Acta Mathematica Academiae Paedagogicae Nyíregyháziensis
20 (2004), 105–113
www.emis.de/journals

STATISTICAL METHODS IN MEASURING SEARCH ENGINE PERFORMANCE

ERZSÉBET TÓTH

Abstract. There is a growing need for a set of benchmarking tests for measuring search engines, because research on search engine evaluation is inconsistent in methods and approaches. Using these tests researchers can make comparisons between different search engines and carry out longitudinal studies on a particular search engine with a degree of consistency. Several problems emerge during the evaluations of search engines which are needed to be solved to measure search engine performance objectively. This paper presents a brief review of the different statistical methods and approaches utilized in search engine evaluations. It describes six significant research projects with their fulfilled activities and statistical findings. It also serves as a guideline for choosing an appropriate statistical method for use in search engine evaluations.

1. General problems related to search engine evaluations

Before discussing issues connected to search engine evaluation, a clear definition of search engine will be given. The term search engine refers to a general class of programs that enable users to search for documents on the World Wide Web. These programs index documents, then attempt to match documents relevant to a user's search requests. In evaluations researchers measure the quality of Web search services which can be determined on the basis of several measures. However, it is extremely difficult to find reliable measures that reflect this quality appropriately. *Oppenheim* [8] suggested that benchmarking tests should include the following measures at the very minimum:

1. precision;
2. relative recall;
3. speed of response;
4. consistency of results over an extended period;
5. proportion of dead or out of date links;
6. proportion of duplicate hits;
7. overall quality of results as rated by users;
8. evaluation of GUI for user friendliness;
9. helpfulness of help and variations of software for new and experienced users;
10. options for display of results;
11. presence of adverts;
12. coverage;
13. estimated/expected search length;
14. length and readability of abstracts;
15. search engine effectiveness.

2000 *Mathematics Subject Classification.* 62P30, 68M20.
Key words and phrases. Performance evaluation, measurement of search engines.

The lack of a standard set of measures causes a great problem in evaluations. Because of this deficiency research on search engine evaluation is inconsistent in methods and approaches. So there is a real need for working out a standard set of measures to evaluate search engines properly. From a statistical point of view it would be also interesting to examine if there is any relationship among these measures.

Evaluations have been carried out mostly on robot engines, but in principle any of the search engines can be measured. Several problems emerge during the evaluations of search engines e.g. all the time their are changing and developing their search mechanisms and user interface. In addition to these problems the constantly changing content of the web has to be taken into account as well. In most cases the results of evaluations remain valid only for a short period of time and indicate only the performance of search engines at that time. Although several difficulties associated with search engine evaluations, we have to make an effort to measure search engines currently in use. However, so far standardized evaluation approaches have not been applied for this purpose. In general experiments report their own idiosyncratic approaches and they mostly avoid the use of standardized evaluation approaches.

According to *Su* [14] researchers do not use a systematic approach in evaluations. They do not have a consistent view about what to measure and how to measure a service. In general they assess the relevancy of the first 10 or 20 hits. He notes that users are left out from most studies as active participants. In most cases the relevance judgments are made by researchers and not by users. Findings of these studies suggest that the differences in performance between the best two or three search services are minimal.

Leighton and *Srivastava* [4] stated that many earlier research studies have arrived at conflicting conclusions as to which services are better at providing superior precision. Most precision studies have had too small test suites, so they are not suitable for carrying out more thorough statistical analysis. During the evaluation of Web pages researchers have to avoid subtly favouring one service over another. Search results can be blinded so that the evaluator does not know from which search service the web page comes. We can use this blinding process only for the initial inspection of the web page, since later checking and updating are carried out with an awareness of the source of the page. Conscious or unconscious bias can enter at many points in the design methodology which must be guarded against, for example we can select a general subject area which we know a given search engine is stronger in than others. Evaluators have to use fair methods for retrieving search services and evaluating search results. They have to develop an unbiased suite of queries to measure search services objectively. This criticism also applies to evaluation tests conducted in traditional information retrieval systems.

2. Research projects focusing on search engine evaluation

2.1. *Sroka* [13] has evaluated the performance of Polish versions of English language search engines and homegrown Polish search engines (Polish AltaVista, Polish Infoseek, Virtual Poland, NEToskop, Onet.pl, WOW). In his evaluation precision was emphasized that he determined on the basis of topical relevance judgements. However, he did not consider the authoritativeness of retrieved web pages. He analysed the first 10 hits retrieved from each search engine with this method. He also studied the overlap of retrieved results and the response time of each search engine. The number of retrieved hits from each search engine was recorded, but he omitted recall as a relevancy criterion. He formulated 10 queries about various topics, four of them required Boolean logic (AND, AND NOT). He conducted searches with and without Polish diacritical marks. He concluded that he retrieved the largest number of documents for each query by using diacritical marks. The average number of relevant documents out of the first ten hits was calculated for each search engine. He determined a mean precision score for each query to find which queries were the most difficult for search engines to handle.

2.2. *Clarke* and *Willett* [3] have carried out 30 searches on three different search engines, such as AltaVista, Excite and Lycos. The relevance of the first 10 retrieved hits for each query was determined on the basis of a three-point scale: a score of 1 was assigned to relevant documents, 0.5 to partially relevant documents and 0 to non-relevant hits. Mean values for precision, recall and coverage were calculated. The significance of the differences in performance among the three search engines were evaluated by using the Friedman two-way analysis of variance test. *Leighton* and *Srivastava* used this test for the analysis of search-engine results.

The purpose of the applied statistical method is to test the null hypothesis that k different samples (corresponding to the three search engines) have been produced from populations having the same median. The data can be seen in a table consisting of N rows (here rows are equivalent to the 27 queries) and k columns. The data in each row are ranked from 1 to k. The Friedman statistic, F_r, is built on the sum of the ranks for each search engine [11].

2.3. *Leighton* and *Srivastava* [5] compared the performance of five search engines, such as AltaVista, Excite, Infoseek, Hotbot and Lycos. They constructed a test suite containing 15 queries that were submitted to search engines. They measured the precision of the first 20 results by taking into account the percentage of relevant hits within the first 20 retrieved. The relevance assessment of hits was based on six different relevance categories. They conducted five experiments for the first 20 precision. In general the *first X precision* reflects the quality of a service that the typical user will experience.

In the calculation of precision a weighing factor was used to increase value for ranking effectiveness. Precision and ranking effectiveness were combined into one metric. This metric incorporates several qualities: first, we have to take into consideration that a link either fulfils the relevance criteria under examination, or it does not. A binary scale of relevance was applied, because the relevance categories were different definitions of relevance. Second, more weight is given to effective ranking of relevant documents. Third, the statistic has to reflect the fact that if the search service retrieves fewer results with the same number of good results, it is easier for the user to find relevant links. At last they had to decide how to handle inactive and duplicate links. In the first three tests duplicates were only deleted from the numerator of the precision ratio, search services were penalized. In the last two tests duplicates were not penalized. In all five experiments the retrieved inactive links were penalized. However, they did not consider two other types of duplicates, such as mirror pages and clusters of pages.

A *formula* calculating the performance of a service on a query is between zero and one. The first 20 hits for each query have been grouped by their status and type of relevance. We begin the formula for this metric with converting the relevance categories into binary values of zero or one. For example, if we perform a test for pages that minimally fulfilled the search expression, then a value of one is assigned to pages in categories one, two and three. A value of zero is given to all other pages. We assign a value to each position on a 20 position linear scale of value in order to weigh ranking. On this scale the first position represents the greatest value and the last position represents the smallest value. In the formula the first 20 hits are divided into three groups. In each group the links receive an equal weight, the weight that the first link in the group would have obtained in a 20 position linear scale of value. The first three links have a weight of 20, the next seven have a weight of 17, the last ten have a weight of 10. They added up these weighted values to produce the numerator of the metric.

For example, if a service retrieved five good links, it would calculate the score of $(3 \cdot 20) + (2 \cdot 17) = 60 + 34 = 94$ if these hits were the first five ranks. If these hits were all between ranks 11 and 20, it would score only $(5 \cdot 10) = 50$. They calculated the denominator of the metric from the number (up to 20) of links retrieved by the search service. If the service retrieved 20 or more links, then the sum of all weights to 20 was applied. $(3 \cdot 20) + (7 \cdot 17) + (10 \cdot 10) = 279$.

The denominator is altered, if we have fewer than 20 links returned. If the denominator were not altered, it would always be 279. If there are no links returned, the denominator would be zero, with the metric undefined. Because of this boundary condition, they calculated the denominator by adding up all of the weights to 20, 279, and then subtracting 10 for each link less than 20 retrieved.

For example, if a service retrieves 15 links, the denominator is $279-(5\cdot 10) = 229$, but if it retrieves one link, the denominator is $279-(19\cdot 10) = 89$. Then they divided the numerator by the denominator to calculate the final metric. They used a *Rube Goldberg machine* as a function that could be described with this complete formula:

$$\frac{(\text{Links } 1 - 3 \cdot 20) + (\text{Links } 4 - 10 \cdot 17) + (\text{Links } 11 - 20 \cdot 10)}{279 - [(20 - \min(\text{No. of links retrived}, 20)) \cdot 10]}$$

They assessed the metrics for each experiment to check if the residuals from an ANOVA [15] were distributed normally. It was found that only in experiment four the residuals were normally distributed. In the tests the Friedmann's randomized block design [10] was utilized, because the normality assumption needed for the ANOVA model was not fulfilled. In the Friedman test the blocks were the queries and the treatment effects were the search services. The Friedman test analyses population medians rather than mean values because of the skewness that is present. In all five experiments they refused the null hypothesis that the search service's medians could be equal. Because they refused null hypotheses, they could make pairwise multiple comparisons between individual services.

They have conducted a correspondence analysis of the queries by search services. For this purpose they used scores from experiment two as weights. They could see how a search service or a query corresponded to the composite score of the whole by using a correspondence analysis. They displayed all of the information in the correspondence relationship by presenting services and queries on a graph in higher dimensional space.

They concluded that Alta Vista, Excite and Infoseek provided more relevant hits than Hotbot and Lycos. The first 20 hits for the top three services included all of the words from the search expression more frequently than the first 20 hits for the lower two services. This metric formula could be tested against other weighing and scoring schemes, such as the traditional Coefficient of Ranking Effectiveness [7]. In the future a study should be made where the test suite is large enough to compare structured search expressions versus unstructured ones.

2.4. *Chignell, Gwizdka and Bodner* [2] have carried out two experiments for studying the performance of commercial search engines, such as Excite, Hotbot and Infoseek. In the first experiment they analysed the effect of time of day and the effect of query strategy on query processing time for each of three search engines. They used nine prespecified queries which could be divided into three different categories: general, higher precision and higher recall. Document relevance was measured by using a 'consensus peer review' procedure. So they chose six other search engines on which the same queries were conducted. They obtained binary judgment of relevance from the hits of six different search engines (AltaVista, Lycos, Northern

Light, Search.com, Web Crawler, Yahoo) to the same query. A hit from one of the search engines was considered to be relevant if it was also retrieved by at least one of the six referee search engines in response to the query. The usage of this method is questioned in relevance assessment, because there is a minimal overlap between the set of hits of the search engines. They also measured the number of broken and duplicate links for each search.

Multivariate analysis of variance (MANOVA) [1] was applied for the analysis of the data. They found a multivariate effect for the two-way interaction of Query Type and Search Engines ($F(20, 604.6) = 10.6, p < 0.001$). A significant univariate interaction for precision caused this effect ($F(4, 186) = 6.9, p < 0.001$). The univariate interactions for the other three dependent variables (corresponding to the three search engines) were not important.

They realized a significant main effect of search engine query time ($F(2, 186) = 65.5, p < 0.001$). A significant difference was shown between the query processing times of Excite and Infoseek ($p < 0.001$), and Hotbot and Infoseek ($p < 0.001$) by post hoc Tukey testing.

They found a significant main effect of search engine on the number of broken links, on the number of duplicate links and on precision. There was a significant main effect of time of day on query processing time. They realized a significant main effect of query type on the precision scores. The three search engines executed best the general queries which were followed by high precision and high recall queries. At this stage of the experiment they have analysed only the effects of the independent variables on a single dependent variable.

We may define a user-oriented composite measure of performance on the basis of four dependent variables. They received the ranking of the three search engines for each dependent variable by performing post hoc Tukey tests. A simple formula was devised by using ranking information: the number of first, second and third place rankings were summed for each search engine. The first ranks were multiplied by a coefficient of 3, the second ranks by a coefficient of 2, and the third rank by a coefficient of 1.

For instance: the composite measure of performance for Excite is:

$$3 \cdot \frac{1}{4} + 2 \cdot \frac{2}{4} + \frac{1}{4} = \frac{2}{3} = 66.7\%.$$

This measure has two boundaries. First, it does not take into account the case where there is no statistical significant difference between two search engines. Second, all dependent variables are treated as being equal.

The second experiment analysed the influence of geographical coverage and Internet domains on search engine performance. They used three search engines (Altavista, HotBot, Infoseek), six Internet domains and four queries in a fully factorial design of the 72 observations. The four queries were translated to three different languages. They measured the precision of the first 20 hits on the basis of human relevance judgments. Altogether 14 dependent measures were used in the second experiment. The following are few measures that indicate significant differences between search engines.

Full precision (PRECFULL): : it takes into consideration the subjective score assigned to each hit. Equality (1) displays how full precision is calculated.

$$\text{precFull} = \frac{\sum_{i=1}^{\text{mF20Hits}} \text{score}_i}{\text{mF20Hits} \cdot \max \text{HitScore}} \tag{1}$$

We mean by score_i the score assigned to ith hit,

$$\text{mF20Hits} = \min(20, \text{hitsReturned}),$$

hits Returned – total number of hits returned, max Hitscore – max score that can be given to one hit which is 3, PRECFULL is determined for hitsReturned > 0.

Best precision (PRECBEST): This measure takes into consideration only those hits that obtained score of 3 (corresponding to the most relevant hits).

$$\text{precBest}(1, \text{mF20Hits}) = \frac{(\text{count\_of\_score}_\text{i} = 3)_{i=1}^{\text{mF20Hits}}}{\text{mF20Hits}} \tag{2}$$

Differential precision (DPOBJ): takes into account the position of relevant hits within the first 20 hits retrieved. It is desirable for users to find more relevant hits in the first 10 hits than in the second 10 hits. This measure is built on objective calculation of precision. It mechanically looks for the presence and absence of required terms and it makes a distinction between good and bad links. *Differential objective precision* can be calculated between the first 10 and the second 10 hits in the following way:

$$\begin{aligned} &\text{DpObj}(1, \text{mF20Hits}) \\ &\qquad = \text{precObj}(1, \text{mF10Hits}) - \text{precObj}(\text{mF10Hits}, \text{mF20Hits}) \end{aligned} \tag{3}$$

Expected search length (FSLENi): was first recommended by Cooper and detailed by Van Rijsbergen in 1979. It takes into consideration how many nonrelevant items a user has to examine on the average to find a given number of relevant documents. So this measure reflects the user effort during information retrieval. They used a modified version of this measure. It considers all the documents (relevant and nonrelevant ones) that a user has to examine and one level of links from the retrieved web pages to relevant items.

This measure is built on the number of web pages that a user has to examine before finding the i most relevant pages. With this measure, the most relevant web pages either obtain a relevance score of 3 or a relevance score of 2 and those pages that contain one or more links to a relevant page or to pages will have a relevance score of 3. All web pages that the user has to examine to find the i most relevant pages are calculated as 1. Except those pages that contain links to the most relevant pages which are counted as 2.

$$\text{fSLen}_i = 1 - \frac{\max \text{Slen}_i - \text{SLen}_i}{\max \text{SLen}_i - \text{bestSlen}_i} \tag{4}$$

We mean by maxSLen_i the maximum search length for i relevant web pages within n retrieved search hits, bestSLen_i represents the best (i.e. the shortest) possible search length for i relevant web pages, SLen_i is a search length for i most relevant web pages. Within the range of function fSLen_i $[0, 1]$ 0 means the best search length which is the shortest one. fSLen_i was calculated for $i = 1$ and 3.

Hits and Hit ratio (HITS, HITRATIO): they recorded the total number of hits retrieved for a query. They calculated hit ratio as the ratio of the total number of hits returned for a query to the total number of hits retrieved by a given search engine in a given domain.

Full factorial multivariate analysis was done, in which search services and domain names were used as the independent factors, with 14 dependent measures. They found a significant multivariate interaction between search engines and domains ($F(221, 425.24) = 1.56, p < 0.001$). Interaction between search engines and domains had significant univariate effects on the number of unique hits, on total number of hits, on the proportion of retrieved hits to each search engine collection size, on the quality of returned results, and it also had a significant effect on search length 1. The domain name and the search engine had a significant multivariate effect separately. They conducted univariate analyses to determine the source of these effects. The search engine had a significant univariate effect on the following measures: differential objective precision, best and full precisions, and search length 1.

2.5. Analyses of search engine transaction logs [12] have shown that the average users formulate term-based queries having approximately two terms and they sometimes use operators. In contrast to this search experts apply more search terms and advanced search operators than the average searchers. *Lucas* and *Topi* [6] conducted a comprehensive study to examine the influence of query operators and term selection on the relevancy of search results. 87 participants involved in the survey formulated queries on eight topics that were used on a search engine of their choice. Besides this search experts constructed queries on each of the topics that were submitted to the eight preferred search engines of participants (AltaVista, AOL, Excite, Go, Google, iWon, Lycos, Yahoo!).

All of the queries were executed and analysed during a 1-day period. A cutoff value of 10 was used in judging the relevancy of pages, because relevant links appearing on the first page of search hits were most likely to be viewed. The relevancy of the first 10 web pages was judged on the basis of a four-category ordinal scale for relevancy. The relevancy criteria associated with each of the relevancy scores for a given search topic were determined independently.

The most important results of the study were associated with the research model and examined by using two different regression analyses, such as *full multiple regression and step-wise hierarchical multiple regression.* Average standardized relevancy was a dependent variable in the regression model that indicated search engine performance. In the research model seven variables were connected to operator usage, and four variables were connected to term usage. All of them were treated as independent variables. If we take together these variables they will explain 31,8 of the variance in the dependent variable. We consider this result both statistically significant and a relevant amount of variance. The first-order Pearson correlations between the research variables were presented in a matrix.

It was found that search term selection and usage are more important than the selection and usage of operators. The two independent variables closely related to dependent variable are term variables. One of them assesses the number of terms (e.g. taking into account the absolute difference between the number of terms in a subject's query and a corresponding expert query). Another assesses the number of subject terms that match with terms in the corresponding expert query. If we take together these two variables they will explain more than 25% of the variance in the dependent variable. The number of misspelled terms is also important which forecasts the search performance. These findings pay our attention to the fact that in the training and support materials developed for search engines more emphasis should be placed on issues connected to search term selection and usage.

We find only one operator variable among the four most significant independent variables, which is the *number of other nonsupported operators*. It actually assesses the number of NOT, OR and (-) operators in contexts where their usage is not supported. However, the usage of nonsupported ANDs and nonsupported (+) operators is positively linked to search performance.

2.6. *Radev*, *Libner* and *Fan* [9] examined how successful the most popular search engines are at finding accurate answers to natural language questions. Altogether 700 queries were submitted to nine different search engines. They downloaded and stored the top 40 documents retrieved by each search engine. It was established that all search engines returned at least one correct answer on more than three-quarters of factual natural language questions. In the experiment they tested three hypotheses which were the following:

1. Search engines are effective at answering factual natural language questions.
2. Certain characteristics of questions forecast the likelihood of retrieving a correct answer across all search engines.
3. Questions with particular characteristics are more likely to draw the correct answer from specific search engines.

A score was assigned to each of the search engines as the sum of the reciprocal ranks of documents containing the correct answer. Then they calculated the mean score across all queries for each search engine to evaluate the first hypothesis mentioned above.

To evaluate hypotheses two and three, they coded the 700 queries on the following four factors:

1. type of answer required
2. presence of proper noun
3. time dependency of the query
4. number of words in query

Corresponding to the three hypotheses, they planned to use an analysis of variance (ANOVA) (1) to compare the general score of the nine search engines, (2) find out the importance of the above four factors in predicting score, and (3) measure differential performance of search engines on each of these factors.

The initial distribution of scores showed a positive skew, because there was a large proportion of zero-value scores. This skew would not fulfil the normality assumption of ANOVA. To solve this problem a two-step analysis was carried out.

1. Scores were converted to value of zero or nonzero. After that a binary logistic regression was conducted. A nonzero score was selected, when the search engine retrieved at least one correct answer in the top 40 documents for a given question. A zero score was chosen, when there was no correct answer retrieved. This analysis looks for a relationship between the four question characteristics and whether a correct answer is retrieved at all.
2. The second part of the analysis primarily dealt with nonzero values - cases where at least one correct answer was returned. The distribution of this restricted dataset still had some positive skew, so square-root transformation was applied to it. Then an analysis of variance (ANOVA) was conducted.

On the basis of the findings they concluded that all the search engines managed to return the correct answer in the top 40 documents 75% of the time or more. This fact suggests that Web search engines can be used effectively as a component of a system that is capable of finding answers to factual natural language questions.

3. Conclusion

Research on search engine evaluation is inconsistent in applied methods and approaches, for this reason there is a growing need for a set of benchmarking tests for search engines. In addition to this a standard set of measures should be worked out for monitoring the performance of search engines. Using these tools researchers can make comparisons between different search engines and carry out longitudinal studies on a particular search engine with a degree of consistency that was not possible earlier. We hope that this summary serves as a guideline for choosing an appropriate statistical method for use in search engine evaluations. In the design of our experiment we have to decide which statistical method would correspond to our research purposes and create a statistical model with the appropriate variables.

References

[1] M. Bilodeau and Brenner D. *Theory of multivariate statistics*. Springer, 1999.

[2] M.H. Chignell, J. Gwizdka, and R.C. Bodner. Discriminating meta-search: a framework for evaluation. *Information Processing and Management*, pages 337–362, 1999.

[3] J. Clarke and P. Willett. Estimating the recall performance of Web search engines. *Aslib Proceedings*, 49(7):184–189, July/August 1997.

[4] H.V. Leighton and J. Srivastava. Precision among World Wide Web search services (search engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. http://www.winona.msus.edu/library/webind2/webind2.htm, 1997.

[5] H.V. Leighton and J. Srivastava. First 20 precision among World Wide Web search services (search engines). *Journal of the American Society for Information Science*, 50(10):870–881, 2000.

[6] W. Lucas and H. Topi. Form and function: the impact of query term and operator usage on web search results. *Journal of the American Society for Information Science*, 53(2):95–108, 2002.

[7] T. Noreault, M. Koll, and M.J. McGill. Automatic ranked output from Boolean searches in SIRE. *Journal of the American Society for Information Science*, pages 333–339, 1977.

[8] C. Oppenheim, A. Morris, C. McKnight, and S. Lowley. The evaluation of WWW search engines,. *Journal of Documentation*, 56(2):190–211, 2000.

[9] D.R. Radev, K. Libner, and W. Fan. Getting answers to natural language questions on the Web. *Journal of the American Society for Information Science*, 53(5):359–364, 2002.

[10] D. Raghavarao. *Constructions and combinatorial problems in design of experiments*. Dover Publications, 1988.

[11] S. Siegal and N.J. Castellan. *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 1988.

[12] C. Silverstein, M. Henzinger, J. Marais, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[13] M. Sroka. Web search engines for Polish information retrieval: questions of search capabilities and retrieval performance. *International Information & Library Review*, pages 87–98, 2000.

[14] L.T. Su. Developing a comprehensive and systematic model of user evaluation of Web-based search engines. In *National Online Meeting: Proceedings*, Information Today, pages 335–345, Medford, NJ, 1997.

[15] H.E.A. Tinsley. *Handbook of applied multivariate statistics and mathematical modeling*. Academic Press, 2000.

Received February 05, 2004.

Institute of Mathematics and Computer Science,
College of Nyíregyháza,
H-4401 Nyíregyháza, P.O Box 166
Hungary
E-mail address: tothe@nyf.hu