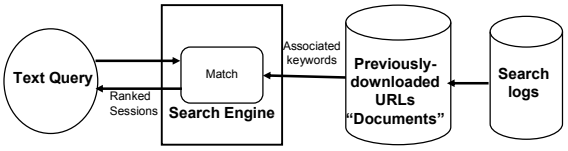
A Search Engine Based on Query Logs, and Search Log Analysis

Michael Oakes & Yan Xu
University of Sunderland, England

A Search Engine Based on Query Logs



The 5 constituent trigrams in "_KATZE_"

Tri-gram	_KA	KAT	ATZ	TZE	ZE_	Overall Prob-ability
English	1.23 E-5	4.72 E-7	4.72 E-7	3.49 E-5	1.70 E-5	1.63 E-27
French	5.00 E-7	5.00 E-7	5.00 E-6	5.00 E-7	1.01 E-5	6.31 E-32
German	5.64 E-4	1.41 E-4	1.41 E-4	1.41 E-4	1.12 E-4	1.77 E-19

Languages of the first 100,000 query lines (excluding null = 9975)

- English 29.69; Italian 13.38; German 12.27;
- French 9.56; Dutch 7.84; Spanish 7.47;
- Finnish 5.61; Portuguese 5.36; Swedish 4.69;
- Greek 0.

Proportion of Sessions Consisting of Zero, One or More than One language

- 0 languages: 4701 sessions (9.66%)
- 1 language: 42354 sessions (87.07%)
- >1 language: 1592 sessions (3.27%)

- Total sessions in the first 100,000 queries = 48647.

Sessions Grouped by Language of the Interface (%)

- EN 84.5; PL 4.14; FR 4.10; DE 2.35;
- IT 1.05; PT 0.73; LV 0.71; SL 0.70;
- HU 0.50; HR 0.21; ET 0.19; SR 0.17;
- SK 0.16; CS 0.15; NL 0.11; LT 0.08;
- EL 0.06; FI 0.03; -- 0.02; DA 0.02;
- MT 0.01

Cross-tabulation for the interface language (v) and query submitted (h)

	null	Dan	Ger	Spa	Fin	Fre	Ita	Dut	Por	Swe	Eng
Ger	380	97	360	167	84	98	203	243	115	100	535
Gre	15	0	0	4	1	2	12	0	4	12	4
Fre	561	115	320	265	266	1021	406	130	148	66	789
Ita	121	14	30	142	65	88	318	22	29	8	152
Dut	10	0	14	9	0	14	9	210	2	4	43
Por	106	11	19	222	19	36	137	64	73	5	51
Swe	4	1	26	0	5	0	6	3	23	25	26
Eng	7856	3316	9109	5496	3970	6985	10107	6014	4126	3650	23892

Conclusions

- Frequency of trigrams: sequences of 3 adjacent characters
- Most searches are in English
- Most searches are in only one language
- Usually the query language is the same as the interface language
- Latin is similar to Italian e.g. “Commentaria in Psalmos Davidicos”