# Latent Session Model for Web User Clustering
# A case study on modeling users of an online real estate website

Haijie Gu
Carnegie Mellon University

Andrew Bruce
Zillow.com

Carlos Guestrin
University of Washington

**Abstract**

We analyze the web access log of Zillow.com – one of the largest real estate website and present a hierarchical mixture model which learns clusters of users and sessions from the combination of web usage and content data. The model is able to exploit the hierarchical structure of the usage data, and learns stereotypical session types and user segments such as high end or low end house buyers. We show that our model produces better clusters both qualitatively and quantitatively comparing to a 2-phase baseline model.

## 1 Introduction

In this paper, we perform clustering analysis on the web usage data and web content data from Zillow.com – one of the biggest real estate web portals in the U.S. This study leads to a novel latent variable model that is able to exploit the hierarchical structure of the web log data, where users have multiple sessions which have multiple pageviews, and automatically learns clusters that represent different stereotypical browsing patterns.

Understanding user's preference is crucial to online business. For web-based companies, user-website interaction contains rich information about customers' depth and range of interest in the product space. A great amount of such interaction are captured by the web server access log. However, such data are often under utilized in many companies possibly due to its high noise to signal ratio as well as the lack of effective processing and modeling tools.

The online real estate market has experienced a rapid growth over the past years: according to latest reports from National Association of Realtors [5], 90% of home buyers use internet for house research, 76% physically visited homes viewed online, and 41% bought homes they found first online; In May 2013, the number of unique visitors to the top 10 online real estate sites is 223 millions, with 74 visits per user [13]. Therefore, a better understanding of such a large web user base is valuable for improving the online house shopping experience.

Learning from web access log is not a new topic and a great amount of work has been developed and applied to various applications such as click stream prediction [6, 10], personalization [11, 4] and recommendation systems [8]. See [14] for an overview of early work. However, these work generally focus on learning from the usage data alone without the considering the web content information. Therefore, even reasonable prediction rate is achieved, the users's underlying interest or intent cannot be fully explained.

The work closest to ours is the one from Jin, Zhou, and Mobasher [9] using pLSI [7] to jointly model the access data and web content data in an unified generative framework. The difference is that their model operates on individual session level and does not model the correlation of sessions for the same user.

Generally speaking, our latent session model clusters pageviews into session based latent classes, and based on which users are further clustered into groups. The model simultaneously learns 1.) latent classes of sessions that represent stereotypical browsing pattern driven by different user interest; and 2.) soft clustering of users over the latent session classes, which can be used as low dimensional features for classifications or recommendations to improving ads targeting and personalization.

# 2 The Zillow Dataset

## 2.1 Raw data and preprocessing

The raw data comprise two sources: the server access log and the home property database. We preprocessed the web logs from Feb 9 to Jun 5, 2013 which contains the users' requests to home detail pages (HDP) in Washington State.[1]. Each row in the data corresponds to one request from a user to an HDP with timestamp and 3 IDs:

- guid: global unique id of a user. Guid is associated with a cookie set to never expire when the user first went to the site.[2]
- session id: identifier of a browser session. Session is set to expire in 30 mins without new requests.
- property id: identifier of a home property.

Figure 1 shows the distribution of the number of pageviews per user (and per session) which follow the powerlaw distribution typically seen in the web traffic [3].

Because of the skewness of the web log data, we restrict our analysis to users who have at least 5 pageviews. The pruned log data has 274636 unique users, 12946030 unique sessions, and 5482595 total requests. User averages with 4.7 sessions and 20 pageviews; each session has 4.23 pageviews.

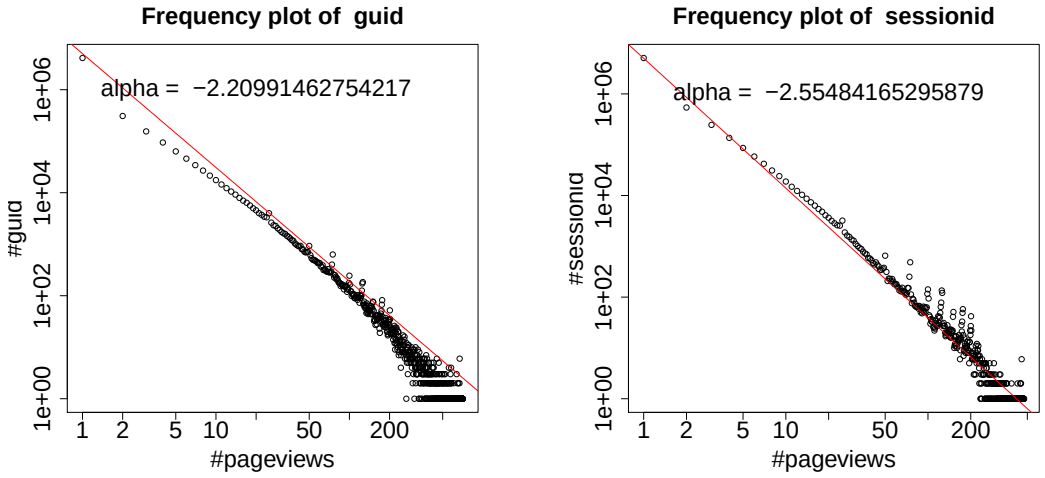



Figure 1: Fit Power-law distribution for number of pageviews per user (left) and per session (right).

The home property data contain features of the home properties such as the size, number of bedrooms, lot size, estimated value, etc. To keep the model general and without relying on heavy feature engineering, we choose 4 most basic features which we believe describes a home property:

1. SQFT, the finished size of in square feet.
2. lotsize, the lot size (if any) in square feet.
3. zestimate, the Zillow estimated market value in dollar.
4. isSFR, whether the home is a single family residence(SFR) or a condo apartment.

Finally, we join the log data and the home property data and collapse the same HDP views in each session. To capture the motivation underlying an HDP view, we also include 2 additional features to the combined dataset:

[1]We filtered out bot traffic which is identified using a blacklist

[2]It is impossible to identify a physical user from the web log because a user can clear cookie anytime or using multiple browsers. Therefore, even though a real user may have more than one "guid", in this analysis we will assume "guid" and "user" are interchangeable.

5. isforsale (Binary), whether the home is listed for sale at the time of the page view;
6. isrepeated (Binary), whether the pageview is repeated during the same session.

The intuition is that "isforsale" indicates user's underlying interest, e.g. to buy or to sell a house, and "isrepeated" provides behavioral evidence of the action. Table 1 lists an example of the combined data for a user with 2 sessions.

| guid | sessionid | sqft | lotsize | zest | issfr | forsale | isrepeated |
|---|---|---|---|---|---|---|---|
| ffa7a546866b467... | 55FE6C8200EE9... | 2550.00 | 8712.00 | 655048.00 | 1 | 0 | 0 |
| ffa7a546866b467... | 55FE6C8200EE9... | 3230.00 | 7725.00 | 681007.00 | 1 | 0 | 0 |
| ffa7a546866b467... | 55FE6C8200EE9... | 2670.00 | 12196.00 | 719513.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 3170.00 | 8712.00 | 705076.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 3700.00 | 14754.00 | 620250.00 | 1 | 0 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 3033.00 | 8712.00 | 663147.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 2220.00 | 11804.00 | 667096.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 2540.00 | 10890.00 | 511360.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 2490.00 | 14374.00 | 800051 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 2430.00 | 9191.00 | 684356.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 3150.00 | 14461.00 | 800065.00 | 1 | 1 | 0 |
| ffa7a546866b467... | 9D478E84AD5E6... | 3170.00 | 9583.00 | 726414.00 | 1 | 1 | 0 |

Table 1: An example user with two sessions in the combined dataset.

## 2.2 Feature Distribution

Among the 2457970 homes properties in the home property data, zestimate ranges from 32950 to 4198000 with an average of 270400; 78 percent of the properties are Single Family Residence. Figure 2 illustrates the estimated density of the home property features fitted with LogNormal distribution, which is widely used for modeling asset prices both theoretically and empirically. The zestimate and SQFT are well fitted with LogNormal distribution except for the outliers at the high end. The heavy tail effect seen at lotsize and the upper end of zestimate can be explained by spatial heterogeneity [12].

The final feature space captures two orthogonal aspects of a pageview: the type of the interest described by SQFT, lotsize, zestimate, isSFR and the intent described by isforsale and isrepeated. Figure 3 shows the density of of each feature averaged by user and by session in the combined dataset. It is important to notice that the density shape of forsale and isrepeated are significantly different in two settings. The session level density are highly concentrated. Forsale is strongly trimodal, meaning sessions are either spent at only forsale home, only forsale homes, or half half; and most of sessions do not repeat any pageviews. However, at user level, density of forsale ratio smooths out across the rest of the domain; isrepeated has a big density region between 0 and 0.5.

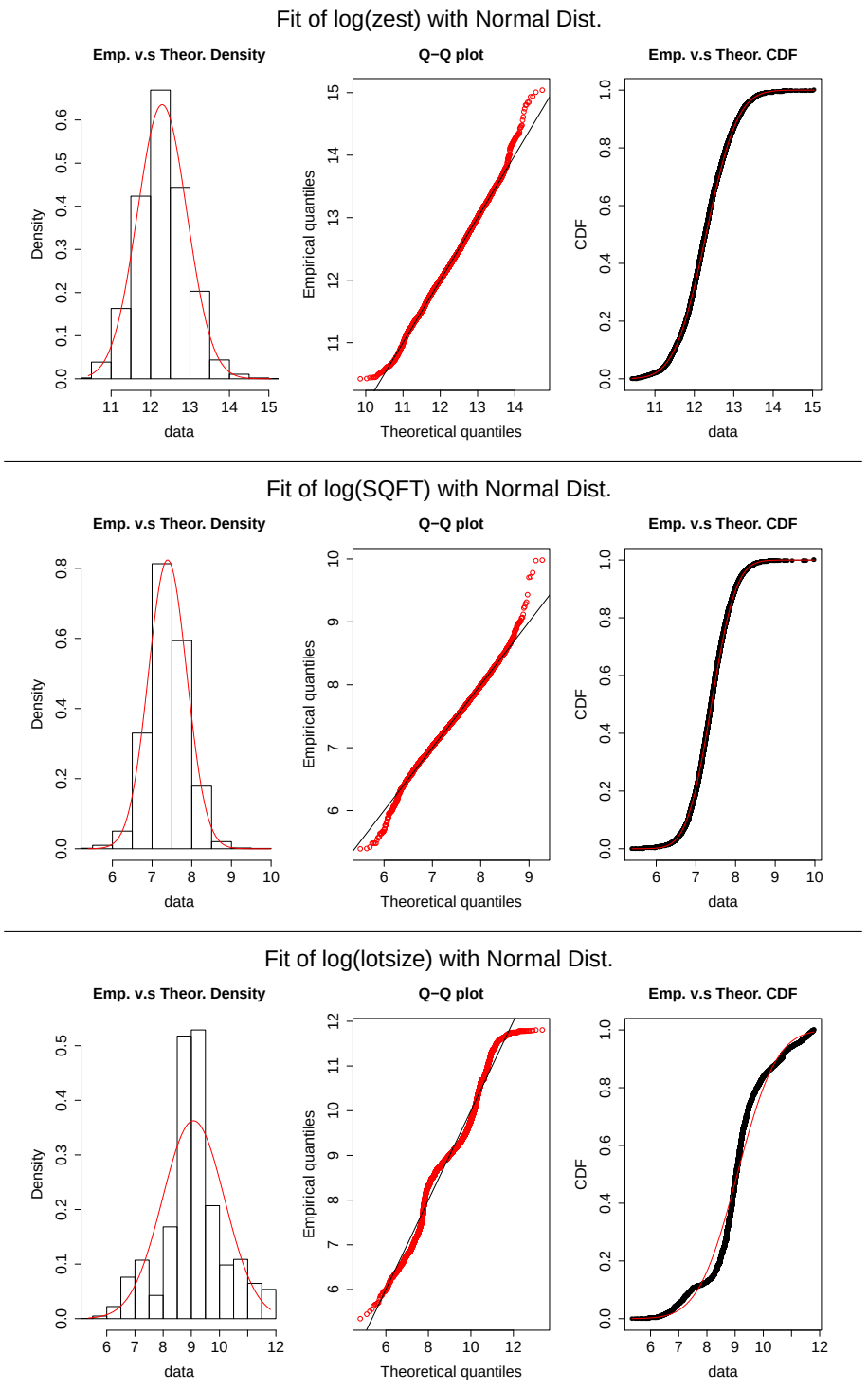
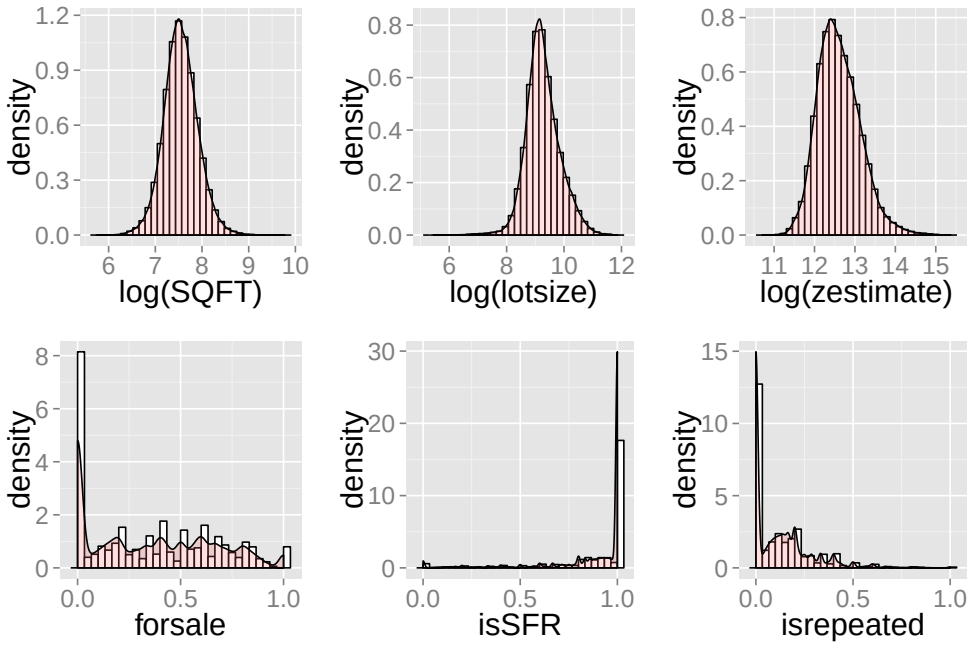
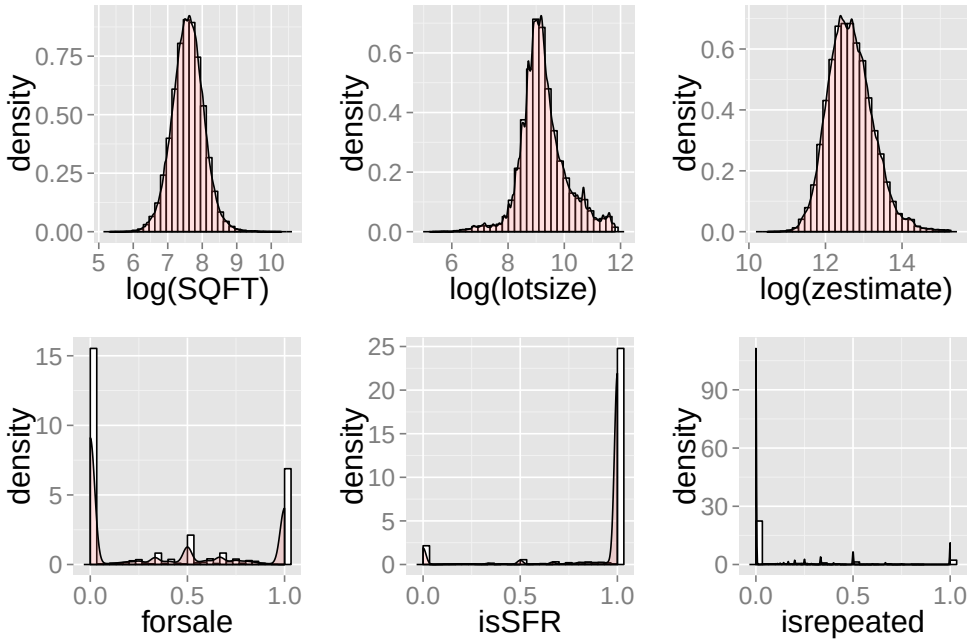



Figure 2: From top to bottom are the estimated density of zestimate, SQFT, and lotsize with LogNormal distribution.




(a) Estimated Density of features averaged by user.




(b) Estimated Density of features averaged by session.

Figure 3: Density plots of the features in the combined dataset averaged by user (a) and by session (b).

# 3 Model

## 3.1 A simple 2-phase baseline model

It is natural to start with a 2-step procedure, where we first cluster all the sessions and then aggregate the counts of the session clusters for each user.

A simple way to create session level feature is to average the pageview level features by sessionid and standardize to have unit variance. Then we can choose one of the cluster algorithms, e.g. K-Means in this case, to obtain a hard clustering of sessions over $K$ latent classes ($z_{ij}$). In the second step, we simply aggregate and normalize the sessions' cluster counts to obtain the user's mixture: $p_{il} = \frac{\sum I(z_{ij}=l)}{\sum_{zij}}$.

Despite the simplicity of K-Means, it performs relatively well at clustering sessions into representative groups [11]. However, when the goal is to cluster the users, the key problem of such ad-hoc procedure is that sessions are clustered independently from the users. As we see in figure 3, the distributions are differently at user level and session level. As an example, there are 7% of the sessions in which all pageviews are repeated; on the contrary, only 2% users have more than half repeated pageviews. Consequently, the session clusters could be "overfitted", e.g. having a cluster with just high ratio of repeated pageviews. On the other hand, when the session clusters aggregate into user level mixture, such high-repeated-ratio cluster is not representative and the corresponding component size is small. In addition, this procedure cannot model the correlation between features, e.g., zestimate, SQFT and lotsize. It also lacks unified probabilistic interpretation, making it difficult to add priors knowledge and tune parameters.

To overcome these problems, we propose a hierarchical mixture model which 1) jointly models users and sessions to share information among sessions of the same user, 2) allows flexible choice of distribution to model different feature types and 3) is able to capture correlation between features.

### 3.2 Latent Session Model

As mentioned in previous sections, the difference between the distribution of session level features and that of user level features suggest that a user's intent within a single session is relatively consistent comparing to those across sessions.

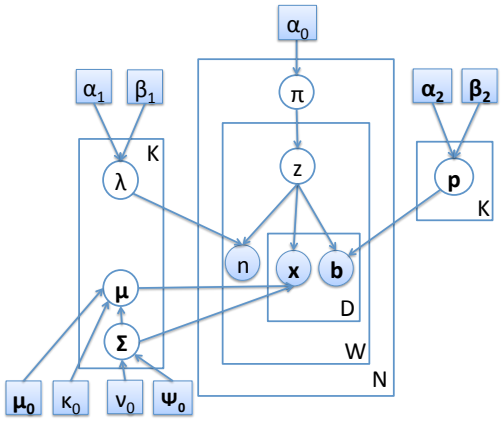
To capture such intuition, we start by modeling sessions with a discrete latent class variable taking values from $1 \cdots K$. The latent class represents the user's goal or intent for a single session. Next, users are modeled as (sparse) multinomial distribution over $K$ latent session class. The reason is that, users could have different intent or goals in different context, but the overall usage should be concentrated within a few goals depending on the long term interest.

Finally, conditioning on the session's topic, pageviews within the same session are assumed to be i.i.d. generated from the topic specific distribution, forming a naive bayes structure. The choice of distribution depends on the actual type of the feature. We use multivariate normal distribution to jointly model the continuous features, i.e. log of zestimate, lotsize, and SQFT, denoted as $\boldsymbol{x}$; Poisson distribution for the number of unique pageviews denoted as $n$; and independent Bernoulli distribution for binary variables denoted as $\boldsymbol{b}$ (e.g. isSFR, isforsale and isrepeated). For simplicity, we use conjugate priors on all model parameters. Specifically, the Dirichlet prior over the user's multinomial distribution can be used to control the sparsity of the session classes at the user level.

Such model can be viewed as a variant of the topic model[2], or an equivalent pLSI model [7], where we can draw analogy between user and document, latent session class and topics, and pageviews and words. The difference lies in the choice of distribution and the naive bayes structure between the session class and the pageviews of that session. More formally, we define the generative model as follows:

$$
\begin{aligned}
\pi_i &\sim Dir(\alpha_0) \\
z_{ij}|\pi_i &\sim Cate(\pi_i) \\
\lambda_l &\sim Gamma(\alpha_1, \beta_1) \\
n_{ij}|z_{ij} &\sim Poisson(\lambda_{z_{ij}}) \\
\boldsymbol{\mu}_l, \Sigma_l &\sim NIW(\boldsymbol{\mu}_0, \kappa_0, \nu_0, \boldsymbol{\Psi}_0) \\
\boldsymbol{x}_{ijk}|z_{ij} &\sim N(\boldsymbol{\mu}_{z_{ij}}, \Sigma_{z_{ij}}) \\
\boldsymbol{p}_l &\sim Beta(\boldsymbol{\alpha}_2, \boldsymbol{\beta}_2) \\
\boldsymbol{b}_{ijk}|z_{ij} &\sim Bern(\boldsymbol{p}_{z_{ij}})
\end{aligned}
$$

Figure 4 lists the notation table and draws the graphical model in the plate diagram.




| | Observed |
|---|---|
| $K, N$ | number of latent classes, users. |
| $W_i$ | number of sessions for user $i$. |
| $n_{ij}$ | Number of unique hdp views for $session_{ij}$. |
| $\boldsymbol{b}_{ij}$ | Binary features of $session_{ij}$: (isforsale, isSFR, isrepeated). |
| $\boldsymbol{x}_{ij}$ | Continuous features of $session_{ij}$: log(zest, lotsize, SQFT). |
| | Latent |
| $\pi_i$ | Mixture coefficient for user $i$. |
| $z_{ij}$ | Latent class assignment of $session_{ij}$. |
| $\lambda_l$ | Poisson parameter for latent class $l$. |
| $\boldsymbol{p}_l$ | Bernoulli parameters for class $l$. |
| $\boldsymbol{\mu}_l, \Sigma_l$ | Multivariate normal parameters for class $l$. |

(a) Model Notations

Figure 4: Latent Session Model

### 3.3 Inference

There are three sources of uncertainties to be inferred from the data:

1. distribution over session classes for each user: $\pi_i$ for $i = 1 : N$.
2. the topic assignment for each session: $z_{ij}$ for $j = 1 : N_i$.
3. the distribution over features conditional on the latent class: $\boldsymbol{\mu}_k, \Sigma_k, \lambda_k, \boldsymbol{p}_k$.

The complete likelihood is:

$$\begin{aligned} & P(\boldsymbol{x}, \boldsymbol{b}, \boldsymbol{n}, \boldsymbol{z}, \boldsymbol{\pi}, \boldsymbol{\lambda}, \boldsymbol{p}, \boldsymbol{\mu}, \Sigma) \\ = \; & P(\boldsymbol{\mu}, \Sigma) P(\boldsymbol{p}) P(\boldsymbol{\lambda}) \Pi_{i=1}^{N} P(\boldsymbol{\pi_i}) \Pi_{j=1}^{N_i} P(z_{ij}|\boldsymbol{\pi_i}) P(n_{ij}|\lambda_{z_{ij}}) \Pi_{k=1}^{n_{ij}} P(\boldsymbol{x}_{ijk}|\boldsymbol{\mu}_{z_{ij}}, \Sigma_{z_{ij}}) P(\boldsymbol{b}_{ijk}|\boldsymbol{p}_{z_{ij}}) \end{aligned}$$

We treat the topic assignments $z_{ij}$ as latent variable $Z$, the rest as parameters $\Theta = \{\pi, \boldsymbol{\mu}, \Sigma, \boldsymbol{p}, \lambda\}$, and apply expectation maximization to learn the model. In the E-step, we fix the estimated $\hat{\Theta}$ and compute $p(Z|\hat{\Theta})$.

$$\log q(z_{ij} = l|X, \hat{\Theta}) \propto \log p(n_{ij}|\lambda_l) + \sum_{k=1}^{n_{ij}} \log p(\boldsymbol{x}_{ijk}|\boldsymbol{\mu}_l, \Sigma_l) + \log p(\boldsymbol{b}_{ijk}|\boldsymbol{p}_l) \tag{1}$$

In the M-Step, we fix $q(Z|\hat{\Theta})$, and find MAP estimation of the parameters.

$$\text{Weighted Sufficient Statistics} \tag{2}$$

$$\overline{\boldsymbol{x}}_l \equiv \sum_{ijk} q(z_{ij} = l) \boldsymbol{x}_{ijk} \tag{3}$$

$$\overline{\boldsymbol{b}}_l \equiv \sum_{ijk} q(z_{ij} = l) \boldsymbol{b}_{ijk} \tag{4}$$

$$\overline{r}_l \equiv \sum_{ij} n_{ij} q(z_{ij} = l) \tag{5}$$

$$\hat{S}_l \equiv \sum_{ij} q(z_{ij} = l) \sum_{k=1}^{n_{ij}} (\boldsymbol{x}_{ijk} - \overline{\boldsymbol{x}}_l)(\boldsymbol{x}_{ijk} - \overline{\boldsymbol{x}}_l)^T \tag{6}$$

$$\overline{n}_l \equiv \sum_{ij} q(z_{ij} = l) n_{ij} \tag{7}$$

$$\tag{8}$$

$$\text{User Level Parameters} \tag{9}$$

$$\hat{\pi}_{il} \propto \sum_j q(z_{ij} = l) - 1 \tag{10}$$

$$\text{Session Level Parameters} \tag{11}$$

$$\hat{\lambda}_l = \frac{\overline{n}_l + \alpha_1 - 1}{\beta_1 + \sum_{ij} q(z_{ij} = l)} \tag{12}$$

$$\text{Pageview Level Parameters} \tag{13}$$

$$\hat{\boldsymbol{\mu}}_l = \frac{\overline{\boldsymbol{x}}_l + \kappa_0 \mu_0}{r_l + \kappa_0} \tag{14}$$

$$\hat{\Sigma}_l = \frac{\Psi_0 + S_l + \frac{\kappa_0 r_l}{\kappa_0 + r_l}(\overline{\boldsymbol{x}}_l - \boldsymbol{\mu}_0)(\overline{\boldsymbol{x}}_l - \boldsymbol{\mu}_0)^T}{\nu_0 + r_k + D + 2} \tag{15}$$

$$\hat{\boldsymbol{p}}_l = \frac{\overline{\boldsymbol{b}}_l + \boldsymbol{\alpha_2}}{\boldsymbol{\alpha_2} + \boldsymbol{\beta}_2 + r_l} \tag{16}$$

$$\tag{17}$$

During the M-Step, we also infer the missing data using the MAP estimate. The model is learned by alternating the E-step and M-step, until the expected log likelihood converges.

In the context of clustering, the model learns 1.) soft clustering of session over classes; 2.) soft clustering of user over session classes; and 3.) soft clustering of features over session classes.

# 4 Evaluation

We initialize the baseline model using K-Means++ [1], which guarantees to produce results close to the global minimum with high probability. After that, the result of the baseline model is used to initialize the full model.

Before jumping into the evaluation, there are two important questions we need to answer: how to choose the number of latent session classes $K$, and how stable is the clustering result. Choosing $K$ is still an active research area, and the actual choice depends various from case to case. Here, we choose the number of clusters which has the best trade-off between stability and objective value. This will also address the second question about the cluster stability.

## 4.1 Sensitivity Analysis and Model Selection

For given $K$, the K-Means algorithm minimizes the total within-cluster sum of squares (TWISS):

$$\min_{c_1,\ldots,c_K} \frac{1}{N} \sum_{i=1}^{N} \min_{1 \le j \le K} \|X_i - c_j\|^2 \tag{18}$$

where $c_j \in R^d, 1 \le j_1 \le K$ are the cluster centers. To measure the distance between the two set of cluster centers, we define the following matching distance:

$$d(C_i, C_j) = \min_{\boldsymbol{\pi} \in \boldsymbol{\Pi_K}} \frac{1}{K} \sum_{k=1}^{K} \|C_i^{\pi_k} - C_j^k\|_1 \tag{19}$$

where $\boldsymbol{\Pi_K}$ is the set all possible permutations of $\{1, \cdots, K\}$. The distance metric $d$ defines the $l_1$ distance between two sets of cluster centers $C_i$ and $C_j$ under the best matching of cluster ids. Figure 5 illustrates the mean and standard deviation of TWISS (top) and matching distance (bottom) for $K \in \{5, 10, 15, 20\}$ and $n = 20$ experiments for each $K$. The sample for the latter measurement is taken from all unique pairs from $1, \cdots, n$. We see the "elbows" for TWISS at $K = 10$ and for the matching distance at $K = 15$.

For the best trade-off between stability and good local minimum, we choose $K = 10$. The rest of the results are based on the experiment for $K = 10$ with the lowest TWISS.

## 4.2 Characteristics of Latent Session Classes

Figure 6 visualizes the latent session classes by plotting the features from each latent session class. For the baseline model (left), we plot the mean and the standard deviation of the features in each cluster. For the latent session model (right), we plot the parameters of the feature distribution. While two results are similar at the first glance, the main difference lies in the forsale and isrepeated where the baseline model learns extreme values. For example, the cluster 5 and 6 in the baseline model has 80% repeated pageviews but the ones in the latent session model only have 30%. Such difference suggest that the baseline model suffers from overfitting due to its 2-phase procedure. There are sessions within which pageviews are always repeated, however, most of them are very small sessions, e.g., sessions with just a click and a refresh. Therefore these sessions alone are not significant enough to form its own cluster in the full model where parameters are shared among other sessions of the same user.

Another way to visualize the latent session classes is via hierarchical clustering of the parameter space, as shown in figure 7a. From the dendrogram, it is easy to read out the characteristics of each session class and assign them the corresponding stereotypes:

- Class 2 and 7 are for Condo researching and shopping.
- Class 1 and 10 are for low end SFR researching and shopping.
- Class 4 and 9 are for the Power Sessions featuring high pageview counts, averaging 150 and 30 unique pageviews respectively.
- Class 5 and 6 are for SFR shopping and SFR researching featuring repeated views on medium value houses.
- Class 3 and 8 are for Special Interests featuring non-repeated pageviews on high value and big lot houses respectively.

The latent session model also enables us to visualize the covariance structure between the features. As is shown in figure 8, clusters have different covariance structures among zestimate, SQFT and lotsize. For example, the size of the house influences zestimate more in the "low end SFR" cluster 1 than it does in the

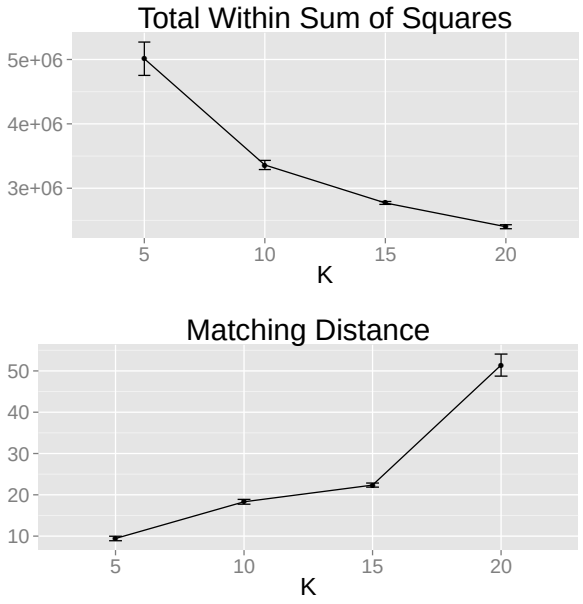



Figure 5: The mean and standard deviation of the total within-cluster sum of square (top), and the matching distance (bottom) for $K = 5, 10, 15, 20$ and over 20 experiments for each K.

"higher end SFR" cluster 3. Another example is that the lot size does not affect zest as much in the "very big lot" cluster 8 as it does in other clusters with normal lot size.

We can also visualize the distribution of session classes within different user groups. We obtain user metadata for a subset of the users and label each user with one of the four groups:

1. Professional: real estate agent, broker etc.
2. Internal Employee: Zillow internal employees identified using IP addresses.
3. Active buyer: users who have taken actions to contact a local agent.
4. Unknown: the rest of the users.

It is worth mentioning that the first three groups are not mutually exclusive as people might take different roles in different times. For simplicity, we only assign one group to each user in the order of Internal > Professional > Active buyer > Unknown. This results in 11327 professionals, 466 internals, 3444 active buyers and 259399 unknowns.

Figure 7b illustrates the distribution of the latent session classes averaged by the group label. It is not surprising to see that internals has the highest probability mass in class 3, which is for high value houses and the lowest mass in class 1, which is the low end and not forsale houses. Also, active buyers have higher mass in class 5 and 7 which are medium SFR shopping and condo shopping.

### 4.3 User Purity

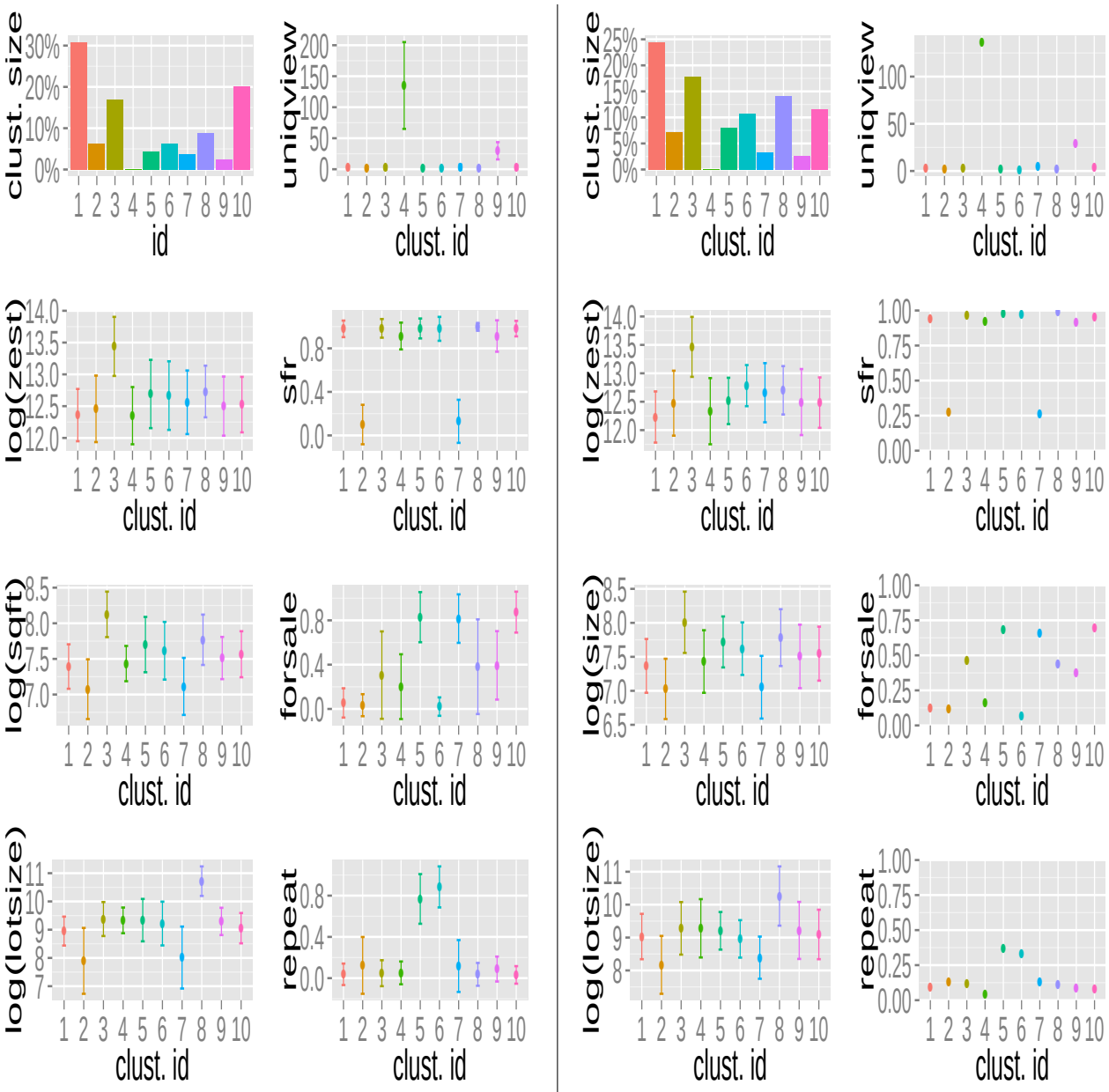
In a 90-day time window, it is reasonable to assume that users only have a few session types, for example, it is less common that a user is simultaneously seriously interested in both high end and low end houses. We define the purity of a user as the number of distinct classes taken for his sessions. For example, the possible purity values for a user with 5 sessions are 1 to 5, the smaller the purer.

Figure 9 summarizes the density of user purity for both models. We omit the trivial case where users have only one session, and combine the cases where the session count is greater than the number of classes. Because the latent session model outputs soft clustering of sessions over latent classes, we plot the sampled mean with standard error.

For the baseline model, the user purity exhibits normal-like density shape because the session classes are learned independent of the user constraint. However, because the full model jointly learns the user cluster and session cluster, the resulting user purity is higher, even for people with 10 sessions, 20 percent take on only one session class and the mode is 2 classes, which is more realistic.

### 4.4 Classifying User Groups

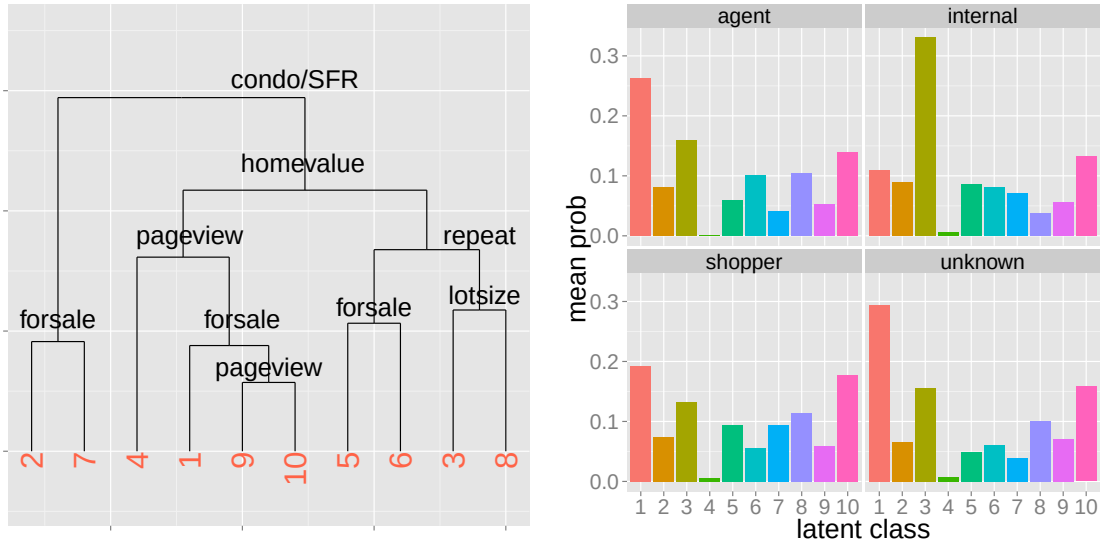
Although the main focus of unsupervised generative models is to assign data points into groups and enable exploratory analysis, predictive tasks can still be useful for model selection and assessment. For both models, we treat the user's distribution over session class as low dimensional features, and train Support Vector Machines to classify which group the user belongs to. We run the experiments 100 times, each time with 100 users uniformly sampled from each group. In each experiment we train SVM to discriminate all possible pairs of the groups using 5 fold cross validation, and report the mean accuracy per group pair. For comparison, we also show the accuracy of SVM trained using raw features averaged by user. The results are summarized in figure 10. The latent session model significantly outperforms the baseline model in all cases except for agent v.s unknown. The largest improvement in accuracy is 10% for internal v.s unknown. On average, the accuracy of latent session model is 4.5% higher than that of the baseline model. Raw features score the highest accuracy in all cases, and the improvement over the latent session model ranges from 0.7% to 6.74%, averaging 2.8% across all cases. However, the raw features can not be used for clustering and summarizing the data.




(a) Baseline Model

(b) Latent Session Model

Figure 6: Comparison of the cluster specific feature parameters in two models.




(a) (b)

Figure 7: On the left is the dendrogram of session classes. On the right is the mean distribution of the session classes over different user groups.

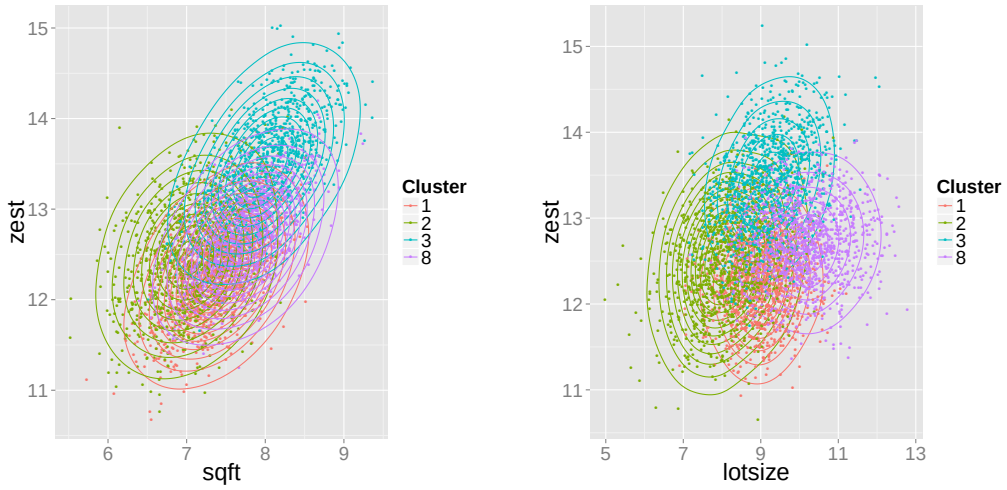



Figure 8: Contours of the estimated covariance between zestimate and SQFT (left), and zestimate and lotsize (right) for session class 1,2,3, and 8.

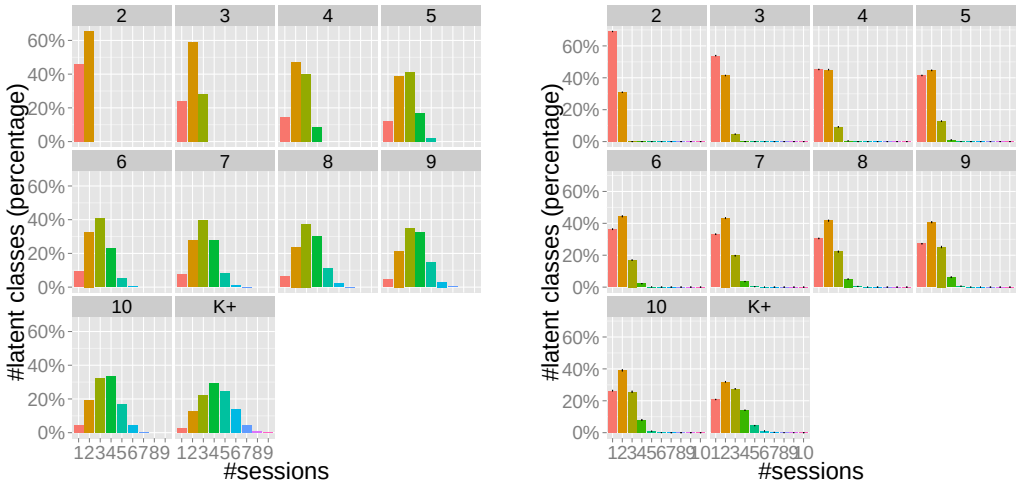



Figure 9: Normalized histogram of user session purity. In latent session model (right), the number of latent session class per user is much smaller than that in in the baseline model with a two-phase learning procedure. Since the full model outputs a soft clustering of session, the sampled mean is used with standard deviation.

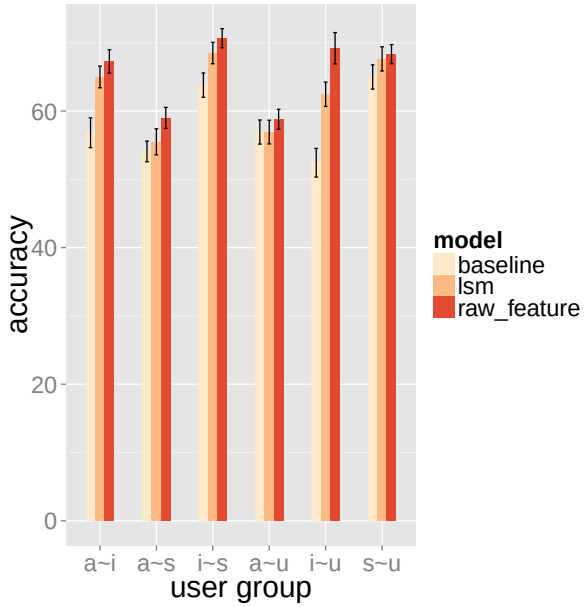



Figure 10: Classification accuracy of SVM using features from baseline model, latent session model, and raw features. The mean accuracy with standard error over 20 runs is plotted for each pair of labels(a: agent/professional, i: internal, s: shopper/active user, u: unknown).

# 5 Conclusion and Future Work

We present a generative model for learning the user web interaction from the combination of web server logs and web contents. We applied our model to a novel dataset obtained from a large online real estate website and automatically learned the stereotypical sessions such as shopping or researching on different types of houses. We evaluated our model by comparing it to a 2 phase baseline model qualitatively at exploratory tasks and qualitatively at classification tasks. The model can be easily generalized to other types of web traffic, e.g., books or travelings to gain a better understanding of the user base.

Current model assumes exchangeability for sessions, and a naive bayes model between sessions and their pageviews. Future work includes modeling the sequence structure of sessions and the correlation among pageviews.

# 6 Acknowledgement

We thank Chunyi Wang and other members at Zillow for providing the data source and Yucheng Low for useful feedback on earlier drafts of this paper.

# References


[1] David Arthur and Sergei Vassilvitskii. "k-means++: The advantages of careful seeding". In: *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics. 2007, pp. 1027–1035.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *the Journal of machine Learning research* 3 (2003), pp. 993–1022.

[3] Mark E Crovella and Azer Bestavros. "Self-similarity in World Wide Web traffic: evidence and possible causes". In: *Networking, IEEE/ACM Transactions on* 5.6 (1997), pp. 835–846.

[4] Magdalini Eirinaki and Michalis Vazirgiannis. "Web mining for web personalization". In: *ACM Transactions on Internet Technology (TOIT)* 3.1 (2003), pp. 1–27.

[5] *Field Guide to Quick Real Estate Statistics*. 2013. URL: `http://www.realtor.org/field-guides/field-guide-to-quick-real-estate-statistics`.

[6] Şule Gündüz and M Tamer Özsu. "A web page prediction model based on click-stream tree representation of user behavior". In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2003, pp. 535–540.

[7] Thomas Hofmann. "Probabilistic latent semantic indexing". In: *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM. 1999, pp. 50–57.

[8] Xin Jin, Yanzan Zhou, and Bamshad Mobasher. "A maximum entropy web recommendation system: combining collaborative and content features". In: *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*. ACM. 2005, pp. 612–617.

[9] Xin Jin, Yanzan Zhou, and Bamshad Mobasher. "A unified approach to personalization based on probabilistic latent semantic models of web usage and content". In: *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP'04)*. 2004.

[10] Sung-Hae Jun. "Web usage mining using support vector machine". In: *Computational Intelligence and Bioinspired Systems*. Springer, 2005, pp. 349–356.

[11] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. "Automatic personalization based on Web usage mining". In: *Communications of the ACM* 43.8 (2000), pp. 142–151.

[12] Takaaki Ohnishi et al. "On the evolution of the house price distribution". In: (2011).

[13] *REALTOR.COM TRAFFIC*. 2013. URL: `http://www.realtor.org/sites/default/files/reports/2013/nar-website-traffic-stats-may-2013-08.pdf`.

[14] Jaideep Srivastava et al. "Web usage mining: Discovery and applications of usage patterns from web data". In: *ACM SIGKDD Explorations Newsletter* 1.2 (2000), pp. 12–23.