UCLA
Department of Statistics Papers

Title
Bayesian Hierarchical Model of the Browsing Behavior of World Wide Web Users

Permalink
https://escholarship.org/uc/item/75q8s003

Authors
Sanchez, Juana
Liu, Ching-Ti

Publication Date
2004

Bayesian Hierarchical Model of the Browsing Behavior of World Wide Web Users[*]

Juana Sanchez and Ching-Ti Liu
UCLA Department of Statistics

Abstract

We consider the case of surfing within a single large Web site, which is important from the point of view of site design, web server proxy efficiency and search engine optimal ranking of pages. The site used as an example to illustrate a method for clustering user sessions that we propose is msnbc.com. We use a random sample from a publicly available server log data on the Web pages chosen by 989818 users in a twenty-five hour period, where the response measure for each user is an ordered sequence of choices among 17 categories (UCI KDD Archive). A common way to model the browsing behavior of users is to assume that the decision of users is a random walk with a probability distribution of first passage time to a threshold that is a two-parameter inverse-gaussian distribution. Another hypothesis examined in the literature is that users at each page conduct an independent Bernoulli trial to make a stopping decision, which implies a geometric distribution. Mixtures of first-order Markov processes or model-based clustering with and without a Bayesian flavor have offered very useful exploratory data analysis. All these studies have shown evidence that web-surfing behavior may be non-Markov in nature and have illustrated how hard it is to capture dependencies in the data. The performance of the models over a wide range of Web Site formats is still inconclusive. This performance has been measured by the ability to predict page hits, by the resulting distribution of page hits, and by the contribution to efficient web caching schemes. Some models have been tested with server log data of AOL or similar Sites and others have been tested within a single Web site like msnbc.com. The levels of aggregation of pages and clustering of user behavior have also varied within studies. In this paper, we assume that for the case of browsing within a news portal like msnbc.com, where contents are continually changing, the server-log data is only meaningful when categories are aggregated, like they are for the msnbc.com data set, and the order of the browsing may not be relevant. We use a Bayesian hierarchical model of the page counts per user to obtain posterior distributions of page access frequency that allow us to cluster user sessions in a relatively small number of groups. The model has the ability to have enough parameters to fit the data well, while using a population distribution that can structure dependence in the parameters. The model can be generalized to different types of Web sites, different levels of aggregation of pages and different clustering schemes.

1. Introduction

Web usage mining, an area of web mining, focuses on analyzing visiting information from server-log data in order to understand the browsing behavior of users. Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis.

Past studies have focused on finding patterns in the sequences of pages visited by a user and on the prediction of future moves of users based on those patterns. For example, Cadez et al. (2003) use web-server logs of individual browsing records for a sample of users at the msnbc.com site. This is the same data set that we analyze in this paper. They model the data as having been generated in the following fashion: (a) A user arrives at the web site and is assigned to a particular cluster with some probability, and (2) the behavior of that user is then generated from a markov model with parameters specific to that cluster. They pretend that this model generates the web data observed, and that we only see the user behaviors and not the actual cluster assignments. Then they use a standard learning technique, the Expectation-Maximization (EM) algorithm, to learn the proportion of users assigned to each cluster as well as the parameters of each Markov model. This is a model based approach to clustering also known as mixture model. Its advantage

---

[*]Address all correspondence, including requests for data or programs, to Juana Sanchez at `jsanchez@stat.ucla.edu`.

is that sequences of different lengths may be assigned to the same cluster. The performance of the model is measured by how accurately it assigns out of sample users to their cluster using the maximum posterior probability of a sequence belonging to a cluster. Different number of clusters are used, and the number that gives the best out of sample prediction as measured by a log-score is the chosen one. Once the users have been assigned to clusters by the model, the clusters are visualized using software they designed. The aim is mostly exploratory.

Sen and Hansen (2003) explore first and second-order Markov models, mixtures of first order Markov models with pre-specified number of clusters, and a Bayesian first order Markov model for all users with a prior distribution for the transition matrix. The authors find the latter approach to be the most promising. They use link information to model navigation. The posterior probabilities are then used to predict a visitor's next request. The web site used to test the models is cm.bell-labs.com.

Huberman et al. (1997) explored also the suitability of the first order Markov model for log data from several web sites. They were trying to determine whether the same web surfing law, namely the first order Markov model, applies to all sites. Their paper concludes that it does, but their results are not conclusive.

Each of the authors mentioned above have different objectives. Cadez et al want to cluster users, Sen and Hansen want to predict a user's next move, and Huberman et al. want to find a law to characterize users' behavior. Other authors, too many to mention them all here, have been concerned with other aspects of web browsing.

The objective of our paper is to present a simple methodology that allows us to discover clusters of users by matching posterior distributions of page access frequencies. In the spectrum of previous research conducted by other authors, our scope is closer to that of of Cadez et al. We take into account the enormous variability across users by introducing a random effect for each user and another random effect for user interacting with page. The homegeneity among users is expressed in the common posterior distributionof of page access frequency for users belonging to known groups. Groups whose posterior probabilities of page frequencies overlap are considered to belong to the same cluster.

We use publicly available pre-processed server-log data from the msnbc.com site. Thus our paper is not concerned with the many engineering issues involved in the preparation of the server-log data for analysis.

The random sample and the programs used for the analysis that we present below are available from the first author upon request.

2. The Data

The original data comes from Internet Information Server (IIS) logs for msnbc.com and news-related portions of msn.com for the entire day of September 28, 1999 (Pacific Standard Time) (Heckerman, UCI KDD Archive). The reader will understand the description that follows better if s/he opens the msnbc.com web Site before reading. Keep in mind, though, that since 1999 the structure of the Site has changed slightly. The representation of the processed server-log data is fairly abstract: (a) the server-log files have been converted into a set of sequences, and one sequence for each user session, (b) each sequence is represented as an ordered list of discrete symbols (numbers), and (c) each symbol represents one of 17 categories of web pages requested by the user. The 17 categories correspond to sets of Uniform Resource Locators (URLs) on the site. Each sequence in the dataset corresponds to page views of a user during that twenty-four hours period. Since there are 989818 users, there are 989818 sequences. Each symbol in a sequence corresponds to a user's request for a page. Requests are not recorded at the finest level of detail–that is, not at the level of URL, but rather, they are recorded at the levels of page category (as determined by a site administrator). The categories (and their corresponding symbols) are 1="frontpage", 2="news", 3="tech", 4="local", 5="opinion", 6="on-air", 7="misc", 8="weather", 9="health", 10="living", 11="business", 12="sports", 13="summary", 14="bbs (bulletin board service)", 15="travel", 16="msn-news", and 17="msn-sports". The number of URLs per category ranges from 10 to 5000. Page requests served via a caching mechanism were not recorded in the server logs, and hence, are not present in the data (Heckerman, UCI KDD Archive). As an example, we write below the sequences for the first three user sessions in the data set (one line per user):


```
User session 1: frontpage, frontpage
User session 2:  news
```

Depth	Percentage of user sessions
1	60.75%
2	21.66
3	9.56
4	4.37
5	1.99
more than 5	1.63

Table 1: Depth of a user session within the msnbc.com web site

```
User session 3:   tech,news,news,local,news,news,news,tech,tech
```

or, symbolically,

```
User session 1:   1, 1
User session  2:   2
User session 3:   3, 2, 2, 4, 2, 2, 2, 3, 3
```

The above is saying that the first user session entered the msnbc.com web site via the frontpage, and hit two links in the frontpage. User session 3, however, started in the tech page, moved to the news page and hit two links there, then to local, then to news again and hit two links there, then to tech again, and hit two links there. The reader should visit the msnbc.com web site to understand that the user can move from any page category to another due to the frame structure of that web Site. This aspect makes this web Site different from Sites used in other papers.

Previous cluster visualization of these data by Cadez et al.(2003) have considered the last category in the sequence the "end state." Following Sen and Hansen (2003) we attach an additional state at the end of each sequence, the "exit" state, with symbol 18. Once in it, the user can not return to any of the other states unless it starts another sequence. For the user sessions above then, the sequences are:

```
User session 1:   frontpage, frontpage, exit
User session 2:   news, exit
User session 3:   tech,news,news,local,news,news,news,tech,tech, exit
```

or, symbolically,

```
User session 1:   1, 1, 18
User session 2:   2, 18
User session 3:   3, 2, 2, 4, 2, 2, 2, 3, 3, 18
```

We distinguish in our paper between the length of a user session and the depth of a user session. We define the depth of a user session as the number of different page categories per session, excluding the exit. For example, user session 1 above only has 1 unique page category so its depth is 1, user session 2 also has one unique page category, with depth=1, and user session 3 has 3 distinct page categories and has depth=3. Table 1 summarizes the most common values for the depth variable. Most user sessions (60.75%) visit only one page category. 98.33% of user sessions visited 5 or less different pages. Thus visitors that browse through more than 5 pages in one session are rare but nevertheless they comprise 1.63%.

The length of a visit, on the other hand, is defined as the total number of words per session, including repeats but excluding the exit. For example, user session 1 has length 2, and user session 2 length 1, whereas user session 3 has length 9. The average length per session is 4.747, the median is 2, the 1st quartile is 1, the third quartile is 5, the 80th percentile is 7, the 90th percentile is 11, the 97th percentile is 20, the 99th percentile is 31. So only 1 percent of the users have length larger than 31. However, beyond this number there are some very high values of length that make the distribution very skewed to the right. Table 2 summarizes the distribution of the length variable.

The users tend to have vastly different web-surfing patterns in terms of the length and the depth of a session. Similarly, they tend to enter the Web site from very different page categories. We call the entry point "gate." Table 3 summarizes features of each page category, i.e., for how many users was that the

Length (L)	Percentage of user sessions
1	36.91%
$1 < L \leq 2$	15.51
$2 < L \leq 5$	22.62
$5 < L \leq 7$	7.88
$7 < L \leq 11$	8.14
$11 < L \leq 20$	6.06
$20 < L \leq 31$	1.85
$L > 31$	0.99

Table 2: Distribution of the length of a user session within the msnbc.com web site

Page Categories	Percentage entering msnbc via	Percentage of sessions who visited	Percentage of total length (total=4698794)
frontpage (1)	28.02	34.15	20.01
news (2)	7.79	26.67	9.62
tech (3)	6.72	19.84	4.41
local (4)	5.15	23.04	8.21
opinion (5)	0.42	5.08	3.22
on-air (6)	16.47	22.08	8.83
misc (7)	0.35	9.06	6.50
weather (8)	7.25	9.65	9.35
msn-news (9)	6.59	9.11	4.18
health (10)	1.29	5.11	2.80
living (11)	1.44	5.81	2.06
business (12)	5.73	11.33	5.63
msn-sports (13)	6.36	7.77	4.59
sports (14)	5.41	12.36	8.42
summary (15)	0.74	2.95	1.20
bbs (16)	0.04	0.21	0.53
travel (17)	0.15	1.11	0.36

Table 3: Relative importance of each page category according to different variables

gate of entry into the msnbc.com site, the percentage of visitors that visited the page at least once, and the percentage of the total length that comes from that page viewing. For example, 28.02% of user sessions entered the msnbc.com web site through the front page. The front page was also the one most frequented.

It is also interesting to see the first order Markov transition matrix derived from the data. It helps see the frequencies with which user sessions move from the page they are at to the next page. This can be seen on table 4.

We can see in the transition matrix that users tend to move from a link in a page category to another link in the same page category more often than any other move. It is also worth mentioning that regardless of the page category in which a user finds her or himself, the next most preferred page category is page 1, the front page.

If, with this matrix, we simulate sequences of page visits or sessions, we find that the distribution of the length of a session derived from this matrix, does not fit very well the distribution of the length of a session for the data, particularly on the tails. We take this as an indication that a Markov model of first order is not a good model for these data. This is consistent with the findings of Cadez et al.(2003) for these data and the findings of Sen and Hansen (2003) for the Bell Labs data.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	44.5	7.8	2.5	2.9	0.9	3.3	3.0	0.9	0.3	2.0	2.9	4.4	0.2	3.8	0.6	0.1	0.5	19.3
2	10.1	49.1	2.4	2.5	0.7	2.1	1.4	1.5	0.5	1.7	0.8	2.2	0.1	1.9	1.3	0.0	0.2	21.6
3	8.0	3.8	34.7	1.7	0.7	1.4	0.4	0.6	2.3	1.2	1.4	2.0	0.4	1.3	1.3	0.0	0.3	38.5
4	5.6	2.9	0.9	58.2	0.4	1.3	5.6	1.0	1.9	0.7	0.6	0.8	0.6	1.1	0.2	0.0	0.1	18.0
5	3.8	1.8	0.6	0.5	78.6	1.5	0.2	0.6	0.9	0.4	0.5	0.4	0.1	0.4	1.2	0.2	0.0	8.3
6	4.4	2.6	1.4	1.7	0.7	34.7	8.9	1.1	1.8	1.6	0.7	1.0	0.7	0.8	2.8	0.1	0.0	35.1
7	8.3	1.1	0.3	10.8	0.1	9.6	58.1	0.3	2.4	0.7	0.1	0.3	2.6	0.9	0.3	0.0	0.0	4.1
8	1.4	1.9	0.3	0.8	0.2	0.7	0.6	74.9	0.7	0.2	0.2	0.3	0.4	0.5	0.1	0.0	0.0	16.6
9	5.1	2.1	2.7	5.1	1.4	1.9	4.9	2.4	41.0	0.6	1.0	2.1	2.6	0.3	0.0	0.0	0.0	26.8
10	9.1	5.5	2.1	1.4	0.6	3.5	1.5	0.6	0.9	51.5	1.2	1.7	0.1	0.9	1.7	0.0	0.2	17.5
11	19.6	4.1	1.9	2.4	1.6	3.1	0.7	0.7	1.3	1.7	31.5	1.5	0.3	1.8	1.5	0.0	3.0	23.2
12	10.3	3.2	2.1	1.3	0.4	1.1	0.6	0.5	2.1	0.9	0.7	49.0	0.4	2.0	0.5	0.0	0.1	24.8
13	1.3	0.1	0.3	0.8	0.0	0.7	3.9	1.1	1.7	0.0	0.1	0.3	55.1	8.4	0.0	0.0	0.0	26.0
14	6.7	1.2	0.6	0.8	0.1	0.6	0.8	0.4	0.4	0.2	0.3	0.7	2.5	63.8	0.2	0.1	0.1	20.4
15	6.2	8.8	4.3	1.8	4.4	11.6	3.6	0.8	0.4	5.0	3.8	2.0	0.1	2.4	23.8	0.1	0.3	20.8
16	2.7	0.9	0.1	0.2	0.6	1.1	0.1	0.1	0.1	0.2	0.1	0.2	0.0	1.6	0.2	87.7	0.0	4.1
17	21.8	4.6	2.2	2.0	1.1	2.2	3.2	0.8	0.4	1.7	9.5	1.6	0.1	1.5	0.9	0.0	26.3	20.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Table 4: First Order Markov transition matrix (Percent)

Minimum	1
1st quartile	1
Median	2
Mean	4.632
3rd Quartile	6
Maximum	99

Table 5: Summary of the length of a session in the random sample

2.1. Simple random sample of sessions

We process the data set summarized above further to make it suitable for the model that we want to fit. Then we draw a simple random sample of 5000 sessions from the newly processed data set. The random sample used for the analysis that follows is available from the first author upon request.

We show below the new structure of the data that we use for the model fit. The user sessions in the example below are the first three user sessions described earlier:

```
User session 1: 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 2 1
User session 2: 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 1 1
User session 3: 0 5 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 3 9 3
```

The first 17 columns are the length in each of the 17 page categories or number of times during the session that the visitor hit that page category; notice that there are many zeros to reflect that some user sessions did not visit that particular page category. The last three columns are, respectively, gate, total length for the session and depth.

The structure of the data set that we have now requires, at the time of estimation of statistical models, that we use a method that takes that sparsity into account. Next section explains that model.

The total length of a session in the random sample is summarized in Table 5.

For purposes of illustrating our methodology for clustering users, a sample size of 5000 is sufficient. However, to simplify the illustration, in this paper we discretize the variables more and reduce the final data matrix to the following structure:

```
1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1  1
0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0  1
0 5 3 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0  3
```

As before, the first 17 columns are the length for each page category. Column 18 is a dummy variable with 1 representing that the user entered msnbc.com via the frontpage and a 0 representing that the visitor entered via some other page. Column 19 represents the depth of the visit, and we keep it in the range 1-17. We don't use the total length column in the model described below.

3. Statistical Model

In the model that we propose in this paper, a unit of observation is a visitor's session in the msnbc.com web site. We have 17 measurements for each user session, so we have a multivariate clustered response for each user session, $y_i = (y_{i1}, y_{i2}, ..., y_{i17})$. Each measurement y_{ij} in the clustered response is the number of times the visitor i hit the jth page category in the msnbc.com web site, and is modeled as a **Poisson** with mean μ_{ij} that depends on several covariates and random effects through the log link function.

The responses observed in our data are more heterogeneous than the Poisson model assumptions, which leads us to believe that the systematic differences between visitors' sessions are not attributable to random variation only. That is, there is overdispersion due to these systematic (non-random) differences that needs to be taken into account in our model. The overdispersion could also be caused by the tendency of the measurements from the same visitor to be correlated. And that also needs to be taken into account.

Thus we consider a Generalized Linear Mixed Model (GLIMM) with random effects for each visitor's session and page-session interaction where the distribution for the random effects incorporates the extra

variability between sessions and within each session's clustered responses, while common regression parameters reflect that all visitors have something in common. This makes prediction possible.

Random effects modeling allows estimation of visitor's session specific parameters which borrow strength from data on all sessions, yet do not constrain every visitor to follow an identical trajectory.

The model we use for the msnbc.com data does not make all the parameters in the regression model random effects. Instead, we provide a model with both fixed and random effects. The log link function is

$$\log(\mu_{ij}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + b1_i + b_{ij}$$

where the x_i's are gate (gate=1 for frontpage and 0 for other pages), depth (1-17) and a dummy equal to 1 for $j = 1$, the frontpage, and equal to 0 for $j \neq 1$, the other pages. This is justified by the prominent position that the front page occupies in the browsing behavior of users. All the descriptions done in the data summary point to the conclusion that the front page is the most visited, it is also the most common gate, and has been visited by more visitors than any other page.

For random effects, we have $b1_i$ for extra variability between individual visitors'sessions, and b_{ij} for a random interaction term of visitor's session times page to model extra-Poisson variability within a session. The only difference betwen the fixed effect and random effect parameters is that the former's covariances are constants expressing subjective prior knowledge, while the latter's covariance matrix depends on unknown hyperparameters $\theta = (\tau_{b1}, \tau_b)$ that must be estimated from the data. A hyper-prior distribution of the θ must then be specified. This makes our model a hierarchical model. We adopt independent noninformative priors for the fixed effects parameters and for the θ. With these specifications, the complete model is then

$$y_{ij} = \frac{\mu_{ij}^y}{y_{ij}!} e^{-\mu_{ij}}$$

$$log\mu_{ij} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + b1_i + b_{ij}$$

To make the prior distributions of the fixed effects noninformative, we assign a Normal prior distribution with prior precision very small, close to 0.

$$\beta_i \sim N(0.0, 0.0001)$$

The prior distributions of the random effects parameters are:

$$b1_i \sim N(0.0, \tau_{b1})$$

$$b_{ij} \sim N(0.0, \tau_b)$$

Then τ_{b1} and τ_b, the precision parameters of the random effects, are also given noninformative priors (flat priors) which we achieve by assigning very small value to the Gamma parameters.

$$\tau_{b1} \sim \Gamma(0.001, 0.001)$$

$$\tau_b \sim \Gamma(0.001, 0.001)$$

$$\sigma_{b1} = \frac{1}{\sqrt{\tau_{b1}}}$$

$$\sigma_b = \frac{1}{\sqrt{\tau_b}}$$

Notice that we assume that there is no correlation between the parameters of the linear model at all. That assumption can be broken in more complex expressions of this model. Models similar to the above one have been applied to other data sets (Breslow and Clayton (1993), Clayton (1996)). The theoretical Bayesian foundations of these models and the MCMC computation for them can be found in Zeger and Karim (1991) and Clayton (1996). In this paper, we use the software BUGS to estimate the model. BUGS is very slow for this kind of models. Therefore, finding our results took one day of computation with 3000 runs burn in and 6000 iterations. Obviously, more efficient computational approaches need to be obtained

Coeff	Mean	Sd	2.5 %	97.5 %	Median
Intercept	-5.136	0.045	-5.219	-5.046	-5.138
Gate slope	-0.083	0.035	-0.153	-0.014	-0.084
Depth slope	0.646	0.011	0.624	0.667	0.646
page1 slope	2.150	0.051	2.049	2.250	2.149
sd b1 r.e	0.030	0.008	0.017	0.048	0.029
sd b r.e	2.331	0.020	2.291	2.371	2.332

Table 6: Fixed effects and random effects coefficients. Posterior distribution characteristics.

if we want to handle large samples like this and scalability issues have to be addressed to handle even larger samples.

We are interested in the posterior probability distributions for the fixed effects parameters and the random effects parameters. These can be seen in table 5 and Figure 1. For the purposes of predicting behavior of users, the most interesting posterior distributions are those for the mean μ, for different values of the covariates.

The random effects allow us to decompose the variability observed in the data into (a) variability in the counts per page due to the variability within a visitor's session cluster and (b) variability between clusters of individual visitors' sessions. This is the variability around the overall mean.

With the above model specification, the main questions we try to answer with the posterior distributions for the μ of different groups are: Do distributions of counts (or length) for the frontpage and other pages differ across those who enter via the frontpage and those who don't, and across different depths of sessions? How much group overlap is there in the distribution of counts for page 1 and other pages? Can we cluster different groups according to the overlap of their posterior distributions for counts?

4. Results

The results of running the Bayesian model described above are given in Table 6, which contains the estimates of the model parameter estimates. Their posterior distributions are plotted in Figure 1.

As we can see in Table 6, the posterior mean for the standard deviation of the variability across sessions (sd b1) is almost zero, indicating that, overall, there is no significant heterogeneity among sessions. But the interaction random effect variability (sd b) is pretty high, 2.331 suggesting that page by page, there is a significant difference across sessions. All parameter estimates have very small Sd, suggesting strong evidence for the effects observed. Notice that the posterior distribution for the b1 random effect is skewed right.

For the purpose of clustering groups of users into homogeneous clusters, the most important distributions are those that predict the length of visit to frontpage for different groups, and the length of visit to "other pages." That is, we need the posterior distributions for the mean length μ of the Poisson model for different levels of the covariates. We present the results for these in Table 7. In this table, the groups are characterized by the covariates on the left hand column. The dependent variable by which we are clustering these groups are mean length for "frontpage" or mean length for "other page," and we indicate these in the middle of the top row for each group. Figure 2 summarizes the posterior 95% intervals for the μ, mean length, for all the groups we consider in Table 7 (in total, 20 groups). Notice that we are only including groups with depth up to 5, to illustrate our method and to make the graph easier to interpret. The depth variable ranges from 1 to 17.

From the posterior 95% intervals for the different groups (Figure 2, Table 7), we can predict that those who enter the msnbc.com site via the front page (the two groups of five confidence intervals on the left), and those who enter via other pages (the two groups of five confidence intervals on the right), have similar posterior distributions for the average length of stay on the front page and the average length of stay at other pages. This is the case for each of the depth levels considered. There are very extreme values of frontpage visits for those entering through other pages and depth 5, indicating that some of those entering msnbc via other pages and browsing longer do so mostly on page 1. On the other hand, the 95% posterior intervals for average at other pages is similar for the groups entering the site via the front page and via other pages.
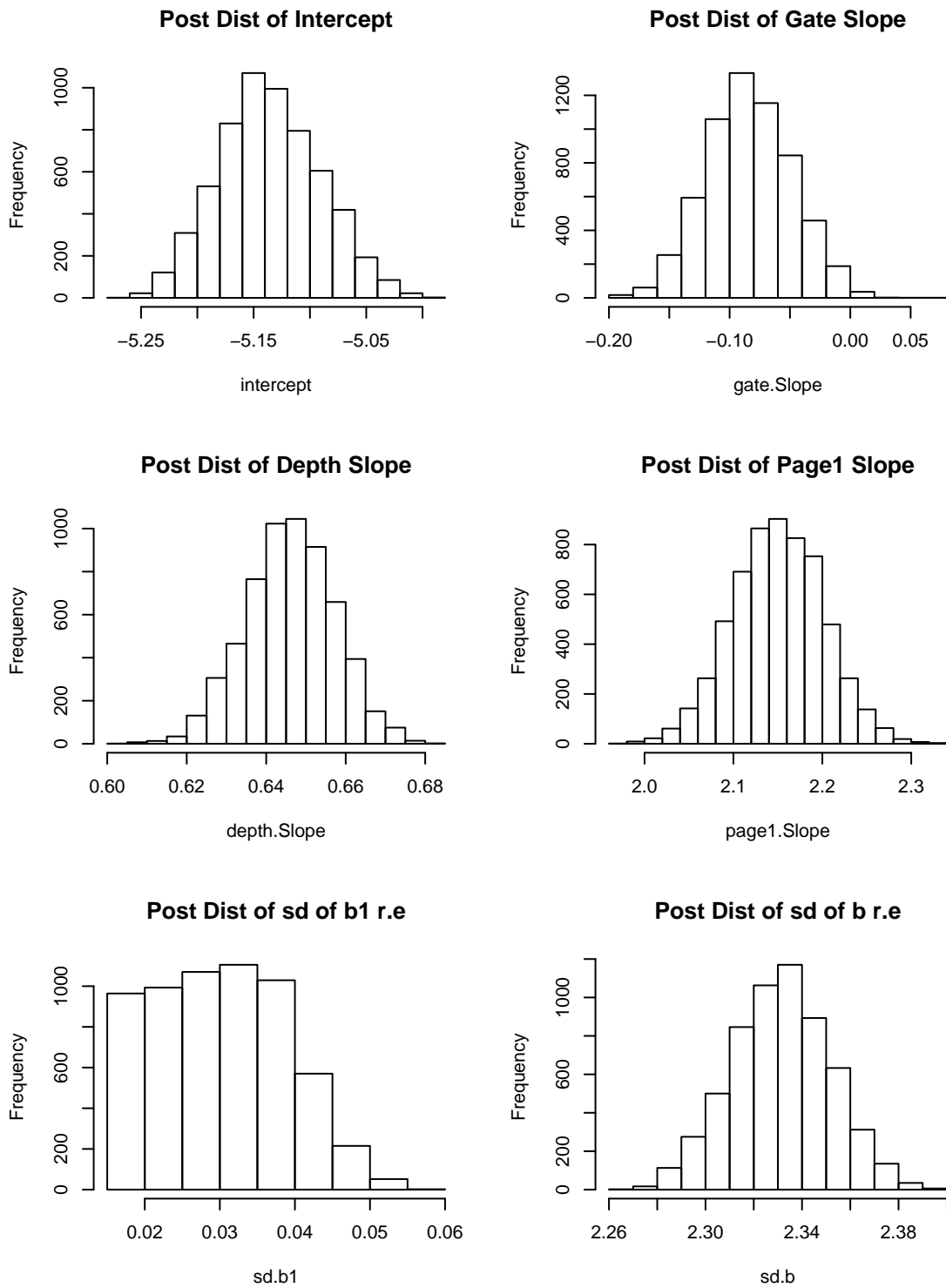
Figure 1: Posterior distributions of the fixed effects and random effects parameters

	Mean	Sd	2.5 perc	97.5 perc	median
gate 1			frontpage		
depth 1	1.753	38.06	0.0006	7.315	0.066
depth 2	2.010	14.01	0.0015	13.45	0.137
depth 3	3.540	22.21	0.0026	25.32	0.024
depth 4	7.534	105.6	0.0039	46.72	0.462
depth 5	12.32	77.13	0.0092	75.83	0.910
gate 1			Other pages		
depth 1	9.895E-2	6.105E-1	7.693E-5	6.978E-1	7.644E-3
depth 2	2.488E-1	1.991E+0	1.270E-4	1.566E+0	1.492E-2
depth 3	6.513E-1	2.204E+1	2.621E-4	2.624E+0	2.925E-2
depth 4	7.062E-1	5.346E+0	5.962E-4	4.953E+0	5.386E-2
depth 5	2.021E+0	3.449E+1	9.485E-4	8.967E+0	1.022E-1
gate ≠ 1			frontpage		
depth 1	1.067E+0	6.259E+0	8.382E-4	7.515E+0	7.688E-2
depth 2	1.965E+0	2.348E+1	1.382E-3	1.332E+1	1.403E-1
depth 3	3.468E+0	1.860E+1	2.576E-3	2.609E+1	2.563E-1
depth 4	7.306E+0	4.902E+1	5.267E-3	5.060E+1	5.350E-1
depth 5	1.688E+1	1.794E+2	9.841E-3	9.925E+1	9.679E-1
gate ≠ 1			Other pages		
depth 1	1.243E-1	9.085E-1	8.535E-5	7.782E-1	8.493E-3
depth 2	3.792E-1	1.118E+1	1.510E-4	1.619E+0	1.644E-2
depth 3	4.494E-1	3.274E+0	3.567E-4	2.771E+0	3.086E-2
depth 4	1.186E+0	1.832E+1	6.254E-4	6.331E+0	6.062E-2
depth 5	1.575E+0	1.302E+1	1.289E-3	1.085E+1	1.189E-1

Table 7: Predictions of length of visit to the frontpage and "other pages" for different groups, where a group is characterized by the gate through which the user session starts and by the depth. The numbers in the table summarize the posterior distributions of average length for the different groups depicted.
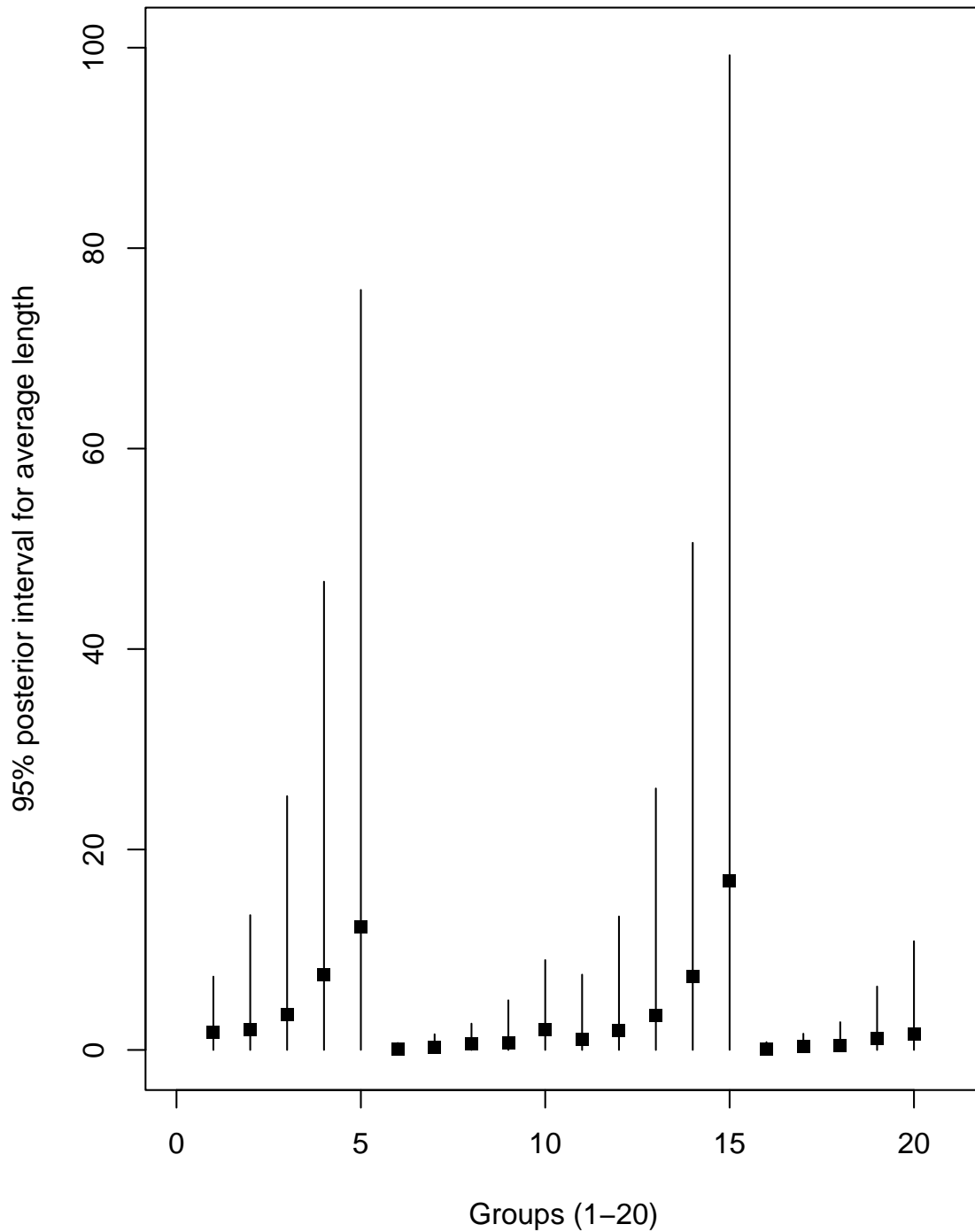
Figure 2: 95 percent posterior intervals for average length. From left to right, the first group of five intervals correspond to length of visits to frontpage for the group with gate=1, depths 1(first interval)–5(last interval); the second group of five is length of visits to "other pages" when gate= 1, depth 1–5; the third group of five is length of visits to frontpage when when gate ≠ 1, depth 1–5; last group of five is length for other pages, when gate ≠ 1, depth 1–5. See Table 7 for the numerical details of each interval.

Visitors who tend to visit a lot of pages (high depth) spend much more time in the front page than in other pages, regardless of which is their gate of entry into msnbc.com.

How can we use information like this to cluster user sessions? The criterion we propose is to cluster sessions according to the posterior distributions of the average length like those above. If the information presented here was the only information we had, we could do different levels of clustering. A very broad clustering, i.e., two clusters only, would contain in one cluster all user sessions that enter msnbc.com via the frontpage,and in the other cluster all user sessions that enter msnbc.com via other page categories (sport, local,...etc.). But the posterior distributions for average length included in these clusters vary by depth, suggesting that a finer level of clustering could be achieved by getting 17 cluster, one for each depth level.

As we can see in Figure 2, the cluster with depth=1 has four posterior distributions, one for sessions entering via frontpage and visit to frontpage, another for sessions entering via frontpage and visits to other pages; a third for sessions entering via other pages and visit to frontpage and a fourth for sessions entering via other pages and visit to other pages. This includes a lot of different behaviors of users within one cluster, which is consistent with the findings of Cadez et al.

Of course, the clustering could be even more specific and more interesting if we had included in the model specific indicators for each of the "other pages" (1-17) and if we had included indicators for each of the "other gates" (1-17). The number of variables in the model would increase considerably and the amount of time that obtaining the results would take would be unbearable using BUGS. Much faster algorithms and scalability issues need to be solved to do that.

5. Conclusions and Ideas for Future Work

The methodology we have presented in this paper to cluster visitors to a web site based on the posterior distributions of the length of a user session for different groups, allows us to derive very general clusters and conclusions about the behavior of users.

Our conclusions allow us to explain, using covariates, why Cadez et al.. got clusters that included session of very different lengths. Having the covariates to explain the difference in the posterior distributions of average length allows us to cluster sessions into more homogeneous groups and to understand what explains the heterogeneity within a cluster.

The results also explain indirectly why Markov models may not be good models for predicting sessions' moves. These models fail to account for the extraordinary variability reflected in the posterior distributions of the average length of sessions.

We illustrated our method in this paper by using a simplified version of the hierarchical model and using a relatively small random sample. We have indicated how to make the model more detailed by including more variables. The model can also be expanded to account more in detail for the long tails of the distributions. For example, the priors for the random effects could be made t-distributions with low degrees of freedom, or alternatively, they could be mixtures.

To use larger samples and bigger models the computational issues have to be solved first. Faster algorithms and scaling need to be resolved for this method to become of practical use by web administrators. We are working on these two issues and our results will be presented in a future paper.

Web administrators can use a model like this to determine the groups of users in the site, and to design the web page accordingly. The posterior distributions of one day could be use as the prior distributions of the next day to update the information daily and see how users' behavior changes over time.

References

[1] Breslow, N.E. and Clayton, D.G. (1993). Approximate Inference in Generalized Linear Mixed Models. JASA March 1993, Vol. 88, No. 421. p 9-25.

[2] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S. (2003). Model-based clustering and visualization of navigation patterns on a Web site. Accepted for publication on Journal of Data Mining and Knowledge Discovery, 7(4).

[3] Clayton, D.G. (1996). Generalized Linear Mixed Models. Markov Chain Montecarlo in Practice. Editors: W.R. Gilks, S. Richardson, and D.J. Spiegelhalter. Chapman and Hall. p. 275-298.

[4] Hansen, M.H. and Sen, R.(2003). Predicting Web User's next access based on log data. Journal of Computational and Graphical Statistics, Vol 12, No. 1, March, p. 143-155.

[5] Heckerman, D. The UCI KDD Archive (`http://kdd.ics.uci.edu`) Irvine, CA: University of California, Department of Information and Computer Science. The URL for the data used in this paper is `http://kdd.ics.uci.edu/databases/msnbc/msnbc.data.html`

[6] Huberman, B.A., Pirolli, P.L.T., Pitkow, J.E., Lukose, R.M. (1998). Strong Regularities in World Wide Web Surfing. Science, Vol. 280, 3 April.

[7] Spiegelhalter D., Thomas A., Best N. and Gilks W. (1996). BUGS 0.5 Bayesian Inference Using Gibbs Sampling. Manual (version ii).

[8] Zeger, S.L. and Karim, M.R. (1991). Generalized Linear Models with Random Effects; A Gibbs Sampling Approach. JASA March 1991, Vol. 86, No. 413. p 79-85.