Investigating COVID-19-Related query logs of Chinese search engine users

Ziyi Ye | Jiaxin Mao | Yiqun Liu | Min Zhang | Shaoping Ma

Tsinghua University, Beijing, China

Correspondence
Ziyi Ye, Tsinghua University, Beijing, China.
Email: yeziyi1998@gmail.com

Abstract
Query logs include valuable information for understanding user intent and behavior in Web search. In this article, we investigate COVID-19-related query logs by dividing search sessions into different intent and analyzing the user behavior of groups and individuals. We believe it important to learn about the epidemic's influence on users' search behavior and refine search engine to confront similar epidemic outbreaks in the future.

KEYWORDS
COVID-19, query log, user behavior

1 | INTRODUCTION

With the rapid development of the Internet, many people tend to look for health-related information and issues online. There are approximately 80% of Internet users who searched for health information online, 66% of whom began their last online health inquiry at a search engine (S. Fox, 2006). Previous studies built connection between Internet activity and the development of an epidemic. The query logs of search engine are considered to be informative for forecasting the epidemic because users explicitly express their epidemic-related search intent in their queries. Xu et al. (2011) showed that online surveillance systems based on query logs can detect the outbreak of epidemic earlier than traditional surveillance system based on clinical reports.

During the last several months, coronavirus disease 2019(COVID-19) has fundamentally changed people's life. Chinese search engine users are among the first ones to generate millions of related search records due to the early impact of the epidemic. In those health-related search, understanding information goals and patterns of user behavior is essential for user psychological analysis and giving correct medical guidance (White & Horvitz, 2013). However, search intent and behavior during an epidemic are rarely studied. In this work, using query logs from a large number of Chinese search engine users, we attempt to understand users' search intent during the epidemic and characterize users' search behavior. By developing an intent classification model, we monitor the changes of the focus of public attention during the development of the epidemic. The findings may be useful for designing a better Web search system to help people pursue epidemic-related information.

2 | METHODOLOGY

We conduct our experiments on Sogou.com, a widely used web search engine in China. Our experiments are divided into four steps. Firstly, we generate a manual-rule-based model to identify a session's relevancy to COVID-19. Secondly, we sample a certain number of users and extract epidemic-related search sessions. Thirdly, we design an intent classification taxonomy and train a supervised model for intent prediction. Finally, we analyze the users' search behavior for each intent in related sessions.
According to the epidemic-related query extraction algorithm based on the click-through graph (Xu et al., 2011), we extract 1,492 epidemic-related queries and segment them into terms. Within all the terms, we manually select hundreds of epidemic-related terms and obtain search sessions containing at least one of these terms. To obtain a more accurate dataset, we label 1,200 queries (split into train set and valid set) and design manual rules\(^1\) to identify its relevancy to COVID-19. The recall of our manual-rule-based model in valid set achieves 91%.

After that, we randomly sample 7,967,953 sessions of 102,485 users according to users’ IP and search frequency. The manual-rule-based model is applied to each query in the session, and if a third of queries in the session are related to COVID-19, we regard the whole session as COVID-19-related.

Thirdly, we manually label each session with five labels (see Table 1). Labels are not mutually exclusive and each of them is a binary decision. Each session is labeled by three annotators and the Fleiss’ \(\kappa\) achieves 0.88, indicating a high inter-annotator agreement. Based on these annotated data, we use BERT (Jacob 2019) with a linear layer as our encoder and fine-tune the pretrained bert-based-Chinese model.\(^2\) A ten-fold cross validation is used to evaluate the BERT-based model. The results shown in Table 2 indicate that the proposed model achieves a reliable intent classification performance.

Finally, we analyze the frequency of COVID-19-related sessions and user behavior measures including queries per session, queries per second, click probability, click position, and time to first click for each intent.

## 3 RESULT AND DISCUSSION

Figure 1 shows the frequency trends of COVID-19-related sessions among the selected users. Due to the influence of the outbreak, the frequency of sessions related to COVID-19 increased sharply in mid-to-late January and gradually declined thereafter. The frequency of real-time outbreak sessions is relatively high, indicating that a large proportion of the population paid continuous

### TABLE 1  The intent classification taxonomy for COVID-19 related search

| Intent                     | Description                                      | Example query in related session |
|----------------------------|--------------------------------------------------|----------------------------------|
| Real-time outbreak         | The development of the epidemic in time and space | COVID-19 in Beijing              |
| Related news and information | COVID-19 related news and information             | Taiwan bans the export of gauze masks |
| Related transaction        | The impact of the epidemic on its own affairs and related policies | Schools’ opening date in Beijing |
| Medical science popularization | Basic medical knowledge and prevention method of COVID-19 | How to wear a mask? |
| Symptoms, diagnosis and treatment | Pneumonia-related symptom, diagnosis and treatment | Fever; treatment procedures for pneumonia |

### TABLE 2  Result of fine-grained intent classification experiment in the ten-fold cross validation

| Task                                | Tag       | Precision | Recall  | F1-score | Number |
|-------------------------------------|-----------|-----------|---------|----------|--------|
| Relevance                           | Irrelevant| 0.6348    | 0.8588  | 0.73     | 255    |
| Real-time outbreak                  | Irrelevant| 0.9688    | 0.8986  | 0.9323   | 1242   |
| Related news and information        | Irrelevant| 0.9474    | 0.9016  | 0.9239   | 1199   |
|                                    | Relevant  | 0.7882    | 0.8798  | 0.8314   | 499    |
| Related transaction                 | Irrelevant| 0.9334    | 0.8401  | 0.8843   | 1351   |
|                                    | Relevant  | 0.5519    | 0.7666  | 0.6417   | 347    |
| Medical science popularization      | Irrelevant| 0.9391    | 0.8287  | 0.8805   | 1378   |
|                                    | Relevant  | 0.5104    | 0.7688  | 0.6135   | 320    |
| Symptoms, diagnosis and treatment   | Irrelevant| 0.9455    | 0.8138  | 0.8747   | 1407   |
|                                    | Relevant  | 0.4620    | 0.7732  | 0.5784   | 291    |
attention to the spread of COVID-19. The frequency change of related transaction sessions is the least obvious because people are continuously facing new impacts of COVID-19 (such as rent change, schools’ opening date, etc.). The sessions of medical science popularization and related news and information intentions declined rapidly after the peak, indicating that the public’s curiosity on epidemic-related information cooled quickly after its sudden eruption.

Table 3 shows the distribution of the clicked websites. Chinese search engine users tend to use Sogou, Weibo, and other portals other than official websites like Chinese centers for disease control that provide first-hand information. With a more user-friendly service, these commercial portals attract more users. The distribution of clicked websites also reflect users’ main demand. The most popular websites include epidemic maps, news platforms (Sogou, Weibo), online medical consulting platforms (Sogou online medical consulting), knowledge-sharing platforms (Zhihu, Douban), and online shopping platforms for medical goods.

Tables 4 and 5 summarizes the short-term user behavior characteristics in COVID-19-related sessions compared to other sessions and sessions with different fine-grained intentions respectively. We find that the average click position of COVID-19-related search is lower, while time to first click is longer and queries per second are lower. It’s possible to cost more time to digest professional medical information. Moreover, sessions with symptoms, diagnosis, and treatment intention take much more time to click or reformulate their query, which suggests that users are more careful when they have this kind of search intent.

Our findings indicate long-term search frequency changes and significant differences between user behavior of different search intent. A commercial search engine system might benefit from our quantitative analysis, that is ambiguous queries can be understood according to epidemic development and special user behavior, and search engine result page might be tailored for users who want

![FIGURE 1 Frequency trends of COVID-2019 related sessions](image)

**TABLE 3** Distribution of partial popular clicked websites

| Website                          | Description                                                                 | Fraction  |
|---------------------------------|-----------------------------------------------------------------------------|-----------|
| Sogou’s epidemic map            | An integrated service platform fouse on the spread of epidemic               | 20.8%     |
| Sohu                            | An online media sires                                                       | 6.2%      |
| Weibo                           | A microblogging and social networking service website                       | 4.1%      |
| Zhihu                           | An online knowledge-sharing platform                                         | 4.7%      |
| Sogou Q&A                       | An online general Q&A platform                                              | 2.0%      |
| Encyclopedia website of famous doctor | An online health-related Q&A platform                                      | 1.1%      |
| Youlai doctor                   | An online health-related Q&A platform                                       | 1.1%      |
| Chinese CDC(Center for Disease Control) | Chinese center for disease control website                              | 5 * 10^-3% |
| Chinese government website      | Chinese local government website                                            | 1.0%      |

**TABLE 4** Short-term user behavior characteristics in COVID-2019 related sessions compared to other

| Characteristics                             | Other sessions | COVID-2019 related sessions |
|---------------------------------------------|----------------|-----------------------------|
| Queries per session                        | 4.3878         | 4.098                       |
| Queries per second (10^-3s^-1)              | 9.90           | 18.68                       |
| Average click position                      | 2.7983         | 3.2347                      |
| Average click probability                   | 0.476          | 0.5271                      |
| Time to first click(s)                      | 17.511         | 13.976                      |
to learn about professional medical information. Additionally, our work gives us new insights into the information needs of Chinese people in the context of COVID-19. Thus, it is helpful to characterize information environments in a global health crisis.

4 | CONCLUSION

Based on the query logs collected from a Chinese web search engine amid COVID-19, we define a fine-grained search intent taxonomy for COVID-19-related searches and propose methods for extracting related search sessions and identifying users' intent. After finding the related sessions and identifying the intent behind them, we analyze user behavior within all related search sessions and that for each intent, respectively. We find that the change of search frequency during epidemic time varies under different search intent. We also investigate the distribution of clicked websites and the difference in short-term user behavior characteristics. Our work helps to understand users' search intent and behavior during an epidemic outbreak and thus promotes the Internet to provide better service. Future work will include further investigation based on those processed search records and comparison of other epidemics in different periods.

ENDNOTES

1 Manual Rules is designed by matching against special keywords that generate from training set
2 The input sequence of our model is concatenation of queries in a session, which are separated by “[SEP]” token.

REFERENCES

Fox, S. (2006). Online health search 2006. Pew Internet & American Life Project.
Xu, D., Liu, Y., Zhang, M., Ma, S., Cui, A., & Ru, L. (2011, June). Predicting epidemic tendency through search behavior analysis. In Twenty-Second International Joint Conference on Artificial Intelligence.
White, R. W., & Horvitz, E. (2013). Captions and biases in diagnostic search. ACM Transactions on the Web (TWEB), 7(4), 1–28.
Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

How to cite this article: Ye Z, Mao J, Liu Y, Zhang M, Ma S. Investigating COVID-19-Related query logs of Chinese search engine users. Proc Assoc Inf Sci Technol. 2020;57:e424. https://doi.org/10.1002/pra.2424

### TABLE 5 Short-term used behavior characteristics with different fine-grained intentions

| Characteristics                  | Real-time outbreak | Related news and information | Related transaction | Medical science popularization | Symptoms, diagnosis and treatment |
|----------------------------------|--------------------|-------------------------------|---------------------|-------------------------------|----------------------------------|
| Queries per session              | 3.145              | 4.870                         | 5.103               | 4.574                         | 3.316                            |
| Queries per second ($10^{-3} s^{-1}$) | 12.75             | 9.53                          | 13.03               | 11.73                         | 9.68                             |
| Average click position           | 2.1814             | 2.8279                        | 3.2322              | 3.3019                        | 3.5039                           |
| Average click probability        | 0.5204             | 0.4774                        | 0.5519              | 0.5924                        | 0.5742                           |
| Time to first click (s)          | 20.23              | 16.50                         | 16.51               | 22.72                         | 26.12                            |