Modeling Cultural Evolution with Metadata Collections

Nic Weber
Information School, University of Washington, nwweber@uw.edu

Andrea Thomer
School of Information Sciences, University of Illinois at Urbana-Champaign, thomer2@illinois.edu

Metadata analytics is an emerging area of study concerned with the application of statistical, quantitative, and qualitative methods to the evaluation of structured data about digital objects.

In this poster, we demonstrate that metadata analytics can also be a valuable way of inferring historical relationships between digital objects, and understanding the cultures, organizations, and institutions responsible for their creation.

Specifically, we use phylomemetic techniques (Howe & Windram 2014) to visualize the evolution of Debian, a popular distribution of Linux.

Debian is one of the most successful distributions of Linux. Each release of Debian contains over four hundred individual different software packages (e.g. Word Processing apps like AbiWord, image editing apps like gimp, etc.). As with many digital artifacts, changes to Debian over time are summarized in metadata records about each release:

<table>
<thead>
<tr>
<th>Package</th>
<th>unstable</th>
<th>std.</th>
<th>testing</th>
<th>8.0</th>
<th>testing</th>
<th>7.0</th>
<th>unstable</th>
</tr>
</thead>
<tbody>
<tr>
<td>current</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>app</td>
<td>1.1</td>
<td>1.1</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>cmake</td>
<td>1.1</td>
<td>1.1</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>boost</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>boost</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.9</td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
</tbody>
</table>

These changes follow a pattern of “descent with modification,” in which the distribution is replicated and altered in each new version. We take inspiration from recent work in phylomemetics, in which analyses from evolutionary biology are applied to the study of cultural artifacts, and ask,

**Can we infer history of software from metadata records alone?**
**What can we learn from analysis of collections of metadata records that we can’t easily learn from individual records alone?**
**What can analysis of collections of metadata records tell us about the metadata quality?**

Method: Phylomemetics for Metadata Analytics

- Used PAUP (Phylogenetic Analysis Using Parsimony *and other methods) version 4.0a150 (Swofford, 2016) to infer the most parsimonious phylomemetic tree.
- Characters with explicit timestamps such as “Release Date” were excluded to test whether the metadata was rich enough to retrieve the history on its own.

Results

The phylomemetic approach successfully recovered the known chronology and evolutionary history of the Debian package; this implies that the Distrowatch metadata records are indeed sufficiently informative of the changes made to Debian over time. This echoes prior findings that this method can be used as a rough metric of metadata quality (Thomer & Weber 2014).

Future work:

- Can disparities between time and number of changes in a data product be used to identify points of social stress within development communities?
- Can this method be used to help identify significant properties about a digital object – and thereby guide metadata creation?

References:
