

A Study of Mobile Search Queries in Japan

R. Baeza-Yates G. Dupret J. Velasco

Yahoo! Research Latin America

{rby|gdupret|jvelasco}@yahoo-inc.com

Query Log Analysis: Social and Technological Challenges



Outline

- 1 Context & Motivation
- 2 Characteristics of the Sample Query Log.
- 3 Scripts
- 4 Query Categories



Context & Motivations.

- An order of magnitude larger than the study by Kamvar & Baluja (1,000,000 page view requests in Google)



Context & Motivations.

- An order of magnitude larger than the study by Kamvar & Baluja (1,000,000 page view requests in Google)
- Basis for subsequent analysis for other countries or languages



Context & Motivations.

- An order of magnitude larger than the study by Kamvar & Baluja (1,000,000 page view requests in Google)
- Basis for subsequent analysis for other countries or languages
- Important potential advertising business:
 - What do people search?
 - How important is the interface?



Outline

- 1 Context & Motivation
- 2 Characteristics of the Sample Query Log.
- 3 Scripts
- 4 Query Categories



Query Log Characteristics.

- 1,000,000 *distinct* mobile queries (821,264 after normalization, 1,214,251,125 searches)
- 100,000 *distinct* desktop queries (92,001 after normalization, 896,071,277 searches)

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



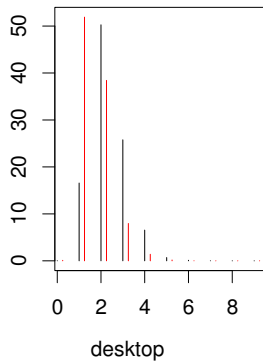
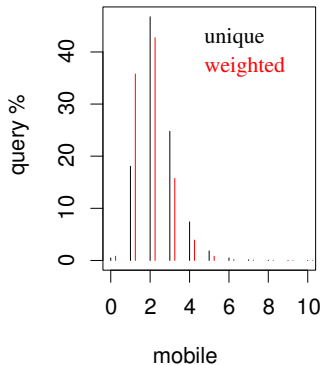
Query Log Characteristics.

- 1,000,000 *distinct* mobile queries (821,264 after normalization, 1,214,251,125 searches)
- 100,000 *distinct* desktop queries (92,001 after normalization, 896,071,277 searches)

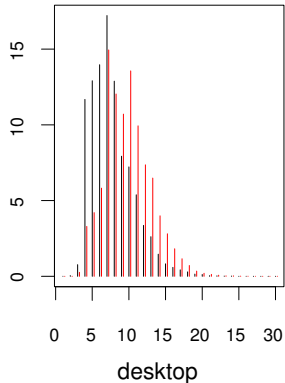
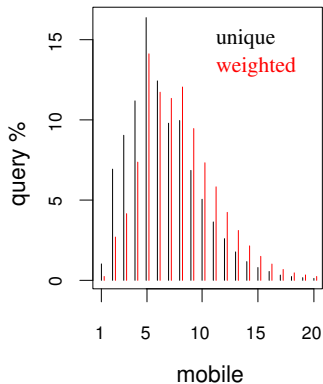
	Japan	US	
Mobile		(Kamvar & Baluja)	
# tokens	2.293	XHTML: 2.3	PDA: 2.7
# chars	7.928	XHTML: 15.5	PDA: 17.5
Desktop			
# tokens	2.249	2.3	
# chars	9.623		



Number of Terms per Queries

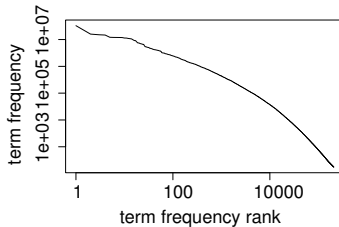
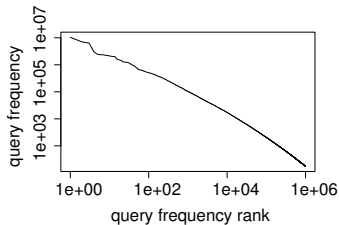


Number of Characters per Queries

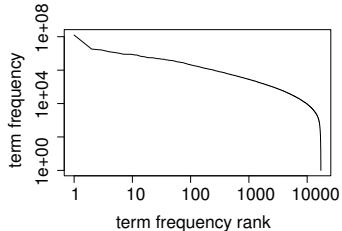
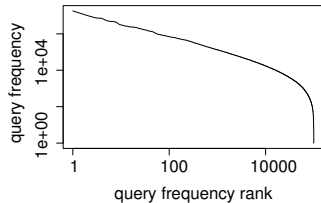


Query & Term Frequency Distributions

Mobile



Desktop



Outline

- 1 Context & Motivation
- 2 Characteristics of the Sample Query Log.
- 3 Scripts**
- 4 Query Categories



Distinct Mobile & Desktop Queries

Desktop	Katakana	Hiragana	Kanji	Romaji
Contains ¹	52,075 52.1%	8,716 8.7%	72,729 72.6%	10,466 10.5%
contains only ²	18,629 18.6%	1,013 1.0%	34,661 34.6%	4,472 4.5%
2 / 1	35.8%	11.6%	47.7%	42.7%
Mobile	Katakana	Hiragana	Kanji	Romaji
Contains ¹	465,550 46.6%	132,783 13.3%	662,981 66.3%	165,305 16.5%
contains only ²	174,823 17.5%	19,842 2%	312,220 31.2%	94,741 9.5%
2 / 1	37.6%	15%	47.1%	57.3%



Repeated Mobile & Desktop Queries

Desktop	Katakana	Hiragana	Kanji	Romaji
Contains ¹	1,946	247	2,831	388
	46.2%	5.9%	67.2%	9.2%
<i>Jones</i>	45.7%	18.7%	63.1%	22.6%
contains only ²	1,032	45	1,779	202
	24.5%	1.0%7	42.2%	4.8%
<i>Jones</i>	>8.7%	>0%	>17%	>8.6%
2/1	53.0%	18.3%	62.8%	52.1%
Mobile	Katakana	Hiragana	Kanji	Romaji
Contains ¹	676	175	864	263
	44.8%	11.6%	57.3%	17.4%
contains only ²	352	44	469	188
	23.3%	2.9%	31%	12.4%
2/1	52.0%	25.0%	54.2%	71.5%

units: million queries



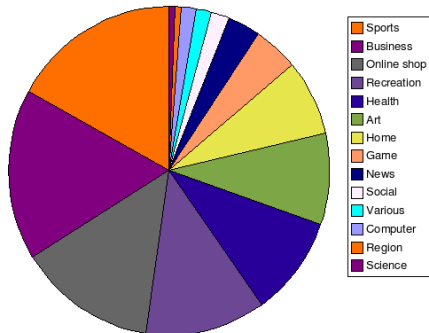
Outline

- 1 Context & Motivation
- 2 Characteristics of the Sample Query Log.
- 3 Scripts
- 4 Query Categories**

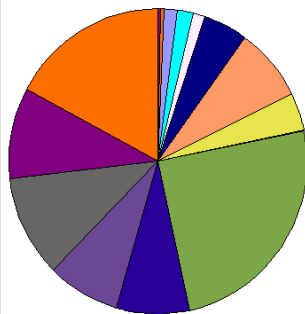


Query Topics

Mobile Search



Desktop Search



Comparison

Yahoo!	Mobile	Desktop	Google	Google Category
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



Comparison (continued)

Yahoo!	Mobile	Desktop	Google	Google Category
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



Discussion

- Improve sampling
- Improve categorization:
 - reliability
 - detail

