A Study of Mobile Search Queries in Japan

Ricardo Baeza-Yates, Georges Dupret, Javier Velasco
Yahoo! Research Latin America
Santiago, Chile

ABSTRACT

In this paper we study the characteristics of search queries on mobile phones in Japan, comparing them with previous results of generic search queries in Japan and mobile search queries in the USA. We confirm some results while find some interesting differences on the query distribution, use of the different script languages and query topics.

1. INTRODUCTION

At the end of 2006 the number of mobile subscribers in the world reached 2.7 billion (ITU), with almost 40% concentrated in the Far East. Today, China has surpassed the 400 million, duplicating the number in the USA (just above 200 million). Behind, India and Russia have surpassed the 120 million and Brazil and Japan the 100 million barrier. Market predictions put the total number of mobiles for 2010 over 4 billion. For that time, mobiles might become the most common device to access the Internet. For this reason mobile search is an important potential advertising business and it is important to know what people search for and what should be the best interface (see for example [3]). Hence, to exploit this potential, we need to understand the mobile search usage in the world. Last year, in [5], a study of mobile search usage of one million USA queries was presented.

In this paper we present a study of mobile search usage in Japan, where the mobile Web has a high level of penetration, perhaps the largest in the world. Our study is one order of magnitude larger than the study mentioned in the previous paragraph [5], and can serve as a basis for future work on other countries and/or languages. We also compare our results with the query log used by Jones *et al.* [4] for the automatic generation of related queries in Japanese.

In the next section we study the characteristics of the sample query log used (from 2006). In the third section we do a topic classification, a difficult task when there is more than one script. In fact, in Japan normally four scripts are used: Kanji, ideograms of Chinese origin indispensable in written Japanese, where each symbols represents a concept; Hiragana and Katakana, phonetics syllabaries of 48 letters; and Romaji, the normal Latin alphabet. We end with some final remarks.

WWW2007, May 8–12, 2007, Banff, Canada.

2. QUERY LOG CHARACTERISTICS

We use two samples from Yahoo! Japan query logs of 2006, using one million mobile and one hundred thousand desktop *unique* queries. That is, we collected the frequency of one million and one hundred thousand different queries in each case. Our sample is hence much larger than the previous mobile search study [5].

We start with an analysis of the top-level statistics, by comparing the summaries of characters and tokens for Desktop and Mobile queries (Table 1). The identification of the number of characters is not particularily difficult but the extraction of tokens is complicated by the fact that Japanese language does not require spaces between ideograms belonging to different semantical units. For this, we used Chasen, an open source Japanese morphological analysis system [1] to tokenize the queries.

Although the mean number of tokens (terms) is similar for mobiles (2.29) and desktop (2.25), when we look at the mean number of characters used, mobile queries (7.9) are shorter than desktop queries (9.6).

This could be attributed to differences in the physical interface for text input, where is more difficult to type, and could suggest that people tend to use more specific vocabulary in their mobile queries, as people do not want to have to refine a query to obtain the desired answer.

Our results on token length for queries are consistent with the results obtained by Kamvar and Baluja [5]: 2.3 in XHTML devices, 2.3 for desktops and 2.7 for PDAs. However, if we compare the number of characters from such study (15.5 in XHTML, 17.5 for PDAs), there's an important difference: our queries are much shorter. This can be attributed to the language difference, in Japan, people choose from four scripting languages to input their queries, some of which are ideogram-based, thus allowing to express comparable ideas with much shorter words. This is consistent with the analysis of almost 100 million Japanese queries by Jones *et al.* [4], where the average number of tokens is around 2.5. In fact, in that study Kanji queries have average length 2.5, Katakana 2.7, Romaji 1.9 and Hiragana 0.6.

2.1 Queries

As shown in Figure 1 and Table 1, most unique queries are composed of two terms, being similar in mobile and desktop queries alike, and is consistent with previous studies on the subject on other languages [1].

Single term queries are important when queries are weighted by frequency, and in the case of desktop queries, these ex-

[1] `http://chasen.naist.jp/hiki/ChaSen/`

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Mobile queries						
Number of tokens	0	2	2	2.293	3	183
Number of chars	1	5	7	7.928	10	999
Desktop queries						
Number of tokens	0	2	2	2.249	3	9
Number of chars	1	7	9	9.623	12	44

Table 1: Query Log Summary.

ceed those with two terms.

It is interesting to note that two-term queries are more frequent than one-term queries on mobile devices, considering the fact that this usually requires a larger input effort by the user. Again the main hypothesis is that users prefer to be more precise to avoid refining the query and thus having to input more text.

With respect to character length in mobile and desktop (Figure 2) queries, we can see that desktop queries can be much longer than mobile queries, the bulk of these queries are concentrated between 5 and 10 characters in mobile devices and between 7 and 12 on desktop computers.

In terms of character length, desktop queries tend to be longer than mobile queries. This behavior is easily predictable considering the ease of use on full keyboards allows people to write more complex queries. As shown in table 1, mobile queries have a mean of 7.9 versus 9.6 for desktop queries, and the medians are 7 and 9.

When we analyze the frequency distribution for queries on mobiles and desktops (Figure 3 at the top), we can notice a power-law distribution for mobile queries. However, there is a sudden drop on our data for desktop queries, due to the size of the sample.

2.2 Terms

Frequency distribution for terms on mobiles and desktops (Figure 3 at the bottom) is similar to the analysis of query distribution, denoting a power law in the case of mobile devices, and a sudden drop for the desktop queries, again due to the sample size.

2.3 Characters

2.3.1 Distinct Queries

When looking at the ratio given by the percentage of unique vs mixed queries for each script (written language), the script with the highest ratio of exclusivity will imply a more specific use of such (Tables 2, 3, 4, 5). We can notice a difference in this factor across both platforms: Romaji is the most specifically used script on mobile devices, while the highest ratio on desktop computers is found for Kanji.

A possible explanation is a trade-off between complexity of input and semantic content. Kanji is more complex than Hiragana, Katakana or Romaji to input because there are more characters, around 2000 in common use. Text input on Japanese systems involve selecting the desired script first, later typing in the pronunciation, to later select from a list of available ideograms.

Regarding semantic content, Kanjis directly represent a meaning and can have various pronunciations. On the other hand, Hiragana and Katakana are phonetic scripts, also used in written Japanese, that are used either to transcript phonetically foreign words or to represent grammatical words that have no meaning per se. Because Hiragana and Katakana are phonetic scripts, they can be used either during the input process of a Kanji symbol or to represent it phonetically, a little bit like we could write Japanese using Roman letters by reproducing how it sounds. Finally, Romaji are the Roman letters as used by the English language. They are rarely used to transcript phonetically Kanji, although they can and maybe they are; we should study this in more details. Usually Romaji have a "fashion" aspect and can be used for brand names. In some cases, a Japanese brand can only have a Romaji name, like "Sony" or "Bathing Ape" that don't have a Japanese equivalent. Also, Katakana, Hiragana and Romaji have approximately the same level of inputting complexity.

2.3.2 Repeated Queries

Observing script usage behavior for frequency weighted queries in mobile and desktop (Tables 4 , 5), the data suggests that Kanji suffers a drop when changing from desktop to mobile devices, this drop in use of Kanji seems to be replaced by similar ideas expressed in Hiragana or Romaji. This holds especially true for queries with only one script. This drop in Kanji use is somewhat lower in the case of distinct queries (Tables 2, 3).

Looking at the ratio of script used exclusively in a query out of the total present in queries (Line 5 in these tables), we see in the case of frequency weighted queries that there's an increase for Romaji when migrating from desktop (52) to mobiles (71), Hiragana has the same behavior only lower (18, 25), a drop in Kanji (63, 54), while Katakana remains quite stable (53, 52).

Compared to Jones *et al.* [4], there are clear differences in the use of Hiragana and Romaji. As they used Japanese queries used worldwide, Romaji could be higher due to the use of English-based keyboards. However, we do not have a reasonable explanation for the Hiragana difference, but could be partly due to queries by non-Japanese people.

3. QUERY TOPIC

In order to analyze queries by topical categories, we took the taxonomy found on DMOZ Open Directory[2] and analyzed the language present on the linked pages to create a standard binomial language model [2] for each category. We also create a language model from the whole set of documents to smooth the class language models. We then classify a query to the class which generated it with the highest probability. To test the model, we selected 100 queries in each class and asked a translater to check whether the classification was acceptable. The results obtained where good enough (over 90%) to justify the use of this crude model,

[2]See `http://www.dmoz.org/`.

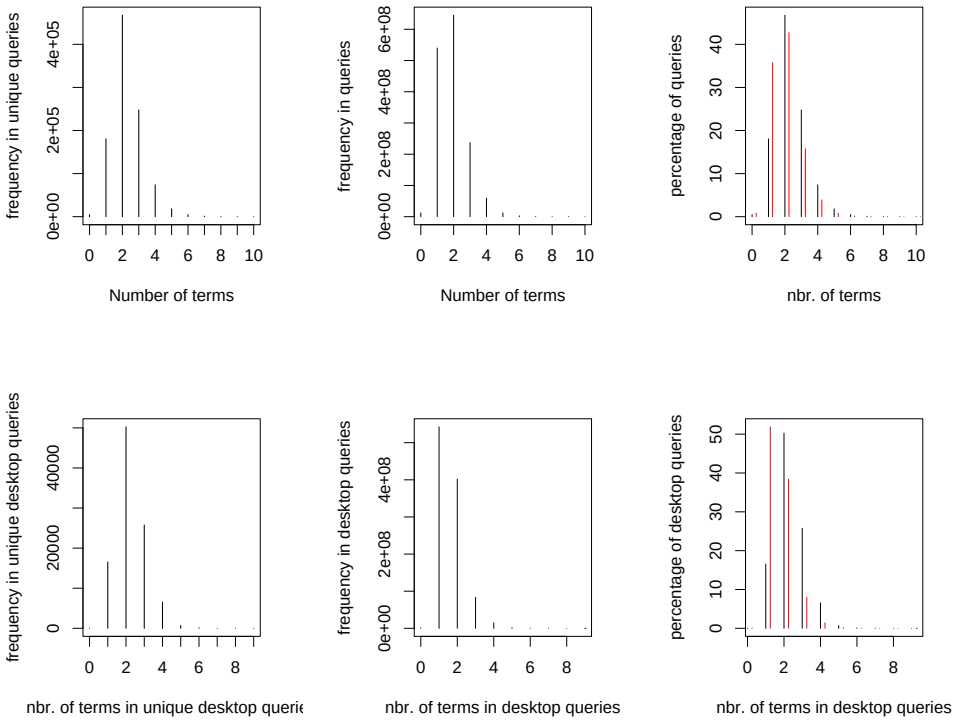

Figure 1: Number of terms per mobile (top) and desktop (bottom) query. The first column displays frequencies for unique queries, second column includes repeat queries and the third column superimposes both previous graphs with a slight shift for visualization purposes. Graphs on the third column use percentage scale in the Y-axis for comparison purposes.

	Katakana	Hiragana	Kanji	Romaji
(1) Query Contains	465,550	132,783	662,981	165,305
(2) Percentage	46.6	13.3	66.3	16.5
(3) Query contains only	174,823	19,842	312,220	94,741
(4) Percentage	17.5	2	31.2	9.5
(5) (3) / (1)	37.6	15	47.1	57.3

Table 2: When mobile queries are counted once (1,000,000 distinct queries).

	Katakana	Hiragana	Kanji	Romaji
(1) Query Contains	52,075	8,716	72,729	10,466
(2) Percentage	52.1	8.7	72.6	10.5
(3) Query contains only	18,629	1,013	34,661	4,472
(4) Percentage	18.6	1.0	34.6	4.5
(5) (3) / (1)	35.8	11.6	47.7	42.7

Table 3: When desktop queries are counted once (100,000 distinct queries).

	Katakana	Hiragana	Kanji	Romaji
(1) Query Contains	676,447,242	175,162,721	864,690,883	262,751,332
(2) Percentage	44.8	11.6	57.3	17.4
(3) Query contains only	351,598,815	43,761,189	468,753,904	187,897,193
(4) Percentage	23.3	2.9	31	12.4
(5) (3) / (1)	52	25	54.2	71.5

Table 4: When mobile queries are weighted by their frequency.

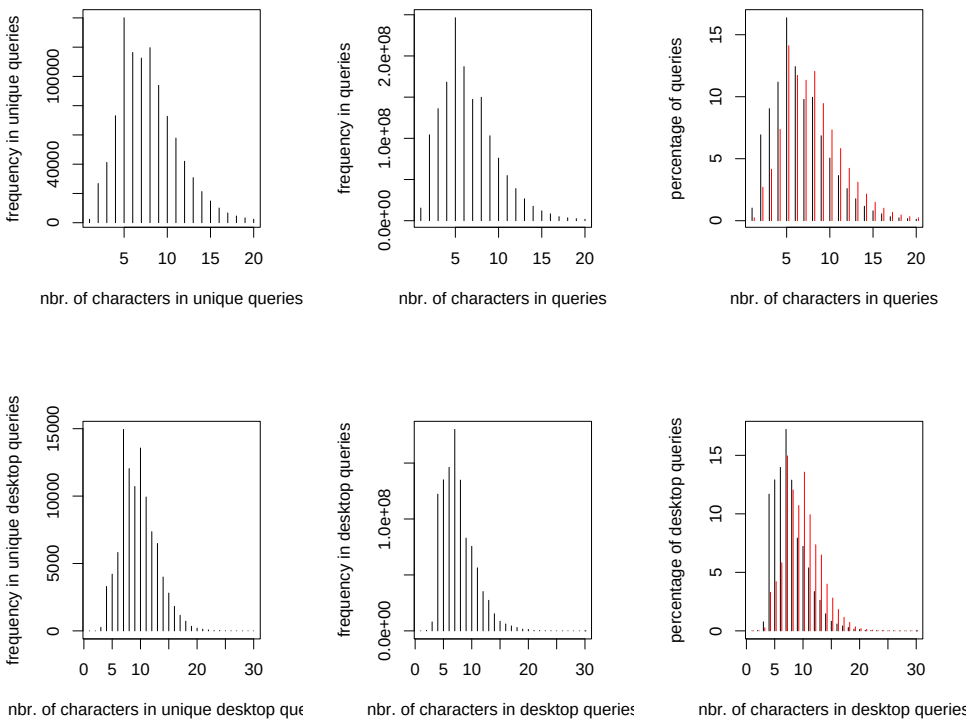

Figure 2: Number of characters per mobile (top) and desktop (bottom) queries. The first column displays frequencies for unique queries, second column includes repeat queries and the third column superimposes both previous graphs with a slight shift for visualization purposes. Graphs on the third column use percentage scale in the Y-axis for comparison purposes.

	Katakana	Hiragana	Kanji	Romaji
(1) Query Contains	1,945,799,846	247,045,743	2,830,690,148	387,532,243
(2) Percentage	46.2	5.9	67.2	9.2
According to [4]	45.7	18.7	63.1	22.6
(3) Query contains only	1,032,269,352	45,122,660	1,778,584,491	201,712,005
(4) Percentage	24.5	1.07	42.2	4.8
According to [4]	>8.7	>0	>17	> 8.6
(5) (3) / (1)	53.0	18.3	62.8	52.1

Table 5: When desktop queries are weighted by their frequency.

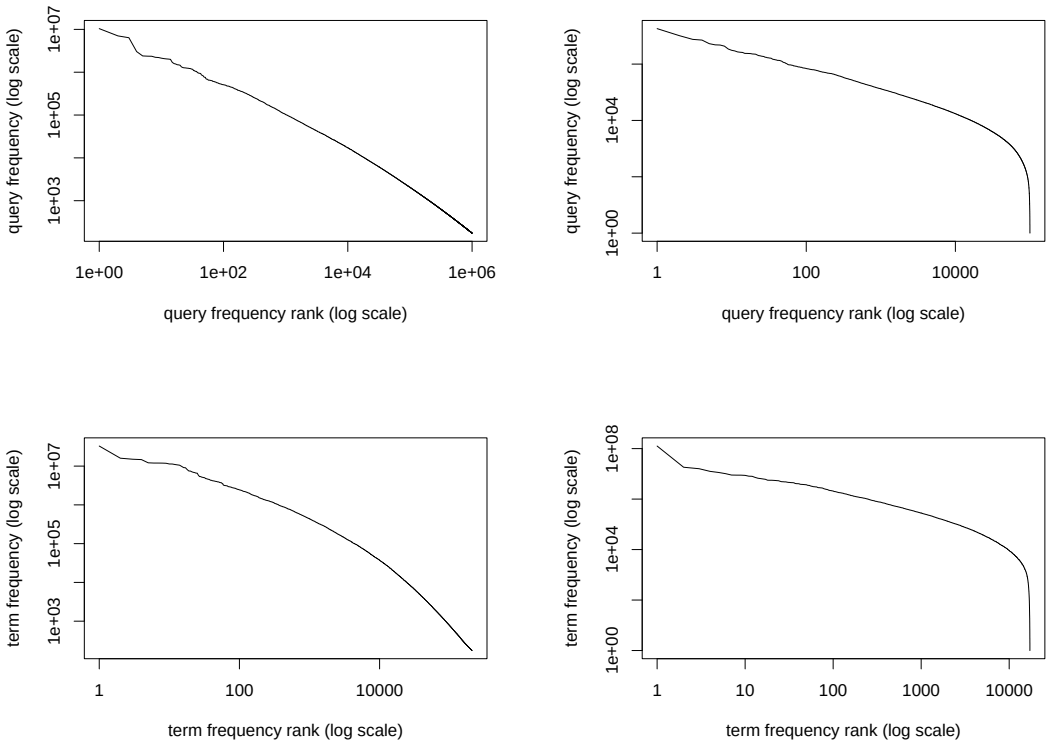

Figure 3: Mobile (left) and Desktop (right) query (top) and terms (bottom) frequency distribution.

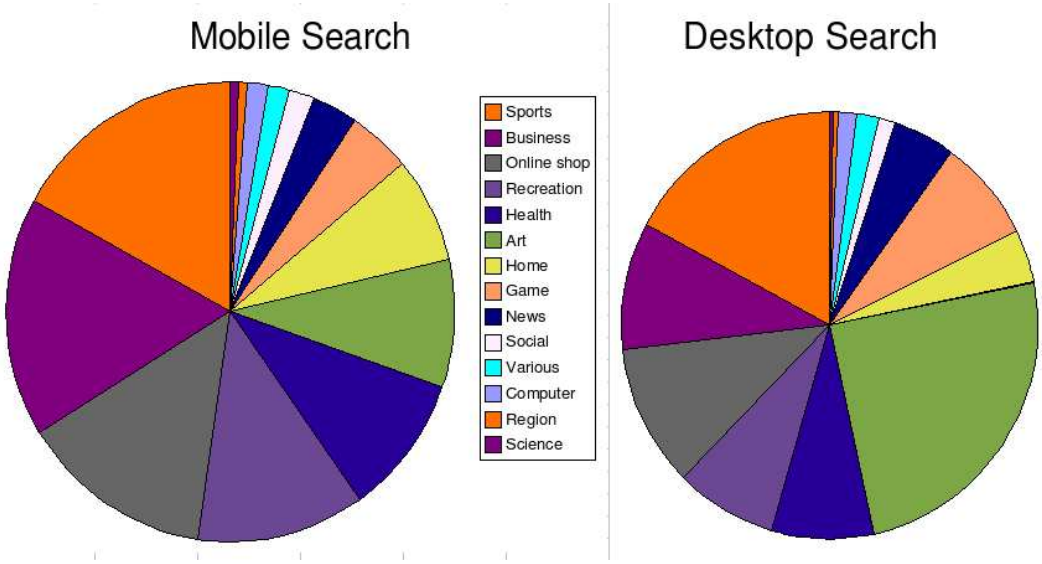

Figure 4: Mobile and Desktop volume according to categories.

Category	Mobile	Desktop	Google	Category[5]
Business*	2.0	0.6	<2	Business
Business*	0.03	0.01	<2	Food & Drink
Business*	0.02	0.01	<2	Shopping & Consumer services
Games	4.6	8.0	>2	Games
Health	10.0	7.7	>2	Health & Beauty
Online shop	14.0	10.9	> 5	Internet & Telecom
Recreation*	5.6	3.6	>2	Travel & Recreation
Recreation*	0.3	0.1	<2	Automotive
Science	0.5	0.2	<2	Science
Sports	17.1	17.2	>2	Sports
Art	8.8	24.8	< 2	Arts & Literature
Computer	1.5	1.4	>2	Computers & Technology
Home	7.6	4.1	<2	Home & Garden
News	3.3	4.8	<2	News & Current Events
Recreation*	5.8	4.1	>10	Entertainment
Social	1.8	1.3	>2	Society

Table 6: Comparison with USA mobile search study (* = subcategories were used).

although it is clear that a better method should be designed for more fine grained categorization. Pie charts (Figure 4) for query volume distributed by topic for mobile and desktop queries show much similarity in general terms. The largest differences are seen on the Art and Business categories. We also compared our categorization with the one in [5], although different and imprecise (only bounds on the percentage of each class are given), matching categories or subcategories. In addition they are divided in XHTML and PDA cases, without giving the relative proportion of each case, so we use the XHTML case for the comparison. In Table 6 we give the classes where our results agree in most cases (top) and disagree (bottom). The differences are probably cultural between Japan and the U.S.A. Other eight categories are omitted as there is not enough information to do the comparison or cannot be matched.

4. FINAL REMARKS

Our analysis is by not means complete, as contains some noise on the sample and the categorization process can be improved (see for example [6, 7]). In particular queries are not uniformly distributed on the DMOZ taxonomy, and a more adequate categorization should be used, like in [5], adding for example an Adult category. Hence, we are currently improving the classification and studying other characteristics of the query logs.

Acknowledgements

We would like to thanks the help of Bill Michels and Dash Gopinath from Yahoo! Search Marketing International in obtaining the data used for this study.

5. REFERENCES

[1] Ricardo Baeza-Yates. Query Usage Mining in Search Engines. In *Web Mining: Applications and Techniques*, Anthony Scime, editor. Idea Group, 2004, 307–321.

[2] W. Bruce Croft and John Lafferty, editors. Language Modeling for Information Retrieval, Kluwer Publishers, 2004.

[3] Karen Church, Mark T. Keane1 and Barry Smyth. An Evaluation of Gisting in Mobile Search. In Advances in Information Retrieval, editors D. Losada and J.M. Fernández-Luna. Lecture Notes in Computer Science 3408, Springer, 546-548, 2005.

[4] Rosie Jones, Kevin Bartz, Pero Subasic and Benjamin Rey, "Automatically Generating Related Queries in Japanese", Language Resources and Evaluation, special issue on Asian Language Technology, 2007 (to appear).

[5] Maryam Kamvar, Shumeet Baluja. A large scale study of wireless search behavior: Google mobile search. ACM Conference on Human Factors in Computing Systems, Montral, Qubec, Canada, 701-709, 2006.

[6] Dou Shen, Rong Pan, Jian-Tao Sun, Jeffrey J. Pan, Kangheng Wu, Jie Yin, Qiang Yang. Q^2C@UST: Our winning solution to query classification in KDDCUP 2005. ACM SIGKDD Explorations Newsletter 7 (2), 100-110, Dec. 2005.

[7] Dou Shen, Jian-Tao Sun, Qiang Yang, Zheng Chen. Building bridges for web query classification Proceedings of the 29th ACM SIGIR conference on Research and development in information retrieval, 131-138, 2006.