# Visualizing Clickstream Data with Multidimensional Scaling

Jaron Azaria, Wim Pijls, and Michiel van Wezel

Econometric Institute, Faculty of Economical Sciences, Erasmus University, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.

Econometric Institute Report EI 2005-51

**Abstract.** We visualize a a web server log by means of multidimensional scaling. To that end, a so-called dissimilarity metric is introduced in the sets of sessions and pages respectively. We interpret the resulting visualizations and find some interesting patterns.

## 1 Introduction.

This paper describes the investigation of the data of a Web server log, as part of the ECML/PKDD 2005 Discovery Challenge [1]. The Web server data comprise a listing of page requests. In each request a session-id and a page-id can be identified. A set of requests corresponding to the same session-id defines a session. We made two analyses, based on sessions and pages respectively.
The technique used in our analyses is MDS, short for multidimensional scaling. MDS is a technique suitable for visualizing objects that are not given by numerical coordinates. When only a distance measure between objects is defined, MDS is able to embed the objects in an $n$-dimensional Euclidean space. If this embedding is conducted in a 2-dimensional space, the resulting picture might reveal some interesting patterns to a human observer. The patterns can be investigated further, possibly with different tools. In this way MDS is applied as exploratory data analysis.
In the current situation, we focus on the interdependencies between sessions and pages. To visualize the interdependencies using MDS, we introduce distance measures between sessions as well as between pages. Two sessions are close to each other if they have many pages and many mutual transitions in common. Likewise, two pages are similar if they have may sessions in common.

This paper is structured as follows. Section 2 describes the preprocessing of the data. Here we show, how the sets of objects which are subject to MDS, is obtained. In Section 3 we introduce the distance measures between those objects. In MDS a distance is also called a dissimilarity metric. Section 4 gives an introduction to multidimensional scaling. The results of our analysis are presented in Section 5. Finally, we give some conclusions in Section 6.

## 2 Preprocessing the Data.

As is the case in nearly any data mining project, we need to preprocess the data set. The data set has been provided as a large collection of web-server log files of some web-shop. Each log file comprises the clicks of one hour on the site of that shop. The name of the log-files conforms to the following template _YYYY_MM_DD_HH_stream.log. Each line in the log-files has the following format.

[shopNo];[timestamp];[IP-address];[sessionId];[page];[referrer]

We call such a line a *visit*. To illustrate this format, we randomly selected a line, including the following values:

- shopNo=12, the number of the department or subshop, the web-shop is divided into multiple subshops;
- timestamp=1074592807, the time stamp according to the UNIX standard;
- IP-address=80.188.69.175, the IP-address of the visitor;
- sessionID=1b835c20458a82ac6e26ac8015f6eddc;
- page=/dt/?c=11642, the tail of the URL of the requested page; "dt" is a page reference and "$c = 11642$" is an optional parameter value representing the id of a product or product type;
- referrer=/ls/?id=3, the previous page that requested the current page.

We have used two types of objects in our analysis, viz. sessions and pages. The precise description of them is discussed below.

**Sessions.** First of all, the set of visits is partitioned into subsets. The session-id is the criterion for this partitioning, i.e., the visits with identical session-id values make up a subset. A session is defined as a set of visits with the same session-id. As mentioned before, a page name may contain a parameter $c$ indicating a product id. Next to the log files a list of products with corresponding parameter values was provided in [2]. Since the range of products is very large, we have defined our own set of product groups. Five groups are distinguished:

1 Computers,
2 Audio,
3 Video,
4 White goods,
5 Mobile.

The second preprocessing step is renaming the pages by replacing the product id with the group name the product belongs to. In some cases the product id of a page could not be identified. Those pages are not considered.
In order to take the transitions between pages into account, the next step is transforming sessions into graphs. A graph $G$ is associated with each session $S$. Each page name in $S$ is a vertex in $G$. The edges of $G$ are defined as follows. We order the visits in session $S$ by time stamp. An edge from a vertex or page $p_1$ to

a vertex $p_2$ is included in $G$ if and only of $p_1$ and $p_2$ are pages in two consecutive visits in the ordering by time stamp of $S$.
The collection of graphs obtained in this way is subject to MDS, to be described in Section 4. We found out that the total number of sessions in the data set is approximately $500,000$. This number is far too large for MDS, since, when applying MDS to a set of $n$ objects, an $n \times n$ matrix must be processed. Therefore, we randomly selected one sessions from each log file, resulting in only 576 objects, a feasible number.
Note that the *referrer* attribute was neglected, since the information included was too poor.

**Pages.** We have considered only the pages including descriptions of products or product categories. These pages are identified by the denotation "dt" or "ct" in their URL, indicating that the page concerns a product or product category respectively. The dt-pages far dominate the ct-pages. We merged all 576 files into one database. We defined two sets of pages, viz., the set of *ct*-pages with a frequency of $\geq 100$ in the database and the set of *dt*-pages with a frequency of $\geq 250$. The first set contains 63 unique pages, the second set 940.

## 3 Dissimilarities

As mentioned before, a dissimilarity metric between objects is needed in order to apply MDS to a set of objects. Therefore, we introduce dissimilarity metrics for the set of graphs obtained by the preprocessing as well as for the set of pages. First of all, we present some definitions. The first one is the common definition of a directed graph.

**Definition** A directed graph $G$ is a tuple $< V, E >$ where

- $V$ is the finite set of vertices
- $E \subseteq V * V$ is the finite set of directed edges

Since the vertices in our graphs represent web pages, we may say that the vertices have unique labels. In case of unique labels the definition of a common subgraph reduces to the following simple one.

**Definition** Given two graphs $(V_1, E_1)$ and $(V_2, E_2)$, a *common subgraph* $G_0$ of $G_1$ and $G_2$, is a graph $G_0 = (V_0, E_0)$ such that $V_0 \subseteq V_1 \cap V_2$ and $E_0 \subseteq E_1 \cap E_2$.

The size of a graph $G$, denoted by $|G|$, is defined as $|G| = |V| + |E|$. A maximal common subgraph of $G_1$ and $G_2$, denoted by $\text{MCS}(G_1, G_2)$, is defined as a common subgraph of $G_1$ and $G_2$ with maximal size. For finding a distance or a metric between graphs, we rely on [4] and [5]. A slight modification of their definitions is utilized. Given two non-empty graphs $G_1$ and $G_2$, the distance between $G_1$ and $G_2$ is defined as:

$$d(G_1, G_2) = 1 - \frac{|\text{MCS}(G_1, G_2)|}{\text{MAX}(|G_1|, |G_2|)}. \quad (1)$$

The following properties hold for this distance measure $d$:

a) $0 \leq d(G_1, G_2) \leq 1$
b) $d(G_1, G_2) = 0 \Leftrightarrow G_1$ and $G_2$ are identical
c) $d(G_1, G_2) = d(G_2, G_1)$

In [4] it is proved that this distance measure satisfies the triangle inequality.

Analogously, we define the dissimilarity between two pages. A page $P$ is identified with the set of sessions that include page $P$. Hence, by $|P|$ we mean the number of sessions including $P$. Accordingly $P_1 \cap P_2$ is defined as the set of sessions including both $P_1$ and $P_2$.

$$d(P_1, P_2) = 1 - \frac{|P_1 \cap P_2|}{\text{MAX}(|P_1|, |P_2|)}. \tag{2}$$

Again the above three properties hold for this metric.

## 4 Multidimensional Scaling

MDS [3] is a collection of techniques for embedding objects in a space with a chosen dimensionality based on dissimilarities between these objects. MDS provides insight in the underlying structure of relationship between the objects that are embedded. The MDS techniques can be subdivided in two groups, metric MDS and non-metric MDS. Metric MDS assumes the dissimilarities between objects to be proportional to Euclidean distances, non-metric MDS only assumes the dissimilarities to be related to Euclidean distances by an unknown monotone transformation.

| | A | C | D | H | L | M | NY | SF | S | W | A | C | D | H | LA | M | NY | SF | S | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Atlanta | 0 | | | | | | | | | | 0 | | | | | | | | | |
| Chicago | 58 | 0 | | | | | | | | | 4 | 0 | | | | | | | | |
| Denver | 121 | 92 | 0 | | | | | | | | 22 | 13 | 0 | | | | | | | |
| Houston | 70 | 94 | 87 | 0 | | | | | | | 8 | 15 | 12 | 0 | | | | | | |
| Los Angeles | 193 | 174 | 83 | 137 | 0 | | | | | | 34 | 31 | 11 | 24 | 0 | | | | | |
| Miami | 60 | 118 | 172 | 96 | 233 | 0 | | | | | 6 | 21 | 29 | 18 | 39 | 0 | | | | |
| New York | 74 | 71 | 163 | 142 | 245 | 109 | 0 | | | | 10 | 9 | 27 | 25 | 42 | 20 | 0 | | | |
| San Francisco | 213 | 185 | 94 | 164 | 34 | 259 | 257 | 0 | | | 35 | 32 | 16 | 28 | 2 | 44 | 43 | 0 | | |
| Seattle | 218 | 173 | 102 | 189 | 95 | 273 | 240 | 67 | 0 | | 36 | 30 | 19 | 33 | 17 | 45 | 40 | 7 | 0 | |
| Washington D.C. | 54 | 59 | 149 | 122 | 230 | 92 | 20 | 244 | 232 | 0 | 3 | 5 | 26 | 23 | 37 | 14 | 1 | 41 | 38 | 0 |

**Fig. 1.** Dissimilarity matrices containing distances (l) and ranks (r).

Figures 1 and 2 show an example application of MDS. The left matrix in Figure 1 is a distance matrix (measured in units of 100 miles) of inter-city distances

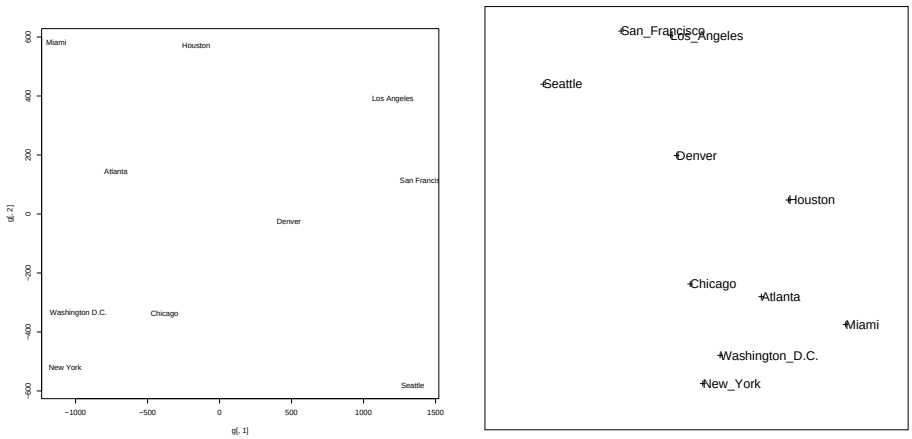


**Fig. 2.** Embedding for matrices in Figure 1. Left: metric MDS (left matrix embedded). Right: non-metric MDS (right matrix embedded).

in the USA. The left plot in Figure 2 shows the metric MDS embedding for this matrix. This embedding closely resembles the map of the USA. In the right matrix, the distances are replaced by their rank order. (This is a monotone transformation.) The plot on the right shows the embedding obtained with non-metric MDS, which is a rotated version of the USA-map. Surprisingly, non-metric MDS is able to recover the relative positions of the objects using these distorted distances.

A nice feature of MDS, that makes it especially useful for clickstream analysis, is that it does not require vectors of numerical values in order to create an embedding. It merely requires dissimilarity values, and these are often more easily obtained than numerical vectors.

The quality of an MDS embedding is usually measured using a criterion called stress. This expresses the discrepancy between the dissimilarities $\delta$ between the objects in the dissimilarity matrix and the distances $d$ between the objects in the embedding:

$$\text{stress} = \sum_{ij, i \neq j} (\delta_{ij} - d_{ij})^2,$$

where $i, j$ sum over the objects. In the case of non-metric MDS the dissimilarities are replaced by transformed dissimilarities, often called disparities. See [3] for details. Below, we report *relative* stress values w.r.t. the embedding in 0-dimensional space, where all $d$'s are 0.

## 5 Experiment & Results

As stated before, we ran two sets of experiments: Embedding http-sessions and embedding http page-requests. We briefly describe both experiments below.

### 5.1 Embedding the Sessions

We constructed two sets of 576 random sessions in which one session was selected randomly from each log file. We refer to these session sets as `subset1` and `subset2` For each subset we computed a dissimilarity matrix using the graph distance metric described in Section 3.
These dissimilarity matrices were subsequently embedded in two-dimensional space with MDS. The MDS analysis was performed using the public domain software package `R` [6]. Library `MASS` contains the routine `isoMDS`, which is an implementation of non-metric MDS.

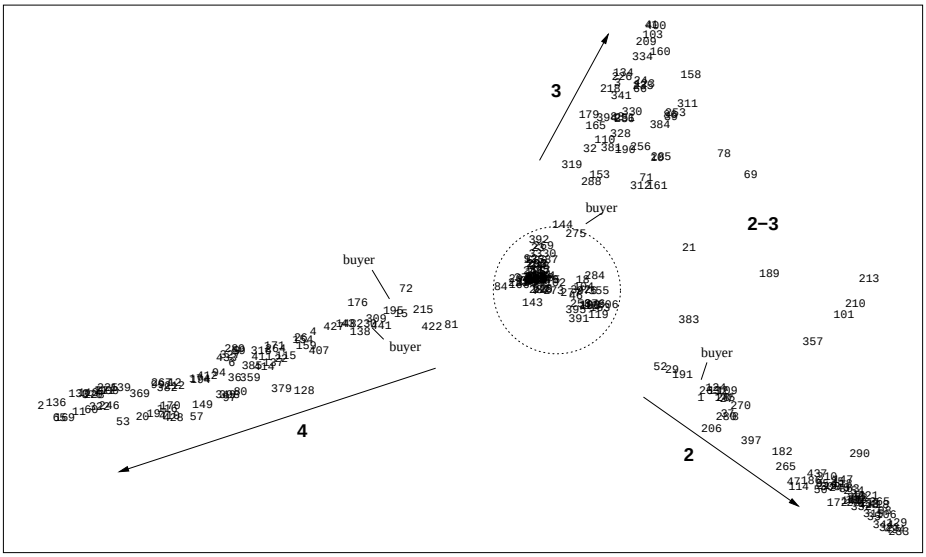


**Fig. 3.** Embedding obtained for `subset1`. 1= PC, 2= Video, 3= Audio, 4= White goods, 5= Mobile. The stress of this 2-dimensional solution is approximately 47%.

Figure 3 shows the resulting embedding for *subset1*. The numbers in the plot refer to the number of the sessions in the subset. This helps to interpret clusters and other patterns in the embedding, since we are able to identify a session and see what page requests it contains.
The first thing we notice when looking at Figure 3 is the tripod-like structure of the plot. It looks asif we can identify 4 different clusters: the center (where the sessions are plotted very close to each other), and the clusters with the 4-arrow, 2-arrow and 3-arrow. The cluster for group 4 (White goods) seems to be the biggest, followed by group 2 (Video) and group 3 (Audio). Next, we notice that groups 1 and 5, respectively PC and Mobile, are not represented by a distinct

cluster. In fact, it turns out that sessions with page request to these groups are very few in number.

**Observation** *The most predominant product categories on this web site are 2 (Video), 3 (Audio) and 4 (White goods).*

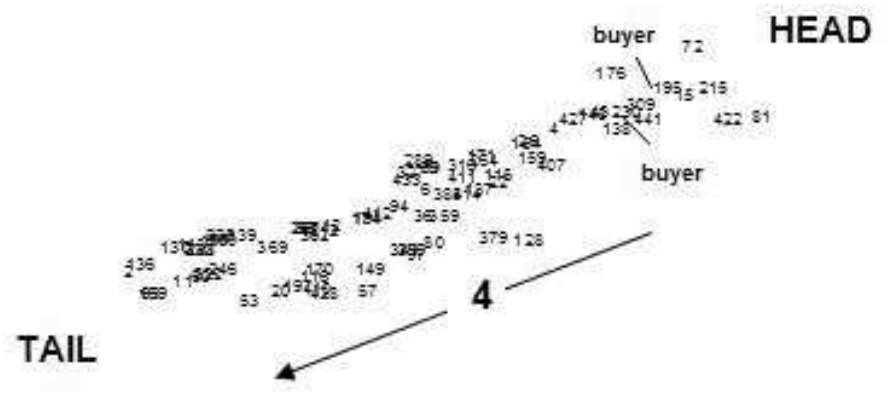


**Fig. 4.** Cluster group 4, example of head and tail.

If we take a closer look at a cluster, for example cluster group 4, we can identify a head and a tail, as shown in Figure 4. When analyzing the sessions in the group 4 cluster we notice that the more one follows the arrow and analyzes the tail, the more sessions only exist out of page request of products from group 4 (White goods). But when we analyze sessions located towards the head of the cluster we notice more sessions with mixed page request. Although page request to products of group 4 are still present, page requests to things like search pages, contact pages, mail pages, advice pages and shopping carts pages are present as well. This is here where we also find the buyers, these are the sessions with page request to `udaje.php` (filling-in details of contract). The same head/tail analogy can be followed if we look at the other two clusters, group 2 and 3 (not the center cluster). Note that also in these clusters the buyers are in the head area. Furthermore, if we follow the arrow to the tail, again we will find that in this area only sessions with page request to their respective product groups exists. Like we saw in group 4 cluster.

**Observation** *A fairly large proportion of site visitors interested in categories 2, 3 and 4 (Video, Audio and White Goods) are only interested in pages directly related to the products in the group, presumably product information. The effect appears to be strongest for Video products (category 2). This may indicate that*

*the site serves as an information source for customers making the actual purchase elsewhere.*

A big difference between clusters for groups 4, 3 and 2, is that groups 2 and 3 seem to have an overlap, which is absent between 4 and 3, or 4 and 2. A possible explanation is that people who are looking for video products will also look for audio products in the same session, and vice versa. It is less likely that people who are looking for white goods (freezers, washing machines etc.) will also look for audio or video in the same session.

**Observation** *It frequently happens that pages from the Video and Audio department are visited in the same session. These are the only two product categories that have an association. This information could, e.g., be used to generate cross category recommendations.*

An interesting question arises when we look at non-buyer sessions very close to buyer sessions. Could these sessions be potential buyer sessions? This is a possibility, but it does not necessarily have to be the case. Sessions in the head area of clusters are closest to the center. The center contains only the sessions that have no page request to any product group. Sessions in the center have page request of mail pages, advice pages, search pages etc. So, sessions in close vicinity of buyer sessions may include a request to one or more of these 'non product pages' as well.

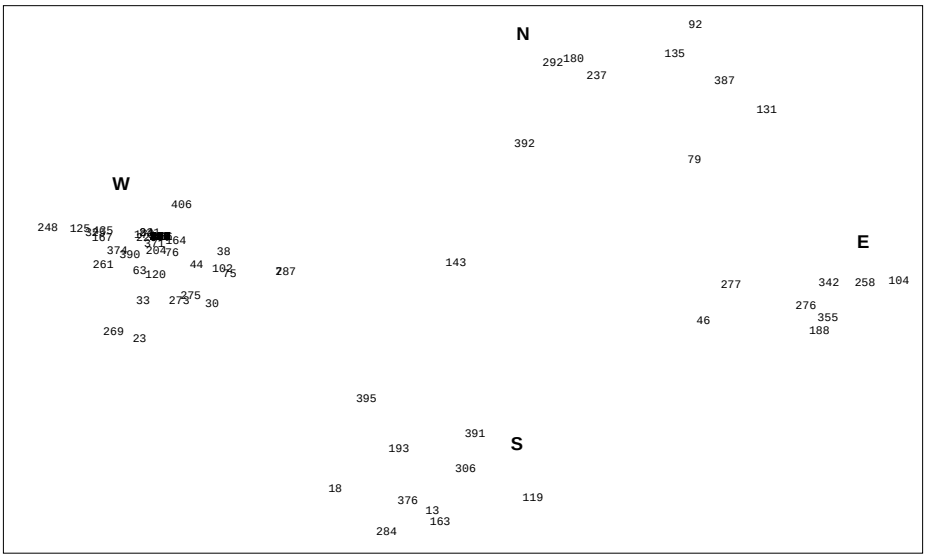


**Fig. 5.** Separate MDS plot for center cluster `subset1`.

As stated earlier the center cluster consists of sessions plotted very close to each other. This makes it difficult to analyze them. Therefore, we created a separate MDS plot for sessions located in the center cluster. The resulting plot is shown in Figure 5. We can clearly identify four (sub)clusters, named North(N), East(E), South(S) and West(W).
When analyzing these clusters, we found that sessions in cluster N and E consist of page request to product group 1. Because of the small amount of sessions involved in this cluster, compared to the other product groups clusters (4,3 and 2), it becomes clear again that product group 1 (PC's) is not popular in the web-shop. Cluster S consists of small sessions only visiting the main page, and sometimes another non-product page.
Cluster W actually has a few little sub-clusters. One sub-cluster consists of sessions with page request to product group 5. Note that this product group is even less popular than group 1.
Another sub-cluster contains sessions with page request to `udaje.php`. This is the page where the actual purchase transaction takes place. The strange thing is that no products where visited in the session. A possible explanation is that the products were already chosen and stored in a cookie on the customers computer.
Finally, there is a small cluster with sessions that consists out of page request to advice pages only.

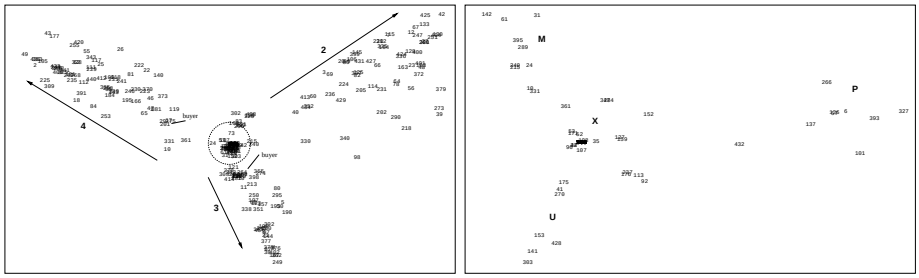


**Fig. 6.** Embedding obtained for `subset2`. Left: 1= pc, 2= video, 3= audio, 4= white goods, 5= mobile. Right: separate embedding for center.

The embeddings found for `subset2` look highly similar to those found earlier, apart from rotation. The embeddings are shown in Figure 6.

### 5.2 Embedding the Pages

As stated before, we embedded both the *ct*-page requests and de *dt*-page requests. Ct refers to product categories, dt to product details.

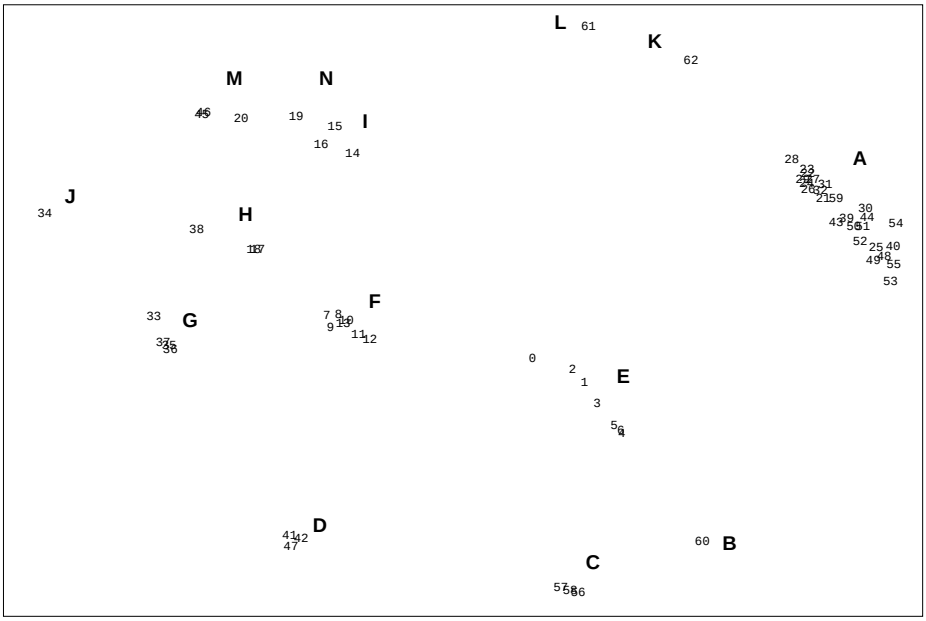


**Fig. 7.** Embedding for ct-pages. Clusters of related product categories are clearly visible. The relative stress-value for this embedding was 21%. See text for more comments.

Figure 7 shows the embedding obtained for the *ct*-page requests. If two *ct* requests appear close together in the embedding it means that the frequently appear together in a session. This embedding clearly displays a number of clusters, each with a clear interpretation: A contains only ct requests related to white goods. K,L concern cookers and dishwashers respectively. B concerns digital cameras. C concerns computers. D concerns PDA's. E concerns film cameras and lenses. F concerns audio. G concerns car audio. H concerns home cinema. I,M,N concern television, plasma & projection TV's and camcorders respectively. J concerns MP3 players.

**Observation** *The MDS embedding displays a natural clusting of product categories based only on co-occuring page requests. Moreover, clusters with 'similar' product categories appear in eachother's vicinity in the embedding. E.g., car audio appears next to audio.*

The embedding for the dt-page requests is shown in Figure 8. This embedding clearly shows some clusters and other structure as well, but unfortunately we

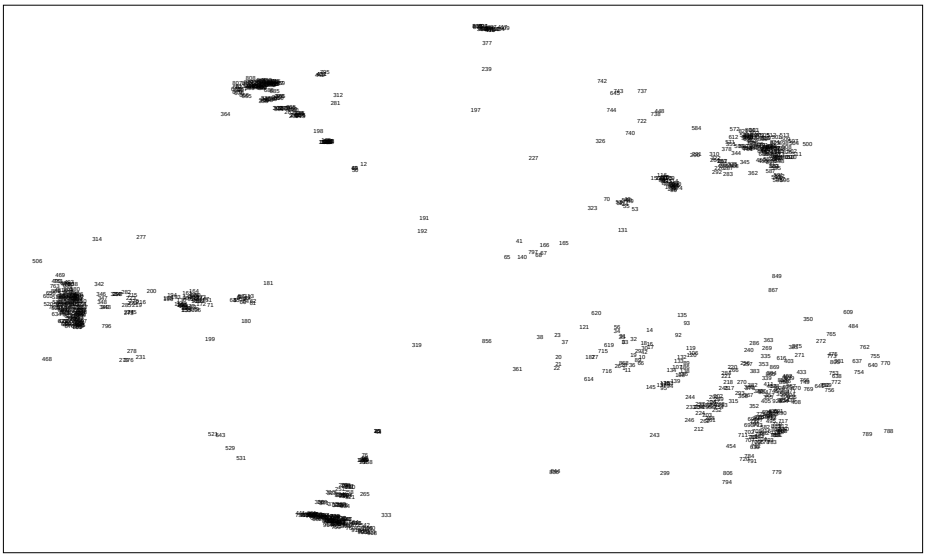

**Fig. 8.** Embedding for dt-pages. The relative stress-value for this embedding was 31%.

were not able to interpret this plot because the meaning of most (app. 95%) of dt-codes was unknown.

## 6 Conclusion and future research

In this paper we performed exploratory data analysis by MDS on clickstream data. We embedded both sessions and page-requests. MDS was used because of the non-numerical nature of the http-requests, for which feature vectors can not be constructed easily, but dissimilarities *can*.
The dissimilarity matrix for the embedding of sessions was constructed by applying a graph distance metric to graphs representing http sessions. For the embedding of pages it was based on counting co-requests within sessions.
The resulting embedding of sessions put us on track of some interesting patterns, e.g., people often visit pages from product categories 2 (video) and 3 (audio) in one session, whereas category 4 (white goods) appears in isolation. In the video product category there are many browsers that do not make any purchases.
Furthermore, we looked at the possibility to identify a group of potential buyers. It could be possible that sessions that not made a contract, but are plotted near the buyer-sessions, are potential buyers. Unfortunately we could not give an definite answer to this question. Future research is needed for this matter.
The embedding of product categories revealed a number of clusters, which all had a clear interpretation. Clusters were positioned in a natural way relative to eachother. The embedding of separate products showed structure as well, but we were unable to interpret it due to missing meta-data.

Although these conclusions might seem trivial, keep in mind that they were obtained using exploratory analysis. In constrast, most learning- and statistical methods either test a specific hypothesis or attempt to find a specific input/output relationship, thus they require a-priori assumptions. Exploratory data analysis does not require the user to make these assumptions — Rather, they help the user in formulating suitable topics for further investigation.
We plan to do a number of additional analyses in the near future. We observed that the resulting projections are highly dependent on dissimilarity measure that is used — we plan to experiment with alternative dissimilarity measures. (In the case of session dissimilarities we experimented with an alternative graph distance without positive results.) Dissimilarity measures that focus on different aspects may lead to different embeddings and consequently new interpretations. Finally, we would like to scale up our analyses to a larger number of sessions.

## References


1. Petr Berka and Bruno Cremilleux. ECML/PKDD 2005 discovery challenge. Web-Site, 2005. `http://lisp.vse.cz/challenge/CURRENT/`.
2. Petr Berka and Bruno Cremilleux. Faq clickstream data. Web-Site, 2005. `http://lisp.vse.cz/challenge/ecmlpkdd2005/CLICK05-FAQ.htm`.
3. I. Borg and P. Groenen. *Modern multidimensional scaling. Theory and applications.* Springer Series in Statistics. Springer-Verlag, New York, 1997.
4. H. Bunke and K. Shearer. A graph distance metric based on the maximal common subgraph. *Pattern Recognition Letters*, 19, Nos 3 - 4:255–259, 1998. http://dx.doi.org/10.1016/S0167-8655(97)00179-7.
5. Mirtha-Lina Fernández and Gabriel Valiente. A graph distance measure combining maximum common subgraph and minimum common supergraph. *Pattern Recognition Letters*, 22(6–7):753–758, 2001.
6. W.N. Venables, D.M. Smith, and the R Development Core Team. *An introduction to R – Notes on R: A programming environment for data analysis and graphics.* The R Foundation for Statistical Computing, 1.6.2 edition, 2003. R is free software, available from `http://www.r-project.org/`.