Micro-interactions and Macro-observations

Deciding Between Competing Model

Steffen Staab
Slides by Klaas Dellschaft
Details of Model-based Research

- How to represent observables?
  - Distribution functions

- How to compare simulation and reality?
  - Analytical evaluation
  - Visual comparison
  - Goodness-of-fit tests

- How to decide between competing models?

Today!
Use Case: Dynamics in Tagging Systems

- How do users influence each other in tagging systems?
- Is this influence positive or negative for the indexing quality?
WeST

Method of Model-based Research

**Reality**
- Unknown Model
- Unknown rules of interaction

**Model**
- Stochastic Model
- Assumed rules of interaction

**Micro-interactions**

**Macro-observations**
- Observed Properties
- Simulated Properties

Compare
Co-occurrence Streams

- Co-occurrence Streams:
  - All tags co-occurring with a given tag in a posting
  - Ordered by posting time

- Example tag assignments for 'ajax':
  - \{mackz, r1, \{ajax, javascript\}, 13:25\}
  - \{klaasd, r2, \{ajax, rss, web2.0\}, 13:26\}
  - \{mackz, r2, \{ajax, php, javascript\}, 13:27\}

- Resulting co-occurrence stream:

  javascript  rss  web2.0  php  javascript

  time
Tag Frequencies

![Graph showing tag frequencies](image)

- Relative Tag Frequency
- Tag Rank

**Average**
Vocabulary Growth

![Graph showing the growth of vocabulary over the number of tag assignments. The graph displays multiple lines representing different datasets, with a red line indicating the average growth. The x-axis represents the number of tag assignments, and the y-axis represents the number of known tags.](image-url)
Method of Model-based Research

Micro-interactions

- Unknown Model
- Unknown rules of interaction

Macro-observations

- Observed Properties
- Simulated Properties

Reality

Model

- Stochastic Model
- Assumed rules of interaction

Compare
Both are heavy-tailed distributions (cf. power-law)
Distributions differ in their average exponent
Higher variance in the exponents for the tag frequencies
Different distributions
But: Possibly a common pattern of explanation
Hypothesis:
- If users would only have the free input field, then tag frequencies and word frequencies would look alike
- Tag recommendations change the original frequency distribution
Selection probabilities for 25 resources from a larger Delicious data set
Less than 50% of the tags do not occur in the recommendations
Method of Model-based Research

Micro-interactions

Reality

Unknown Model

Unknown rules of interaction

Model

Stochastic Model

Assumed rules of interaction

Macro-observations

Observed Properties

Simulated Properties

Compare

WeST

Klaas Dellschaft
klaasd@uni-koblenz.de

Introduction to Web Science
12 of 32
The Epistemic Model

- $P_{BK}$: Probability of entering a tag in the free input field
  - $p(W|t)$: Probability distribution of words $w \in W$ in context of topic $t$
  - $p(W|t)$ is modeled by the word frequencies in a text corpus about topic $t$

- $P_I = 1 - P_{BK}$: Probability of imitating a tag from the recommendations
  - Only modeling the selection from the popular tags
  - $n$ corresponds to the number of recommended tags

http://www.uni-koblenz.de/~staab/Research/Publications/2008/DellschaftStaabHypertext08.pdf
Alternative Models / Hypotheses
Yule-Simon-Model with Memory

- $p$: Probability of adding a new tag
- $1-p$: Probability of an already existing tag
- Probability $Q_t(x)$ of selecting a tag decreases with its age $x$
  - Only models the imitation of tags
  - Extension of linear preferential attachment

[Link to article: http://www.pnas.org/content/104/5/1461.short]
Semantic Walker Model

- Tag assignments of a user:
  - Random walk through a graph (=semantic space)
- Different models for generating the graph:
  - Watts-Strogatz model
  - Erdös-Rényi model
  - ...

- Only models the entering of tags in the free input field

http://www.pnas.org/content/106/26/10511.short
Deciding between Alternative Models
Problem of Induction (I)

- Logical problem:
  - Can the claim that a universal theory is true be justified by empirical reason?
  - **No!!!**

- Reformulation of the problem:
  - Can the claim that a universal theory is true or false be justified by empirical reason?
  - **Yes!!!** We can justify that it is false!
  - We should prefer theories which have not be shown to be false

---

David Hume

Karl Popper
Problem of Induction (II)

- Popper’s Critical Method:
  - We should try to falsify theories
  - If a theory is falsified, …
    - … we should define a succeeding theory
    - … which explains the same observations as the predecessor
    - … and which is successful where the predecessor failed

- Following this method, we may hit upon a true theory. But in no case the method establishes its truth, even if it is true.
- All theories must be regarded as hypothetical
### Comparison of Tagging Models

<table>
<thead>
<tr>
<th>Model</th>
<th>Imitation Processes</th>
<th>Background Knowledge</th>
<th>Tag Frequencies</th>
<th>Vocabulary Growth</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yule-Simon Model</td>
<td>✓</td>
<td>✗</td>
<td>power-law with cut-off</td>
<td>linear</td>
</tr>
<tr>
<td>Sem. Walker Model</td>
<td>✗</td>
<td>✓</td>
<td>power-law with cut-off</td>
<td>sublinear</td>
</tr>
<tr>
<td>Epistemic Model</td>
<td>✓</td>
<td>✓</td>
<td>power-law with cut-off</td>
<td>sublinear</td>
</tr>
</tbody>
</table>

- **Tag Frequencies**
  - Visual comparison: None of the models can be falsified
  - Apply goodness-of-fit tests for deciding between models
- **Vocabulary Growth**
  - Visual comparison: Yule-Simon Model is falsified!
  - Imitation is not sufficient for explaining the dynamics in tagging systems!
  - Is imitation part of a minimal model of tagging systems?
Method of Model-based Research

**Micro-interactions**
- **Reality**
  - Unknown Model
  - Unknown rules of interaction
- **Model**
  - Stochastic Model
  - Assumed rules of interaction

**Macro-observations**
- **Reality**
  - Observed Properties
- **Model**
  - Simulated Properties

**Compare**
Evaluation Setup

- Comparing simulations with real co-occurrence streams
  - Delicious: 10 streams
  - Bibsonomy: 5 streams

- Tag Frequencies
  - Measure distance between simulation and reality
  - Apply significance test whether it is a valid model (Kolmogorov-Smirnov Test)

- Vocabulary Growth
  - Minimize distance between simulation and reality
  - Significance tests not applicable because growth functions can not be represented as random variables
    - Allow maximal distance of ±10% to observed vocabulary growth
## Tag Frequencies – Results

<table>
<thead>
<tr>
<th>Stream</th>
<th>Epistemic Model</th>
<th></th>
<th>Semantic Walker</th>
<th></th>
<th>Yule-Simon Model</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ringtones</td>
<td>0.050</td>
<td>0</td>
<td>0.101</td>
<td>0</td>
<td>0.148</td>
<td>0</td>
</tr>
<tr>
<td>setup</td>
<td>0.029</td>
<td>0.03</td>
<td>0.133</td>
<td>0</td>
<td>0.147</td>
<td>0</td>
</tr>
<tr>
<td>boat</td>
<td>0.018</td>
<td>0.37</td>
<td>0.160</td>
<td>0</td>
<td>0.128</td>
<td>0</td>
</tr>
<tr>
<td>historical</td>
<td>0.012</td>
<td>0.86</td>
<td>0.176</td>
<td>0</td>
<td>0.093</td>
<td>0</td>
</tr>
<tr>
<td>messages</td>
<td>0.015</td>
<td>0.85</td>
<td>0.138</td>
<td>0</td>
<td>0.110</td>
<td>0</td>
</tr>
<tr>
<td>decorative</td>
<td>0.055</td>
<td>0.03</td>
<td>0.065</td>
<td>0.01</td>
<td>0.157</td>
<td>0</td>
</tr>
<tr>
<td>costs</td>
<td>0.012</td>
<td>0.99</td>
<td>0.196</td>
<td>0</td>
<td>0.051</td>
<td>0</td>
</tr>
<tr>
<td>ff</td>
<td>0.024</td>
<td>0.72</td>
<td>0.176</td>
<td>0</td>
<td>0.071</td>
<td>0</td>
</tr>
<tr>
<td>checkbox</td>
<td>0.049</td>
<td>0.51</td>
<td>0.080</td>
<td>0.09</td>
<td>0.175</td>
<td>0</td>
</tr>
<tr>
<td>datawarehouse</td>
<td>0.069</td>
<td>0.06</td>
<td>0.106</td>
<td>0</td>
<td>0.090</td>
<td>0</td>
</tr>
<tr>
<td>tools</td>
<td>0.063</td>
<td>0</td>
<td>0.212</td>
<td>0</td>
<td>0.062</td>
<td>0</td>
</tr>
<tr>
<td>social</td>
<td>0.057</td>
<td>0</td>
<td>0.243</td>
<td>0</td>
<td>0.066</td>
<td>0</td>
</tr>
<tr>
<td>design</td>
<td>0.039</td>
<td>0.04</td>
<td>0.181</td>
<td>0</td>
<td>0.084</td>
<td>0</td>
</tr>
<tr>
<td>analysis</td>
<td>0.081</td>
<td>0</td>
<td>0.056</td>
<td>0.04</td>
<td>0.156</td>
<td>0</td>
</tr>
<tr>
<td>blogs</td>
<td>0.037</td>
<td>0.09</td>
<td>0.210</td>
<td>0</td>
<td>0.057</td>
<td>0</td>
</tr>
</tbody>
</table>

Sign. ≤ 0.1: Significantly different distributions
Tag Frequencies – Messages Stream

\[ S(x) : \text{Fraction of Tags with Frequency } x \text{ or Higher} \]

- **messages**
- **Semantic Walker**
- **Epistemic Model**
- **Yule-Simon Model**

Tag Frequency vs. Fraction of Tags with Frequency x or Higher
### Vocabulary Growth – Results

<table>
<thead>
<tr>
<th>Stream</th>
<th>Epistemic Model</th>
<th>Semantic Walker</th>
</tr>
</thead>
<tbody>
<tr>
<td>ringtones</td>
<td>0.48</td>
<td>-0.21</td>
</tr>
<tr>
<td>setup</td>
<td>-0.09</td>
<td>-0.10</td>
</tr>
<tr>
<td>boat</td>
<td>-0.10</td>
<td>0.11</td>
</tr>
<tr>
<td>historical</td>
<td>0.10</td>
<td>-0.04</td>
</tr>
<tr>
<td>messages</td>
<td>0.07</td>
<td>-0.09</td>
</tr>
<tr>
<td>decorative</td>
<td>-0.20</td>
<td>-0.25</td>
</tr>
<tr>
<td>costs</td>
<td>0.09</td>
<td>-0.11</td>
</tr>
<tr>
<td>ff</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>checkbox</td>
<td>0.64</td>
<td>0.20</td>
</tr>
<tr>
<td>datawarehouse</td>
<td>-0.10</td>
<td>0.13</td>
</tr>
<tr>
<td>tools</td>
<td>0.08</td>
<td>-0.17</td>
</tr>
<tr>
<td>social</td>
<td>0.09</td>
<td>0.12</td>
</tr>
<tr>
<td>design</td>
<td>0.10</td>
<td>0.14</td>
</tr>
<tr>
<td>analysis</td>
<td>0.34</td>
<td>-0.45</td>
</tr>
<tr>
<td>blogs</td>
<td>0.10</td>
<td>-0.20</td>
</tr>
</tbody>
</table>
Vocabulary Growth – Ringtones Stream (I)

Where is this “step” coming from?
Step is caused by User 157 and 455 who are spammers.

Step is caused by User 10 who is a spammer.
Vocabulary Growth – Setup Stream

The diagram illustrates the growth of distinct tags over the number of tag assignments. Two lines are plotted:
- The black line represents the Semantic Walker.
- The red line represents the Epistemic Model.

The shaded area between the lines indicates the range of values for the setup scenario.
The Influence of Spammers (I)

- Epistemic Model
  - Based on assumptions about tagging behavior of regular users
  - Difficulties to reproduce observable properties, if spammers are present in the data set
    - Spammers and regular users have different tagging models?

- Compare properties between streams before removing spammers (unfiltered) and after removing spammers (filtered)
  - Kolmogorov-Smirnov Test should detect significant differences!
The Influence of Spammers (II)

The diagram shows the distribution of tag frequencies across different categories:

- **ringtones (filtered)**
- **ringtones (unfiltered)**
- **social (filtered)**
- **social (unfiltered)**

The x-axis represents the tag frequency, ranging from 1 to 10000, and the y-axis represents the fraction of tags with frequency x or higher, ranging from 1 to 0.0001.

The graph illustrates the comparative distribution of tag frequencies for filtered and unfiltered categories, highlighting the differences in tag distribution across ringtones and social content.
The Influence of Spammers (III)

- Compare difference between tag frequencies in filtered and unfiltered co-occurrence streams
- Goodness-of-fit Test: Kolmogorov-Smirnov Test

- Ringtones Stream:
  - Maximal distance $D$: 0.4972
  - Significance: 0.0 (= significant differences exist)

- Social Stream:
  - Maximal distance $D$: 0.0589
  - Significance: 0.0 (=significant differences exist)

- Spammers and regular users have different tagging models
Possible reasons for failed evaluations

1. The model is completely wrong

2. The model is missing influence factors
   - Semantic Walker Model: Missing influence of imitation
   - Yule-Simon Model: Missing influence of free tag input

3. The data contains noise
   - Epistemic Model: Influence coming from spammers