How Users Search WWW.HR Web Directory

Gordan Gledec, Igor Ljubi
Department of Telecommunications
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia

6th Carnet Users Conference
Zagreb, September 27-29th, 2004.

Summary

- About WWW.HR project
  - Web directory
- Indexing submitted sites
- Searching the directory
- Analysis of user queries
- Results
- Conclusion

WWW.HR

- Web based information service supported by CARNet
  - established in 1994.
  - thematic portal, providing regional information concerning Croatia
  - two services:
    - Facts about Croatia
    - Web directory

WWW.HR Search Index

- site name in Croatian and English
- site description in Croatian and English
- site URL
- category names in Croatian and English
- META keywords extracted from the submitted page

Supported queries

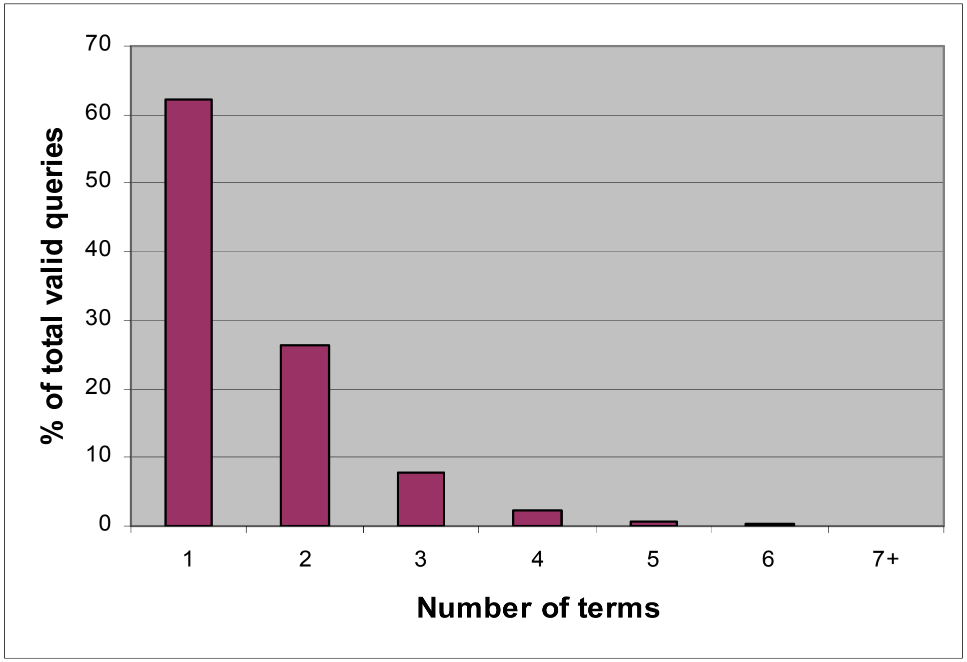
- all keywords
  - default query, eg. "zagreb live"
- logical expressions
  - eg. "+zagreb -live", "zagreb or live"
- wildcard queries
  - eg. "europe*"
- phrases
  - eg. "history of shipbuilding"

Analysis

- Monitoring period: May 6th - June 16th, 2004
  - 2,000,000 pageviews, 450,000 user sessions
  - 75,000 sessions directed to search engine
- What was analysed:
  - number of terms per query
  - use of advanced search features
  - query spelling
  - frequency of queries
  - distribution of queries in time
  - query results returned to the user

Query terms analysis

- ~185,000 total queries
- ~67,000 unique queries
- 2,42 % invalid queries
  - <3 characters
- 287,021 query terms
  - only 44,185 unique
- average length of the query:
  - 1,54 terms



Advanced search features usage

- advanced features:
  - logical expressions
  - stemming
  - phrases
- only 2390 queries (1,28%) used advanced features
  - 614 queries contained the '*' wildcard
    - 25,69% of advanced, but only 0,33% of all queries
    - mostly after the initial query yielded no results

Query spelling

- Hacheck - Croatian academic spelling checker
  - groups suspicious terms into 4 groups
  - spelling mistake in ~7% of all query terms:
    - eg: *raifassen*, *raifeissen*, *raiffaisan*, *raiffaisen*, *raiffeisenbank*, *raiffeissen*
    - support for diacritical letters, QWERTZ vs. QUERTY

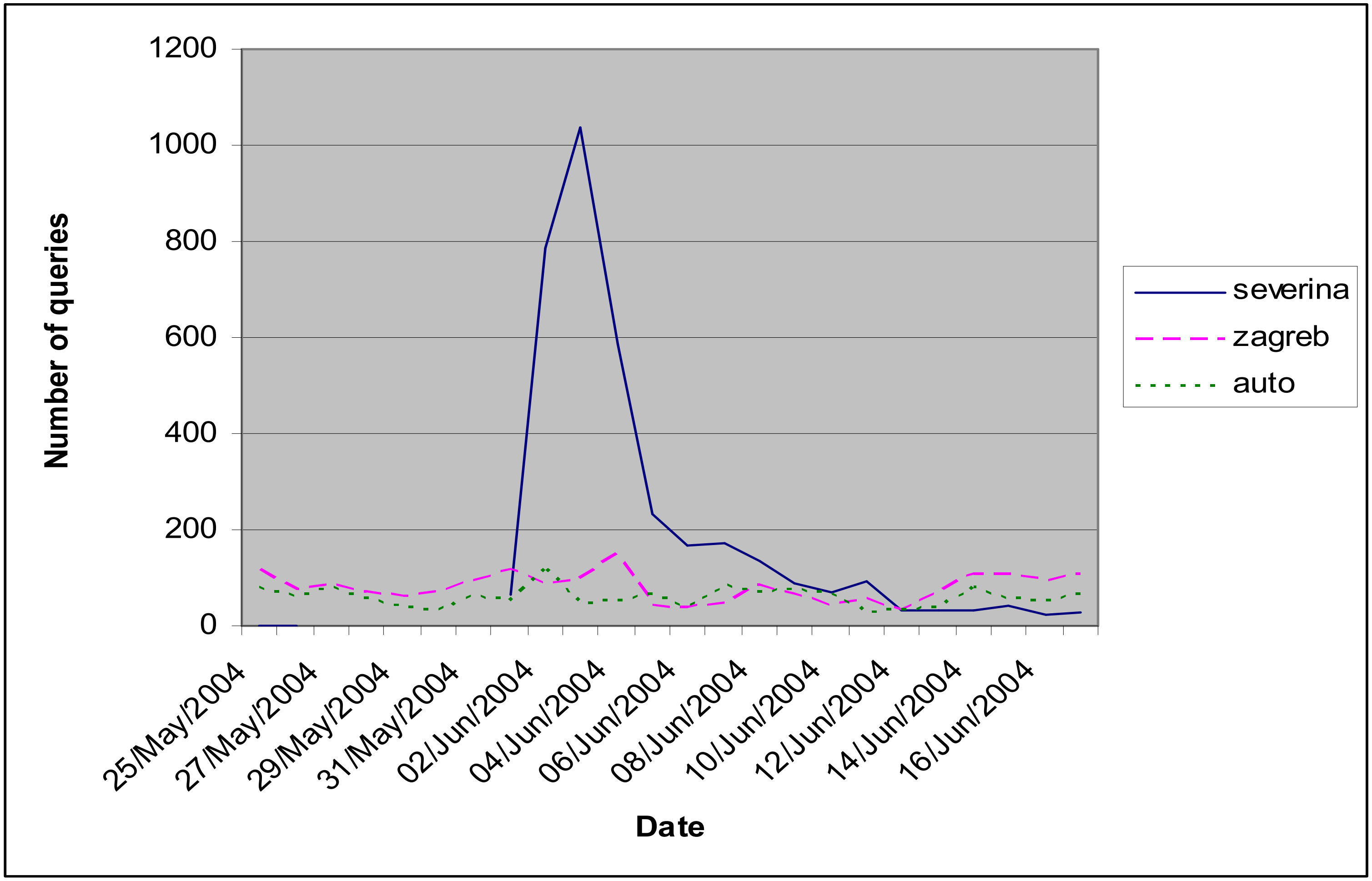
	unique	# occurances
Number of query terms	44185	287021
Extremely suspicious	4996	8722
Very suspicious	4321	9588
Moderately suspicious	4426	9797
Almost insuspicious	2697	6742
Total	16440	34849

Frequency of queries

- top 10 queries and top 10 query terms
  - 20 most frequently used terms found in 12% of all queries

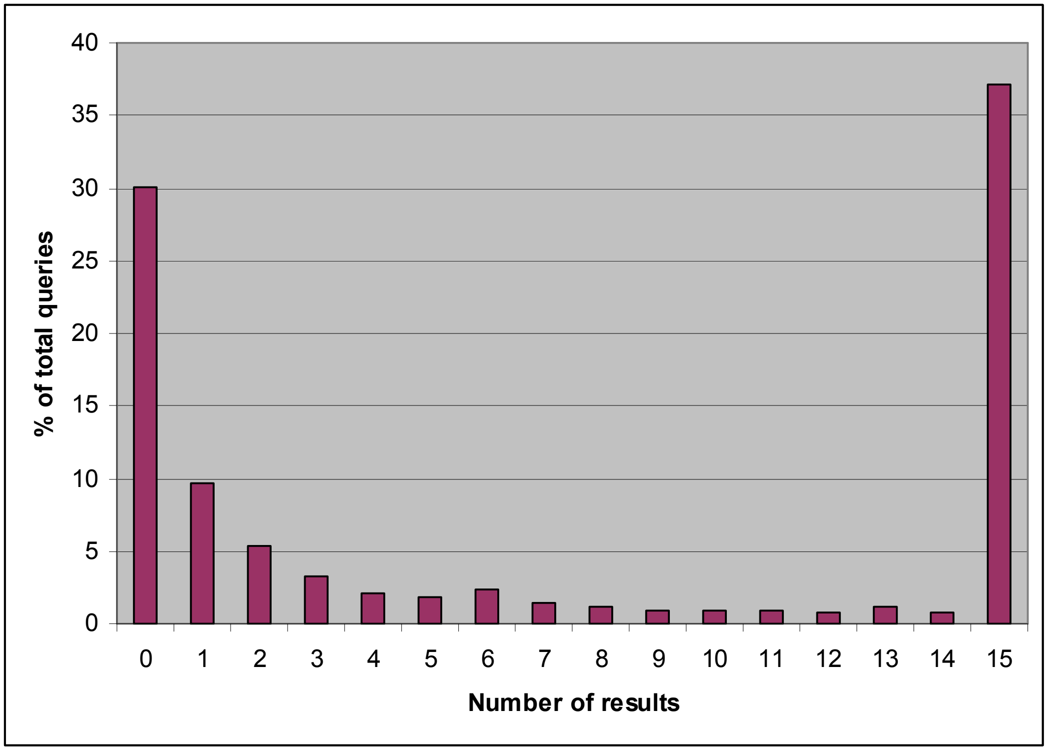
queries	#	terms	#
severina	2613	severina	3538
nekretnine	1813	zagreb	2902
sex	1680	auto	2302
auto	1319	apartmani	2289
split	1259	nekretnine	2258
zagreb	948	sex	2094
chat	886	split	2001
zadar	739	hotel	1636
telefonski imenik	707	dubrovnik	1243
apartmani	684	list	1174

Time distribution of queries



Results returned to the user

- 30% of queries returned no matching results
  - more than 35% queries returned too many results



Conclusion

- almost 300,000 user queries analysed
- most users use 1 or 2 query terms and make a fairly large amount of spelling errors
- results in compliance with similar research on other search engines
    - users fail to recognize the difference between directories and search engines which crawl the Web
- high percentage of queries that return no results asks for new search mechanisms
    - ontologies, spelling suggestions...

Thank you for your attention.

... and don’t forget to visit Croatian Homepage

at www.hr ☺