VISUAL ANALYSIS OF MASSIVE WEB SESSION DATA

Jack Shen
Jack Shen

- Architect @ Behavior and Product Experience, eBay
- Previously
  - Research Scientist @ eBay Research Labs
  - Ph.D. in Information Visualization @ UC Davis
- zqshen@gmail.com
Visual Analysis of Massive Web Session Data
Zeqian Shen, Jishang Wei, Neel Sundaresan, and Kwan-Liu Ma,
In IEEE Symposium on Large Data Analysis and Visualization 2012
Connecting buyers and sellers worldwide

- Hundreds of millions of people visit every day
- Millions of items are posted and purchased every day

Understanding user experience is crucial for improving our service.
Analyzing Web Session Data

- Sessions and events

- Visual analytics enables exploration and discovery.

- Challenges:
  - Large-scale data: hundreds of millions of sessions, 2TB per day.
  - Structural complexity
Real-world Scenarios for Analyzing User Experience

- **Goal:** Improve the success rates of the key flows
- **Study the correlations between user behavior patterns in the flow and the success rates.**
  - The success rates of various event sequence patterns, i.e., ordering of events and elapsed time.
  - The inbound/outbound, elapsed time and other statistics of an event and its correlation with the success rate.
  - Analysts conduct the analysis iteratively.
System Design

Two-tier visualization system

Tier 1: Data filtering on Hadoop

Tier 2: In-memory visualization, real-time interactive exploration

Web Session Data

Data Filtering

Data Visualization

Hypothesis, Analysis, and Verification

User Interaction
Data Filtering

Using Mobius, a Hadoop-based large scale data analytics platform, to extract event sequences based on user input.
Visualization

Grouping the event sequences by ordering of the events and elapsed time
Data Sizes at Different Tiers

- **Tier 1**: 2TB on Hadoop, hundreds of millions of sessions.
- **Tier 2**: ~50MB, 200K sequences (Checkout)
  - In memory tree: ~30MB, bucket size for elapsed time: 1 second.
  - Visualizing: ~5MB, 67K sequences, bucket size for elapsed time: 10 seconds.
User Interface
User Interactions: Highlighting
User Interactions: Aligning

Rebuilding the tree is needed
User Interactions: Zooming

Changing the bucket size for elapsed time to do semantic zooming
Case Study: Category Selection in Listing Process

Process for sellers listing an item on eBay

Category Selection
- Browse Category
- Search Category
- Catalog Search
- Describe Item
- Review Enhance
- Congratulations
Case Study: Category Selection in Listing Process

Event: SearchCategory
Elapsed Time (second):
- avg: 40.00 median: 40.00
- min: 40.00 max: 40.00
Traffic:
- 8.30% of the total
- Success Rate: 68.84%

Event: SearchCategory
Elapsed Time (second):
- avg: 12.00 median: 12.00
- min: 12.00 max: 12.00
Traffic:
- 47.26% of the total
- Success Rate: 80.60%

Cumulative Distribution
- 0% 20% 40% 60% 80% 100%

- Event: DescribeItem
  - Elapsed Time (second):
    - avg: 445.07 median: 224.00
    - min: 0 max: 36180
  - Traffic:
    - 84.02% of the total
    - Success Rate: 88.05%
Case Study: Category Selection in Listing Process

Understand behavior patterns of sellers who chose to find a category through browsing.
Case Study: Category Selection in Listing Process

Understand the impact of catalog based category selection on user experience in listing.
Feedback

- The system is useful for both novice and expert users.
- The response latency in data filtering is acceptable.
- The visualization and interaction are intuitive and useful.
Summary

- Introduced a two-tier visual analytics system for analyzing user temporal behavior patterns.

- Discussed the process and principles in designing such large scale visualization system:
  - Design the system in tiers based on real-world scenarios.
  - As moving up the tiers, less data and computation, more interactive.
Questions?
zqshen@gmail.com