain



CLASS

e.g. *Mixed Alpha/Beta* depth = 1

ARCHITECTURE

e.g. *Alpha/Beta Barrel* depth = 2

TOPOLOGY
(fold family)

e.g. *TIM Barrel* depth = 3

**HOMOLOGOUS
SUPERFAMILY**

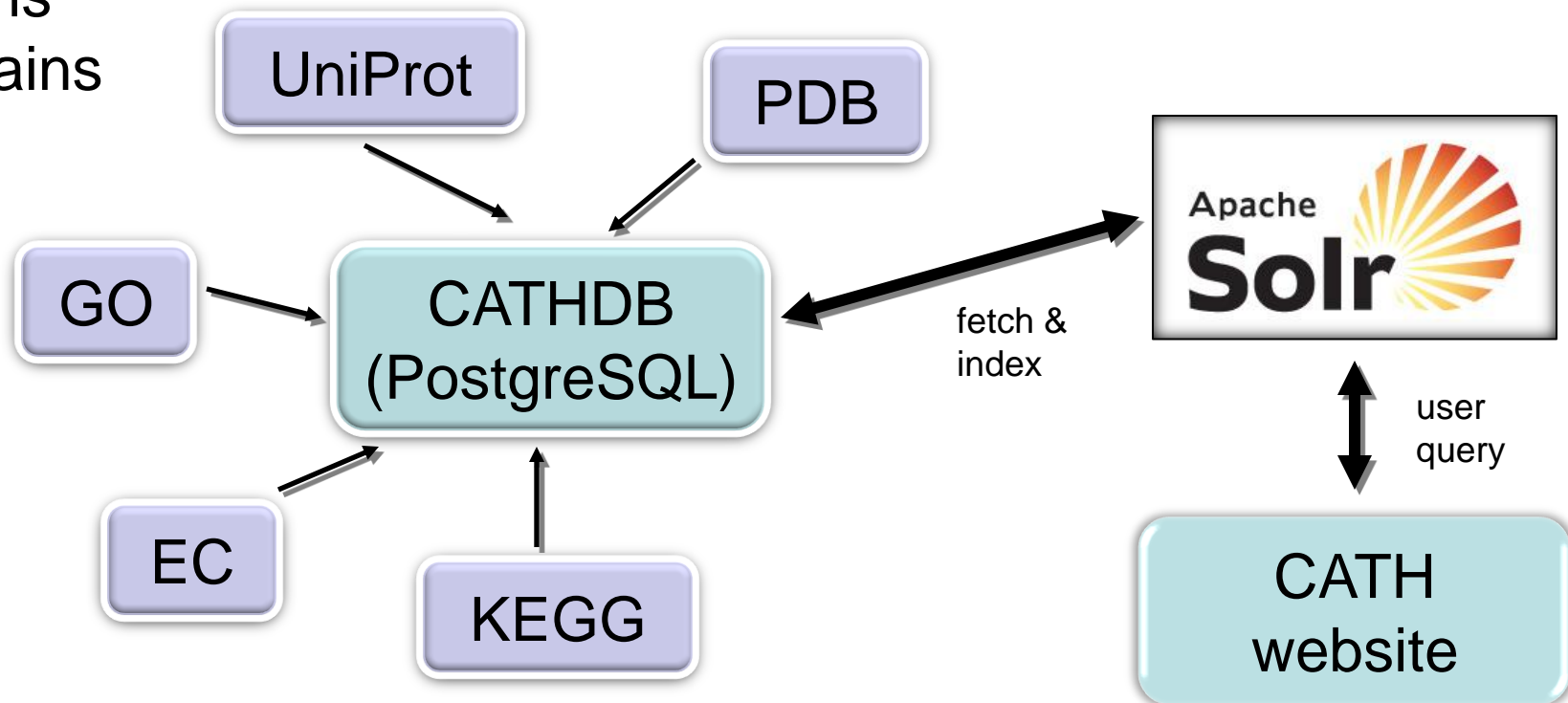
e.g. *Aldolase class I* depth = 4

Previous CATH search box used plain SQL –
lots of LIKE>equals matches against relevant fields

- Slow
- Poor coverage (exact string match needed)
- No relevance ordering
- Slow
- Hard to debug unexpected results
- Can't be exposed as a service
- Did we mention slow...?

Entities to index:

- CATH nodes, esp. superfamilies
- Proteins
- Chains
- Domains



Fetching/indexing data needs *no coding* – just XML and a bit of SQL

CATH Search: "aldolase"

Search **Result Summary** Cathnodes (11) Domains (778) Chains (864) Pdb (267)



Query 'aldolase' matched 1920 entries in the CATH database. The top hit for each type of entry given below. Click on the 'more' link or the tabs to see all the results.

**Cathnode: Aldolase class I**

Related Keywords: Pyrimidine nucleotide/nucleoside/nucleobase anabolism Uridine 5'-monophosphate synthase Biosynthesis of histidine Pyruvate carboxylase activity Glycosphingolipid catabolic process Salmonella typhi Geobacillus stearothermophilus Malate dehydrogenase Clostridium botulinum Paracoccus versutus L-lactate dehydrogenase (cytochrome) activity Hahella chejuensis KCTC 2396 NADPH dehydrogenase 2 Mll9387 protein Dihydroorotate dehydrogenase B, catalytic subunit Lipid, fatty acid and isoprenoid metabolism Jasmonic acid biosynthetic process 12-oxophytodienoate reductase activity Tryptophan biosynthesis protein trpCF Dihydrodipicolinate synthase Sus scrofa Soluble fraction Orotidine-5'-phosphate decarboxylase activity Drosophila melanogaster Escherichia coli O157:H7 Phosphoglycerate kinase Peroxisome Glycolysis / Gluconeogenesis Indole-3-glycerol phosphate synthase Bacillus halodurans Fructose-bisphosphate aldolase, glycosomal Identical protein binding Neisseria gonorrhoeae FA 1090 Fructose-bisphosphate aldolase (S)-mandelate dehydrogenase Fructose and mannose metabolism Dihydropyrimidine dehydrogenase (NADP+) ATP binding Pyrobaculum aerophilum Vitamin B6 metabolism Borrelia burgdorferi Porphyrin and chlorophyll metabolism Thermus thermophilus Thermoproteus tenax Thermotoga maritima KHG/KDPG aldolase Biotin synthase Metabolism of porphyrins 2-dehydro-3-deoxyphosphogluconate aldolase Tagatose-1,6-bisphosphate aldolase kbaY

<http://www.cathdb.info/cathnode/3.20.20.70> - [View all 11 matching cathnodes](#)

**Domain: PDB code 1dxs, chain A, domain 00**

Related Keywords: Non-polymer CLASS II ALDOLASE Water Polymer Polypeptide(L) Escherichia coli K12 Hexameric 2-dehydro-3-deoxyglucarate aldolase 2-DEHYDRO-3-DEOXY-GALACTARATE ALDOLASE PHOSPHATE ION Ascorbate and aldarate metabolism MAGNESIUM ION

<http://www.cathdb.info/domain/1dxsA00> - [View all 778 matching domains](#)

**Pdb: PDB code 1dxs**

Related Keywords: Non-polymer CLASS II ALDOLASE Water Polymer Polypeptide(L) Escherichia coli K12 Hexameric 2-dehydro-3-deoxyglucarate aldolase 2-DEHYDRO-3-DEOXY-GALACTARATE ALDOLASE PHOSPHATE ION Ascorbate and aldarate metabolism MAGNESIUM ION

<http://www.cathdb.info/pdb/1dxs> - [View all 267 matching pdbs](#)

**Chain: PDB code 1dxs, chain A**

Related Keywords: Non-polymer CLASS II ALDOLASE Water Polymer Polypeptide(L) Escherichia coli K12 Hexameric 2-dehydro-3-deoxyglucarate aldolase 2-DEHYDRO-3-DEOXY-GALACTARATE ALDOLASE PHOSPHATE ION Ascorbate and aldarate metabolism MAGNESIUM ION

<http://www.cathdb.info/chain/1dxsA> - [View all 864 matching chains](#)

Fun with indexes

- *“Give me the 10 most similar domains to this one”*

Based on term similarity.

Useful for quickly finding functional analogues.

- *“Give me the top 10 keywords for this superfamily”*

Can be used to generate tag clouds & summaries.

Useful for data mining too.

Some facts and figures

- Indexing takes ~ 6 hours (dedicated box)
- Index is currently 2.0GB
- 357,850 entities in index (some are 1-3MB)
- Typical single query time in Solr: 1 **ms** – 200 **ms**
- Total web query time: ~ 3s

Any questions?

