Dimensional Reduction of Web-Traffic Data

Vladimir Nikulin

Computer Sciences Laboratory, RSISE, Australian National University, Canberra, Australia

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 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. Above strategy will prevent creation of one super-big cluster, and will help to reduce number of non-significant clusters. The proposed method was 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 was generated using specially designed web-traffic simulator.

Keywords: 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 bulk 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, Ref.3 considered grouping according to the logically sensible approach. Another approach may be based on the statistical methods: for example, we can consider projection pursuit with such special cases as principal component, discriminant and factor analysis.\(^4\) Methods based on the projection pursuit approach optimize in some sense linear transformation from the given to the known low-dimensional space. However, in practice, the dimension or number of clusters may not be known.\(^5\) This model is called unsupervised clustering.

Traditional web-clickstreams data-structure\(^6\) represents a sequence of web-pages which client visited during particular session (variable length data, see, for example, Ref.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 until we do not know how much time user spent considering corresponding web-pages (the time-range may vary from a few seconds to several minutes).

Ref.7, 8 used collaborative filtering 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.

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. 3(d) and Sect. 4).

As a next step after dimensional reduction we can consider the problem of predicting 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. 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.

2. THE MODEL

Suppose we have a dataset $\mathbf{X} := \{x_1, \ldots, x_n\}$ of $n$ records of web areas (classified into $m$ different areas or vroots) which users visited during one session: $x_j := \{x_{ij}, i = 1..m\}$ where $x_{ij} = 1$ if $j$-user visited area $i$, alternatively, $x_{ij} = 0$.

\[ S = \sum_{j=1}^{n} x_j \cdot x_j^T = \{s_{ik}, i, k = 1..m\} \]

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

\[ s_{ii} = \max_k s_{ik} \geq 1 \quad \forall i = 1..m. \]

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 equivalent if $s_{ii} = s_{ij}$. Figure 1(a) illustrates an example of $vlink$ matrix where first three rows/columns represent equivalent vroots.

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

\[
p_{ik} = \begin{cases} 
0 & \text{if } i = k \text{ or } C_i = \sum_{k=1}^{m} s_{ik} = 0; \\
\frac{s_{ik}}{C_i} & \text{otherwise.}
\end{cases}
\]
REMARK 2.1. The probabilistic component \( p_{ik} \) indicates similarity between rows (or corresponding \( \text{vroots} \) \( i \) and \( k \)). As a result of the setting \( p_{ii} = 0 \) we will exclude from the definition of the following below target function (5) weights of the clusters.

Table 1. First example with an ideal initial \( \text{vlink} \) matrix \( S_0 \) (see Figure 1(a)).

<table>
<thead>
<tr>
<th>Step</th>
<th>( \mathcal{D}(S, \alpha, \beta) )</th>
<th>( m )</th>
<th>Attractor</th>
<th>2nd ( \text{vroot} )</th>
<th>Step</th>
<th>( \mathcal{D}(S, \alpha, \beta) )</th>
<th>( m )</th>
<th>Attractor</th>
<th>2nd ( \text{vroot} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.007076</td>
<td>30</td>
<td></td>
<td></td>
<td>15</td>
<td>0.493699</td>
<td>15</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>1</td>
<td>0.006892</td>
<td>29</td>
<td>2</td>
<td>1</td>
<td>16</td>
<td>0.598496</td>
<td>14</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>2</td>
<td>0.043250</td>
<td>28</td>
<td>2</td>
<td>3</td>
<td>17</td>
<td>0.623234</td>
<td>13</td>
<td>26</td>
<td>25</td>
</tr>
<tr>
<td>3</td>
<td>0.044397</td>
<td>27</td>
<td>5</td>
<td>4</td>
<td>18</td>
<td>0.736680</td>
<td>12</td>
<td>26</td>
<td>27</td>
</tr>
<tr>
<td>4</td>
<td>0.086353</td>
<td>26</td>
<td>5</td>
<td>6</td>
<td>19</td>
<td>0.752875</td>
<td>11</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>5</td>
<td>0.089262</td>
<td>25</td>
<td>8</td>
<td>7</td>
<td>20</td>
<td>0.848217</td>
<td>10</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>0.137851</td>
<td>24</td>
<td>8</td>
<td>9</td>
<td>21</td>
<td>0.815838</td>
<td>9</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td>0.143080</td>
<td>23</td>
<td>11</td>
<td>10</td>
<td>22</td>
<td>0.763540</td>
<td>8</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>8</td>
<td>0.199645</td>
<td>22</td>
<td>11</td>
<td>12</td>
<td>23</td>
<td>0.698259</td>
<td>7</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>9</td>
<td>0.207907</td>
<td>21</td>
<td>14</td>
<td>13</td>
<td>24</td>
<td>0.624053</td>
<td>6</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>10</td>
<td>0.274143</td>
<td>20</td>
<td>14</td>
<td>15</td>
<td>25</td>
<td>0.545006</td>
<td>5</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>11</td>
<td>0.286260</td>
<td>19</td>
<td>17</td>
<td>16</td>
<td>26</td>
<td>0.466740</td>
<td>4</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>0.364011</td>
<td>18</td>
<td>17</td>
<td>18</td>
<td>27</td>
<td>0.388467</td>
<td>3</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>0.380830</td>
<td>17</td>
<td>20</td>
<td>19</td>
<td>28</td>
<td>0.000000</td>
<td>2</td>
<td>5</td>
<td>26</td>
</tr>
<tr>
<td>14</td>
<td>0.471866</td>
<td>16</td>
<td>20</td>
<td>21</td>
<td>29</td>
<td>0.000000</td>
<td>1</td>
<td>5</td>
<td>29</td>
</tr>
</tbody>
</table>

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

\[
d_{ik} = \sum_{v=1}^{m} \xi_{ikv} \tag{3}
\]

where

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

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

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

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. Note that \( \mathcal{D}(S, \alpha, \beta) = 0 \), \( 1 \leq m \leq 2 \), according to the definition (3).

REMARK 2.2. In the above definition we excluded probabilities with small value considering them as a noise.

Figures 1(b), 1(f) and 3(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 $S$ defined in (1).
2: Find preferable pair for merging according to the maximum of

$$\max \left\{ \frac{s_{k_i,k_j}}{s_{k_i,k_j}} \cdot (z_{k_i,k_j} + \varphi), \frac{s_{k_i,k_j}}{s_{k_i,k_j}} \cdot (z_{k_i,k_j} + \varphi) \right\}, i, j = 1 .. m, i \neq j, \quad z_{k_i,k_j} = \frac{s_{k_i,k_j}}{\sqrt{s_{k_i,k_i}s_{k_j,k_j}}}, \quad (7)$$

where $\tau, \gamma$ and $\varphi$ are positive regulation parameters.
3: Suppose that $s_{k_i,k_i} \geq s_{k_j,k_j}$. Then,

$$s_{k_i,k_i} := s_{k_i,k_i} + s_{k_j,k_j}, v = 1 .. m; s_{k_i,k_i} := s_{k_i,k_i} + s_{k_i,k_j}, v = 1 .. m, v \neq i; k_v = k_v + 1, v = j.m.$$ 

In the alternative case ($s_{k_j,k_j} > s_{k_i,k_i}$)

$$s_{k_i,k_i} := s_{k_i,k_i} + s_{k_i,k_i}, v = 1 .. m; s_{k_i,k_i} := s_{k_i,k_j} + s_{k_i,k_i}, v = 1 .. m, v \neq j; k_v = k_v + 1, v = i.m.$$ 

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

REMARK 2.3. The main target of the parameter $\varphi$ is to link small and isolated web-areas to other web-areas.

DEFINITION 2.1. We denote the size of the 1) initial vlink matrix $S_0$ by $m_0$, 2) current vlink matrix $S$ by $m(S)$ or simply $m$.

Essentially, the Algorithm 1 is based on the original indexes $k_i$ which may not be sequential as a result of the merging process (in difference to the sequential secondary index $i = 1 .. m$). These indexes may be seen in the columns "Attractor" and "2nd vroot" of the Tables 1, 2 and 4.

### Table 2. Second example with an initial vlink matrix $S_0$ (see Figure 1(e)), which was generated using Algorithm 2.

<table>
<thead>
<tr>
<th>Step</th>
<th>$D(S, \alpha, \beta)$</th>
<th>$m$</th>
<th>Attractor</th>
<th>2nd vroot</th>
<th>Step</th>
<th>$D(S, \alpha, \beta)$</th>
<th>$m$</th>
<th>Attractor</th>
<th>2nd vroot</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.017415</td>
<td>30</td>
<td>15</td>
<td>0.585646</td>
<td>15</td>
<td>19</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.021450</td>
<td>29</td>
<td>17</td>
<td>18</td>
<td>16</td>
<td>0.726832</td>
<td>14</td>
<td>19</td>
<td>21</td>
</tr>
<tr>
<td>2</td>
<td>0.045804</td>
<td>28</td>
<td>17</td>
<td>16</td>
<td>17</td>
<td>0.773128</td>
<td>13</td>
<td>15</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>0.051563</td>
<td>27</td>
<td>26</td>
<td>25</td>
<td>18</td>
<td>0.850579</td>
<td>12</td>
<td>15</td>
<td>14</td>
</tr>
<tr>
<td>4</td>
<td>0.076634</td>
<td>26</td>
<td>26</td>
<td>27</td>
<td>19</td>
<td>0.869567</td>
<td>11</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>0.093552</td>
<td>25</td>
<td>22</td>
<td>23</td>
<td>20</td>
<td>0.906736</td>
<td>10</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>0.131689</td>
<td>24</td>
<td>22</td>
<td>24</td>
<td>21</td>
<td>0.857120</td>
<td>9</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>0.176062</td>
<td>23</td>
<td>2</td>
<td>3</td>
<td>22</td>
<td>0.812840</td>
<td>8</td>
<td>4</td>
<td>22</td>
</tr>
<tr>
<td>8</td>
<td>0.198241</td>
<td>22</td>
<td>2</td>
<td>1</td>
<td>23</td>
<td>0.766944</td>
<td>7</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>9</td>
<td>0.232214</td>
<td>21</td>
<td>7</td>
<td>8</td>
<td>24</td>
<td>0.647712</td>
<td>6</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>10</td>
<td>0.264306</td>
<td>20</td>
<td>7</td>
<td>9</td>
<td>25</td>
<td>0.541396</td>
<td>5</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>11</td>
<td>0.289534</td>
<td>19</td>
<td>11</td>
<td>10</td>
<td>26</td>
<td>0.471386</td>
<td>4</td>
<td>4</td>
<td>26</td>
</tr>
<tr>
<td>12</td>
<td>0.376485</td>
<td>18</td>
<td>11</td>
<td>12</td>
<td>27</td>
<td>0.390505</td>
<td>3</td>
<td>4</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>0.416400</td>
<td>17</td>
<td>29</td>
<td>30</td>
<td>28</td>
<td>0.000000</td>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>14</td>
<td>0.520363</td>
<td>16</td>
<td>29</td>
<td>28</td>
<td>29</td>
<td>0.000000</td>
<td>1</td>
<td>4</td>
<td>29</td>
</tr>
</tbody>
</table>

3. ILLUSTRATION OF THE MAIN IDEA USING AN IDEAL SYNTHETIC EXAMPLE

In order to simplify notations and without loss of generality we will assume that 1) clusters have equal size, and 2) all vroots within any particular cluster have sequential indexes.
Definition 3.1. Let us denote by $Q(v, k)$ the following 2D set of indexes:

$$
\begin{align*}
&i = v \cdot h + u, \\
&j = i - u + 1..i - u + v
\end{align*}
$$

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

Definition 3.2. We call squared matrix $G$ as $(a, b)$-diagonal if

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

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

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

Note that $(v, a, b)$-diagonal matrix where value of $a$ is significantly bigger comparing with value of $b$ 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 $q_{v, b}$ 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 $S_{0}$ correspond to the value $b = 1$.

Proposition 3.1. Suppose that $S$ is $(v, a, b)$-diagonal matrix, $v \geq 2$, and

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

Then

$$
D(S, \alpha, \beta) = -\frac{(v - 1)(v - 2)}{(m - 1) \log m} [Z_m \log Z_m + 0.5\beta] + \frac{(m - 2)\beta}{2 \log m}, \quad Z_m = \frac{a}{(v - 1)a + (m - v)b}.
$$

Proof. By definition $D$ 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 $-2 \cdot Z_m \log Z_m$ and 2) noise components (NC) with value $\beta$.

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

Proposition 3.2. Suppose that $S$ is $(a, b)$-diagonal matrix, $m \geq 2$, and

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

Then

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

Proof. Similar to the proof of the Proposition 3.1, $D$ 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 2 values multiplied by the norm coefficient (6).

Proc. of SPIE Vol. 6241 62410J-5
3.1. The main idea.

Let us consider simplified ideal case. Suppose that vlink matrix may be effectively approximated by \((v; a, b)\)-diagonal matrix. Then, we can use formula (10) for averaged divergence \(D\) which includes 2 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 \(\beta\) is small enough or equal to zero (means \(NC\) component is much smaller comparing with \(SC\) component) divergence \(D\) will grow as a result of the sequence of merging operations. The growing process will continue until the corresponding vlink matrix \(S\) will take \((a, b)\)-diagonal shape (means \(S\) will be closed to \((a, b)\)-diagonal shape). Figure 1(a) may be used as an illustration of the initial vlink matrix \(S\) 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 \((a, b)\)-diagonal matrix (see Figure 1(c)). Figure 1(g) represents a very important case just one step before the peak (this case is not covered by the Proposition 3.2 and will be considered in the following Proposition 3.3). After the peak the target function \(D\) will decline according to the Proposition 3.2.

3.2. Some properties of the model and web-traffic simulator.

In an ideal case: \(S = S_0\), \(m = 30\), \(v = 3\), \(k = 10\), we have \(m\) equivalent options for the first step of the merging process. Suppose that 1st and 2nd \(vroots\) were merged as it is displayed in the Figure 1(b). Then, first two lines of the new vlink matrix \(\tilde{S}\) will not contribute any SCs to the value of the target function assuming that \(\alpha \ll \frac{b - 2m}{m + 1} < \alpha\). Therefore,

\[
D(\tilde{S}, \alpha, \beta) = -2 \cdot (m - 2) \tilde{Z}_m \log \tilde{Z}_m + \beta [0.5m(m - 1)(m - 2) - m + 2] \cdot \log m
\]  

(13)

where \(\tilde{Z}_m = \frac{a}{2a + (m - 1)b}\) (note that \(m\) in (13) represents size of the squared matrix \(\tilde{S}\) which is one unit smaller comparing with size of the initial vlink matrix \(S_0\), for example, \(\tilde{Z}_{m-1} = Z_m\)).

Let us compare SCs of the bounds (10) and (13) assuming that \(v = 3\):

\[
B(m) = -\frac{2Z_m \log Z_m}{(m - 1) \log m}, \quad \tilde{B}(m - 1) = -\frac{2(m - 3)Z_m \log Z_m}{(m - 1)(m - 2) \log (m - 1)}, \quad Z_m = \frac{a}{2a + (m - 3)b}.
\]

The following relation take place

\[
-\frac{m - 1}{2Z_m \log Z_m} [B(m) - \tilde{B}(m - 1)] = \frac{1}{\log m} - \frac{m - 3}{(m - 2) \log (m - 1)} > 0 \quad \forall m \geq 3.
\]  

(14)

Step 1 of the Table 1 may be regarded as an illustration of the above property (see, also, Figure 2(g)). At the same time we can make a conclusion that \(B(m)\) and \(\tilde{B}(m - 1)\) are asymptotically equivalent (see Figure 2(h)).

Considering graph of the target function (see Figure 1(d)) we note interesting feature which relates to the interval \(10 \leq m \leq 30\): odd steps will not make significant changes in difference to the following even steps. This fact has simple explanation. After second step the vlink matrix \(\tilde{S}\) will have the only one significant (diagonal) component in the first row/column. Therefore, all elements of the first row will be connected to any other \(SC\) under condition (11) (see additional term \(H_m\) in the following below (15)). As a result the value of the target function will be increased sharply. Also, we note that graphs Figures 1(h) and 3(d) are much more smoothed compared with Figure 1(d). This property may be caused by the uniformly distributed random noise.

Let us consider the target function after second step

\[
D(\tilde{S}, \alpha, \beta) = \frac{(m - 1)(L_m + 2H_m) + \beta [0.5m(m - 1)(m - 2) - 3(m - 1)]}{m(m - 1) \log m} = \frac{L_m + 2H_m + \beta [0.5m(m - 2) - 3]}{m \log m}
\]  

(15)

where

\[
L_m = -2 \tilde{Z}_m \log \tilde{Z}_m, \quad H_m = -\log \tilde{Z}_m - \frac{1}{m - 1} \tilde{Z}_m \log (m - 1), \quad \tilde{Z}_m = Z_{m+2} = \frac{a}{2a + (m - 1)b}.
\]
The SC for the bound (15) is

\[ \hat{B}(m) = -2 \hat{Z}_m \log \hat{Z}_m - \log \frac{\hat{Z}_m}{m-1} + \hat{Z}_m \log (m-1) \frac{m \log m}{m} . \]

Figure 2. (a-b) bounds \( B_1(m) \) (solid black), \( B_2(m+1) \) (blue dashed) and \( B_4(m+2) \) (red dot-dashed); (c-d) \( B_1(m) - B_2(m+1) \) (solid black), \( B_2(m+1) - B_4(m+2) \) (blue dashed); (e) \( \Lambda(m) = (m+3)(m+2) \log (m+3) [B_2(m+1) - B_4(m+3)] / 6; \) (f) \( \log \Lambda(m) \); (g) \( \frac{1}{\log m} \frac{m-3}{(m-2) \log (m-1)}; \) (h) \( \frac{m-3}{(m-2) \log (m-1)} , m \geq 3, \) see (14).

It may be demonstrated easily that \( B(m) < \hat{B}(m-2) \) \forall m \geq 4. 

EXAMPLE 3.1. The exact values of the target function \( D(S, \alpha, \beta) \) which are presented in the Table 1 (4 values: initial and for the steps NN1-2, 20 where the last one corresponds to the peak of the graph Figure 1(d)) may be computed using formulas (10), (13), (15) and (12) with the following parameters \( v = 3, m = 30, a = 5000, b = 1, \) \( \alpha = 0.005, \beta = 0.00001. \)

Initial value (before first step) of the target function:

\[ D(S_0, \alpha, \beta) = \frac{1}{\log 30} \left[ -2(Z \log Z + 0.5\beta) + \frac{28\beta}{2} \right] \approx 0.007076, \quad Z = \frac{5000}{10027} . \]
Figure 3. (a) vlink matrix $S_{mw}$ 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): $m_0 = 285, \gamma = 0.2, \tau = 3, \alpha = 0.005, \beta = 0.00001, \varphi = 0.0001$.

Formula for the peak is a particular simple: $8\log 9 - \log 10 \approx 0.848217$.

**DEFINITION 3.3.** We call squared matrix $G$ as $(v_1, v_2 \parallel a, b)$-diagonal if

$$g_{ij} = \begin{cases} a & \text{if } i = v_1, j = v_2 \text{ or } i = v_2, j = v_1 \text{ or } i = j; \\ b & \text{otherwise.} \end{cases}$$

Similarly, we can define $(v_i, i = 1..u \parallel a, b)$-diagonal matrix, $u \leq m$.

The case of $(v_i, i = 1..u \parallel a, b)$-diagonal matrix was mentioned in the Sect. 3.1, and corresponds to the case “one step before the peak”.

**PROPOSITION 3.3.** Suppose that $S$ is $(v_1, v_2 \parallel a, b)$-diagonal matrix where $a > b, \beta = 0$, and

$$\frac{b}{a + (m - 2)b} < \alpha \leq \frac{1}{m - 1}, \quad m \geq 2. \quad (16)$$

Then

$$D(S, \alpha, \beta) = B_2(m) + \frac{2(m - 2) \left[ -\frac{1}{m-1} \log Z_m + Z_m \log (m-1) \right]}{m(m-1) \log m}. \quad (17)$$
Table 3. List of the most frequent web-areas; $I_O$ - original index; $I_S$ - secondary index (23); column $m$ indicate number of clusters when corresponding $vroot$ appeared in the last time. For example, $vroot$ “Products” was the last in the merging process.

<table>
<thead>
<tr>
<th>Number of repeats</th>
<th>$I_O$</th>
<th>$I_S$</th>
<th>Name of $vroot$</th>
<th>$m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>10837</td>
<td>1008</td>
<td>9</td>
<td>Free Downloads</td>
<td>3</td>
</tr>
<tr>
<td>9383</td>
<td>1034</td>
<td>35</td>
<td>Internet Explorer</td>
<td>4</td>
</tr>
<tr>
<td>8463</td>
<td>1004</td>
<td>5</td>
<td>Microsoft.com Search</td>
<td>5</td>
</tr>
<tr>
<td>5330</td>
<td>1018</td>
<td>19</td>
<td>isapi</td>
<td>7</td>
</tr>
<tr>
<td>5108</td>
<td>1017</td>
<td>18</td>
<td>Products</td>
<td>1</td>
</tr>
<tr>
<td>4628</td>
<td>1009</td>
<td>10</td>
<td>Windows Family of Oss</td>
<td>8</td>
</tr>
<tr>
<td>4451</td>
<td>1001</td>
<td>2</td>
<td>Support Desktop</td>
<td>6</td>
</tr>
<tr>
<td>3220</td>
<td>1026</td>
<td>27</td>
<td>Internet Site Construction for Developers</td>
<td>2</td>
</tr>
<tr>
<td>2968</td>
<td>1003</td>
<td>4</td>
<td>Knowledge Base</td>
<td>9</td>
</tr>
<tr>
<td>2123</td>
<td>1025</td>
<td>26</td>
<td>Web Site Builder’s Gallery</td>
<td>12</td>
</tr>
<tr>
<td>1791</td>
<td>1035</td>
<td>36</td>
<td>Windows95 Support</td>
<td>17</td>
</tr>
<tr>
<td>1506</td>
<td>1040</td>
<td>41</td>
<td>MS Office Info</td>
<td>16</td>
</tr>
<tr>
<td>1500</td>
<td>1041</td>
<td>42</td>
<td>Developer Workshop</td>
<td>11</td>
</tr>
<tr>
<td>1446</td>
<td>1032</td>
<td>33</td>
<td>Games</td>
<td>14</td>
</tr>
<tr>
<td>1160</td>
<td>1037</td>
<td>38</td>
<td>Windows 95</td>
<td>21</td>
</tr>
<tr>
<td>1115</td>
<td>1030</td>
<td>31</td>
<td>Windows NT Server</td>
<td>22</td>
</tr>
<tr>
<td>1110</td>
<td>1038</td>
<td>39</td>
<td>SiteBuilder Network Membership</td>
<td>24</td>
</tr>
<tr>
<td>1087</td>
<td>1020</td>
<td>21</td>
<td>Developer Network</td>
<td>10</td>
</tr>
<tr>
<td>912</td>
<td>1000</td>
<td>1</td>
<td>regwiz</td>
<td>25</td>
</tr>
<tr>
<td>865</td>
<td>1007</td>
<td>8</td>
<td>International IE content</td>
<td>13</td>
</tr>
<tr>
<td>842</td>
<td>1052</td>
<td>52</td>
<td>MS Word News</td>
<td>18</td>
</tr>
<tr>
<td>759</td>
<td>1036</td>
<td>37</td>
<td>Corporate Desktop Evaluation</td>
<td>28</td>
</tr>
<tr>
<td>749</td>
<td>1002</td>
<td>3</td>
<td>End User Produced View</td>
<td>19</td>
</tr>
<tr>
<td>728</td>
<td>1014</td>
<td>15</td>
<td>Office Free Stuff</td>
<td>15</td>
</tr>
<tr>
<td>716</td>
<td>1295</td>
<td>285</td>
<td>Training</td>
<td>20</td>
</tr>
</tbody>
</table>

where

$$B_2(m) = \frac{(m - 2)^2(m - 3) \log (m - 1)}{m(m - 1)^2 \log m}, \quad Z_m = \frac{a}{a + (m - 2)b}. \quad (18)$$

Proof. Without loss of generality we will assume that $v_1 = 1$ and $v_2 = 2$. Then, first and second rows in the $vlink$ matrix $S$ will not have any mutual contribution to $SCs$. We have $\alpha \leq (m - 1)^{-1} \leq Z_m$ (the last inequality will be strong if $m \geq 3$). In total, first and second rows will contribute $2(m - 2)$ terms to the value of $D$. Each term has value

$$- \frac{1}{m - 1} \log Z_m + Z_m \log (m - 1).$$

The contribution of other rows in (17) may be found using method of the Proposition 3.2 with $m - 2$ (without first 2 lines). Denominator represents a standard norm coefficient (6). \qed

Remark 3.1. The following property take place $B_2(m + 1) \leq B_1(m) \forall m \geq 2$. Note that above inequality will be strong $\forall m \geq 3$.

Proposition 3.4. Suppose that $S$ is $(v_1, v_2, v_3||a, b)$-diagonal matrix, $a > b$, $\beta = 0$, and

$$\frac{b}{2a + (m - 3)b} < \alpha \leq \frac{1}{m - 1}, m \geq 3. \quad (19)$$

Then

$$D(S, \alpha, \beta) = B_3(Z_m, m) + \frac{6(m - 3)\left[\frac{1}{m - 1} \log Z_m + Z_m \log (m - 1)\right]}{m(m - 1) \log m} \quad (20)$$

Proc. of SPIE Vol. 6241  62410J-9
Firstly, we simulated \( n = 5000 \) web-traffic records. Then, we computed \( S^\star \) according to (1).

\[
B_3(Z_m, m) = B_4(m) + \frac{-6Z_m \cdot \log Z_m}{m(m-1)\log m};
\]

\[
B_4(m) = \frac{(m-2)(m-3)(m-4)\log(m-1)}{m(m-1)^2\log m}; \quad Z_m = \frac{a}{2a + (m-3)b}.
\]

**Proof.** We have \( \alpha \leq (m-1)^{-1} \leq Z_m \) (the last inequality will be strong if \( m \geq 4 \)). The required result may be derived using methods of the Proposition 3.1 and 3.3.

**Remark 3.2.** Considering asymptotical relations \( (m \to \infty) \) we can easily find structures of the bounds (17) and (20):

\[
O(1) + O(m^{-2}) \quad \text{and} \quad O(1) + O(m^{-3}) + O(m^{-2})
\]

where terms \( B_2(m) \) and \( B_4(m) \) correspond to \( O(1) \).

Note that \( (m+3)(m+2)\log(m+3) [B_2(m+2) - B_4(m+3)] / 6 \geq 0.4487 > \exp(-1.0) \forall m \geq 3 \) (see Figure 2(e-f)). Therefore, \( B_3(Z, m+1) < B_2(m) \) for all \( 0 \leq Z \leq 1 \) and \( m \geq 5 \).

**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) or \( t = m \), 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.

\( Vlink \) matrix \( S \) (see Figure 1(e)) for the second experiment was produced using Algorithm 2 with \( T = S_0 \) and \( E = 500 \). Firstly, we simulated \( n = 5000 \) web-traffic records. Then, we computed \( S^\star \) according to (1).
4. EXPERIMENTS ON THE MSWEB DATASET

Msweb dataset includes 32711 records and 294 vroots. Table 3 represents the most frequent vroots.

The following procedure was used in order to produce secondary indexes \( I_S \) out of original indexes \( I_O \):

\[
I_S = \begin{cases} 
I_O - 999 & \text{if } 1000 \leq I_O \leq 1046; \\
I_O - 1000 & \text{if } 1048 \leq I_O \leq 1284; \\
I_O - 1002 & \text{if } 1287 \leq I_O \leq 1295; \\
I_O - 1003 & \text{if } I_O = 1297.
\end{cases}
\] (23)

We reduced 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 deviation 2.5 and maximum 35. Figure 3(a) illustrates vlink matrix \( S_{msweb} \), which was computed according to msweb data.

<table>
<thead>
<tr>
<th>Table 4. Example with initial vlink matrix ( S_{msweb} ) which was computed according to the msweb dataset.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step</td>
</tr>
<tr>
<td>246</td>
</tr>
<tr>
<td>247</td>
</tr>
<tr>
<td>248</td>
</tr>
<tr>
<td>249</td>
</tr>
<tr>
<td>250</td>
</tr>
<tr>
<td>251</td>
</tr>
<tr>
<td>252</td>
</tr>
<tr>
<td>253</td>
</tr>
<tr>
<td>254</td>
</tr>
<tr>
<td>255</td>
</tr>
<tr>
<td>256</td>
</tr>
<tr>
<td>257</td>
</tr>
<tr>
<td>258</td>
</tr>
<tr>
<td>259</td>
</tr>
<tr>
<td>260</td>
</tr>
<tr>
<td>261</td>
</tr>
<tr>
<td>262</td>
</tr>
<tr>
<td>263</td>
</tr>
<tr>
<td>264</td>
</tr>
<tr>
<td>265</td>
</tr>
</tbody>
</table>

As it was mentioned in introduction, given vroots may be grouped using different methods. For example, Ref.3 identified eight groups: “Products”, “Catalogue”, “Internet”, “Entertainment”, “Office”, “Development”, “Windows”, “Initials”. Then, Ref.3 developed three the most likely graphical models, which represent relations between above eight groups. Note, also, correspondence between Figure 3(a) and Figure 13 “Frequency distribution of web-pages”.

REMARK 4.1. The structure of the graph Figure 3(d) is remarkably similar comparing with graphs Figure 1(d) and (h). Although, image Figure 3(c) is much more “smoothed” comparing with Figure 1(c), see, also the corresponding 3D surface, Figure 4(g).

Note that all vroots after step 264, Table 4, may be found in the Table 3 of the most popular web-areas.

A Pentium 4, 2.8GHz, 512MB RAM, computer was used for the computations which were conducted according to the special program written in C. The overall complexity of the algorithm is \( O(n^3) \). The total computation time was less than 1 sec. for the synthetic example with \( m = 30 \), and about 32 sec. for the msweb dataset with \( m = 285 \).
5. CONCLUDING REMARKS

The proposed method was tested successfully against ideal synthetic vlink matrix with known solution. As a next step we considered more complex and realistic case: we generated synthetic web-traffic data and computed corresponding vlink matrix. Again, the automatical system produced correct answer considering 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 (an averaged log-likelihood divergence) grows initially to the point \( m \approx 25 \), then, it declines to zero. In line with main computations the system produced transformation (merging) function. Using this function we can compress 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\(^{9}\) 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 in order to detect number of significant clusters, and analyze stability of the clustering configuration in a large datasets of internet users.

ACKNOWLEDGMENTS

This work was supported by a grant of the Australian Research Council.

REFERENCES