Contribution Profiles of Voluntary Mappers in OSM

Renate Steinmann, Simon Gröchenig, Karl Rehrl und Richard Brunauer
The crucial question is: „Who contributes what to a VGI project?“
Related Work

- Studies which assess contributor activity of VGI projects

- **Neis and Zipf (2012)**
  - Analysis of contributor activity using OSM history dump data
  - 38% of registered mappers performed only one edit
  - 5% are continuous mappers
  - Proposal of a classification of contribution profiles based on the number of nodes

- **Mooney and Corcoran (2012)**
  - Focus on characteristics of heavily edited objects (edited 15 or more times)
  - 11% of mappers contributed 87% of the heavily edited objects

- **Budhathoki and Haythornthwaite (2012)**
  - Distinguish between serious and casual mappers
  - Mapper classification is based on the number of contributed nodes, and/or the longevity of the contribution and/or the number of contribution days during active mapping periods
Related Work

- **Liu and Ram (2009)**
  - Analysis of collaboration patterns and the impact on the quality of Wikipedia articles
  - Found out relationships between contribution profiles and data quality

- **West et al (2012)**
  - Intend to draw a data-driven portrait of Wikipedia editors
  - Investigation of how online behaviour of Wikipedia editors can be distinguished from other users‘ behaviour

- **No satisfying answer** to the question **how user profiles can be identified** and described in VGI projects
Research questions

- How can different contribution profiles be identified from the OSM Full Planet History File?

- Which different mapping styles are reflected by the revealed profiles?
Example - Action Profile

- Action Profile of the user Steve Coast

<table>
<thead>
<tr>
<th>Action type</th>
<th>Number of actions (absolute)</th>
</tr>
</thead>
<tbody>
<tr>
<td>create</td>
<td>15,095</td>
</tr>
<tr>
<td>update</td>
<td>4,091</td>
</tr>
<tr>
<td>delete</td>
<td>1,255</td>
</tr>
</tbody>
</table>

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of days</td>
<td>181</td>
</tr>
<tr>
<td>Number of actions</td>
<td>20,441</td>
</tr>
</tbody>
</table>
Action Profile

- Profile calculation
  - Count all **create, update and delete actions** which a user has performed since registration
  - Calculation of **relative values** of create, update and delete actions
  - Additional profile attribute: **total number of contributed actions**
    - Decadic Logarithm
    - normalized values 0-1
  - Additional profile attribute: **number of mapping days**
    - Decadic Logarithm
    - normalized values 0-1
Example - Feature Profile

Feature Profile of the user
Steve Coast

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Number of actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>highway</td>
<td>1.438</td>
</tr>
<tr>
<td>building</td>
<td>1.095</td>
</tr>
<tr>
<td>amenity</td>
<td>345</td>
</tr>
<tr>
<td>natural</td>
<td>69</td>
</tr>
<tr>
<td>landuse</td>
<td>23</td>
</tr>
<tr>
<td>waterway</td>
<td>19</td>
</tr>
<tr>
<td>railway</td>
<td>11</td>
</tr>
<tr>
<td>place</td>
<td>4</td>
</tr>
<tr>
<td>power</td>
<td>-</td>
</tr>
<tr>
<td>other</td>
<td>277</td>
</tr>
</tbody>
</table>
Feature Profile

- Profile calculation
  - Count the **number of actions** of each user for each **primary feature**
  - Calculate **relative share** of edited map features by primary key
  - **Normalize** values between 0 and 1
Method – Identification of contribution profiles

Method in 5 steps

1. Statistical analysis – Visual inspection of data
2. Profile definition
3. Profile calculation
4. k-means clustering
5. Interpretation of results
### Results – Statistical analysis

<table>
<thead>
<tr>
<th></th>
<th>Sum</th>
<th>Average</th>
<th>Median</th>
<th>Maximum</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Actions</strong></td>
<td>2,992,949,966</td>
<td>11,314</td>
<td>23</td>
<td>185,645,045</td>
<td>550,200</td>
</tr>
<tr>
<td><strong>Create</strong></td>
<td>2,217,711,607</td>
<td>8,384</td>
<td>11</td>
<td>184,879,749</td>
<td>420,300</td>
</tr>
<tr>
<td><strong>Update</strong></td>
<td>504,506,387</td>
<td>1,907</td>
<td>5</td>
<td>169,404,872</td>
<td>330,468</td>
</tr>
<tr>
<td><strong>Delete</strong></td>
<td>270,731,972</td>
<td>1,023</td>
<td>1</td>
<td>18,276,275</td>
<td>46,638</td>
</tr>
</tbody>
</table>

- Number of users: 264.531
Results – Contribution Metrics

With an increasing number of mapping days the number of contributors decreases.

- Just 3% of users are active on more than 100 days.
Results – Contribution Profiles

- $k$-means clustering ($k=10$) on create, update and delete actions (relative values), action count (decadic logarithmic) and mapping days (decadic logarithmic), overview of cluster centroids
Creators contribute **most actions** (cluster 1 and 7)

**Deleters** (cluster 3) and **Premium All-Rounders** (cluster 6) are underrepresented.
Results – Contribution Profiles

Action shares per feature type (derived from OSM primary keys) per year (2005-2013)
Results – Contribution Profiles

$k$-means clustering ($k=10$) on primary keys, overview of cluster centroids
By far the biggest number of users fall into cluster 1 and cluster 9 ("Highway Mapper").

Cluster 3 (Power-Highway Mapper) and 6 (Place-Highway Mapper) have very small numbers of users.
Conclusions

- Adoption of \textit{k-means clustering} with centroids is gainful to identify and describe different mapper types.

- Most of the \textit{clusters are rather distinct} whereas some show a fluent transition (cluster calculation per year).
Conclusions

- **Repeat method** with different parameter sets to answer specific research questions.

- It would be worth to have a closer look on **relationships** between the two different profiling contexts „action types“ and „feature types“.

- Further research could examine **relationships** between **contribution profiles** and **quality of contribution**.
Thank you for your attention!