

How Users Search WWW.HR Web Directory?

Gordan Gledec, Igor Ljubi
Department of Telecommunications
Faculty of Electrical Engineering and Computing, Zagreb, Croatia
E-mail: gordan.gledec@fer.hr, igor.ljubi@fer.hr

1. INTRODUCTION

Two dominant ways of finding information on the Web are through the use of Web search engines and Web directories. As the Web became the major source of information for many users, the way users search the Web became a crucial issue. Studies on the Web search appear regularly [1], reporting user search patterns and effectiveness of the search engines. This paper analyses how users search WWW.HR – the Croatian Web directory, and its effectiveness in providing relevant response to users' queries.

2. CASE STUDY

The WWW.HR is a Web-based information service supported by the Croatian Academic and Research Network – CARNet. Established in 1994, WWW.HR tends to be a thematic portal, providing regional information specifically concerning Croatia. WWW.HR consists of two services: general facts about Croatia and a bilingual Web directory. The directory is a hierarchically organized, fully searchable catalogue of Croatian or Croatia-related Web sites. Its top level contains 14 categories and an ever growing number of subcategories [3].

Sites are submitted by their authors. Upon the submission, the site index is created based on the information provided by submitters:

- 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 page.

Based on user's input, the directory index database is queried and matching results returned to the user, containing the list of submitted sites that match user's query. Several types of queries are supported:

- All keywords (default query),
- Logical expression (using and, or, not, +, - and parenthesis),
- Queries with wildcards ('*' representing any keyword suffix),
- Phrase (any text in quotation marks).

3. ANALYSIS

In order to assess users' behaviour, the query statistics were generated from site access log during the period of May 6th – June 16th, 2004. During that period, the directory received more than 2,000,000 pageviews in almost 450,000 sessions, with almost 10% of pageviews directed to search facility

(session is a set of requests coming from the same IP address in the period of 15 minutes). From the data from the log file, the following 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.

4. RESULTS

The results show that the average length of the query is 1,54 terms, more than 60% of the queries contain only one query term, while slightly more than 3% contain more than 4 query terms. Advanced query features such as logical operators or stemming are very seldom used (1,28% of all valid queries), but more than 7% of total query terms were misspelled (according to the Hascheck spelling checker [4]) Also, the analysis showed that 20 most frequently used query terms are found in 12% of queries. As to the results returned to the user, 30% of queries yielded no results at all, and 35% of queries returned more than 15 results.

5. CONCLUSION

In the study of users' search patterns, we analysed almost 300,000 user queries from to the biggest Croatian Web directory. We found that most users use one or two query terms, make a fairly large amount of spelling errors and seldom use advanced search features. Although the results show that users' search patterns are in compliance with similar research on other search engines [1], [2], they clearly indicate that users fail to recognize the difference between the search engine which crawls the Internet and indexes the pages it encounters, and the web directory. They obviously expect to see the same results from their queries. The high percentage of queries that return no matching results or too many matching results indicates the need for new search mechanisms on the directory and thus presents a challenge which is being addressed through the use of ontologies [5]. One of the ways to define an ontology is by analysing which query terms return the same or similar results and combining such query terms into one ontology.

6. REFERENCES

- [1] Amanda Spink, Dietmar Wolfram, B.J.Jansen, Tefko Saracevic: "Searching the Web: the public and their queries", JASIS (Journal of the American Society for Information Science), 2000, 52(3):226-234
- [2] Steen Christensen, Hans J. Skovgaard: "Web Site Usability Metrics: Search Behavior – Search Trends", Mondosoft, URL: <http://www.mondosoft.com/>, June 2004.
- [3] Maja Matijašević, Gordan Gledec: "Implications of Web resource composite for Web server workload characteristics", Proceedings of NEW2AN 2004, Sankt Petersburg, Russia, February 2004, pp. 189-193.
- [4] Hascheck – Croatian Academic Spelling Checker – <http://hacheck.tel.fer.hr/>, June 2004.
- [5] Damir Jurić, Maja Matijašević, Gordan Gledec: "Using Ontologies To Improve Search On WWW.HR Web Directory", to be presented at CUC 2004.