Merging Algorithm to Reduce Dimensionality in Application to Web-Mining

Vladimir Nikulin[1] and Geoffrey J. McLachlan[2]

[1] Airservices Australia,
{vladimir.nikulin@airservicesaustralia.com}
[2] Department of Mathematics, University of Queensland
{gjm@maths.uq.edu.au}

Abstract. Dimensional reduction may be effective in order to compress data without loss of essential information. Also, it may be useful in order to smooth data and reduce random noise. The model presented in this paper was motivated by the structure of the *msweb* web-traffic dataset from the *UCI* archive. It is proposed to reduce dimension (number of the used web-areas or *vroots*) as a result of the unsupervised learning process maximizing a specially defined average log-likelihood divergence. Two different web-areas will be merged in the case if these areas appear together frequently during the same sessions. Essentially, roles of the web-areas are not symmetrical in the merging process. The web-area or *cluster* with bigger weight will act as an attractor and will stimulate merging. In difference, the smaller cluster will try to keep independence. In both cases the powers of attraction or resistance will depend on the weights of the corresponding clusters. The above strategy will prevent creation of one super-big cluster, and will help to reduce number of non-significant clusters. The proposed method is illustrated using two synthetic examples. The first example is based on an ideal *vlink* matrix, which characterizes weights of the *vroots* and relations between them. The *vlink* matrix for the second example is generated using specially designed web-traffic simulator.

Key words: distance-based clustering, data compression, log-likelihood, web-traffic data

1 Introduction

A general problem faced in computer science is to reduce the dimensions of a large datasets in order to make sense of the information contained in them [1].

The main model and approach of this paper were motivated by the *msweb* dataset that corresponds to the visits to a set of areas (*vroots*) of the Microsoft corporate web-site. This dataset is publicly available through the *UCI KDD* Archive at the University of California [2]. Given a significantly high number of *vroots* and low average number of different pages visited during one separate session, we are interested to group pages into relatively homogeneous clusters in

order to avoid sparse tables. For example, [3] considered grouping according to the logically sensible approach. Another approach may be based on statistical methods: for example, we can consider projection pursuit with such special cases as principal component, discriminant, and factor analyses [4]. The corresponding methods optimize in some sense the linear transformation from the given to the known low-dimensional space. However, in practice, the dimension or number of clusters may not be known [5].

Traditional web-clickstreams data-structure [5] represents a sequence of web-pages, which clients visited during particular sessions (variable length data, see, for example, [6]). Note that the structure of the *msweb* dataset is essentially different: for any particular session each *vroot* was characterized as being visited (vote one) or not visited (vote zero). It appears to be reasonable not to make two different clicks equivalent as we do not know how much time a user spent considering corresponding web-pages (the time-range may vary from a few seconds to several minutes).

According to [7] and [8], collaborative filtering may be useful in order to predict the utility of *vroot* to a particular user based on a database of user votes considering vote zero in *msweb* dataset as a hidden or missing. For example, singular value decomposition is regarded as one of the most popular tools of collaborative filtering.

The proposed unsupervised clustering approach is based on the *vlink* matrix (1), and is presented in the following Sect. 2. Section 3 illustrates the main idea behind the proposed method using two synthetic examples. Importantly, further application of the same algorithm with the same settings against *msweb* dataset produced the same graphical structure of the target function (see Fig. 2(d) and Sect. 4).

As a next step after dimensional reduction we can consider the problem of predicting a user's behavior on a web-site, which has gained importance due to the rapid growth of the world-wide-web and the need to personalize and influence a user's browsing experience [9]. Markov models and their variations have been found well suited for addressing this problem. In general, the input for these problems is the sequence of web-pages that were accessed by a user and the goal is to build Markov models that can be used to model and predict the web-page that the user will most likely access next. This study will help to explore and understand human behavior within internet environment [10], [11].

2 The model

Suppose we have a dataset $\mathbf{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of n records of web areas (classified into m different areas or *vroots*), which users visited during one session: $\mathbf{x}_j := \{x_{ij}, i = 1..m\}$ where $x_{ij} = 1$ if j-user visited area i, alternatively, $x_{ij} = 0$.

Assuming that $\mathbf{x}_j$ is a vector-column, we form *vlink* matrix

$$\mathcal{S} = \sum_{j=1}^{n} \mathbf{x}_j \cdot \mathbf{x}_j^T = \{s_{ik}, i, k = 1..m\} \tag{1}$$

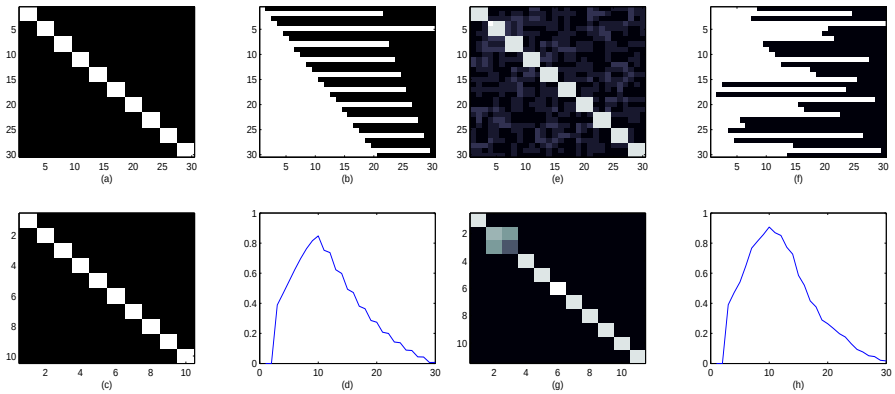

Fig. 1. *vlink* matrices (a) $\mathcal{S}_0$ and (e) $\mathcal{S}_\star$; (c) *vlink* matrix, which corresponds to the peak of d); (g) *vlink* matrix, which corresponds to the 19th step -one step before the peak of (h) (note that dark area of the image (g) is not uniform as it may be understood); (d, h): behavior of (5) as a function of the number of clusters, the following parameters were used $m_0 = 30, \gamma = 0.2, \tau = 3, \alpha = 0.005, \beta = 0.00001, \varphi = 0.0001$. In (b) and (f) vertical axis represents *vroot*, horizontal axis represents step of the merging process: light colour was changed to dark colour when the corresponding *vroot* was absorbed by another cluster.

where high values of $z_{ik} = s_{ik} \left(s_{ii} \cdot s_{kk}\right)^{-0.5}, i \neq k$, indicate higher similarity between areas i and k, the value of s_{ii} may be used as a measurement of the weight of the area i, and we employ the assumption

$$s_{ii} = \max_k s_{ik} \geq 1 \quad \forall i = 1..m. \tag{2}$$

We can make a conclusion that i-vroot was always accompanied by the $j-vroot$ if $s_{ii} = s_{ij}$. Accordingly, we will call i and $j-vroots$ as equivalent if $s_{ii} = s_{ij} = s_{jj}$. Figure 1(a) illustrates an example of *vlink* matrix where first three rows/columns represent equivalent *vroots*.

We form the matrix of probabilities $P = \{p_{ik}, i, k = 1..m\}$, where

$$p_{ik} = \begin{cases} 0 \quad \text{if} \quad i = k \quad \text{or} \quad C_i = \sum\limits_{\substack{k=1 \\ k \neq i}}^{m} s_{ik} = 0; \\ \dfrac{s_{ik}}{C_i}, \quad \text{otherwise.} \end{cases}$$

Remark 1. The probabilistic component p_{ik} indicates similarity between rows (or corresponding *vroots*) i and k. As a result of the setting $p_{ii} = 0$, we exclude from the definition of the following below target function (5) weights of the clusters.

Table 1. Merging process with an ideal initial *vlink* matrix $\mathcal{S}_0$ (see Figure 1(a)).

Step	$\mathcal{D}(\mathcal{S},\alpha,\beta)$	m	Attractor	2nd *vroot*	Step	$\mathcal{D}(\mathcal{S},\alpha,\beta)$	m	Attractor	2nd *vroot*
	0.007076	30			15	0.493699	15	23	22
1	0.006892	29	2	1	16	0.598496	14	23	24
2	0.043250	28	2	3	17	0.623234	13	26	25
3	0.044397	27	5	4	18	0.736680	12	26	27
4	0.086353	26	5	6	19	0.752875	11	29	28
5	0.089262	25	8	7	20	**0.848217**	10	29	30
6	0.137851	24	8	9	21	0.815838	9	5	2
7	0.143080	23	11	10	22	0.763540	8	5	8
8	0.199645	22	11	12	23	0.698259	7	5	11
9	0.207907	21	14	13	24	0.624053	6	5	14
10	0.274143	20	14	15	25	0.545006	5	5	17
11	0.286260	19	17	16	26	0.466740	4	5	20
12	0.364011	18	17	18	27	0.388467	3	5	23
13	0.380830	17	20	19	28	0.000000	2	5	26
14	0.471866	16	20	21	29	0.000000	1	5	29

We are interested to maximize information (or minimize similarity) per unit cluster independently on the cluster's weights using an average symmetrical log-likelihood divergence (5). We will use the log-likelihood function in order to measure distance between i and k web-areas,

$$d_{ik} = \sum_{\substack{v=1 \\ v\neq i,k}}^{m} \xi_{ikv}, \tag{3}$$

where

$$\xi_{ikv} = \begin{cases} -p_{iv}\cdot\log p_{kv} - p_{kv}\cdot\log p_{iv} & \text{if } \; p_{iv}, p_{kv} \geq \alpha; \\ \beta & \text{otherwise,} \end{cases} \tag{4}$$

and where $\alpha > 0$ and $\beta \geq 0$ are regulation parameters. Accordingly, the average distance will be defined as

$$\mathcal{D}(\mathcal{S},\alpha,\beta) = A(m)\sum_{i=1}^{m-1}\sum_{k=i+1}^{m} d_{ik}, \tag{5}$$

where

$$A(m) = \frac{1}{m(m-1)\log(m)}, m \geq 3, \tag{6}$$

is a norm coefficient. Note that the multiplier $\log(m)$ in the denominator of (6) corresponds directly to the maximum value of the Entropy function; $\mathcal{D}(\mathcal{S},\alpha,\beta) = 0, 1 \leq m \leq 2$, according to the definition (3).

Remark 2. In the above definition we excluded probabilities with small values considering them as a noise.

Figures 1(b), 1(f) and 2(b) illustrate merging process: the absorbed cluster changed color from light to dark.

Algorithm 1. Merging process.

1: Initial setting: $k_i = i, i = 1..m$, where m is a size of the squared matrix $\mathcal{S}$ defined in (1).
2: Find preferable pair for merging maximizing

$$\max\{\frac{s_{k_ik_i}^{\gamma}\cdot(z_{k_ik_j}+\varphi)}{s_{k_jk_j}^{\tau}}, \frac{s_{k_jk_j}^{\gamma}\cdot(z_{k_ik_j}+\varphi)}{s_{k_ik_i}^{\tau}}\}, i, j = 1..m, i \neq j \qquad (7)$$

where $z_{k_ik_j} = \frac{s_{k_ik_j}}{\sqrt{s_{k_ik_i}s_{k_jk_j}}}$; τ, γ and φ are positive regulation parameters.
3: Suppose that $s_{k_ik_i} \geq s_{k_jk_j}$. Then,

$$s_{k_ik_v} := s_{k_ik_v} + s_{k_jk_v}, v = 1..m; s_{k_vk_i} := s_{k_vk_i} + s_{k_vk_j}, v = 1..m, v \neq i;$$

$$k_v = k_v + 1, v = j..m.$$

In the alternative case ($s_{k_jk_j} > s_{k_ik_i}$)

$$s_{k_jk_v} := s_{k_jk_v} + s_{k_ik_v}, v = 1..m; s_{k_vk_j} := s_{k_vk_j} + s_{k_vk_i}, v = 1..m, v \neq j;$$

$$k_v = k_v + 1, v = i..m.$$

4: $m := m - 1$, and go to the Step 2 if $m \geq 3$.

Remark 3. The main target of the parameter φ is to link small and isolated web-areas to other web-areas.

Definition 1. *We denote the size of the 1) initial vlink matrix $\mathcal{S}_0$ by m_0, 2) current vlink matrix $\mathcal{S}$ by $m(\mathcal{S})$ or simply m.*

Essentially, Algorithm 1 is based on the original indices k_i which may not be sequential as a result of the merging process (in difference to the sequential secondary index $i = 1..m$). These indices may be seen in the columns "Attractor" and "2nd *vroot*" of the Tables 1.

3 Illustration of the main idea using an ideal synthetic example

In order to simplify notations and without loss of generality we assume that (1) clusters have equal size, and (2) all *vroots* within any particular cluster have sequential indices.

Definition 2. *Let us denote by $Q(v,k)$ the following 2D set of indices:*

$$\begin{cases} i = v \cdot h + u, \\ j = i - u + 1..i - u + v \end{cases} \tag{8}$$

where $u = 1..v$ and $h = 0..k-1$.

Definition 3. *We call squared matrix G as (a,b)-diagonal if*

$$g_{ij} = \begin{cases} a \quad if \quad i = j; \\ b \quad otherwise. \end{cases}$$

We call m-dimensional squared matrix G as $(v;a,b)$-diagonal if $m = v \cdot k$ where k is a natural number, and

$$g_{ij} = \begin{cases} a \quad if \quad i \in Q(v,k); \\ b \quad otherwise. \end{cases}$$

Note that in the $(v;a,b)$-diagonal matrix a value of a significantly larger compared to b represents an ideal case of vlink matrix. Figure 1(a) represents an illustration of $(3,5000,1)$-diagonal matrix, which corresponds to the case of $k = 10$ clusters. In more details, $k = 10$ small white squares $q_{v,h}$ with size $v = 3$ (see definition (8): $Q(v,k) = \cup_{h=0}^{k-1} q_{v,h}$) correspond to the value $a = 5000$; all other black elements of the matrix $\mathcal{S}_0$ correspond to the value $b = 1$.

Proposition 1. *Suppose that $\mathcal{S}$ is $(v;a,b)$-diagonal matrix, $v \geq 2$, and*

$$\frac{b}{(v-1)a + (m-v)b} < \alpha \leq \frac{a}{(v-1)a + (m-v)b}. \tag{9}$$

Then

$$\mathcal{D}(\mathcal{S},\alpha,\beta) = -\frac{(v-1)(v-2)\left[\psi(Z_m) + 0.5\beta\right]}{(m-1)\log m} + \frac{(m-2)\beta}{2\log m} \tag{10}$$

where $\psi(Z_m) = Z_m \log Z_m$, $Z_m = \frac{a}{(v-1)a+(m-v)b}$.

Proof. By definition $\mathcal{D}$ represents a sum with $0.5m(m-1)(m-2)$ terms. These terms may be split into 2 parts: (1) significant components (SC) with value $-2 \cdot \psi(Z_m)$ and (2) noise components (NC) with value β.

The size of the first group is $0.5m(v-1)(v-2)$. Similarly, the second group includes $0.5m((m-1)(m-2) - (v-1)(v-2))$ elements. □

Proposition 2. *Suppose that $\mathcal{S}$ is (a,b)-diagonal matrix, $m \geq 2$, and*

$$\alpha \leq \frac{1}{m-1}. \tag{11}$$

Then

$$\mathcal{D}(\mathcal{S},\alpha,\beta) = B_1(m) = \frac{(m-2)\log(m-1)}{(m-1)\log m}. \tag{12}$$

Proof. Similarly, as in the proof of the Proposition 1, $\mathcal{D}$ represents a sum with $0.5m(m-1)(m-2)$ uniform terms. The value of one particular term is $\frac{2\log(m-1)}{m-1}$. The required formula will be obtained as a product of the above two values multiplied by the norm coefficient (6). □

3.1 The main idea

Let us consider the simplified ideal case. Suppose that the *vlink* matrix may be effectively approximated by the $(v; a, b)$-diagonal matrix. Then, we can use formula (10) for divergence (5), which includes two terms (subject to the condition $v \geq 3$): 1) SC, which represents a decreasing function of m; 2) NC, which represents an increasing function of m. Assuming that the parameter β is small enough or equal to zero (means NC component is much smaller compared to the SC component) the divergence $\mathcal{D}$ will grow as a result of the sequence of merging operations. The growing process will continue until the corresponding *vlink* matrix $\mathcal{S}$ will take the (a, b)-diagonal shape (means $\mathcal{S}$ will be close to the (a, b)-diagonal shape). Figure 1(a) illustrates the initial *vlink* matrix with $(v; a, b)$-diagonal structure, and Figure 1(e) illustrates a matrix, which is close to the $(v; a, b)$-diagonal structure. As a result of the sequence of merging operations, these matrices will be transformed to the (a, b)-diagonal matrix (see Figure 1(c)). Figure 1(g) represents a very important case just one step before the peak. After the peak, the target function $\mathcal{D}$ will decline according to Proposition 2.

3.2 Web-traffic simulator

Vlink matrix $\mathcal{S}_\star$ (see Figure 1(e)) for the second experiment was produced using Algorithm 2 with $T = \mathcal{S}_0$ and $E = 500$. Firstly, we simulated $n = 5000$ web-traffic records. Then, we computed $\mathcal{S}_\star$ according to (1).

Algorithm 2. Web-traffic simulator (repeats of *vroots* within any particular record are not allowed).

1: Order m - number of web-areas; T - *vlink* matrix (squared matrix with size m and non-negative elements) and E - exit weight.
2: Form vector of prior probabilities $q_i \propto T_{ii}, i = 1..m$, and draw initial web-area j_1 according to q_i using uniformly distributed random variable.
3: Draw second web-area $j_t, t = 2$, according to the probabilities proportional to the j_1 row of the matrix T where j_1 *vroot* was excluded, and exit weight was added as a last element of the vector.
4: Stop the algorithm if $j_t = m - t + 2$ (exit index), alternatively, go to the next step.
5: $t := t + 1$; form vector of probabilities proportional to the *minimal* values of rows $j_k, k = 1..t$, where columns $j_k, k = 1..t$, are excluded (no repeats are allowed), and exit weight is added as a last element of the vector.
6: Draw web-area j_t and go to the step 4.

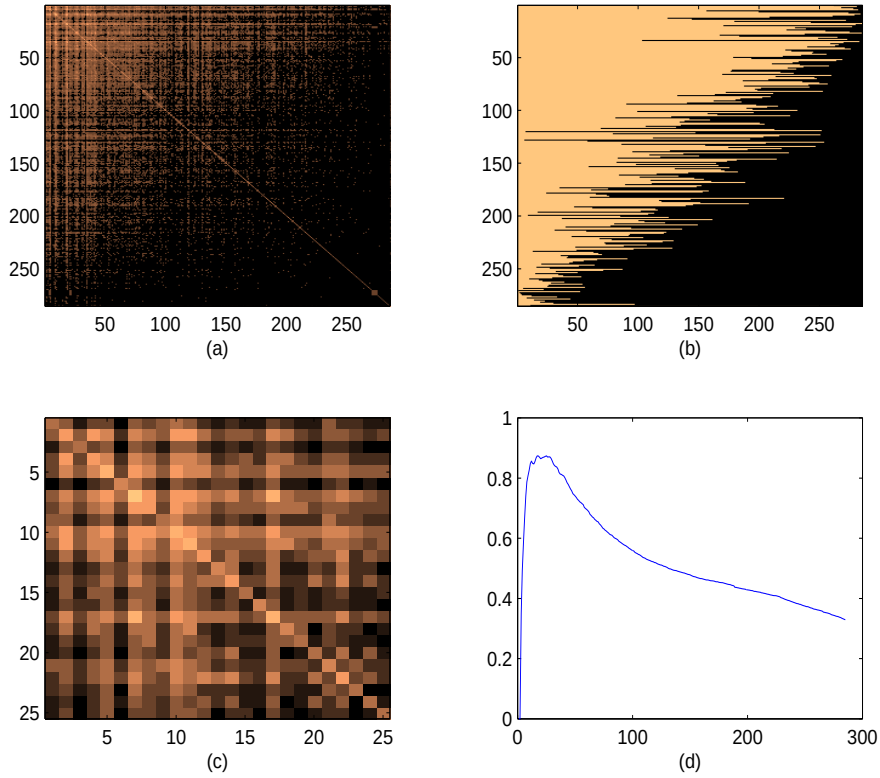

Fig. 2. (a) *vlink* matrix $\mathcal{S}_{mw}$ for *msweb* dataset; (b) merging process (see, also, Figure 1); (c) *vlink* matrix which corresponds to $m = 25$ - peak of the graph d); the following parameters were used in (d): $m_0 = 285, \gamma = 0.2, \tau = 3, \alpha = 0.005, \beta = 0.00001, \varphi = 0.0001$.

4 Experiments on the *msweb* dataset

Msweb dataset [2] includes 32711 records and 294 *vroots*. Table 2 represents the most frequent *vroots*.

The following procedure was used in order to produce sequential secondary indices I_S out of original indices I_O:

$$I_S = \begin{cases} I_O - 999 & \text{if} \quad 1000 \le I_O \le 1046; \\ I_O - 1000 & \text{if} \quad 1048 \le I_O \le 1284; \\ I_O - 1002 & \text{if} \quad 1287 \le I_O \le 1295; \\ I_O - 1003 & \text{if} \quad I_O = 1297. \end{cases} \tag{13}$$

We reduced the number of *vroots* to 285 because 9 *vroots* (NN285-292 and N294) were not used. In average, there are 3.016 *vroots* per one record with standard

Table 2. List of the most frequent web-areas; I_O -original index; I_S -secondary index (13); column m indicate number of clusters when corresponding *vroot* appeared in the last time. For example, *vroot* "Products" was the winner of the merging process.

Number of repeats	I_O	I_S	Name of *vroot*	m
10837	1008	9	Free Downloads	3
9383	1034	35	Internet Explorer	4
8463	1004	5	Microsoft.com Search	5
5330	1018	19	isapi	7
5108	1017	18	**Products**	1
4628	1009	10	Windows Family of Oss	8
4451	1001	2	Support Desktop	6
3220	1026	27	Internet Site Construction for Developers	2
2968	1003	4	Knowledge Base	9
2123	1025	26	Web Site Builder's Gallery	12
1791	1035	36	Windows95 Support	17
1506	1040	41	MS Office Info	16
1500	1041	42	Developer Workshop	11
1446	1032	33	Games	14
1160	1037	38	Windows 95	21
1115	1030	31	Windows NT Server	22
1110	1038	39	SiteBuilder Network Membership	24
1087	1020	21	Developer Network	10
912	1000	1	regwiz	25
865	1007	8	International IE content	13
842	1052	52	MS Word News	18
759	1036	37	Corporate Desktop Evaluation	28
749	1002	3	End User Produced View	19

deviation 2.5 and maximum number of *vroots* 35. Figure 2(a) illustrates the *vlink* matrix $\mathcal{S}_{mw}$, which was computed according to *msweb* data.

Remark 4. The structure of the graph Figure 2(d) is remarkably similar compared with graphs Figure 1(d) and (h). Although, image Figure 2(c) is much "smoother" compared to Figure 1(c).

5 Concluding Remarks

The proposed method was tested successfully against an ideal synthetic *vlink* matrix with known solution. As a next step, we considered a more complex and realistic case: we generated synthetic web-traffic data and computed the corresponding *vlink* matrix. Again, the automatical system produced the correct answer considering the inverse task.

Then, we applied the same system with identical regulation parameters to the real *msweb* dataset. As a result of the merging process the target function (5) grows initially to the point $m \approx 25$, then it declines to zero. This is in line

with the main computations for the system produced transformation (merging) function. Using this function we can compress the original dataset with 285 *vroots* to the size of only 25.

The presented system is general and may be used elsewhere. For example, we can consider such areas as author-topic [12] or movie [8] classification/clustering.

Dimensional reduction will open prospects to conduct further research using such sophisticated and computationally expensive techniques as *variational inference* [13] or *universal clustering* [14], which could be effective for detecting the number of significant clusters, and for analyzing the stability of the clustering configuration in large datasets of internet users.

References

[1] Botelho, S., Lautenschlger, W., de Figueiredo, M.B.: Dimensional reduction of large image datasets using non-linear principal components. In Gallagher, M., Hogan, J., Maire, F., eds.: IDEAL 2005. Volume LNCS 3578., Springer-Verlag (2005) 125–132

[2] Msweb: msweb anonymous web data. In: UCI Knowledge Discovery in Databases Archive: http://kdd.ics.uci.edu/databases/msweb/msweb.html. (1998)

[3] Giudici, P., Castelo, R.: Association models for web mining. Data Mining and Knowledge Discovery **5** (2001) 183–196

[4] Huber, P.: Projection pursuit. The Annals of Statistics **13** (1985) 435–475

[5] Nasraoui, O., Cardona, C., Rojas, C.: Using retrieval measures to assess similarity in mining dynamic web clickstreams. In: KDD'05, August 21-24, Chicago, Illinois, USA. (2005)

[6] Msnbc: msnbc.com anonymous web data. In: UCI Knowledge Discovery in Databases Archive: http://kdd.ics.uci.edu/summary.data.type.html. (1999)

[7] Breese, J., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of 20th Conference on Uncertainty in Artificial Intelligence (UAI). (1998)

[8] Zitnick, C., Kanade, T.: Maximum entropy for collaborative filtering. In: Proceedings of 14th Conference on Uncertainty in Artificial Intelligence (UAI). (2004)

[9] Deshpande, M., Karypis, G.: Selective markov models for predicting web-page accesses. In: Proceedings of the First SIAM International Conference on Data Mining, Chicago, USA, April 5-7, 2001, SIAM (2001)

[10] Cadez, I., Heckerman, D., Meek, C., Smyth, P., White, S.: Model-based clustering and visualization of navigation patterns on a web site. Data Mining and Knowledge Discovery **7** (2003) 399–424

[11] Nikulin, V., Smola, A.: Parametric model-based clustering. In Dasarathy, B., ed.: Data Mining, Intrusion Detection, Information Assurance, and Data Network Security, 28-29 March 2005, Orlando, Florida, USA. Volume 5812., SPIE (2005) 190–201

[12] Smyth, M.S.P., Griffiths, T.: Probabilistic author-topic models for information discovery. In: KDD'04, August 22-25, Washington, USA. (2004)

[13] Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocations. Journal of Machine Learning Research **3** (2003) 993–1022

[14] Nikulin, V.: On the universal clustering under a broad class of loss functions. International Journal of Neural Systems **16** (2006) 329–339