A Time-Aware Exploration of RecSys15 Challenge Dataset

Humberto Jesús Corona Pampín
Zalando Ireland Ltd.
Dublin, Ireland
humberto.corona@zalando.ie

Ana Peleteiro
Zalando Ireland Ltd.
Dublin, Ireland
ana.peleteiro@zalando.ie

ABSTRACT

E-commerce is currently one of the main applications of recommender systems, since it generates vast amounts of data that can be used to make predictions and analyse users's behaviour. In this paper we present an overview of the public dataset used for the *RecSys Challenge 2015*. We describe the basic statistical properties of this dataset and how events (clicks and purchases) are distributed over products (items) and users (sessions). We also present a time-aware analysis of the dataset, with the aim to better understand the change of user behaviour within time cycles, and how it affects the activity in user purchases. We further study the relation between categories, the solely type of metadata present in this completely anonymised dataset. We are interested both in how these categories are distributed and how users and items interact with them. Finally, along the paper we explain the implications that the results obtained from our analysis may have when building models for the challenge.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

Recommender Systems, Dataset exploration, E-commerce

1. INTRODUCTION

E-commerce is one of the main applications of recommender systems [9]. It drives content discovery and helps to increase revenue and diversity of sales in online shopping platforms such as Amazon.com [6] or XBOX Live [4]. However, research on this field has been limited due to the lack of public available datasets. Thus, the release of the *Yoochoose Dataset* for the *RecSys Challenge 2015* [2] presents a major step forward on the research in e-commerce scenarios, being the first dataset of this nature.

The *Yoochoose Dataset* is provided by Yoochoose[1], a company that powers recommendations for online shopping platforms, news and media. The dataset includes 33 million clicks and 1.1 million purchases from 9 million sessions over circa 53 thousand items, collected in a period of six months, from April to September 2014. The goal of the *RecSys Challenge 2015* was to accurately predict how many and which of the items clicked in a session will be purchased.

In this paper we study and provide insights from the *Yoochoose Dataset*. We aim to better understand the user behaviour in e-commerce platforms, and also to better asses what type of features should be included when building recommender systems for them. Moreover, we pay special attention to the temporal characteristics of this dataset.

The remainder of this paper is organised as follows. First, Section 2 presents a general analysis of the dataset. Section 3 describes the dataset from a temporal perspective. Section 4 focuses on the analysis of the categories and the interaction between categories, items and users. Finally, in Section 5, conclusions are presented.

2. A GENERAL ANALYSIS

In this section we present an analysis of the main properties of the *Yoochoose Dataset*. We focus on the distribution of clicks across items and users, as well as the session length and the purchases actions that occur on each session.

Table 1 shows some properties of this dataset, where the clicks and purchases for the same session are aggregated together. We measure the session length in number of events (clicks and purchases) and in time (seconds), as well as the number of purchase events per session (of those sessions that have at least one purchase) and the number of interactions per item, which follows a long-tailed distribution. We see that 90% of the sessions have 7 events or less. Thus, sessions tend to be short and only around 5% of the sessions include a purchase. Moreover, 90% of those sessions (where a purchase is made) include 4 or less purchases.

	Min.	Median	90 %	Max.
Session length (events)	1	2	7	262
Session length (seconds)	0	134	17.9 (mins)	2.8 (days)
Interactions (per item)	1	23	978	162622
Purchases per session	1	2	4	144

Table 1: Summary statistics.

The median session length in time is slightly higher than 2 minutes. However, we also find sessions of up to 2.8 days and 262 events, which highly deviate from the expected user

[1]Yoochoose, http://www.yoochoose.com

CERI'16 June 16–18, 2016, Granada, Spain

DOI:

behaviour. This anomalous behaviour might be related to bots visiting the platform, users who do not close the navigation tabs or internal sessions.

We further investigate the distribution of interactions per item shown in Table 1 by plotting the item popularity distribution in Figure 1, which shows the distribution of clicks across items. The horizontal axis represents each of the items sorted by the number of clicks involving this item, while the vertical axis represents the number of clicks that particular item received. Both axis are in presented in logarithmic scale. Here, we see that this dataset is very skewed towards popular items, as $\sim$ 0.03% of the items receive more than 50,000 clicks, and around 10% of the items (the 5200 most popular) receive circa 90% of the clicks, while the remaining 90% of items only receive 10% of the clicks. This distribution of item popularity is very common in e-commerce scenarios [1].

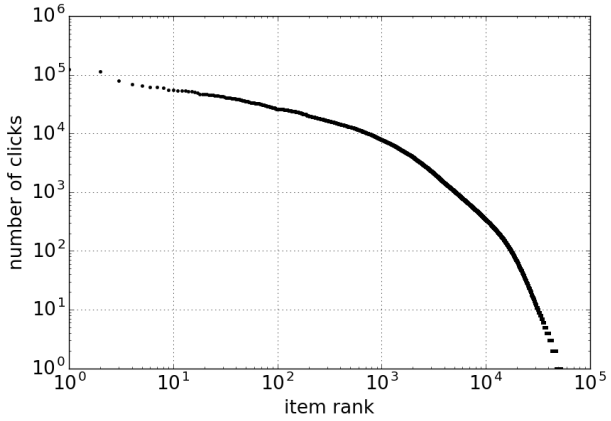

Figure 1: Popularity distribution of items in number of clicks.

Figure 2 shows the distribution of sessions by their length. The horizontal axis represents the length (measured in number of clicks), while the vertical axis represents the number of sessions with that particular length. This figure shows that $\sim$ 51% of the sessions have less than three clicks. Thus, sessions tend to be short and $\sim$ 90% of the sessions have 7 clicks or less. Moreover, there are only a few longer sessions with more than 200 clicks. We also saw similar patterns in the case of session length distribution calculated by time. This is an expected result, as clicks and time must be highly correlated, and very high clicks to time ration could indicate anomalies in the user behaviour as the one previously highlighted.

The distribution of session length in this dataset(measured in events and time) also follows a long-tailed distribution [1]. Thus, a small subset of items - the popular ones - are often clicked and purchased, while most of the items in the catalog are rarely clicked or purchased. The distribution of the data motivates the use of recommender systems in this scenario, as they are known to enhance content discovery, and subsequently purchase, of items in the long tail [3].

Finally, Figure 3 shows the number of clicks per session (horizontal axis), against the ratio of purchases over clicks (vertical axis). As seen on the graph, these two dimensions are highly correlated (Pearson correlation between number

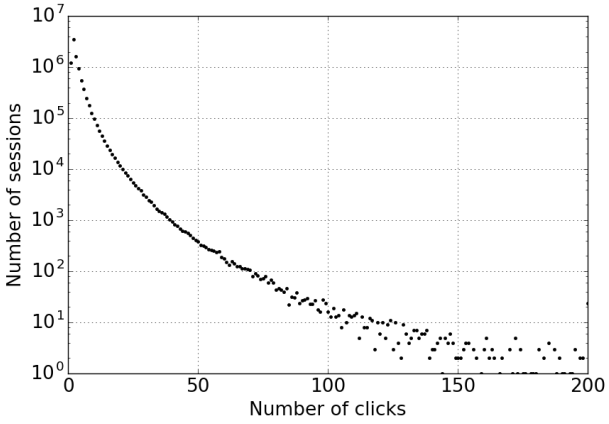

Figure 2: Session length (in number of clicks).

of purchases and number of clicks $\sim$ 0.697) , showing that longer sessions lead to a higher probability of purchasing at least one item. However, a very small number of sessions ($\sim 10,000$) have length $\geq$ 40 clicks, introducing noise over the trend for these large sessions, as the figure shows.

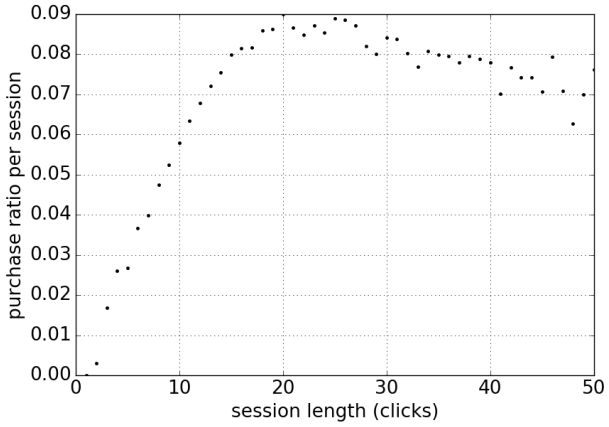

Figure 3: Click distribution sorted by total number of sessions with purchases.

The analysis presented in this section has showed interesting trends that depict user behaviour in online e-commerce platforms. Exploiting this characteristics of user behaviour is key to build recommender systems that can accurately model and predict purchase actions. Moreover, as in many recommender systems scenarios, the dataset is extremely sparse (99.993% of the clicks are missing) and the distribution of clicks over items follows a long tail. Thus, the *popularity bias* in this dataset is very large, and a popularity-based solution[2] can perform relatively well in comparison with the winnig approach proposed by Romov et. al. [8], which scored $\sim$ 63,000.

[2] *How to build a naive (very naive) system*, by Wen Chen http://playwithnlp.blogspot.jp (December 2014), scored over 30,000 in *RecSys Challenge 2015*

3. A TIME-AWARE ANALYSIS

Exploiting both short-term, long-term and cyclic temporal features is known to improve the accuracy of recommendations in different scenarios such as music [7], where user behaviour is highly cyclic, or movies [5], where different concept drifts act at the same time. In this section, we explore the temporal characteristics of the dataset in order to asses the feasibility of using time-aware models to predict shopping behaviour. The ultimate goal is to understand when, why and how people browse and purchase products.

	number of purchases	number of clicks	ratio
April	178719	5973371	0.0299
May	181213	5544429	0.0327
June	171271	5127025	0.0334
July	158516	4434056	0.0350
August	257910	6646541	0.0388
September	203124	5278522	0.0384

Table 2: Monthly clicks, purchases, and purchases per clicks ratio distribution.

We first analyse monthly trends. Table 2 shows the number of events per month over a six months period, from April to September 2014. The table shows the number of purchases and clicks, as well as the ratio of purchases per click. We can highlight three main points. First, the total number of clicks per month is an order of magnitude higher than the number of purchases. This showcases the very small clicks to purchases ratio in e-commerce scenarios. In this table we can also see how the activity on the site was decreasing from April to July, with the browsing activity decreasing at a lower pace than the purchasing activity, and thus, the ratio increases from ~ 0.030 in April to ~ 0.035 in July. Moreover, there is a 50% increase in the site activity in August. This might indicate a major change on the site, or a heavy seasonality in the items sold.

Next, we present a finer granularity temporal analysis, studying the weekly and daily cycles of activity in the site. Figure 4 represents the number of purchases (vertical axis, left) and clicks (vertical axis, right dashed line) on each of the 186 1-hour long time slots which represent a full week starting on Monday. This figure shows an interesting weekly cycle. Sundays and Mondays have the highest activity of all days in the week. This is somehow expected if we assume people do not work on Sundays, and in many countries in Europe shops are closed that day. Moreover, if we take into account the delivery process, most delivery companies do not deliver on weekends, it is expected that orders are not placed Friday or Saturday. The rest of the week shows a more stable daily activity of circa 150 to 180 thousand purchases per day. However, the minimum activity found on Tuesdays does not correlate to any known human behaviour pattern.

Figure 5 shows the average number of purchases and clicks per hour, in a 24 hour cycle. As expected, the circadian cycle is very strong. We find a minimum in the activity in the site around 2h, and two peaks of activity during the day; a local maximum in the morning circa 9h; and a global maximum in the evening at around 18h. These maximums in daily activity are most probably related to the beginning and end of the working day. However, Figure 4 shows that the pattern on Saturdays is very different, with the site activity growing with time until 19h.

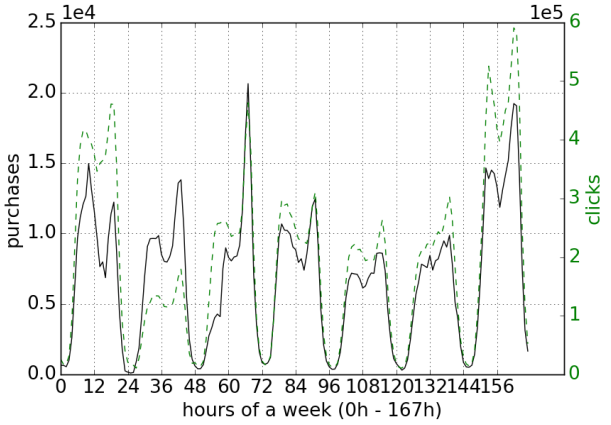

Figure 4: Weekly clicks and purchases.

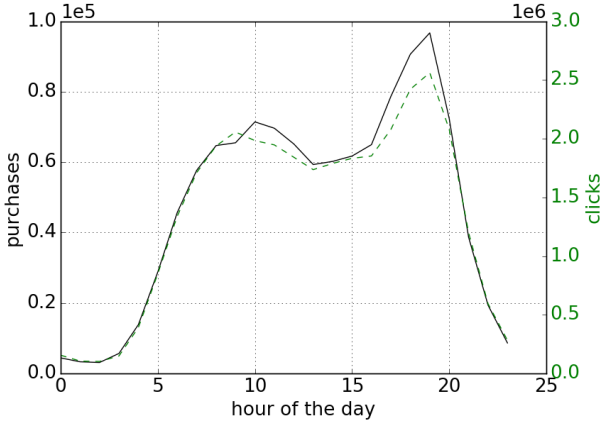

Figure 5: daily clicks and purchases.

Moreover, the most interesting pattern seen here is how the conversion rate (ratio of number purchases per clicks) varies with time. Not surprisingly, the highest conversion rates are found in Saturday (max ~ 0.050 at circa 21h) and Sunday (max ~ 0.048 at circa 24h). Thus, the probability of a customer to purchase an item on the weekend is higher than during the weekdays.

From this analysis, we can infer that user behaviour is not time independent, and it varies with month, day and hour of the day. This was an expected result, since the average person's schedule on Monday is not the same as Saturday and Sunday, and weeks are usually divided into weeks and weekends most of the countries. Therefore, in this section we have shown that temporal features are key to predict user behaviour, and thus, the models developed for this purpose should include them. Apart from the results presented, we also explored other patterns, such as the distribution of activity across each month, but no clear trend resulted from the analysis, thus we do not include those figures.

4. A CATEGORIES ANALYSIS

In this section we study the categories and their relationships. There is a total of 339 anonymised categories recorded in this dataset. Twelve of them are categories of items and 327 refer to product brands. However, the category data is not complete, and circa 49% of the click events have no associated categories. Moreover, *sales* is encoded as an special category, and $\sim$ 32% of the clicks correspond to items in sale. For our analysis, we focus on the 12 most relevant categories, excluding brands and sale items.

The category distribution across items and clicks follows a long tail distribution, similar to the distribution of items across clicks (Figure 1). Thus, the most popular category accounts for $\sim$ 5% of the clicks, while the following most popular category accounts for 10 times less clicks ($\sim$ 0.05%).

Since items can be found in more than one category, we can study the correlation between them by analysing the co-ocurrence of categories across items. Figure 6 shows the normalised co-ocurrence of items among categories. As this results on a symmetric matrix, we only represent the upper triangular part. Even when the number of categories in the dataset is small, we see that some categories are correlated. For example, categories 8 and 10 are highly correlated, similarly to categories 4 and 8 or 1 and 5. This could indicate either very similar categories, or an existing underlying categories taxonomy. However, as the data is completely anonymised, it is difficult to draw further conclusions.

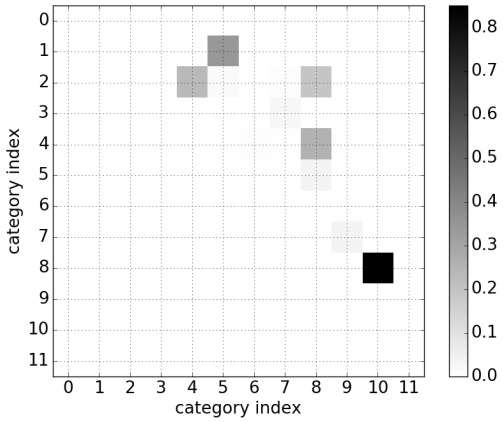

Figure 6: Normalised category co-ocurrence.

Finally, in our experiments we have also observed that even if users tend to click on a very small number of items, those items tend to be in the same category. For example, $\sim$ 89% of the sessions do all clicks on a single category, while only $\sim$ 0.11% click on two categories. A low number of categories within a session could explain a shopping intention, wether sessions with higher number of clicks may be more *explorative* sessions. To draw further characterise user intent, more domain knowledge is needed. However, as the dataset is anonymised, this analysis is not possible.

5. CONCLUSIONS

In this paper we have performed an analysis of the *Yoochoose dataset*. We have presented a summary of the main statistical properties of the dataset, its users and items. We have also exposed the main temporal characteristics, paying special attention to the most prominent cycles in the data. Finally, we have presented a brief analysis on the interactions between user, items and categories.

We have shown that the dataset has similar characteristics to most of the active recommender systems problems. It is very sparse, and both user behaviour and item behaviour follow a long-tailed distribution. This highlights the suitability of building recommender systems to enhance content discovery as well as increase and diversify sales.

We have shown interesting trends that describe user behaviour across different time granularities and cycles. We have observed that, as expected, users' behaviour is not time independent, and that the behaviour follows a circadian cycle. Finally, we have found interesting correlation between categories that might motivate the creation of a taxonomy, even if we cannot further investigate this since categories are anonymized.

As for future work, we are currently developing a solution based on a supervised classification approach that includes temporal and category-based features and we plan to further experiment with those features in future work in different datasets.

Furthermore, to enable reproducibility and comparability of this work, we plan to share the source code used to generate all the tables and figures shown here in a publicly available github repository https://github.com/hcorona/ceri-2016

6. ACKNOWLEDGMENTS

This work has been partially inspired by the discussions with the Recommender Systems research group in the Insight Centre for Data Analytics, as well as comments and discussions in the three original blogposts we published during the length of the *RecSys Challenge 2015* in http://hcorona.github.io/blog.html/ (with circa 2500 visits in total)

7. REFERENCES

[1] C. Anderson. The Long Tail. *Wired Mag.*, 12, 2004.

[2] D. Ben-Shimon, A. Tsikinovsky, M. Friedmann, B. Shapira, L. Rokach, and J. Hoerle. Recsys challenge 2015 and the yoochoose dataset. In *Proceedings of the 9th ACM Conference on Recommender Systems*, pages 357–358. ACM, 2015.

[3] O. Celma. *Music recommendation and discovery in the long tail*. PhD thesis, 2008.

[4] N. Koenigstein, N. Nice, U. Paquet, and N. Schleyen. The Xbox Recommender System. In *Proc. Sixth ACM Conf. Recomm. Syst.*, pages 281–284, 2012.

[5] Y. Koren. Collaborative Filtering with Temporal Dynamics. *Commun. ACM*, 53(4):89–97, 2010.

[6] G. Linden, B. Smith, and J. York. Amazon.com Recommendations Item-to-Item Collaborative Filtering. *IEEE Internet Comput.*, 1(February):76–80, 2003.

[7] M. K. Park, Chan Ho. Temporal Dynamics in Music Listening Behavior: A Case Study of Online Music Service. In *Proc. Ninth IEEE Int. Conf. Comput. Inf. Sci.*, pages 573–578, 2010.

[8] P. Romov. RecSys Challenge 2015 : ensemble learning with categorical features. 1, 2015.

[9] J. B. Schafer, J. Konstan, and J. Riedl. Recommender Systems in e-commerce. *Proc. 1st ACM Conf. Electron. Commer.*, pages 158–166, 1999.