Insight into User on E-commerce Search

dgliu

September 26, 2018

Part I
Data Preprocess and Statistics

We use the following procedure to generate the experimental data used in this paper. The statistical information of the data set can be found in Tab 1.

(i) Collect all query and action data from 20180726 to 20180804 and remove invalid data.

(ii) Divide the session by determining whether the time interval between adjacent queries is greater than 30 minutes.

(iii) Remove user with more than 14 sessions.

(iv) Remove data with the length of session, keyword and click more than 10, 32 and 100 respectively.

Table 1: The statistics of the data set

Timestamp	20180726-20180804
Session	926,850,016
User	247,790,305
Item	168,960,322
First level category	158
Leaf category	14,194

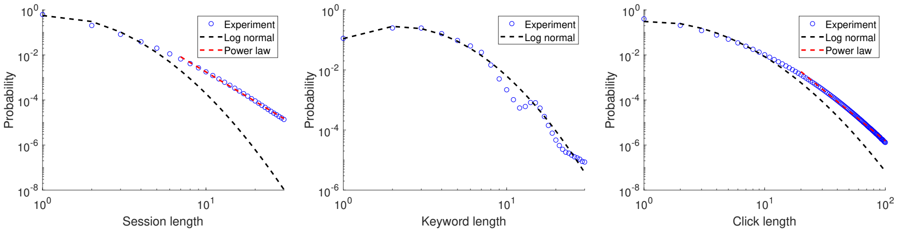
The question we are interested in is whether users follow certain commonalities in query. To answer this question, we first start with the length of the query and the action, then analyze the modification pattern of the query, and finally give the relationship between the query and the action pattern.

Part II
The Length

Considering the characteristics of query in E-commerce, we give the distribution of session length, click length and keyword length and fit them in turn. We will also explain the possible reasons with the optimal distribution.

The distribution of **Session length** are shown in Fig 1(a). The length follows distribution that is like a log-normal at low and intermediate values, with a characteristic peak and turnover, but transitions to a power-law distribution at high values. Similarly, we also observed that the distribution for **Click length** follows a modified log-normal with a power-law (MLP) distribution in Fig 1(c). This shows that in e-commerce search, users usually want to capture the target through fewer queries or clicks. But there are still small groups that are willing to spend more time inquiring and clicking to find satisfactory products.

From Fig 1(b), we can see that **Keyword length** follows a log-normal distribution, perhaps because users usually use modifiers and targets as query keywords, and have a restricted length.



(a) Session (b) Keyword (c) Click

Figure 1: The distribution of length

Further, we want to know how users combine these three aspects to produce personalized strategies, or what mainstream strategies exist in e-commerce search scenarios. We used binned scatter plots to examine correlations between variables. The result is shown in Fig 2, Fig 3 and Fig 4, where the overall keyword length and click length are expressed as mean values for *session length* $>$ 1.

Part III
The Modification

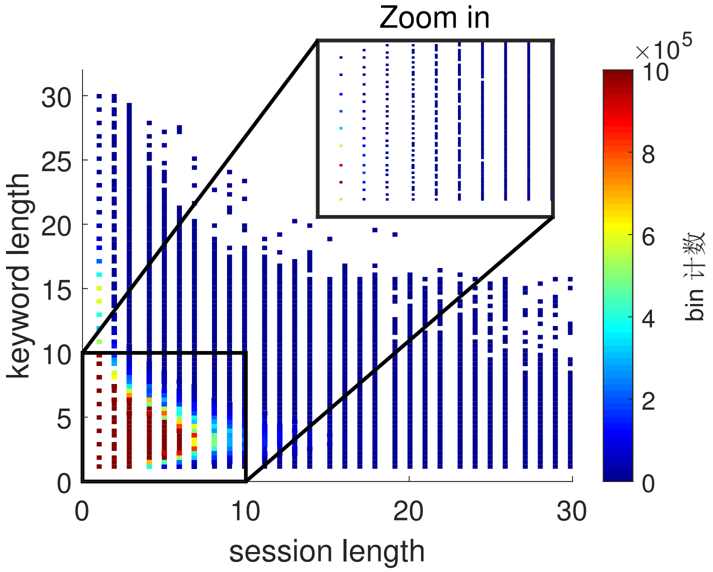


Figure 2: Session vs Keyword

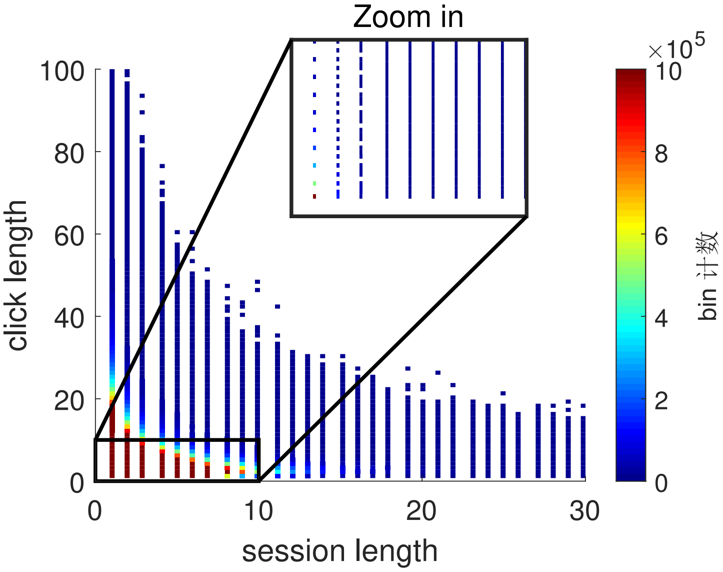


Figure 3: Session vs Click

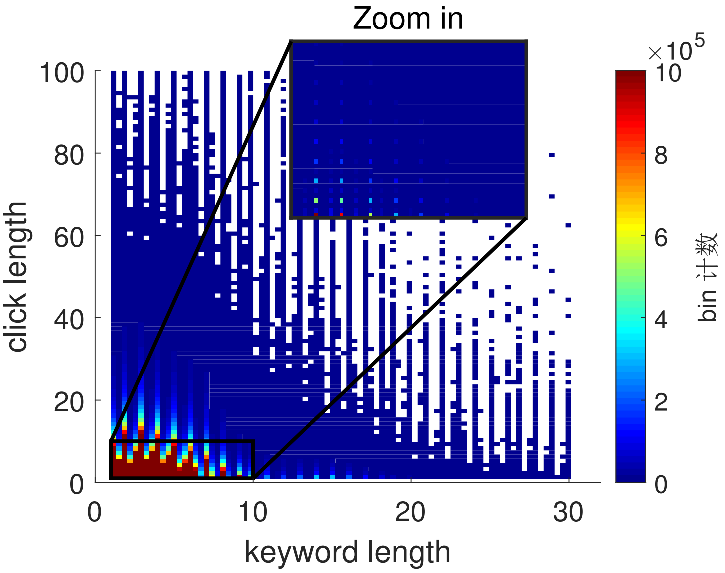


Figure 4: Keyword vs Click