A Large Scale Study of English-Chinese Online Dictionary Search Behavior

Yong Liu and Jianmiao Fan

Indiana University, Bloomington, Indiana, USA
{yonliu,jfan}@dict.cn

Abstract. This paper presents a large scale study of user search behavior on a popular online dictionary website. Our goal is to understand the current status of online dictionary, especially English-Chinese dictionary user search behavior by analyzing 10 million queries on Dict.cn website during the period form late 2006 to early 2007. We believe our findings will help traditional dictionary publishers and online dictionary providers to improve their existing products and services, and further develop innovative services to better serve the unique needs of their users.

Keywords: Online dictionary, search interface, English-Chinese dictionary.

1 Introduction

Hundreds of millions of people in China are learning and using English on a daily basis. Many of them happen to be Internet users. The huge demand on English learning and the well established Internet infrastructure make online English-Chinese dictionary services more and more popular today. The goal of this study is to better understand the search behavior of online dictionary users. The study results can be used to help paper dictionary publishers and online dictionary providers to improve the existing products and services, and hence meet the unique needs of their users.

Dict.cn is a leading English-Chinese online dictionary website offering the largest English-Chinese dictionary database (over 3 million terms) to Internet users for free. It ranks among top 3000 most popular websites globally and serves over 1.5 million queries worldwide daily. In this study, we present analysis on Dict.cn's search logs collected through its Web interface, Web API, MSN bot, and WAP interface. The search log used in this study consists of 10 million queries randomly sampled during a three month period in late 2006 and early 2007. Both English and Chinese queries are included in this data set. Each search log entry contains fields including the user No., the query, the corresponding translation and a timestamp. All of our data is strictly anonymous, which means we maintain no data that can match a user with an identity. All of the results we report in this paper are aggregate statistics.

This study covers basic query statistics such as query length, queries per day, and repetitive queries. More sophisticated analysis is also conducted to examine the relation between users and their search patterns. For instance, one question this study tries to answer is how users forget what they learnt from a dictionary in a certain

C. Stephanidis (Ed.): Universal Access in HCI, Part III, HCII 2007, LNCS 4556, pp. 389–398, 2007.

period. Another important question is that if it is possible to classify users according to their search behavior. By answering these research questions, we believe this study will help paper dictionary publishers and online dictionary providers to achieve a better understanding of their customers' search behavior and hence improve their existing products or services.

The rest of this paper is organized as follows. Section 2 discusses some related work in user search behavior analysis field. Section 3 introduces multiple search interfaces provided by Dict.cn. Section 4 presents our dictionary query analysis based on a large scale search log data set. Finally section 5 gives a conclusion.

2 Related Work

A lot of research has been done on user search behavior analysis on general purpose search engines [2, 3, 6, 7]. These research efforts extracted user search patterns by studying a large set of search logs. The analysis results can be later used to improve the service quality of search engines. However, only a limited number of literatures discussed online dictionary user search behavior in the past decade.

In [8], Bergenholtz, H. et al. discussed using search logs to analyze user behavior in using an online Danish dictionary and how the results could be used to improved online dictionaries. In [9], De Schryver, et al. showed how a combination of search log analysis and the processing of formal online feedback forms may lead to improving dictionary contents. The same authors also studied the relation between search frequency and corpus frequency of online dictionary lemmas, and claimed their findings can be used to guide the compilation of future dictionaries [10]. However, most of these previous research efforts were based on a limited scale of online dictionary search logs involving only several thousand website visitors and tens of thousands queries. Such limitation made the reported statistical analysis results less convincing. Another thing the previous research failed to examine is the similar behavior pattern among multiple individual online dictionary users. Compared to traditional paper dictionaries, online dictionaries have great potential to facilitate the interactions among users during their searching and learning process. Study on the relations among individual users may help online dictionary providers to develop innovative services to better meet the unique needs of their users.

Different from previous research efforts on online dictionary search log analysis, our study is based on a large scale search log data set which came from a popular online dictionary website extensively used in real world. We focused not only on query statistics analysis, but also on user clustering by similar search behavior.

3 Dict.cn Search Interfaces

Dict.cn features four major dictionary search interfaces, which are Web interface, Web API, MSN bot, and WAP interface. The Web interface (a in Fig. 1) is Dict.cn's most visited search interface which Internet users access directly through a Web browser. Dict.cn also provides a MSN bot (b in Fig. 1), which users can add to their MSN messengers as a friend, to enable dictionary service via a chat interface. Dict.cn's WAP interface (c in Fig. 1) serves queries sent from cell phones while the

users are on the go. Last but not least, the Web API (d in Fig. 1) provided to third party developers by Dict.cn allows users to access Dict.cn's dictionary service via desktop software, toolbars, and browser plugins.

In the entire search log used in this study, 88.1% of queries came in through the Web interface. Besides, 8.7% of queries were received from the Web API and 2.6% from the MSN bot. The WAP interface contributed least to Dict.cn, which only received 0.4% of the total queries.

Fig. 1. Dict.cn Query Interfaces

4 Query Analysis

In this study we examined characteristics of user search behavior on Dict.cn. The six aspects we covered were query length, query difficulty, query distribution, queries per day, repetitive queries, and query based social network analysis.

4.1 Query Length

The entire search log consists of two types of queries – the English queries which were submitted by users looking for the corresponding Chinese interpretations and the Chinese queries, vice versa. For the English queries, we found the average number of words per query to be 1.67 (median = 1, max = 61, standard deviation = 1.4), with an average number of 10.87 (median = 9, max = 100, standard deviation = 7.79)

characters per query[1]. For the Chinese queries, the average number of Chinese characters per query was 5.39 (median = 4, max = 50, standard deviation = 4.92) [2]. Figure 2 shows the length distributions of both English and Chinese queries.

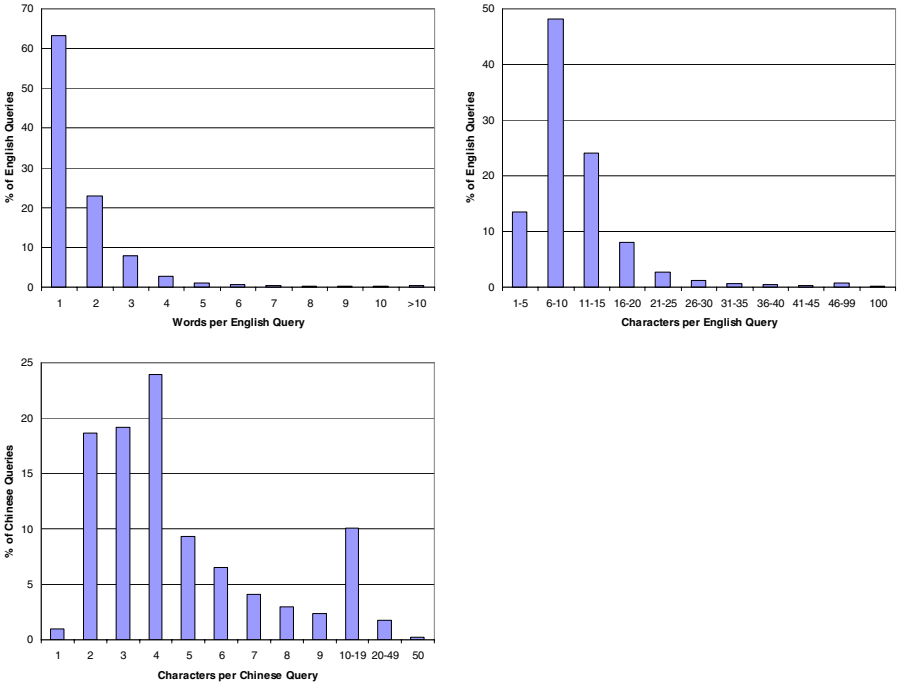

Fig. 2. Distribution of number of words and characters per English query and number of characters per Chinese query for Dict.cn queries received from all interfaces

It is interesting to compare the average query length between Dict.cn and general purpose search engines. Some published statistics suggested the average number of words per English query for search engines to be 2.35 [2] and 2.6 [3]. It is reasonable that the dictionary queries are much shorter on average since different from a search engine, words, not phrases, are usually the building blocks of an online dictionary.

Four-character query is the most popular Chinese queries according to Figure 2. However, previously reported average word lengths in various Chinese corpora are all less than 3 characters [4, 5]. This discrepancy suggests that online dictionary users tend to use combinations of two or more Chinese words to form search queries.

4.2 Query Difficulty

Previous research suggested that frequency of use is a valid measure of English word difficulty and can be utilized in vocabulary test construction and translation [1]. In

[1] The maximum number of characters in an English query allowed by Dict.cn is 100.
[2] The maximum number of characters in a Chinese query allowed by Dict.cn is 50.

this study we used this metric to examine the difficulty distribution of English queries on Dict.cn. The query difficulty measure *D(x)* used in this study is defined as following. As can be seen, the smaller *D(x)* is, the harder query *x* is.

$$D(x) = \text{sqrt}(\text{times of query } x \text{ submitted/number of total queries}). \quad (1)$$

In the search log used in this study, 60.3% of the English queries were submitted by users located in China, and 39.7% were from users outside of China. The average query difficulty of China based English queries is 8.22×10^{-3}(median = 0.0081, max = 0.00246, standard deviation = 0.0028). And the average query difficulty of non-China based English queries is 7.95×10^{-3} (median = 0.0076, max = 0.0246, standard deviation = 0.003). This means that users outside of China tend to submit more difficult English queries. This result is not surprising considering the fact that these users usually have better English language capability than Chinese domestic users.

Figure 3 shows the distribution of averaged English query difficulty for both China and non-China based user groups. This again demonstrates that Chinese domestic users tend to submit easier English queries to Dict.cn. For the better understanding of the meaning of query difficulty, here are some examples on query difficulty: *D*("imprimatur") is 4.5×10^{-3}, *D*("privacy") is 8.3×10^{-3}, and *D*("next") is 1.8×10^{-2}.

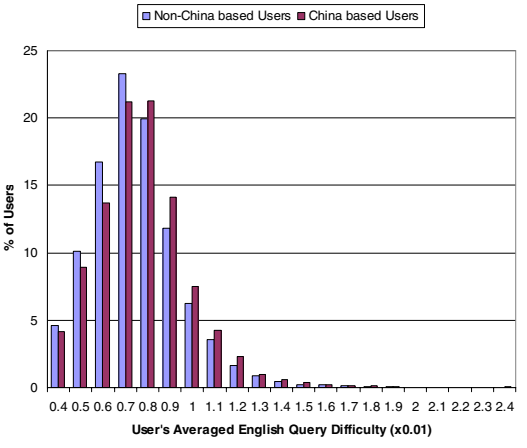

Fig. 3. Difficulty level of English queries by users

4.3 Query Distribution

More than one million queries are submitted to Dict.cn daily. However, the number of unique queries submitted is much smaller. To study the variation in the queries, we examined the percentage of the total query volume that was covered by top-N unique queries. We picked the mostly used 1000 English and Chinese queries from the dictionary search log and calculated the distribution of top 1 to 1000 queries.

As Figure 4 shows, English queries vary more than Chinese queries do. The percentage of total searches covered by the top 100 Chinese queries is 4.3% while the percentage covered by the top 100 English queries is only 2.4%. However, the difference between the coverage of top 1000 English and Chinese queries is not that much (13.1% vs. 15.9%). This suggests that Chinese queries focus more heavily on the top couple of hundred unique queries.

Previously a similar study reported that top 1000 search engine queries accounted for approximately 6% of all queries [6]. The same study also reported that top 1000 mobile search queries accounted for around 22% of all mobile queries. These findings suggest that general purpose search engine queries are much more diversified than dictionary queries, whereas mobile search queries are less diversified. This suggests that the usable contents on the existing mobile Internet is still not as rich as the contents provided by a large scale online dictionary.

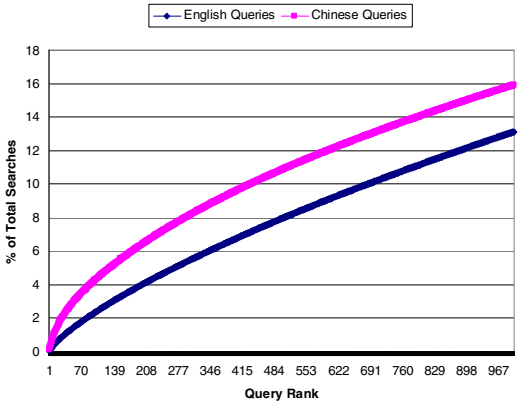

Fig. 4. Cumulative percentage of total searches accounted for by the top 1000 English and Chinese queries

4.4 Queries Per Day

As part of our study, we also examined the distribution of the number of queries submitted by users per day. Figure 5 illustrates the study results for both China based user groups and non-China based user groups. As can be seen, the largest user portions in both groups are those who submitted only one query per day. However, both groups feature a long tail in the figure. There are a fairly large portion of users who submitted more than 20 queries per day in both groups. The more queries a user submits, the better his or her search behavior can be profiled. Based on a user's profile, more customized services can be provided in the future.

Figure 5 also shows that non-China based users tended to submit more queries per day. The reason might be that these users had stronger needs to use a dictionary since they stayed in a non-Chinese language environment.

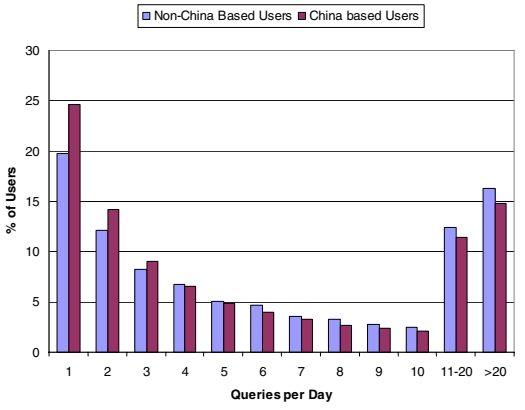

Fig. 5. Queries per Day by China and Non-China based Users

4.5 Repetitive Queries

People forget things. And this happens a lot during the learning process. To study how fast our users forget their newly learnt words, we examined the number of repetitive queries a user launched in a 10 day period. Figure 6 shows the result.

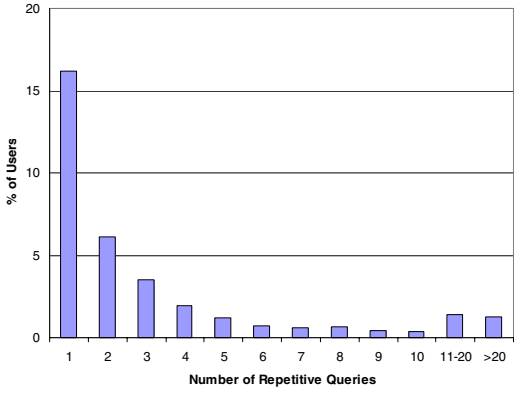

Fig. 6. Number of repetitive queries by one user in 10 days

We found that 34% of users at least searched for an identical word or phrase twice in 10 days, and 2.7% of users searched for more than 10 identical words. This finding explains why dictionaries always sell, and why a dictionary website can survive without any updates for a long period.

4.6 Query Based User Clustering

An identical query may be submitted to an online dictionary by different users during a certain period. The users who did this may share the same background or the same

interests. For instance, users who searched for the phrase "hyper threading" have a high possibility of either having computer science background or being interested in computer technology. Also, users who submitted the same queries may have similar language capability. If we can find those similar users (in terms of background, interest, language capability, etc.) based on the search logs, we will be able to provide innovative services to online dictionary users such as collaborative study.

In this study we randomly picked 1022 users who submitted at least 30 queries in a one month period. We examined the number of users who shared a certain number of queries with at least one other user. As can be seen from Figure 7, more than 20% of users shared at least 10 queries with one or more other users. Around 10% users shared at least 15 queries. This result demonstrates the feasibility of finding a similar peer for at least 10% of total users since 15 identical queries can largely guarantee that two users have some similarity between them. Given a longer period of the search log collection, the matching of two similar users can be more accurate.

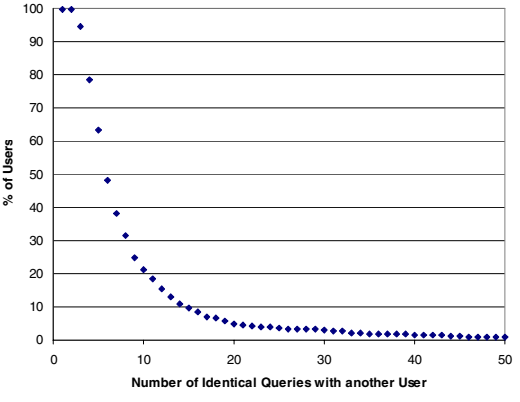

Fig. 7. Number of identical queries with at least one other user

We also examined the possibility of clustering users into separate user groups to maximize the similarity within each group. By applying a fast clustering algorithm published in [11], we clustered the before mentioned 1022 users into 125 groups. Figure 8 (a) shows a query similarity matrix of users where each spot denotes the corresponding pair of users who had at least 3 identical queries. Since the users are randomly arranged along two axes, the similarity matrix looks totally random. Figure 8 (b) shows the 3D similarity matrix in a diagonal perspective. As can be seen, the majority of the similarity pairs have 3 or 4 identical queries. Figure 8 (c) shows what the similarity matrix is like after users are re-arranged along two axes according to the 125 clustered groups (users in the same group are placed adjacent to each other on x and y axis). We can see now most points are gathered along the principal diagonal of the similarity matrix. Figure 8 (d) provides a better view where we clearly see the similarity points, especially those with 5 or more identical queries, are now gathered tightly along the diagonal. This demonstrates that the clustering algorithm works very well on the online dictionary users based on their queries.

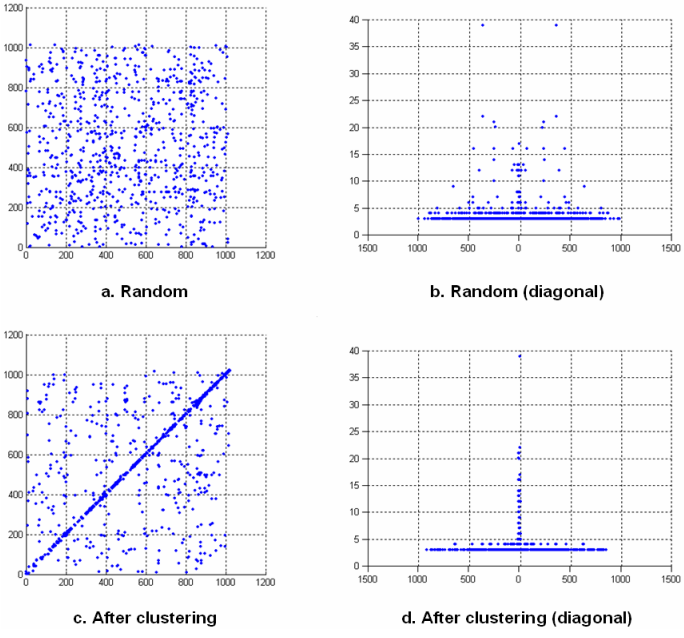

Fig. 8. Query based user clustering

5 Conclusions

In this paper we examined online dictionary user search behavior based on a large scale search log data set of a leading online English-Chinese dictionary website. The study results suggest that:

- On average, online dictionary queries are shorter than search engine queries.
- English queries submitted by Chinese domestic dictionary users are less difficult than those submitted by non-China based users.
- Online dictionary queries are less diversified than search engine queries.
- Non-China based users query more per day on average.
- A large number of users submit repetitive queries in a short time period.
- Online dictionary users can be well clustered into separate groups according to the similarity of their search behavior.

Although we have presented some findings on dictionary query based user clustering in this paper, we believe there are still many interesting open research issues in this topic. This will be the subject of our next large scale study on online dictionary user search behavior.

Acknowledgments. Here we would like to gratefully acknowledge the help of all Dict.cn users. This paper would not be possible without their input.

References

1. Tamayo, J.M.: Frequency of Use as a Measure of Word Difficulty in Bilingual Vocabulary Test Construction and Translation. Educational and Psychological Measurement 47(4), 893–902 (1987)
2. Silverstein, C., Henzinger, M., Marais, H., Moricz, M.: Analysis of a Very Large Web Search Engine Query Log. SIGIR Forum 33(1), 6–12 (1999)
3. Spink, A., Jansen, B., Wolfram, D., Saracevic, T.: From E-Sex to E-Commerce: Web Search Changes. IEEE Computer 35(3), 107–109 (2002)
4. Nie, J.Y., Gao, J., Zhang, J., Zhou, M.: On the Use of Words and N-grams for Chinese Information Retrieval. In: Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages, Hong Kong, China, pp. 141–148 (2000)
5. Xu, J., Zens, R., Ney, H.: Do We Need Chinese Word Segmentation for Statistical MachineTranslation? In: Proceedings of the Third SIGHAN Workshop on Chinese Language Learning,Barcelona, Spain, pp. 122–128 (2004)
6. Kamvar, M., Baluja, S.: A Large Scale Study of Wireless Search Behavior: Google Mobile Search. In: Proceedings of the SIGCHI conference on Human Factors in computing systems (2006)
7. Wen, J.R., Nie, J.Y., Zhang, H.J.: Clustering User Queries of a Search Engine. In: Proceedings of International WWW Conference (10). Hong-Kong (2001)
8. Bergenholtz, H., Johnsen, M.: Log Files as a Tool for Improving Internet Dictionaries. Journal of Linguistics 34, 117–141 (2005)
9. De Schryver, G.M., Joffe, D.: On How Electronic Dictionaries are Really Used. In: Proceedings of the Eleventh EURALEX Interna-tional Congress, Lorient, France, pp. 187–196 (2004)
10. De Schryver, G.M., Joffe, D., Joffe, P., Hillewaert, S.: Do Dictionary Users Really Look Up Frequent Words? – On the Overestimation of the Value of Corpus-based Lexicography. Lexikos 16, AFRILEX-reeks/series 16, 67–83 (2006)
11. Frey, B., Dueck, D.: Clustering by Passing Messages between Data Points. Science DOI: 10.1126/science.1136800 (2006)