Analysis of Performance Variation Using Query Expansion

Nega Alemayehu
National Institute of Standards and Technology, 100 Bureau Drive, Stop 8940, Githersburg, MD 20899.
E-mail: nega.alemayehu@nist.gov

Information retrieval performance evaluation is commonly made based on the classical recall and precision based figures or graphs. However, important information indicating causes for variation may remain hidden under the average recall and precision figures. Identifying significant causes for variation can help researchers and developers to focus on opportunities for improvement that underlay the averages. This article presents a case study showing the potential of a statistical repeated measures analysis of variance for testing the significance of factors in retrieval performance variation. The TREC-9 Query Track performance data is used as a case study and the factors studied are retrieval method, topic, and their interaction. The results show that retrieval method, topic, and their interaction are all significant. A topic level analysis is also made to see the nature of variation in the performance of retrieval methods across topics. The observed retrieval performances of expansion runs are truly significant improvements for most of the topics. Analyses of the effect of query expansion on document ranking confirm that expansion affects ranking positively.

Introduction

Query expansion (QE) has been studied in a variety of ways. Researchers have tried to show the performance improvements of automatic query expansion techniques. For instance, substantial improvements in performance due to QE are reported for the INQUERY (Xu & Croft, 1996, 2000) and the SMART (Buckley, Singhal, & Mitra, 1997; Mira, Singhal, & Buckley, 1998) retrieval systems. It is argued that expansion generally improves performance in large-scale collections (Beaulieu, Robertson, & Rasmussen, 1997; Kekalainen & Jarvelin, 1998; Mitra, Singhal, & Buckley, 1998; Xu & Croft, 1996). Most of the studies, however, are based on comparison of average performance figures, such as precision and recall. Although such methods may show the improvements, such observed differences are not always significant. Moreover, important information that can be of help in the design process may remain hidden beyond the average figures. Little has been done in the detailed analysis of performance data and the search for causes of variation compared to designing new expansion techniques. This is a common problem in information retrieval (IR) research, as indicated by Hull (1996).

The application of statistical techniques to retrieval performance data can provide information that is hidden beyond mean average precision and recall. Retrieval performance depends on different factors, such as, system (indexing and searching), difficulty of information need, and quality of query. The knowledge on how these factors and their interactions affect retrieval performance will help researchers and developers in designing or modifying methods. In this article, an analysis of query expansion is presented. The analysis method shows how the factors affecting performance can be tested for significance.

A brief discussion on the subject of QE is given in the next section as background information. The reader, however, must not expect this article to be a work on query expansion. It is rather an article that looks at the variations in the performances of systems and runs with and without query expansion. A statistical analysis of variance on the performances of a few retrieval systems who took part in the TREC-9 Query Track, in general, and query expansion runs against nonexpansion runs, in particular, is presented.

Query Expansion: The Problem

The increasing complexity of information items, on one hand, and the changing needs of users, on the other, require researching and designing more effective methods of storage and retrieval. IR addresses problems in the areas of information need, on one side, and information items, on the other. Representations of information need and information items affect the whole retrieval process. Accordingly, information need representation has been and is still one of the

Received October 13, 2001; revised October 7, 2002; accepted October 7, 2002

TABLE 1. Text collection used in the TREC-9 query track.

Title	Size (MB)	#Docs	#Words/Doc	
			Median	Mean
Wall Street Journal, 1987–1989	267	98,732	245	434.0
Associated Press newswire, 1989	294	84,678	246	473.9
Computer Select Articles, Ziff-Davis	242	75,180	200	473.0
Federal Register, 1989	260	25,960	391	1315.9
Abstracts of U.S. DOE publications	184	220,087	120	120.4

most important problems in IR. Mismatch between words representing information need and words that the authors use in their documents causes a major problem. Query expansion and/or reformulation is undertaken when an original query fails to retrieve wanted information items, that is, when the original query fails to incorporate terms which authors use to discuss concepts (Efthimiadis, 1996).

On-line information users mostly tend to use few words to express their information needs. Spink, Wolfram, Jansen, & Sarecevic (2000), for instance, show that Web users tend to express their information needs with a few words, mostly two or three. The shorter a query, the less probable it is to match against words of documents. Therefore, there is a need to support such queries to increase the likelihood of matching.

Query expansion can be manual, interactive (semimanual), or automatic (Efthimiadis, 1996). In manual procedures, a user adds and/or removes terms to/from a query. In the semimanual expansion, queries are modified based on judged relevant and/or irrelevant documents—a process known as relevance feedback. On the other hand, automatic query expansion procedures do not involve the user once a query is submitted—it assumes that the first few documents in the ranked result set are relevant (pseudofeedback). Harman (1992a) gives a detailed account of relevance feedback and other query reformulation techniques.

Additional information on approaches to QE can be found in the literature (Beaulieu et al., 1996; Carpineto, de Mori, Romano, & Big, 2001; Efthimiadis, 1996; Ekmekcioglu, Robertson, & Willett, 1992; Harman, 1988, 1992a, 1992b; Mandala, Tokunaga, & Tanaka, 1999; Mitra et al., 1998; Peat & Willett, 1991; Qiu & Frei, 1993; Xu & Croft, 1996; *inter alia*). The application of relevance feedback for query expansion has been experimented within different retrieval systems.

There have been many more efforts on designing and/or improving QE methods than detailed analyses of performance data as indicated above. It should be noted, however, that there were some efforts to analyze QE performance data. Kekalainen and Jarvelin (1998), for example, use a nonparametric statistical method for the analysis of expansion and nonexpansion performance data. Nonparametric methods are preferred when assumptions on the nature of the distribution of the study variable could not be met. However, the method is also susceptible to type II errors, that is, rejecting true differences as nonsignificant (Tague-Sutcliff & Blustein, 1995). The works by Banks, Over, and Zhang (1999), Buckley and Walz (2000a), Hull (1996), Tague-Sutcliff and Blustein (1995), although not directly related to QE performance data analysis, are worth mentioning. They use statistical analysis of variance techniques to show retrieval performance differences and their approach can be used to analyze QE performance data.

In this article, analyses of the retrieval performance of systems in the TREC-9 Query Track is presented. A parametric analysis of variance method is used. The following section describes the Track: its design, test collection, retrieval systems participated and the performance measure used. Later, results of the analysis experiments, including the analysis method used, are presented. Then the conclusions are presented.

The Query Track at TREC-9

In the language of the query track, a TREC topic is equivalent to a statement of information need (Buckley & Walz, 2000a). A study of performance of retrieval systems based on a single query for an information need may be misleading as a topic can be represented in so many ways. It is therefore important to have a sample of possible representations (queries) for each information need for a sensible study of the performances of systems (Buckley & Voorhees, 2000).

The Track Design

Information retrieval experiments involving many participants (systems) have been the tradition of TREC. The query track setup with respect to document collection, information needs (or topics), queries, and retrieval systems was as follows.

The document collection for the TREC-9 query track was the TREC Disk 1 texts [see Voorhees & Harman (2000) for a detailed description of the TREC experiments]. The collection is multidisciplinary in nature and is believed to be a representation of a realistic problem as has been the case in TREC. Table 1 shows the breakdown of the collection by title (Source: Voorhees & Harman, 2000).

The total number of topics (information needs) used in the experiment was 50. Each topic was represented in 43 different ways, that is, a sample of 43 was considered out of the many possible ways of information need representation.

TABLE 2. Runs Submitted to the TREC-9 Query Track.

Part. group	System	Run ID.	Description
U. Massachusetts	INQUERY	IN7a	Queries used as is
U. Massachusetts	INQUERY	IN7e	Queries are fully expanded
U. Massachusetts	INQUERY	IN7p	Queries preprocessed
Sun Microsystems	nova	SUN	Queries used as is
Sun Microsystems	nova	SUN1	Queries preprocessed
Sabir Research	SMART	Saba	Queries used as is
Sabir Research	SMART	Sabe	Queries are fully expanded
Sabir Research	SMART	Sabm	Queries are modestly expanded
U. Melbourne	MG	UoMd	Document-based expansion
U. Melbourne	MG	UoMl	Locality-based expansion
Hummingbird	SearchServer	hum4	Used new linguistic expansion package
Hummingbird	SearchServer	humA	Added fixes for spelling errors
Hummingbird	SearchServer	humB	Baseline
Hummingbird	SearchServer	humD	Document length deemphasized—set to low
Hummingbird	SearchServer	humI	Terms in more than 15% of rows not discarded
Hummingbird	SearchServer	humK	Keyword fields not indexed
Hummingbird	SearchServer	humV	IDF not squared
Microsoft	OKAPI	ok9u	Queries used as is

This makes the total number of queries 2,150 (50 × 43). (It also creates 43 query sets—each query set containing 50 queries each of which is a representation of a topic.) Twenty-one of the query sets were from TREC-8 (Buckley & Walz, 2000a) and the other 22 were added for TREC-9. The query sets were contributed by six groups: John Hopkins University, University of Massachusetts, Sabir, Acsys, Queens College and University of Melbourne. Thirty of the query sets were produced by randomly selected students at the University of Massachusetts, and the rest were by experts from the other groups (Allan, Connell, Croft, Feng, Fisher, & Li, 2000; Buckley, 2000).

Six retrieval systems, with variable numbers of runs, participated in the track. The total number of runs or retrieval methods submitted is 18, out of which only about a third are considered in the analysis of performance variation for the reasons mentioned in a later section. The different runs (or retrieval methods) submitted to the query track are shown in Table 2. Details of the individual runs can be found in the TREC-9 proceedings (Allan et al., 2000; Buckley & Walz, 2000b; D'Souza et al., 2000; Robertson & Walker, 2000; Tomlinson & Blackwell, 2000; Woods, Green, Martin, & Houston, 2000).

Participating groups used their retrieval systems to search the collection for each query and submitted a ranked list of 1,000 documents, which is called a run, to NIST for evaluation. A pool of documents from the 18 runs, top 100 from each, was used for assessment. The relevance judgments originally made for these topics in TREC-1 were used as the official set of relevant documents.

There exist several methods for measuring IR performance (Buckley & Voorhees, 2000). For the purpose of this study, average precision is used as the measure of a run's performance. Average precision reflects the performance of a system over all relevant documents by taking the average of the precision values at each relevant document. If run k retrieves r relevant documents for query j of topic i, average precision (Y_{ijk}) is given by:

$$Y_{ijk} = \frac{1}{r} \sum_{d \in D} \frac{\#\text{rel}_{ijk}}{\#\text{ret}_{ijk}} I(d)$$

where (rel_{ijk}) is the number of relevant documents retrieved and (ret_{ijk}) is the number of documents retrieved by run k for query j of topic i at or before document d; D is the set of all documents; $I(d)$ is 1 if d is relevant or 0 otherwise. This measure favors systems that retrieve relevant documents at the top of the rank.

Experiments

Introduction

Table 3 shows precisions at 30 and 200 documents, mean, standard deviation, and median average precision of the 2,150 queries for the 18 runs. [The table entries are sorted in descending order of mean average precision (MAP).] Note that even though the differences between MAP and median show that the distributions of the performance data are skewed, both mean and median have similar patterns across the runs. Therefore, use of average precision to compare performance of runs for the rest of this analysis is appropriate.

The SMART full and modest expansion methods have the highest performance, whereas variants of the Sun systems (SUN and SUNl) have the lowest performances. Although Hummingbird has submitted six runs, the table shows that all have similar average performance; that is there was no noticeable variation between the performances of the Hummingbird runs (hum4, humA, humB, humD, humI, humK, humV). The variability in the performance of

TABLE 3. Mean of precision @30, @2000 documents and mean, standard deviation median average precision.

RunID	*P@30*	*P@200*	MAP	Std. Dev.	Median
Sabe	0.415	0.262	0.252	0.244	0.194
Sabm	0.382	0.250	0.232	0.239	0.166
IN7e	0.406	0.249	0.229	0.228	0.153
Saba	0.352	0.223	0.192	0.210	0.108
ok9u	0.368	0.217	0.192	0.210	0.107
IN7p	0.354	0.214	0.185	0.203	0.101
IN7a	0.352	0.212	0.180	0.195	0.097
HumD	0.337	0.205	0.177	0.206	0.095
HumA	0.331	0.202	0.174	0.202	0.091
HumI	0.331	0.201	0.174	0.202	0.091
HumB	0.329	0.201	0.173	0.202	0.091
hum4	0.330	0.201	0.171	0.200	0.087
humK	0.328	0.199	0.171	0.203	0.080
humV	0.330	0.197	0.165	0.193	0.083
UoMd	0.313	0.198	0.161	0.183	0.086
UoMl	0.306	0.196	0.154	0.191	0.073
SUNl	0.183	0.106	0.068	0.103	0.021
SUN	0.151	0.091	0.057	0.109	0.010

queries is higher for those systems who do well; that is, the higher the MAP the higher is the standard deviation. The standard deviation refers to the variability in the performance of queries for a given run, i.e., the standard deviation of the average precision of the 2,150 queries. Low deviations show that the methods are performing more or less in a similar fashion irrespective of query type, making runs with high variance much more important for analysis.

It may be difficult to compare the performance of systems and reach any conclusions on how well systems perform in all topics across all queries based on the figures from Table 3. The table, however, gives basic information for selection of runs for further detailed analysis of performance variation.

The performance analysis experiments are presented in the following sections. The experimental methodology is presented first. Then results of the analysis of variance experiment testing the significance of various factors in performance are presented, followed by a section that narrows the focus of analysis to expansion versus nonexpansion of queries. This article then continues the investigation of query expansion but analyzes this on a per-topic basis. Then it examines the effect of query expansion on ranking.

Experimental Methodology

Selection of runs. The application of statistical techniques can show the significance and nature of performance variations due to retrieval method and topic. Some of the runs showing apparent performance differences are excluded in the analyses. Two criteria, in addition to expansion factor, were used to select systems for further analysis: better overall observed performance and high variability. Low deviation shows that a run's performance is more or less similar irrespective of query type. The INQUERY and the SMART runs satisfy the criteria and are the main ones selected for detailed analysis. All topics are included in the analysis irrespective of high or low MAP.

The analysis of performance variation has two phases. The first phase deals with variations due to run, topic, and run–topic interaction, and considers only seven runs (i.e., IN7a, IN7e, IN7p, Saba, Sabe, Sabm, and ok9u). The OKAPI run is included for the first part of the analysis because it has high MAP and variance. The performances of the seven runs are not only better but also highly variable for the collection of queries in the experiment compared to the other runs. The null hypothesis tested is that there are no run, topic, and run–topic interaction effects, that is, variations due to method, due to topic, and due to run–topic interaction are not significant.

The second stage of the analysis looks at performance differences between expansion and nonexpansion runs. The SMART and INQUERY runs are analyzed separately in two groups. The OKAPI run is not included in the second stage of the analysis because there is no expansion run submitted against which the nonexpansion could be compared. The expansion versus nonexpansion comparison allows us to see what is happening when a query is automatically expanded, that is, to measure the effect of expansion on retrieval performance. In other words, the gain/or loss in performance due to automatic expansion of queries. The hypothesis is that there is no significant performance difference between the performances of runs of a given system. That is, the performances of runs using queries that are subject to expansion and a run using nonexpanded original queries are the same. Before moving to the first phase of performance variation analysis, we discuss how the data are analyzed.

Method of data analysis. The observed performance variation shown in Table 3 could statistically be either significant or insignificant. A statistical analysis of variance method can help us identify true differences. Before getting into the details of the statistical tests, it is worth mentioning the possible factors contributing for the variation. In statistical terms:

$$\textit{Total Variation} = \textit{Variation due to Known Effects} + \textit{Error Variance}$$

This model helps to factor out the variation in performance to its component sources. Analysis of variance techniques have been in use for IR performance evaluation data (Banks, Over, & Zhang, 1999; Hull, 1996; Tague-Sutcliff & Blustein, 1995) to measure variation due to different factors. In TREC context, the variation due to known effects include the fixed effects under consideration (system/run and topic) and possible interaction effects that are expected to have effect on performance. The statistical models of performance studied for the analysis of variance include:

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \varepsilon_{ij} \quad (1)$$

TABLE 4. Layout of the experimental data.

Topic	Query	Run_1	Run_2	. . .	Run_7
1	101	$y_{1,1,1}$	$y_{1,1,2}$	. . .	$y_{1,1,7}$
1	102	$y_{1,2,1}$	$y_{1,2,2}$	. . .	$y_{1,2,7}$
⋮	⋮	⋮	⋮	⋮	⋮
1	143	$y_{1,43,1}$	$y_{1,43,2}$	. . .	$y_{1,43,7}$
⋮	⋮	⋮	⋮	⋮	⋮
50	5001	$y_{50,1,1}$	$y_{50,1,2}$	. . .	$y_{50,1,7}$
⋮	⋮	⋮	⋮	⋮	⋮
50	5043	$y_{50,43,1}$	$y_{50,43,2}$	. . .	$y_{50,43,7}$

↔ Within Query (or run) Effect

$$Y_{ij} = \mu + \alpha_i + \gamma_j + \alpha_i\gamma_j + \varepsilon_{ij} \quad (2)$$

where μ is the grand mean, α_i is the System(run) effect, γ_j is the topic effect, $\alpha_i\gamma_j$ is the System-Topic interaction effect, and ε_{ij} is the random error (individual query effect).

Equation 1 assumes simple additive effects of System and Topic, whereas Equation 2 has an additional interaction effect. Banks et al. (1999) further test an extended model of 2, with a multiplier for the interaction effect ($\lambda\alpha_i\gamma_j$) to capture specific forms of interaction.

The basic assumptions for these models include independence between runs, normality of the sampling distributions and equality of variances of runs. These assumptions may be violated to a different extent like any other practical data analysis. However, due to the robustness of ANOVA, the procedure can be applied (Jackson & Brashers, 1994) with some departures from the assumptions.

All the runs use all 50 topics, each having 43 queries (also known as 43 query sets); that is, all the queries (subjects) are used for each run. The layout of the experimental data is shown in Table 4. In a repeated measure environment (Crowder & Hand, 1990; Everitt & Der, 1998), the experimental subjects (queries) fall under all treatments (runs or retrieval methods). The method allows us to study the effects that interest us: between query effects (those whose values change from query to query); within query effects or run effects (those whose values may differ from run to run); between query, and within query (or run) interaction effects.

When repeated measure analysis is used, each query acts as its own control. The normal query-to-query variation can thus be removed from the error sum of squares. This is particularly important to analyze performance data where query-to-query variability within a topic is high. The experimental data shows not only variability between topics but also within a topic. Performance of queries of topic 78, for example, has a standard deviation of 0.3259 and a range of 0.8574. This indicates high variability between queries within a topic. Controlling for between-query variability can greatly reduce the error term in the analysis of variance and allow us to identify smaller system, topic and interaction differences.

The model for analysis of variance of the repeated measures is:

$$Y_{ijk} = \mu + \alpha_i + \gamma_k + \tau_{jk} + \varepsilon_{ijk} \quad (3)$$

where Y_{ijk} is the k^{th} measurement [run k, k = 1, . . .,7] of query j [j = 1, ...,43] of topic i [i = 1, . . .,50], μ is the grand mean, α_i is the topic effect, γ_k is run effect, τ_{jk} is the query-run interaction effect, and ϵ_{ijk} is random error. The total sum of squares in the repeated measurement data can be divided into:

(1) the sum of squares between individual queries; (a) the sum of squares between *topics*; and (b) the sum of squares between *queries* within *topics*;
(2) the sum of squares between *runs*;
(3) the sum of squares for the interaction (*run*query*); (a) the sum of squares for *run*topic* interaction; (b) the sum of squares for *run*query* within topic (*run*query(topic)*).

We are interested in testing the significance of the *run*, *topic*, *run*topic* sources of variation indicated in model (3). However, the query within topic (*query(topic)*) effect is an important element in a repeated measurement that is worth testing. It also reduces the sum of square that goes to the error term unnecessarily.

Results of Analysis of Variance

The SAS ANOVA (Analysis of Variance) and GLM (Generalized Linear Model) procedures are used for the analysis. Analysis of variance shows that the system, topic, and system–topic interaction effects are all significant (see Table 5a). A similar result was found by Banks et al. (1999). In fitting model (3), our data has 43 replicates or queries under each topic that makes it different from the data used by Banks et al. (1999).

The significant probabilities for all the effects in the model are much less than 0.05 (i.e., Pr > F is 0.0001 for all effects). It should be noted that the system–topic interaction effect is lower compared to the additive effects, however. This is shown by the magnitude of the mean square of the variance components, labeled "Mean Square," that is, 0.067 (interaction) against 1.701 (run) and 8.549 (topic). The table shows that topic is the number one important factor, followed by retrieval method (run), which contributes to the variation in performance. The value of R^2 (variation explained by the model) in our analysis is 0.9487, which is a better fit. (Arcsine transformations to stabilize average precision data were used but the results of the analyses were the same apart from improving the value of R^2 as shown by Tague-Sutcliff & Blustein, 1995).

When there is an interaction effect, it is of primary importance. The interaction term tells if the run changes/differences are the same for all topics. The nature of the run–topic interaction can be clearly seen from the interac-

TABLE 5. Analysis of variance (dependent variable: average precision).

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
a) All Seven Runs					
Model	2449	694.5274	0.2836	95.24	0.0001
Error	12600	37.5185	0.0030		
Corrected total	15049	732.0459			
	R-Square	C.V.	Root MSE		AVP Mean
	0.9487	26.1391	0.0546		0.2088
Source	*DF*	Anova SS	Mean square	*F*-Value	Pr > *F*
TOPICID	49	418.8830	8.5486	2870.93	0.0001
QID (TOPICID)	2100	245.7583	0.1170	39.3	0.0001
RUNID	6	10.2069	1.7011	571.30	0.0001
TOPICID*RUNID	294	19.6792	0.0669	22.48	0.0001
b) The INQUERY Runs					
Model	2249	277.4743	0.1234	66.63	0.0001
Error	4200	7.7771	0.0019		
Corrected total	6449	285.2514			
	R-Square	C.V.	Root MSE		AVP Mean
	0.9727	21.7492	0.0430		0.1979
Source	*DF*	Anova SS	Mean Square	*F*-Value	Pr > *F*
TOPICID	2100	104.6689	0.0498	26.92	0.0001
QID(TOPICID)	2	3.1132	1.5566	840.65	0.0001
RUNID	49	166.1079	3.3900	1830.74	0.0001
TOPICID*RUNID	98	3.584	0.0366	19.75	0.0001
c) The SMART runs					
Model	2249	340.9093	0.1516	79.94	0.0001
Error	4200	7.9637	0.0019		
Corrected total	6449	348.8730			
	R-Square	C.V.	Root MSE		AVP Mean
	0.9772	19.3221	0.0435		0.2254
Source	DF	Aova SS	Mean Square	*F*-Value	Pr > *F*
TOPICID	2100	123.0345	0.0586	30.90	0.0001
QID(TOPICID)	2	3.9209	1.9605	1033.93	0.0001
RUNID	49	209.6213	4.2780	2256.18	0.0001
TOPICID*RUNID	98	4.3326	0.0442	23.32	0.0001

tion graph (Fig. 1). The interaction plots are not parallel, indicating that the differences in performance of runs are not the same across topics. The magnitude of variation between the performances of runs for a query depends on which topic a query is in (topics 70 and 93 have the highest variation between runs). For difficult topics (low MAP), the differences between performances of runs are very low (see later section for the discussion on topic level analysis).

In general, the analysis shows that the observed system and topic variations are significant. That is, the performance variations between the seven runs are not by chance. The variation of performance due to topics is also not by chance, and in fact, the variance component due to topics is much greater than the variation attributable to the retrieval methods. The next section deals with analysis of variation between expansion and nonexpansion runs.

Expansion versus Nonexpansion Experiments

Introduction. Only performances of the runs of INQUERY and SMART are investigated in this section. IN7e and Sabe are the two full expansion runs of INQUERY and SMART, respectively. The scatter plots of Figure 2 show the performances of INQUERY (IN7e and IN7p) and SMART (Sabe and Sabm) runs against their own and the OKAPI nonexpansion runs. The plots show strong correlation between the runs. The correlation coefficients between IN7e and IN7a, IN7p and IN7a, IN7e and ok9u, Sabe and Saba, Sabm and Saba, Sabe and ok9u are 0.9247, 0.9543, 0.8732, 0.9233, 0.9598, and 0.8497, respectively. (Note that these coefficients measure the strength of relationship between the runs, not the agreements between them. Agreement measures the degree to which two systems assign similar ranks to documents; whereas relationship shows that the degree to which a change in rank assigned by a system is associated with a change in rank by another.) The coefficients show that the expansion runs are more closely related with their own nonexpansion runs compared to their relationship with the OKAPI nonexpansion run. In such data, a repeated-measure approach, discussed above, is a better choice to account for query-to-query differences. To see the effect of query expansion on performance, separate comparisons are made for the runs of each system.

The INQUERY runs. The three INQUERY runs are treated as if they are measurements at three different times (applying a repeated-measures analysis). Table 5b shows the

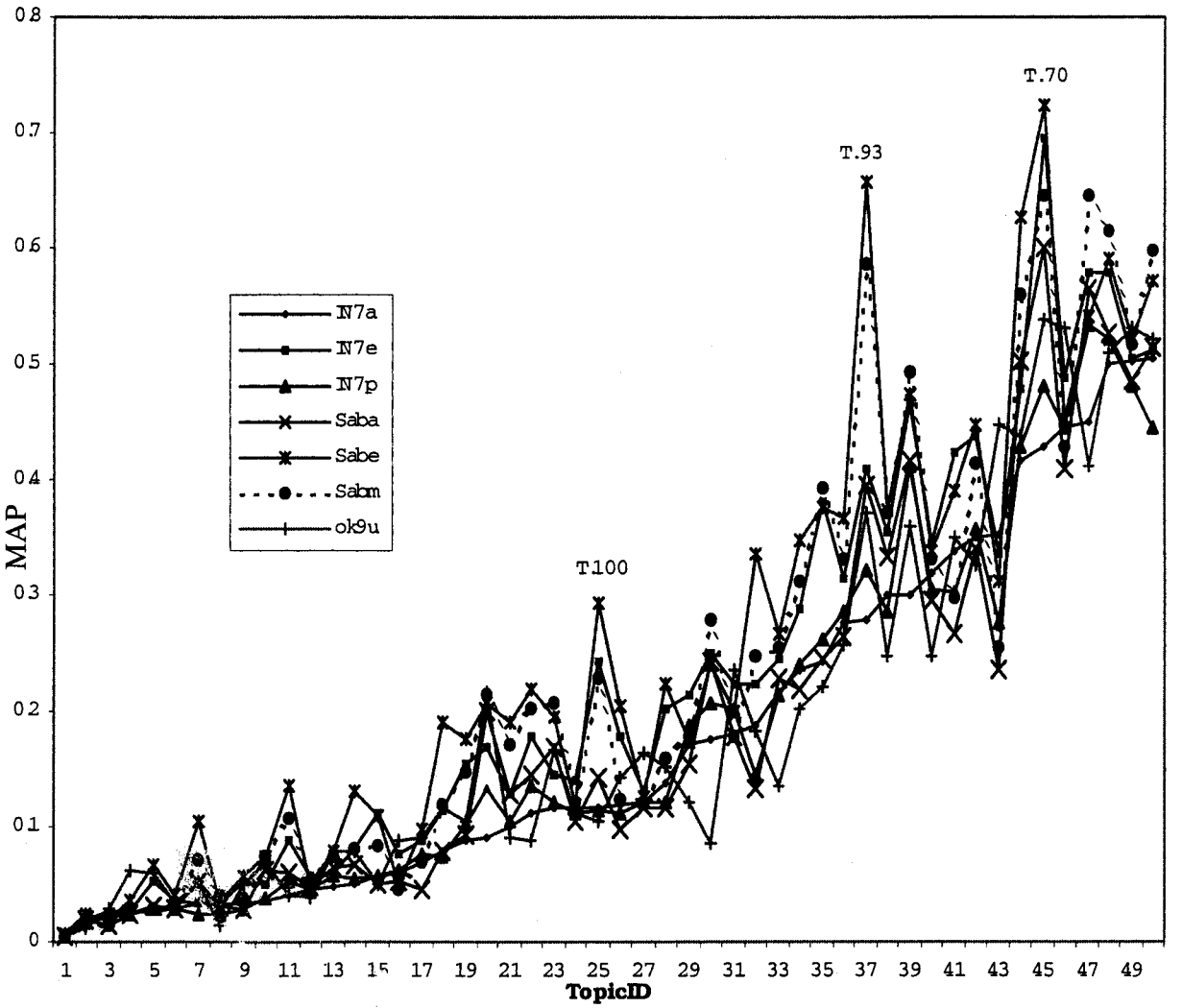

FIG. 1. System–topic interaction effect for all runs (topics (TopicID) are sorted by MAP values of IN7a.

ANOVA by fitting model (3) for the INQUERY runs. It shows that both the main and interaction effects are significant with R^2 value of 0.9727. It should be, however, noted that the variation due to topics (3.39) is higher than variation due to runs (1.56) and the interaction (0.04). The interaction effect contributes the least to the variation.

A plot of the interaction effect (Fig. 3a) shows that the size of differences in performance of the three runs vary from topic to topic. Apart from two topics, where the observed nonexpansion is slightly better, the full-expansion run is consistently above the nonexpansion run in all other topics. The interaction effect shows that the magnitude of the differences vary from topic to topic. In some cases, the difference is higher and in other cases the difference is lower. Otherwise, query expansion (IN7e) nearly always improves performance. [The third run (IN7p) is closer to the nonexpansion (IN7a) run.]

The SMART runs. A similar method is used to analyze the SMART runs (Table 5c shows the ANOVA). The fixed effects, run, topic, and their interaction, are significant with R^2 value of 0.9772. The interaction plot (Fig. 3b) shows that the magnitude of the differences between the runs is variable similar to the results of the INQUERY runs. The ANOVA table shows the variance components attributable to the run, topic, and interaction effects as 1.96, 4.28, and 0.04, respectively, making topic the highest variance contributing factor. The interaction plot, however, shows both expansion methods (full expansion and modest expansion) are above the nonexpansion method except for three topics where the latter is slightly above one of the two runs (see later section for topic level analysis). Looking at the performances of the two expansion runs (Sabe and Sabm), the full expansion (Sabe) performs better in most of the topics.

Even though the magnitude of difference in performance varies from topic to topic, expansion (both full and modest expansions) mostly improves performance. In general, the performance variation between expansion and nonexpansion runs is not by chance. That is, a run using a query

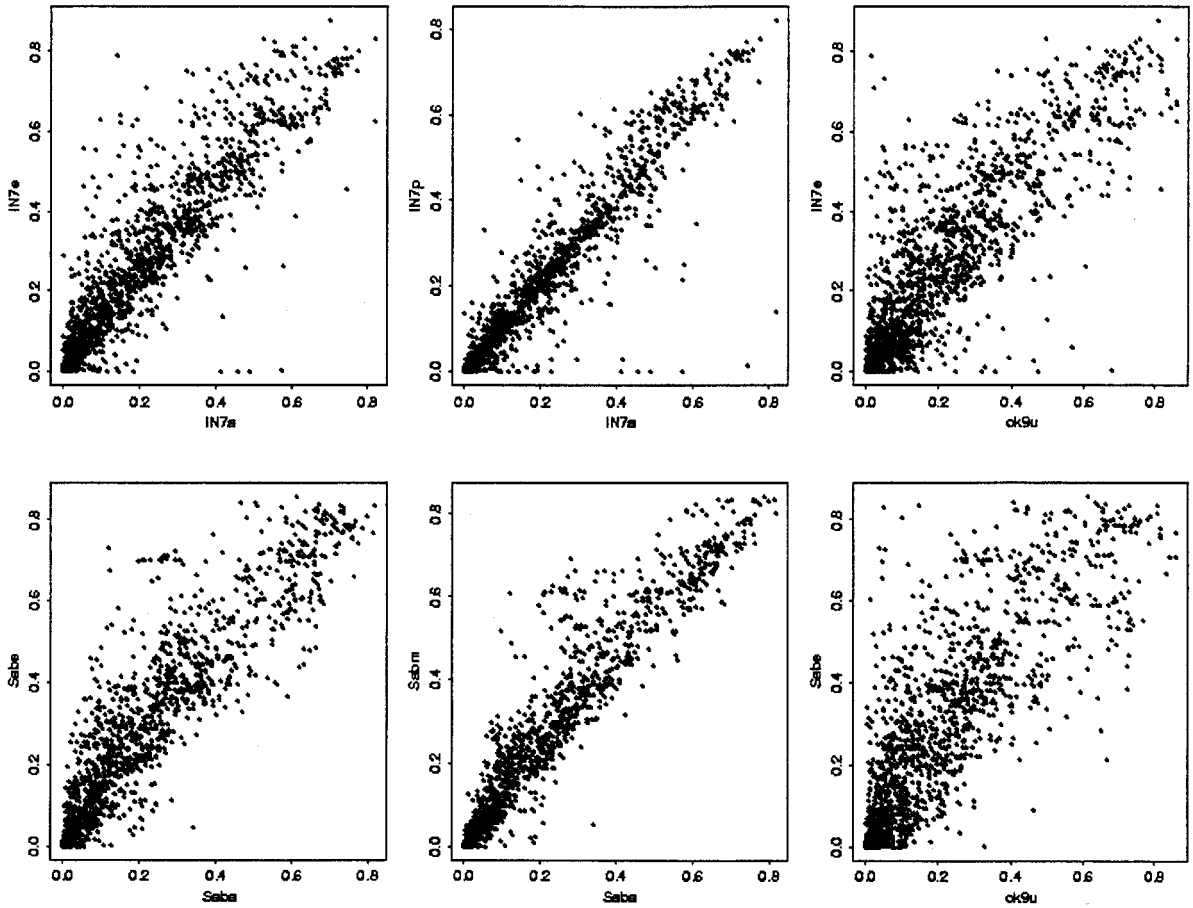

FIG. 2. Scatter plots of average precision for pairs of runs.

subject to automatic expansion has a better average precision than a run using a query as is.

Analysis of Performance by Topic

Introduction. Expression of information needs in the right words affects the performance of systems. The more abstract the information need is, from the point of view of topic expression, the more difficult it is to express in a form of a query (see Table 6 for the grouping of topics by MAP—topics are sorted by MAP). Although there may be a good number of relevant documents for a topic in a given collection, retrieval performance may be low because of the difficulty in information need expression. Voorhees and Harman (1997) note a low correlation between topic difficulty and number of relevant documents.

For instance, there are 275 relevant documents for topic 74 compared to 74 documents for topic 78. However, the difficulty in expressing topic 74 in queries makes retrieving relevant documents very hard. (The title of topic 74 is "Conflicting Policy.") The description of the topic reads as "Document will cite an instance in which the U.S. government propounds two conflicting or opposing policies." This topic has a clear information need marker, "conflicting or opposing," but lacks known elements on which there exist conflicting policies. It is typical of a topic (information need) that is difficult to express in a form of query and the difficulty is shown by very low average precision for all the queries. On the other hand, most queries of topic 78 were able to retrieve relevant documents with a better average precision. (The title of topic 78 is "Greenpeace.") It should be noted, however, that there were queries for topic 78 that did not retrieve even a few relevant documents. For instance, no run retrieved a single relevant document at the 1,000 cutoff point for query 13 of topic 78 "Will the delirious effects of emission and pollution be stopped?"

Based on the analysis of performance data for each topic, an effort is made to classify topics into two groups. The first group contains those topics where runs have only minimal observed variation, and the second group is a set of discriminating topics. The second group is further divided in to three sets: topics where nonexpansion runs do better than expansion; topics where paired runs are different but there is not sufficient evidence to claim one is better than the other; and topics where expansion runs do better than nonexpansion (see Tables 6 and 7 for the distribution of topics based on these grouping).

In the comparison of runs for each topic, the primary hypothesis tested is that retrieval method does not affect retrieval performance, that is, the mean performances of the

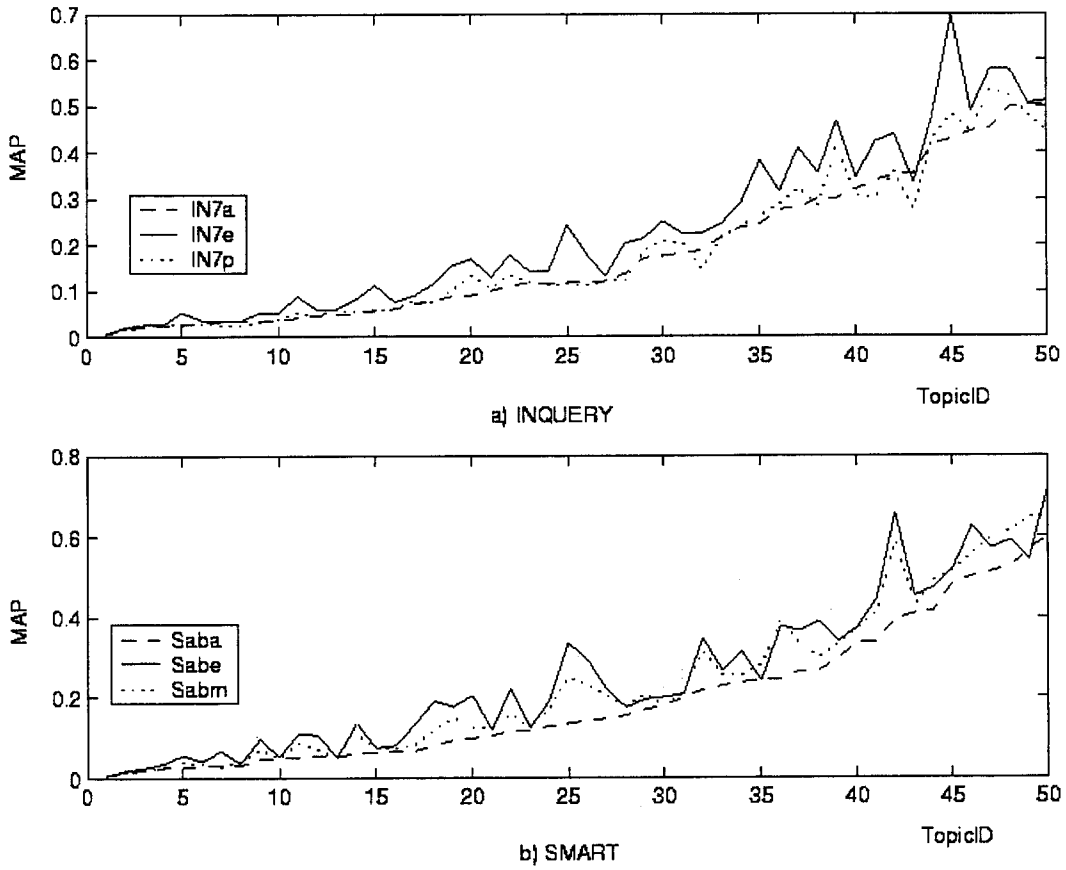

FIG. 3. System–topic interaction effect for INQUERY and SMART runs: [topics (TopicID) are sorted by MAP values of the respective nonexpansion runs.]

three runs are equal. Post hoc paired comparison hypotheses are tested whenever the primary hypothesis is rejected—that is, when there is performance difference between runs. Results of analysis of each of the systems is presented in the following sections.

The INQUERY runs. The preliminary investigation was whether the INQUERY runs have variations in performance or not. (This is conducted for each topic.) The analysis of variance for each topic shows that the run effect is significant in all except two topics (topics 81 and 95), that is, the observed performance differences are true. Paired comparisons are undertaken for these topics and show that most of the significant variations are between the full expansion (IN7e) and nonexpansion (IN7a). Expansion shows significant improvement over nonexpansion in 41 topics. The performances of the two runs were not significantly different in 7 topics including topics 61 and 91 where the observed performance of the nonexpansion is slightly better.

As shown in Table 8, comparison of IN7a (query as is) and IN7p (query preprocessed) shows that IN7a performs significantly better in five topics, whereas IN7p does better in 12 topics, making discriminating topics 17. It is, therefore, difficult to say one is better than the other. (See Fig. 4 for the plot of performance differences *d*.) The plotted points are performance differences against means of paired runs. The variable *d* for each plot is the difference in average precision between pairs of runs; and

TABLE 6. Topic groupings by MAP.

MAP interval	No. topics	Topics
$0.0 < \text{MAP} \leq 0.1$	($n_1 = 17$)	{74, 84, 75, 91, 95, 92, 87, 80, 96, 60, 73, 89, 97, 67, 72, 71, 83}
$0.1 < \text{MAP} \leq 0.2$	($n_2 = 13$)	{59, 68, 76, 81, 94, 88, 63, 79, 85, 65, 69, 100, 98}
$0.2 < \text{MAP} \leq 0.3$	($n_3 = 5$)	{64, 66, 53, 90, 62}
$0.3 < \text{MAP} \leq 0.4$	($n_4 = 6$)	{86, 77, 61, 57, 99, 54}
$0.4 < \text{MAP} \leq 0.5$	($n_5 = 4$)	{56, 93, 51, 55}
$0.5 < \text{MAP}$	($n_6 = 5$)	{52, 78, 82, 58, 70}

TABLE 7. Distribution of topics by MAP and power of discrimination.

MAP	Nondiscriminating topics		Discriminating topics	
	IN7a–IN7e	Saba–Sabe	IN7a–IN7e	Saba–Sabe
$0.0 < \text{MAP} \leq 0.1$	3	3	14	14
$0.1 < \text{MAP} \leq 0.2$	1	2	12	11
$0.2 < \text{MAP} \leq 0.3$	0	1	5	4
$0.3 < \text{MAP} \leq 0.4$	1	0	6	6
$0.4 < \text{MAP} \leq 0.5$	0	0	4	4
$0.5 < \text{MAP}$	2	1	3	4

$$IN7apMean = (IN7a + IN7p)/2,$$

$$IN7aeMean = (IN7a + IN7e)/2,$$

$$IN7peMean = (IN7p + IN7e)/2,$$

$$SabamMean = (Saba + Sabm)/2,$$

$$SabaeMean = (Saba + Sabe)/2 \text{ and}$$

$$SabmeMean = (Sabm + Sabe)/2.$$

The differences in performance, *d*s, are expected to be distributed evenly around zero; in other words, the mean of the differences is zero, if two pairs of retrieval methods (runs) are similar. The figure shows that the numbers of plotted points, *d*s, above zero are high for expansion and nonexpansion runs (e.g., *d* versus *IN7aeMean* and *d* versus *SabaeMean*), indicating that expansion methods are better. Although it is not possible to say the points are evenly distributed, the plot of differences between IN7a and IN7p are around zero as the performance of the two runs do not show significant differences in 33 topics (66 %).

Looking at the distribution of topics (Table 7) where runs do not show any significant variation in performance, most of them are from low and high MAP groups. This shows that difficult topics remain difficult with no significant improvement due to expansion. On the other hand, nonexpansion runs perform equally well for well-formed queries.

The SMART runs. The run effect was not significant in four topics (69, 81, 91, and 96). (The MAP for these four topics is below 0.2.) The paired comparison show that the observed differences between Saba and Sabe are not significant in three more topics (66, 82, and 89). In these three topics the observed performances of Saba are slightly higher than Sabe. The tests show significant run effects for these topics because of the differences between modest expansion and nonexpansion. In topic 89, Saba does better than Sabm, whereas in the other two topics Sabm is above Saba. In all other 43 topics, the performance of Sabe is higher than Saba (see Table 8 and Fig. 3b). For all topics with MAP between 0.3 and 0.5, runs show significant differences. The SMART third run (modest expansion) shows significant differences with the nonexpansion in 37 topics. The full expansion does well for nearly all the topics where there is a significant difference.

From the topic level analysis of performance data, it can be concluded that automatic query expansion does not do any damage, even if no improvement is seen. In topics where nonexpansion shows better observed performances, the statistical tests show that the variations are not significant. The plots of Figure 4 show that most of the *d* values of Sabe and Sabm against Saba fall above 0. This is another indication that the two runs are better than the nonexpansion run.

Both systems (INQUERY and SMART) show more or less similar pattern in terms of the distribution of topics in which runs show significant differences. In topics within the medium MAP group, runs have significant performance differences. Table 8 shows the distribution of topics by category of paired comparison results. The expansion runs of both systems have true differences against their respective nonexpansion runs in at least 84% of the topics (42 and 43 for INQUERY and SMART, respectively). The performance of the INQUERY query preprocessing run (IN7p), however, does not show true differences in 66 % of the topics. This is an indication that the current query preprocessing does not guarantee consistent improvement in performance over nonexpansion.

Effect of Query Expansion on Ranking of Relevant Documents

It is important that relevant documents are retrieved at the top of the ranked list. Analysis of the data on document ranks assigned by the systems using query expansion and nonexpansion shows that query expansion not only retrieves more relevant documents but also alters the ranking by improving the rank order of relevant documents.

Investigation of the rank data was made on 33 topics (whose $\text{MAP} > 0.1$). Percentages of relevant documents retrieved at different rank levels are used to compare runs. The rank levels used are variable, except the pool depth (1,000), and are related to the number of relevant documents of a given topic because the number of relevant documents vary from topic to topic. If a topic has *n* number of relevant documents, the rank levels are *n*, 150% *n* and the pool depth 1,000. The percentage at rank *n* is *R*-precision. For instance, topic 78 has 74 pooled relevant documents. The rank level information is "how many of the documents retrieved

TABLE 8. Distribution of topics by run pair (SMART runs are on the right inside brackets).

Run pair	Results of paired comparisons				Run pair
	Not sig. diff.	Sig. diff.[a]	Sig. less	Sig. greater	
IN7a–IN7p	33 (12)	0 (0)	12 (37)	5 (1)	(Saba–Sabm)
IN7a–IN7e	7 (7)	1 (0)	42 (43)	0 (0)	(Saba–Sabe)
IN7p–IN7e	3 (13)	1 (3)	46 (31)	0 (3)	(Sabm–Sabe)

[a] Significantly different but does not show direction of difference.

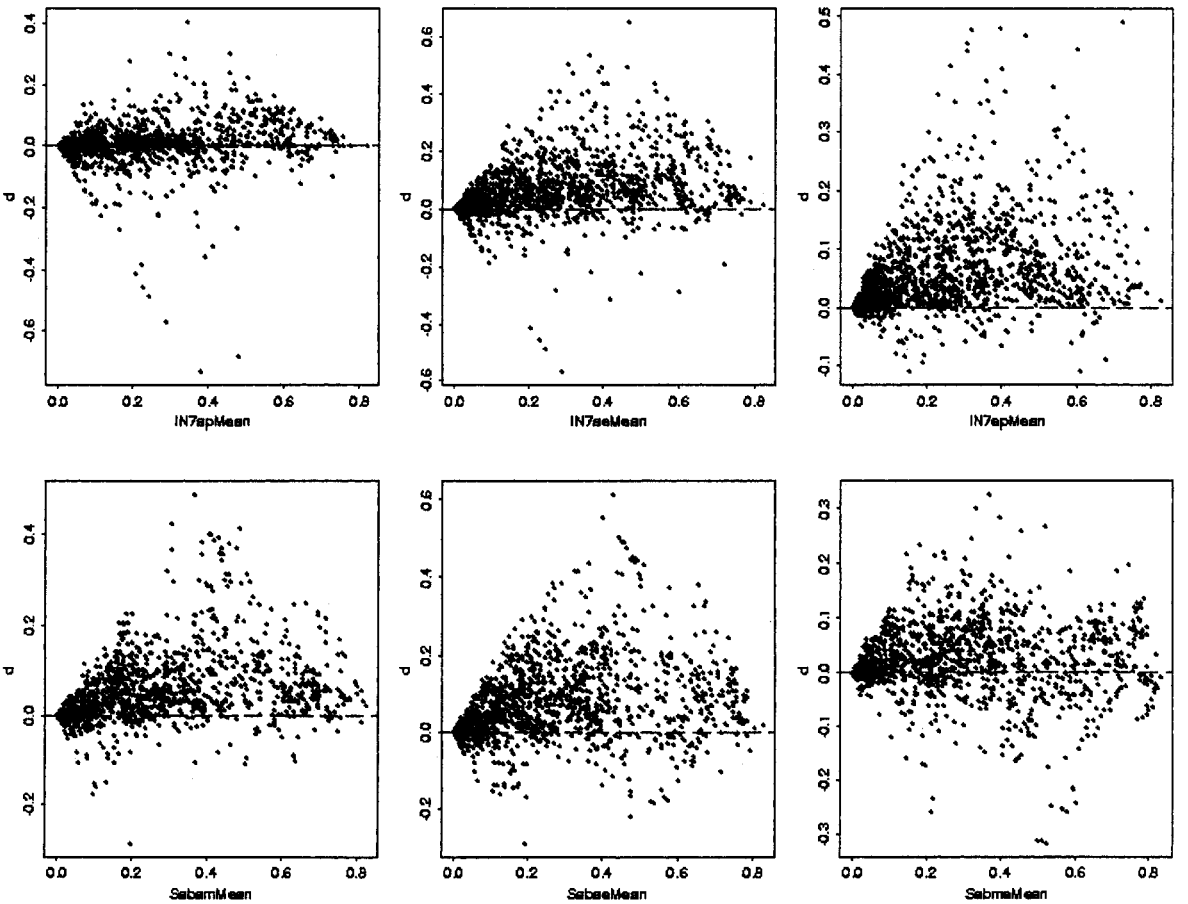

FIG. 4. Plots of performance (average precision) differences of paired runs against their mean.

within ranks 74, 114 (74 + 50% of 74, i.e., 150% of 74) and 1,000 are relevant?" These are basically precision figures at variable ranks—the first being *R*-precision. The percentage of relevant documents that are not retrieved within 1,000 pool depth is also used as a measure of performance.

Based on the rank data of the 33 topics, *R*-precision for the SMART full expansion run (Sabe) is 39.5%, i.e, Sabe retrieves on average about 39.5% of the relevant documents within rank *n*. The proportion of relevant documents retrieved within rank 150% *n* is 56% (39.51 + 16.62). The corresponding figures for the nonexpansion run (Saba), however, are 35 and 50%, respectively. Similarly, the INQUERY expansion run (IN7e) improves the performance figures from 32 to 38% and from 47 to 53%, respectively).

The expansion runs also reduce the percentage of documents that are not retrieved within the 1,000 pool depth. Table 9 shows the percentage figures.

Conclusion

Evaluation of performance is one of the difficult (Banks et al., 1999) and yet important activities in IR. This article reports a study that explored the performance of systems of the Query Track at TREC-9. It specifically investigates the performance of query expansion against nonexpansion runs of same systems (INQUERY and SMART). The analyses test the significance of factors affecting retrieval performance. In addition to looking at the average performance differences in terms of average precision, it also analyses the ranking order improvements due to expansion. The conclusions that can be made from the analyses of the experimental data are:

(1) All run (retrieval method), topic, and run–topic interaction effects are significant.

TABLE 9. Percentage of relevant documents retrieved.

Rank	IN7a	IN7e	IN7p	Saba	Sabm	Sabe
Within # rel. docs (*R*-precision)	32.38	37.60	33.36	34.66	39.51	38.52
Within 150% of # rel. docs	14.87	15.13	14.68	15.72	16.62	16.71
Within 1,000	7.45	8.55	7.69	7.91	9.48	8.83
Not ret. within 1,000 docs.	45.30	38.72	44.26	41.71	34.38	35.94
	100	100	100	100	100	100

(2) The variation due to topic is much greater than the variation due to retrieval method and the interaction effects. The performance of a retrieval method is highly variable within a topic making information need expression an important factor.
(3) The significant interaction effect does not make the main effects less important. Detailed topic level analysis shows that the magnitudes of the improvements by the expansion runs over the nonexpansion runs vary from topic to topic. In some topics, the variation is high and in others the variation is low. For topics with low MAP, variation is low in general.
(4) The SMART and the INQUERY full expansion methods perform significantly better in 43 and 42 topics, out of 50, respectively. The differences in performances between the two pairs of runs of the two systems are not significant for 7 topics.
(5) MAP for nondiscriminating topics is either low or high. This indicates that difficult topics remain difficult with no significant improvement due to expansion, on one hand, and nonexpansion methods perform equally well for well formed queries, on the other.
(6) Analyses of document ranks show that QE not only improve average precision but also retrieves relevant documents at top ranks providing a user a better ranking order of documents. The SMART and the INQUERY expansion runs retrieve about 39.5 and 38% of the relevant documents, respectively, within the limits of the number of relevant documents for a given topic compared to their respective nonexpansion method where the corresponding figures are 35 and 32%, respectively.

In this article, a case study showing the potential of statistical analysis of variance method is presented. The performance data of systems, using same experimental collection, can be analyzed by applying repeated-measures analysis of variance approach. The article shows how the within- and between-query effects can be tested by accounting for query-to-query variability. The results can help IR researchers and system designers to focus on opportunities for improvement that are hidden within the average figures. The method of analysis used in this article can be used in any other retrieval experiment involving different factors that would possibly affect performance.

Acknowledgments

I thank Donna Harman for her guidance in the preparation of the article and her specific comments and suggestions, and Paul Over for his comments on the draft article. I also thank the anonymous reviewers for their comments and suggestions.

References

Allan, J., Connell, M.E., Croft, W.B., Feng, F.-F., Fisher, D., & Li, X. (2000). INQUERY and TREC-9. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Banks, D., Over, P., & Zhang, N.F. (1999). Blind men and elephants: Six approaches to TREC data. Information Retrieval, 1, 7–34.

Beaulieu, M., Robertson, S., & Rasmussen, E. (1996). Evaluating interactive systems in TREC. Journal of the American Society for Information Science, 47, 85–94.

Buckley, C. (2000). The TREC-9 query track. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Buckley, C., & Voorhees, E. (2000). Evaluating evaluation measure stability. In: Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval, Athens, Greece (pp. 33–40).

Buckley, C., & Walz, J. (2000a). The TREC-8 query track. In: E.M. Voorhees & D. Harman (Eds.), Proceedings of the eighth Text REtrieval Conference (TREC-8) (NIST Special Publication 500-246).

Buckley, C., & Walz, J. (2000b). SabIR research at TREC-9. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Buckley, C., Singhal, A., & Mitra, M. (1997). Using query zoning and correlation within SMART: TREC 5. In: E.M. Voorhees & D. Harman (Eds.), Proceedings of the fifth Text REtrieval Conference (TREC-5) (NIST Special Publication 500-238, pp. 105–118).

Carpineto, C., de Mori, R., Romano, G., & Bigi, B. (2001). An information-theoretic approach to automatic query expansion. ACM Transactions on Information Systems, 19(1), 1–27.

Crowder, M.J., & Hand, D.J. (1990). Analysis of repeated measures. Monograph on statistics and applied probability 41. London: Chapman.

D'Souza, D., Fuller, M., Thom, J., Vines, P., Zobel, J., de Kretser, O., Wilkinson, R., & Wu, (2000). Melbourne TREC-9 experiments. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Ekmekcioglu, F.C., Robertson, A.M., & Willett, P. (1992). Effectiveness of query expansion in ranked-output document retrieval systems. Journal of Information Science, 18(2), 139–147.

Efthimiadis, E.N. (1996). Query expansion. In: M.E. Williams (Ed.), Annual Review of Information Science and Technology, 31, 121–187.

Everitt, B.S., & Der, G. (1998). A handbook of statistical analyses using SAS. London: Chapman & Hall.

Harman, D. (1988). Towards interactive query expansion. Proceedings of the eleventh international conference on Research and development in information retrieval (pp. 323–331).

Harman, D. (1992a). Relevance feedback revisited. In: Proceedings of the 15th international conference on research and development in information retrieval, Copenhagen, Denmark.

Harman, D. (1992b). Relevance feedback and other query modification techniques. In: W.B. Frakes & R. Baeza-Yates (Eds.), Information retrieval: Data structures and algorithm (pp. 241–263). Englewood Cliffs, NJ: Prentice Hall.

Hull, D.A. (1996). Stemming algorithms: A case study for detailed evaluation. Journal of the American Society for Information Science, 47, 70–84.

Jackson, S., & Brashers, D.E. (1994). Random factors in ANOVA. Quantitative Applications in the Social Sciences series no. 98. Thousand Oaks, CA: Sage Publications.

Kekalainen, J., & Jarvelin, K. (1998). The impact of query structure and query expansion on retrieval performance. Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (pp. 130–137).

Mandala, R., Tokunaga, T., & Tanaka, H. (1999). Combining multiple evidence from different types of thesaurus for query expansion. Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval (pp. 191–197).

Mitra, M., Singhal, A., & Buckley, C. (1998). Improving automatic query expansion. Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval (pp. 206–214).

Peat, H.J., & Willett, P. (1991). The limitations of term co-occurrence data for query expansion in document retrieval systems. Journal of the American Society for Information Science, 42(5), 378–383.

Qiu, Y., & Frei, H. (1993). Concept based query expansion. Proceedings of the sixteenth annual international ACM SIGIR conference on research and development in information retrieval (pp. 160–169).

Robertson, S.E., & Walker, S. (2000). Microsoft Cambridge at TREC-9: Filtering track. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Spink, A., Wolfram, D., Jansen, M.B.J., & Sarecevic, T. (2001) Searching the Web: The public and their queries. Journal of the American Society for Information Science and Technology, 52(3), 227–234.

Tague-Sutcliff, J., & Blustein, J. (1995). A statistical analysis of the TREC-3 data. In: Harman, D.K. (Ed.), Proceedings of the third Text REtrieval Conference (TREC-3) (pp. 385–398).Washington, DC: U.S. Government Printing Office.

Tomlinson, S., & Blackwell, T. (2000). Hummingbird's Fulcrum SearchServer at TREC-9. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Voorhees, E., & Harman, D. (1997). Overview of the Fifth Text REtrieval Conference (TREC5). In: E.M. Voorhees, D.K. Harman, (Eds.), The Fifth Text Retrieval Conference (TREC-5), NIST Special Publication 500-238 (pp. 1–28).

Voorhees, E., & Harman, D. (Eds.). (2000). Overview of the Eighth Text REtrieval Conference. Proceedings of the eighth Text REtrieval Conference (TREC-8). NIST Special Publication 500-246.

Woods, W.A., Green, S., Martin, P., & Houston, A. (2000). Halfway to question answering. Available at: http://trec.nist.gov/pubs/trec9/t9_proceedings.html.

Xu, J., & Croft, B. (1996). Query expansion using local and global document analysis. In: H. Frie, D. Harman, P. Schauble, & R. Witkinson (Eds.). Proceedings of the nineteenth international ACM-SIGIR conference on research and development in information retrieval (pp. 4–11). Zurich, Switherland: ACM Press.

Xu, J., & Croft, B. (2000). Improving the effectiveness of information retrieval with local context analysis. ACM Transactions on Information Systems. 18(1), 79–112.