A Large-Scale Comparison of Boolean vs. Natural Language Searching for the TREC-7 Interactive Track

William Hersh, Susan Price, Dale Kraemer, Benjamin Chan, Lynetta Sacherek, Daniel Olson
Division of Medical Informatics & Outcomes Research
Oregon Health Sciences University
Portland, OR, USA

Studies comparing Boolean and natural language searching with actual end-users are still inconclusive. The TREC interactive track provides a consensus-developed protocol for assessing this and other user-oriented information retrieval research questions. We recruited 28 experienced information professionals with library degrees to participate in this year's TREC-7 interactive experiment. Our results showed that this was a highly experienced group of searchers who performed equally well with both types of systems.

Introduction

The goal of the Oregon Health Sciences University (OHSU) TREC-7 Interactive Track experiment was to continue our investigation of Boolean vs. natural language searching. This year our experiments featured a large study sample as well as the use of experienced information professionals with a library degree.

Previous research in comparing Boolean and natural language systems has yielded conflicting results. The first study to compare Boolean and natural language searching with real searchers was the CIRT study, which found roughly comparable performance between the two when utilized by search intermediaries (Robertson and Thompson 1990). Turtle found, however, that expert searchers using a large legal database obtained better results with natural language searching (Turtle 1994). We have performed several studies of medical end-user searching comparing Boolean and natural language approaches. Whether using recall-precision metrics in bibliographic (Hersh, Buckley et al. 1994) or full-text databases (Hersh and Hickam 1995), or using task-completion studies in bibliographic (Hersh, Pentecost et al. 1996) or full-text databases (Hersh, Elliot et al. 1995), the results have been comparable for both types of systems. Our TREC-6 interactive experiments showed a trend towards better results for natural language searching, though the small sample size precluded achievement of statistical significance (Hersh and Day 1997).

The analysis in this paper focuses on system-oriented comparisons. A subsequent paper will focus on user-oriented factors associated with successful searching. This analysis looks specifically at three categories of data:

1. User characteristics
2. Statistical analysis of instance recall
3. Comparison of systems

Methods

The OHSU TREC-7 experiments were carried out according to the consensus protocol as described elsewhere in the proceedings. We used all of the instructions, worksheets, and questionnaires developed by consensus, augmented with some additional instruments. Our experiments compared Boolean and natural language searching systems as used by experienced librarian searchers.

Performance measures

The performance measures used in the TREC-7 interactive track were instance recall and instance precision. The searcher was instructed to look for instances of each topic (e.g., the number of discoveries by the Hubble telescope). Relevance assessors at NIST defined the instances from pooled searching results from all experimental groups. Instance recall was defined as the proportion of true instances identified during a topic, while instance precision was defined as the number of true instances identified divided by all instances identified by a user.

Systems

Both the Boolean and natural language systems were accessed via Web-based interfaces as shown in Figures 1 (Boolean) and 2 (natural language). There was a common IR system behind both interfaces, MG, a publicly available system with Boolean and natural language features (Witten, Moffat et al. 1994). MG was run on a Sun Ultrasparc 140 with 256 megabytes of RAM running the Solaris 2.5.1 operating system. Each interface accessed MG via CGI scripts which contained JavaScript code for logging search strategies, documents viewed (title displayed to user), and documents seen (all of document displayed by user). Searchers accessed each system with either a Compaq DeskPro 200 MHz Pentium Pro machine running Windows 95 and Netscape Navigator 3.0 or an Apple PowerMac 9600 with a 180 MHz PowerPC 604 running System 7.5.5 and Netscape Navigator 3.0. Figure 3 shows the documents displayed for viewing by both interfaces.

Experiments

Subjects were recruited by advertising over the American Society for Information Science Pacific Northwest Chapter listserv. The advertisement explicitly stated that participation would be limited to information professionals with a library degree and participants would be paid a modest remuneration for their participation. As subjects confirmed their participation, they were classified by type of library setting in which they worked: special (e.g., corporate, professional, or scientific), academic, and public.

The experiments took place in a computer lab at OHSU in the first half of August, 1998. An experimental session took four hours, with the first half used for personal data and attributes collection and the second half used for searching. All instructions, worksheets, and questionnaires developed by the consensus process are underlined in the remainder of this section.

The personal data and attributes collection consisted of the following steps:
1. Orientation to experiment (10 minutes)
2. Turn in Pre-Search Questionnaire (distributed by mail and completed before session)
3. Cognitive test administration (40 minutes)
4. Meyers-Briggs Personality Test (15 minutes)
5. Orientation to searching session and both retrieval systems, giving subjects the Searcher Instructions and demonstrating a search to them with each system (20 minutes)
6. Practice search using Hubble telescope search on second page of Searcher Instructions with both systems (10 minutes)

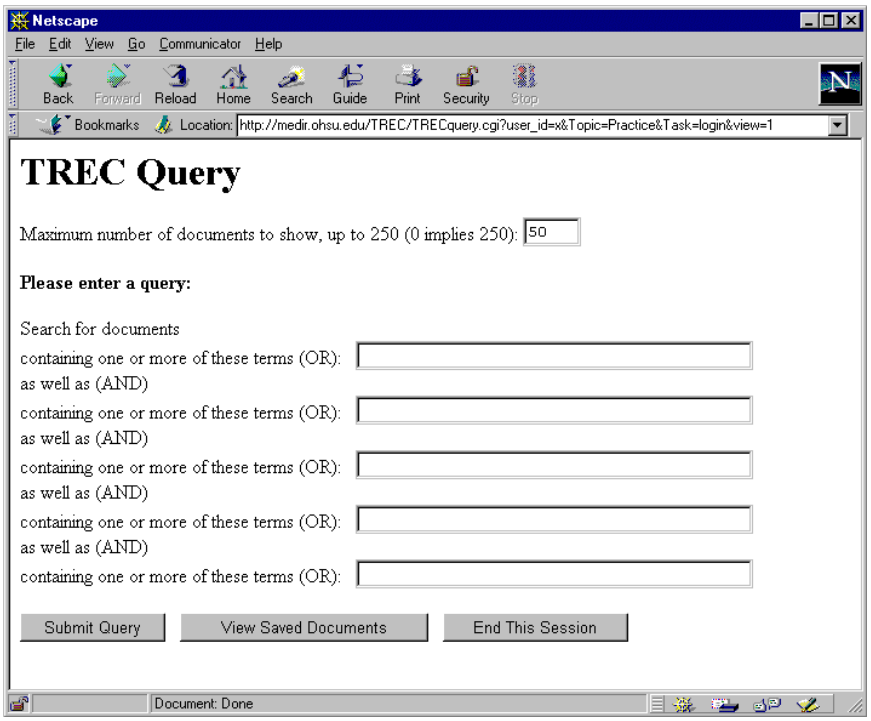

Figure 1 – Boolean searching interface.

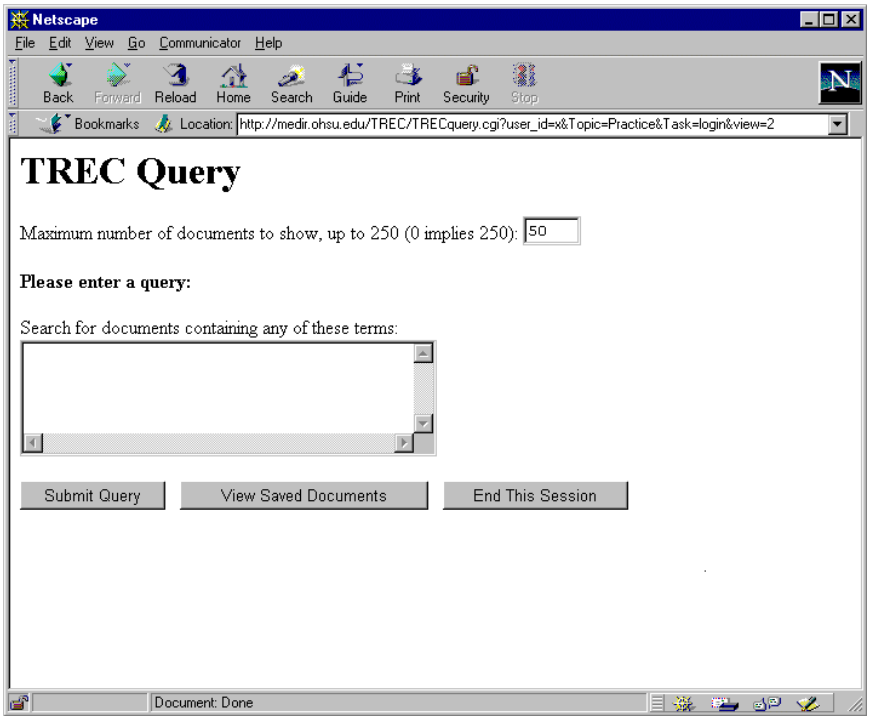

Figure 2 – Natural language searching interface.

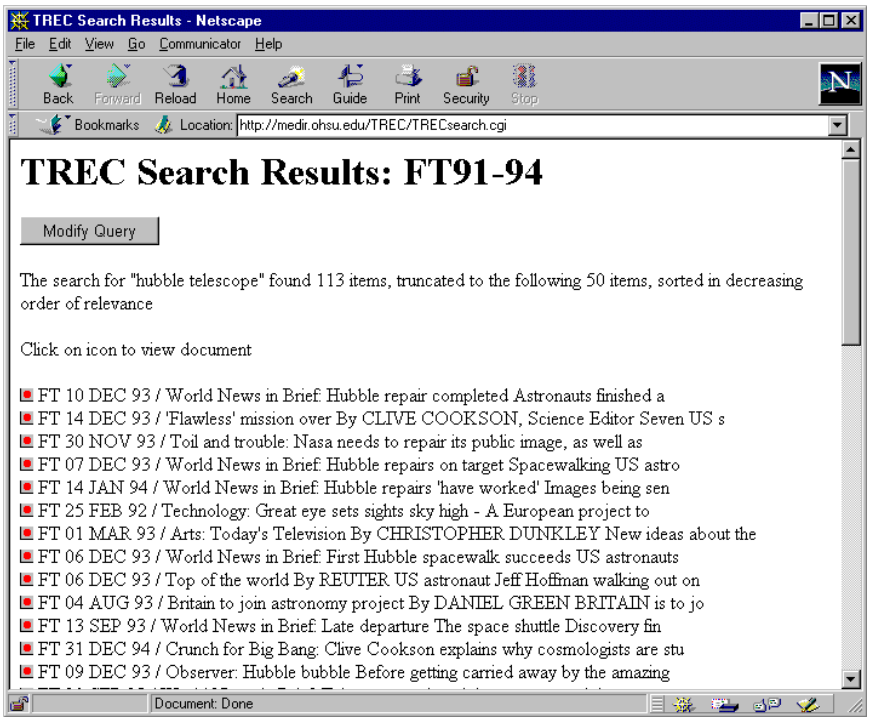

Figure 3 – Documents displayed for viewing.

The cognitive test administration consisted of four cognitive tests from the Educational Testing Service shown in past IR research to be associated with some aspect of successful searching. They included (ETS mnemonic in parentheses):

1. Paper folding test to assess spatial visualization (VZ-2)
2. Nonsense syllogisms test to assess logical reasoning (RL-1)
3. Advanced vocabulary test I to assess verbal reasoning (V-4)
4. Controlled associations test to assess associational fluency (FA-1)

The FA-1 test was used by all TREC-7 interactive groups per the consensus protocol.

The personal data and attributes collection was followed by a 10-15 minute break. The searching portion of the experiment consisted of the following steps:

1. Searching on first 4 topics with assigned system using Searcher Worksheet and Post-Topic Questionnaire (60 minutes)
2. Post-System Questionnaire for system used on first 4 topics and stretch (5 minutes)
3. Searching on second 4 topics with assigned system using Searcher Worksheet and Post-Topic Questionnaire (60 minutes)
4. Post-System Questionnaire for system used on second 4 topics and Exit Questionnaire (5 minutes)

Each subject was randomized into one of four blocks based on system (Boolean vs. natural language) and topic block group. Topic block 1 consisted of topics 365i, 357i, 362i, and 352i. Topic block 2 consisted of topics: 366i, 392i, 387i, and 353i. Users could be assigned as shown in Table 1.

Per the consensus protocol, each subject was allowed 15 minutes per query. Subjects were instructed to identify as many instances as they could for each query. They were also instructed for each query to write

each instance on the Searcher Worksheet and save any document associated with an instance (either by using the "save" function of the system or writing its document identifier down on the Searcher Worksheet).

The Post-System Questionnaire was augmented from the consensus protocol to include the Questionnaire for User Interface Satisfaction (QUIS) 5.0 instrument (Chin, Diehl et al. 1988). QUIS provides a score from 0 (poor) to 9 (excellent) on a variety of user factors, with the overall score determined by averaging responses to each item.

Analysis of Results

As noted above, the analysis of results consisted of three categories of data: user characteristics, statistical analysis of instance recall, and comparison of system factors. User characteristics were obtained from the Pre-Search Questionnaire, cognitive tests, and Meyers-Briggs Personality Test.

The statistical analysis was performed on a modified model of the TREC-6 statistical analysis (Lagergren and Over 1998). The experimental design and, hence, the appropriate analysis of variance model (ANOVA) for our study was more complex than proposed in the TREC-7 instructions and the TREC-6 analyses. One reason for the additional level of complexity was the addition of the factor of type of librarian into our design. In our study, the subjects were recruited from three different classes of librarians (academic, public, and special). Thus, subjects were nested in librarian type.

The design proposed by NIST was also too simple in several other regards. Several additional restrictions were imposed by the way the study was conducted. First, the study design was a modified crossover design. The subject was assigned to use one system on four consecutive topics and then crossed over to the other system on four different topics. The topics were allocated in two "blocks," designated in this analysis as B1 (topics T1 through T4) and B2 (topics T5 through T8). Subjects were also assigned a "sequence." That is, the subject was assigned to do the topics in either B1 or B2 first. The third factor was that of system (Boolean or natural language). The addition of the factors of block and sequence were required by the consensus design defined in the TREC-7 proposal.

There was also an additional restriction imposed by the TREC-7 design. Within each combination of block and sequence, the topics were always assessed in the same order. That is, topic 1 was always followed by topic 2, then 3, and then 4, while topic 5 was always followed by topic 6, then 7, and then 8. Thus there was a particular order of topics. We treated this order as a nested factor within the combination of block by sequence. However, it is important to note that topic was confounded with order. For example, the difficulty of the group of four topics may increase with successive topic number. Thus, the scores could decrease over time. On the other hand, the subjects could become more facile with the system over time and thus the scores might increase over time. Unfortunately, in the design used, topic order is confounded with topic, and thus there is no way to separate the effect of topic order from the effect of topic itself. While the term topic is used in the analyses described below, topic includes both topic itself and topic order.

Table 1 – Block assignment for users.

User	Block search 1	Block search 2
P1	B, Topic Set 1	N, Topic Set 2
P2	N, Topic Set 2	B, Topic Set 1
P3	B, Topic Set 2	N, Topic Set 1
P4	N, Topic Set 1	B, Topic Set 2

The original study design described in the TREC-6 interactive track specified three main effects: search system (SYSTEM), search topic (TOPIC), and study subject (ID) (Lagergren and Over 1998). The highest order model (M4) was defined as follows:

$$Y_{ijk} = \mu_{...} + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + \varepsilon_{ijk}$$

where

Y_{ijk} = recall proportion for subject i, system j, topic k

$\mu_{...}$ = grand mean

α_i = subject effect, $i = 1,\ldots,8$

β_j = system effect, $j = 1,2$

γ_k = topic effect, $k = 1,\ldots,8$

ε_{ijk} = error

For the TREC-7 data, we decided to deviate from the models specified in the report previously cited. The model we developed is much more complex than the model above:

$$\begin{aligned} Y_{ijklmn} = \quad & \mu_{.....} + \alpha_{i(lm)} + \beta_j + \gamma_{k(mn)} + \lambda_l + \eta_m + \theta_n \\ & + (\beta\lambda)_{jl} + (\beta\eta)_{jm} + (\lambda\eta)_{lm} + (\lambda\theta)_{ln} + (\eta\theta)_{mn} \\ & + \varepsilon_{(ijklmn)} \end{aligned}$$

where

Y_{ijklmn} = recall proportion

$\mu_{.....}$ = grand mean

λ_l = librarian type effect, $l = 1,2,3$

η_m = topic sequence effect, $m = 1,2$

$\alpha_{i(lm)}$ = subject effect, nested, random, $i = 1,\ldots,8$

β_j = system effect, $j = 1,2$

θ_n = topic block effect, $n = 1,2$

$\gamma_{k(mn)}$ = topic effect, nested, random, $k = 1,\ldots,8$

$\varepsilon_{(ijklmn)}$ = error

All effects are fixed except subject and topic, $\alpha_{i(lm)}$ and $\gamma_{k(mn)}$, respectively, which are random effects. Table 2 shows an example of the variable coding for four consecutive subjects.

One way to compare the two different models above is to consider the estimable functions for each ANOVA model. The estimable functions are linear combinations of the model parameters (the factor levels in an ANOVA model) which are invariant to the solution of the normal equations (Searle 1971). The model used in TREC-6 includes interactions for which the estimable function includes parameters other than the terms that form the interactions. Specifically, the system-by-topic interaction includes not only system and topic parameters but also subject parameters and the system-by-subject interaction also includes the topic parameters. This means that the test for these interaction terms are, to some extent, confounded with terms not included in the interactions. For example, the system-by-topic interaction is confounded by subject. The estimable functions for interaction terms in the OHSU ANOVA model include only parameters that are included in the interaction terms.

Table 2. Example of variable coding.

Librarian type	Block sequence	Searcher ID	System	Topic Block
Academic	B1-B2	P1-A	Boolean	B1
Academic	B1-B2	P1-A	Nat. Lang.	B2
Academic	B1-B2	P4-A	Nat. Lang.	B1
Academic	B1-B2	P4-A	Boolean	B2
Academic	B2-B1	P2-A	Nat. Lang.	B2
Academic	B2-B1	P2-A	Boolean	B1
Academic	B2-B1	P3-A	Boolean	B2
Academic	B2-B1	P3-A	Nat. Lang.	B1

The system comparison was obtained from collected data. No statistical analysis beyond the above was performed. The factors compared included:

1. Instance recall
2. Instance precision
3. Total number of search terms
4. Documents viewed (system showing title after search)
5. Documents seen (user displaying full document after selecting title)
6. Documents saved (representing an instance)
7. Post-system questionnaire rating of system being easy to learn
8. Post-system questionnaire rating of system being easy to use
9. Post-system questionnaire rating of system begin easy to understand
10. Post-system QUIS for user satisfaction
11. Exit questionnaire of which system was easier to learn
12. Exit questionnaire of which system was easier to use
13. Exit questionnaire of which system was liked better

Results

A total of 24 subjects participated in the study – eight each of special, professional, and academic librarians. All subjects were information professionals with a library degree. All completed the protocol as described above.

Searcher characteristics

The gender breakdown of the 24 subjects analyzed was 16 women and 8 men. The average age of all subjects was 41.1 years. Their average duration they had been doing on-line searching was 7.8 years. Table 3 shows their specific computer experience.

All subjects stated that they searched once or twice daily. All subjects also either agreed (41.7%) or strongly agreed (58.3%) with the statement, “I enjoy carrying out information searches.”

Table 3 – Searchers' computer experience.

Experience with…	1 - No Experience	2	3 - Some experience	4	5 - A great deal of experience
Using a point-and-click interface (e.g., Macintosh, Windows)	0	0	1	2	21
Searching on computerized library catalogs either locally (e.g., local library) or remotely (e.g., Library of Congress)	0	0	0	7	17
Searching on CD-ROM systems (e.g., Encarta, Grolier, Infotrac)	0	1	5	12	6
Searching on commercial on-line systems (e.g., BRS Afterdark, Dialog, Lexis-Nexis)	1	5	6	5	7
Searching on World Wide Web search services (e.g., Alta Vista, Excite, Yahoo, HotBot)	0	1	1	8	14

Statistical analysis of instance recall

With 3 librarian types and 8 subjects per librarian type, there were 24 subjects. The 24 subjects each had 8 observations, one for each topic. Thus the design was balanced, and gave a total of 192 observations. Our model had 43 degrees of freedom, and the r^2 (amount of variance explained) for the model was 0.513.

Librarian type was a marginally significant effect. Pairwise comparisons revealed that special librarians were not significantly different from academic librarians (0.387 versus 0.344, respectively), and academic librarians were not significantly different from than public librarians (0.344 versus 0.302, respectively). However, there were difference between special and public librarians. Topic (nested within block and sequence) was also a significant effect, indicating variation in instance recall across different topics.

Across-system comparison

Table 4 shows comparisons across systems. Instance recall and precision were virtually identical for both systems. The total number of search terms used by subjects was also nearly identical. The number of documents viewed was much larger for the natural language system. The number documents both seen and saved was slightly higher for the Boolean system, though the difference was statistically significant. All of the post-system user satisfaction measures favored the Boolean system at or near statistically significant levels.

Table 5 shows responses from the exit survey. Users were asked their system preference in terms of ease of learning, ease of use, and overall preference. The Boolean system was clearly preferred on all three measures.

Table 4 – System effects ANOVA models.

Factor	Least squares mean		*p*-value	Model r^2
	Boolean	Nat. Lang.		
Instance Recall	0.346	0.342	0.8854	0.513
Instance Precision	0.688	0.698	0.7822	0.458
Total Search Terms	6.59	6.15	0.2815	0.617
Documents Viewed	141.1	241.4	**0.0004**	0.452
Documents Seen	14.68	13.58	**0.0641**	0.590
Documents Saved	5.32	4.65	**0.0543**	0.605
Post-System Easy to Learn (1–5) *	3.45	2.92	**0.0850**	0.808
Post-System Easy to Use (1–5) *	2.95	2.13	**0.0073**	0.868
Post-System Easy Understand (1–5) *	3.42	3.04	0.1310	0.864
Post-System QUIS average *	4.61	4.09	**0.0007**	0.969

* The nature of this varaible requires the exclusion of the Topic effect.

Table 5 - Exit survey responses, *n* = 23.

Factor	Boolean	Nat. Lang.	*p*-value
Users Said Easier to Learn	17	6	0.0347
Users Said Easier to Use	19	4	0.0026
Users Said Liked Better	19	4	0.0026

Discussion

This experiment assessed the ability of highly experienced information professionals to identify instances of topics in an on-line database. The presearch questionnaire showed they had a great deal of searching experience and computer experience in general. They performed online searching and carried it out on a daily basis. These subjects strongly preferred the Boolean searching interface, although they used about the same number of search terms and chose the same number of documents for seeing. Their instance recall and precision were virtually identical, indicating their searching performance with each was comparable.

There were some other observations of interest revealed by this study. First, subjects used slightly more than 6 unique search terms per query. The most notable aspect of this result is that it is about three times higher than the average number of terms used by general users of Web search engines (Jansen, Spink et al. 1998). Second, there were differences based upon the type of library position in which they worked. Those from special libraries did better as a group than those from academic libraries, while the latter outperformed those from public libraries.

Further analysis will change the orientation from a system-oriented to a user-oriented perspective. We will look at association of experience attributes, cognitive factors, personality types, and system operations with performance measures. We will also compare document-oriented recall and precision with the instance-oriented measures used in this study. In the future, we hope to compare other advanced retrieval systems and their features using the TREC-7 interactive track technique.

References

Chin, J., V. Diehl, et al. (1988). Development of an instrument measuring user satisfaction of the human-computer interface. *Proceedings of CHI '88 - Human Factors in Computing Systems*, New York, ACM Press, 213-218.

Hersh, W., C. Buckley, et al. (1994). OHSUMED: an interactive retrieval evaluation and new large test collection for research. *Proceedings of the 17th Annual International ACM SIGIR*, Dublin, Springer-Verlag, 192-201.

Hersh, W. and B. Day (1997). A comparison of Boolean and natural language searching for the TREC-6 interactive task. *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, Gaithersburg, MD, NIST, 585-595.

Hersh, W., D. Elliot, et al. (1995). Towards new measures of information retrieval evaluation. *Proceedings of the 18th Annual International ACM SIGIR*, Seattle, WA, ACM Press, 164-170.

Hersh, W. and D. Hickam (1995). An evaluation of interactive Boolean and natural language searching with an on-line medical textbook. *Journal of the American Society for Information Science* 46: 478-489.

Hersh, W., J. Pentecost, et al. (1996). A task-oriented approach to information retrieval evaluation. *Journal of the American Society for Information Science* 47: 50-56.

Jansen, B., A. Spink, et al. (1998). Real life information retrieval: a study of user queries on the Web. *SIGIR Forum* 32: 5-17.

Lagergren, E. and P. Over (1998). Comparing interactive information retrieval systems across sites: the TREC-6 interactive track matrix experiment. *Proceedings of the 21st Annual ACM SIGIR*, Melbroune, Australia, ACM Press, 162-172.

Robertson, S. and C. Thompson (1990). Weighted searching: the CIRT experiment. *Informatics 10: Prospects for Intelligent Retrieval*, York, ASLIB, 153-166.

Searle, S. (1971). Linear Models. New York, J. Wiley and Sons.

Turtle, H. (1994). Natural language vs. boolean query evaluation: a comparison of retrieval performance. *Proceedings of the 17th Annual International ACM SIGIR*, 212-220.

Witten, I., A. Moffat, et al. (1994). Managing Gigabytes - Compressing and Indexing Documents and Images. New York, Van Nostrand Reinhold.