On Mobile User Behaviour Patterns

Milan Vojnović

Microsoft Research
7 J J Thomson Avenue, CB3 0FB Cambridge
United Kingdom
milanv@microsoft.com

January 2008

Technical Report
MSR-TR-2008-08

Microsoft Research
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052
http://www.research.microsoft.com

Abstract–We present empirical analysis of human searches for information from mobile devices, focusing on temporal dynamics, semantics, and topics of queries. Our analysis is based on a large scale data of mobile search logs over a week period from a major US mobile service provider. We find that human searches appear in bursts over time with the distribution of the query inter arrival time following a power law decay up to a day and decays exponentially beyond. Interestingly, this finding conforms to some other measures of human activity reported in previous studies. We also provide preliminary characterisation results of the semantics and topics of queries, some of which conform to that of previous studies. The results would be of general interest for understanding the dynamics of human activity and, in particular, may be leveraged for the design of mobile services.

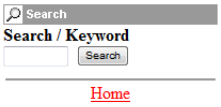

Fig. 1. Search user interface.

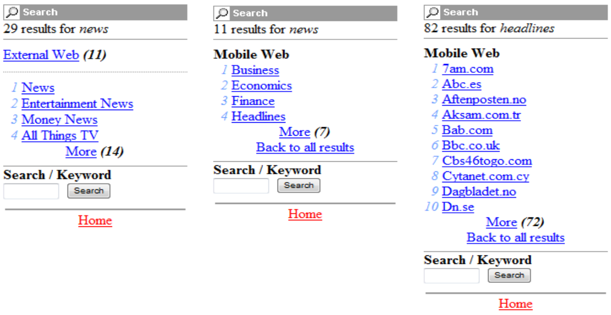

Fig. 2. An example search navigation: (a) the result set for the query "news", (b) the result set by selecting "External Web", (c) the result set by selecting "Headlines". The result sets contain categories of URLs or URLs.

I. Introduction

Search for information from mobile devices may become a mainstream daily human activity as is already an established habit with the search from desktop computers. This would be further catalyzed by continuing increase of the mobile device capabilities, turning a simple cellular phone into a sophisticated personal computing device with many models of smartphones and PDAs already available in the market. It is important to understand the patterns of mobile user behaviour as this may be leveraged in the design of the search service and other applications. To this date, it seems that there has been only a few reported studies on the mobile search usage, see [11], [10], [2]. The various aspects of human dynamics have been a subject of intense research, e.g. in the usage of network services [16], [5], [6], [3], [9], [8], [13] and the human mobility [4], [12]. Understanding of human habits is of general interest and, in particular, may be leveraged for the design of network services.

In this paper, we analyze mobile search data covering queries from about 800,000 distinct users from a major US mobile service provider with about 50M total number of mobile service subscribers. Our goal is to investigate the temporal dynamics of user searches for information. We also analyze the semantics and topics of queries, which follow the previous work [11], [10], [2]. To the best of our knowledge, analysis of the temporal dynamics of mobile searches has not been reported yet. Current mobile search is performed largely by using either mobile versions of the search engines or carrier-supplied search functionalities. Unlike to previous studies [11], [10], [2] that considered queries to mobile versions of search engines, our data is from a federated search service that combines the carrier search results with that of a major search engine provider. The access to the search user interface is available from web portals of the carrier that contain the search query user interface (Fig. 1). The results are hyperlinks to URLs or categories thereof (see Fig. 2 for an example).

We summarise our main findings as follows. We find that human-initiated searches from mobile devices appear bursty in time with the distribution of the query inter arrival time exhibiting the following dichotomy: it approximately follows a power law decay up to a day and decays exponentially beyond. We show data for particular users suggesting that the dichotomy characterises the behaviour of individuals and may not only be an artifact of the aggregation of samples over a population of users. Same dichotomy was already found to characterise the inter contact time of human-carried and vehicle-mounted mobile devices and the return time of a mobile device to a stationary site [12], hence suggesting that this may be a result of some general nature of human activity. We find that a large portion of the observed users ($> 80\%$) issue a query in only one day of the week. A small portion of the observed users ($< 1\%$) issue queries each day of the week. In general, there is a high variability of the queries issued per user over the week. We also observe that the aggregate volume of queries over a day does not significantly differ for a week and a weekend day, which perhaps non-surprisingly, suggest that a small portion of mobile searches are work related. Our analysis suggests that more than half of mobile search sessions consist of one query only and the distribution approximately follows a power law in the range from ten to hundred queries. Our analysis of query semantics and topics parallels that of [11], [10]. In conformance to the previous studies, we find that a large portion of queries are of adult category. In contrast, we find the mean number of words per query to be significantly smaller (1.5 vs. 2.5 [11], [10]). This may be a result of using the lists of top popular queries that are provided by the service.

II. Data

Our data consists of IIS logs of the mobile search service over a week period, from Apr 01, 2007 (Sun, 4:00) to Apr 08, 2007 (Sun, 4:00). The logs are standard IIS with fields including the timestamp, URI, user identifier, query, HTTP status, etc (see [15] for details). We consider a subset of data log

TABLE I
Per day statistics.

Day	1	2	3	4	5	6	7
log records	734,100	702,015	695,306	716,686	706,365	698,449	655,793
with userid (%)	78.69	78.96	78.56	78.50	78.33	78.70	77.46
with query (%)	52.15	52.77	53.78	52.89	54.40	53.72	51.79
distinct userids	188,975	184,541	183,873	186,900	185,881	188,298	178,718

records with non-empty user identifiers, i.e. carrier subscribers. In total, the part of the data that we consider contains 498,872 distinct queries, 278,595 distinct query words, and 865,726 distinct user identifiers. We provide data per day summary statistics in Table I.

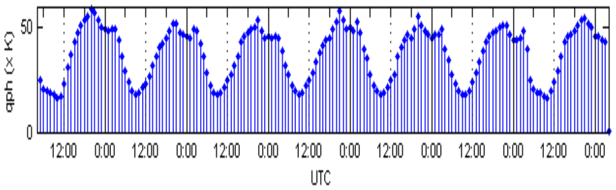

Fig. 3. Queries per hour.

III. Results

We first present our analysis of the temporal dynamics of mobile searches and then of query semantics and topics.

A. *Temporal dynamics of mobile searches*

We consider the aggregate query volume over one hour time slots. From Fig. 3, we observe that the aggregate queries exhibit strong diurnal periodicity, which is indeed to be expected as queries are issued by humans. Within each day, the activity phase is centred on the peak usage at about 20:00 UTC (12:00 PST or 15:00 EST). It is interesting to observe that there is no significant difference in the volume of searches for week and weekend days.

TABLE II
Number of days a user submitted at least one query.

Days	1	2	3	4	5	6	7
Users (%)	81.78	11.73	3.61	1.42	0.85	0.34	0.21

We next observe that the number of queries per user varies widely. The number of queries per user in the week versus the user query frequency rank (Fig. 4) approximately follows a power law for the top 10,000 of users with respect to the number of generated queries (about 1% of the total number of users). A small portion of users generate many queries and a large portion of users generate a few. It is of interest to understand how individual users generate queries over time. Do users typically issue queries daily or sporadically over some larger timescale? To address this question, we consider the histogram of the number of days that individual users issue at least one query (here a day is a calendar day). See Table II. The results suggest that most of users issue queries in one day of a week and the portion of users issuing at least one query in n days of a week is decreasing with n. A small portion of users issued a query in each day of the week (about 0.2% i.e. 1,730 users). This suggests that for most users, mobile search is not a daily activity yet.

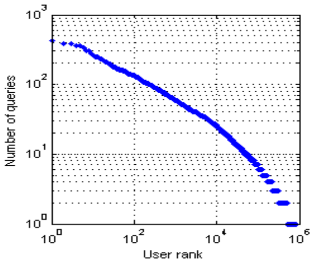

Fig. 4. Number of queries per user.

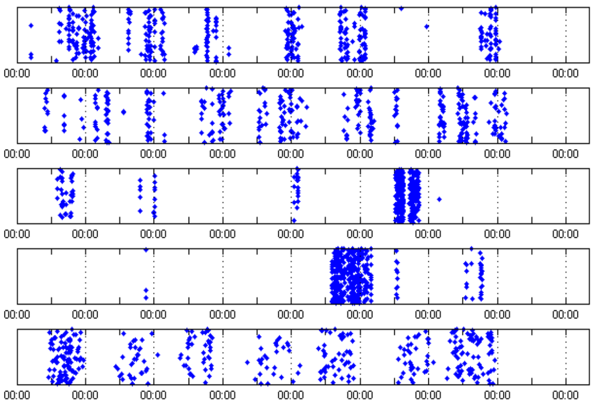

Fig. 5. Query times of users with most queries.

We now consider user inter query time. In Fig. 5, we show samples of query times for a set of users that generated most queries. These samples demonstrate that the user query activity appears in bursts over time, which seems to be governed by diurnal periodicity and bursts over shorter timescale. In Fig. 6–left, we show the complementary cumulative distribution function (CCDF) of the samples of the query inter times aggregated over 2,000 users who generated most queries. This distribution appears to approximately follow a power law for the inter query time of the duration up to about a day and then decays faster onwards. Re-plotting the same graph with

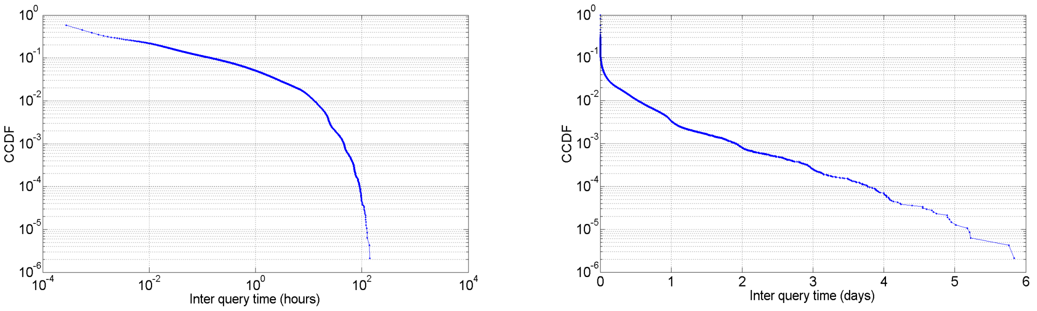

Fig. 6. CCDF of inter query time: (Left) log-log scale, (Right) lin-log scale.

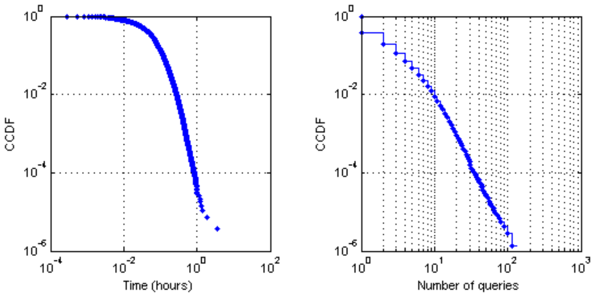

Fig. 7. CCDF of the session activity time: (Left) in real time, (Right) in the number of queries.

x axis in linear scale (Fig. 6–right) confirms that beyond the one day horizon, the distribution is exponentially bounded. These qualitative results are reminiscent of those found to characterise the inter contact time between mobile devices and a mobile device and a stationary site [12]. On the one hand, it is interesting that these seemingly different characteristics of human activity exhibit the same qualitative feature. On the other hand, this may not be entirely surprising as indeed both are a result of human activity.

Sessions. We partition the queries issued by each user into sessions by using the standard threshold heuristics–a query instance is classified as the beginning of a new session if the time elapsed since the last query by this user exceeds a threshold. We set the threshold to 5 min, following the previous studies ([10] and references therein). We consider the time between the first query of a session and the last query of this session, for sessions with at least 2 queries (we call this ”session activity time”). In Fig. 7–left, we show the CCDF of the session activity time with samples from the 2,000 users with most queries. The median session activity time is about 1.2 min with 90% of samples shorter than 5 min. The data suggests a power law decay for the distribution of the session activity time in the range quarter to one hour. We further consider the session activity time but measured in the number of queries (Fig. 7–right). We find that 61.87% of the query sessions are made of one query only. The median number of queries per session is 1 with 90% of sessions containing $\leq$ 3 queries. The decay of the CCDF approximately follows a power law in the range of ten to hundred queries. We finally consider the time per query within a session. Fig. 8 suggests that for most of the queries the time per query is about 1 min. This conforms to previously reported results on the time to enter a query [10].

Individual vs. population. We have observed that the CCDF of the query inter time aggregated over a population of users features a power law decay for values up to a day. We also observed that the data does not support the hypothesis that the query inter arrivals are statistically identical for distinct users. It is legitimate to ask whether the observed power law decay is an invariant property characterising the search activity of a single user or is a result of the aggregation of the samples over a population of users. It is well known that a power law distribution may result from the aggregation of exponentially distributed random variables with appropriately defined means (e.g. power law distributed [7]). In Fig. 9, we show the CCDFs for the session inter time for a sample of individual users. The results suggest that the observed power law for the query inter time may be a feature characterising individual human activity.

B. *Semantics and topics of queries*

In this section, we consider semantics and topics of queries. We find that 72.75% of users issued queries that are all distinct. The median number of distinct queries per query over users is 1 and the mean is 0.89. This is inline with desktop web search where unique queries are prevalent.

Semantics. We first consider the statistics for the corpus of all distinct queries, not accounting for the frequency of the query searches. We find that the mean number of words per query and characters per query are 2.11 and 13.81, respectively. The corresponding quantities obtained per query submitted to the system, are 1.52 and 9.34, respectively. In either case, the queries consisting of one word are most frequent, which conforms to [11]. The difference between the frequencies of one word and two words queries is more emphasized in the distribution obtained by weighting the queries with respect to their popularity (Fig. 10). The mean number of words per query 1.52 is significantly smaller than reported in earlier studies (2.56 [10]), which may be an artifact of the usage of the lists of top popular queries that are provided by the system (this requires further investigation).

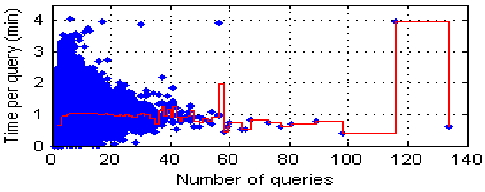

Fig. 8. Time per query within a query session.

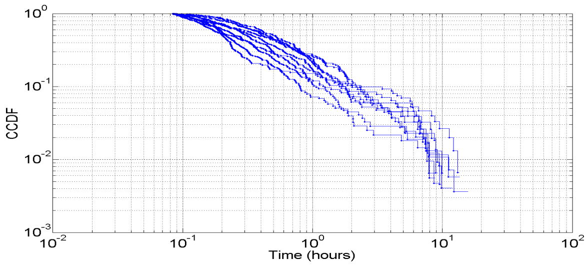

Fig. 9. CCDF of the session inter time for users with largest session frequency.

Topics. We next verify some previously reported statistics [11] on users' interests on our data.[1] In particular, we find that 15.76% of unique queries belong to adult category with 18.74% of all searches being queries of adult category. The latter is close to the reported $>$ 20% [11] for cellular phone queries. We performed similar analysis to evaluate how frequently users use URLs as queries (using the search service as a bookmark service). We find that 28.43% of distinct queries are URLs, which appears large, but we find that substantially smaller portion of all query searches (7.54%) are classified as URLs. This number is substantially smaller than the reported 17% in [11]. We further analyse user's interests over content by conducting the following simple analysis. For each user, we identify the most frequently used queries by this user (we also have done this for query words). We then use such identified summaries of users' interests to classify the users interests. We find that the most frequent queries of 24.23% of users contain a query of adult category. Same analysis but performed on the set of most frequent query words yields 21.67% of users.

IV. Concluding Remarks

We found that human searches for information from mobile devices feature the distribution of the query inter arrival time that follows a power law up to a day and decays exponentially beyond. Similar ("non Poisson") characterisation has been previously reported for other measures of human activity, e.g. email response time [3], [8] and the inter contact time between mobile devices and a mobile device and a stationary site [12]. Future work may consider generative models that would explain the causes of the observed human dynamics.

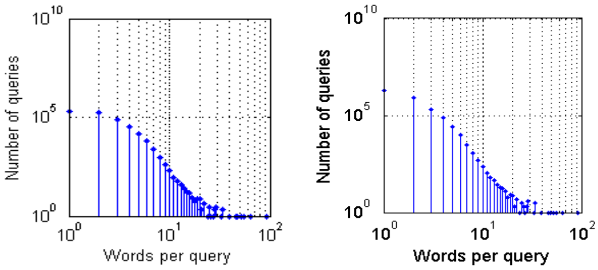

Fig. 10. Words per query: (Left) w/o query popularity weighting, (Right) w query popularity weighting.

This question appears to remain open as the previously proposed models ([3]; see also analyzed in [1]) seem to lack full empirical validation. Indeed, there exist many generative models for power laws (see survey [14]) but a question is which model is best supported by data. Analysis of query semantics and topics and the user interaction with the system deserves a further study.

Acknowledgment

I would like to thank Kaveh Goudarzian and Emmanuel Marot from Microsoft Live Search for providing me with access to data and very useful feedback.

References

[1] A. Vásquez. Exact results for the Barabási model of human dynamics. *Physical Review Letters*, PRL 95:248701–1:4, December 2005.

[2] R. Baeza-Yates, G. Dupret, and J. Velasco. A study of mobile search queries in Japan. In *Proc. of the WWW2007*, May 2007.

[3] A.-L. Barabási. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, May 2005.

[4] A. Chaintreau, P. Hui, J. Crowcroft, C. Diot, R. Scott, and J. Scott. Impact of human mobility on the design of opportunistic forwarding algorithms. In *Proc. of the IEEE Infocom '06*, pages 606–620, 2006.

[5] M. E. Crowella and A. Bestavros. Self-similarity in world wide web traffic: Evidence and possible causes. *IEEE/ACM Trans. on Networking*, 5(6):835–845, December 1997.

[6] C. Dewes, A. Wichmann, and A. Feldmann. An analysis of internet chat systems. In *Proc. of the ACM IMC '03*, October 2003.

[7] S. Jin and A. Bestavros. Sources and characteristics of web temporal locality. In *Proc. of the MASCOTS*, 2000.

[8] A. Johansen. Probing human response times. *Physica A*, 338:286–291, 2004.

[9] A. Johansen. Comment on A.-L. Barabasi, nature 435 207-211 (2005), Feb 4 2006. Preprint, arXiv:physics/0602029v1 [physics.soc-ph].

[10] M. Kamvar and S. Baluja. A large scale study of wireless search behavior: Google mobile search. In *CHI 2006*, April 2006.

[11] M. Kamvar and S. Baluja. Deciphering trends in mobile search. *IEEE Computer Magazine*, 40:58–62, August 2007.

[12] T. Karagiannis, J.-Y. Le Boudec, and M. Vojnović. Power law and exponential decay of inter contact times between mobile devices. In *Proc. of the ACM Mobicom '07*, pages 183–194, 2007.

[13] J. Leskovec and E. Horvitz. Worldwide buzz: Planetary-scale views on an instant-messaging network. Technical Report MSR-TR-2006-186, Microsoft Research, June 2007.

[14] M. Mitzenmaher. A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*, 1(2):226–251, 2003.

[15] Microsoft TechNet. W3c extended log file format (iis 6.0), 2007. http://www.microsoft.com/technet/prodtechnol/WindowsServer2003/-Library/IIS/676400bc-8969-4aa7-851a-9319490a9bbb.mspx.

[16] W. Willinger, M. S. Taqqu, R. Sherman, and D. V. Wilson. Self-similarity through high-variability: Statistical analysis of ethernet lan traffic at the source level (extended version). *IEEE/ACM Trans. on Networking*, 5(1):71–86, February 1997.

[1] We classify a query to adult category if it contains a string from a dictionary. Same for the classification to URL category where the dictionary is ".", "www", "http".