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, **??** 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*, *search strategies* 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.

Since we are interested in the strategy that users take in the process of finding a goal, we need a different definition of the session to ensure the continuity of the search behavior and the consistency of the search target. This is similar to the problem of detecting session boundaries [**? ? ? ? ? ?**]. Most existing work identifies session boundaries using time constraints, where the interval between two adjacent actions cannot exceed a preset threshold θ_t [**? ? ?**]. 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 [**? ? ?**]. An unreasonable threshold would also result in a split session containing multiple goals.

Considering the above shortcomings, [**? ? ?**] 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 instance, [**?**] needs to construct a classifier based on features such as keywords and search results, and [**?**] 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. On the other hand, using the clicked item to determine the consistency of the goal does not lead to additional calculations.

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, c, o and G represent the search session, query keywords, clicked items, operations on the items and a specific goal respectively, the definition can be formalized as $S_i = \{(q_1, < c_1, o_1 >), (q_2, < c_2, o_2 >), ..., (q_n, < c_n, o_n >)\}$, where $|t_{q_j} - t_{q_i}| < \theta_t$ and $\{c_1, c_2, ..., c_n\} \in G$. It should be noted that c_i would contain one or more items, and o_i would be a null value when the corresponding item is not operated.

3.2 Search Strategies and Search Tags

Based on Def. **??**, a search session consists of search keywords, clicks and additional operations. Since the operation implicitly reflects the user's satisfaction with the item and is unknown to the user in advance, the variables that the user directly manipulates in the search session are the remaining two. We define the *search strategies* as follows:

Definition 2. A *Search Strategy* refers to how users coordinate search keywords and click behavior to better hit the goal.

There are two sources of strategy: the experience gained from the long-term interaction between the user and the e-commerce search system, and the other is the user's own style of behavior. In

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

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

addition, an important prerequisite for the strategy is that the user has a strong purpose before starting the search session. It should be emphasized that not all search sessions are strategic, e.g. the user may just pass the time or the user lacks experience in how to improve search results.

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

Definition 3. The *Search Tags* would be divided into four categories based on the user's operation on the items, which are Add+Purchase class (add to favourite or cart before the final purchase), Purchase class (only purchase), Add class (only add to favourite or cart) and Stroll/Fail 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 search sessions based on the Def. **??**. Finally, We label all sessions using Def. **??**. 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 query keywords are pre-processed by removing punctuation (except decimal points and dashes), and word segmentation. 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. **??**[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	84.23%
Multiple Queries	15.77%
Purchase	14.03%
Add	19.07%
Stroll/Fail	66.90%

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. **??**. 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 $\gamma = 6.6809$, $\mu = 0.13046$ and $\sigma = 0.31897$. This observation is very different from web search where the length follows the power law distribution except $length = 1$ and $\gamma > 6.6809$ [**? ?**]. 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. **??** shows the distribution of click lengths, and we observe the same properties as the session lengths, where $\gamma = 4.7678$, $\mu = 0.6329$ and $\sigma = 0.74641$. Most users would use a small number of clicks, while still some users maintain a high yield of clicks.

In Fig. **??**, we observe that the distribution of keyword lengths follows a lognormal distribution, where $\mu = 1.016$ and $\sigma = 0.55551$. This property is slightly different from the previous findings in web search [**? ?**], 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 it is necessary to perform query reformulation, they are more likely to modify according to certain strategies rather than blindly.

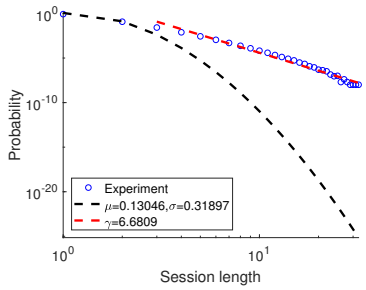

Figure 1: The distribution of session length.

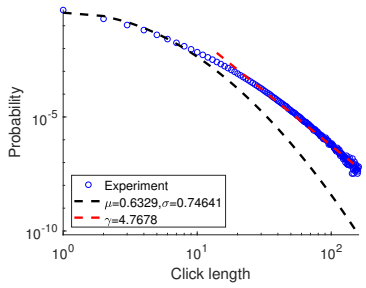

Figure 2: The distribution of click length.

4 SINGLE QUERY

Most of the previous work has focused on sessions with multiple queries. But as shown in Tab. **??**, single query have a very large proportion of all sessions, and we think it is necessary to investigate them separately. In this section, we first combine a variety of features to extract useful search strategies from a single query session. We then connect these strategies and session tags to investigate how different strategies lead to different results.

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

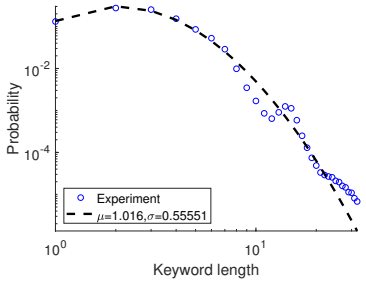

Figure 3: The distribution of keyword length.

Table 2: Centroids of the strategy clusters

Cluster	w_l	c_l	pos	pos_e	seq	seq_e	t	t_e
1	2.715	1.002	1.225	0	2.118	0.001	23.303	0.001
2	4.821	1.000	1.204	0	2.212	0.001	24.501	0
3	2.891	1.007	1.865	0	8.728	0.006	22.397	0
4	2.825	2.653	1.089	0	4.026	0.996	26.067	0.986
5	2.978	3.856	2.869	0.937	5.496	0.970	23.192	0.980
6	3.561	5.172	19.693	0.813	6.204	0.805	19.051	0.830
7	3.458	12.117	5.504	0.909	5.910	0.953	19.525	0.964
Outlier	6.584	3.274	4.121	0.945	4.724	0.934	47.384	0.946

4.1 Strategies Extraction

4.2 Strategies and Tags

Table 3: Proportion of tags for each strategy cluster

Cluster	Add+Purchase	Purchase	Add
Average	0.0739	0.0642	0.1722
1	**0.0463**	0.0693	**0.1079**
2	**0.0510**	0.0755	**0.1106**
3	**0.0343**	**0.0433**	**0.0964**
4	**0.1079**	0.0841	0.1896
5	**0.0925**	0.0507	**0.2477**
6	0.0811	**0.0277**	**0.2640**
7	**0.1271**	0.0426	**0.4258**
Outlier	0.0950	0.0664	0.1853

5 MULTIPLE QUERIES

5.1 Motif Finding

...

5.2 Similarity of Query pair

...

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.
[11] S. Chawla and A. Gionis. k-means--: A unified approach to clustering and outlier detection. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 189–197. Society for Industrial and Applied Mathematics, 2013.