Authors:
Rasul Dent
Justine Diliberto
Anna Nikiforovskaja
Cindy Pereira

Intelligent systems and recommendations project

M2 NLP

Université de Lorraine, IDMC

1 Data exploration and analysis

This first report presents the primary steps of the development of a recommender system, that are data exploration and analyses.

1.1 Presenting dataset

The dataset studied is a part of the Diginetica corpus. It contains 6 dataframes, dealing with e-commerce session-based data. The aim is to implement a recommender system to provide an optimized product ranking. Users have been anonymized, sessions are bounded by periods of inactivity, and queries without clicks are not represented. Furthermore, "a single user might have multiple sessions, and one session can have multiple users (if user re-logins into another account)".

1.2 Primary data attributes

- item id: integer representing a distinct product
- category id: integer representing a set of distinct products
- product name tokens: hash-codes corresponding to words
- query id: serial label for query
- sessionId: serial label for unique session
- time-frame: milliseconds since beginning of session
- priceLog2: price after log transformation
- event dates: what day the click/view/purchase occurred

1.3 Data files and questions they can answer

- products: Which products and hashed keywords correspond with different price brackets?
- product categories: Which items belong to the same categories?
- train purchases: Which items were bought together in the same session?
- train items views: How far into a given session did a user view an item?
- train/test queries: All of the above -> which products should be recommended in a session?

Number of items	184,047
Number of users	204,789
Number of clicks	1,127,764
Number of views	1,235,380
Number of purchases	18,025
Number of categories	1,217

Table 1: Data counts

1.4 Statistics

We have calculated some statistics on the dataset to understand the possible issues while building a recommendation system.

The first thing we can notice just by seeing numbers of items, users and purchases is that our data is extremely sparse. This is a common problem while building a recommendation system, however we will have to choose the recommendation methods with that in mind.

Afterwards we have decided to look at the distribution of ratios between views and purchases. In fig. 1 we can see, that in general the ration is quite small, the most popular ratios are between 0 and 0.1 and there are approximately 10^5 such cases. However, sometimes it is still possible to have for one item the number of purchases to be higher than the number of views. We think it might be because people sometimes need specific items more than once. This can be true for nails, light bulbs or even car tires, which are never bought as one instance.

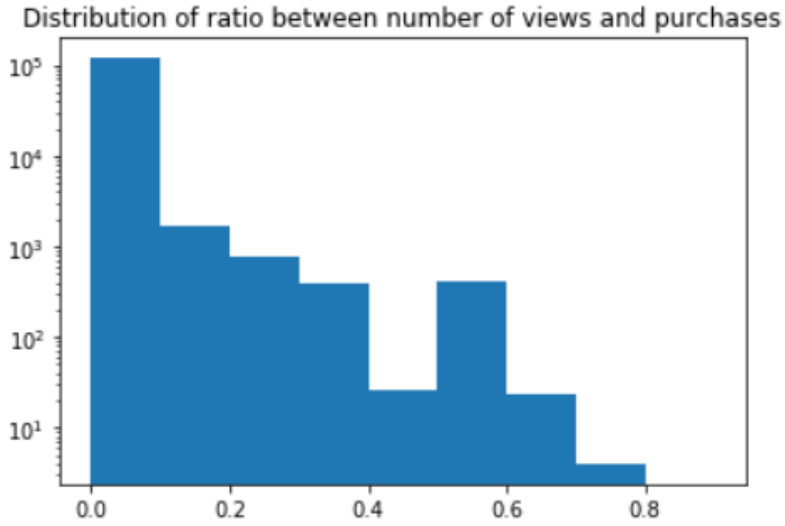

Figure 1: Distribution of ratios between number of views and purchases, on axis x we see the ratio, on axis y the number of items with this ratio.

We also study a distribution of ratio between the number of clicks and number of views, which is shown in fig. 2. An interesting part here is that there is now not a huge difference between different ratios in terms of how many items have the same ratio. And there are also some products which are viewed more than clicked. This might be a sign a user came to the product from outside of the site, as the clicks from outside of the site are not represented in our data. The ratio of products more viewed than clicked is of approximately 0.3069.

In addition to this, the distribution of the number of items per category has been computed and can be viewed in fig. 3. We can clearly notice a regular decrease in the quantity of categories with bigger number of items. This means the categories are not evenly distributed. In other words, some categories are most common than others, and there are many categories that have very few items.

Finally, we studied the price distribution. In fig.4 we can see that the items cost either less than 20¤, or between 45¤ and 100¤. This can be explained by the fact that people often buy cheap products (less than 20¤), or they will choose quality, which will significantly increase the price. A few products will have a price in between, but they are exceptions.

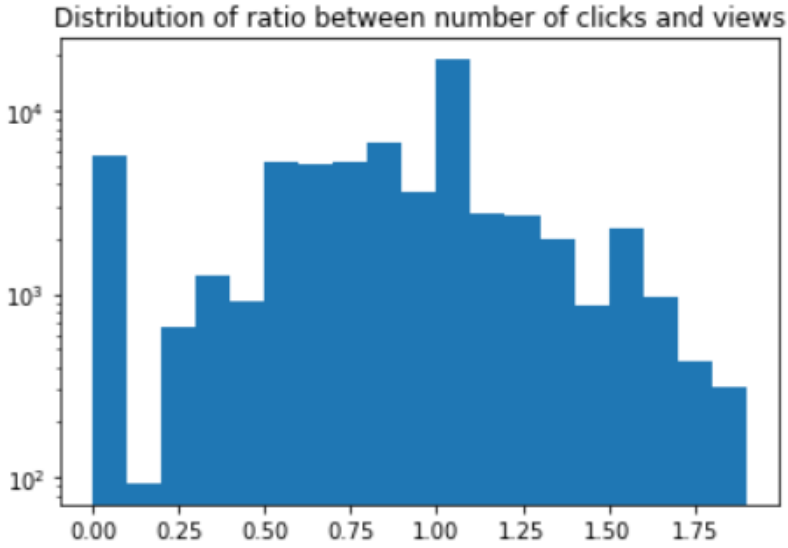

Figure 2: Distribution of ratios between number of clicks and views, on axis x we see the ratio, on axis y the number of items with this ratio.

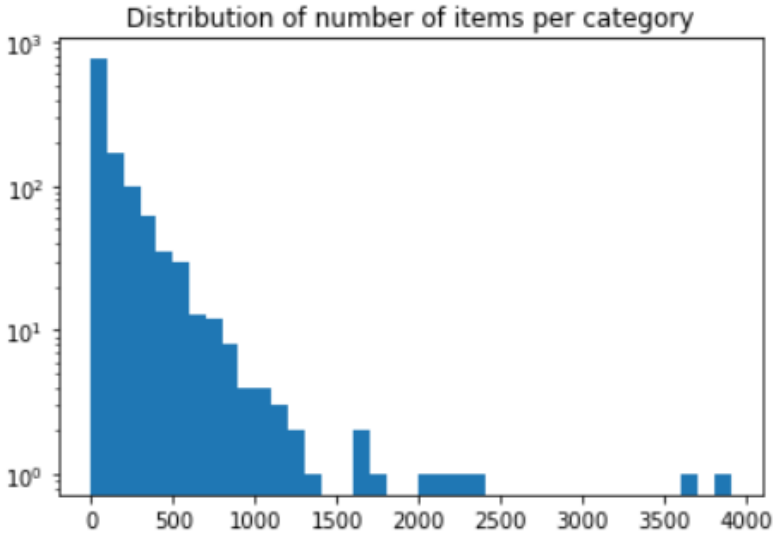

Figure 3: Distribution of number of items per category. On axis x we can see the possible numbers of items in categories, on axis y the amount of categories with this number of items.

1.5 Conclusions

We have a variety of data relating users, products, time, and price. Although we are able to note multiple correlations, several questions remain regarding the distribution of prices and how to account for purchases involving large quantities. As usual, we will have to reduce the dimensionality and redundancy of the dataset.

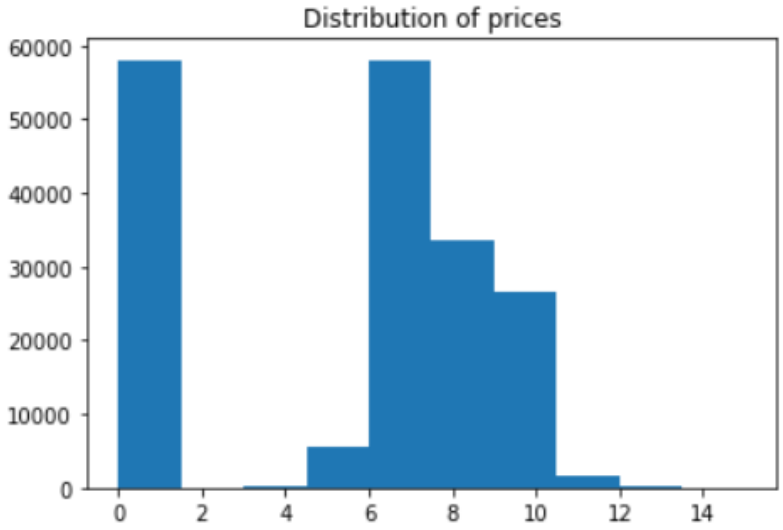

Figure 4: Distribution of the prices. On axis x we can see the log2 price, and on axis y the number of items involved.