Automatic Taxonomy Generation for a News Group

A Case Study

Anna Divoli
Pingar Research
@annadivoli
Why Automatic Generation?
Dynamic
Fast
Cheap
Consistent
RDF / Flexible
...

Why from a Document Collection?
Focused/specific
Optimal for those documents
...
Talk Overview

The Team

The Process

News Group Case Study

Evaluation

Other Use Cases

Summary
Olena Medelyan, Steve Manion, Jeen Broekstra, Anna Divoli, Anna Lan Huang and Ian Witten

Constructing a Focused Taxonomy from a Document Collection
To appear in Proceedings of the Extended Semantic Web Conference 2013, ESWC, Montpellier, France
How Taxonomy Generation Works
**Taxonomy Generation Overview**

**Input:**
Documents stored somewhere

**Analysis:**
Using variety of tools* and datasets, extract concepts, entities, relations

**Grouping & Output:**
A taxonomy is created that groups resulting taxonomy terms hierarchically

* denotes the use of tools like pingar, Freebase, DBpedia, etc.
Taxonomy Generation in 5 Steps!

1. Import & convert to text
2. Extract concepts
3. Annotate with Linked Data
4. Disambiguate clashing concepts
5. Consolidate taxonomy

Input Docs

Document Database
Solr

Concepts & Relations Database
Sesame

Focused SKOS Taxonomy

Preferred top-level terms
Step 1. Document input & conversion

Current input:
• Directory path read recursively

Other possible inputs:
• Docs in a database or a DMS
• Emails +attachments (Exchange)
• Website URL
• RSS feed

External tool to convert different file formats to text

Database to store document content
Step 2. Extracting concepts

Documents Database → 2. Extract concepts → Concepts Database

http://localhost/solr/select?q=path:mycollection\document456.txt

Pingar API:
Taxonomy Terms:
- Climate and Weather
- Leaders
- Agreements

People:
- Yvo de Boer
- Maite Nkoana-Mashabane

Organizations:
- Associated Press
- South African Council of Churches

Locations:
- South Africa

Wikify:
Wikipedia Terms:
- South Africa
- Yvo de Boer
- U.N.
- Climate agreements
- Associated Press

Specific terminology:
green policies; climate diplomacy
Step 3. Annotation with meaning

Concepts Database ➔ 3. Annotate with Linked Data ➔ Annotations Database

Pingar API:
People:
- Yvo de Boer
- Maite Nkoana-Mashabane
Organizations:
- Associated Press
- South African Council of Churches
Locations:
- South Africa

Later this additional info will help create e-Discovery & semantic search solutions

mycollection/document456.txt
Over the past three years, Apple has acquired three mapping companies.

Two concepts were extracted, that are dissimilar. Discard the incorrect one.

For millions of years, the oceans have been filled with sounds from natural sources.

Two concepts were extracted, that are similar. Accept both correct.

Agrovoc term: Marine areas
Step 5a. Group taxonomy

- **Concepts & Relations Database**
- **5a. Add relations**
- **Focused SKOS Taxonomy**

Building the taxonomy bottom up

- **animals**
  - **horse family**
    - **felines**
    - **zebra**
    - **donkey**
    - **tiger**
    - **lizard**
    - **horse**
    - **bird**
    - **pigeon**

Category: Carnivorous animals

Category: Animals

Broader: Squamata/Reptiles/Tetrapods/Vertebrates/Chordates/Animals
Step 5b. Consolidating taxonomy

Films and film making
- Film stars
  - Mila Kunis
  - Daniel Radcliffe
  - Sally Hawkins
  - Julianna Margulies

Association football clubs
- Former Football League clubs
  - Manchester United F.C.
  - Manchester City F.C.

Finance
- Economics and finance
- Personal finance
- Commercial finance
- Tax
  - Capital gains tax
  - Tax
  - Capital gains tax

Concepts & Relations Database

5b. Prune relations

Focused SKOS Taxonomy
**Taxonomy Generation Process**

**Input:**
Documents stored somewhere

**Analysis:**
Using variety of tools* and datasets, extract concepts, entities, relations

**Output:**
A taxonomy is created that groups resulting taxonomy terms hierarchically

---

* Pingar API for People, Organization, Locations & Taxonomy Terms from related taxonomies;
  Wikification for related Wikipedia articles and category relations;
  Linked Data analysis for creating links to Freebase & DBpedia
How Does It Look Like?
Case Study: A News Group

**Fairfax NZ**
This taxonomy was created from 2000 news articles by Fairfax New Zealand around Christmas 2011.

**Taxonomy Statistics**
- Concept Count: 10158
- Edges Count: 12668
- Intermediate Count: 1383
- Leaves Count: 8748
- Labels Count: 11545

**Nesting Counts**

- 0: 27
- 1: 6102
- 2: 2903
- 3: 2891
- 4: 2057
- 5: 1202
- 6: 745
- 7: 354
- 8: 179
- 9: 41
- 10: 10

Average Depth: 2.65
Case Study: A News Group
Case Study: A News Group
Case Study: A News Group

National News (588)
- Agriculture (90)
- Catteries
- Climate and weather (18)
- Education (43)
- Environment (363)
- Flowers (12)
- Seasons (44)
- Shops (2)
- Stables
- Surveying (2)
- Warehousing
- Workshops (1)
- Zoning

Public order, justice and rights (169)
- Arrest
- Civil and human rights
- Court orders (1)
- Crime (29)
- Emergencies (26)
- Fraud (2)
- Hearings
- Judiciary (11)
- Law (71)
- Legislation (1)
- Offenders (4)
- Prisons (8)
- Prosecutions
- Remand
- Sentencing
- Witnesses
Case Study: A News Group

- Business (499)
  - Education and skills (114)
    - Further and higher education (15)
      - Lecturers (1)
    - University (11)
      - Canterbury University
      - Charles University in Prague
      - London University
      - New York University
      - Otago University
      - Oxford University
      - Pontifical Urbanian University
      - Prague University
      - Stanford University
      - Victoria University
      - Waikato University
    - Schools (94)
  - Skills and competences (2)
  - Entertainment (173)
  - Food and drink (398)
  - Government, politics and public administration (222)
  - Health, well-being and care (235)
  - Housing (5)
  - Information and communication (601)
  - International affairs and defence (79)
  - Leisure and culture (568)
  - Life in the community (419)
  - Locations (3346)
  - National News (568)

- Health, well-being and care (235)
  - Benefits
  - Care (9)
  - Diabetes (2)
  - Disability (2)
  - Explosives (1)
  - Health (166)
  - Hospitals (26)
    - Anderson's Veterinary Hospital
    - Auckland Hospital
    - Burnwood Hospital
    - Children's hospital
    - Christchurch Charity Hospital
    - Christchurch Hospital
    - Dunedin Hospital
    - Gold Coast Hospital
    - Hospital
    - Hospital departments (2)
      - Kaikoura Hospital
      - Lady Duncan Hospital
      - Matamata Chronicle Hospital
      - Middlemore Hospital
      - Pioneering Polio Hospital
      - Pittsburgh Children's Hospital
      - Princess Margaret Hospital
      - Santa Casa de Misericordia Hospital
      - Southland Hospital
      - Southland Hospital
      - Stratford Hospital
      - Taranaki Base Hospital
      - Timaru Hospital
      - Waiau Hospital
      - Wellington Hospital
      - Whakatane Hospital
Case Study: A News Group

- People (5023)
- Public order, justice and rights (169)
- Royal Christmas Message
- Science, technology and innovation (560)
  - Astronomy and space (3)
    - Comet (2)
      - C/2006 P1
      - Sungrazing comet
  - Biology and applied life sciences (44)
  - Biotechnology (1)
    - Vaccination
  - Engineering (51)
  - Environmental and earth sciences (93)
  - Mathematics (2)
    - Risk
      - Statistical analysis
  - Physics (24)
  - Research and development (2)
    - Research (1)
      - Atmospheric Research
  - Science (263)
  - Technology (67)

- Science, technology and innovation (560)
- Single Nodes (554)
- Sports and recreation (416)
  - Aeronautical activities (2)
  - Athletics (11)
  - Bowling sports and games (1)
  - Cooling down
  - Crossword (1)
  - Field sports (2)
  - Football (1)
  - Golf
  - Horse racing (12)
  - League (111)
  - Martial arts (3)
  - Motor sports (4)
  - New Zealand national rugby union team (1)
  - Outdoor pursuits (11)
  - Pet
  - Premier League (1)
  - Professional golfer (15)
  - Racing (16)
  - Rugby (7)
  - Snowman
  - Team sports (161)
  - Tennis (9)
  - Water sports (21)
  - Winter sports (2)

- Transport and infrastructure (206)
Case Study: A News Group

Food and drink (398)
- Cuisine (269)
  - Ham (1)
    - Christmas ham
  - Pudding (2)
    - Christmas pudding
- Meal (50)
  - Almond (1)
    - Christmas cake
    - Christmas dinner
- Leisure and culture (568)
  - Music (147)
    - Christmas songs (11)
      - Christmas music
      - I Wish It Could Be Christmas Everyday
      - Last Christmas
      - Snoopy's Christmas
      - Wonderful Christmas Time
  - Locations (3346)
    - Christmas Island
  - National News (568)
  - Seasons (44)
    - Christmas and holiday season

Locations (3346)
- Locations by Geography (922)
  - Europe (334)
    - Eastern Europe (4)
    - Turkey (2)
      - Armenians in Ottoman Turkey
- National News (568)
  - Agriculture (90)
    - Meat (30)
      - Beef (4)
      - Domestic turkey
  - Environment (363)
    - Plants, animals and wildlife (143)
      - Animals (56)
  - Bird (7)
    - Domestic turkey
    - Turkey (bird)
Case Study: A News Group

Fairfax - 4 Days from Sep 2001

Excerpt of the taxonomy generated from:
Fairfax articles taken from
- Sep 9th & 10th (1242 articles) and
- Sep 13th & 14th (1667 articles) NZT!

Colors of terms:
- proposed to group other terms
- found in both document collections
- in 9-10 Sep 2001 docs
- in 13-14 Sep 2001 docs
- search match

Taxonomy Statistics:
Concept Count: 12699
Edges Count: 13755
Intermediate Count: 709
Leaves Count: 11985
Labels Count: 12741
Case Study: A News Group

proposed to group other terms in both document collections in 9-10 Sep 2001 docs in 13-14 Sep 2001 docs
Case Study: A News Group

proposed to group other terms in both document collections in 9-10 Sep 2001 docs in 13-14 Sep 2001 docs
Case Study: A News Group

FairFax NZ - 4 Days from Sep 2001

Excerpt of the taxonomy generated from:
Fairfax articles taken from
- Sep 9th & 10th (1242 articles) and
- Sep 13th & 14th (1667 articles)  NZT!

Colors of terms:
- proposed to group other terms
- found in both document collections
- in 9-10 Sep 2001 docs
- in 13-14 Sep 2001 docs
- search match

Taxonomy Statistics:
Concept Count: 12699
Edges Count: 13755
Intermediate Count: 709
Leaves Count: 11985
Labels Count: 12741
Average Depth: 1.85
(0: 5 - 1: 4082 - 2: 8980 - 3: 755
4: 333 - 5: 132 - 6: 31 - 7: 6 - 8: 1)

Including NZPSV
Taxonomy Statistics:
Concept Count: 13970
Edges Count: 15020
Intermediate Count: 1277
Leaves Count: 12677
Labels Count: 15407
Average Depth: 3
(0: 16 - 1: 10153 - 2: 1888 - 3: 1400
4: 1203 - 5: 1053 - 6: 756 - 7: 427
16: 10)
Case Study: A News Group

September 2001

- Science, technology and innovation (11)
- Environmental and earth sciences (5)
- Geography (10)
- Continents (3)
- Europe (1)
- Eastern Europe (2)
- Island (23)
- Christmas Island

proposed to group other terms in both document collections
in 9-10 Sep 2001 docs
in 13-14 Sep 2001 docs

Christmas 2011

- Life in the community (15)
- Abuse (13)
- Terrorism (5)
- Public order, justice and rights (16)
- Crime (25)
- Organised crime and terrorism
- Single Nodes (109)
- Islamic politics and Islamic world studies (1)
- Islamic fundamentalism (6)
- Al-Qaeda (1)
- War on Terror (4)
Evaluation

Recall: 75%
(comparing with manually generated taxonomy for the same domain)

Precision:
89% for concepts
90% for relations
(15 human judges based evaluation)

<table>
<thead>
<tr>
<th>Type</th>
<th>Number</th>
<th>Errors</th>
<th>Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>People</td>
<td>1145</td>
<td>37</td>
<td>3.2%</td>
</tr>
<tr>
<td>Organizations</td>
<td>496</td>
<td>51</td>
<td>10.3%</td>
</tr>
<tr>
<td>Locations</td>
<td>988</td>
<td>114</td>
<td>11.5%</td>
</tr>
<tr>
<td>Wikipedia named entities</td>
<td>832</td>
<td>71</td>
<td>8.5%</td>
</tr>
<tr>
<td>Wikipedia other entities</td>
<td>99</td>
<td>16</td>
<td>16.4%</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>868</td>
<td>229</td>
<td>26.4%</td>
</tr>
<tr>
<td>DBPedia</td>
<td>868</td>
<td>81</td>
<td>8.1%</td>
</tr>
<tr>
<td>Freebase</td>
<td>135</td>
<td>12</td>
<td>8.9%</td>
</tr>
<tr>
<td>Overall</td>
<td>3447</td>
<td>393</td>
<td>11.4%</td>
</tr>
</tbody>
</table>
Other Use Cases

What’s in these files / emails?

How to refine search by metadata?

What to include into our corporate taxonomy?

How to find all docs on a given topic?

Content Audit

Better search with facets

Information Architecture

Better browsing
Other Use Cases: Discovery

proposed to group other concepts in two or more document collections in the bipolar document collection in the breast cancer document collection in the neither cancer or bipolar doc. collection
Summary

Entity Extraction

Linked Data

Disambiguation

Consolidation

News Group Case Study

Other Use Cases
Focused SKOS Taxonomy Extraction Process (F-STEP) wiki

bit.ly/f-step

pingar.com

@PingarHQ

anna.divoli@pingar.com

@annadivoli