Mining Evolution Data of a Product Family

Michael Fischer, Johann Oberleitner, Jacek Ratzinger
Vienna University of Technology

Harald Gall
University of Zurich
Product Family Evolution

Objective: find relationships between variants of a product line

Approach:
- Use lexical search in change log messages
- Retrieve change dependencies from RHDB
- Visualize findings about change dependencies and information flow between source code directories via Multidimensional Scaling (MDS)

Case study:
- Product variants: FreeBSD, NetBSD, OpenBSD (30,000 - 60,000 files)
- Direct copies of CVS systems
- Used keywords: *freebsd*, *netbsd*, *linux*
Process

FreeBSD

NetBSD

OpenBSD

Consolidated DB

Visualization

import

RHDB

RHDB

RHDB

usr_sbin

usr_bin

usr_sbin

usr_bin

lib

src

lib

sys

man

share

gnu

arch

i386
# Common files

<table>
<thead>
<tr>
<th>Variant</th>
<th>Variant</th>
<th>all modules</th>
<th>src/sys/ only</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeBSD</td>
<td>NetBSD</td>
<td>3810</td>
<td>1333</td>
</tr>
<tr>
<td>FreeBSD</td>
<td>OpenBSD</td>
<td>3839</td>
<td>1079</td>
</tr>
<tr>
<td>NetBSD</td>
<td>OpenBSD</td>
<td>6969</td>
<td>6847</td>
</tr>
</tbody>
</table>
Information flow between variants of the BSD systems based on lexical search

<table>
<thead>
<tr>
<th>Variant</th>
<th>Keyword</th>
<th>all revisions</th>
<th>revision ≥ 1.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>FreeBSD</td>
<td>netbsd</td>
<td>5131</td>
<td>3577</td>
</tr>
<tr>
<td></td>
<td>openbsd</td>
<td>2729</td>
<td>1353</td>
</tr>
<tr>
<td></td>
<td>linux</td>
<td>1791</td>
<td>1387</td>
</tr>
<tr>
<td>NetBSD</td>
<td>freebsd</td>
<td>2852</td>
<td>2186</td>
</tr>
<tr>
<td></td>
<td>openbsd</td>
<td>2679</td>
<td>2224</td>
</tr>
<tr>
<td></td>
<td>linux</td>
<td>1547</td>
<td>1125</td>
</tr>
<tr>
<td>OpenBSD</td>
<td>freebsd</td>
<td>2406</td>
<td>1933</td>
</tr>
<tr>
<td></td>
<td>netbsd</td>
<td>16802</td>
<td>7423</td>
</tr>
<tr>
<td></td>
<td>linux</td>
<td>775</td>
<td>463</td>
</tr>
</tbody>
</table>
Global view: OpenBSD
Abstracted view: *OpenBSD*
## Referenced files: OpenBSD

- Topmost referenced files with one of the given keywords in the change logs of OpenBSD

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Count</th>
<th>Path</th>
</tr>
</thead>
<tbody>
<tr>
<td>freebsd</td>
<td>59</td>
<td>src/sys/dev/pci/files.pci</td>
</tr>
<tr>
<td>.</td>
<td>52</td>
<td>src/sys/dev/pci/pciide.c</td>
</tr>
<tr>
<td>.</td>
<td>52</td>
<td>src/sys/dev/pci/pcidevs</td>
</tr>
<tr>
<td>netbsd</td>
<td>45</td>
<td>src/sys/arch/i386/i386/machdep.c</td>
</tr>
<tr>
<td>.</td>
<td>43</td>
<td>src/sys/dev/pci/pciide.c</td>
</tr>
<tr>
<td>.</td>
<td>39</td>
<td>src/sys/conf/files</td>
</tr>
<tr>
<td>linux</td>
<td>14</td>
<td>src/sys/compat/linux/linux_socket.c</td>
</tr>
<tr>
<td>.</td>
<td>14</td>
<td>src/sys/compat/linux/syscalls.master</td>
</tr>
<tr>
<td>.</td>
<td>5</td>
<td>src/sys/dev/ic/if_wireg.h</td>
</tr>
</tbody>
</table>
Detailed change analysis

- Change in FreeBSD file `ufs_quota.c`
  ```c
  < sleep((caddr_t)dq, PINOD+2);
  ---
  > (void) tsleep((caddr_t)dq, PINOD+2, "dqsync", 0);
  ```

- Change in NetBSD file `ufs_quota.c` (six years later)

- Change in OpenBSD file `ufs_quota.c` (eight years later)
  ```c
  < sleep((caddr_t)dq, PINOD+2);
  ---
  > (void) tsleep(dq, PINOD+2, "dqsync", 0);
  ```
Conclusions & Possible Extension

- Lexical search indicates increasing information flow between product variants
- Recovered high level view indicates high coupling and wide spread of “alien” source code
- Text mining in modification reports
  - Vector space search (latent semantic indexing)
- Code clone detection
  - Clone detection in source code
  - Track clone propagation through version history
  - Clustering
    - Measuring Similarity of Large Software Systems Based on Source Code Correspondence [Yamamoto et al.]
    - CCFinder [Kamiya et al.]