Boolean System Revisited:
Its Performance and its Behavior

X. Allan Lu*, John D. Holt, David J. Miller

Lexis-Nexis, a division of Reed Elsevier Inc.
9555 Springboro Pike
Miamisburg, Ohio 45342

Telephone: 513-865-6800 ext. 5404
Fax: 513-865-1655
alan, johnh,davidm@lexis-nexis.com

ABSTRACT

This experimental study attempts to provide a general conclusion to the Boolean information retrieval system regarding its performance on retrieval effectiveness and its behavior on retrieval of different relevant documents. Specifically a representative commercial Boolean system is compared to the best ranking systems reported in TREC4. The Boolean system delivered a comparable performance and retrieved a set of rather unique relevant documents. The study also justifies its methodology for comparing non-ranking and ranking systems.

1.0 Introduction

Boolean search has been dominating the commercial information retrieval for over 30 years and a large number of professional searchers have mastered the Boolean search tool. The size of the Boolean searcher community forces the latest information retrieval software vendors to provide this search feature. But the Boolean search is an aged search technique. The criticism to the Boolean search system range from its requirement of intensive user training to its awkwardness in expressing complex conceptual relationships in queries, and to the absence of relevance ranking. The newly commercialized natural language search systems, like Freestyle™, are free from these problems and enable the professionals as well as the end-users to access the large commercial full-text databases.

An interesting question arises: should the Boolean search system be replaced entirely by the natural language search system that has relevance ranking capability? A positive answer to this question requires that the natural language search system is better than the Boolean system in terms of retrieval effectiveness, and that the natural language search system is capable of finding a similar set of relevant documents that the Boolean system tends to find. The motivation for having a better performance is obvious, the motivation for retrieving a similar set of relevant documents is to demonstrate to the user the redundancy of the Boolean system.

The information retrieval literature in the past suggests that in terms of search query formulation the simple relevance ranking system that accepts natural language queries usually performs as good as the Boolean system that provides complex querying language [1,2,3,4,5,6,7]. However these studies used very small test databases and their results suffered the lack of generalizability. A more recent study [8] using much larger test database suggested that the relevance ranking system could perform better than the Boolean system. Unfortunately the study used only legal data instead of more general materials, and employed only one type of relevance ranking system. In addition the study had the methodological issue of comparing relevance ranking to chronological ordering that definitely favored the relevance ranking system. Subject relevance and subject currency are two different relevance factors. The Boolean searcher attempts to incorporate both factors in the search in order to read recent and relevant documents first. By contrast the ranking system can only handle the single factor--subject relevance, indicating that the traditional ranking principle is rather rigid and inflexible in the environment of manmachine interaction. As a result, the ranking system may maintain a better performance by ranking the old and relevant document higher than the recent and relevant while the Boolean system may suffer given the same ranking sequence. For a better experimental design it is prudent to retreat to only compare the density of retrieved relevant documents in the Boolean and ranking systems, leaving the issue of determining which relevance factor, subject relevance or subject currency, is more important to the searcher.

These reported studies also failed in looking into the retrieved documents to examine whether the retrieved from different systems are similar or significantly different regardless their recall and precision performances. When there are many relevant documents available the systems with comparable performance may have different retrieval behavior, that is, they retrieve different sets of the relevant documents. Specifically, the Boolean searcher would be reluctant to migrate to the natural language search system if the Boolean system tends to find different relevant documents. Thus it is desirable to look beyond the traditional effectiveness measures when comparing the Boolean system and the ranking system.

TREC4 [9] presents an excellent opportunity to conduct a thorough comparison of the Boolean system and the relevance ranking system. First of all, the test data is sufficiently large and diverse, putting adequate stress on the involved systems. Particularly neither the Boolean search subject nor the ranking system designer is able to guess the amount of relevant documents in the test databases. Second, the test queries are much shorter than those used in the first three TRECs, with the average size (i.e., number of searchable words) close to the query size in the commercial retrieval environment [10]. Without the short queries the experimental design will favor the relevance ranking system and punish the Boolean system. Third, the search instructions designed for the interactive search track in TREC4 prevent any Boolean search subject from creating an answer set that is either very large to boost recall to an artificially high level or extremely small to ensure unrealistically high precision. The instructions are recall as well as precision neutral. The search time of thirty minute also prevent any Boolean search subject from constructing very knowledgeable Boolean queries, making the Boolean search similar to those performed in the real situations. Fourth and finally, the good systems identified in the ad hoc tests represent a group of different relevance ranking systems, making more general observations feasible. Better yet, the ad hoc search with manual query expansion in TREC4 is very similar to the Boolean search from the search procedure point of view.

2.0 Experiment Design

The Boolean search results are the results that Lexis-Nexis submitted to the interactive search track in TREC4. Briefly, each of the 10 Boolean search subjects, the Boolean search experts from the customer services department, was given 5 of the 50 TREC4 ad hoc topics in the sequence from 201 to 250. The topic assignment ignores any specific subject expertise on the part of any given subject. The databases for the ad hoc tests in TREC3 and TREC4 were loaded into the Lexis-Nexis online system for the subjects to access them through a Boolean querying interface. The interface provides every Boolean search feature that one can find in the commercial online systems, including word truncation and proximity. Some subjects interacted with the TREC3 queries and data for getting familiar with the TREC environment. The subjects were instructed to complete each search topic within 30 minutes and to create an answer set that is fairly precise and comprehensive as they usually do for their customers. The test searches were carried out by the subjects, in their offices, on a time available basis. In most cases, the topics were begun and finished without interruption. However, in some cases, a topic was saved, while business demands had to be met. This environment accounts for some of the lengthy times associated with a few topics.

For the purpose of this study the Lexis-Nexis entry for the interactive track did a few tasks that were neither required nor instructed. For maintaining the sample size the Lexis-Nexis entry did all 50 topics instead of the required 25 for the interactive search task. Next, unlike other interactive search entries in TREC4 the final answer set for every test topic represents the complete document set from a single Boolean search without reviewing each individual retrieved document. That is, the subjects did not manually construct the final answer set to include those documents deemed relevant by them, or to exclude those documents deemed not relevant by them. This is the major reason why the sizes of the answer sets have a wider range (Table 1) than those from other interactive systems. This is also the reason that the Boolean search results are more comparable to those of the manual ad hoc in TREC4 where the designer of the ranking system manually built the final search queries.

Finally the number of interactions with the system is very limited and in some cases is zero. The limited number of interaction with the system enhance the comparability of the Boolean search results to those from the ranking systems.

Topic	Retrieved	Topic	Retrieved
201	87	226	17
202	103	227	59
203	19	228	223
204	111	229	42
205	34	230	14
206	26	231	88
207	42	232	33
208	46	233	34
209	76	234	16
210	47	235	99
211	150	236	16
212	63	237	181
213	31	238	65
214	10	239	40
215	18	240	288
216	32	241	17
217	19	242	61
218	5	243	21
219	21	244	77
220	18	245	23
221	247	246	6
222	56	247	53
223	148	248	28
224	47	249	20
225	127	250	164

TABLE 1. Boolean Answer Set Sizes

The Boolean search results can be summarized as follows: The range of search time is from 4 and half minutes to 143 minutes, the average search time is 38 minutes, and the median search time is 30 minutes. The average micro-precision and micro-recall [11] are 0.4 and 0.17, respectively, and the average macro-precision and macro-recall [9] are 0.38 and 0.22, respectively. The composed E measure with alpha=0.5 is 0.76.

The relevance ranking systems in this study are the four best performers in the automatic and manual ad hoc tests [6]. Since both the Boolean system and the manual ad hoc systems had human involvement, their results are more comparable. The automatic ad hoc systems were included and compared for drawing a general performance conclusion.

The four best automatic ranking systems are “CrnlAE” from Cornell University, “pircs1” from Queen’s College, “citya1” from City University of London, and “INQ201” from University of Massachusetts. Later another system “CrnlAL” from Cornell University was added to this group. The four best manual systems are “CnQst2” from Excalibur Technologies Inc., “pircs2” from Queens College, “uwgcl1” from University of Waterloo, and “INQ202” from university of Massachusetts.. For each of the selected ranking systems its 50 rank lists submitted to NIST for scoring were used for comparison. The query relevance table was also used for extracting the relevant documents from the rank lists.

To make the Boolean search results comparable to the ranking results, this study turned every relevance rank list into a document set using the size of the corresponding Boolean answer set generated for the same query [5]. The major benefit of this method is its focus on the process of search, leaving out the process of presenting search results and the associated relevance factors such as subject relevance and subject currency, and allowing a more static comparison with a better experimental control. The other arguments in the introduction section regarding the favorable TREC4 environment further reduce the potential bias usually associated with this method.

The actual design can be easily represented in a form of a matrix with its rows as queries and its columns as various systems, Boolean or ranking. Each cell in the matrix has two items, precision ratio and recall ratio calculated for a particular system and a particular query. Since there are 49 (Topic 201 was removed from the test by NIST) queries and 10 systems, the matrix has 490 cells or 490 pairs of precision and recall ratios. The populated matrix invites a typical one-way ANOVA analysis with both precision and recall as dependent variables and system as independent variable. The results from the one-way ANOVA test should indicate whether there is any significant performance difference among the 10 systems. When such difference is detected, the further analysis like Tukey's Studentized Range (HSD) Test or pairwise T-test can be employed to find out the system(s) that is different from the rest. MANOVA test can also be applied to this matrix to obtain the multivariate statistics of the two dependent variables.

This study takes an additional step to investigate whether the different systems tend to retrieve different documents or more specifically different relevant documents in a quantitative term. For this task a matrix with the same dimensions as the precision-recall matrix described in the preceding paragraph is populated with the number of unique documents retrieved and with the number of unique relevant documents retrieved in each cell. Once the populated matrix is ready, the same one-way ANOVA test and the same MANOVA test can be applied to it, with both the number of unique documents and the number of unique relevant documents as dependent variables and system as independent variable. The test results should provide the insight as to whether the systems tend to retrieve a similar amount of unique documents and unique relevant documents.

A definition for unique documents and unique relevant documents is in order. Initially the unique document to a system was defined as a document that was only retrieved by that system. The same definition applied to the unique relevant documents. Since the Boolean system is more comparable to the manual ad hoc systems, this study splits the matrix into two, the first behavior matrix having the Boolean system and the 4 manual ad hoc systems, and the second behavior matrix having the Boolean system and the 4 automatic ad hoc systems.

Note that a system with a capability of retrieving many unique documents is not necessarily a better system in terms of precision and recall. A system with such a behavior is merely a different system, and it may be helpful to the searchers in their recall-oriented research work.

3.0 Results

3.1 Retrieval Effectiveness

The one-way ANOVA test results from analyzing the precision-recall matrix suggest that there is no statistical difference among the 10 systems in terms of either precision measure or recall measure [Table 2]. The ANOVA test was repeated with the data transformed first using square root function and second using the inverse trigonometric sine function to take care the potential problem of unequal variances usually associated with percentage figures. The results from the repeated test do not suggest anything new. The results from

CLASS	**LEVELS**				
system	**10**				
Dependent:	precision				
Source	**DF**	**Anova SS**	**mean sqr**	**F value**	**Pr > F**
system	9	.507059	.056340	.80	.6198
Dependent:	recall				
Source	**DF**	**Anova SS**	**mean sqr**	**F value**	**Pr > F**
system	9	.131306	.014589	.39	.9408
Manova test:	no system effect				
		S = 2	M = 3	N = 238.5	
Statistic	**Value**	**F**	**Num DF**	**Den DF**	**Pr > F**
Wilks'	.982010	.4853	18	958	.9649
Roy's	.016554	.8829	9	480	.5404

TABLE 2. ANOVA analysis on variable Precision and variable Recall

multivariate analysis of variance that took both precision and recall into account also suggest that there is no statistical difference among the 10 systems.

The 2 Tukey Groupings generated by Tukey's Studentized Range Test for the variable precision and the variable recall contain only one group that covers every system, suggesting, again, that there is no significant difference among the 10 systems. The mean precisions and the mean recalls copied from the 2 Tukey Groupings are listed in [Table3] for the 10 systems.

System	N	Precision	Recall	System	N	Precision	Recall
CnQst2	49	.2763	.4672	uwgcl1	49	.2305	.4199
INQ202	49	.2489	.4226	Boolean	49	.2281	.4139
pircs2	49	.2431	.4000	citya1	49	.2223	.3690
CrnlAE	49	.2404	.4026	INQ201	49	.2201	.3481
CrnlAL	49	.2329	.3951	pircs1	49	.2195	.3676

TABLE 3. Mean precisions and Mean recalls from the 2 Tukey's tests

3.2 Retrieval Behavior

The results from analyzing the retrieval behavior of the Boolean and automatic ad hoc systems suggest that the systems are different. As illustrated in [Table 4] some systems in the first behavior matrix retrieved more uniquely relevant documents as well as more unique documents than other systems in the group. The examination to [Table 5] reveals that it is the Boolean system that made the difference. The Boolean system is statistically different from any automatic ad hoc system in the group in retrieving more uniquely relevant documents. Read the mean column in [Table5] for a quick review.

CLASS **system**	**LEVELS** **5**				
Dependent:	UNIQUELY REL				
Source	**DF**	**Anova SS**	**mean sqr**	**F value**	**Pr > F**
system	4	1827.73	456.93	8.91	.0001
Dependent:	UNIQUE				
Source	**DF**	**Anova SS**	**mean sqr**	**F value**	**Pr > F**
system	4	15677.41	3919.35	2.76	.0284
Manova Test:	no system effect				
		S = 2	M = 0.5	N = 118.5	
Statistic	**Value**	**F**	**Num DF**	**Den DF**	**Pr > F**
Wilks'	.854631	4.8821	8	478	.0001
Roy's	.149153	8.9492	4	240	.0001

TABLE 4. ANOVA analysis on variable Uniquely Relevant Document and variable Unique Document in the first behavior matrix

Alpha = 0.5	**DF = 240**	**MSE = 51.27**	
Critical value of studentized range = 3.887	**Minimum significant difference = 3.976**		
Means with the same letter are not significantly different			
Tukey Grouping	**Mean**	**N**	**System**
A	9.612	49	boolean
B	3.837	49	CrnlAE
B	3.224	49	pircs1
B	2.714	49	citya1
B	2.020	49	INQ201

TABLE 5. Tukey's Studentized range test for variable Uniquely Relevant Document

For the variable unique document, a similar examination to [Table 6], however, reveals a slightly different picture. It is that the Boolean system only retrieved more unique documents than "CrnlAE" did, the Boolean system retrieved a similar number of unique documents as the three other systems, namely, "INQ201" "cityal" and "pircs1."

Alpha = 0.5		**DF = 240**	**MSE = 1419.1**	
Critical value of studentized range = 3.887		**Minimum significant difference = 20.919**		
Means with the same letter are not significantly different				
Tukey Grouping		**Mean**	**N**	**System**
	A	39.878	49	boolean
B	A	25.939	49	INQ201
B	A	22.061	49	citya1
B	A	21.776	49	pircs1
B		16.122	49	CrnlAE

TABLE 6. Tukey's Studentized range test for variable Unique Document

Turning to the second behavior matrix consisting of the Boolean system and the manual ad hoc systems, the picture is more complicated than the one described in the preceding paragraph for the automatic ad hoc systems. As before the results from the general one-way ANOVA analysis reported in [Table 7] for the variable uniquely relevant document and the variable unique document suggest that there is at least one system in the current matrix that is significantly different from the rest.

The voting in multiple comparison in [Table 8] is that the Boolean system retrieved more uniquely relevant documents than "INQ202" did, and both "uwgcl1" and "CnQst2" retrieved more uniquely relevant documents than "INQ202" did. "pircs2" is apparently neutral and is indifferent from any other system. In terms of retrieving just unique documents the Tukey grouping in [Table 9] fails in indicating which system(s) contributed to the significant difference detected in [Table 7]. It is only the pairwise T-tests in [Table 10] that suggest that the Boolean system is more productive in finding unique documents than "pircs2" and "INQ202", and "uwgcl1" is more productive than "INQ202". The tests in [Table 9-10] still suggest that the Boolean system is different from at least some of the manual ad hoc systems in retrieving unique documents for the TREC4 ad hoc queries.

CLASS	**LEVELS**				
system	**5**				
Dependent:	UNIQUELY REL				
Source	**DF**	**Anova SS**	**mean sqr**	**F value**	**Pr > F**
system	4	1000.68	200.136	4.29	.0009
Dependent:	UNIQUE				
Source	**DF**	**Anova SS**	**mean sqr**	**F value**	**Pr > F**
system	4	10719.82	2143.964	2.43	.0353
Manova Test:	no system effect				
		S = 2	M = 1	N = 142.5	
Statistic	**Value**	**F**	**Num DF**	**Den DF**	**Fr > F**
Wilks'	.918748	2.4844	10	547	.0064
Roy's	.079987	4.6073	5	288	.0005

TABLE 7. ANOVA analysis on variable Uniquely Relevant Document and variable Unique Document in the second behavior matrix

			Alpha = 0.5	**DF = 240**	**MSE = 46.676**
Critical value of studentized range = 4.057			**Minimum significant difference = 3.96**		
Means with the same letter are not significantly different					
Tukey Grouping			**Mean**	**N**	**System**
	A		6.429	49	boolean
B	A		6.245	49	uwgcl1
B	A		6.082	49	CnQst2
B	A	C	3.735	49	pircs2
		C	2.082	49	INQ202

TABLE 8. Tukey's Studentized range test for variable Uniquely Relevant Document

	Alpha = 0.5	**DF = 240**	**MSE = 882.37**
Critical value of Studentized range = 4.057	**Minimum significant difference = 17.218**		
Means with the same letter are not significantly different			
Tucky Grouping	**Mean**	**N**	**System**
A	32.020	49	boolean
A	30.796	49	uwgcl1
A	24.245	49	CnQst2
A	19.000	49	pircs2
A	16.959	49	INQ202

TABLE 9. Tukey's Studentized range test for variable Unique Document

Alpha = 0.5; Confidence = 0.95; DF = 240	**MSE = 883.37**	
Critical value of T = 1.96824		
Least significant difference = 11.812		
Comparisons significant indicated by * **	**Difference between Means**	
boolean - uwgcl1	1.224	
boolean - CnQst2	7.776	
boolean - pircs2	13.020	***
boolean - INQ202	15.061	***
uwgcl1 - CnQst2	6.551	
uwgcl1 - pircs2	11.796	
uwgcl1 - INQ202	13.837	***
.... no other comparisons are significant		

TABLE 10. T test (LSD) on variable Unique Document

4.0 Conclusion

A study of comparing the Boolean retrieval system to the general ranking systems was carried out during and after TREC4. The conference provided a rare opportunity for constructing a sound experiment design to compare these two types of retrieval systems. The design benefited from 1) the large and diverse test databases, 2) the short query topics, 3) the interactive search instructions for Boolean search, and 4) the list of the best performed ranking systems for comparison. The design used the existing cutoff method, based on the size of a Boolean document set to turn the ranking document list to a document set for comparing it to the Boolean set. This method separates retrieval and presenting since the former can be studied statically while the latter needs a situation of information need to justify the effectiveness of a particular presentation scheme such as subject relevance or subject currency. The discussion of the design suggests that for the purpose of presenting retrieved documents subject relevance is merely one of many different relevance dimensions, and that with the possibility of presenting the retrieved along different relevance dimensions, the traditional relevance ranking principle appears rigid and inflexible.

The test results from this static design suggest that the Boolean system could have a similar performance as the best relevance ranking systems in the environment of TREC4. A closer examination on mean precision and mean recall, however, suggest that the selected ranking systems, especially the manual ad hoc systems did slightly better, but not statistically significant. The results from the one-way ANOVA tests and the associated MANOVA tests are very consistent throughout.

Unlike the other reported research work, this study took a step further to examine the retrieval behavior of the selected systems. The behavior was defined as the capability of a system in retrieving uniquely relevant documents and unique documents for a query. The Boolean searcher may hesitate to only rely on the ranking system if the Boolean system maintains such a capability. The results from the behavior tests do suggest that the Boolean system tends to retrieve a set of relevant documents that is significantly different from the relevant set retrieved by the selected ranking systems. On the other hand, the unique document set that the Boolean system tends to create is not that distinct from the unique set that the ranking systems tend to build, especially when the ranking systems are those manual ad hoc systems in TREC4.

5.0 References

[1] Aitchison, T. M. Et al. (1970) "Comparative evaluation of indexing languages, Part II: Results," Report R70/2, INSPEC, Institution of Electrical Engineers, London.

[2] Cleverdon, C. W. (1970) An investigation into a suitable mechanised information retrieval system at the defense operational analysis establishment, Cranfield Institute of Technology.

[3] Miller, W. L. (1971) "The efficiency of MEDLARS titles for retrieval," Journal of the American Society for Information Science, Vol.22, 318-321.

[4] Miller, W. L. (1971) "A probabilistic search strategy for MEDLARS," Journal of Documentation, Vol.27, 254-266.

[5] Salton, G. (1972) "A new comparison between conventional indexing (MEDLARS) and automatic text processing (SMART)," Journal of the American Society for Information Science, Vol.23, 75-84.

[6] Evans, L. (1975) "Search strategy variations in SDI profiles," Report R75/21, INSPEC, Institution of Electrical Engineers, London.

[7] Evans, L. (1975) "Methods of ranking SDI and IR outputs," Report R75/23, INSPEC, Institution of Electrical Engineers, London.

[8] Turtle, H. (1994) "Natural language vs. Boolean query evaluation: A comparison of retrieval performance," Proceedings of the 17th Annual International ACM-SIGIR conference, 212-221, Dublin, Ireland.

[9] Harman, D. K., editor. (1996) Overview of the Fourth Text REtrieval Conference (TREC-4), to be published by NIST.

[10] Lu, X. A. and Keefer, R. B. (1995) "Query expansion/reduction and its impact on retrieval effectiveness," Overview of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225, edited by D. K. Harman, 231-240.

[11] Tague, J. M. (1981) "The pragmatics of information retrieval experimentation," Information Retrieval Experiment, Chapter 5, 59-102, edited by K. S. Jones, Butterworth.