# How to Find Goals? Insight into the Wisdom of User Based on E-commerce Search Sessions

Anonymous Author(s)

## ABSTRACT

...

## KEYWORDS

...

**ACM Reference Format:**
Anonymous Author(s). 2018. How to Find Goals? Insight into the Wisdom of User Based on E-commerce Search Sessions. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

...

## 2 RELATED WORK

...

## 3 DEFINITIONS AND DATA

In this section, we first introduce the definition of the *search sessions* and *search tags* in e-commerce. We then describe how to construct session data from a e-commerce search log and give statistics for the data set. Finally, we present the distribution of data set and compare them to traditional web search.

### 3.1 Search Sessions

In the field of e-commerce, the system typically creates, stores and terminates a new active session with simple triggering rules. This is, the user opens the page (or application) as a signal to create, and closing the web page (or application) indicates termination. All actions during this time period would be stored, including the search logs we are interested in. The actions involved in the search include entering query keywords, clicking on items, and operating on the items, i.e. add to favourite, add to cart and purchase. However, this means that a user continuously searches for different goals when he does not close the page, such as T-shirts and jeans, and only one actual session is counted. Direct analysis on raw data can lead to ambiguous results.

To accurately learn hidden search strategies, we need to properly define the *Search Sessions* to ensure continuity of search behavior and consistency of search goals. This is similar to the problem of detecting session boundaries [1–6]. Most existing work identifies session boundaries using time constraints, where the interval between two adjacent actions cannot exceed a preset threshold $\theta_t$ [1–3]. This method is simple and intuitive, because if the user's next action is too far from the previous one, his intentions are likely to have changed. But this method suffers from the question of determining reasonable parameters. As the previous work presented, when the threshold was set to 5, 10 and 30 minutes, the best results were obtained for different experimental data [1–3]. An unreasonable threshold would also result in a split session containing multiple goals.

Considering the above shortcomings, [4–6] proposes to use *tasks* to characterize a user's goal and further extract a task-based session. This method guarantees the consistency of the session to a certain extent, but requires a lot of extra calculations. For example, [4] needs to construct a classifier based on features such as keywords and search results, and [5] needs to combine text and semantic features to calculate similarity. Fortunately, in an e-commerce scenario, the item clicked by the user provides a new perspective to discover the user's goal. Assuming that a user continues to search and click on different T-shirts, we naturally think that his goal is to buy a T-shirt.

Based on above discussion, it is sound to define session by constrain both time and goal. We define *Search Sessions* in the following way:

Definition 1. *A search session is a sequence of actions that are generated by a user for a specific goal over a continuous period of time.*

Let $S$, $Q$, $I$, and $O$ represent the search session, query keywords, clicked items, and operations on the items respectively, the definition can be formalized as $S = \{(Q_1, <I_1, O_1>), (Q_2, <I_2, O_2>), ..., (Q_n, <I_n, O_n>)\}$. It should be noted that $I$ would contain one or more items, and $O$ would be a null value when the corresponding item is not operated.

### 3.2 Search Tags

After discovering search strategies, we also care about what search results these strategies bring, such as success or failure. Considering the characteristics of e-commerce, we use the *Session Tags* to represent the search results and define it in the following way:

Definition 2. *The tags would be divided into three categories based on the user's operation on the items, which are Purchase class, Add class (add to favourite or cart) and Stroll class (null).*

### 3.3 Data

The search log we use in this work was sampled over a month's period from Taobao[1]—China's largest C2C e-commerce platform. We construct search session data in three steps. First, We extract the historical behavior of each individual user from the search log in chronological order. Second, we segment each user's stream into

[1] www.taobao.com


*Conference'17, July 2017, Washington, DC, USA*

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

search sessions based on the Def. 1. Finally, We label all sessions using Def. 2. We also discard all invalid data through some rules, such as crawling behavior.

The data set consists of information including query keyword, clicked items and additional operation. The metadata for the item includes the title, price, category, and so on. All interactions are timestamped. The statistics of the data set are shown in Tab. 1[2]

**Table 1: Statistics of the data set**

| | |
|---|---|
| Timestamp | July 2018 |
| Session | 926,850,016 |
| User | 247,790,305 |
| Item | 168,960,322 |
| Category (First Level) | 158 |
| Category (Leaf Level) | 14,194 |
| Single Query | 84.21% |
| Multiple Queries | 15.79% |
| Purchase | 14.15% |
| Add | 19.09% |
| Stroll | 66.76% |

## 3.4 Distribution

After building the data set, we investigate the properties of the search session in e-commerce and show how they differs from traditional web search.

The distribution of session lengths are presented in Fig.1. The length follows the lognormal distribution $p(\theta) \propto \theta^{-\gamma}$ at the head ($length < 3$), and the power-law distribution $p(\theta) = \frac{1}{\sqrt{2\pi}\sigma\theta} e^{\frac{-(\ln\theta-\mu)^2}{2\sigma^2}}$ at the tail ($length \geq 3$), where $\gamma = 6.5464$, $\mu = 0.1306$ and $\sigma = 0.3190$. This observation is very different from web search where the length follows the power law distribution except $length = 1$ and $\gamma > 6.5464$[7, 8]. This means that in e-commerce, users have a greater probability of taking two or even longer queries. Users use more energy to find goals, which also gives us more opportunities to analyze hidden search strategies. Fig.2 shows the distribution of click lengths, and we observe the same properties as the session lengths, where $\gamma = 4.8034$, $\mu = 0.6337$ and $\sigma = 0.7463$. Most users would use a small number of clicks, while still some users maintain a high yield of clicks.

In Fig.3, we observe that the distribution of keyword lengths follows a lognormal distribution, where $\mu = 1.0145$ and $\sigma = 0.5554$. This property is slightly different from the previous findings in web search [9, 10], where the distribution of keyword lengths is fitted with a power-law and Poisson distributions (not including partial header data). But in both, users tend to use simple keywords to describe goals or needs. When users need to perform query reformulation, they are more likely to modify according to certain strategies rather than blindly.

# 4 SINGLE QUERY

Previous work has focused on sessions with multiple queries. But as shown in Tab.1, single query have a very large proportion of all sessions, so we think it's necessary to analyze them independently.

[2]For the sake of simplicity, the last five items are shown as a percentage.

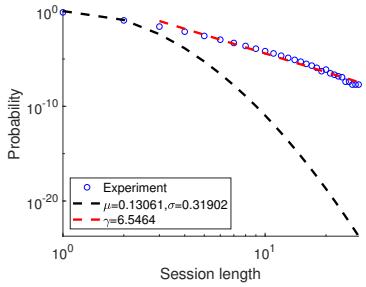


**Figure 1: The distribution of session length.**

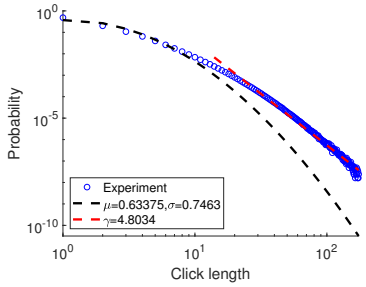


**Figure 2: The distribution of click length.**

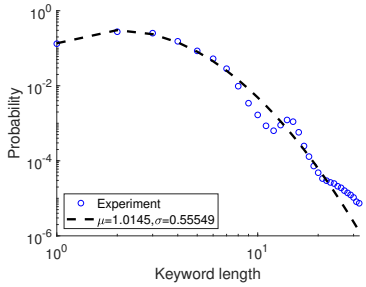


**Figure 3: The distribution of keyword length.**

In this section, we first combine a variety of features to find useful search strategies from a single query session. Then we connect these strategies and session tags to investigate how different strategies lead to different outcomes.

## 4.1 Search Strategies

# 5 MULTIPLE QUERIES

...

# 6 PREDICTING SEARCH SUCCESS

...

# 7 EXPERIMENTS

...

# 8 CONCLUSION

...

## REFERENCES


[1] C. Silverstein, H. Marais, M. Henzinger and M. Moricz. Analysis of a very large web search engine query log. *ACm SIGIR Forum*, 33(1):6-12, 1999.

[2] D. He, A. Göker and DJ. Harper. Combining evidence for automatic web session identification. *Information Processing & Management*, 38(5):727-742, 2002.

[3] BJ. Jansen, A. Spink and V. Kathuria. How to define searching sessions on web search engines. In *International Workshop on Knowledge Discovery on the Web*, pages 92–109. Springer, 2006.

[4] R. Jones and KL. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 699–708. ACM, 2008.

[5] C.Lucchese, S. Orlando, R. Perego, F. Silvestri and G. Tolomei. Identifying task-based sessions in search engine query logs. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 277–286. ACM, 2011.

[6] Z. Liao, S. Yang, L. He and Y. Huang. Evaluating the effectiveness of search task trails. In *Proceedings of the 21st international conference on World Wide Web*, pages 489–498. ACM, 2012.

[7] H. Cao, DH. Hu, D. Shen, D. Jiang, JT. Sun, E. Chen and Q. Yang. Context-Aware Query Classification. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 3–10. ACM, 2009.

[8] H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen and H. Li. Context-Aware Query Suggestion by Mining Click-Through and Session Data. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 875–883. ACM, 2008.

[9] A. Arampatzis and J. Kamps. A study of query length. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 811–812. ACM, 2008.

[10] T. Kramar, M. Barla and M. Bieliková. Personalizing Search Using Socially Enhanced Interest Model Built from the Stream of User's Activity. *Journal of Web Engineering*, 12(1&2):65-92, 2013.