Domain-specific modeling: Towards a Food and Drink Gazetteer

Authors: Andrey Tagarev, Laura Tolosi, and Vladimir Alexiev
Presenter: Andrey Tagarev

ontotext

Overview

1. Motivation
2. The Goal
3. Development
4. Results

Europeana Foundation

Europeana: think culture initiative by the Europeana Foundation collects cultural heritage objects:

- From all European countries
- From many sources: museum, galleries, archives and museums
- In many media: images, text, sounds, video
- On many different topics

Food and Drink Project

The Europeana Food and Drink (EFD) project is aimed at cultural heritage objects in the domain of food and drink.

Contributors participate in these tracks:

- Content track: collect 50-70k high quality digital assets and associated metadata about FD
- Public Engagement Track: engage public in the collection and use of the data
- Creative Applications Track: develop innovative products with data

Food and Drink Project

Our application is aimed at categorizing food and drink (FD) related concepts in order to facilitate search and semantically enrich Europeana cultural heritage objects (CHOs).

It can be used both on the heritage items collected for the Europeana Food and Drink project, and the larger body (over 40 million) of previously aggregated CHOs (metadata).

The Challenge

Semantic enrichment of a huge quantity of diverse data to allow searching and sorting by non-expert users.

The Tool

Ontotext automatic concept extraction tool.

Capable of:

- General concept extraction (based on DBpedia and WikiData)
- Named Entity Recognition and Linking
- On-the-fly Relationship extraction between Entities
- Entity Disambiguation

The Goal

Build a Food and Drink gazetteer to serve in classification of general FD-related concepts to be used in automated semantic enrichment and efficient faceted search.

The gazetteer is to be built with a minimal amount of manual work.

The Goal (2)

Desirable features of the solution:

- A generalized approach that can be applied to other topics of interest.
- A scalable approach that can be applied to other topics with minimal additional work.
- An encyclopedic approach that can be applied to topics which cannot be strictly or exhaustively defined (e.g. Sports, Arts, Food and Drink, History).

Wikipedia

We selected Wikipedia as the base knowledge set from which we extract our gazetteer for a number of reasons:

- A diverse collection of general knowledge
- A large number of existing concepts (~35 million articles)
- A strong multilingual element (articles in over 240 languages)
- A hierarchical organization of articles.

Wikipedia Stats (2014-12)

Lang	Articles	Cats	Art->Cat	Cat per art	Cat->Cat	Cat per cat
English	4,774,396	1,122,598	18,731,750	3.92	2,268,299	2.02
Dutch	1,804,691	89,906	2,629,632	1.46	186,400	2.07
French	1,579,555	278,713	4,625,524	2.93	465,931	1.67
Italian	1,164,000	258,210	1,597,716	1.37	486,786	1.89
Spanish	1,148,856	396,214	4,145,977	3.61	675,380	1.7
Polish	1,082,000	2,217,382	20,149,374	18.62	4,361,474	1.97
Bulgarian	170,174	37,139	387,023	2.27	73,228	1.97
Greek	102,077	17,616	182,023	1.78	35,761	2.03

Wikipedia Statistics Per Language.
Wide variation in number of cats and cats per art (density of categorization)

ontotext

The Algorithm

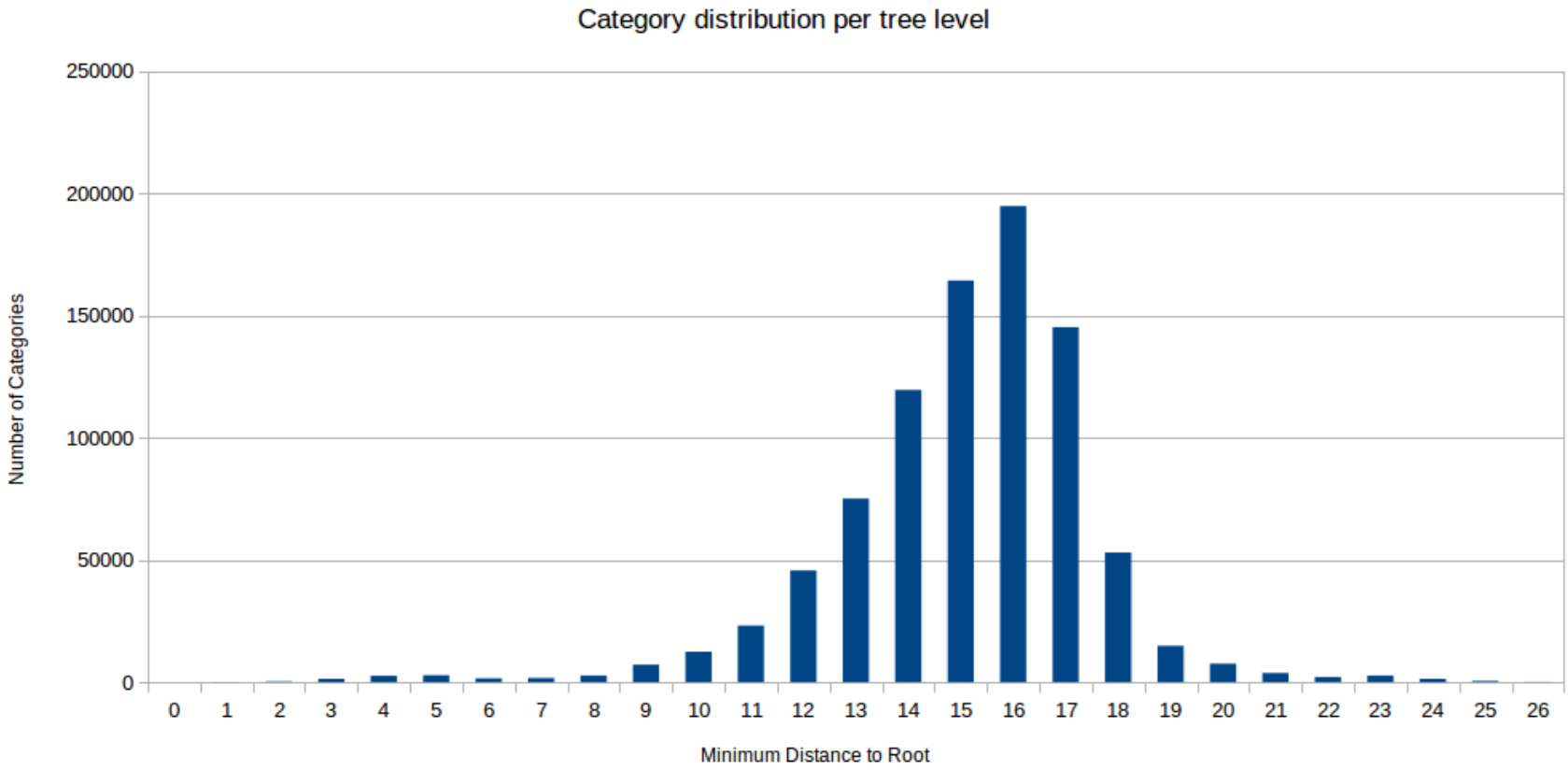
1) Select the maximally general Wikipedia category that best describes the domain (dbc:Food_and_drink) as the *root*.
2) Starting at the *root*, build a tree by following *skos:broader*$^{-1}$ connections to subcategories and removing cycles.
3) Perform manual curation by an expert to prune incorrect paths from the tree.
4) Bottom up enrichment by enlarging the tree using articles that are “certainly” domain-relevant (eg class dbo:Food)

Initially Constructed Tree

The initially constructed tree before manual annotator work contained:

- 26 levels
- 887523 categories (80% of all categories in the English Wikipedia)
- Essentially **useless**

Initially Constructed Tree



Category distribution by level in initially constructed tree: median 15 levels

Superfluous Categories

Examples of irrelevant categories in tree:

- Due to wrong hierarchy.

  Food and drink → Food politics → Water and politics →Water and the environment → Water management → Water treatment →Euthenics → Personal life → Leisure → Sports → Sports by type → Team sports→ Football.

- Due to partial inclusion.

  The subcategory *Animal_products* has some children relevant to FD (*Animal-based seafood, Dairy products, Eggs (food), Fish products, Meat)* and some that are not *(Animal dyes, Animal hair products, Animal waste products, Bird products, Bone products, Coral islands, Coral reefs, Hides)*.

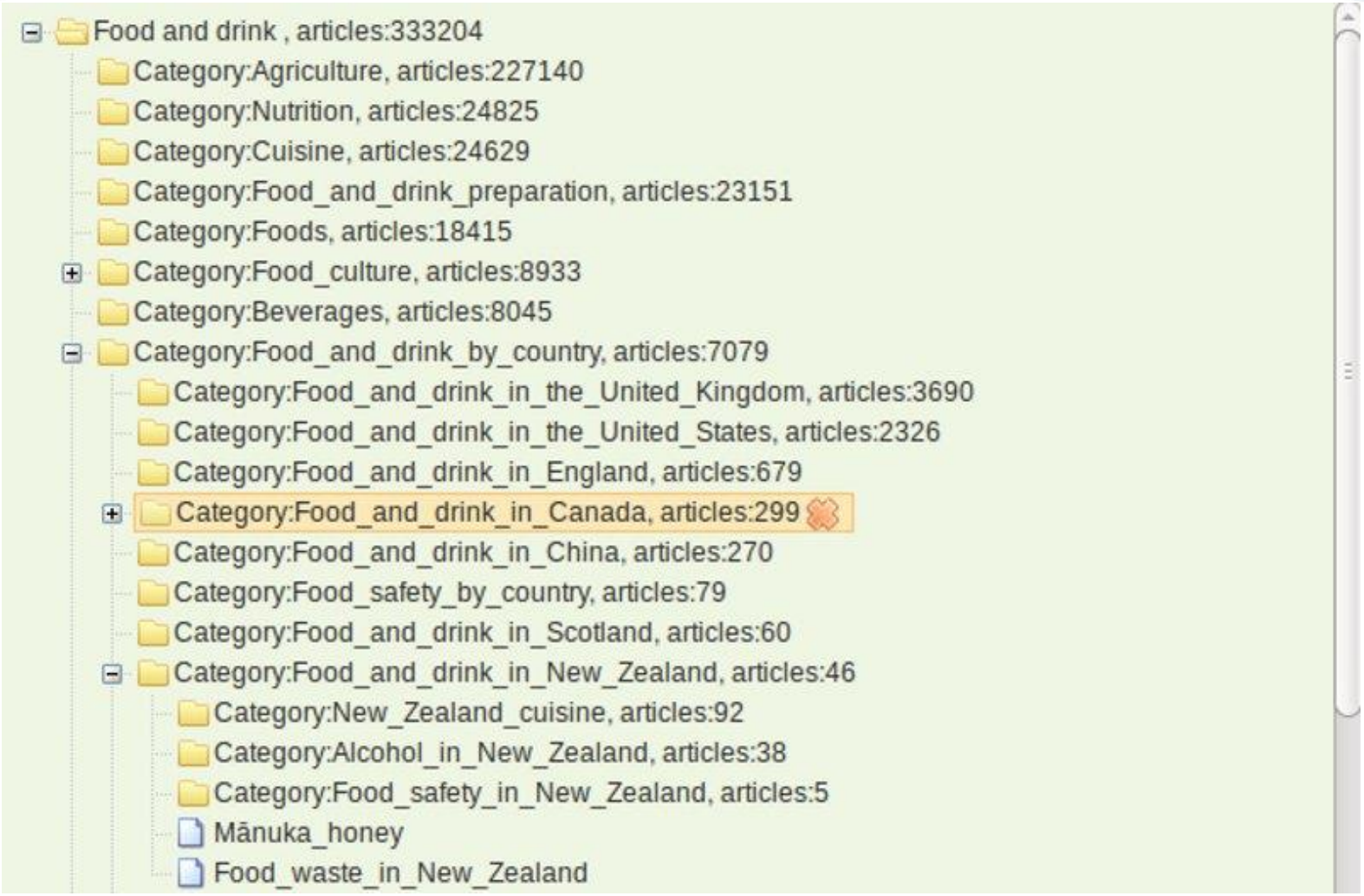
- Due to non-human food and eating.

  The subcategory *Eating behaviors* has some appropriate children, e.g. *Diets, Eating disorders*, but has also some inappropriate children, e.g. *Carnivory, Detritivores*.

- Due to semantic drift

  The farther away from the root, the vaguer is the relevance

Manual Pruning



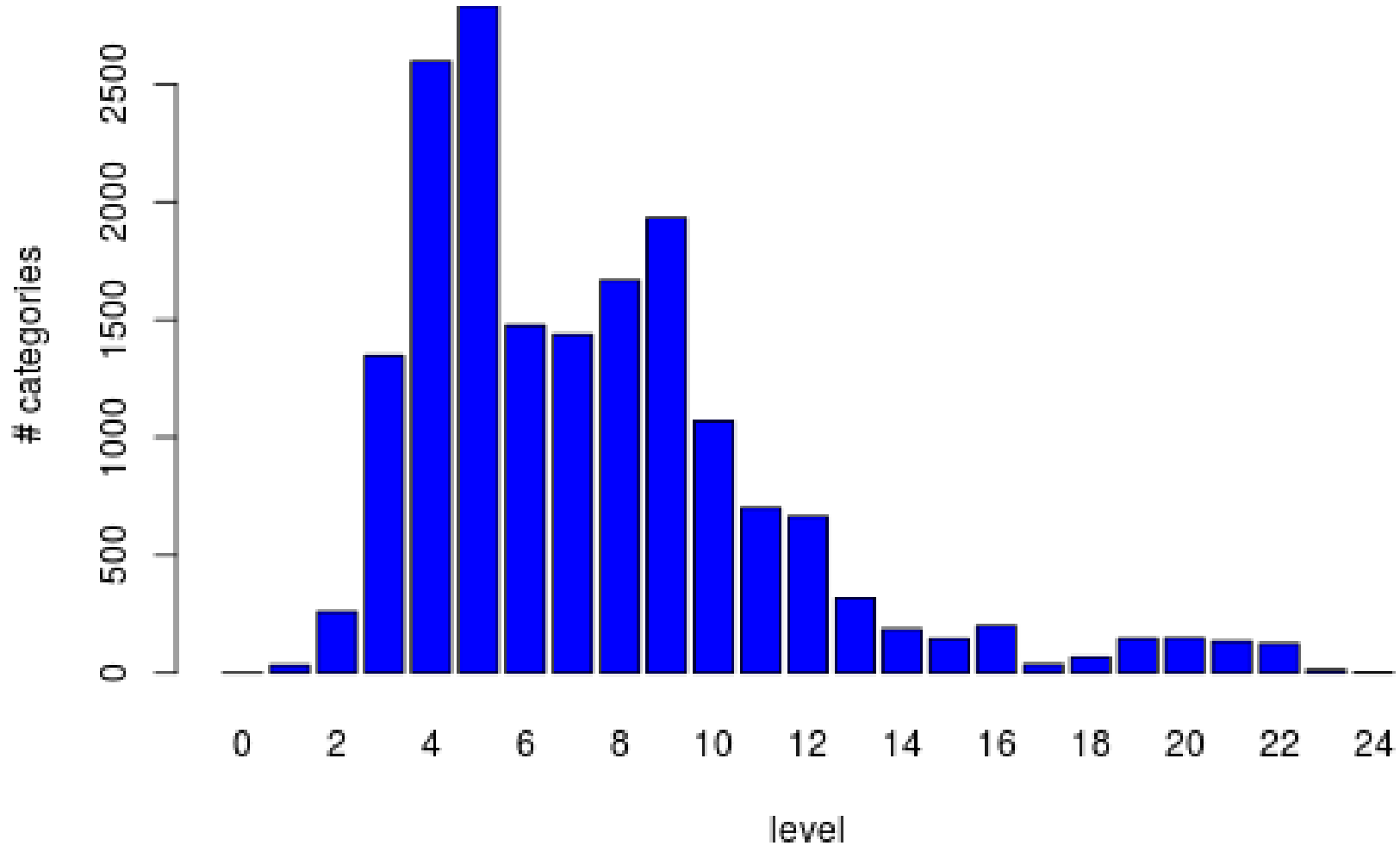
User Interface For Top Down Pruning By Experts

ontotext

Effects of Pruning

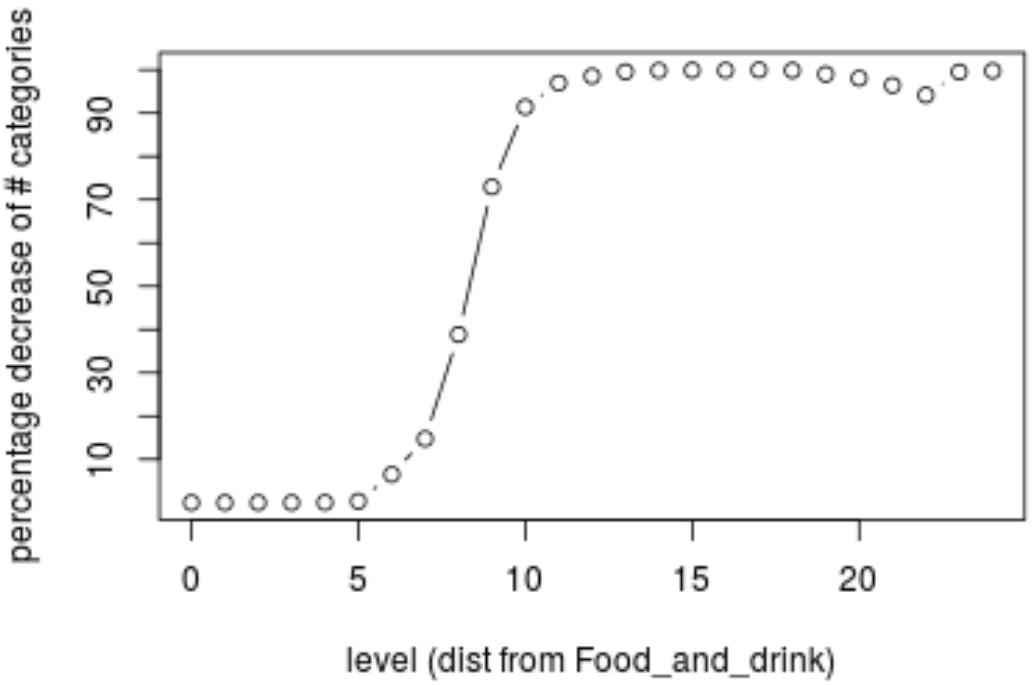
- Select 250 “top” categories by heuristic
- Mark 239 as irrelevant to the topic
- Initial tree size: 887523 unique categories
- New tree size: 17542 unique categories
- Effects: 50-fold decrease in tree size
- Reduce median levels from 16 to 6

Pruned Tree



Tree after pruning 239 of the top 250 categories: median 6 levels

Pruned Tree



Percentage of categories removed per level after pruning

ontotext

Evidence and Scoring

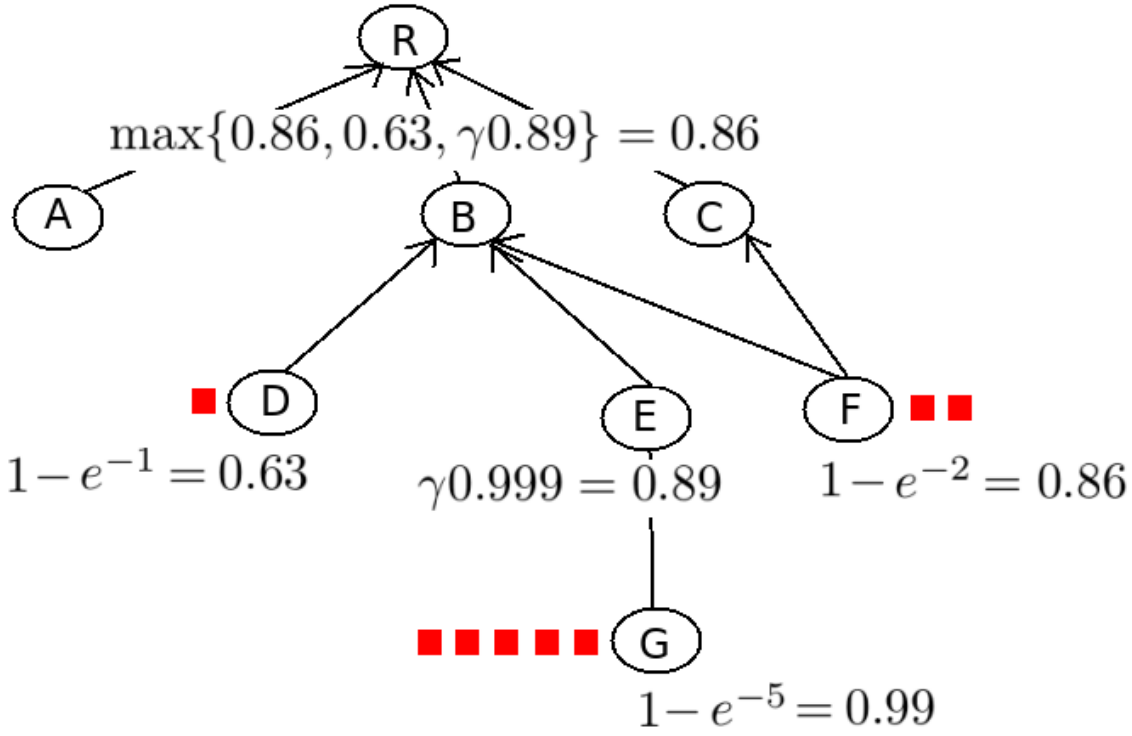
- Automatic tree testing and refinement
- Bottom-up approach
- Driven by enrichment data
- Complementary to top-down expert working with the drill-down UI

Evidence and Scoring

The first approach is based on the use of a decay factor to propagate a diminishing category relevance to parent categories.

$$score(Y) = \max\{1 - e^{-t}, \max_{i=1}^{n}\{\gamma score(Y_i)\}\}$$

Evidence and Scoring



Example of first approach to scoring

Evidence and Scoring

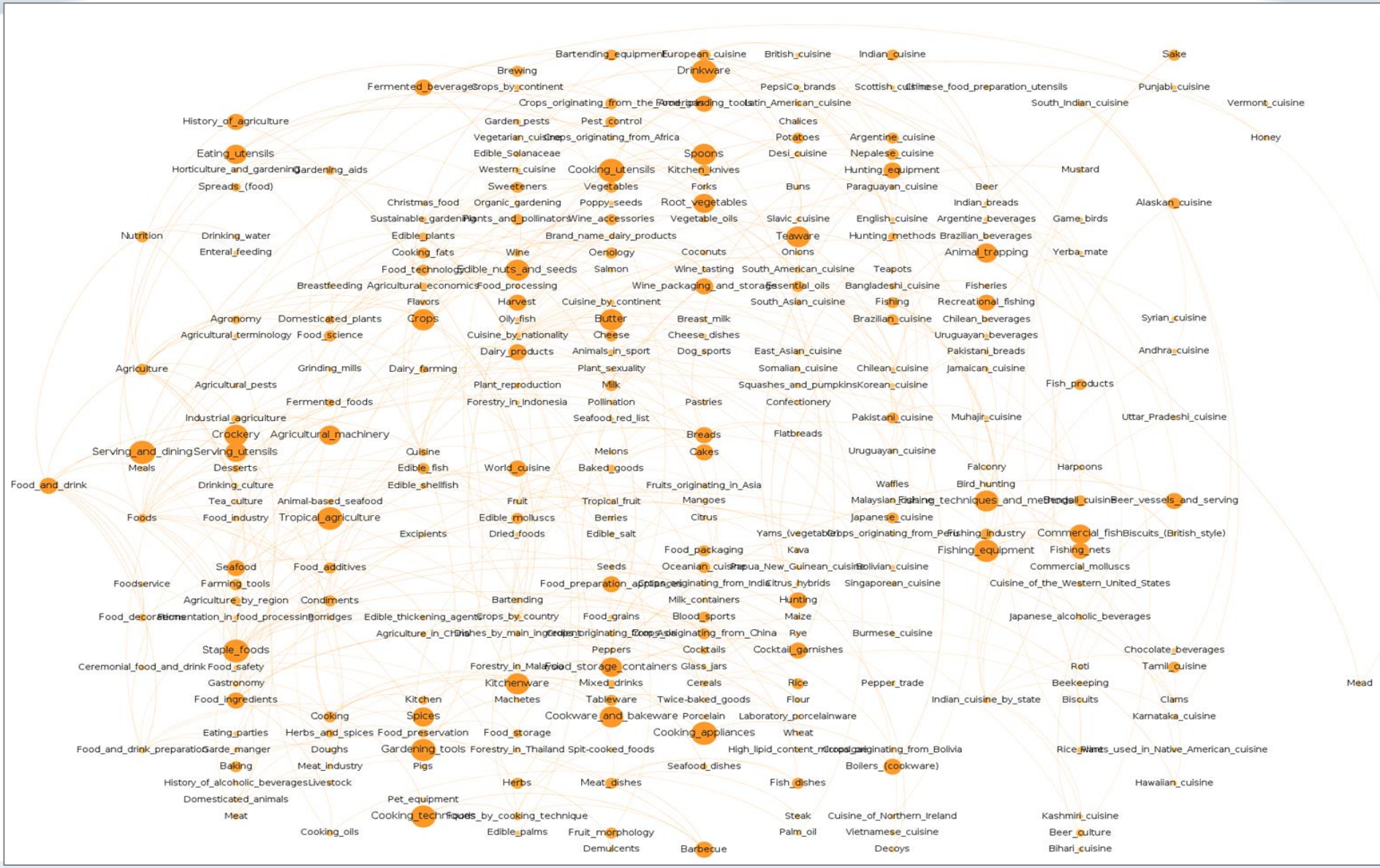
The second approach is based on an additive propagation of evidence scores.

Given child category A with a piece of evidence and its parent category B:

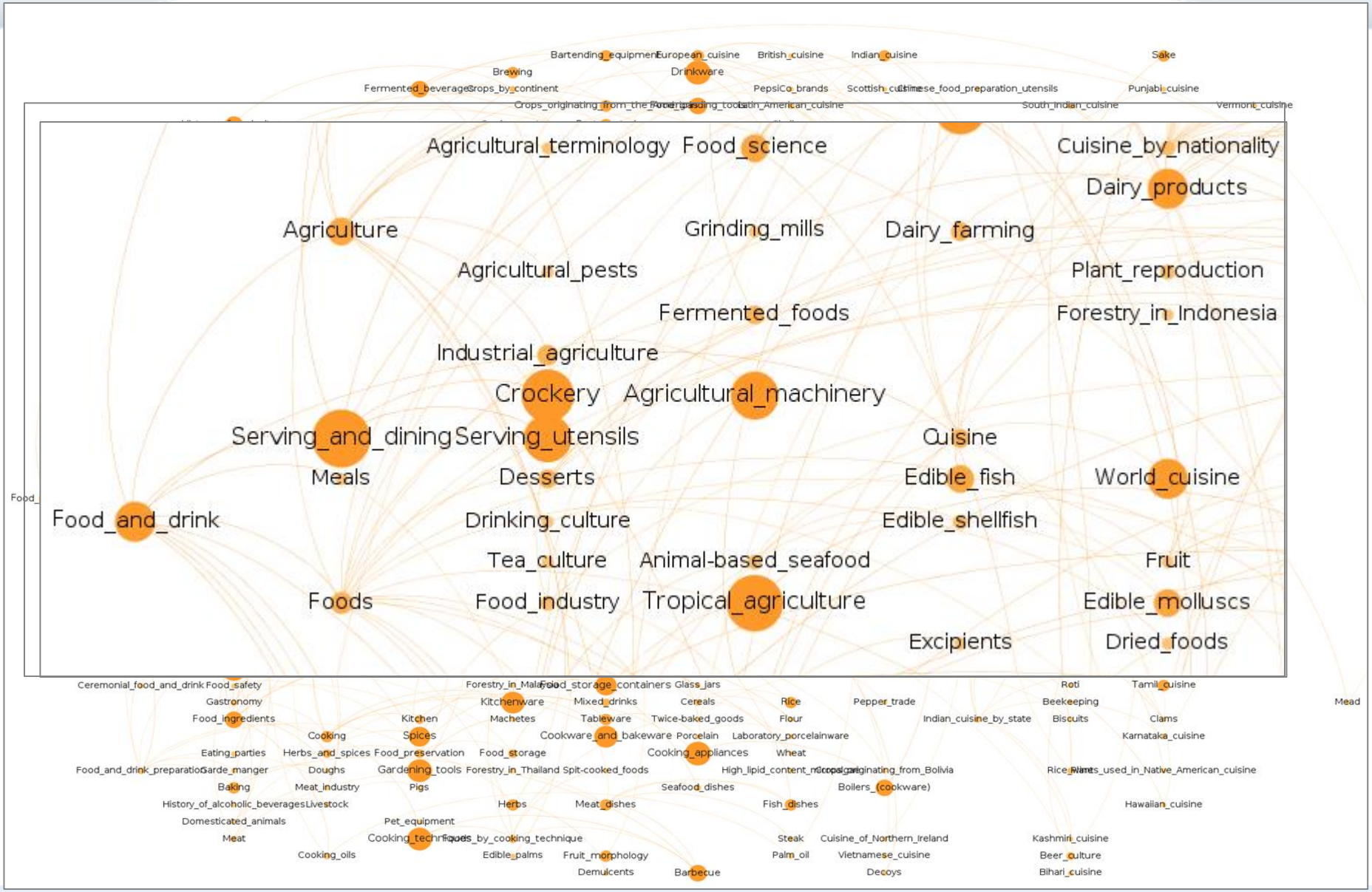
- If *level(A) < level(B)*, increase score of B by one and propagate evidence.
- If *level(A) = level(B)*, propagate evidence.
- If *level (A) < level(B)*, do nothing.
  (How can child have smaller level? It's a poly-hierarchy)

ontotext

Evidence Propagated



Evidence Propagated



Result: A Tasteful Tagger

europeana food and drink

http://foodanddrinkeurope.eu

Europeana Food and Drink

- Enrichment of cultural objects
- ...related to Food and Drink
- ...also Place enrichment
- ...upcoming: Cultures

Eg. CHO from Horniman M

Description: Beer horn made from a cow's horn. Made by elders.
Collector: Rose, Cordelia
Culture: Samburu
Maker: elder
Theme: Food and Feasting
Classification: horn (narcotics & intoxicants: drinking). drinking containers (food service). Horn material).
Place: Lariak Orok, near Kisima, Kenya, Africa.

Result: A Tasteful Tagger

europeana food and drink
http://foodanddrinkeurope.eu

Europeana Food and Drink

- Enrichment of cultural objects
- ...related to Food and Drink
- ...also Place enrichment
- ...upcoming: Cultures

Eg. CHO from Horniman M

Description: **Beer horn** made from a **cow**'s horn. Made by elders.
Collector: Rose, Cordelia
Culture: **Samburu**
Maker: elder
Theme: **Food** and **Feasting**
Classification: **horn** (narcotics & intoxicants: **drinking**). **drinking** containers (**food** service). Horn material).
Place: **Lariak Orok**, near **Kisima**, **Kenya**, **Africa**.

Result: A Tasteful Tagger

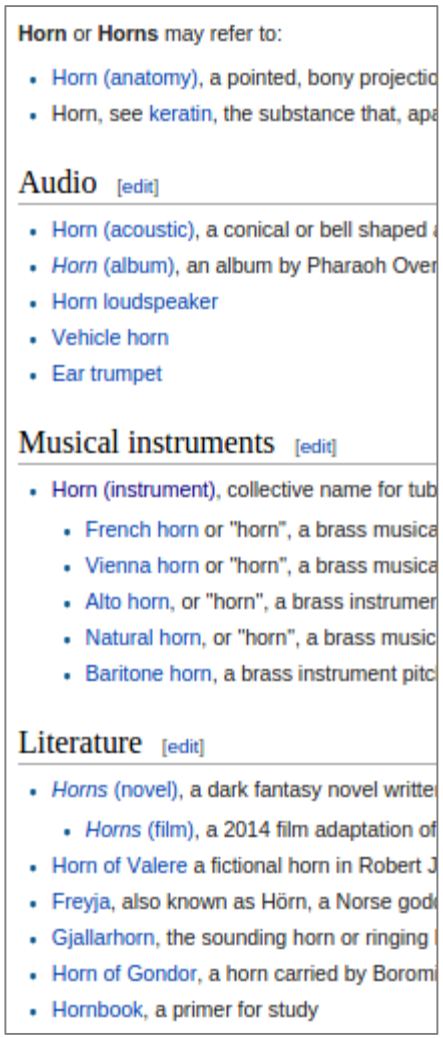
Description: **Beer horn** made from a **cow**'s horn. Made by elders.
Collector: Rose, Cordelia
Culture: **Samburu**
Maker: elder
Theme: **Food** and **Feasting**
Classification: **horn** (narcotics & intoxicants: **drinking**).
drinking containers (**food** service). Horn material).
Place: **Lariak Orok**, near **Kisima**, **Kenya**, **Africa**.

Result: A Tasteful Tagger

Description: **Beer horn** made from a **cow**'s horn. Made by elders.
Collector: Rose, Cordelia
Culture: **Samburu**
Maker: elder
Theme: **Food** and **Feasting**
Classification: **horn** (narcotics & intoxicants: **drinking**).
drinking containers (**food** service). Horn material).
Place: **Lariak Orok**, near **Kisima**, **Kenya**, **Africa**.

candidates

https://en.wikipedia.org/wiki/Horn is a disambiguation page:



Horn or **Horns** may refer to:

- Horn (anatomy), a pointed, bony projectio
- Horn, see keratin, the substance that, apa

Audio [edit]

- Horn (acoustic), a conical or bell shaped
- *Horn* (album), an album by Pharaoh Over
- Horn loudspeaker
- Vehicle horn
- Ear trumpet

Musical instruments [edit]

- Horn (instrument), collective name for tub
  - French horn or "horn", a brass musica
  - Vienna horn or "horn", a brass musica
  - Alto horn, or "horn", a brass instrumer
  - Natural horn, or "horn", a brass music
  - Baritone horn, a brass instrument pitc

Literature [edit]

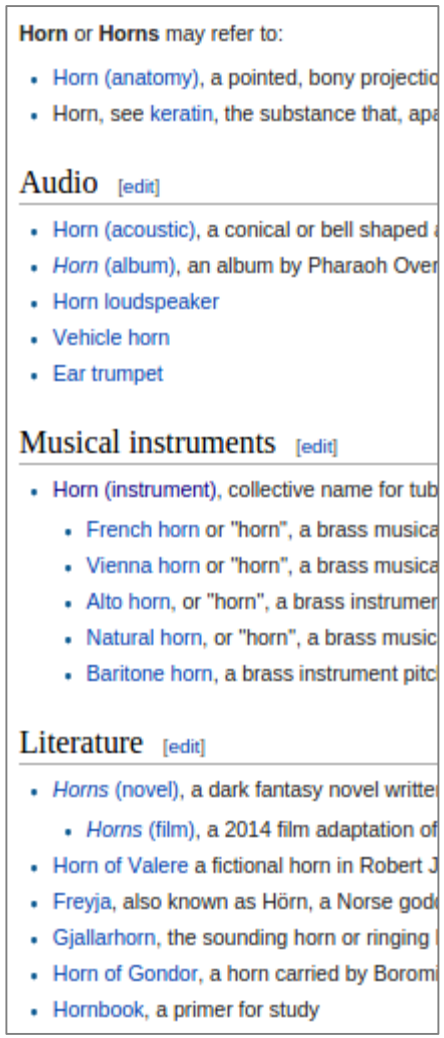
- *Horns* (novel), a dark fantasy novel writte
  - *Horns* (film), a 2014 film adaptation of
- Horn of Valere a fictional horn in Robert J
- Freyja, also known as Hörn, a Norse god
- Gjallarhorn, the sounding horn or ringing
- Horn of Gondor, a horn carried by Boromi
- Hornbook, a primer for study

Result: A Tasteful Tagger

Description: **Beer horn** made from a **cow**'s horn. Made by elders.
Collector: Rose, Cordelia
Culture: **Samburu**
Maker: elder
Theme: **Food** and **Feasting**
Classification: **horn** (narcotics & intoxicants: **drinking**).
drinking containers (**food** service). Horn material).
Place: **Lariak Orok**, near **Kisima**, **Kenya**, **Africa**.

candidates

https://en.wikipedia.org/wiki/Horn is a disambiguation page:



Horn or **Horns** may refer to:

- Horn (anatomy), a pointed, bony projectio
- Horn, see keratin, the substance that, apa

Audio [edit]

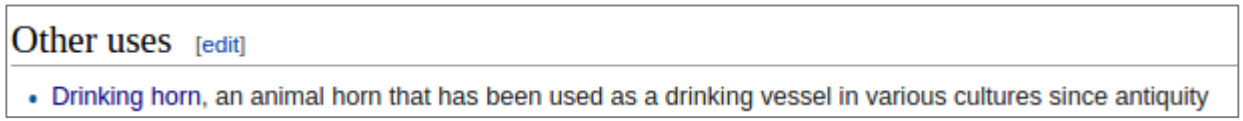
- Horn (acoustic), a conical or bell shaped
- *Horn* (album), an album by Pharaoh Over
- Horn loudspeaker
- Vehicle horn
- Ear trumpet

Musical instruments [edit]

- Horn (instrument), collective name for tub
  - French horn or "horn", a brass musica
  - Vienna horn or "horn", a brass musica
  - Alto horn, or "horn", a brass instrumer
  - Natural horn, or "horn", a brass music
  - Baritone horn, a brass instrument pitc

Literature [edit]

- *Horns* (novel), a dark fantasy novel writte
  - *Horns* (film), a 2014 film adaptation of
- Horn of Valere a fictional horn in Robert J
- Freyja, also known as Hörn, a Norse god
- Gjallarhorn, the sounding horn or ringing
- Horn of Gondor, a horn carried by Boromi
- Hornbook, a primer for study



Other uses [edit]

- Drinking horn, an animal horn that has been used as a drinking vessel in various cultures since antiquity

After scrolling over 40 meanings, the correct match appears

ontotext

References

- Vladimir Alexiev. Europeana Food and Drink Classification Scheme. Deliverable D2.2, Europeana Food and Drink project, February 2015. http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Classification-Scheme-(D2.2).pdf
- Vladimir Alexiev. Europeana Food and Drink Semantic Demonstrator Specification. Deliverable D3.19, Europeana Food and Drink project, March 2015. http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-Specification-(D3.19).pdf
- Vladimir Alexiev. Europeana Food and Drink Semantic Demonstrator M18 Progress Report. Progress Report D3.20a, Europeana Food and Drink project, June 2015. http://vladimiralexiev.github.io/pubs/Europeana-Food-and-Drink-Semantic-Demonstrator-M18-Report-(D3.20a).pdf

Conclusion

Questions?