Sankey Path Analysis of Web Button Clicks and Page Views

Nancy Abramson
Product Data Scientist
nabramson@ieee.org
720-468-1796
Customer Application Fails

Understand which customer actions create a poor experience

- Replicate for testing
- Focus for programming fix
- Redirect user away from an error to safe haven
A Sankey diagram is a map showing how information flows through events
Sankey Diagram Power

Flow diagram
- Width proportional to flow
- Dense Data Display


© 2015 Nancy Abramson, Abramson Ventures, LLC
Analysis Overview

- Combine Web and Server Data Sources
  - Best Practices to Blend Data Sources
  - Method for Disparate Session
  - SQL Sample to Fill in Sparse Data
  - Aggregate Values from Dual Sources

Result: Sankey Visual Analysis

Application: Automation and Targeted Debug
Adobe Omniture Web
- Third party service
- Collects website clicks made by visitors
- Organized by events in a session

Apache Tomcat Server
- System health
- System errors
- Log report
- Aggregate Values from Dual Sources

Union data together within a database table
Data Blending Difficulties

Each data source often has its own...

- Surrogate Key
- Session Definition
- Data Elements
- Server Location
- Reporting Speed
- Sparse Data

*Benefit of using different data sets is seeing the data from multiple perspectives*
**Harmonize**

**Impose structure to blend**

- Need common natural or passed key between data sets (e.g. account, device id, user id)
- Store timestamps without time zone (UTC)
- Synchronize servers with Network Time Protocol (NTP)
- Combine data sources into conformed columns to preserve the richness of the user experience

**Insert metadata columns**

- Source of data for traceability
- Insert date timestamp
- Session using *your own business* definition
Partition Window Over Natural Key

- Order by time
- Use business rules for max/min session times
- Require session events such as login/logout
- Select values in window (min, max, first, last)

Conform Sparse Events Across Columns

```
COALESCE(post_pagename,'') ||
COALESCE(post_buttonclick,'') ||
COALESCE(post_togglename,'') as behavior
```
Replace NULL Row Values within Session

```
UPDATE event_table set user_id = v.user_id
FROM
(SELECT max(user_id) as user_id, session_id
 FROM event_table
 WHERE session_id IN
   ( SELECT distinct session_id FROM event_table WHERE source = 'omniture' AND user_id is NULL )
 GROUP BY session_id ) v
WHERE v.session_id = event_table.session_id
AND event_table.user_id IS null AND source = 'omniture';
```
Use Business for Multiple Column Compression

CASE

WHEN (event is NULL OR event LIKE 'APP') AND channel is null THEN error_code

WHEN (event is NULL OR event LIKE 'APP') AND (error_code is null OR error_code LIKE 'APP') THEN channel

WHEN (error_code is NULL OR error_code LIKE 'APP') AND channel is NULL THEN event

WHEN (event is NULL OR event LIKE 'APP') THEN error_code

WHEN channel is NULL THEN error_code

WHEN (event is NULL OR event LIKE 'APP') THEN channel

ELSE event

END AS click
### Example Row Count Aggregation

<table>
<thead>
<tr>
<th>common_id</th>
<th>event</th>
<th>source</th>
<th>time</th>
</tr>
</thead>
<tbody>
<tr>
<td>100ASD39Q</td>
<td>Your Account</td>
<td>omniture</td>
<td>2015-07-23 16:00:02</td>
</tr>
<tr>
<td>100ASD39Q</td>
<td>Message</td>
<td>omniture</td>
<td>2015-07-23 16:00:45</td>
</tr>
<tr>
<td>100ASD39Q</td>
<td>CLOSE</td>
<td>omniture</td>
<td>2015-07-23 16:01:54</td>
</tr>
<tr>
<td>100ASD39Q</td>
<td>ERR-3005</td>
<td>tomcat</td>
<td>2015-07-23 16:03:19</td>
</tr>
<tr>
<td>100ASDAFX</td>
<td>CLOSE</td>
<td>omniture</td>
<td>2015-07-16 22:35:09</td>
</tr>
<tr>
<td>100ASDAFX</td>
<td>Your Account</td>
<td>omniture</td>
<td>2015-07-16 22:41:15</td>
</tr>
<tr>
<td>100ASDAFX</td>
<td>CLOSE</td>
<td>omniture</td>
<td>2015-07-16 22:42:38</td>
</tr>
<tr>
<td>100ASDAFX</td>
<td>ERR-3005</td>
<td>tomcat</td>
<td>2015-07-16 22:50:26</td>
</tr>
<tr>
<td>111C320L20</td>
<td>CLOSE</td>
<td>omniture</td>
<td>2015-07-23 15:50:56</td>
</tr>
<tr>
<td>111C320L20</td>
<td>Your Account</td>
<td>omniture</td>
<td>2015-07-23 15:51:03</td>
</tr>
<tr>
<td>111C320L20</td>
<td>CLOSE</td>
<td>omniture</td>
<td>2015-07-23 15:55:23</td>
</tr>
<tr>
<td>111C320L20</td>
<td>ERR-3018</td>
<td>tomcat</td>
<td>2015-07-23 16:00:12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>common_id</th>
<th>error_code</th>
<th>any_event</th>
<th>time_start</th>
<th>time_end</th>
</tr>
</thead>
<tbody>
<tr>
<td>100ASD39Q</td>
<td>ERR-3005</td>
<td>[Your Account, Message, CLOSE, ERR-3005]</td>
<td>2015-07-23 16:00:02</td>
<td>2015-07-23 16:03:19</td>
</tr>
</tbody>
</table>
Row Event Compression

- 1,118,512 Adobe rows
- 4,635,269 Apache rows
- 5,753,781 Error events
- 58,360 Aggregate events

Dozen event paths represent 75% of data
Results

Manual
- Business Review
- Targeted Development Debug
- Focused Testing and Replication

Automatic
- Top error path re-routes to working page
More Information

- Free Classes: https://www.coursera.org/

- *The Visual Display of Quantitative Information*
  Edward Tufte

- Teradata Aster Big Data Analytics