

# Partitioning instead of Clustering:

An alternative approach for mining structured content in folksonomies

Student: Waltl Bernhard

Advisor: Steinhoff Alexander

Supervisor: Prof. Matthes Florian

# Agenda

- Introduction
- Motivation
  
- Partitioning Algorithm
  - Idea
  - Implementation and Cost Functions
- Results
  
- Conclusion

# Tagging Systems

- Tag: unstructured keyword
- Folksonomies: Folk & Taxonomy



# Tagging Systems

art beach blue bw california canada canon china christmas city concert de  
england europe family festival film flower flowers food france friends green  
instagramapp iphoneography italy japan live london music  
nature new newyork night nikon nyc paris park party people  
photography portrait red sky snow square squareformat street summer  
sunset travel trip uk usa vacation water wedding white winter

[www.flickr.com](http://www.flickr.com)

- Freely chosen (no dictionary)
- No clear structure  
→ can be chaotic!

# Browsing & Navigating

- To improve the access and navigation additional aid required
- Two common techniques
  - Clustering
    - Grouping tags with high Co-Occurrence
    - Determining similarity
  - Hierarchical Relationships
    - Assuming hierarchical relationships between tags
    - Subsumption

# Clustering

Explore / Tags / rain / clusters



storm, clouds, sky, weather, cloud, dark, landscape, rainbow, lightning, thunder

➔ See more in this cluster...



water, bw, raindrops, window, droplets, white, black, red, light, canon

➔ See more in this cluster...



wet, green, street, drops, leaves, night, reflection, umbrella, bokeh, city

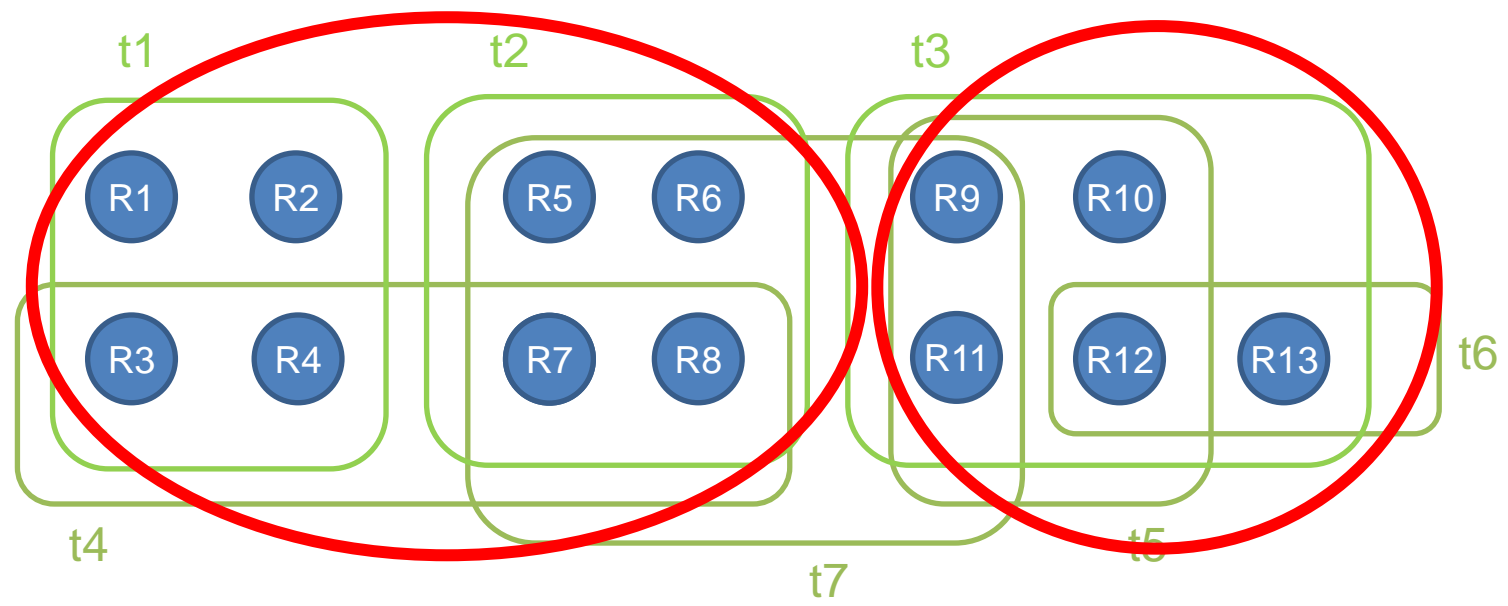
➔ See more in this cluster...



macro, flower, nature, drop, leaf, flowers, garden, rose, closeup, purple

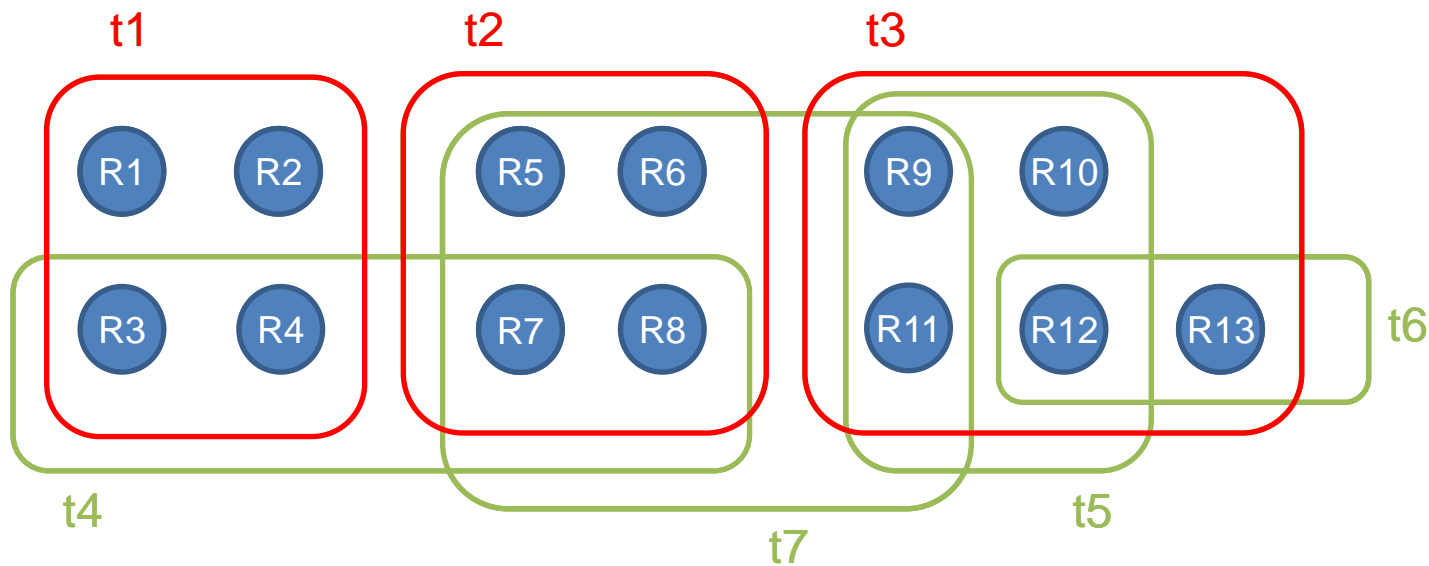
➔ See more in this cluster...

# Looking for Structure



# Discovering Structure

- How can structure be discovered?





# Partitioning Algorithm

- Idea
  - Find patterns within a folksonomy
    - NO clustering
    - NO hierarchical relationship
  - Providing facets
    - Represent different categories of one dimension
    - Faceted search provide a flexible and intuitive way to explore large collections of items (Yee et al., 2003)

# Faceted Search

**Nobel Prize Winners**  
1901 to 2004

Powered by Flamenco

Show tooltip previews of subcategories
 
 Username  Password    
[Create a New Account](#)

**GENDER**

|                             |                            |
|-----------------------------|----------------------------|
| <a href="#">female</a> (33) | <a href="#">male</a> (698) |
|-----------------------------|----------------------------|

**COUNTRY**

|  |  |
|--|--|
| <a href="#">Argentina</a> (5)<br><a href="#">Australia</a> (8)<br><a href="#">Austria</a> (12)<br><a href="#">Belgium</a> (11)<br><a href="#">Burma</a> (1)<br><a href="#">Canada</a> (9)<br><a href="#">Chile</a> (2) | <a href="#">China</a> (2)<br><a href="#">Colombia</a> (1)<br><a href="#">Costa Rica</a> (1)<br><a href="#">Czechoslovakia</a> (2)<br><a href="#">Denmark</a> (13)<br><a href="#">more...</a> |
|--|--|

**PRIZE**

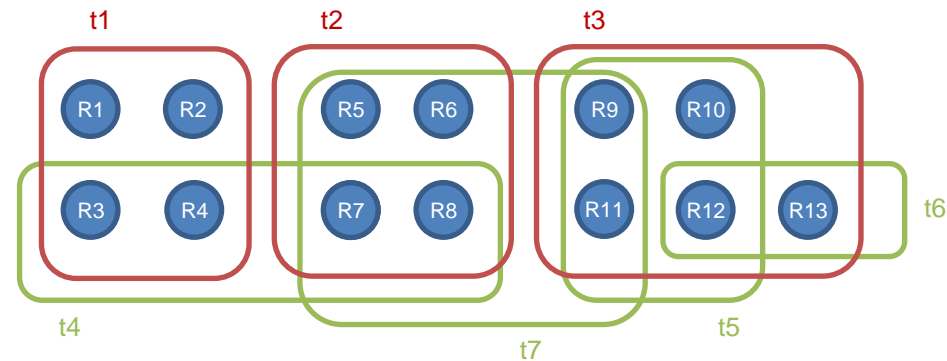
|   |  |
|---|--|
| <a href="#">chemistry</a> (138)<br><a href="#">economics</a> (55)<br><a href="#">literature</a> (101) | <a href="#">medicine</a> (182)<br><a href="#">peace</a> (108)<br><a href="#">physics</a> (168) |
|---|--|

**YEAR**

|  |   |
|--|---|
| <a href="#">1900s</a> (57)<br><a href="#">1910s</a> (40)<br><a href="#">1920s</a> (54)<br><a href="#">1930s</a> (58)<br><a href="#">1940s</a> (43)<br><a href="#">1950s</a> (72) | <a href="#">1960s</a> (79)<br><a href="#">1970s</a> (103)<br><a href="#">1980s</a> (97)<br><a href="#">1990s</a> (98)<br><a href="#">2000s</a> (58) |
|--|---|

# Discovering Facets

- A facet is a set of tags
  - Example:  $\{t1, t2, t3\}$



- Tags within a facet have to satisfy specific constraints
  - No overlap regarding to their extensions
  - Two tags that are labeled on the same resources must not appear within the same facet
- Determination using Linear Programming

# Linear Programming

- Optimization with linear relationships

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

|                               |  |
|-------------------------------|--|
| $x$ with $x_i \in \{0,1\}$    | Vector that represents the tag selection                   |
| $c$ with $c_i \in \mathbb{R}$ | „Value“ vector of the tags („ <i>objective function</i> “) |
| $A$                           | A matrix representing the constraints                      |
| $b$                           | Coefficients of the inequalities                           |

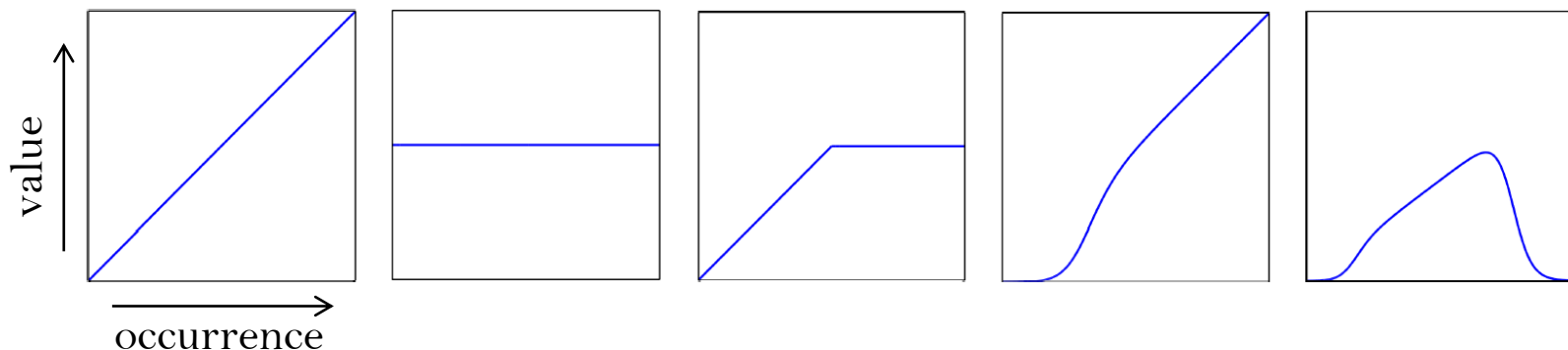
# Cost Function

$$\begin{aligned} \max \quad & c^T x \\ \text{s.t.} \quad & Ax \leq b \\ & x \geq 0 \end{aligned}$$

- Since linear programming maximizes outcome, cost function has huge effect!
- A value is assigned to each tag, depending on the number of their frequency

# Cost Function

- 5 different cost functions are implemented
  - Number of Occurrences
  - Uniformly
  - Capped
  - Logistic
  - Logistic with Cutoff



# Results

- Dataset
  - Photos from the online photo sharing platform Flickr
  - Public Group “Munich, Germany”

|                   |           |
|-------------------|-----------|
| Number of Photos  | > 30.000  |
| Number of Tags    | > 30.000  |
| Number of Members | ca. 3.700 |

Stand: 20 August, 2012



# Partitioning Facets

- Some very representative facets could be determined

| „Canon“              |       | „Olympus“            |      | „Museum“             |       |
|----------------------|-------|----------------------|------|----------------------|-------|
| ▼ Partitioning Facet |       | ▼ Partitioning Facet |      | ▼ Partitioning Facet |       |
| 1000d                | (37)  | e3                   | (20) | altepinakothek       | (38)  |
| 300d                 | (57)  | e30                  | (10) | artgallery           | (34)  |
| 350d                 | (50)  | e410                 | (18) | bmw                  | (148) |
| 400d                 | (120) | e420                 | (38) | brandhorst           | (137) |
| 40d                  | (37)  | e500                 | (43) | deutsches            | (67)  |
| 450d                 | (44)  | e510                 | (26) | glyptothek           | (39)  |
| 500d                 | (135) | ep2                  | (9)  | musée                | (40)  |
| 550d                 | (87)  | equirectangular      | (16) | neuepinakothek       | (18)  |
| 5dmarkii             | (108) | trip                 | (13) | residenz             | (17)  |
| 7d                   | (60)  | xa                   | (42) | theatermuseum        | (41)  |
| ixus                 | (39)  | xa2                  | (47) |                      |       |
| moment               | (41)  | xa4                  | (26) |                      |       |
| powershot            | (85)  |                      |      |                      |       |
| robert               | (35)  |                      |      |                      |       |

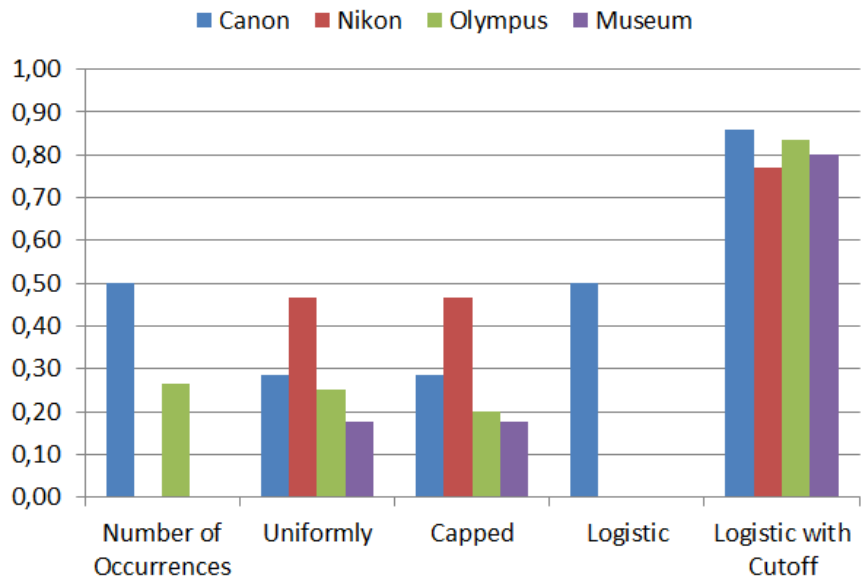


# Results

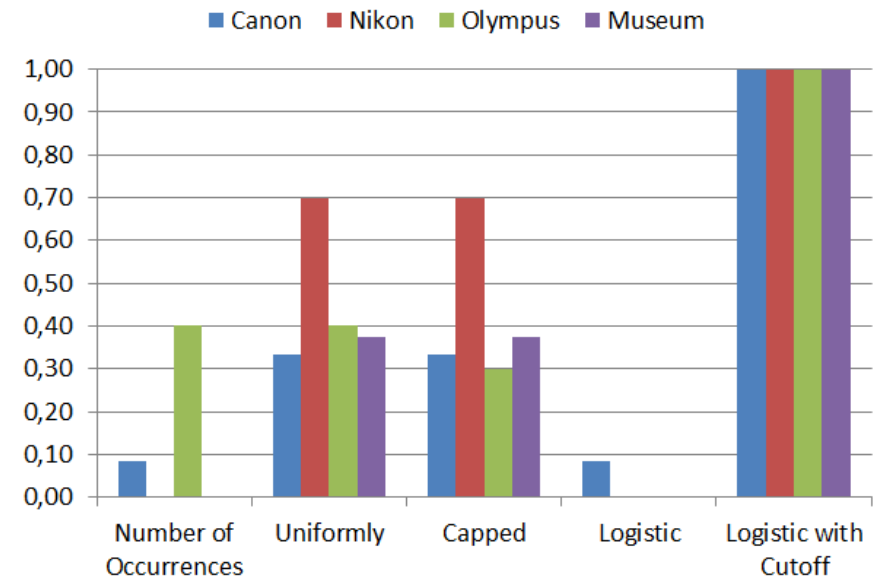
- Although some very meaningful facets could be extracted, noisiness is challenging!

| Tag     | Number of Occ. | Uniformly | Capped | Logistic | Logistic w. Cutoff |
|---------|----------------|-----------|--------|----------|--------------------|
| Canon   | 1 / 2          | 4 / 14    | 4 / 14 | 1 / 2    | 12 / 14            |
| Nikon   | 0 / 3          | 7 / 15    | 7 / 15 | 0 / 2    | 10 / 13            |
| Olympus | 4 / 15         | 4 / 16    | 3 / 17 | 0 / 1    | 10 / 12            |
| Museum  | 0 / 1          | 3 / 17    | 3 / 17 | 0 / 1    | 8 / 10             |

# Results



Precision



Recall

# Conclusion

- The partitioning algorithm works well on consistently tagged resources
  - Noisiness and ambiguity impair the result
- Cost function has huge impact
  - Further research necessary!
- However, it provides a new method to discover the latent structure of text-labeled objects!

# Thanks for your attention!

Questions?!

