Mobile Query Reformulations

Milad Shokouhi
Microsoft
Cambridge, UK
milads@microsoft.com

Rosie Jones
Microsoft
Cambridge, MA
rojone@microsoft.com

Umut Ozertem
Microsoft
Sunnyvale, CA
umuto@microsoft.com

Karthik Raghunathan
Microsoft
Sunnyvale, CA
kraghu@microsoft.com

Fernando Diaz
Microsoft
New York, NY
fdiaz@microsoft.com

ABSTRACT

Users frequently interact with web search systems on their mobile devices via multiple modalities, including touch and speech. These interaction models are substantially different from the user experience on desktop search. As a result, system designers have new challenges and questions around understanding the intent on these platforms.

In this paper, we study the query reformulation patterns in mobile logs. We group query reformulations based on their input method into four categories; text-text, text-voice, voice-text and voice-voice. We discuss the unique characteristics of each of these groups by comparing them against each other and desktop logs. We also compare the distribution of reformulation types (e.g. adding/dropping words) against desktop logs and show that there are new classes of reformulations that are caused by errors in speech recognition.

Our results suggest that users do not tend to switch between different input types (e.g. voice and text). Voice to text switches are largely caused by speech recognition errors, and text to voice switches are unlikely to be about the same intent.

Categories and Subject Descriptors

Information Systems [**Information Retrieval**]: Information retrieval query processing—*Query reformulation, query log analysis*

Keywords

Speech retrieval, speech recognition, mobile search

1. INTRODUCTION

People increasingly rely on mobile devices for various tasks, including web search. The volume of mobile queries has grown exponentially in recent years and is expected to exceed the number of queries submitted from desktop devices by next year.[1] In addition to traditional keyboard based input, mobile devices provide users with additional modalities such as gestures and speech. The ability to interact with software through speech has resulted in an increase in the number of queries spoken to search systems. For example, Google reported in 2010, that 25% of queries on Android devices are submitted by voice.[2]

In this paper, we study the query reformulation patterns in mobile logs. We conduct a large-scale analysis over the query logs of Bing search engine and demonstrate the differences in query distribution conditioned on the input method (voice or text). We show that there are fundamental differences in the query distributions and characteristics of query-reformulations in mobile and desktop logs. We verify and confirm some of the patterns reported in recent lab studies [7] over a large-scale dataset. We also compare various statistics such as overall reformulation effectiveness per *input type* that cannot be collected at scale in the lab due to natural limitations of such labs studies. Our results suggested that users rarely switch between different input methods. That is, if the submitted query is typed, the reformulation is also likely to be typed and if the submitted query is by voice, then the reformulation is likely to be submitted as a voice query too.

In addition, we compare the distribution and taxonomy of mobile reformulations against desktop logs. Our comparative analysis reveals that text reformulations are similar across desktop and mobile logs. However, when users switch from voice to text, they are relatively more likely to substitute their query words to correct the speech recognizer errors. Switches from text to voice often happen when the user is searching for a new intent, and overall users are less likely to use abbreviations in their mobile queries.

2. RELATED WORK

Query Reformulation & Rewriting. Users frequently reformulate queries throughout sessions to complete their search tasks [8]. Several taxonomies of query reformulation has been suggested in the past [1, 5]. However, all these studies focused on desktop queries and were published before mobile devices became popular. The information retrieval literature is rich with related techniques that leverage query reformulations and clicks in the past user logs, however, to the best of our knowledge, this is the first large-scale study on mobile query reformulations.

Users tend to reformulate their queries when they are not happy with search results [4]. Hence, query-reformulation patterns can be used to guide query re-writing. In such applications, query

[1] http://selnd.com/1maZlXC, last accessed 6 May 2014.

[2] http://bit.ly/Ljpy8A, last accessed 6 May 2014.

SIGIR'14, July 6–11, 2014, Gold Coast, Queensland, Australia.

http://dx.doi.org/10.1145/2600428.2609497.

reformulations are either mined and aggregated over the entire query logs [9] or they are alternatively captured in context to personalize search [17] and query suggestions [14].

Mobile Search. Kamvar and Baluja [10] presented one of the earliest large-scale analyses of mobile search queries. Compared to desktop queries, they found mobile searches to be topically less diverse, although similar in terms of the number of words and character in queries. They reported that users take substantially longer time to issue their queries on mobile devices and they suggested that this extra "effort" may explain why they observed on average fewer queries per session in their logs. The authors later conducted a similar analysis on more recent logs collected in 2007 (compared to 2005 for the original study) [11]. They found that in more recent mobile searches, users were typing faster and clicking on more results, and sessions were longer and less homogeneous.[3]

Church et al. [3] studied the mobile internet requests of more than 600,000 European mobile subscribers over a day in 2005.[4] Their results suggested that browsing happens substantially more often than searching on mobile devices, although search sessions were generally found to be longer than browsing sessions. They also found mobile queries to be on average shorter and more repetitive compared to desktop. The authors also reported that users are more likely to reformulate their mobile queries, making mobile search sessions on average longer than those on desktop.

Kamvar and Baluja [12] showed that contextual signals such as the application being used, and the user's location improve the ranking of query auto-completion suggestions on mobile devices. Kamvar and Beeferman [13] reported that users are more likely to use voice as their input method for local queries particularly on smaller keyboards, and are less likely do so for longer queries.

Most related to our work, Jiang et al. [7] conducted a lab experiment with 20 participants to study the query reformulation patterns of users when searching with mobile devices. They reported that speech recognition errors happen in about half of voice queries. In such cases, performance is often significantly affected, and users tend to reformulate their queries by removing and substituting words, which on average leads to slightly better queries and performance. In a follow up study, Jeng et al. [6] reported that users usually find voice search more challenging due to voice input errors caused by misrecognition of acronyms, pronunciations and other factors.

In another similar study, Schalkwyk et al. [16] compared the topical categories of mobile queries versus desktop. This paper however focuses on mobile query reformulations specifically.

3. QUERY REFORMULATIONS

Huang and Efthimiadis [5] categorized query reformulations in desktop search logs into 12 groups; word reorder, whitespace and punctuation, add words, remove words, url stripping, stemming, acronym, substring, superstring, abbreviation, word substitution, and spelling correction. Jiang et al. [7] noted that in addition to these 12 groups of *lexical* reformulations, there are other types of *phonetic* reformulations that can be observed when people search by voice. This latter group is largely caused by misrecognition of query words by the speech recognizer, and the repeat of the same (or a similar) query by the user. They reported based on a lab study that users put more stress on the whole or certain parts of the query when they reformulate their queries by voice.

[3]The query logs used in both studies only contained text queries.

[4]Note that their analysis pre-dates the smart-phone era, hence some of the findings may no longer be representative.

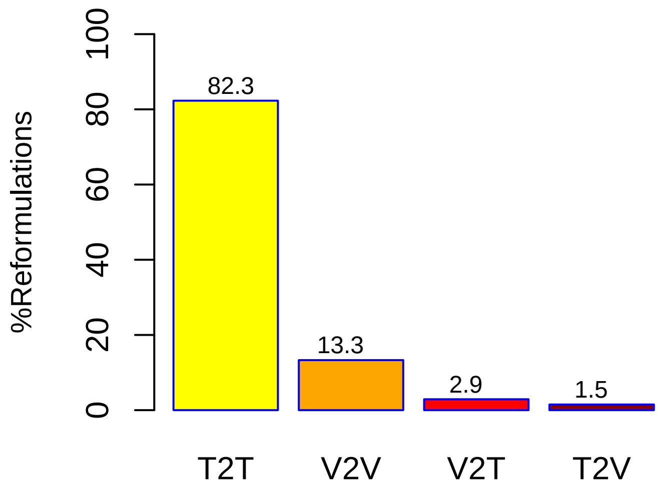

Figure 1: Distribution of reformulation modalities in the mobile search logs of a commercial search engine; text to text (T2T), voice to voice (V2V), voice to text (V2T) and text to voice (T2V).

Huang and Efthimiadis [5] focused entirely on desktop search queries and did not verify if query reformulations in mobile logs follow a similar distribution. Jiang et al. [7] focused on mobile reformulations but specifically *voice to voice* reformulations. While they noted that there are different types of reformulations (e.g. add words, remove words) in mobile search as well, they did not provide any statistics about the distribution of these reformulations. In addition, none of the previous studies investigates the properties of query reformulations with respect to their input method (voice to voice, text to text, text to voice, and voice to text).

The analyses in this paper address this gap. We investigate the distribution of different types of reformulations on mobile devices and compare them against a desktop query log. We also study the properties of query reformulation separately for each input method.

Modality in mobile search logs. On mobile devices, people can choose to search by voice or text. As a result, in a search session, not all queries are necessarily submitted with the same input method. Here, we present a break-down of query reformulations based on their input method. For our analysis, we sample the mobile search query logs of Bing search engine between 1 October 2013 and 31 October 2013 (1 month). In total there are 2,643,283 multi-query sessions containing 7,668,809 distinct queries from 706,763 unique users in our sample dataset.

We follow the common definition of a search session [2, 19] and define each session as a sequence of search actions (e.g. queries, clicks) with no periods of inactivity longer than 30 minutes. Investigating the most effective techniques for classifying session boundaries in mobile logs is itself an interesting area which is left for future work. For our study, we decided to go with the common choice of 30-minute inactivity windows to make our results comparable with previous work.

Figure 1 illustrates the distribution of reformulation modalities in our mobile dataset computed over all successive query pairs. Each query reformulation pair (q_1, q_2) consists of a pair of consecutive queries that appeared in the same search session. Here, text to text, voice to voice, voice to text and text to voice reformulations are respectively denoted by T2T, V2V, V2T, and T2V and represent cases where the first query was submitted by one input method, and the following query by another. There are at least two key trends that immediately stand out when looking at these distributions: (1)

users do not tend to switch between different input methods. If they have typed their first query, their next query is also likely to be typed and if they have issued a voice query, their next query is also likely to be submitted by voice. (2) Switching between voice to text – while infrequent – is still substantially more likely than the reverse (text to voice). Further investigations – covered later in this section – suggest that when users switch between text to voice, it is often a new search with a new intent, while switches from voice to text happen frequently when the original speech query is misrecognized by the engine.

Query reformulation types. We follow the steps of Huang and Efthimiadis [5] in categorizing query reformulations and comparing their properties. They categorized query reformulations in search logs into the following 12 groups:

1. Word reorder: query and reformulation share the same words but in different order (e.g. *quicktime download* → *download quicktime*).
2. Whitespace and punctuation: The reformulation is different from the query only in whitespace and punctuations (e.g. *quick time* → *quicktime*).
3. Remove words: The reformulation consists of a subset of query words (e.g. *quicktime download* → *quicktime*).
4. Add words: The query consists of a subset of reformulation words (e.g. *quicktime* → *quicktime download*).
5. URL stripping: The query and reformulation are the same after removing ".com", "www" and "http:" from both sides (e.g. *apple* → http://www.apple.com).
6. Stemming: Query and reformulation share the same stem according to the Porter stemmer [15] (e.g. *apple* → *apples*).
7. Form acronym: The reformulation is the acronym form of the query (e.g. *information retrieval* → *ir*).
8. Expand acronym: The query is acronym form of reformulation (e.g. *ir* → *information retrieval*).
9. Substring: The reformulation is a substring of the query (e.g. *quicktime* → *quick*).
10. Superstring: The query is a substring of the reformulation (e.g. *quick* → *quicktime*).
11. Abbreviation: The corresponding words from the query and reformulation are prefixes of each other (e.g. *greenleaf co* → *greenleaf colorado*).
12. Word substitution: When one or more words in the query are replaced with semantically related words in the reformulation (e.g. *obama bio* → *obama biography*). The semantic relatedness of words are determined according to the Wordnet database.[5]

We adopt the same taxonomy of Huang and Efthimiadis [5] and use their classifier to categorize query reformulations. When a query reformulation cannot be assigned to any of these groups it is called *new*, and it is assumed that it does not share the same intent as last query. This happened respectively for 64%, 72%, 74% and 91% of T2T, V2T, V2V and T2V mobile reformulations in our dataset. As a reference point, we also compare our results against a similar set of desktop sessions sampled from Bing search logs between 1 June 2014 and 31 August 2013 (two months). In total, 94% of reformulations in our desktop logs are categorized as new. Given that desktop searches have been reported to be more successful than mobile [3, 18], users are expected to satisfy their intents with fewer queries, and the higher ratio of new reformulations is not surprising. Similarly, the relatively high percentage of new T2V reformulations suggests that when users switch from text to voice, they often form a new session in terms of intent.

[5] http://wordnet.princeton.edu

Table 1: Distribution of reformulation types in mobile (🔊) query logs, by modality. The last column (⌨) represents the distribution over a typical sample of desktop query logs. Voice and text queries are respectively denoted by V and T. The reformulation categories presented here are based on the taxonomy suggested by Huang and Efthimiadis [5] (excluding reformulations that were categorized as *new*). For brevity we do not present those classes that accounted for less then 1% of cases across all types of reformulations. Specifically, these were *form acronym*, *expand acronym*, and *word reorder*.

Reformulation	🔊 T2T	🔊 V2T	🔊 V2V	🔊 T2V	⌨ DES
Abbreviation	8%	4%	4%	3%	12%
Add words	48%	38%	57%	47%	49%
Remove words	17%	16%	15%	24%	12%
Spell correct.	10%	23%	10%	8%	9%
Stem compare	2%	2%	1%	2%	1%
Sub-string	1%	1%	1%	1%	1%
Sup-string	2%	2%	2%	4%	2%
Space punct.	0%	0%	1%	2%	2%
URL Strip	2%	3%	1%	1%	3%
Word Subst.	8%	11%	7%	7%	9%

Table 1 contains the classification results on the mobile and desktop query logs excluding those reformulations that were categorized as new. The mobile and desktop reformulations are respectively specified by (🔊) and (⌨) symbols. Overall, there is a higher rate of spell correction in mobile reformulations which is somewhat expected given that the typing is generally easier on desktop devices. Word substitution reformulations occurs most frequently in mobile searches when switching from voice to text (V2T). The substituted words in many cases are the wrong predictions by the speech recognizer that had to be fixed by typing. In addition, URL stripping occurs more often when the reformulated query is submitted via text. People are more likely remove words in their mobile reformulations, and are unlikely to use abbreviations unless in their T2T reformulations.

While the numbers in Table 1 provide an overview of the distributions of query reformulations in mobile and desktop logs, they do not capture the effectiveness of such reformulations. Huang and Efthimiadis [5] defined *successful reformulations* as those cases in which the results for the original query received no clicks, but those returned for the reformulations had at least one click (SkipClick). Similarly, ClickSkip can be defined as those reformulations in which the original query has a click but the reformulation does not, and can be regarded as *unsuccessful* reformulations. Figure 2 illustrates the distribution of ClickSkip and SkipClick reformulations in desktop and mobile logs for different switches in modality. For comparison purposes, we also present the results for ClickClick scenarios in which both query and its reformulation had clicks on their results. The SkipClick ratio is highest among voice to text (V2T) searches. This again can be explained by speech correction errors that needed to be fixed by typing. Text to voice (T2V) reformulations have the lowest ratio of unsuccessful searches, and this again can be

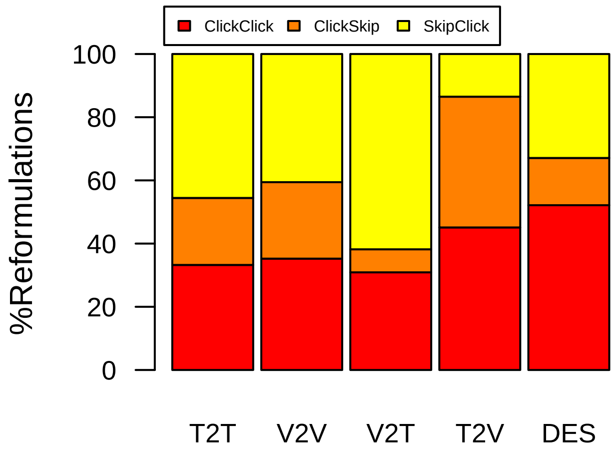

Figure 2: The percentage of query reformulations with at least one click with respect to different input types; text to text (T2T), voice to voice (V2V), voice to text (V2T) text to voice (T2V), and desktop (DES). The colors represent the recorded clicks in a query-reformulation pair; ClickClick (red) cases, both original query and its following reformulation were clicked, ClickSkip (Orange) the query had a click but its reformulation did not (*unsuccessful reformulation*), SkipClick (yellow) the query had no clicks but there was a click for its reformulation (*successful reformulations*).

explained by the fact that most of these reformulations are about *new* intents and hence are expected to perform relatively better.[6] Desktop reformulations (DES) have a higher ClickClick rate overall. Given that desktop search sessions have been reported to be longer and to have 20% higher clickthrough [18] these results are expected.

4. CONCLUSIONS

We presented a large-scale study of query reformulations in mobile search logs. Our results suggested that users rarely switch between different input types unless they are searching for a new intent or correcting speech recognizers errors. We showed that compared to searches on desktop, users are more likely to drop words from and spell-correct their mobile queries and less likely to use abbreviations.

In future work, we intend to investigate the effectiveness of using previous reformulations for improving the accuracy of speech retrieval in mobile searches.

References

[1] P. Anick. Using terminological feedback for web search refinement: A log-based study. In *Proc. SIGIR*, pages 88–95, Toronto, Canada, 2003.

[2] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui. Modeling the impact of short- and long-term behavior on search personalization. In *Proc. SIGIR*, pages 185–194, Portland, OR, 2012.

[3] K. Church, B. Smyth, P. Cotter, and K. Bradley. Mobile information access: A study of emerging search behavior on the mobile internet. *ACM Transactions on the Web (TWEB)*, 1 (1), 2007.

[4] A. Hassan, X. Shi, N. Craswell, and B. Ramsey. Beyond clicks: Query reformulation as a predictor of search satisfaction. In *Proc. CIKM*, pages 2019–2028. ACM, 2013.

[5] J. Huang and E. N. Efthimiadis. Analyzing and evaluating query reformulation strategies in web search logs. In *Proc. CIKM*, pages 77–86. ACM, 2009.

[6] W. Jeng, J. Jiang, and D. He. Users' perceived difficulties and corresponding reformulation strategies in voice search. In *Proc. HCIR*, Vancouver, Canada, 2013.

[7] J. Jiang, W. Jeng, and D. He. How do users respond to voice input errors?: Lexical and phonetic query reformulation in voice search. In *Proc. SIGIR*, pages 143–152, 2013.

[8] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *Proc. CIKM*, pages 699–708, Napa Valley, CA, 2008.

[9] R. Jones, B. Rey, O. Madani, and W. Greiner. Generating query substitutions. In *Proc. WWW*, pages 387–396, Edinburgh, UK, 2006.

[10] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *Proc. SIGCHI*, pages 701–709, Montreal, Canada, 2006.

[11] M. Kamvar and S. Baluja. Deciphering trends in mobile search. *Computer*, 40(8):58–62, Aug. 2007. ISSN 0018-9162.

[12] M. Kamvar and S. Baluja. The role of context in query input: Using contextual signals to complete queries on mobile devices. In *Proc. MobileHCI*, pages 405–412, Singapore, 2007.

[13] M. Kamvar and D. Beeferman. In T. Kobayashi, K. Hirose, and S. Nakamura, editors, *Proc. INTERSPEECH*, pages 1966–1969, Makuhari, Japan, 2010. ISCA.

[14] Z. Liao, D. Jiang, E. Chen, J. Pei, H. Cao, and H. Li. Mining concept sequences from large-scale search logs for context-aware query suggestion. *ACM Transactions on Intelligent Systems and Technology*, 3(1):17:1–17:40, Oct. 2011.

[15] M. F. Porter. An algorithm for suffix stripping. In *Readings in Information Retrieval*, pages 313–316. Morgan Kaufmann Publishers Inc., 1997.

[16] J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope. Your word is my command?: Google search by voice: A case study. In *Advances in Speech Recognition*, pages 61–90. Springer US, 2010.

[17] X. Shen, B. Tan, and C. Zhai. Context-sensitive information retrieval using implicit feedback. In *Proc. SIGIR*, pages 43–50, Salvador, Brazil, 2005.

[18] Y. Song, H. Ma, H. Wang, and K. Wang. Exploring and exploiting user search behavior on mobile and tablet devices to improve search relevance. In *Proc. WWW*, Rio de Janeiro, Brazil, 2013.

[19] R. White and S. M. Drucker. Investigating behavioral variability in web search. In *Proc. WWW*, pages 21–30, Banff, Alberta, Canada, 2007.

[6] Note that users are unlikely to reformulate unless they are unhappy with the original results [4]. Therefore relatively lower performance of text to text (T2T) and voice to voice (V2V) cases is expected as these are more likely to be searches about the same intent.