


A Model of Web Site Browsing Behavior Estimated on Clickstream Data

Randolph E. Bucklin
Catarina Sismeiro

The Anderson School at UCLA



Overview

- 
- Introduction/Clickstream Research
 - Conceptual Framework
 - Within-site browsing behavior for one web site
 - Examine aspects related to stickiness
 - Modeling Approach
 - Page request decision (binary probit)
 - Page view duration (proportional hazard)
 - Data and Estimation
 - Results and Implications
 - Next Steps

“Clickstream” Research

- Internet Customization
 - Recommendation systems (Ansari et al 2000)
 - Customization of email (Ansari and Mela 2000)
- Site Visit Behavior (Media Metrix)
 - Johnson et al (2000)
 - Moe and Fader (2000, 2001)
- Site Navigation and Browsing
 - Huberman et al (1998)
 - Adar and Huberman (1999)
- Purchasing
 - Moe and Fader (2000)
 - Montgomery et al (2001)

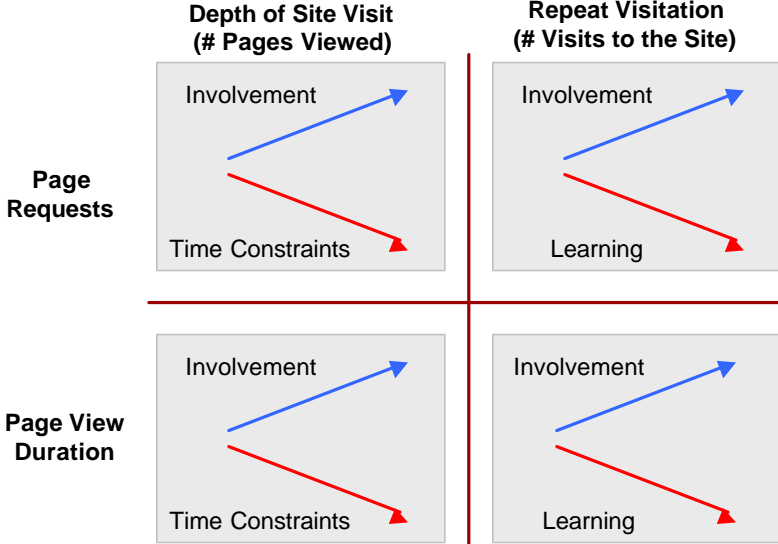
3

Objectives

- Use clickstream data to understand within-site browsing behavior and the implications for site performance (e.g., for site stickiness)
- How is within-site browsing behavior influenced by the following factors:
 - Depth of visit
 - Repeat visitation
 - Other covariates (content, response time, errors)
- What implications might be drawn?
 - Behavioral: Evidence for lock-in and learning
 - Managerial: Site metrics, design and customization

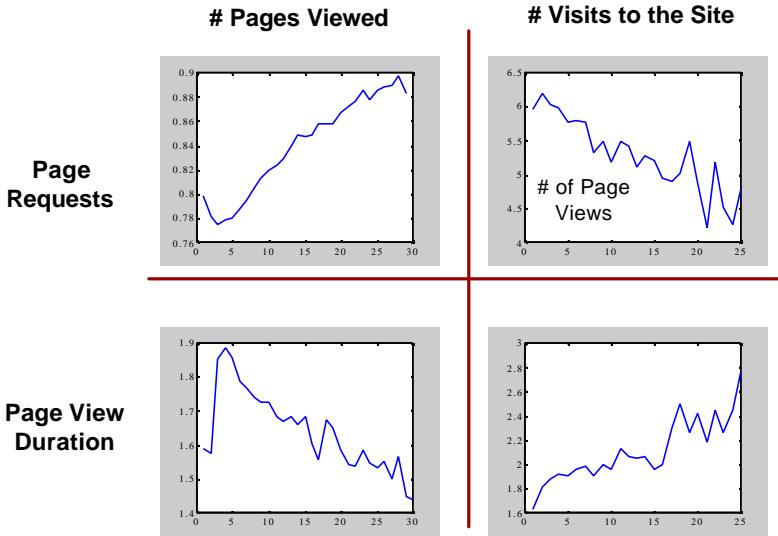
4

Understanding Browsing Behavior



5

Aggregate Statistics



6

Conceptual Modeling Framework

- Model Two Elements of Browsing Behavior at the Individual Level
 - Additional page request versus site exit
 - Page view durations
- Estimate Two Models
 - Page request model: binary probit for probability of staying or exiting the site
 - Duration model: proportional hazard
 - Include heterogeneity and covariates
- What We are *Not* Modeling...
 - Link choices and navigation pathway
 - Effects of specific page content on browsing

7

Page Requests Model

- Binary probit with possible visitor heterogeneity
- The utility of an additional page request j for visitor i is given by:

$$y_{ij} = X_{ij} \mathbf{b}_i + \mathbf{e}_{ij} \quad \forall j = 1 \dots J_i$$

$$\mathbf{e}_{ij} \sim N(0,1) \quad \forall i = 1 \dots N$$

- Where:
 - $y_{ij} > 0 \Rightarrow C_{ij} = 1$ (*additional page request*)
 - $y_{ij} < 0 \Rightarrow C_{ij} = 0$ (*website exit*)
- Random effects approach to heterogeneity

8

Duration Model

- Conditional proportional hazards model with possible visitor heterogeneity
- Weibull baseline hazard

$$h_o(t_{ij}) = I a t_{ij}^{a-1}, \quad t_{ij} > 0, \quad a > 0, \quad I > 0$$

- | | |
|--|---------|
| ▶ Positive duration dependence | $a > 1$ |
| ▶ Negative duration dependence | $a < 1$ |
| ▶ No duration dependence (exponential) | $a = 1$ |

- Gamma heterogeneity (“frailties”) with unit mean

$$\mathbf{w}_i \sim \text{Gamma}(\mathbf{k}, \mathbf{k}) \quad \forall i = 1 \dots N$$

9

Duration Model (cont'd)

- Hazard depends on covariates and expected page view utility
- The hazard of a page view j for visitor i is then given by:

$$h(t_{ij} | \mathbf{w}_i, \bar{z}_{ij}, \bar{y}_{ij}) = h_o(t_{ij}) \exp(\bar{z}_{ij} \mathbf{f}_1 + g(\bar{y}_{ij}, \mathbf{f}_2)) \mathbf{w}_i$$

- ▶ $g(\cdot)$ is a polynomial of the expected utilities
- ▶ Covariates do not include intercept

10



Data

- Online reseller of automotive vehicles
- Data covers month of October 1999
- W3C Extended Log File format
 - Includes cookies - for first hit, cookie was inferred using IP-address and time lag
 - Cookies identify returning visitors and contain information on previous vehicle configurations
- “Hits” aggregated to page views
- Log files also record
 - Bytes transferred and server response time
 - Errors and reloads

11



Data Filtering

- Only page requests with cookies to identify:
 - Unique users
 - Multiple sessions per user (Repeat)
 - Cars configured
- Only sessions with more than one page view
 - Session defined as sequence of page views from a unique cookie
 - If next page requested by user is more than 30 minutes apart, it is assumed to start new session
- Eliminate Web Crawlers, Company Web Administrators, etc.

12

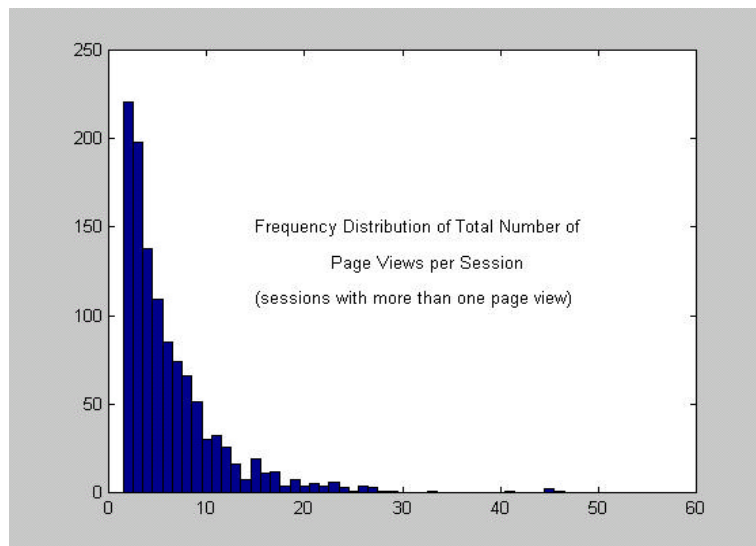
Summary Statistics for Browsing

	Random Sample	
	October 1999	October 15 - 31
Number of Visitors	169,910	5,000
Number of Sessions	245,782	6,630
Number of Requests	1,462,457	40,560
Sessions per Visitor	1.447	1.326
Pages per Visitor	8.607	8.112
Pages per Session	5.950	6.118
Average Page View Duration*	102.692	117.156

*in seconds

13

Distribution of Page Views



14

Predictor Variables

- Visit Depth
 - Cumulative number of page views in the visit at the current stay/exit decision point
- Repeat Visitation
 - Number of previous visits to the site
- Additional Covariates
 - Information sent (bytes between server and client)
 - Server response time
 - Errors occurring on a page download (0/1)
 - Reload occurrence (0/1)
 - Presence of dynamic content (0/1)
 - Previous product configuration (0/1)
 - Previous page view duration (page request model only)
 - Expected utility from page request model (duration only)

15

Summary Statistics for the Covariates

	Mean	Std. Dev.
BYTES	0.275	0.238
CMAKE	0.201	0.817
CPAGE	5.886	6.186
CSESSION	1.921	2.208
DYNAMIC	0.126	0.332
ERROR	0.081	0.273
FIRSTMAKE	0.146	0.354
ORDER	0.005	0.069
PREVDUR	1.633	2.909
RELOAD	0.233	0.423
SRVRESP	0.003	0.041

16

Estimation

- Bayesian estimation with diffuse priors
- Gibbs sampling and Metropolis-Hastings
 - Check for convergence
 - Eliminate first 20,000 draws
 - Save each 4th draw of remaining 8,000
- Log of pseudo Bayes factors for model comparison and selection
 - Computed using the saved 2,000 draws
 - Intercept only model as base case

17

Model Fits

Probit Model			
	Model 1	Model 9	No Heter. Model 9
INTERCEPT	x	x	x
BYTES		x	x
SRVRESP		x	x
RELOAD		x	x
ERROR		x	x
LN(PREVDUR +1)		x	x
LN(CPAGE)		x	x
LN(CSESSION)		x	x
LN(CMAKE + 1)		x	x
Contribution to Log PSBF	-18,111.6	-16,703.8	-17,453.6
Number of Covariates	1	9	9
Log PSBF		1,407.8	658.0

18

Model Fits (cont.)

Duration Model

	No Covar.		Joint Choice & Duration	
	Model 1	Model 7	Model 9	No Heter. Model 10
BYTES		x	x	x
SRVRESP		x	x	x
RELOAD		x	x	x
ERROR		x	x	x
LN(CPAGE)		x	x	x
DYNAMIC		x	x	x
EXPECTED UTILILITY			x	x
Contribution to Log PSBF	-193,627	-192,933	-192,917	-194,013
Number of Covariates	--	6	7	7
Log PSBF		694	710	-386

19

Model Parameters

	Choice ¹		Duration ²	
	95% Probability		95% Probability	
	Mean	Interval	Mean	Interval
BYTES	0.211	[0.126 , 0.308]	-0.351	[-0.400 , -0.287]
LN(CMAKE + 1)	0.186	[0.105 , 0.269]	n.s.	n.s.
LN(CPAGE)	-0.711	[-0.755 , -0.668]	-0.027	[-0.046 , -0.008]
LN(CSESSION)	-0.194	[-0.236 , -0.151]	n.s.	n.s.
DYNAMIC	n.s.		0.544	[0.479 , 0.598]
ERROR	0.427	[0.317 , 0.542]	-0.246	[-0.303 , -0.188]
LN(PREVDUR + 1)	-0.220	[-0.250 , -0.189]	n.a.	n.a.
RELOAD	-0.050	[-0.102 , -0.006]	-0.241	[-0.274 , -0.208]
SRVRESP	-8.485	[-9.434 , -7.445]	-1.883	[-2.072 , -1.686]
INTERCEPT	1.885	[1.834 , 1.943]	n.a.	n.a.
EXP UTILITY	n.a.	n.a.	0.071	[0.047 , 0.093]
?	n.a.	n.a.	0.988	[0.980 , 0.996]
?	n.a.	n.a.	4.993	[4.635 , 5.360]
?	n.a.	n.a.	0.012	[0.012 , 0.013]

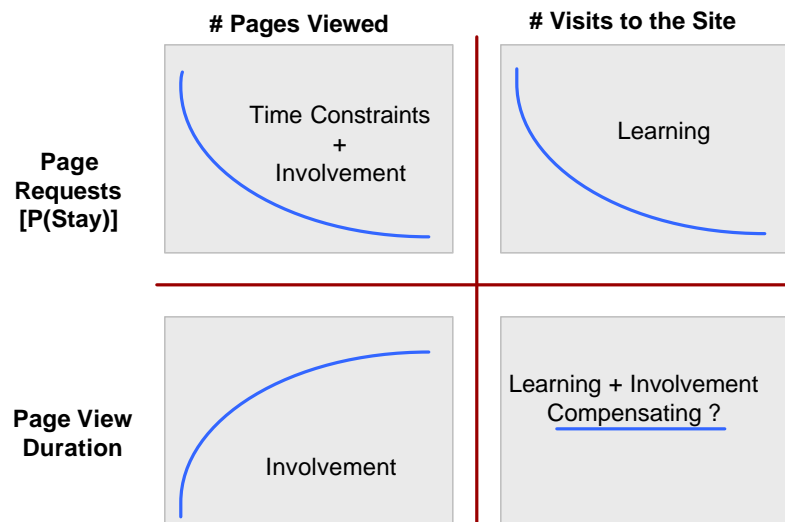
¹ cross-sectional means; ² model parameters; n.a. = not applicable; n.s. = not significant ²⁰

Individual Level Elasticities

	Choice			Duration - Direct			Duration - Total		
	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
BYTES	-0.184	0.136	0.010	0.001	0.665	0.105	0.001	0.667	0.101
CMAKE	-0.064	0.374	0.067	---	---	---	-0.049	0.007	-0.016
CPAGE	-1.212	-0.000	-0.238	0.026	0.026	0.026	0.024	0.109	0.076
CSESSION	-0.358	0.027	-0.067	---	---	---	-0.012	0.039	0.015
DYNAMIC	---	---	---	-0.562	0.000	-0.094	-0.562	0.000	-0.094
ERROR	-0.009	0.127	0.002	0.000	0.257	0.025	0.000	0.245	0.023
PREVDUR	-0.524	0.077	-0.090	---	---	---	-0.020	0.046	0.016
RELOAD	-0.166	0.050	-0.016	0.000	0.239	0.057	0.000	0.260	0.058
SRVRESP	-8.418	0.027	-0.044	0.000	0.914	0.007	0.000	0.970	0.008

21

Findings for CPAGE and CSESSION



22

Additional Behavioral Findings

- A Possible Dynamic Tradeoff Between Page Views and Duration
 - Higher expected utility from another page view implies shorter expected duration of next page view
 - Higher previous page view duration lowers page request probability
- Cross Sectional Patterns
 - Users who are more sensitive to CPAGE and PREVDUR also have shorter page view durations (as revealed by correlations of the individual-level parameter estimates with the frailties)
- Costs vs Benefits

23

Implications for Managers

- Customization
 - Significant interest in customization of Internet experiences among managers and academics (e.g., Ansari and Mela 2000, Adar and Huberman 1999)
 - Effects consistent with across-site learning suggest limits on the benefits of site customization
- Site Design
 - Question of clutter versus clarity
 - Information per page
 - Number of page views needed to complete transaction
 - Findings for bytes per page and cumulative page views supports move to fewer pages and more information per page (site condensation)

24



Implications for Managers (cont'd)

- Site Metrics
 - Aggregate-level metrics of site stickiness differ from individual-level modeling results
 - Mix of new and returning visitors can change site usage and affect metrics

25



Next Steps

- Combine analysis of browsing behavior with prediction of purchase
- Diagnose site-related drivers of order decisions
- Develop a sequential model approach
 - Order conditional upon product configuration
 - Product configuration conditional upon search
 - Search conditional upon initial browsing experience
- Preliminary analysis reveals differences between those that configure or order a car and those that do not

26