Analysing Clickstream Data: From Anomaly Detection to Visitor Profiling

Peter I. Hofgesang and Wojtek Kowalczyk

Free University of Amsterdam, Department of Computer Science,
Amsterdam, The Netherlands
{hpi,wojtek}@few.vu.nl

Abstract. The paper presents results of analysis of clickstream data in the context of the ECML/PKDD Challenge. We focused on two aspects: detection of anomalies and profiling visitors of internet shops. Several unusual patterns were discovered with the help of simple tools such as frequency tables, histograms, etcetera. Some insight into click-behaviour of web visitors was obtained with the help of finite mixtures of multinomial distributions and a graphical representation of navigation trees.

1 Introduction

We believe that the main objective of an online shop is to provide useful and relevant product and shopping information in a clear and well-structured way to retain regular customers and to attract new ones, increasing the profit in this way. On the internet each shop is just "one click away". If a user is not satisfied with the service he/she just goes to a next one and will likely never return.
To increase the profitability one must constantly monitor the behaviour of real customers on the website and detect eventual changes in the market, to adapt to the requirements of their customers.

In this paper we focus on anomalies detection and profiling of users. The exploratory data analysis of the Challenge data revealed numerous anomalies that require further investigations. Also, profiling of users of a web site is essential for all online service providers to better understand their (potential) clients' behaviour.

We used standard statistical and data mining techniques for exploratory data analysis and anomalies detection. A model of finite mixtures of multinomial distributions [1] was applied to the session data to extract user profiles for further analysis. Furthermore we used a tree-like visualisation of profile sequences to reveal hidden sequential information about the user sessions.

The rest of this paper is organised as follows. In Section 2 we review the essential results of the exploratory data analysis and describe the anomalies found in the data and the necessary steps performed for the further analysis. In Section 3 we present the process of identification of user profiles and a tree-like visualisation of profile sequences. Finally, Section 4 contains our conclusions.

2 Exploratory Data Analysis

The original data set contains about 3.5 million records that reflect internet traffic registered during a period of 25 days by websites of 7 internet shops. Records can be grouped with help of the provided session identifier field (SID) into 522410 unique sessions that originate from 79526 different IP-addresses. The distribution of sessions and visited pages over seven shops is shown in Table 1:

ShopId	Assortment	#sessions	#pages	#pages/session
10	Cameras	80740	509688	6.3
11	Audio	57086	400045	7.0
12	TV's, Video	78845	645724	8.2
14	Home Appliances	156285	1290870	8.3
15	MP3 players	67987	308367	4.5
16	Mobile Phones, PDA's	53305	298030	5.6
17	Computers, Software	28179	164447	5.8

Table 1. Distribution of sessions and pages per shop.

As we can see the most frequently visited shop was the one with home appliances, perhaps due to the broadest selection of offered product categories, whereas the shop with computers was least popular. It is also interesting to notice that the shop selling only MP3 players (thus a product from a single product category) had a relatively large number of visits.

To get an insight into the geographical location of visitors we translated IP-addresses into country codes and looked at their distribution. As expected, it turned out that most IP-addresses are from the Czech Republic (50238) and neighbouring countries: Slovakia (1991), Poland (492), Germany (343), The Netherlands (240), Italy (231), and Hungary (210). In general, about 10% of the traffic is generated from abroad.

2.1 Anomalies in the data

While analysing some basic properties of sessions (length, duration, internal consistency, etcetera) we encountered a number of problems. Some of these problems were minor and could be easily fixed or ignored, others were so serious that we decided to redefine the concept of a session.

Multiple IP-addresses per session It is logical to expect that every session involves exactly one IP-address ("IP"). However, we spotted 3690 sessions with more than one IP-address. Most of these sessions involve 2, 3 or 4 different IPs (3051, 362, 113, respectively), but there are sessions with more than 20 IPs (for example, the session with `SID = 9f412dc3a7bd7681e29125c66da3318f` refers to 22 different IP-addresses). More interestingly, some sessions use IP-addresses from several countries (e.g. `SID = 4f840dcfbd33f940cee24e2d7d352ec7`).

Multiple shops per session Although most sessions refer to 1 shop only (in 522398 cases) there are sessions that refer to 2, 3, or 5 shops (there are 9, 2 and 1 such sessions, respectively).

Very long sessions In the data there are 476 sessions that lasted more than 24 hours; e.g., the session with `SID = 35c97651004351765628dffe50209a18` lasted more than 18 days! (a closer look at this session revealed that it consists of two "normal" 2-3 minutes long sessions that were separated by a break of 18 days).

Very intensive sessions Some sessions were very intensive: they contain many pages that are visited in a relatively short time. There are 2865 sessions with more than 100 pages, 19 sessions contain more than 1.000 pages, and there are 2 sessions with more than 10.000 visited pages. The longest session (with `SID = 35c97651004351765628dffe50209a18`) contains 15454 pages that were visited in less than 7 hours. Most likely such sessions were generated by robots or spiders.

Frequent IP-addresses with short sessions Some IP-addresses occur in numerous (short) sessions. For example, there are 29320 sessions, all taking place in less than 20 hours, that were originated from a single IP-address: `147.229.205.80`. Moreover, 5 IP-addresses were used in more than 10.000 sessions each, and another 372 IPs occurred in more than 100 sessions. Most probably, all these sessions were not generated by a human, but by a program that was fetching large collections of pages.

Sequences of short sessions Almost 61% of the sessions are of length 1. However, when we order these sessions by the corresponding IP-address and time we will notice that in some cases they form larger sessions (visits from the same IP-address to the same shop, within a reasonably short time interval, for the same type of pages).

Parallel sessions In some cases it seems that several sessions could be combined into a longer one. We have the impression that when a user opens a URL in a new window a new session ID is generated. This might lead to some errors in the modelling of the visitor's behaviour. For example, we noticed that users often select a product they want to buy and then open a new window to complete the transaction. Unfortunately, in that case the transaction is recorded with a new session identifier and the context is lost.

It is likely that some of the above anomalies were caused by errors made during data preparation. Others might signal some security problems. Further investigation of these issues could be quite useful, provided more "background information" were available.

2.2 Data cleaning and pre-processing

In order to avoid problems listed in the previous paragraph we decided to redefine the concept of a session in the following way: a session is a chronologically ordered sequence of visits (records) that originate from the same IP-address, are related to the same shop, and do not contain gaps longer than 30 minutes (see [2]). Under this definition the number of sessions of length 1 was reduced from about 300.000 to 65.000, whereas the frequency of longer sessions remained almost the same (see Table 2).

Length	Count (old)	Count (new)
1	318523	65258
2	24762	31821
3	17353	18828
4	15351	16332
5	15361	15509
6	13455	13448
7	10958	10883
8	9045	9095
9	7939	8070
10	7028	7091
...	...	...

Table 2. Distribution of the length of sessions before (old) and after recoding (new).

For further analysis we decided to delete all sessions of length 1 as they would obfuscate the modelling process. Additionally, we removed about 12.000 sessions that were longer than 50 pages, because such sessions were most likely generated by robots, spiders, or special programs.

For the purpose of modelling sequences of visited pages we mapped the different page categories into integers. Table 3 shows the mapping of the 13 most frequent categories. The category 13 ("the rest") refers to unknown URLs—URLs that were not specified in the description of the challenge.

3 Identification and analysis of user profiles

One of our main goals is to provide an insight into users' behaviour on the websites. Therefore we clustered the users using a finite mixture model proposed by Cadez et al. (2001) [1]. We used the resulting clusters to establish different profiles for users and to label all visits by one of the profiles.

Furthermore we analysed the sequences of profiles for each user over the whole period to get a better insight into how users behave throughout their visits. We present a transition matrix and a tree-like visualisation of the profile sequences. The tree visualisation shows the most typical profile paths over the whole dataset.

pageId	URL	description
1	/	main page
2	/ct/	product category
3	/ls/	product sheet
4	/dt/	detail of product
5	/znacka/	brands
6	/akce/	actual offers
7	/df/	comparision
8	/findf/	fulltext search
9	/findp/	parameters based search
10	/setp/	setting displayed parameters
11	/poradna/	on-line advice
12	/kosik/	shopping cart
13	the rest	the unknown URLs

Table 3. Mapping of content categories

3.1 Profiles of users

In their paper Cadez et al. (2001) [1] proposed a generative mixture model for predicting user profiles and behaviours based on historical transaction data. A mixture model is a way of representing a more complex probability distribution in terms of simpler models. It uses a Bayesian framework for parameter estimation and the mixture model addresses the heterogeneity of page visits.
Cadez et al. (2001) presented both a global and an individual model. In this paper we apply the global mixture model because of the relatively small number of sessions per users and because of the uncertainty of the IP addresses as unique user identifiers.

After the analysis of different setups for modelling we chose the four component mixture model as the most suitable for the given data. Each cluster is characterised by a vector of frequencies with which members of such a cluster visit specific pages. These frequencies can be visualised in a bar chart (see Figure 1)—a kind of a "group profile". Additionally, the parameter alpha can be interpreted as the size of the cluster.

To interpret the clusters it is necessary to look at the page id – URL mapping table (Table 3). In Figure 1 we can identify four different user profiles. Three of these profiles cover some kind of product information acquisition while the fourth one describes a behaviour of (potential) buyers. Profiles, in order of appearance, are:

- Profile 1. General overview of the products. Users within this category are interested in a particular product category but they don't have any specific product or brand preferences or they would like to take an overview of all the products in the category.
- Profile 2. Focused search. These users know exactly which products they are looking for. They search for specific products and brands.

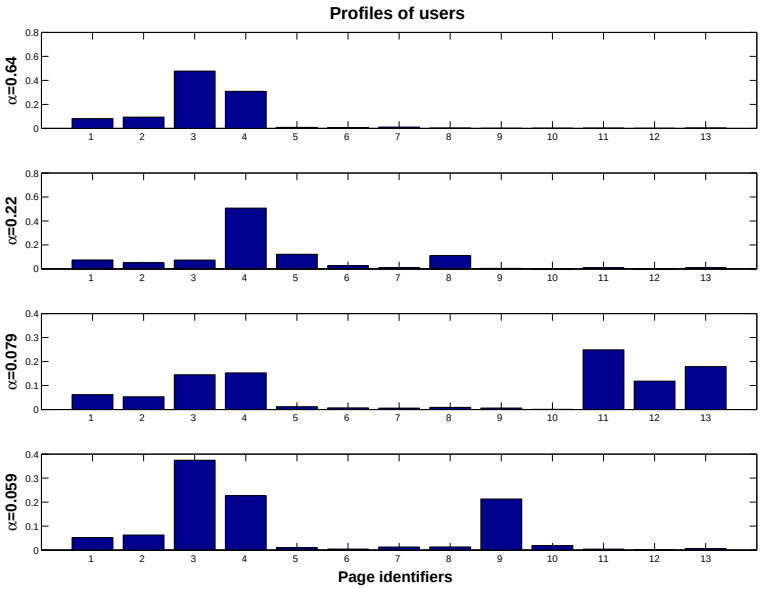

Fig. 1. User profiles of the visitors of the web shop

- Profile 3. Potential buyers. This profile describes potential customers that already placed some products in their shopping cart. The profile also includes customers that actually finalise their transactions, i.e., buy the selected products.
- Profile 4. Parameter based search. Users within this category look for products with specific expectations in mind. They use parameter based search and compare different products.

It is interesting to look at (dynamic) relations between the identified profiles. In particular, we are interested in the following questions:

- How do these profiles relate to each other over sessions from the same user?
- Did buyers buy before already?
- Do buyers come back after a purchase? If so, how do they behave afterwards?

To answer these questions we assigned the appropriate profile identifier to each session of the users and analysed the resulting profile sequences.

3.2 The transitions of profiles

Given the sequences of profiles for each user we can create a transition matrix of the profiles. We count the frequencies of all possible transitions given all the consecutive profile pairs of the sequences. By normalising the resulting frequency

matrix we get the transition probabilities (Table 4). Each x_{ij} element of the table represents the probability of the transition from profile i to profile j ($i, j = 1 \ldots 4$).

	P1	P2	P3	P4
P1	0.7208	0.1592	0.0621	0.0579
P2	0.5908	0.2828	0.0710	0.0553
P3	0.5022	0.1616	0.2873	0.0489
P4	0.6000	0.1702	0.0685	0.1613

Table 4. Transition probabilities of profiles

The most significant transitions are transitions to Profile 1. The high proportions of self-transitions show that many users do not change profile during their visits.

3.3 A tree-like visualisation of profile sequences

To get an insight into how users switch between the identified profiles we replaced their sessions by the corresponding profile types. In other words we created, for every user, a sequence of profile types. Then we analyzed these sequences with help of navigation trees that were introduced in [3].

Figure 2 shows the frequent profile sequences of users. The tree model consists of nodes with their profile labels in specific colours. There is a special virtual node—the root of all sequences. It contains additional information (name of the tree, support rate). Nodes are connected with lines (edges) in different thickness marking the frequencies of the given paths. Edges are labelled with the percentage of (sub)sequences crossing the given node. The absolute number of sequences finished at the given node is given in parentheses. The size of the tree is controlled by the so-called support threshold. The tree contains only (sub)sessions that are more frequent than this given threshold.

Figure 2 shows that most of the users (68.1%) started their visits exhibiting the patterns of Profile 1 (general overview). Furthermore most of these users did not change profile during their visits. The second most relevant profile is Profile 2 (focused search), 20.6% of the users started their visits according to the patterns of this profile.

Figure 3 presents the frequent profile paths of users who had at least one "Potential buyers" profile during their visits. We see again that the general overview and focused search profiles are dominant in these sequences as well. Surprisingly, most of these users (52.9%) started their visits according to the potential buyer profile. This assumes that they bought products (or used the shopping cart) immediately during their first visit. Moreover, most of the users turned into a potential buyer after several visits of overviewing the products (Profile 1).

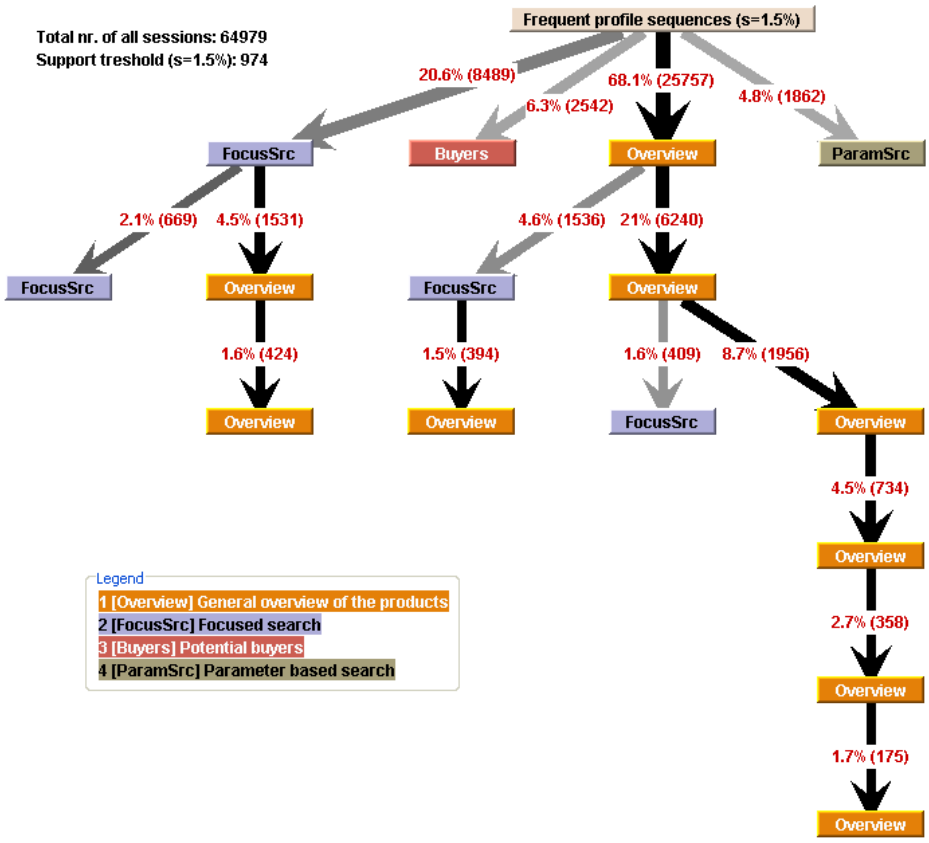

Fig. 2. Tree-like visualisation of frequent profile sequences

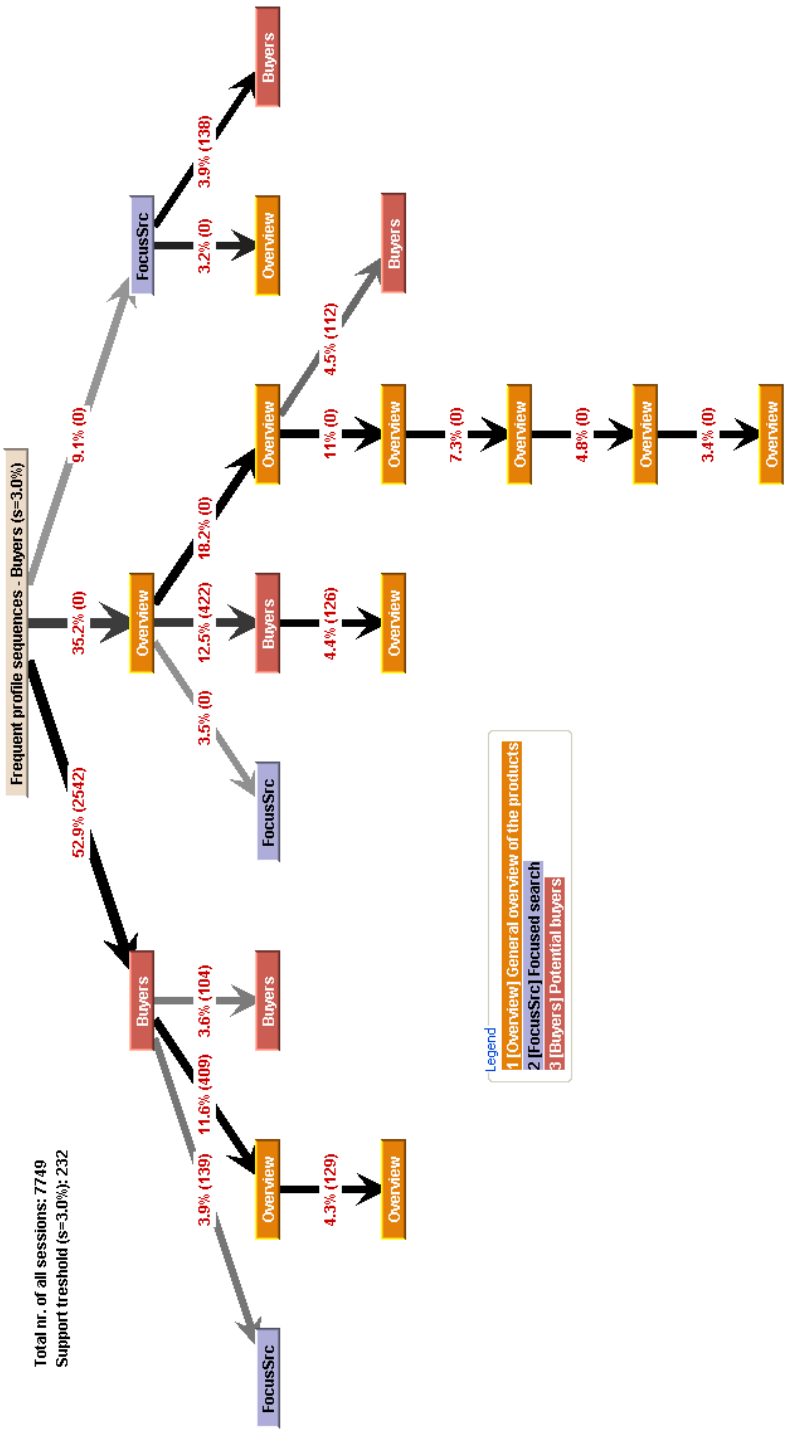

Fig. 3. Tree-like visualisation of frequent profile sequences – Buyers

4 Conclusions

In the paper we applied several data mining techniques to the Challenge data. We used conventional tools (frequency tables, counts, distributions, etcetera) to develop an overall picture of the data and to spot several anomalies. We subsequently used mixture models to group visitors into 4 groups. Finally, we generated navigation trees to visualise typical navigation patterns of clients that approached the most interesting phase of any session: finalisation of the transaction. There are three main conclusions that can be drawn from our results.

First, using a simple mapping of IP-addresses to countries, we discovered that about 10% of the traffic was generated from countries around the Czech Republic. Therefore, it is worthwhile to consider internationalisation of the analysed web shops.

Second, we discovered several anomalies in the data. Some of them could be easily explained (e.g., by the fact that some sessions were generated by robots), others might be symptoms of more serious problems. For example, some anomalies may be caused by errors in scripts that are embedded in pages (a page may for example refer to itself), malicious attacks from outside, fraud attempts etcetera. Perhaps it would be meaningful to run systematic analysis of log files along the lines presented in Section 2.

Finally, we have identified 4 groups of visitors. The most significant group includes users that search for a general information on products from a specific product category. The second group of users is focused on details of concrete products. The third group includes users that use the parameter-based search engine to find products they are looking for. Finally, potential buyers that actually put products into their shopping carts and in some cases purchase them form the fourth group.
Knowledge of these groups may help in redesigning shops' websites or just in understanding the different paths followed by visitors.

References

1. Igor V. Cadez, Padhraic Smyth, and Heikki Mannila. Probabilistic modeling of transaction data with applications to profiling, visualization, and prediction. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 37–46, New York, NY, USA, 2001. ACM Press.
2. Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999.
3. Peter I. Hofgesang. Web usage mining—structuring semantically enriched clickstream data. In *MSc. Thesis, Vrije Universiteit*, Amsterdam, The Netherlands, 2004.