FastCCI
Taming the Commons Category Tree
[[User:Dschwen|Daniel Schwen]]
Commons category tree?

It it really a tree?

- Made-up example
- Diamond relations
- It gets worse...
- Loops!

It is a directed graph!

- Let's be careful with the arrow direction
- Well-defined element relation "is subcategory of"

The actual meaning is very loosely defined:

- **Hyponymy**: a sort/kind/type of...
- **Meronymy**: a part of..., a member of...
- **Attributes**: color, shape, size, nationality, technique, quality...
- **Location**: where, in..., from... (place/event, place/building, place/people, country/language...)
- **Timing**: when (time/event, time/depicted situation, time of birth, inception or construction...)
- **Agentive and influence relations**: creator/work, device/product, company/product...
- **Modification**: original/modified or modified/original (avoid cyclic structure)
Loops!

Number of clusters / loops

Category loops
Strongly connected clusters

Self-referencing categories

Loop / Cluster size

Supercluster

http://www.schwen.de/jside/fast/cc.html#_print_
Bundle similar images

Media in category "Mister Softee trucks in Hong Kong"

The following 14 files are in this category, out of 14 total.

Category: Mister Softee trucks in Hong Kong

'So happy smiling cat' by Yumi Kimura from Yokohama, JAPAN. Licensed under CC BY-SA 2.0 via Wikimedia Commons.
Categories hide content

Pages in category "Paris"

This category contains only the following page.

- Paris

Media in category "Paris"

The following 5 files are in this category, out of 5 total.

- **Avenue Victoria and Hotel de Ville, Paris 2014.jpg** (CopyVio)
  - 736 KB
  - 1.04 MB

- **Dans les rues de Paris, Paris 2014.jpg** (CopyVio)
  - 876 KB

- **Place d'Italie, Paris 2014.jpg** (CopyVio)
  - 1.87 MB

- **The Education of the Princess in the Louvre, Paris 2014.jpg** (CopyVio)
  - 659 KB


Categories (+4): Cities and villages in France by department (-) (±) (1)  Cities and villages in Île-de-France (-) (±) (1)  Paris urban area

Non-topical/index: Categories with a gallery for a better choice of sub-categories (-) (±) (1)

Diffusion into countless sub-categories makes discovering images hard.
Categories *can* encode complex relations

- More than just a simple *and*:
  - Category: Star symbols *on top of* objects
  - Category: People *inside* flowers
    - Category: People *holding* flowers
    - Category: People *with* flowers
  - Category: Nandi sculptures *outside* India
    - 6 levels below *Category: India*
  - Category: Sundials *without* sun
    - 5 levels below *Category: Sun*
Categories *do* encode *complex* relations

Category: Demographic maps of 15-17 year old dependent children whose fathers speak another language and did not state proficiency in English and whose mothers speak another language and speak English not well or not at all in Australia by state or territory

And this monster has *sub-categories!*
Some categories are natural intersections

\[ \text{Art of Japan} \]
\[ \text{Metalwork in Japan} \]
\[ \text{Sculptures in Japan} \]
\[ \text{Metal sculptures in Japan} \]

= Japan & Metalworks & Sculptures

2.850.563 category diamonds found (not all permutations!)

./fastcci_diamond(<3s)
What about using tags?

- Tags or "Topics" are commonly used in popular Photo communities (Flickr, 500px)
- No hierarchy
- We may get tags (through WikiData), however:
  - Don't hold your breath!
  - Community consensus?
  - Additional maintenance burden?
  - Can not replace all Category applications
  - Won't solve today's problems
Category intersection

- Treat categories as Tags

- Certainly works for higher level categories
  Kittens, Wyoming, Bicycles, Green

- Consider all images in and below all sub-categories

- Meaning
  - All images which both labels apply to

- Problem
  - Higher up categories lose meaning
    the further "down" in the graph we go
The longest Category chain

A metric to weigh intersection results:

Path lengths in the graph
Subcategory depth

How many categories do we have on commons with 1, 2, 3,..n levels below?

Ye olde intersection (The SQL way) = recursive queries = painfully slow!
FastCCI

Category intersection needs...

- ...to be **fast**

- ...an easy interface
  - shallow learning curve
  - powerful options available
FastCCI's Killer Application

- Parallel category trees for ⭐ FP, 🌱 QI, 🍋 VI
  - ⭐ Category: Featured picture
    640 × Category: Featured pictures of...
  - 🌱 Category: Quality image
    1825 × Category: Quality images of...
  - 🍋 Category: Valued image
    578 × Category: Valued Images of...

- Ideal target for intersection
- Easy interface → one button!
FastCCI Interface
FastCCI Interface
FastCCI Interface

![Image of FastCCI Interface](https://commons.wikimedia.org/wiki/File:Glacier_Point_at_Sunset,_Yosemite_NP,_CA,_US_-_Diliff.jpg)

**File:Glacier Point at Sunset, Yosemite NP, CA, US - Diliff.jpg**

*Idaho → Geography of Idaho → Great Basin (USA) → Mountain ranges of the Great Basin → Sierra Nevada (United States) → Mountains of the Sierra Nevada (U.S) → Mountains in Yosemite National Park → Half Dome*

From Wikimedia Commons, the free media repository

- Nominate this image for QI
- File
- File history
- File usage on Commons
- File usage on other wikis
- Metadata

---

"Glacier Point at Sunset, Yosemite NP, CA, US - Diliff" by Diliff - Own work. Licensed under CC BY-SA 3.0 via Wikimedia Commons.
FastCCI Advanced Interface

Category: Wyoming


In this category and in...
- Good pictures
- Featured pictures
- Quality images
- Valued images

In this category but not in...
- All images

English: Wyoming is a state in the western United States.

Subcategories:
This category has the following 36 subcategories, out of 36 total.

- A
  - Architecture of Wyoming (5 C, 2 F)
- C
  - Cities in Wyoming (20 C, 4 F)
  - Counties in Wyoming (27 C)
  - Culture of Wyoming (8 C, 1 F)
- D
- G cont.
  - Geology of Wyoming (11 C, 10 F)
  - Government of Wyoming (3 C)
  - Gravestones in Wyoming (5 C)
- H
  - History of Wyoming (29 C, 18 F)
- L
  - Landmarks in Wyoming (1 C, 2 F)
  - Luke Reiner (1 F)
- M
- N cont.
  - Nature of Wyoming (9 C, 17 F)
  - Newspapers of Wyoming (7 F)
- P
  - Parks in Wyoming (9 C, 4 F)
  - People of Wyoming (9 C, 19 F)
  - Photographs of Wyoming (11 C)
  - Populated places in Wyoming (4 C)
- S
  - Postcards of Wyoming (1 C, 5 F)
  - Pronunciation of Wyoming (2 F)
FastCCI Advanced Interface
FastCCI Advanced Interface
Work with Page IDs

- Unique across namespaces
- Just integers (small storage requirement)
  - 15 million files → 60Mb
  - 300 million category relations → 1.5GB
  - Fits into RAM
- Use the API to look up the names (client side)

<table>
<thead>
<tr>
<th>ID</th>
<th>Category</th>
<th>File Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>2729056</td>
<td>Nikola Šubić Zrinski</td>
<td></td>
</tr>
<tr>
<td>2729058</td>
<td>File: BA_roof.jpg</td>
<td></td>
</tr>
<tr>
<td>2729059</td>
<td>Category: Zrinski_family</td>
<td></td>
</tr>
<tr>
<td>2729060</td>
<td>File: BlenderExtrusion3ErrCor.jpg</td>
<td></td>
</tr>
<tr>
<td>2729061</td>
<td>Category: Nobility_of_Croatia</td>
<td></td>
</tr>
<tr>
<td>2729062</td>
<td>File: Hadik_janos.jpg</td>
<td></td>
</tr>
<tr>
<td>2729063</td>
<td>File: Ordine_di_consegna_a_d'Annunzio.JPG</td>
<td></td>
</tr>
<tr>
<td>2729064</td>
<td>File: Teheranosign.jpg</td>
<td></td>
</tr>
<tr>
<td>2729065</td>
<td>File: Perperikon_Bulgaria_1994_CIA_map.jpg</td>
<td></td>
</tr>
<tr>
<td>2729066</td>
<td>File: Emblema_SKVO.jpg</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Servers

FastCCI uses two labs instances

- Master to re-build the FastCCI database
- Worker to perform the queries
Master

- DB Format
- Two tables

  - `fastcci.tree` (category relations)

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>cend</td>
<td>Pointer to end of Subcategories</td>
</tr>
<tr>
<td>1</td>
<td>cfile</td>
<td>Pointer to end of File list</td>
</tr>
<tr>
<td>2</td>
<td>page_id</td>
<td>1st Subcategory</td>
</tr>
<tr>
<td>2</td>
<td>page_id</td>
<td>2nd Subcategory</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td>Category</td>
</tr>
<tr>
<td>cend-1</td>
<td>page_id</td>
<td>Last Subcategory</td>
</tr>
<tr>
<td>cend</td>
<td>page_id</td>
<td>1st File</td>
</tr>
<tr>
<td>cend+1</td>
<td>page_id</td>
<td>2nd File</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cfile</td>
<td>page_id</td>
<td>Last Subcategory</td>
</tr>
</tbody>
</table>

  - `fastcci.cat` (index into the tree)
Queries on the category graph

- Generating a category subgraph
  - Walk the graph breadth first
  - Maintain a visitation map
    (one byte per page_id, avoids loops, set to depth)
  - Maintain a list of categories in the subgraph
    (and store depth with each category)

- Intersect by:
  - Intersect by taking two subgraphs
Queries on the category graph

- Pregenerate a union subgraph for FP/QI/VI at server start
- Store bitmask for awards (FP/QI/VI) for each image
FastCCI Server Internals

- [https://github.com/Commonists/fastcci](https://github.com/Commonists/fastcci)
- Stream updates and results via websockets (immediate feedback: waiting, working, result)
- [http://fastccil.wmflabs.org/status](http://fastccil.wmflabs.org/status)
- [https://github.com/davidmoreno/onion](https://github.com/davidmoreno/onion)
Strength of match slider
Getting the Label Data

- **F** When a new tile comes into view the *Frame script* looks in the `sessionStorage` for cached label data.

- **F** If found it is immediately displayed.

- **TS** Otherwise an Ajax request is made to the label server script on the toolserver.

- **TS** A local database is queried for the label data.

- **TS** The label data is encoded as a JSON string.

- **F** Label data is received, cached in `sessionStorage`, and displayed.

```
{"label":{"style":"7", "lang":"en", "page":"Alexandria%2C_Virginia", "tx":8,"ty":93, "name":"Alexandria", "dx":"4829", "dy":"2198"}, "style":"0", "lang":"en", "page":"Fort_Greble", "tx":78, "ty":44, "name":"Fort Greble", "dx":"4829", "dy":"2198"}, {"style":"5", "lang":"en", "page":"Forest_Highs%2C_Maryland", "tx":117, "ty":78, "name":"Forest Heights", "dx":"4829", "dy":"2198"}, {"style":0, "lang":"en", "page":"United_States_Naval_Research_Laboratory", "tx":62, "ty":51, "name":"United States Naval Research Laboratory", "dx":"4829", "dy":"2198"}, "z":10}
```

```json
{
    "style":"7", "lang":"en", "page":"Alexandria%2C_Virginia",
    "tx":8,"ty":93, "name":"Alexandria", "dx":"4829", "dy":"2198"
}
```

http://www.schwen.de/jside/fastcci.html#_print_
Thank you!

Contact [[User: Dschwen]] with suggestions,
https://github.com/Commonists/fastcci