THESES OF Ph.D. DISSERTATION

NEW METHODS IN WEB INFORMATION
RETRIEVAL EFFECTIVENESS

Adrienn Skrop

Supervisor: Dr. Sándor Dominich

University of Pannonia

Faculty of Information Technology

Information Science and Technology Ph.D. School

2006

1. MOTIVATION AND AIM

In Information Retrieval (IR) the evaluation of IR systems plays an essential role. The most important type of evaluation of IR systems is retrieval effectiveness evaluation. Retrieval effectiveness evaluation measures how well a given system or algorithm can match, retrieve and rank documents that are relevant to user's information need (Mizzaaro, 1997). Laboratory testing of IR algorithms is based on the Cranfield paradigm. The Cranfield paradigm is based on a test collection, and retrieval effectiveness is measured using the standard measures Precision and Recall (Spark Jones, 1981).

Information retrieval on the Web is different from retrieval in traditional document collections. Thus, the Cranfield type evaluation of Web IR systems is usually not possible: the standard measures cannot be calculated. New or revised evaluation measures are required (Gwizdka et. al., 1999; Agosti et. al., 2001; Bar–Ilan, 2005; Sufyan–Beg, 2005; Wang et. al., 2006).

Traditional information retrieval evaluations and early Web experiments evaluated retrieval effectiveness according to how well methods can find documents that contain relevant text. Recent research suggests, however, that this kind of task is not a typical WWW search task (Broder, 2002). Three WWW-based retrieval tasks can be identified: navigational, informational, and transactional. One of the most important navigational tasks is the home page finding task. The home page finding problem is one where the user wants to find a particular site and the query

names the site. Home page finding queries typically specify entities such as people, companies, departments and products.

In the dissertation the Home Page Finding problem is addressed not from an algorithmic (retrieval method) point of view but from a user's viewpoint. I elaborated two new measures – Pseudo Precision and Mean Pseudo Rank – to evaluate the home page identification capability of queries on the Web. The Pseudo Precision and Mean Reciprocal Rank measures were elaborated using the hazard rate function of the Mathematical Theory of Reliability and the Mean Reciprocal Rank measure of traditional home page finding.

2. METHODS USED IN THE THESIS

In the home page finding task the search engine had to find the only one correct answer. The home page finding effectiveness of search engines is evaluated using the Mean Reciprocal Rank measure. The Mean Reciprocal Rank of a query q_i is the reciprocal $1/r_i$ of the rank r_i at which the correct answer v_i was retrieved. The score of a sequence of queries q_1, ..., q_i, ..., q_n is the mean of the individual query's reciprocal ranks: $\dfrac{1}{n}\sum_{i=1}^{n}\dfrac{1}{r_i}$.

The Mathematical Theory of Reliability is the overall scientific discipline that deals with general methods and procedures during the planning, preparation, acceptance and testing of devices (Gnedenko et. al., 1969). These methods and procedures ensure the maximum effectiveness of

devices during use. The Mathematical Theory of Reliability develops methods of evaluating the reliabilities of devices and introduces various quantitative indices for measures of devices performance. The reliability of a device is defined to be the probability of performing its purpose adequately for the time intended. The most important measure of reliability is the failure rate function $\lambda(t)$. The empirical value of failure rate is the number of failures that can be expected to take place over a given unit of time. The failure rate is determined as follows. Perform experiments with N copies of a device. Let $n(t)$ be the number of surviving devices at time t. Then the failure rate is: $\lambda(t) = \dfrac{\Delta n}{\Delta t n(t)}$, where Δn is the number of failures in $(t, t + \Delta t)$. By calculating the failure rate for smaller and smaller intervals of time Δt, the interval becomes infinitely small. This results in the hazard function $h(t)$, which is the instantaneous failure rate at any point in time.

3. NEW SCIENTIFIC RESULTS

3.1   Theses T1

I proposed the Pseudo Precision and Mean Pseudo Rank measures to evaluate the home page identification capability of queries on the Web. [Chapter 4] [SKROP 4, SKROP 7]

3.2 Theses T2

I proposed the MICQ method to measure the home page identification capability of search queries on the Web. [Chapter 5] [SKROP 4, SKROP 3, SKROP 7, SKROP 6]

3.3 Theses T3

Using the MICQ method it was showed that:

a) the home page identification capability of the acronyms of Hungarian government offices: average Pseudo Precision is 0.74, average Mean Pseudo Rank is 0.55. [Section 6.5]

b) the home page identification capability of the acronyms of higher educational institutions in Hungary: average Pseudo Precision is 0.55, average Mean Pseudo Rank is 0.55. [Section 6.6.2] [SKROP 3, SKROP 5, SKROP 6]

c) the home page identification capability of the acronyms of higher educational institutions in Denmark: average Pseudo Precision is 0.86, average Mean Pseudo Rank is 0.74. [Section 6.7.2]

d) the home page identification capability of the acronyms of Hungarian parties: average Pseudo Precision is 0.72, average Mean Pseudo Rank is 0.68. [Section 6.8]

e) the home page identification capability of the acronyms of Danish parties: average Pseudo Precision is 0.57, average Mean Pseudo Rank is 0.39. [Section 6.8]

4. PUBLICATIONS RELATED TO THE THESIS

[SKROP 1]   DOMINICH, S., **SKROP, A.,** and TUZA, ZS. (2006). Formal Theory of Connectionist Web Retrieval. In: Crestani, F. et al. (eds.) *Soft Computing in Web Information Retrieval*, Springer Verlag, pp. 161-194. ISBN: 3-540-31588-8.

[SKROP 2]   DOMINICH, S., and **SKROP, A.** (2005). PageRank and Interaction Information Retrieval. *Journal of the American Society for Information Science and Technology*. vol. 56, no. 1, pp: 63-69. John Wiley & Sons , ISSN 1532-2882, IF = 2.086.

[SKROP 3]   **SKROP, A.** and DOMINICH, S. (2004). Measuring the Identification Capability of Acronyms on the World Wide Web: a Comparative Study. *Journal of Web Engineering*. Vol. 3. No. 3 & 4, pp. 200-215. Rinton Press**,** ISSN 1540-9589.

[SKROP 4]   DOMINICH, S., GÓTH, J., and **SKROP, A.** (2003). A Study of the Usefulness of Institutions' Acronyms as Web Queries.*: In: Sebastiani, F. (Ed.), Advances in Information Retrieval, Lecture Notes in Computer Science, LNCS 2633.* Springer Verlag, pp. 580-587. ISSN 0302-9743, IF=0.513.

[SKROP 5]   **SKROP, A.** (2004). Do Acronyms of Higher Educational Institutions in Hungary Identify Their Own Institutions on the Web? *Proceedings of WWW2004 Workshop on Measuring Web Search Effectiveness: The User Perspective*, New York, NY, USA, May 17–22.

[SKROP 6]   **SKROP, A.** (2004). Do acronyms of institutions in Hungary identify their own institutions on the Web?. *Proceedings of the 2nd PhD Mini-Symposium,* University of Veszprem, Veszprem, Hungary, June 21. pp. 75–77. ISBN 9639495573.

[SKROP 7]   **SKROP, A.** (2003). A Study of the Usefulness of Institutions' Acronyms as Web Queries. *Proceedings of the 1st PhD Mini-Symposium,* University of Veszprem, Veszprem, Hungary, June 16. pp. 58–60. ISBN 9639495271.

[SKROP 8]   GÓTH, J., and **SKROP, A.** (2005). Varying Retrieval Categoricity Using Hyperbolic Geometry. *Information Retrieval*. Vol. 8. No. 2. pp. 265-283. IF = 1.231

[SKROP 9]   **SKROP, A.** (2002). Measuring Relevance Effectiveness in Information Retrieval. *Conference of PhD Students in Math, Physics and Informatics.* University of Veszprem, Veszprem, Hungary, May 30. (in Hungarian)

[SKROP 10]   HORVÁTH, M. és **SKROP, A.** (2002). Searching the Hungarian Web. *Conference of INFO Savaria 2002*, Szombathely, Hungary, April 4-6. (in Hungarian)

[SKROP 11] **SKROP, A.** (2001). Content Exploration by Automatic Classification. *International Conference on Cultural Innovation and Information Retrieval*, National Széchenyi Library, Budapest, Hungary, September 14. (in Hungarian)

[SKROP 12] DOMINICH, S., HORVÁTH, M., and **SKROP, A.** (2001). Evaluations of Relevance Effectiveness in Interaction Information Retrieval. *Proceedings of the 23rd European Colloquium on Information Retrieval Research*, Springer Verlag, eWic, British Computer Society Information Retrieval Specialist Group, GMD IPSI, Darmstadt, Germany, April 4-6, pp. 208-221.

5. REFERENCES

Agosti, M. and Melucci, M. (2001). Information Retrieval on the Web. *In M. Agosti, F. Crestani, G. Pasi (Eds.),Lecture Notes in Computer Science ,Volume 1980 / 2001 Lectures on Information Retrieval:* Third European Summer-School, ESSIR 2000, Varenna, Italy, September 11-15, 2000. Revised Lectures, pp. 242-285, Springer Berlin / Heidelberg ISSN 0302-9743.

Bar-Ilan, J. (2005). Comparing rankings of search results on the Web. *Information Processing & Management*, Volume 41, Issue 6, December 2005, pp. 1511-1519.

Broder, A. (2002). A taxonomy of Web search. *ACM SIGIR Forum 36, 2,* pp. 3-10.

Gnedenko, B. V., Belyayev, Y. K., and Solovyev, A. D. (1969). Mathematical Methods of Reliability Theory, Academic Press, New York.

Gwizdka, J. and Chignell, M. (1999). Towards Information Retrieval Measures for Evaluation of Web Search Engines. *Unpublished manuscript.* Available at http://www.imedia.mie.utoronto.ca/~jacekg/pubs.html

Mizzaro, S. (1997). Relevance: The Whole History. *JASIS 48, 9.* pp. 810-832.

Sparck Jones, K. (1981) The Cranfield tests. In: *Information Retrieval Experiment.* (K. Sparck Jones, ed.). London: Butterworth.

Sufyan Beg, M. M. (2005). A subjective measure of web search quality. *Information Sciences*, Volume 169, Issues 3-4, 1 February 2005, pp. 365-381.

Wang, Y. D. and Forgionne, G. (2006). A decision-theoretic approach to the evaluation of information retrieval systems. *Information Processing & Management*, Volume 42, Issue 4, July 2006, pp. 863-874.