Automatic Construction of a semantic graph

EDAN70 Project in Computer Science

O. Chabrol D. Norrestam

May,23rd 2016

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

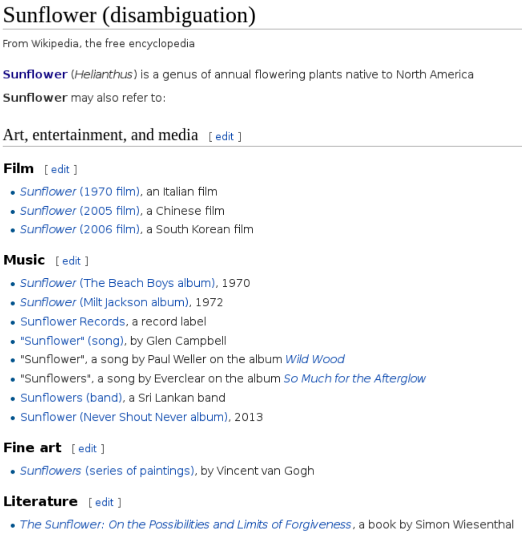
Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Entity Disambiguation

- Basic problem : different entities can have the same name



Sunflower (disambiguation)

From Wikipedia, the free encyclopedia

Sunflower (*Helianthus*) is a genus of annual flowering plants native to North America

Sunflower may also refer to:

Art, entertainment, and media [edit]

Film [edit]

- *Sunflower* (1970 film), an Italian film
- *Sunflower* (2005 film), a Chinese film
- *Sunflower* (2006 film), a South Korean film

Music [edit]

- *Sunflower* (The Beach Boys album), 1970
- *Sunflower* (Milt Jackson album), 1972
- Sunflower Records, a record label
- "Sunflower" (song), by Glen Campbell
- "Sunflower", a song by Paul Weller on the album *Wild Wood*
- "Sunflowers", a song by Everclear on the album *So Much for the Afterglow*
- Sunflowers (band), a Sri Lankan band
- Sunflower (Never Shout Never album), 2013

Fine art [edit]

- *Sunflowers* (series of paintings), by Vincent van Gogh

Literature [edit]

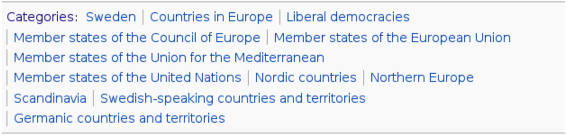
- *The Sunflower: On the Possibilities and Limits of Forgiveness*, a book by Simon Wiesenthal

Outline

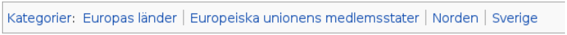
- Project background
  - Entity Disambiguation
  - The Sunflower Algorithm
  - Q-numbers
- Implementation
  - Overview
  - Extract information
  - Parsing the dumps
  - Merging data
  - Depth
- Demonstration

Goal of the algorithm

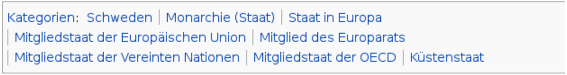
- Find concepts linked to every entity
- Using categories linked to article in database



Categories: Sweden | Countries in Europe | Liberal democracies | Member states of the Council of Europe | Member states of the European Union | Member states of the Union for the Mediterranean | Member states of the United Nations | Nordic countries | Northern Europe | Scandinavia | Swedish-speaking countries and territories | Germanic countries and territories

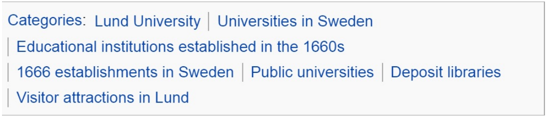


Kategorier: Europas länder | Europeiska unionens medlemsstater | Norden | Sverige



Kategorien: Schweden | Monarchie (Staat) | Staat in Europa | Mitgliedstaat der Europäischen Union | Mitglied des Europarats | Mitgliedstaat der Vereinten Nationen | Mitgliedstaat der OECD | Küstenstaat

- Find the "best" categories



Categories: Lund University | Universities in Sweden | Educational institutions established in the 1660s | 1666 establishments in Sweden | Public universities | Deposit libraries | Visitor attractions in Lund

- We take the most recurrent categories as best categories

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

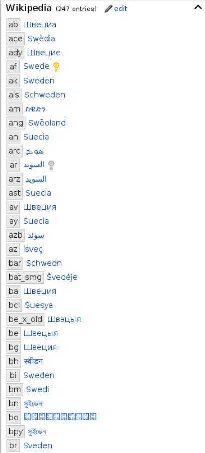
Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Q-numbers

- Allows the link between languages

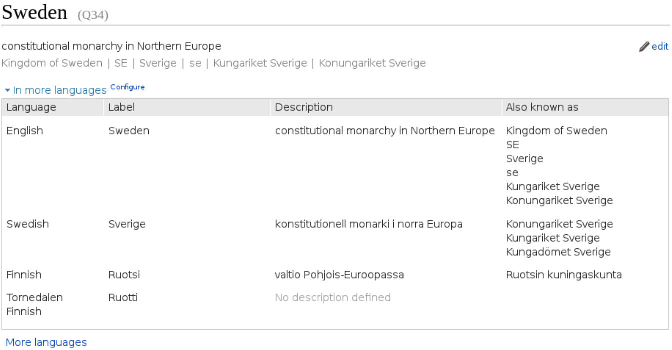


Wikipedia (247 entries) edit

ab	Швециа
ace	Swèdia
ady	Швецие
af	Swede
ak	Sweden
als	Schweden
am	ስዊድን
ang	Swēoland
an	Suecia
arc	ܣܘܝܕ
ar	السويد
arz	السويد
ast	Suecia
av	Швеция
ay	Suecia
azb	سوئد
az	İsveç
bar	Schwedn
bat_smg	Švedėjė
ba	Швеция
bcl	Suesya
be_x_old	Швэцыя
be	Швецыя
bg	Швеция
bh	स्वीडन
bi	Sweden
bm	Swedi
bn	সুইডেন
bo	[illegible]
bpy	সুইডেন
br	Sveden

Q-numbers

- Unique identifier for entities
- Every entity (articles and categories) has one universal Q-number



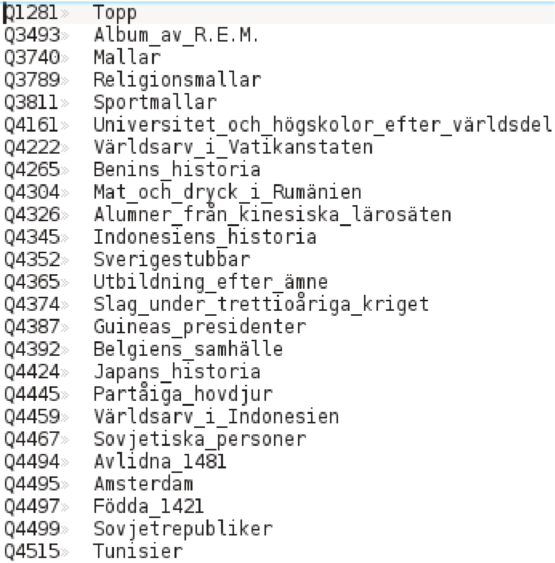
Sweden (Q34)

constitutional monarchy in Northern Europe

Kingdom of Sweden | SE | Sverige | se | Kungariket Sverige | Konungariket Sverige

Language	Label	Description	Also known as
English	Sweden	constitutional monarchy in Northern Europe	Kingdom of Sweden SE Sverige se Kungariket Sverige Konungariket Sverige
Swedish	Sverige	konstitutionell monarki i norra Europa	Konungariket Sverige Kungariket Sverige Kungadömet Sverige
Finnish	Ruotsi	valtio Pohjois-Euroopassa	Ruotsin kuningaskunta
Tornedalen Finnish	Ruotti	No description defined	

Q-numbers examples



```
Q1281	Topp
Q3493	Album_av_R.E.M.
Q3740	Mallar
Q3789	Religionsmallar
Q3811	Sportmallar
Q4161	Universitet_och_högskolor_efter_världsdel
Q4222	Världsarv_i_Vatikanstaten
Q4265	Benins_historia
Q4304	Mat_och_dryck_i_Rumänien
Q4326	Alumner_från_kinesiska_lärosäten
Q4345	Indonesiens_historia
Q4352	Sverigestubbar
Q4365	Utbildning_efter_ämne
Q4374	Slag_under_trettioåriga_kriget
Q4387	Guineas_presidenter
Q4392	Belgiens_samhälle
Q4424	Japans_historia
Q4445	Partåiga_hovdjur
Q4459	Världsarv_i_Indonesien
Q4467	Sovjetiska_personer
Q4494	Avlidna_1481
Q4495	Amsterdam
Q4497	Födda_1421
Q4499	Sovjetrepubliker
Q4515	Tunisier
```

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Overview

The algorithm consists of the following steps:

1. Extract information from Wikipedia
2. For each language
    2.1 Parse information
    2.2 Create data structures
3. Merge languages and create semantic graph

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Extract information

- Have two options
  1. Parse Wikipedia
  2. Use existing dumps
- Chose existing dumps
  - DBpedia (2015-10)

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

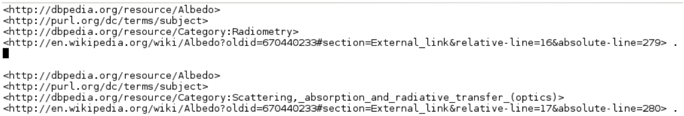
Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Parsing the dumps

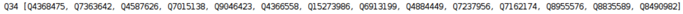
- for every language we have :



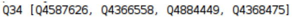
```
<http://dbpedia.org/resource/Albedo>
<http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Radiometry>
<http://en.wikipedia.org/wiki/Albedo?oldid=670440233#section=External_link&relative-line=16&absolute-line=279> .

<http://dbpedia.org/resource/Albedo>
<http://purl.org/dc/terms/subject>
<http://dbpedia.org/resource/Category:Scattering,_absorption_and_radiative_transfer_(optics)>
<http://en.wikipedia.org/wiki/Albedo?oldid=670440233#section=External_link&relative-line=17&absolute-line=280> .
```

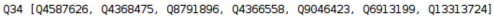
- and we create :



```
Q34 [Q4368475, Q7363642, Q4587626, Q7015138, Q9046423, Q4366558, Q15273986, Q6913199, Q4884449, Q7237956, Q7162174, Q8955576, Q8835589, Q8490982]
```



```
Q34 [Q4587626, Q4366558, Q4884449, Q4368475]
```



```
Q34 [Q4587626, Q4368475, Q8791896, Q4366558, Q9046423, Q6913199, Q13313724]
```

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

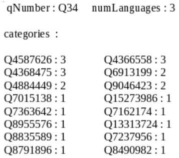
Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Merging data

- after merging the three languages:

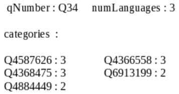


```
qNumber : Q34    numLanguages : 3

categories  :

Q4587626 : 3        Q4366558 : 3
Q4368475 : 3        Q6913199 : 2
Q4884449 : 2        Q9046423 : 2
Q7015138 : 1        Q15273986 : 1
Q7363642 : 1        Q7162174 : 1
Q8955576 : 1        Q13313724 : 1
Q8835589 : 1        Q7237956 : 1
Q8791896 : 1        Q8490982 : 1
```

- we can select the five "best" categories

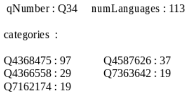


```
qNumber : Q34    numLanguages : 3

categories  :

Q4587626 : 3        Q4366558 : 3
Q4368475 : 3        Q6913199 : 2
Q4884449 : 2
```

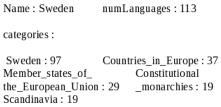
- with 123 languages



```
qNumber : Q34    numLanguages : 113

categories  :

Q4368475 : 97       Q4587626 : 37
Q4366558 : 29       Q7363642 : 19
Q7162174 : 19
```



```
Name : Sweden        numLanguages : 113

categories :

 Sweden : 97            Countries_in_Europe : 37
Member_states_of_              Constitutional
the_European_Union : 29        _monarchies : 19
Scandinavia : 19
```

Outline

Project background

- Entity Disambiguation
- The Sunflower Algorithm
- Q-numbers

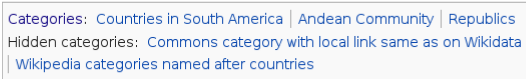
Implementation

- Overview
- Extract information
- Parsing the dumps
- Merging data
- Depth

Demonstration

Depth

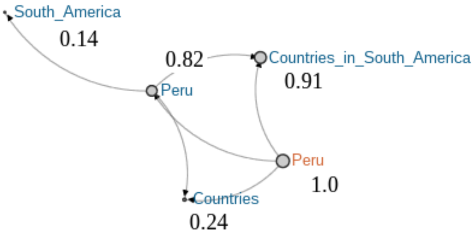
- we already have a width concept
- categories are also part of more general categories



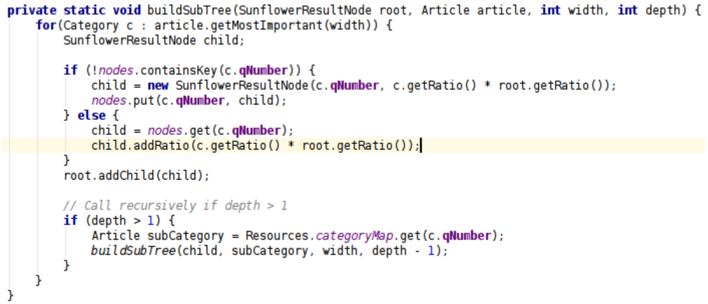
Categories: Countries in South America | Andean Community | Republics

Hidden categories: Commons category with local link same as on Wikidata | Wikipedia categories named after countries

- which leads to depth



Computation of ratios



```
private static void buildSubTree(SunflowerResultNode root, Article article, int width, int depth) {
    for(Category c : article.getMostImportant(width)) {
        SunflowerResultNode child;

        if (!nodes.containsKey(c.qNumber)) {
            child = new SunflowerResultNode(c.qNumber, c.getRatio() * root.getRatio());
            nodes.put(c.qNumber, child);
        } else {
            child = nodes.get(c.qNumber);
            child.addRatio(c.getRatio() * root.getRatio());
        }
        root.addChild(child);

        // Call recursively if depth > 1
        if (depth > 1) {
            Article subCategory = Resources.categoryMap.get(c.qNumber);
            buildSubTree(child, subCategory, width, depth - 1);
        }
    }
}
```

Demonstration

Interface example

End

End of the presentation

Thank you ! Any questions ?