A generic way of log sampling
Motivation

Big event log

Computing time
Computing space

Precision
Comprehensiveness

CRASH?!
Goal

Big event log

Importing plugin for big log
Plugin in ProM to sample log

A efficient way of log sampling

Representative Customized

Sample event log
Case Study

Search engine company
Event log size: 66 million clicks in one month
Case study: Page search (2 million clicks, 4GB/day)

Recorded information:
• Visiting path (Referrer)
• Viewer amount (Hits)
• Page Number (PN)
• Search key words
• TimeStamp
• IP, Browser, ...
Step 1: Import

Current Import Framework loads log into memory.

Import Plugin: CSV Referrer
1 Common format and light weigh (vs XES)
2 Random Log Access:
   1) Sequential reader
   2) Random reader: reader can access random location of file
<table>
<thead>
<tr>
<th>Timestamp</th>
<th>IP</th>
<th>UTP</th>
<th>Browser</th>
<th>UID</th>
<th>Cookit ID</th>
<th>ID</th>
<th>Channel</th>
<th>prodict</th>
<th>Name</th>
<th>If</th>
<th>Referer</th>
<th>Tel</th>
<th>RA</th>
<th>U</th>
<th>RA</th>
<th>AD</th>
<th>PageNum</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-02-02-12:10:10</td>
<td>192.168.0.1</td>
<td>10.0.0.1</td>
<td>Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:65.0) Gecko/20100101 Firefox/65.0</td>
<td>123456789</td>
<td>01234567890123</td>
<td>123</td>
<td>test1</td>
<td>page</td>
<td>www</td>
<td>123</td>
<td><a href="http://example.com">http://example.com</a></td>
<td>1234</td>
<td>5678</td>
<td>321</td>
<td>4567</td>
<td>890</td>
<td>1234567</td>
<td>8901</td>
</tr>
</tbody>
</table>

- **Save Separator**
- **Read More**
- **Random Read**
Step 2: Mapping

Import

Mapping

Log Simplification

Decision Support

Sampling

Export
Step 2: Mapping

Search EngineLog
Trace: UserID
Event: Referer
Attributes:
Hits,
PageNumber,
SearchKeyWord

→ Remove irrelevant columns
Step 3: Log Simplification

Import

Mapping

Log Simplification

Decision Support

Sampling

Export
Step 3: Log Simplification

- Log size is decided by trace number and event number.
- Log complexity is related to the number of different activates and their relations.
- Simplify log by properties analysis and editing, in order to reduce data set.
Log properties analysis

Abstract Character

Event Execution Character

- Log Summary
- Trace Character
- Tandem Array Pattern
- Tandem Repeat Pattern
- Event Value Type
- Event String Character
- Event Value Gradation
- Event Value Similarity

Examples:

- eabcabcf
- eabcabc
- eaabbccaaaabccf
- eaabbccaaaabcc
- eabcabeabcabef
- eabcaabbeabcaabceef

- eabcf
- eabc
- eabcf
- eabce
- eabcef
Log properties

- Abstract Character
  - LogSummary
  - TraceCharacter
  - EventExecutionCharacter
    - TandemArrayPattern
    - TandemRepeatPattern
  - EventValueType
    - Event value gradation
  - EventStringCharacter
    - Event Value Similarity

www.web.com
www.web.com/music
www.web.com/news
www.web.com/news/international
Log properties

SameValueType: Hits
Event Value: 1, 2, 3 ... 100

Edit:
(1,2..33)->small
(34,35..66)->medium
(67,68..100)->large
Log properties

Search key word value:
Halloween party, Christmas recipe, romantic flower, valentines’ day, Halloween costume, Romantic song, Christmas tree

Read similarity edit threshold:
Romantic song, romantic flower, valentines’ day, Halloween party, Halloween costume, Christmas tree, Christmas recipe

Halloween, Christmas, romantic, romantic, Halloween, Romantic Christmas
Search Engine Log

www.web.com/music, hits: 400, sk: Romantic songs
www.web.com/news/international.html?/newsid=1&cid=11, hits: 5, sk: Christmas ceremony

Reduce data set

music, hits: medium, sk: Romantic
news, hits: medium, sk: Halloween
news, hits: high, sk: Halloween
news, hits: low, sk: Christmas
Extensible Framework
Step 4: Log Sampling

Import
Mapping
Log Simplification
Decision Support
Sampling
Export
Step 4: Log Sampling by two dimensions

- Horizontal Sampling
- Vertical Sampling
Step 5: Decision Support

Import

Log Simplification

Decision Support

Sampling

Export

CSV

Mapping

properties analysis

external sorting

case clustering

Horizontal Sampling

Vertival Sampling

Sample CSV
External Sorting

Glue logs with same CaseID, and sort events by TimeStamp.

- Case1,Time2,C
- Case2,Time2,C
- Case1,Time1,B
- Case1,Time3,B
- Case3,Time1,A
- Case2,Time1,B

- Case1,Time1,B
- Case1,Time2,C
- Case1,Time3,B

- Case2,Time1,B
- Case2,Time2,C

- Case3,Time1,A
Sampling Criteria

- Simplicity: \{eaabbcabcf\}, \{eabcf\}
- Completeness: \{aaced\}, \{ac\}, \{ed\}
- Proportion

Sample size:
- Sample size is depends on the similarity among traces.
- More similar traces, smaller sample need to be preventative for original log

Others
Remove the case that do not contains new patterns
- \{ABCABC\} \{AABBCABCCC\}
### Iterative Random Sampling

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>300</td>
<td>85</td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>52</td>
<td></td>
<td></td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>58</td>
<td>10</td>
<td></td>
<td>453</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>88</td>
<td></td>
<td></td>
<td>553</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>53</td>
<td></td>
<td></td>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>30</td>
<td>8</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>b</td>
<td>5</td>
<td></td>
<td></td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>c</td>
<td>6</td>
<td>1</td>
<td></td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>d</td>
<td>9</td>
<td></td>
<td></td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>e</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

Sample size: 10%.
Iteratively Random Sample log until it fills up sample matrix.
Bigger the frequency the trace has, more likely it can be chosen.
Conclusion:
1 Sample is not unique
2 By sampling, log sacrifice irrelevant or less important criteria to reduce the size and magnifier characteristics.

Future work:
1 Evaluation of sampling
2 Experiment on efficiency
Questions:
1 Does the sampling criteria meet requirement of further miners.

2 What has to be contained in the sample is the execution pattern, in order to make sure that the sample and original log has similar discovered model further on.
Thank you!