What are the important variables in the evaluation of information retrieval systems?

Michael J. Nelson
Graduate School of Library and Information Science
University of Western Ontario
<mnelson@julian.uwo.ca>

Although there is a long history of the evaluation of retrieval systems, in many cases we are still working with the original paradigms of queries and documents and using recall and precision as the main measures. A more basic framework is needed in order to evaluate the evaluations and to decide which are the most important results. Some researchers have concentrated on the language aspects of evaluation (Blair 1990). Others, such as Salton concentrated on the automated indexing aspects. More recently, many studies have concentrated on the user interface characteristics. How can the results from all these studies be integrated and recommendations for the design and use of retrieval systems be made? This paper will categorize the variables involved in information retrieval evaluation and look at what earlier research studies can tell us about the relative importance of these variables. For example, previous research on the Text Retrieval Conference (TREC) results has shown that statistically the queries contribute more to the total variance of recall and precision than the differences in retrieval systems. Finally some recommendations for testing the relative importance of variables in information retrieval will be proposed.

Introduction

Information retrieval (IR) testing has been carried out for about forty years now. What have we learned about retrieval systems that has helped to do better searches of databases? How can we design information retrieval tests to gain the most information? This question arose out of previous work on a statistical analysis of the results of the TREC series of tests (Nelson 1995). One of the findings from the TREC tests (Tague-Sutcliffe 1995) is that the variance over queries is much more than the variance over systems. Thus the variables related to the query should be investigated more rigorously. This paper will review some of the previous research which supports this view and will go on to investigate some of the characteristics of queries in the TREC retrieval tests and interpret some of the statistical results. Throughout this paper the work 'query' will be used to describe

both the problem statement and the actual search statement put to the system, although in many other circumstances one should make a clear distinction between the two concepts.

Variables in information retrieval

The basic research model looks at the relationship between independent (or predictor) variables and dependent (or response) variables. Although the choices of variables measured in a particular study depend largely on the objectives and research questions, there have been several papers which have ambitiously tried to delineate all the variables involved in a typical IR study. One of the earliest papers to list the variables involved in information retrieval was by Saracevic and Rees (1967). They divided the variables into two categories, purpose components and function components which are summarized in table 1.

Table 1 (adapted from Saracevic and Rees 1967)

Purpose components	Function components
1. Type of users Identification of user population Specification of user characteristics Establishment of the size of user population Elaboration of users' information needs 2. Subject discipline(s) Definition of the subject boundaries Establishment of the input criteria 3. File size Establishment of size of file to yield satisfactory output	1. Acquisition 2. Indexing language 3. Coding 4. File organization: (i) Documents (ii) Document representations 5. Question analysis 6. Searching procedures 7. Dissemination

Within the function components they concluded there are three classes of variables: system mechanics, human factors and system-human interaction. Many variables are compiled for each of the seven function components sub-classed under the three Purpose components. In particular, four common factors are listed within every system-human interaction class: time in performing the task, work motivation (including policies), environmental conditions and interpretation by personnel of system rules and policies. This gives a large number of variables to consider when looking for possible explanations for differences in IR experiments. They promote the use of control variables throughout IR experiments to discover differences. One of the conclusions is "If a retrieval system is studied as a whole the model must

broad enough to incorporate the large number of variables which operate within any and all retrieval systems."(Saracevic and Rees 1967, 13). Another conclusion is that "By far the greatest source of variation is the human factor. Since the human factor is the basic component under study, it is impossible to eliminate it from *any* information retrieval experiment." (Saracevic and Rees 1967, 13). This comment seems to be based on past general experience with retrieval tests and not on any particular data.

Fidel and Soergel (1983) produce an several impressive tables of variables affecting the online search process. They divide the variables into eight categories plus the interactions between the categories. These categories are: (1) the setting, (2) the user, (3) the request, (4) the database, (5) the search systems, (6) the searcher, (7) the search process, and (8) the search outcome. Even with such a large list the conclude ". . . it is almost impossible to create an exhaustive list of variables affecting online retrieval since this process involves human elements . . ." (Fidel and Soergel 1983, 169).

Tague-Sutcliffe (1992) also suggests a long list of variables which can be investigated under the following categories: database, information representation, users, queries, search intermediaries, retrieval process, and retrieval evaluation. She advocates the use of analysis of variance and other multivariate methods to compare treatments or effects of independent variables in experimental designs.

Keen (1992) concentrates on presenting the results of experiments in IR and discusses many of the problems of data analysis. In particular, the problem of assessing performance differences is discussed. He notes that percentage increases in scores of any kind can be very misleading and that the researcher should be testing for statistical significance. Even if we have statistical significance, is the difference consistent under a variety of conditions and is the difference of practical significance? Small statistically significant findings may not translate into real world practical differences.

It is impossible for any research study to investigate all the hundred's of variables listed, let alone look at the interactions between many different variables. Therefore studies, whether surveys, experimental or observational have restricted their scope to a small number of variables. This makes it more difficult to see which variables are the most important.

One of the main methods advocated by the research community to simplify the problem is to have a standard database, set of queries and relevance judgements, then vary other variables such as indexing methods or searching methods to see which was best. For a number of years the databases used by Salton's group were used as standards. Currently the databases used by the TREC series of tests seem to be the most popular. This is related to a comment by Blair (1996) "If we do not

make substantial progress in finding a standard of document retrieval performance, then we will not be able to distinguish between less effective and more effective retrieval techniques.".

Another problem area not addressed in this paper is the problem of output measures, the dependent variables. Recall and precision are still the most used measures but these are based on a very difficult to define concept of relevance. Others have played with variations on the theme such as utility or user satisfaction, but there are still many unanswered questions in this area and most of the IR test results must be qualified by saying they are based on judgements of relevance.

Some results and what they say about queries

Saracevic and Kantor (1988a, 1988b, 1988c) conducted one of the most comprehensive IR tests in terms of scope of variables involved. In particular they used judges to measure five different characteristics of queries: domain, clarity, specificity, complexity and presupposition which is the presence of implied concepts. Only three variables were significantly different on precision (as measured by log-odds ratios). These were specificity of subject part on a scale of 1 to 5, complexity by two measures, on a scale of 1 to 5 and by number of concepts. There were no significant differences on recall odds. It is informative to rank the variables involved in Saracevic's study (1988c, 209 table 32) by the increase in precision odds (Table 2). Note that variables are always stated in such a way as to give a positive precision odds.

Table 2

Variable	Increase in precision odds (by a factor)
1. No. of Not relevant items was low	5.88
2. No. of Relevant items was high	4.43
3. High contribution to problem resolution	3.21
4. No. of Partially relevant items was low	2.90
5. High level of overall satisfaction	2.49
6. Results worth more time than it took	2.40
7. Question of high complexity (scale)	2.27
8. Question of high complexity (no. of concepts)	2.16
9. Question of low specificity of subject part	2.13
10. Time limit placed on years searched	2.00
11. Existing public knowledge was high	1.87
12. Language was restricted to English	1.75
13. Dollar value for results was high	1.69

Variables 1, 2 and 4 are related to the basic definition of precision and variables 3, 5, 6 and 13 are dependent variables of evaluation. The 7, 8, 9 and 11 variables are properties of the query. Variables 10 and 12 are tactics to reduce the size of the retrieved set. This supports the hypothesis that variables related the query play a very large role in the outcome of a retrieval.

Another interesting method is used by Shaw, Burgin and Howell (1997a) to analyse some previous results in document clustering. They use random graph theory to set a basic low performance level and go on to show that document clustering does not provide significantly different results from this base line in most experiments. In a second article the same researchers (1977b) develop a low performance standard base on the hypergeometric distribution for other retrieval models. Their conclusion in this case is, in part, "Interactions of types of queries and retrieval techniques influence retrieval performance; understanding these interactions and influences can be expected to reveal opportunities for improving the effectiveness of retrieval systems." (Shaw, Burgin and Howell 1997b, 31).

One of the few studies to study query characteristics as the main objective is the research on Medline searching by Heine (1995). He found that when searching using MeSH terms "The variable 'number of search terms' is likely to dominate search performance more than is either informativeness (i.e. the topic's 'generality'), ore mean term specificity." (Heine 1995, 184). Of course as with all of the results reported here, one should read the original papers to see the limitations and conditions of these studies.

The TREC experiments

Sparck Jones (1995) reviews the TREC series of tests and in part of her description discusses the "performance factors" involved in the tests. She divides these into two categories, the environment variables and the system parameters which are the settings made by the various participants in their systems. She goes on to say:

> These heterogeneous relations between performance factors as represented across TREC approaches make it very difficult to assess the general implications of the results for particular approaches or even types of approach, because it may be hard to attribute responsibility for these results to any particular features (i.e. choices of setting for system parameters) of the approaches involved. (Sparck Jones 1995, 300)

Some of the results from the TREC tests are also listed and for the query processing the only technique which seems to offer a clear improvement is relevance feedback.

As two of her final eight conclusions Harmon says: ". . . 7. that concentrating effort on the request is much more effective and efficient than working (a priori) on document descriptions; 8. that query modification through expansion and reweighting is valuable" (Sparck Jones 95, 309).

For the TREC-3 tests, Tague-Sutcliffe (1995) used an analysis of variance model, where the runs were performed over the same set of queries:

$$Y_{ij} = \mu + \alpha_i + B_j + \epsilon_{ij}$$

Where Y_{ij} is the score for the ith participant on the jth query,
μ is the overall mean score,
α_i is the effect of the ith run,
β_j is the effect of the jth query,
ϵ_{ij} is the random variation about the mean.

The data showed that queries accounted for much more of the variance (mean square of 0.94) compared to the variance over systems (mean square of 0.38) as measured on average precision which in ranked retrieval is the average of the precision calculated at each relevant retrieved document. Furthermore, the Scheffé test produced large groups of systems which were not significantly different. Since the systems used are from a wide variety of organizations, many experimental, some operational, this seems to indicate that the query is even a more important component of the retrieval process than many researchers realized.

Further analysis of the TREC data

Note that there is no interaction term in the model to measure the interaction between queries and systems. The model assumes that there any interaction is additive only. This is because each query is searched by each system only once so we have only one observation in each cell. Fortunately, Tukey (Berenson 1983, 161) has devised a statistical test for interaction in this case. This was calculated for the TREC-3 data and the mean square was 2.57 with a very significant F value of 266.9. In some cases the data can be transformed to make the interaction effect additive. For this data the standard suggested transformations did reduce the effect but it was still statistically significant. This means that some systems perform better on some types of queries than on other types. The problem is to find a way to characterize the queries in order to help explain the variance in system performance. Previous work (Nelson, 1995) did not find any strong correlations between the performance of the queries and basic, easily measured properties of the original query statements such as word counts and number of relevant documents. Another way to analyze the data is look at the average precision for individual queries. For

example, one of the worst performing queries was number 151 with an average precision of 0.07 with a standard deviation of 0.01 and a maxim of 0.13. This shows that none of the forty-two systems performed very well on this query. What would be useful is a failure analysis carried out by each system to try to explain why each system failed. Thus, although the interaction effect is statistically significant, there are many queries were the differentiation amongst all the systems was very small as is shown in the original analysis of variance. One possible explanation for this is that the TREC tests used very long descriptions of the topics to be searched and whether or not the system used automatic or manual methods of constructing queries from these descriptions, approximately the same basic vocabulary was being used.

In order to explore the patterns of query performance further the queries were clustered using the cosine similarity coefficient and complete link clustering. This was base on Tague-Sutcliffe (1996) who approached the problem from the systems point of view and tried to cluster the systems based on there similarity of performance on the queries. She then coded the characteristics of the systems and techniques used in the systems and compared these to the clusters formed. No patterns were discerned in the clusters based on these characteristics. For the queries the characteristics investigated were subject topic (as given) and length of query statement. Again no obvious explanation for the clusters that occurred were found. Depending on the cutoff level there were one or two large clusters plus many small ones. Note that the same matrix of correlations is used whether the clustering is done on queries or systems.

Conclusion

The evidence give here supports the view of Robertson and Beaulieu (1997) who in a review of research and evaluation in information retrieval state "It is a commonplace that, for example, different retrieval techniques may be more or less good at dealing with different types of query, though the field is signally lacking in methods of classifying queries in order to assign each to the best technique."(Robertson and Beaulieu 1997, 55). They summarize the future research as:

> What we would really like is a model of the circumstances in which a particular mechanism is likely to be most useful to the searcher . . . Such a model may have a theoretical component (based on, for example, a cognitive view of the search process); but in any case, it should be possible to devise diagnostic experiments to help us formulate it, and to elucidate the

> relationship between the various senses of 'most useful'. (Robertson and Beaulieu 1997, 56).

What is needed in future research is more concentration on the queries and their characteristics by using multivariate statistical techniques to discover which combinations give the user better retrieval such as the work by Heine (1995) cited earlier. There should be a better integration of the experimental tradition as in TREC and the interactive user aspects. Eventually the retrieval software should help the user choose the best strategy depending on characteristics of the query, the database and the user instead of trying to be all things to all users.

References

Berenson, Mark L., David M. Levine, and Matthew Goldstein. 1983. *Intermediate statistical methods and applications: A computer package approach.* Prentice-Hall: Englewood Cliffs, N.J.

Blair, David C. 1990. *Language and representation in information retrieval.* Amsterdam: Elsevier.

Blair, David C. 1995. STAIRS redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science* 47: 4-22.

Boyce, Bert R., Charles T. Meadow and Donald H. Kraft. 1994. *Measurement in information science.* San Diego: Academic.

Fidel, Raya, and Dagobert Soergel. 1983. Factors affecting online bibliographic retrieval: A conceptual framework for research. *Journal of the American Society for Information Science* 34: 163-180.

Heine, M.H. 1995. An investigation of the relative influences of database informativeness, query size and query term specificity on the effectiveness of Medline searching. *Journal of Information Science* 21: 173-185.

Keen, E. Michael. 1992. Presenting results of experimental retrieval comparisons. *Information Processing and Management* 28: 491-502.

Nelson, Michael J. 1995. The effect of query characteristics on retrieval results in the TREC retrieval tests. In *Connectedness: Information, systems, people, organizations: CAIS/ACSI 95, Edmonton, Alberta, Vol. 23*, ed. Hope A. Olson, and Dennis B. Ward, 156- 163. University of Alberta.

Robertson, S.E., and M. Beaulieu. 1997. Research and evaluation in information retrieval. *Journal of Documentation* 53: 51-57.

Saracevic, Tefko. 1991. Individual differences in organizing, searching and retrieving information. *Proceedings of the Annual Meeting of the American Society for Information Science: 1991*, 82-86.

Saracevic, Tefko, and Paul Kantor. 1988a. A study of information seeking and retrieving. II. Users, questions and effectiveness. *Journal of the American Society for Information Science* 39: 177-196.

Saracevic, Tefko, and Paul Kantor. 1988b. A study of information seeking and retrieving. III. Searchers, searches and overlap. *Journal of the American Society for Information Science* 39: 197-216.

Saracevic, Tefko, Paul Kantor, Alice Y. Chamis, and Donna Trivision. 1988c. A study of information seeking and retrieving. I. Background and methodology. *Journal of the American Society for Information Science* 39: 161-176.

Saracevic, Tefko, and Alan M.Rees. 1967. Towards the identification and control of variables in information retrieval experimentation. *Journal of Documentation* 23: 7-19.

Shaw, W.M., Robert Burgin, and Patrick Howell. 1997a. Performance standards and evaluations in IR test collections: Cluster based retrieval models. *Information Processing and Management* 33: 1-14.

Shaw, W.M., Robert Burgin, and Patrick Howell. 1997b. Performance standards and evaluations in IR test collections: Vector space and other retrieval models. *Information Processing and Management* 33: 15- 36.

Sparck-Jones, Karen. 1995. Reflections on TREC. *Information Processing and Management* 31: 291-314.

Tague-Sutcliffe, Jean. 1992. The pragmatics of information retrieval experimentation, revisited. *Information Processing and Management* 28: 467-490.

Tague-Sutcliffe, Jean. 1995. A statistical analysis of the TREC-3 data. In *Overview of the Third Test Retrieval Conference (TREC-3)*, ed. Donna Harman, 385- 398. Gaithersburg, MD: National Institute of Standards and Technology. Special Publication 500-225.

Tague-Sutcliffe, Jean. 1996. *Statistical analysis of TREC-type information retrieval tests: A report to the National Institute of Standards and Technology*. University of Western Ontario, London, Ontario.