Relational Clustering Based on a New Robust Estimator with Application to Web Mining

Olfa Nasraoui
Comp. Engg. & Comp. Sc.
University of Missouri
Columbia, MO 65211
olfa@ece.missouri.edu

Raghu Krishnapuram
Math & Comp. Sc.
Colorado School of Mines
Golden, CO 80401
rkrishna@mines.edu

Anupam Joshi
Comp. Sc. & Elec. Engg.
University of Maryland
Baltimore, MD 21250
joshi@cs.umbc.edu

Abstract

Mining typical user profiles and URL associations from the vast amount of access logs is an important component of Web personalization. In this paper, we define the notion of a "user session" as being a temporally compact sequence of Web accesses by a user. We also define a dissimilarity measure between two Web sessions that captures the organization of a Web site. To cluster the user sessions based on the pair-wise dissimilarities, we introduce the Relational Fuzzy $C-$Maximal Density Estimator (RFC$-$MDE) algorithm. RFC$-$MDE is robust, and can deal with outliers that are typical in this application. We show real examples of the use of RFC$-$MDE for extraction of user profiles from log data, and and compare its performance to the standard Non Euclidean Fuzzy $C-$ Means.

1 Introduction

Personalization deals with tailoring a user's interaction with the Web information space based on information about him/her. It can be accomplished either via search engines such as Lycos, or by making Web sites adaptive. Initial work in this area such as the Firefly system [1] concentrated on creating recommender systems. More recent systems that deal with this idea are W^3IQ [2], *PHOAKS* [3], and the Webwatcher project [4]. Mining typical user profiles from server or access logs is a possible approach to personalization [5, 6]. So far, most efforts have relied on relatively simple non-fuzzy techniques which can be inadequate for real user profile data. In this paper, we define the notion of a "user session" as being a temporally compact sequence of Web accesses by a user. We also define a new distance measure between two Web sessions that captures the organization of a Web site.The goal of our Web mining is to categorize these sessions. In this light, Web mining can be viewed as a special case of the more general problem of knowledge discovery in databases [7, 8].

Most data mining applications involve data that is corrupted with noise. Existing robust estimators suffer from their strong dependence on a known or assumed amount of noise (contamination rate), or equivalently an estimated scale value or inlier bound. The Maximal Density Estimator (MDE) [12] yields a robust estimate of the parameters without assuming that the contamination rate is known. An extension of this estimator, called the Fuzzy $C-$ Maximal Density Estimator (F$C-$MDE), can perform fuzzy clustering [12]. Since Web sessions are too complex to convert to simple numerical features, we extend the F$C-$MDE so that it can work on non-Euclidean relational data. The Web sites at the University of Missouri, Columbia, and the University of Maryland at Baltimore County were used as testbeds for the algorithm.

2 Defining Similarity Between User Sessions

2.1 Preprocessing and Segmentation of the access log data into sessions

Each access log entry consists of :(i) User's IP address, (ii) Access time, (iii) Request method ("GET", "POST", $\cdots$, etc), (iv) URL of the page accessed, (v) Data transmission protocol (typically HTTP/1.0), (vi) Return code, and (vii) Number of bytes transmitted. First, we filter out log entries that are not germane for our task. These include entries that: (i) result in any error (indicated by the error code), (ii) use a request method other than "GET", or (iii) record accesses to image files (.gif, .jpeg, , $\cdots$, etc.), which are embedded in other pages. Next, analogous to [6], the individual log entries are grouped into user sessions. Since Web servers do not typically log user names (unless *identd* is used), we define a user session as accesses from the same IP address such that the duration of elapsed time between

two consecutive accesses in the session is within a prespecified threshold. Each URL in the site is assigned a unique number $j \in \{1, \ldots, N_U\}$, where N_U is the total number of valid URLs. The i^{th} user session is encoded as an N_U-dimensional binary vector $\mathbf{s}^{(i)}$ with the property

$$s_j^{(i)} = \begin{cases} 1 & \text{if user accessed } j^{th} \text{ URL during } i^{th} \text{ session} \\ 0 & \text{otherwise} \end{cases}$$

Our scheme will map one user's multiple sessions to multiple user sessions. This notion of multiple user sessions enable us to better capture the situation when the same user displays a few (different) access patterns on this site.

2.2 Adaptation of Session Data to Clustering

A simple measure of similarity between sessions $\mathbf{s}^{(k)}$ and $\mathbf{s}^{(l)}$ is given by:

$$S_{1,kl} = \frac{\sum_{i=1}^{Nu} s_i^{(k)} s_i^{(l)}}{\sqrt{\sum_{i=1}^{Nu} s_i^{(k)}} \sqrt{\sum_{i=1}^{Nu} s_i^{(l)}}} \quad (1)$$

The problem with this similarity measure is that it completely ignores the hierarchical organization of the Web site. For example, the session pair {/courses/cecs345} and {/courses/cecs343}, as well as the session pair {/courses/cecs345} and {/research/grants} will receive a 0 similarity score according to S_1. This leads us to define an alternative similarity measure that takes into account the syntactic representation of two URLs as follows.

$$S_u(i,j) = \min\left(1, \frac{|(p_i \cap p_j|}{\max(1, \max(|p_i|, |p_j|) - 1)}\right), \quad (2)$$

where p_i denotes the path traversed from the root node to the node corresponding to the i^{th} URL, and $|p_i|$ indicates the length of this path or the number of edges included in the path. Now the similarity on the session level which incorporates the syntactic URL similarities is defined by correlating all the URL attributes and their similarities in two sessions as follows:

$$S_{2,kl} = \frac{\sum_{i=1}^{N_U} \sum_{j=1}^{N_U} s_i^{(k)} s_j^{(l)} S_u(i,j)}{\sum_{i=1}^{N_U} s_i^{(k)} \sum_{j=1}^{N_U} s_j^{(l)}} \quad (3)$$

Unlike S_1, this similarity uses soft URL level similarities. When the two sessions are identical, $S_{2,kl}$ simplifies to $S_{2,kk} = \frac{1}{\sum_{i=1}^{N_U} s_i^{(k)}}$, which can be considerably small depending on the number of URLs accessed. Besides identical sessions, this similarity will generally be underestimated for session pairs who share some identical URLs while the rest of the unshared URLs have low syntactic similarity. In general for such sessions $S_{1,kl}$ provides a higher and more accurate session similarity. Therefore, we define a new similarity between two sessions that takes advantage of the desirable properties of S_1 and S_2 as follows:

$$S_{kl} = \max(S_{1,kl}, S_{2,kl}) \quad (4)$$

For relational clustering, this similarity is mapped to the dissimilarity measure $d_s^2(k,l) = (1 - S_{kl})^2$.

3 MDE and Extensions

3.1 The Maximal Density Estimator (MDE)

MDE [12] is a new robust estimator that is free of any presuppositions about the noise proportion. Let $\mathcal{X} = \{\mathbf{x}_j \mid j = 1, \ldots, N\}$ be a set of feature vectors in an $n-$dimensional feature space. Let θ represent the parameters to be estimated, and let σ denote the scale of the data set. MDE is given by:

$$\min_{\theta,\sigma} \left\{ J = \sum_{j=1}^{N} w_j \frac{d_j^2}{\sigma} - \alpha \sum_{j=1}^{N} w_j \right\}. \quad (5)$$

In (5), d_j^2 is the square of the residual of data point $\mathbf{x}_j$ with respect to the fit computed from the parameters θ, and w_j is a positive weight associated with point $\mathbf{x}_j$. The weight w_j can be considered as the degree of membership of data point $\mathbf{x}_j$ in the inlier set or the set of good points. The first term of this objective function tries to minimize the scaled residuals of the good points. The second term of this objective function tries to use as many good points (inliers) as possible in the estimation process, via their high weights. Thus, the combined effect is to optimize the density. To balance the two terms, in 2-D, we choose $\alpha = 1$. For the general n-dimensional case α should be close to n, since the ratio of the first term to the second term approaches the average of a χ^2 distribution for Gaussian data. Also, we choose to use the Gaussian weight function $w_j = \exp\{-d_j^2/2\sigma\}$. Finally, we should note that d_j^2 should be a suitable distance measure.

If the weights are fixed, then the update equations prototype parameters are found by setting $\frac{\partial J}{\partial \theta} = (1/\sigma) \quad \sum_{j=1}^{N} w_j \frac{\partial d_j^2}{\partial \theta} = \mathbf{0}$. Similarly, the update equation for scale can be derived by fixing the prototype parameters and setting $\frac{\partial J}{\partial \sigma} = 0$. MDE consists of iterative updates of the prototype parameters, followed by updates of the scale parameter and the weights until convergence, or until a fixed maximum number of iterations is reached.

3.2 Fuzzy Clustering with MDE

To perform robust fuzzy clustering, we generalize the objective function of MDE as follows:

$$\min_{\Theta,\sigma_i,u_{ij}} \left\{ J = \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m w_{ij} \frac{d_{ij}^2}{\sigma_i} - \alpha \sum_{i=1}^{C} \sum_{j=1}^{N} u_{ij}^m w_{ij} \right\} \quad (6)$$

Here $\Theta = (\theta_1, \ldots, \theta_C)$ represents the C-tuple of prototype parameters to be estimated, $m \in [1, \infty)$ is a

fuzzifier parameter that controls the degree of fuzziness of the resulting partition, and w_{ij} is given by $w_{ij} = \exp\{-d_{ij}^2\}/(2\sigma_i)$. The memberships u_{ij} in (6) are subject to the following constraints: $0 \leq u_{ij} \leq 1$; $\sum_{i=1}^{C} u_{ij} = 1 \, \forall j = 1, \ldots, N$; and $0 \leq \sum_{j=1}^{N} u_{ij} \leq N \, \forall i = 1, \ldots, C$.

Since each cluster is independent of the rest, it is easy to show that the optimal update equations are similar to the ones obtained for estimating the parameters for one cluster. The scale parameter of the i^{th} cluster is given by

$$\sigma_i = \frac{1}{(2+\alpha)} \frac{\sum_{j=1}^{N} u_{ij}^m w_{ij} d_{ij}^4}{\sum_{j=1}^{N} u_{ij}^m w_{ij} d_{ij}^2}. \quad (7)$$

When d_{ij}^2 is the Euclidean distance $d_{ij}^2 = \|\mathbf{x}_j - \mathbf{c}_i\|^2$, the update equation for the center $\mathbf{c}_i$ is given by

$$\mathbf{c}_i = \frac{\sum_{j=1}^{N} u_{ij}^m w_{ij} \mathbf{x}_j}{\sum_{j=1}^{N} u_{ij}^m w_{ij}} \quad (8)$$

To simplify the algorithm, we decouple the robust parameter estimation process via the robust weights w_{ij}, from the partitioning process via the memberships u_{ij}, and use the simpler Fuzzy c-Means [9] objective function to update the memberships u_{ij} in each iteration as follows:

$$u_{ij} = \frac{\left(\frac{1}{d_{ij}^2}\right)^{\frac{1}{m-1}}}{\sum_{k=1}^{C} \left(\frac{1}{d_{kj}^2}\right)^{\frac{1}{m-1}}}. \quad (9)$$

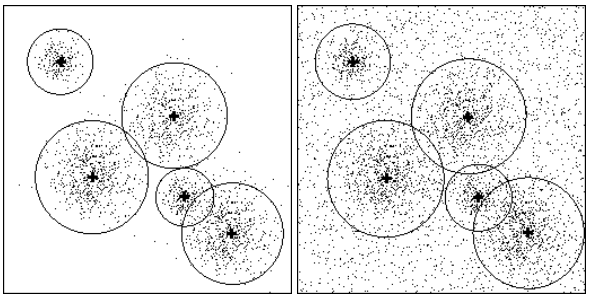
Therefore, the optimization process will consist of alternating updates of the memberships, as given by (9), and the cluster centers and scale parameters as given by (8) and (7). We call the resulting clustering algorithm the Fuzzy $C-$ Maximal Density Estimator (F$C-$MDE).

Fig. 1 illustrates the performance of F$C-$MDE on a noisy data set containing clusters of different sizes and densities. Here F$C-$MDE was applied with $C = 5$ and $m = 2.2$. It was initialized using FCM with $m = 3$. The initial scale parameters were all initialized to the fuzzy average intracluster distance, $\sigma_i = \frac{\sum_{j=1}^{N} u_{ij}^m d_{ij}^2}{\sum_{j=1}^{N} u_{ij}^m}$. Figs. 1(a) and (b) show the centers and the inlier bounds found by F$C-$MDE for the clean data set and the data set contaminated with 52% noise respectively.

3.3 The Relational F$C-$MDE

Since our application deals with similarities between user sessions, we formulate the relational dual of F$C-$MDE, by following the model established by Hathaway et al. [10] and write: $d_{ik}^2 = (\mathbf{R}\mathbf{v}_i)_k - \mathbf{v}_i \mathbf{R} \mathbf{v}_i / 2$, where $\mathbf{R} = [R_{jk}]$ is the dissimilarity between $\mathbf{x}_j$ and $\mathbf{x}_k$, and $\mathbf{v}_i$ is the membership vector defined by

$$\mathbf{v}_i = \frac{(u_{i1}^m w_{i1}, \ldots, u_{iN}^m w_{iN})^t}{\sum_{j=1}^{N} u_{ij}^m w_{ij}}. \quad (10)$$



(a) (b)

Figure 1. Performance of F$C-$MDE on contaminated data. (a) Result on clean data set. (b) Result on data set contaminated with 52% uniform noise

A relational dual of F$C-$MDE exists when there exists a set of N points in $\mathcal{R}^{N-1}$, called a realization of $\mathbf{R}$, satisfying $\mathbf{R} = [R_{jk}] = \|\mathbf{x}_j - \mathbf{x}_k\|^2$. When this is not the case, some of the distances computed using $\mathbf{R}$ may be negative. To overcome this problem, we use the β-spread transform [11] to convert a non-Euclidean matrix $\mathbf{R}$ into an Euclidean Matrix $\mathbf{R}_\beta$. The resulting Relational F$C-$MDE algorithm (RF$C-$MDE) can deal with complex and subjective dissimilarity/similarity measures which are not restricted to be Euclidean or metric.

4 Experimental Results

4.1 Measures for Evaluation of Results

We interpret the results of RF$C-$MDE on the user-session data by using the following quantitative measures. First, the user sessions are assigned to the closest clusters based on the distances. This creates C clusters $\mathcal{X}_i$, for $1 \leq i \leq C$. The sessions in cluster $\mathcal{X}_i$ are summarized in a typical session "profile" vector $\mathbf{P}_i = (P_{i1}, \ldots, P_{i_{N_U}})^T$, where the components of $\mathbf{P}_i$ are URL weights that represent the probability of access of each URL during the sessions of $\mathcal{X}_i$ as follows:

$$P_{ij} = p\left(\mathbf{s}_j^{(k)} = 1 | \mathbf{s}^{(k)} \in \mathcal{X}_i\right) = \frac{|\mathcal{X}_{i_j}|}{|\mathcal{X}_i|}, \quad (11)$$

where $\mathcal{X}_{i_j} = \{\mathbf{s}^{(k)} \in \mathcal{X}_i \mid s_j^{(k)} > 0\}$. The URL weights P_{ij} measure the significance of a given URL to the i^{th} profile. Besides summarizing profiles, the components of the profile vector can be used to recognize an invalid profile which has no strong or frequently accessed pattern. For such a profile, all URL weights will be low. We also use the intra-cluster or within-cluster distance given by $\overline{D}_{Wi} = \frac{\sum_{\mathbf{s}^{(k)} \in \mathcal{X}_i} \sum_{\mathbf{s}^{(l)} \in \mathcal{X}_i, l \neq k} d_{kl}^2}{|\mathcal{X}_i|(|\mathcal{X}_i|-1)}$, the total average pairwise distance of all sessions, and the inter-cluster or between-cluster distance given by $\overline{D}_{Bij} = \frac{\sum_{\mathbf{s}^{(k)} \in \mathcal{X}_i} \sum_{\mathbf{s}^{(l)} \in \mathcal{X}_j} d_{kl}^2}{|\mathcal{X}_i||\mathcal{X}_j|}$ to

evaluate the results. It is important to recall that all distances are in $[0, 1]$.

4.2 Examples of Profile Extraction

We applied RFC-MDE to log data of the Web sites of two departments at two different universities. For space reasons, we only describe the results on the log data of the Web site for the Computer Science department at the University of Maryland, Baltimore County. We used data during a period of 6 hours in the late evening. After filtering out irrelevant entries, the data was segmented into 678 sessions. The maximum elapsed time between two consecutive accesses in the same session was set to 45 minutes. The number of distinct URLs accessed in valid entries was 1681. After clustering the relational data with RF$C-$MDE and $C = 35$, the sessions were assigned to the clusters in a minimum distance classifier sense. Only 12 clusters that had cardinalities exceeding 10 were kept, i. e., those making a sufficiently strong profile. Table 1 illustrates four profiles computed using (11), where only the significant URLs ($P_{ij} \geq 0.15$) are displayed, and the individual components are displayed in the format $\{P_{ij}$ - j^{th} URL$\}$. Table 2 lists the cardinality and the intra-cluster distance for the clusters.

The results show that all clusters correspond to real profiles reflecting leisurely users of the World Wide Web in the late evening, as most profiles reflect an interest in free internet games offered by some of the department's graduate students (Profiles No. 2, 5, 6, and 8). Also profile No. 10 reflects an interest in a large set of pictures of a famous supermodel, posted on a graduate student's homepage. Many sessions that do not belong to any profile are lumped in the 11^{th} profile which is easily recognized as a spurious cluster by using the quantitative evaluation measures. This particular cluster had no significant URLs ($P_{ij} < 0.15$ for all j) and its intra-cluster distance (0.99) was higher than the total average pairwise distance of all sessions (0.98). In addition to the intra-cluster distance, the robust weights, w_{ij}, are extremely useful for extracting the core members of each class, as well as for distinguishing between strong and spurious profiles. For example when only sessions with weights exceeding 0.029 are considered, only profiles Nos. 5, 6, 8, 9, 10, and 12 make strong profiles with more than 10 members.

The Non Euclidean Relational Fuzzy $C-$Means (NERF) [11] was also used to cluster the sessions relation matrix, and resulted in only 4 significant profiles, as shown in Table 3. RF$C-$MDE fared better than NERF in the sense that the spurious cluster (No. 3) found by NERF, which corresponds to the 11^{th} profile found by RF$C-$MDE, had 264 more sessions. Also, NERF completely missed the clusters corresponding to the 3^{rd}, 4^{th}, 7^{th}, 8^{th}, 9^{th}, and 10^{th} profiles, and lumped the 2^{nd}, 6^{th} profiles found by RF$C-$MDE into a single cluster (this corresponds to the 3^{rd} profile found by NERF). Moreover, as can be inferred from the higher cardinality and average intra-cluster distance for comparable profiles (such as the 5^{th} and 1^{st} profiles found by RF$C-$MDE and NERF respectively), NERF's clusters tend to contain more irrelevant sessions or noise. This is mainly because the data is noisy, and NERF is not designed to handle such data.

We note that as a by-product of the clustering process, associations between different URL addresses on the UMBC site can easily be inferred from the resulting robust profiles. In general, the URLs that are present in the same profile tend to be visited together in the same session. For example, by looking at the 8^{th} profile in Table 1, we can deduce that the URLs making up that profile tend to co-occur to form a large *item set* [7].

5 Conclusion

In this paper, we have presented a new approach for automatic discovery of user session profiles in Web log data. We defined the notion of a "user session" as being a temporally compact sequence of Web accesses by a user. A new similarity measure to analyze session profiles is presented which captures both the individual URLs in a profile as well as the structure of the site. The RF$C-$MDE algorithm was successfully used to cluster the sessions extracted from real server access logs into typical user session profiles, and even to identify the noisy sessions and profiles. As a by-product of our clustering process, associations between different URL addresses on a given site can easily be inferred from the resulting robust profiles.

Acknowledgments

This work was partially supported by cooperative NSF awards (IIS 9801711 and IIS 9800899) to Joshi and Krishnapuram respectively, and an IBM faculty development award to A. Joshi.

References

[1] Firefly, "http://www.firefly.com"

[2] A. Joshi, S. Weerawarana, and E. Houstis, "On disconnected browsing of distributed information," *Proc. Seventh IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE)*, pp. 101-108, 1997.

[3] L. Terveen, W. Hill, and B. Amento, "PHOAKS - A system for sharing recommendations," *Comm. ACM*, **40**:3, 1997.

[4] R. Armstrong, D. Freitag, T. Joachims, and T. Mitchell, "WebWatcher: A learning apprentice for the World Wide Web," *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, March, 1995.

[5] C. Shahabe, A. M. Zarkesh, J. Abidi and V. Shah, "Knowledge discovery from user's Web-page navigation," *Proc. Seventh IEEE Intl. Workshop on Research Issues in Data Engineering (RIDE)*, pp. 20-29, 1997.

[6] B. Mobasher, N. Jain, E-H. Han, and J. Srivastava "Web mining: Pattern discovery from World Wide Web transactions," *Technical Report 96-050, University of Minnesota*, Sep, 1996.

[7] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," *Proc. of the 20th VLDB Conference*, pp. 487-499, Santiago, Chile, 1994.

[8] U. Fayad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, ed. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1996.

[9] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981.

[10] R. J. Hathaway, J. W. Davenport and J. C. Bezdek, "Relational duals of the c-means algorithms," *Pattern Recognition*, vol. 22, pp. 205-212, 1989.

[11] R. J. Hathaway and J. C. Bezdek, "NERF c-Means: Non-Euclidean relational fuzzy clustering," *Pattern Recognition*, vol. 27, No. 3, pp. 429-437, 1994.

[12] O. Nasraoui and R. Krishnapuram, "A robust estimator based on density and scale optimization, and its application to clustering," *Proc. FUZZIEEE*, New Orleans, 1996, pp. 1031-1035.

Table 1. Profile Examples

i	$\mathbf{P}_i$
2	{.25 - ∼mshadl1/Other_Links.html} {.75 - ∼mshadl1/profiler.html} {.17 - ∼mshadl1/Episode_Guide.html}
4	{.24 - /courses/undergraduate/201/fall98/lectures/ index.shtml} {.26 - /courses/undergraduate/201/fall98/ projects/index.shtml} {.21 - /courses/undergraduate/ 201/fall98/projects/p4/index.shtml} {.21 - /courses/undergraduate/201/fall98}
8	{.68 - ∼sli2/tetris} {.77 - ∼sli2/directory.html} {.67 - ∼sli2/tetris/content.html} {.19 - ∼sli2/cube/content.html}
9	{.45 - /agents} {0.19 - /agents/news}
12	{.98 - /} {.19 - /people/faculty/faculty.shtml}

Table 2. A subset of user sessions clusters

i	$\lvert\mathcal{X}_i\rvert$	description	$\overline{D}_{Wi}$
1	13	Graduate course and degree enquiries (/www/courses/graduate and /www/graduate)	0.78
2	48	Accesses to ∼mshadl1 pages (graduate student offering a page about the Hit TV series "Profiler")	0.49
3	20	Accesses to ∼sletsc1 pages (a page about model ceramic cottages)	0.64
4	58	Inquiries about undergraduate course No. 201	0.81
5	27	Accesses to ∼sli2 pages, particularly (∼sli2/cube) (a free game offered by a graduate student)	0.43
6	20	Accesses to (∼etoton1/games) (a free game offered by another graduate student)	0.27
7	15	Accesses to ∼ebert (professor) pages	0.68
8	79	Accesses to ∼sli2 pages, particularly (∼sli2/tetris) (another free game offered by a graduate student)	0.55
9	58	Inquiries about agents (/agents) (The Intelligent Software Agents page)	0.71
10	19	Accesses to (∼rmobar/Claudia.html) (Supermodel pictures offered by a graduate student)	0.35
11	200	Mixture of unrelated accesses that don't make a strong profile	0.99
12	42	Short sessions mostly limited to main page and faculty list	0.66

Table 3. A subset of user sessions clusters

i	$\lvert\mathcal{X}_i\rvert$	description	$\overline{D}_{Wi}$
1	13	Graduate enquiries (/www/courses/graduate) and (/www/graduate)	0.78
2	112	Accesses to ∼sli2 pages, particularly (∼sli2/cube)	0.62
3	66	Accesses to (∼mshadl1) and (/ etoton1/games)	0.67
4	464	Mixture of unrelated accesses that don't make a strong profile	0.99
5	35	Short sessions mostly limited to main page	0.58