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: \( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{r_i} \).

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: 

\[
\lambda(t) = \frac{\Delta n}{\Delta n(t)},
\]

where \( \Delta n \) is the number of failures in \((t, t + \Delta t)\). By calculating the failure rate for smaller and smaller intervals of time \( \Delta t \), 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


5. REFERENCES


