



Interactive Path Analysis of Web Site Traffic

Pavel Berkhin

pavelb@accrue.com

Presenting Author

Jonathan D. Becher

becher@accrue.com

Dee Jay Randall

drandall@accrue.com

Accrue Software, Inc.

www.accrue.com

510-580-4500



Problem Space Overview

- Goals

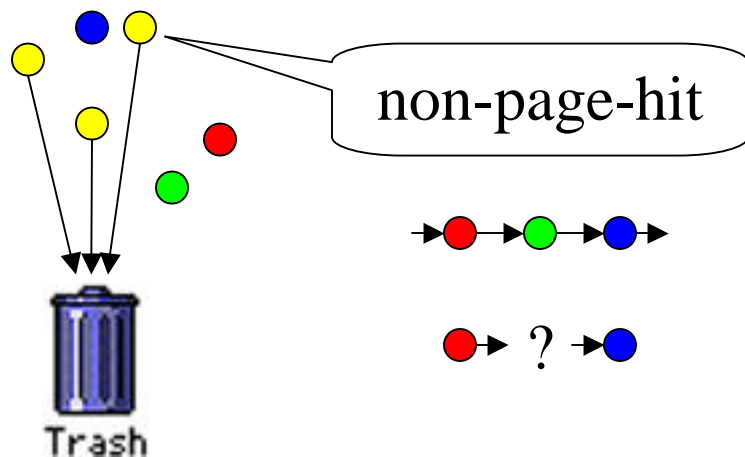
- Learn **sequential patterns** of visitor's sessions
- What sessions convert, why and how
- What is typical, what is interesting
- To devise effective representations of click stream data

- Data

- Accrue G2 data repository
- Sessionizing
- Robots
- Extract relevant information and eliminate the extraneous

- Basic Objects

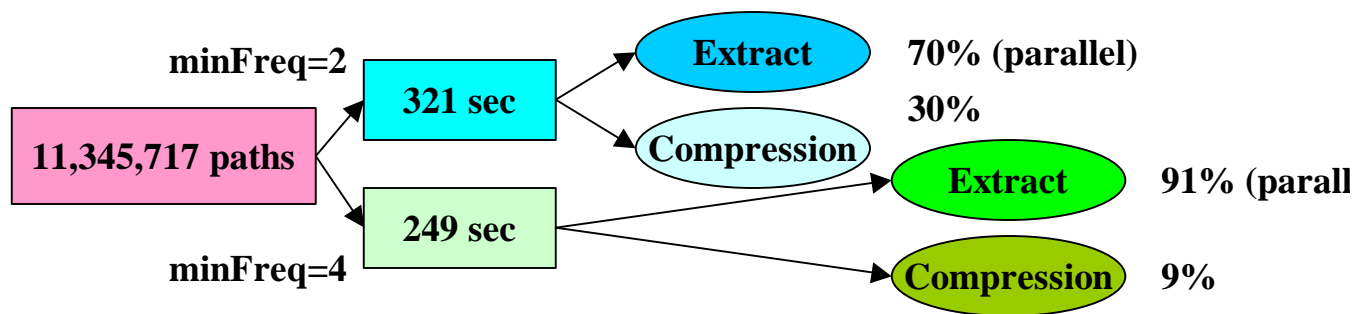
- Elements
- Paths
- Couples



Strategy

- Pre-processing Phase

- Data extraction, compression and repackaging
- Computationally intensive
- User configurable (set once before a run)
- Slow: **Precook frozen chicken with rice**



- Interactive Phase

- Analysis and exploration based on pre-computed structures
- User driven (changed interactively through the GUI)
- 3-tier distributed
- Fast: **microwave a chicken for a lunch**

~~Is it about visualization?~~

Visualization is important, but sequential analysis is *data mining*

Pre-processing Phase

Data Extraction step

- Extraction
- Sampling
- Mapping
- Preprocessing
- Filtering

Data Preparation Step

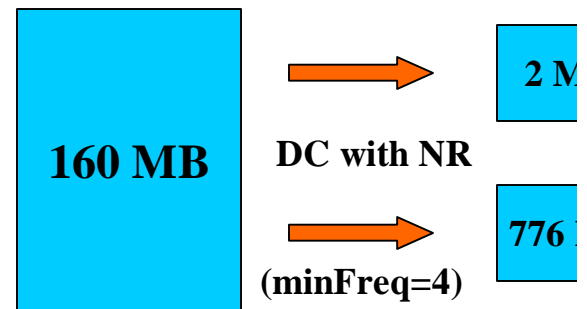
- Noise Reduction (**NR**)
- Data Compression (**DC**)
- Building special data structures

Descriptive Statistics

NR with minFreq=2

Data set	Path count	DC	DC with NR
1	11,341,943 (minFreq=4)	7.1 7.1	77.3 211.
2	4,592,033	7.1	86.2
3	934,162	3.9	40.0
4	165,109	2.3	22.5
5	1,162,135	4.1	51.3

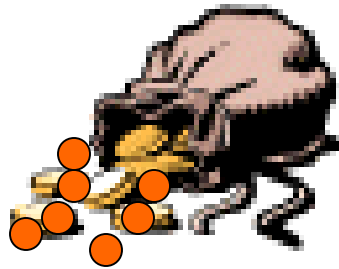
- Media Site
- Educational Site
- Retailer Site



Interactive Phase

- Element
 - Analyzer
 - Explorer

Analyzer



- Path
 - Analyzer
 - Explorer

Explorer



- Couple
 - Analyzer
 - Explorer

- Descriptive Statistics

~~Is it about most frequent objects?~~

Popularity alone is seldom informative. You probably already know what the overall most popular paths are. The *unexpectedly* popular paths are informative.

Elements

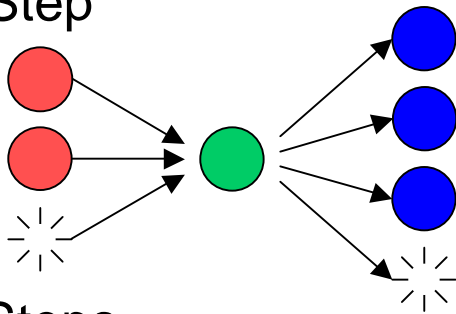
- Analyzer
 - Frequent elements
 - Frequent *entries, exits, 1-hits*
- Explorer
 - k-th step predecessors, successors
 - Composition, convergence
 - ✍ **Q1.** What are the chances that **y** is a **k**-step successor of **a**?
 - ✍ **Q2.** What are the chances of reaching **y** from **a** in **k** steps the first time?
 - ✍ **Q3.** What are the chances of reaching **y** from **a** in no more than **k** steps?

✍ **Example:** what preceded and what followed registration page?

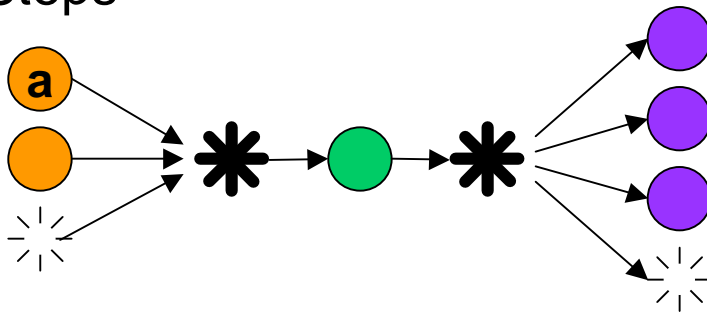
Element Explorer

- Butterfly Graph

- 1 Step



- k Steps



Q1. * = k-1 Steps

Q2. * = k-1 Steps, ? a

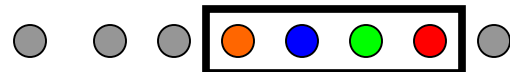
Q3. * ? k-1 Steps

Path Analyzer

● Frequent Subsequences

– Filtering

- ✍ By starting elements
- ✍ By ending elements
- ✍ By including elements, etc.



✍ **Misperception:** Path analysis means examining full paths

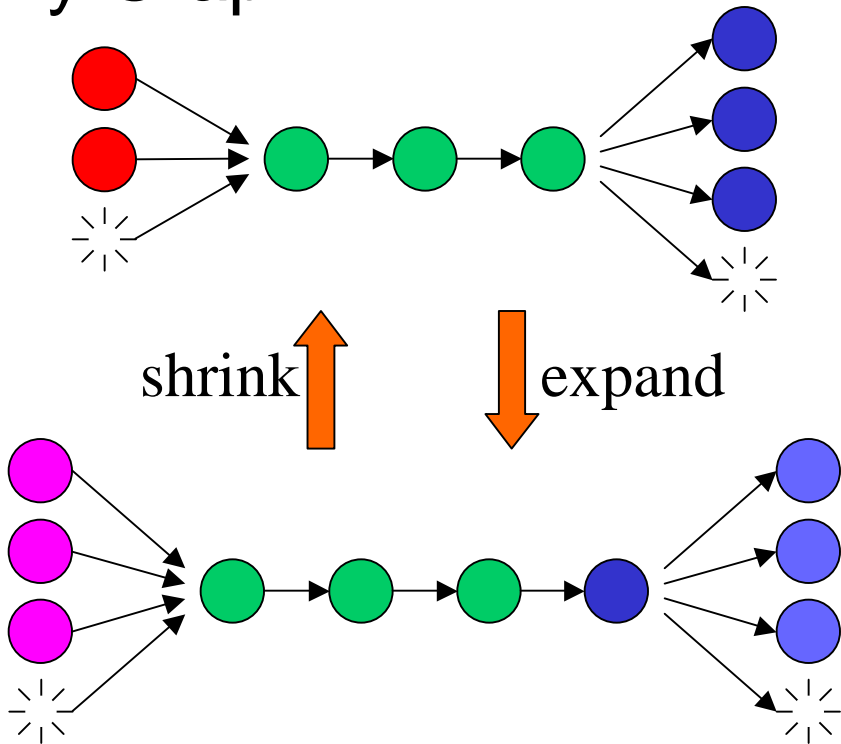
✍ **Truth:** Full paths can be useful, but often far more insight is gained from examining specific length subpaths

– Coverage

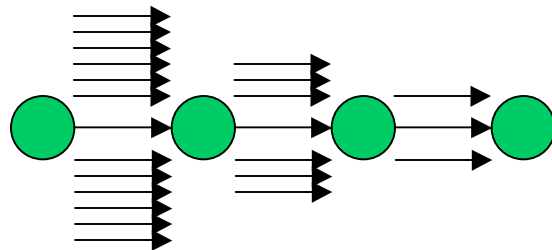
	Education Site	Finance Portal	Computer Vendor
Total number of 4-long paths	4,081,707	8,336,165	2,526,607
N=32	15.20%	78.50%	48.90%
N=64	20.60%	81.10%	52.90%
N=128	26.40%	83.60%	57.20%
N=256	33.30%	85.20%	61.40%

Path Explorer

- Butterfly Graph

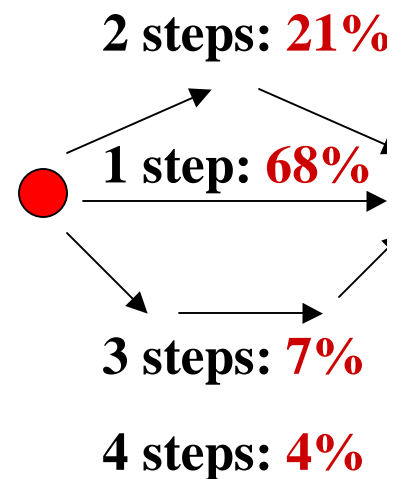


- Flow Drop-off



Couple Analyzer

- Measures
 - Frequency
 - %-age of frequency by steps
 - Average steps
 - Association measures
 - ✍ Confidence
 - ✍ Similarity
- Filtering
 - By starting, ending
 - By measures



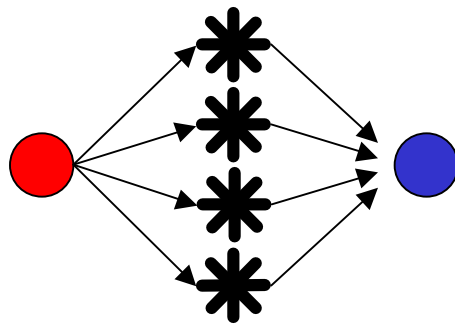
Average steps: 1.

✍ Examples:

- Find couples with strong association
- Find predominantly distantly related couples (high # of average steps, not a physical link)

Couple Explorer

- Frequent Connecting Paths
- Proper Paths
- Filtering
 - Length
 - Frequency
 - Including / Non-Including elements

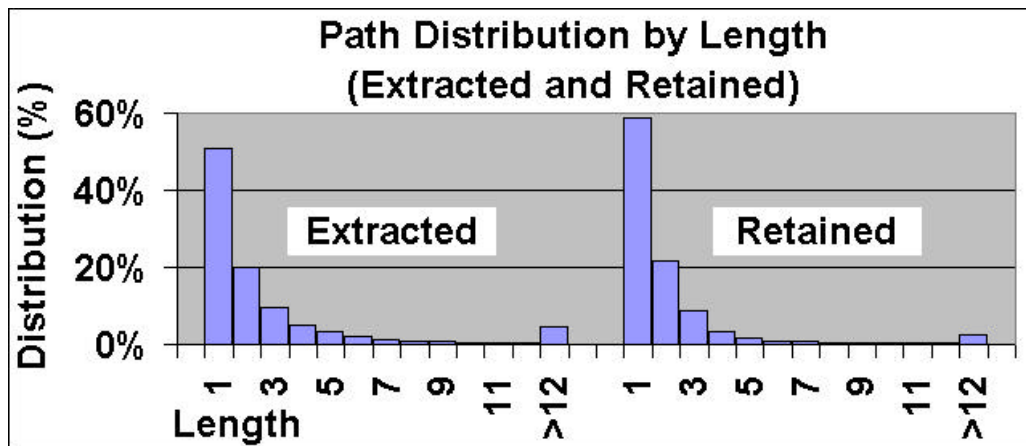


* = k Steps, any elements

Descriptive Statistics

Example: E-Commerce site

	minFreq=2	minFreq= 4	
• Discarded			
🗑️ Robots	22,784	22,784	
🗑️ NFP (404-error) path	84,483	84,483	
🗑️ Infrequent path	1,801,737	2,051,630	
• Extracted full paths	8,134,946	8,134,945	
• Retained full paths	6,333,209	6,083,315	
• Unique full path	158,473	60,620	
• Compression coefficient	51.3	134.2	
• Coverage Ratio	77.8 %	74.8 %	
• Memory usage	19.5 MB	6.55 MB	(1GB original)



Conclusions

- Path Analysis Infrastructure
 - ✍ High Compression
 - ✍ Effective Data Structures
 - ✍ Fast Interactive Data Access
- Answers To Important Business Questions
 - ✍ **Increase conversion**
Identify precisely where and why customers “fall off” in a search or registration process
 - ✍ **Understanding advertising redirects**
Redirect URLs can be essential path elements
 - ✍ **Make promotions more effective**
Discover how to align the site with a given promotion in order to increase the productivity of the promotion
 - ✍ **Discover valuable affinities**
Identify “cross-selling” opportunities -- for services and content as well as products
 - ✍ **Optimize site structure**
Examine how different segments of customers respond to different navigation objects and options presented to them

Note: all algorithms described are available commercially with Accrue G2