On Search Topic Variability in Interactive Information Retrieval

Ying-Hsang Liu
School of Information Studies
Charles Sturt University
Wagga Wagga NSW 2678, Australia
+61 2 6933 2171
yingliu@csu.edu.au

Nina Wacholder
School of Communication and Information
Rutgers University
New Brunswick NJ 089091, USA
+1 732 932 7500 ext. 8214
ninwac@rutgers.edu

ABSTRACT
This paper describes the research design and methodologies we used to assess the usefulness of MeSH (Medical Subject Headings) terms for different types of users in an interactive search environment. We observed four different kinds of information seekers using an experimental IR system: (1) search novices; (2) domain experts; (3) search experts and (4) medical librarians. We employed a user-oriented evaluation methodology to assess search effectiveness of automatic and manual indexing methods using TREC Genomics Track 2004 data set. Our approach demonstrated (1) the reusability of a large test collection originally created for TREC, (2) an experimental design that specifically considers types of searchers, system versions and search topic pairs by Graeco-Latin square design and (3) search topic variability can be alleviated by using different sets of equally difficult topics and well-controlled experimental design for contextual information retrieval evaluation.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval–*query formulation*, *search process*

General Terms
Measurement, Human Factors

Keywords
Information retrieval evaluation, Search topic variability, interactive information retrieval

1. INTRODUCTION
The creation and refinement of test design and methodologies for IR system evaluation have been one of the greatest achievements in IR research and development. In the second Cranfield project [6], the main purpose is to evaluate the effectiveness of indexing techniques at a level of abstraction where users are not specifically considered in a batch mode experiment.

Appears in the Proceedings of The 2nd International Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (CIRSE 2010), March 28, 2010, Milton Keynes, UK.
http://www.irit.fr/CIRSE/

The test design and methodology following the Cranfield paradigm culminated in the TREC (Text REtrieval Conference) activities since the 1990s. TREC has provided a research forum for comparing the search effectiveness of different retrieval techniques across IR systems in a laboratory and controlled environment [30]. The very large test collection used in TREC provided a test bed for researchers to experiment the scalability of retrieval techniques, which had not been possible in previous years. However, how we specifically take into account different aspects of user contexts within a more realistic test environment has been challenging in part because it is difficult to isolate the effects of user, search topic and system in IR experiments (see e.g., [7, 17] for recent efforts).

In batch experiments the search effectiveness of different retrieval techniques is achieved by comparing the search performance of queries. IR researchers have widely used the micro-averaging method of performing statistics on the queries in summarizing precision and recall values for comparing the search effectiveness of different retrieval techniques in order to meet the statistical requirements (see e.g., [25, 27]). The method of micro-averaging is intended to obtain reliable results in comparing search performance of different retrieval techniques by giving equal weights to each query.

However, within an interactive IR search environment that involves human searchers, it is difficult to use a large set of search topics. Empirical evidence has demonstrated that the search topic set size of 50 is necessary to determine the relative performance of different retrieval techniques in batch evaluations [3], because the variability of search topics has an overriding effect on search results. Another possible solution is to use different sets of topics in a non-matched-pair design [5, 21, 22], but theoretically it requires a very large sample of independent searches.

This problem has been exacerbated by the fact that we have little theoretical understanding about the nature and properties of search topics for evaluation purposes [20]. From a systems perspective, recent in-depth failure analyses of variability in search topics for reliable and robust retrieval performance (e.g., [11, 28]) have contributed to our preliminary understanding of how and why IR systems fail to do well across all search topics. It is still elusive what kinds of search topics can be used to directly control the topic effect for IR evaluation purposes.

This study was designed to assess the search effectiveness of MeSH terms by different types of searchers in an interactive search environment. By an experimental design that controls searchers, system versions and search topic pairs and the use of a relatively large number of search topics, we were able to demonstrate an IR user experiment that specifically controls the

search topic variability and assesses the user effect on search effectiveness within the laboratory IR framework (see e.g., [14, 15] for recent discussions).

2. METHOD

Thirty-two searchers from a major public university and nearby medical libraries in the northeast area of the US participated in the study. Each searcher belonged to one of four groups: (1) Search Novice (SN), (2) Domain Experts (DE), (3) Search Experts (SE) and (4) Medical Librarians (ML).

The experimental task was to conduct a total of eight searches to help biologists conduct their research. Participants searched either using a version of the system in which abstracts and MeSH terms were displayed (MeSH+) or another version in which they had to formulate their own terms based only on the display of abstracts (MeSH–). Participants conducted four searches each with two different systems: in one, they browsed a displayed list of MeSH terms (MeSH+) and in the other (MeSH–). Half the participants used MeSH+ system first; half used MeSH– first. Each participant was allowed to conduct searches on eight different topics.

The experimental setting for most searchers was a university office; for some searchers, it was a medical library. Before they began searching participants were briefly trained in how to use the MeSH terms. We kept search logs that recorded search terms, a ranked list of retrieved documents, and time-stamps.

2.1 Subjects

We used the purposive sampling method for recruiting our subjects since we were concerned with the impact of specific searcher characteristics on search effectiveness. The key searcher characteristics were different levels of domain knowledge in the biomedical domain and whether they had substantial search training. The four types of searchers were distinguished by their levels of domain knowledge and search training.

2.2 Experimental design

The experiment was a 4×2×2 factorial design with four types of searchers, two versions of an experimental system and controlled search topic pairs. The versions of a system, types of searchers (distinguished by levels of domain knowledge and search training) and search topic pairs were controlled by a Graeco-Latin square balanced design [8]. The possible ordering effects have been taken into account by the design. The requirement for this experimental design is that the examined variables do not interact and each variable has the same number of levels [16]. The treatment layout of a 4×4 Graeco-Latin square design is illustrated in Figure 1.

1	2	3	4	5	6	7	8
SN	DE	SE	ML	DE	SN	ML	SE
38	12	29	50	38	12	27	45
12	38	50	29	12	45	38	27
29	50	12	38	27	38	45	12
50	29	38	12	45	27	12	38
42	46	32	15	9	36	30	20
46	42	15	42	36	9	20	30
32	15	42	46	30	20	9	36
15	32	46	32	20	30	36	9

9	10	11	12	13	14	15	16
SE	ML	SN	DE	ML	SE	DE	SN
29	50	27	45	42	46	9	36
50	29	29	27	46	36	42	9
27	45	45	50	9	42	36	46
45	27	50	29	36	9	46	42
2	43	1	49	2	43	33	23
43	1	49	2	43	2	23	33
1	49	2	43	33	23	2	43
49	2	43	1	23	33	43	2

Note. Numbers 1-16 refers to participant ID; SN, DE, DE and ML refer to types of searchers, SN=Search Novices, DE=Domain Experts; SE=Search Experts; ML=Medical Librarians; Shaded and non-shaded blocks refer to MeSH**+** and MeSH**–** versions of an experimental system; Numbers in blocks refer to search topic ID number from TREC Genomics Track 2004 data set; 10 search topic pairs, randomly selected from a pool of 20 selected topics, include (38, 12), (29, 50), (42, 46), (32, 15), (27, 45), (9, 36), (30, 20), (2, 43), (1, 49) and (33, 23).

Figure 1. 4×4 Graeco-Latin square design

Because of the potential interfering effect of search topic variability on search performance in IR evaluation, we used a design that included relatively large number of search topics. In theory, the effect of topic variability and topic-system interaction on system performance could be eliminated by averaging the performance scores of the topics (micro-averaging method), together with the use of very large number of search topics. The TREC standard ad hoc task evaluation studies ([1, 3]) and other proposals of test collections (e.g., [20-22, 24, 29]) have been concerned with the large search topic variability in batch experiments. However, in a user-centered IR experiment it is not feasible to use as many as 50 search topics because of human fatigue.

We controlled search topic pairs by a balanced design in order to alleviate the overriding effect of search topic variability. We assumed that all the search topics are equally difficult, since we do not have a good theory about what makes some search topics more difficult than others. By design we ensured that each search topic pair was assigned to all types of searchers and was searched at least two times by the same type of searchers. This design required a total of 10 search topic pairs and a minimum of 16 participants.

2.3 Search tasks and incentive system

The search task was designed to simulate online searching situations in which professional searchers look for information on behalf of users. We decided to use this relatively challenging task for untrained searchers because choosing realistic tasks such as this one would enhance the external validity of the experiment. Considering the relatively difficult tasks, we were concerned that searchers may have problems completing all searches. Because research literature has suggested that the motivational characteristics of participants are possible sources of sample bias [23], we designed an incentive system to motivate the searchers.

We promised monetary incentives according to the participant's search effectiveness. Each subject was paid $20 for

participating and was also paid up to $10.00 dollars more based on the average number of relevant documents in the top ten search results across all search topics; on average each participant received an additional $4.40, with a range of $2.00 - $8.00.

2.4 Experimental procedures

After signing the consent form, the participant filled out a searcher background questionnaire before the search assignment. After a brief training session, they were assigned to one of the arranged experimental conditions and conducted search tasks. They completed a search perception questionnaire and were asked to indicate the relevance of two pre-judged documents when they were done with each search topic. A brief interview was conducted when they finished all search topics. Search logs with search terms and ranked retrieved documents were recorded.

The MeSH Browser [19], an online vocabulary look-up aid, prepared by U.S. National Library of Medicine, was designed to help searchers find appropriate MeSH terms and display hierarchy of terms for retrieval purposes. The MeSH Browser was only available when participants were assigned to the MeSH+ version of an experimental system; in the MeSH– version, participants had to formulate their own terms without the assistance of MeSH Browser and displayed MeSH terms in bibliographic records.

Because we were concerned that the topics were so hard that even the medical librarians would not understand them, we used a questionnaire regarding search topic understanding after each topic. The testing items of two randomly selected pre-judged documents, one definitely relevant and the other definitely not relevant, were prepared from the data set [26].

Each search topic was allocated up to ten minutes. The last search within the time limit was used for calculating search performance. To keep the participants motivated and reward their effort, they were asked to orally indicate which previous search result would be the best answer when the search task was not finished within ten minutes.

2.5 Experimental system

For this study, it was important for participants to conduct their searches in a carefully controlled environment; our goal was to offer as much help as possible while still making sure that the help and search functions did not interfere with our ability to measure the impact of the MeSH terms. We built an information retrieval system based on the Greenstone Digital Library Software version 2.70 [9] because it provides reliable search functionality, customizable search interface and good documentation [31].

We prepared two different search interfaces using a single system using Greenstone: MeSH+ and MeSH– versions. One interface allowed users to use MeSH terms; the other required them to devise their own terms. One interface displayed MeSH terms in retrieved bibliographic records and the other did not. Because we were concerned that the participant responds to the cue that may signal the experimenter's intent, the search interfaces were termed 'System Version A' and 'System Version B' for 'MeSH+ Version' and 'MeSH– Version' respectively (see http://comminfo.rutgers.edu/irgs/gsdl/cgi-bin/library/). The MeSH– version was used as baseline system for an automatic indexing system, whereas the MeSH+ version served as performance of a manual indexing system. That is, MeSH terms added another layer of document representation to the MeSH+ version.

The experimental system was constructed as Boolean-based system with ranked functions by the TF×IDF weighting rule [32]. More specifically, MGPP (MG++), a re-implementation of the mg (Managing Gigabytes) searching and compression algorithms, was used as indexing and querying indexer. Basic system features, including fielded searching, phrase searching, Boolean operators, case sensitivity, stemming and display of search history, were sufficient to fulfill the search tasks. The display of search history was necessary because it provided useful feedback regarding the magnitude of retrieved documents for difficult search tasks that usually required query reformulations.

Since our goal was specifically to investigate the usefulness of displayed MeSH terms, we deliberately refrained from implementing certain system features that allow users to take advantage of the hierarchical structures of MeSH terms, such as the hyperlinked MeSH terms, explode function that automatically includes all narrower terms and automatic query expansion (see e.g. [13, 18]) available on other online search systems. The use of those features would have invalidated the results by introducing other variables at the levels of search interface and query processing, although a full integration of those system features would have increased the usefulness of MeSH terms.

2.6 Documents

The experimental system was set up on a server, using bibliographic records from the 2004 TREC Genomics document set [26]. TREC Genomics Track 2004 Data Set document test collection was a 10-year (from 1994 to 2003) subset of MEDLINE with a total of 4,591,108 records. The test collection subset fed into the system used 75.0% of the whole collection, a total of 3,442,321 records, excluding the records without MeSH terms or abstracts.

We prepared two sets of documents for setting up the experimental system: MeSH+ and MeSH– versions. One interface allowed users to use MeSH terms; the other did not provide this search option. The difference was also reflected in retrieved bibliographic records.

2.7 Search topics

The search topics used in this study were originally created for TREC Genomics Track 2004 for the purpose of evaluating the search effectiveness of different retrieval techniques (see Figure 3-9 for an example). They covered a range of genomics topics typically asked by biomedical researchers. Besides a unique ID number for each topic, the topic was constructed in a format that included the title, need and context fields. The title field was a short query. The need field was a short description of the kind of material the biologists are interested in, whereas the context field provides background information for judging the relevance of documents. The need and context fields were designed to provide more possible search terms for system experimentation purposes.

ID: 39
Title: Hypertension
Need: Identify genes as potential genetic risk factors candidates for causing hypertension.
Context: A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.

Figure 2. Sample search topic

Because of the technical nature of genomics topics, we wondered whether the search topics could be understood by

human searchers, particularly for those without advanced training in the biomedical field. TREC search topics were designed for machine runs with little or no consideration for searches by real users. We selected 20 of the 50 topics using the following procedure:

1. Consulting an experienced professional searcher with biology background and a graduate student in neuroscience, to help make a judgment as to whether the topics would be comprehensible to the participants who were not domain experts. Topics that used advanced technical vocabulary, such as specific genes, pathways and mechanisms, were excluded;
2. Ensuring that major concepts in search topics could be mapped to MeSH by searching the MeSH Browser. For instance, topic 39 could be mapped to MeSH preferred terms hypertension and risk factors;
3. Eliminating topics with very low MAP (mean average precision) and P10 (precision at top 10 documents) score in the relevance judgment set because these topics would be too difficult;

The selected topics were then randomly ordered to create ten search topic pairs for the experimental conditions (see Figure 1 for search topic pairs).

2.8 Reliability of relevance judgment sets

We measured search outcome using standard precision and recall measures for accuracy and time spent for user effort [6] because we were concerned with the usefulness of MeSH terms on search effectiveness by using TREC assessments [12].

Theoretically speaking, the calculation of recall measure requires relevance judgments from the whole test collection. However, it is almost impossible to obtain these judgments from a test collection with more than 3 million documents. For practical reasons the recall measure used a pooling method that created a set of unique documents from the top 75 documents submitted by 27 groups participated in the TREC 2004 Genomics Track ad hoc tasks [26]. Empirical evidence has shown that recall calculated with a pooling method provides a reasonable approximation, although the recall is likely to be overestimated [33]. But as a result of this approach, there was an average pool size of 976 documents, with a range of 476-1450, which had relevance judgments for each topic [12].

It was quite likely that some of the participants in our experiment would retrieve documents that had not been judged. The existence of un-judged relevant documents, called sampling bias in pooling method, is concerned with the pool depth and the diversity of retrieval methods that may affect the reliability of relevance judgment set [2]. The assumption that the pooled judgment set is a reasonable approximation of complete relevance judgment set may become invalid when the test collection is very large.

To ensure that the TREC pooled relevance judgment set was sufficiently complete and valid for the current study, we analyzed top 10 retrieved documents from each human runs (32 searchers × 8 topics = 256 runs). Cross-tabulation results showed that about one-third of all documents retrieved in our study had not been judged in the TREC data set. More specifically, for a total of 2277 analyzed documents, 762 (33.5 %) had not been assigned relevant judgments. There existed large variations in percentage of un-judged documents for each search topic, with a range of 0–59.3%.

To assess the impact of incomplete relevance judgments, we compared the top 10 ranked search results between the judged document set and the pooled document set for each topic. The judged document set was composed of the documents that matched TREC data, i.e., combination of judged not relevant and judged relevant. The un-judged documents, added to the pooled document set, were considered 'not relevant' in our calculations of search outcome. We used precision oriented measures, MAP (mean average precision), P10 (precision at top 10 documents) and P100 (precision at top 100 documents) to estimate the impact of incomplete judgments.

The paired t-test results by search topic revealed significant differences between the two sets in terms of MAP ($t(19) = -3.69$, $p < .01$), P10 ($t(19) = -3.89$, $p < .001$) and P100 ($t(19) = -3.95$, $p < .001$) measures. The mean of the differences for MAP, P10 and P100 was approximately 2.7%, 9.9% and 4.9% respectively. We concluded that the TREC relevance judgments are applicable to this study.

2.9 Limitations of the design

This study was designed to assess the impact of MeSH terms on search effectiveness in an interactive search environment. One limitation of the design was that participants were a self-selected group of searchers that may not be representative of the population. The interaction effects of selection biases and the experimental variable, i.e., the displayed MeSH terms, were another possible factor that limits the generalizability of this study [4]. The use of relatively technical and difficult search topics in the interactive search environment posed threat to external validity, since those topics might not represent typical topics received by medical librarians in practice.

The internal validity of this design was enhanced by specifically considering several aspects: We devised an incentive system to consider the possible sampling bias of searchers' motivational characteristics in experimental settings. Besides levels of education, participants' domain knowledge was evaluated by a topic understanding test. The variability of search topics was alleviated by using a relatively large number of search topics by experimental design. Selected search topics were intelligible in consultation with domain expert and medical librarian. A concept analysis form was used to help searchers recognize potentially useful terms. The reliability of relevance judgment sets was ensured by additional analysis of top 10 search results from our human searchers.

3. DISCUSSION AND CONCLUSION

The Cranfield paradigm has been very useful for comparing search effectiveness of different retrieval techniques at the level of abstraction that simulates user search performance. Putting users in the loop of IR experiments is particularly challenging because it is difficult to separate the effects of systems, searchers and topics and the search topics have had dominating effects [17]. To alleviate search topic variability in interactive IR experiments, we consider how to increase the topic set size by experimental design within the laboratory IR framework.

This study has demonstrated that a total of 20 search topics can be used in an interactive experiment by Graeco-Latin square balanced design and using different sets of carefully selected topics. We assume that the selected topics are equally difficult since we do not have a good theory of search topics that can directly control the topic difficulty for evaluation purposes. Recent attempts to use reduced topic sets and use non-matched topics (see e.g., [5, 10]) indirectly support our experimental

design considerations of search topic variability and topic difficulty. However, an important theoretical question remains. How can we better control the topic effects in batch and user IR experiments?

4. ACKNOWLEDGMENTS

This study was funded by NSF grant #0414557, PIs. Michael Lesk and Nina Wacholder. We thank anonymous reviewers for their constructive comments.

5. REFERENCES

[1] Banks, D., Over, P. and Zhang, N.-F. 1999. Blind men and elephants: Six approaches to TREC data. Inform Retrieval, 1, 1/2 (April 1999), 7-34. DOI=http://dx.doi.org/10.1023/A:1009984519381

[2] Buckley, C., Dimmick, D., Soboroff, I. and Voorhees, E. 2007. Bias and the limits of pooling for large collections. Inform Retrieval, 10, 6 (December 2007), 491-508. DOI=http://dx.doi.org/10.1007/s10791-007-9032-x

[3] Buckley, C. and Voorhees, E. M. 2005. Retrieval system evaluation. In Voorhees, E. M. and Harman, D. K. (Eds.), TREC: Experiment and Evaluation in Information Retrieval, The MIT Press, Cambridge, MA, 53-75.

[4] Campbell, D. T., Stanley, J. C. and Gage, N. L. 1966. Experimental and Quasi-Experimental Designs for Research. R. McNally, Chicago.

[5] Cattelan, M. and Mizzaro, S. 2009. IR evaluation without a common set of topics. In Proceedings of the 2nd International Conference on the Theory of Information Retrieval (Cambridge, UK, September 10-12, 2009). ICTIR 2009. Springer, Berlin, 342-345. DOI=http://dx.doi.org/10.1007/978-3-642-04417-5_35

[6] Cleverdon, C. W. 1967. The Cranfield tests on index language devices. Aslib Proc, 19, 6 (1967), 173-193. DOI=http://dx.doi.org/10.1108/eb050097

[7] Dumais, S. T. and Belkin, N. J. 2005. The TREC Interactive Track: Putting the user into search. In Voorhees, E. M. and Harman, D. K. (Eds.), TREC: Experiment and Evaluation in Information Retrieval, The MIT Press, Cambridge, MA, 123-152.

[8] Fisher, R. A. 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.

[9] Greenstone Digital Library Software (Version 2.70). 2006. Department of Computer Science, The University of Waikato, New Zealand. Available at: http://prdownloads.sourceforge.net/greenstone/gsdl-2.70-export.zip

[10] Guiver, J., Mizzaro, S. and Robertson, S. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. ACM Trans. Inf. Syst., 27, 4 (November 2009), 1-26. DOI=http://doi.acm.org/10.1145/1629096.1629099

[11] Harman, D. and Buckley, C. 2009. Overview of the Reliable Information Access Workshop. Inform Retrieval, 12, 6 (December 2009), 615-641. DOI=http://dx.doi.org/10.1007/s10791-009-9101-4

[12] Hersh, W., Bhupatiraju, R., Ross, L., Roberts, P., Cohen, A. and Kraemer, D. 2006. Enhancing access to the Bibliome: The TREC 2004 Genomics Track, Journal of Biomedical Discovery and Collaboration, 1, 3 (March 2006). DOI=http://dx.doi.org/10.1186/1747-5333-1-3

[13] Hersh, W. R. 2008. Information Retrieval: A Health and Biomedical Perspective. Springer, New York.

[14] Ingwersen, P. and Järvelin, K. 2005. The Turn: Integration of Information Seeking and Retrieval in Context. Springer, Dordrecht.

[15] Ingwersen, P. and Järvelin, K. 2007. On the holistic cognitive theory for information retrieval. In Proceedings of the First International Conference on the Theory of Information Retrieval (ICTIR) (Budapest, Hungary, 2007). Foundation for Information Society.

[16] Kirk, R. E. Experimental Design: Procedures for the Behavioral Sciences. 1995. Brooks/Cole, Pacific Grove, CA.

[17] Lagergren, E. and Over, P. 1998. Comparing interactive information retrieval systems across sites: The TREC-6 interactive track matrix experiment. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne, Australia, 1998). SIGIR '98. ACM Press, New York, NY, 164-172. DOI=http://doi.acm.org/10.1145/290941.290986

[18] Lu, Z., Kim, W. and Wilbur, W. Evaluation of query expansion using MeSH in PubMed. Inform Retrieval, 12, 1 (February 2009), 69-80. DOI=http://dx.doi.org/10.1007/s10791-008-9074-8

[19] MeSH Browser (2003 MeSH). 2004. U.S. National Library of Medicine. Available at: http://www.nlm.nih.gov/mesh/2003/MBrowser.html

[20] Robertson, S. E. 1981. The methodology of information retrieval experiment. In Sparck Jones, K. (Ed.), Information Retrieval Experiment, Butterworth, London, 9-31.

[21] Robertson, S. E. 1990. On sample sizes for non-matched-pair IR experiments. Inform Process Manag, 26, 6 (1990), 739-753. DOI=http://dx.doi.org/10.1016/0306-4573(90)90049-8

[22] Robertson, S. E., Thompson, C. L. and Macaskill, M. J. 1986. Weighting, ranking and relevance feedback in a front-end system. Journal of Information and Image Management, 12, 1/2, (January 1986), 71-75. DOI=http://dx.doi.org/10.1177/016555158601200112

[23] Sharp, E. C., Pelletier, L. G. and Levesque, C. 2006. The double-edged sword of rewards for participation in psychology experiments. Can J Beh Sci, 38, 3 (Jul 2006), 269-277. DOI=http://dx.doi.org/10.1037/cjbs2006014

[24] Sparck Jones, K. and van Rijsbergen, C. J. 1976. Information retrieval test collections. J Doc, 32, 1 (March 1976), 59-75. DOI=http://dx.doi.org/10.1108/eb026616

[25] Tague-Sutcliffe, J. 1992. The pragmatics of information retrieval experimentation, revisited. Inform Process Manag, 28, 4 1992), 467-490. DOI=http://dx.doi.org/10.1016/0306-4573(92)90005-K

[26] TREC 2004 Genomics Track document set data file. 2005. Available at http://ir.ohsu.edu/genomics/data/2004/

[27] van Rijsbergen, C. J. 1979. Information Retrieval. Butterworths, London.

[28] Voorhees, E. M. 2005. The TREC robust retrieval track. SIGIR Forum, 39, 1 (June 2005), 11-20. DOI=http://doi.acm.org/10.1145/1067268.1067272

[29] Voorhees, E. M. On test collections for adaptive information retrieval. Inform Process Manag, 44, 6 (November 2008), 1879-1885. DOI=http://dx.doi.org/10.1016/j.ipm.2007.12.011

[30] Voorhees, E. M. and Harman, D. K. 2005. TREC: Experiment and Evaluation in Information Retrieval. The MIT Press, Cambridge, MA.

[31] Witten, I. H. and Bainbridge, D. 2007. A retrospective look at Greenstone: Lessons from the first decade. In Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries (Vancouver, Canada, June 18-23, 2007). JCDL '07. ACM Press, New York, NY, 147-156. DOI=http://doi.acm.org/10.1145/1255175.1255204

[32] Witten, I. H., Moffat, A. and Bell, T. C. 1999. Managing Gigabytes: Compressing and Indexing Documents and Images. Morgan Kaufmann, San Francisco.

[33] Zobel, J. 1998. How reliable are the results of large-scale information retrieval experiments? In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Melbourne, Australia, 1998). SIGIR '98. ACM Press, New York, NY, 307-314. DOI=http://doi.acm.org/10.1145/290941.291014