TOPIC BY TOPIC PERFORMANCE OF INFORMATION RETRIEVAL SYSTEMS

Walter Liggett
National Institute of Standards and Technology
Gaithersburg, MD 20899
walter.liggett@nist.gov

Abstract

Formulation of topic properties is the goal of this paper. These properties are to be used in judging the difficulty of topics appropriate to the Text REtrieval Conference (TREC) ad hoc task. Applying statistical methods to both TREC-6 and TREC-7 information retrieval results, we identify topic pairs that exemplify topic properties useful in relating topic statements to system performance. From two topics that seem the same with respect to the challenge they provide information retrieval systems, we formulate topic properties by relating the corresponding topic statements to what is known about information retrieval systems. Some properties apparent in the topic pairs identified are linked to topic expansion. These pairs exemplify both the need for expansion and the danger in automatic expansion.

1. Topic Properties

System developers would like to be able to judge topic difficulty from reasoning involving topic properties, but appropriate topic properties have not yet been defined or even conceived. For example, a topic property or perhaps a set of topic properties is needed to describe whether only a few key words or several sentences are needed to narrow the topic sufficiently. With a better understanding of topic properties, a system developer might be better able to instruct users on creation of the topic statement, to build a better user interface, or to configure a system that bases topic expansion on topic type.

There is some doubt that formulation of topic properties for these purposes is even possible. "Little is known about what makes a topic difficult," conclude Voorhees and Harman in their TREC-6 overview (1998). Yet, reading papers in this (TREC-7) Proceedings that describe systems for the ad hoc task, one sees discussions of topic expansion illustrated with specific topic statements. Clearly, the authors of these papers see certain topic statements as examples of the topic properties that underlie their expansion strategies. The Query track in TREC-7 is intended to help with topic property formulation. Our analysis of the ad hoc task shows that one can connect system performance to generic topic properties.

The effect of topic properties on system performance depends on the document collection. One consequence is that our statistical methods might connect two topics not because the topic statements share some topic property but because the document collection has some unexpected characteristic. Another consequence is that topic properties formulated through the TREC ad hoc task might not be as useful with other document collections. It is hoped that one can guard against these consequences by requiring that the topic properties formulated seem independent of document collection.

For each topic in the ad hoc task, the TREC evaluation provides system performances, a collection of numbers that might be termed a performance profile. (As "systems," we include what others might call system variants, alternative runs with different configurations of the software created by some group.) These profiles are a partial answer to the question, "How do topics differ?" However, the information retrieval community needs a more general answer, one that can be applied to any topics in the style of the ad hoc task. The purpose of this paper is display of some TREC data in a way that might crystalize the concepts required for such an answer.

This paper presents a pair of TREC-6 topics and three pairs of TREC-7 topics for the reader to study. It is hoped that the reader will be able to offer an opinion on the topic properties each pair exemplifies. These particular pairs occupy places in the data that seem to recommend them for careful study. First, each pair is unusual in that the two performance profiles differ from the average profile for all topics. Second, within each pair, the two topics have similar profiles so that the question of what the two topics have in common is intriguing. Third, being unusual and similar in these senses holds for both performance measures considered. To the study of these topics, the user must bring knowledge of how information retrieval systems operate. For example, the reader might consider how systems perform topic expansion.

Presentation of the four pairs involves two alternative measures of system performance and a method for decomposing a two-way table, namely, the system-by-topic table of performance measurements. One performance measure considered is average precision, which is familiar to TREC participants. The basis of the other is the depth at 25 percent recall, which is the document rank at which 25 percent of the relevant documents have been found. Use of both of these measures, which are detailed in Section 3, seems beneficial because these measures behave differently. The method for table decomposition is the one that underlies two-way analysis of variance. The components in the decomposition describe performance averaged over topic and over system as well as aspects of the system-topic interaction.

This paper is organized to help the reader focus on the correspondence between topics and observed system performance. In Sections 2 through 4, we present two topic pairs for consideration without going into all the statistical details of their selection. In Sections 5 and 6, we provide these details. In Section 7, we present results for two more topic pairs. Finally, in Section 8, we draw some conclusions about the formulation of topic properties.

2. Topic Pairs for Study

Statistical analysis suggests the following two pairs of TREC-7 topics for careful study, topics 372 and 379 for study with respect to the 40 best systems and topics 372 and 391 for study with respect to the 14 best automatic systems that use the title, description, and narrative parts of the topic statement. These three topics are

Number: 372
Title: Native American casino
Description: Identify documents that discuss the growth of Native American casino gambling.
Narrative: Relevant documents include discussions regarding Native American casino gambling: its social implications, effects on local and Native American economies, and legal aspects related to Native American tribal autonomy.

Number: 379
Title: mainstreaming
Description: Identify documents that discuss mainstreaming children with physical or mental impairments.
Narrative: A relevant document will include the pros and cons of mainstreaming children with physical or mental impairments, the benefits to the impaired child, as well as the attitude, beliefs and concerns of teachers and school administrators with regard to taking time away from the "normal children."

Number: 391
Title: R&D drug prices
Description: Identify documents that discuss the impact of the cost of research and development (R&D) on the price of drugs.
Narrative: Documents that describe how any aspect of the development of a drug affects its price are relevant. Documents that discuss other factors that affect drug prices, such as advertising, without also discussing R&D costs, are not relevant.

Solely on the basis of reading these topics, one might guess that the shared topic property that dominates system performance involves words not in the topic statements that most people would associate with the topics based on their knowledge of current events. For example, association of "Native American" with "reservation" would be helpful in retrieving documents relevant to topic 372. Inclusion of the phrase "public education" might be a useful addition to "mainstreaming" in retrieving documents for topic 379. The words "R&D," "drug," and "price" have variants such as "biotechnology" and "return on investment" that would be useful in a search on topic 391.

The reason that statistical analysis of system performance connects these topics is the appearance of common system successes and common system failures in expansion of these topics. In particular, manual systems seem to do relatively well with topics 372 and 379 whereas automatic systems do relatively poorly.

In trying to conceive of the topic property common to the members of these pairs, one must also recognize other topic properties in which these topics differ. For example, on one hand, "Native American" and "mainstreaming" are terms that must be interpreted only according to their specialized meaning if the search is to be completely successful. On the other hand, to "R&D," "drug" and "price" there correspond equivalent phrases that could be substituted and that must be recognized in a successful search. Conceptualization of a topic property requires attention to such differences.

3. Data Analysis

Comparison of two topics in search of a common topic property involves for each topic, the statement and the performance for a group of systems (or, as some might prefer to say, system variants). We depict performance versus system graphically. As detailed in this section, we use two performance measures, average precision and depth at 25 percent recall. Moreover, we divide performance into components and graph the overall component and the distinctive component separately.

Consider first the determination of average precision. Although the reader may be familiar with this measure through TREC publications, a somewhat different account seems useful because it facilitates comparison with our other measure. Performance for a particular system and topic is based on 1000 documents that the system has identified and ranked according to relevance to the topic. Both performance measures are computed from the ranks of documents deemed relevant by the assessor. In increasing order, we denote these ranks by $r_1, r_2, \ldots, r_i, \ldots$. The ratio i/r_i might be regarded as an estimate of the rate at which relevant documents are discovered. Except for adjustment for the relevant documents not discovered, the average precision is the average of these rate estimates. The adjustment consists of regarding the undiscovered relevant documents as having infinite rank, $r_i = \infty$. Thus, if there are n_R relevant documents of which n are discovered by a system, the average precision for that system is given by

$$P = \frac{1}{n_R} \sum_{i=1}^{n} \frac{i}{r_i}.$$

Measures like depth at 25 percent recall have been suggested for example, by E. M Keen (1997). The determination of this depth involves at most two ranks. Roughly, this depth is the rank r_q, where q is the integer part of $.25n_R$, $q = [.25n_R]$. If $\Delta = .25n_R - q$ is greater than 0 and if $q < n$, then we interpolate using $(1 - \Delta)\, r_q + \Delta\, r_{q+1}$. If $q > n$, then we compute the extrapolation $.25n_R\, r_n / n$, and if its value is greater than 1000, we use it. What we do in other cases is shown below. In addition, we subtract $.25n_R - 1$ so that if 25 percent of the relevant documents are found before any other documents, the measure does not depend on the number relevant for the topic. On the basis of these considerations, our algorithm for depth at 25 percent recall is

$$r_{.25} = \begin{cases} (1-\Delta)\, r_q \; + \; \Delta\, r_{q+1} \; - \; (.25 n_R - 1) & \text{if } 0 < q < n \\ \max[.25 n_R\, r_n / n,\; (1-\Delta)\, r_n \; + \; \Delta\, 1001] \; - \; (.25 n_R - 1) & \text{if } 0 < q = n \\ \max[.25 n_R\, r_n / n,\; 1001] \; - \; (.25 n_R - 1) & \text{if } 0 < n < q \\ \infty & \text{otherwise} \end{cases}$$

Assignment of ∞ to the case in which the system returns no relevant documents, $n = 0$, requires comment. As detailed below, we limit our analysis to the better systems and to easier topics so that $n = 0$ occurs infrequently. In the analysis reported below, a few cases remain and for these we let $r_{.25} = 1500$. This artifice might not be satisfactory were there more than a few. Our analysis is based not on $r_{.25}$ but on $-\log_{10}(r_{.25}) = \log_{10}(1/r_{.25})$. Use of the logarithm reduces the influence of the larger ranks, and use of the minus sign causes larger values to correspond to better performance as is the case with average precision.

These performance measures depend differently on how many relevant documents there are in the collection. If one were to remove half the relevant documents from the collection, the average precision would be more or less cut in half whereas the depth at 25 percent recall would be nearly unchanged. Moreover, undiscovered relevant documents are treated differently in determining each measure. One might argue that one measure is better than the other, but consideration of both seems better than choosing one.

For each performance measure, we compare topics by means of two graphs, one showing the overall component and the other showing the distinctive component. These components are computed from the performances for all topics and systems. One graph compares the two topics in terms of the overall abilities of the systems. The second graph compares the two topics in terms of deviations from these overall abilities. What is perhaps most interesting is the topic-to-topic similarity of the distinctive components as shown in the second graph.

Computation of the overall component requires three steps. Let the number of systems be N_s, the number of topics be N_t, and the performance measure for system i and topic j be y_{ij}. The difficulty of topic j is

$$\hat{\alpha}_j = \frac{1}{N_s} \sum_{i=1}^{N_s} y_{ij}$$

and the (centered) average performance of system i is

$$\hat{x}_i = \frac{1}{N_t} \sum_{j=1}^{N_t} (y_{ij} - \hat{\alpha}_j).$$

In addition, we include a term that describes variation from topic to topic in the effect of overall system abilities:

$$\hat{\beta}_j = \frac{\sum_{i=1}^{N_s} (y_{ij} - \hat{\alpha}_j - \hat{x}_i)\, \hat{x}_i}{\sum_{i=1}^{N_s} \hat{x}_i^2}.$$

One can think of $\hat{\beta}_j$ as representing the degree to which topic j can distinguish the overall abilities of the systems. One can also think of $\hat{\beta}_j$ as representing the degree to which system features that contribute to performance over all topics are effective with topic j. The overall component is given by

$$\hat{\alpha}_j + (1 + \hat{\beta}_j)\, \hat{x}_i.$$

What is portrayed by this component is somewhat familiar to TREC participants.

What is perhaps of more interest is the remainder

$$y_{ij} - \hat{\alpha}_j - (1 + \hat{\beta}_j)\, \hat{x}_i.$$

from which we determine the distinctive component. This remainder reflects interactions, cases where after adjustment for overall performance, one system is better than another for one topic but not for another topic. For example, a system that makes use of only the topic title might do relatively well with titles that adequately delimit the subject but do relatively poorly with diffuse titles. Interactions suggest improvements because they suggest that a system might be assembled from the better parts of existing systems.

The remainder is, however, not easy to interpret because it is noisy. Some appearances of interaction do not reliably predict the results of subsequent evaluations and are not a sound basis for system development. In Section 5, we discuss how we reduce the noise in this remainder and thereby determine the distinctive component. Also, we discuss how we select topics that are worthy of study because their behavior is predictive of future evaluations. Note that adding the remainder to the overall component gives the original performance data. Thus, except for some noise, adding the distinctive component to the overall component gives the original data. The reader will sometimes want to consider the sum of these components.

For our analysis, we select only better systems because differences between good systems are more interesting than other differences. We select the better systems using the median to summarize performance over the topics. The reason for choice of the median is that we would be interested in a generally good system that did poorly on occasion. We adjust the system-topic performance for the overall difficulty of the topic and then find medians over the topics. The algorithm we actually use is Tukey's median polish (Velleman and Hoaglin, 1981). Since we want the same set of systems for analysis by each performance measure, we select the better systems on the basis of depth at 25 percent recall.

Having selected a set of systems, we then select a set of topics. First, we select only topics with more than 10 relevant documents. Second, for the TREC-6 data analyzed in Section 7, we count for each topic the number of selected systems with $r_{.25}$ greater than 2000. We select topics for which this count is no greater than 5. The purpose of eliminating topics is to prevent a few very difficult topics, especially ones for which several systems found no relevant documents, from obscuring more general behavior.

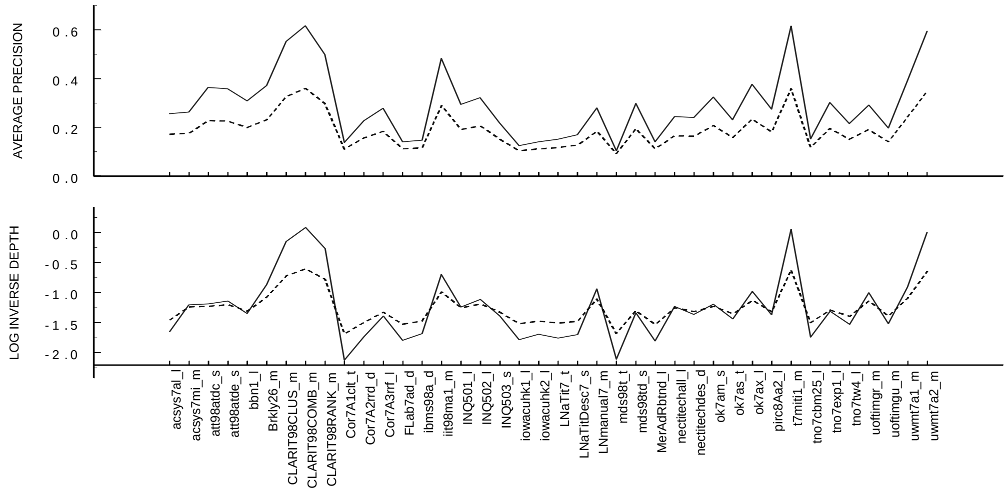

Figure 1. Overall component for topics 372 (solid line) and 379 (dashed line).

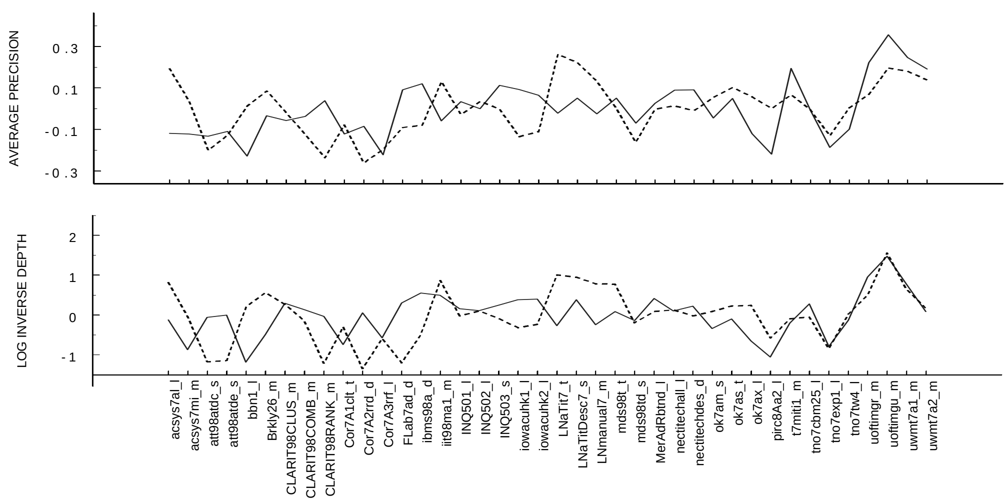

Figure 2. Distinctive component for topics 372 (solid line) and 379 (dashed line).

4. An Interpretation of the Data

Figures 1-6 show performance versus system as detailed in Section 3. In terms of abbreviations explained elsewhere in this Proceedings, system names are given on the horizontal axis with query type appended to each system name: “t” for title only, “d” for description only, “s” for title and description, “l” for title, description and narrative, and “m” for manual. Either average precision or the logarithm of the inverse of the depth is given on the vertical axis. It is the distinctive components shown in Figs. 2, 4, and 6 that suggest that the topic pairs 372-379 and 372-391 share the need for effective topic expansion. Nevertheless, we first consider the overall component shown in Figs. 1, 3 and 5.

In Figs. 1, 3 and 5, we observe that the system-to-system variation in the overall component is larger for one topic than the other. System-to-system variation in the overall component reflects the average performance over all topics. Thus, a topic that varies more with system is related more strongly to average performance. In other words, whatever makes a system perform better or worse for all topics has a greater effect for this topic. Conversely, the topic that varies less with system can be considered to be more singular, less related to all the other topics. In Figs. 1 and 3, we see that topic 379 (mainstreaming) is more singular than topic 372 (Native American casino). In Fig. 5, we see that topic 372 is more singular than topic 391 (R&D drug prices). These observations seem reasonable.

Another thing of note in these figures is that when average precision is considered, one topic has better performance than the other for all systems but when depth at 25 percent recall is considered, this does not hold. A topic that has better performance for all systems can be regarded as an easier topic. The discrepancy between the two measures can be largely explained by the dependence of average precision on the number of relevant documents. The number of relevant documents is 16 for topic 379, 49 for topic 372, and 178 for topic 391. We conclude that there is little in these figures to suggest that one topic in a pair is easier than the other.

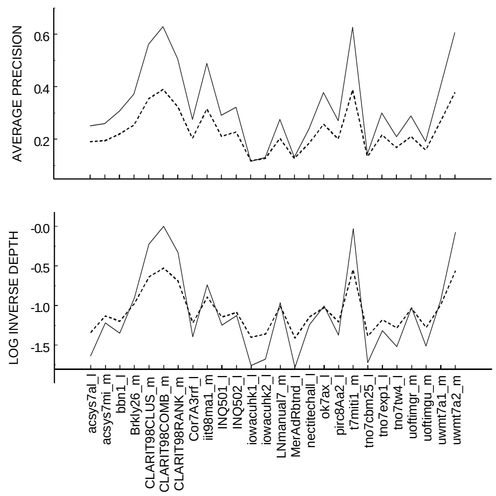

Figure 3. Overall component for topics 372 (solid line) and 379 (dashed line).

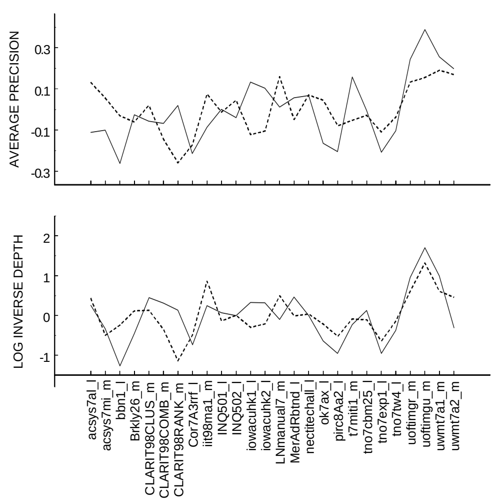

Figure 4. Distinctive component for topics 372 (solid line) and 379 (dashed line).

Figures 2, 4, and 6 show that the topic pairs 372-379 and 372-391 share one or more topic properties related to challenges in topic expansion. This is most clearly shown in Fig. 4 where manual systems largely outperform automatic systems. Presumably, manual systems provide more effective topic expansion. Figure 6, which is based only on automatic systems, shows that some systems perform better than others. One would guess that this is due to better topic expansion.

The figures in this paper reflect the three criteria we use in selecting pairs. Consider Figs. 1 and 2. First, one can see that these topics are unusual by comparing Figs. 1 and 2 and noting that the variation from system to system in the distinctive component is roughly the same size as the variation in the overall component. For a nearly average topic, the variation in the distinctive component would be much smaller. Second, the two topics are at least somewhat similar as shown in Fig. 2 by the fact that the distinctive components largely vary together. Third, this similarity holds for both performance measures as also shown in Fig. 2.

There are further remarks one can make about Figs. 1 and 2. Beyond what has already been said about Fig. 1, note that the two performance measures show the same pattern of system-to-system variation. The agreement between the topics as shown in Fig. 2 is not as compelling as one might like. The agreement seems better on the right than the left. There are particular systems that show agreement such as the system "uoftimgu," which is notable because the distinctive component is high for both topics, and the system "tno7exp1," which is notable because the distinctive component is low for both topics. Also notable are the results for the systems "bbn1" and "CLARIT98RANK" for which there is disagreement. It seems possible to infer the reasons for these disagreements from the papers on these systems elsewhere in this Proceedings. Note in particular the comparison of "CLARIT98CLUS" and "CLARIT98RANK." Generally, one might guess that agreement is not better because all the different query types are included.

We repeated our analysis for just the 26 systems in Figs. 1 and 2 that make use of the entire topic statement, query types "l" and "m." Again, our statistical method chose topics 372 and 379 as different from the rest and similar to each other. The overall component in Fig. 3 is not remarkably different from what is shown in Fig. 1. The distinctive component in Fig. 4 shows a group of manual systems for which relative performance is high for both topics and a group of automatic systems for which relative performance is low for both topics. This seems to be evidence that the need for topic expansion noted in Section 2 can be more effectively achieved with a manual system. Conversely, the topic property that may be inferred from this pair of topics is the need for a type of topic expansion that can be better provided by human interaction with the system than automatically.

Seemingly of interest would be the automatic systems that use the entire topic statement. We repeated the analysis for the 14 systems in Figs. 1 and 2 denoted "l". The two topics that emerged from this analysis are 372 and 391. One would expect a different pair of topics because the choice of pair depends on the systems that enter the analysis. Figures 5 and 6 show the overall and distinctive components for these two topics. Observations on Fig. 5 have largely been covered by our general remarks on the overall component. Figure 6 shows 6 systems, "INQ501", "INQ502", "iowacuhk1", "iowacuhk2", "MerAdRbtnd", and "nectitechall", that did relatively well with both of these topics. Rather than guessing what the underlying

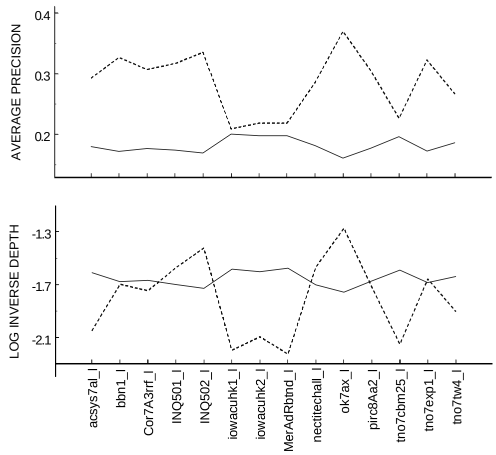

Figure 5. Overall component for topics 372 (solid line) and 391 (dashed line).

topic property is, we ask the following interesting question: What features of these systems causes them to favor topics 372

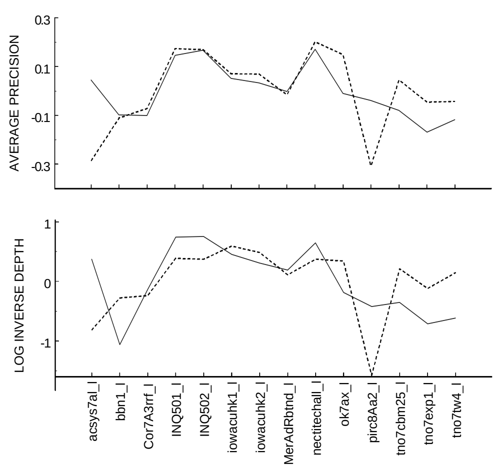

Figure 6. Distinctive component for topics 372 (solid line) and 391 (dashed line).

and 391 when overall the performance of these systems is variable, both better and worse than the average? This is the kind of connection between topic and system that could lead to system improvements. Figure 6 also shows two systems, "acsys7al" and "pirc8Aa2", that did relatively better with topic 372 than topic 391. Did the topic expansion strategies used by these systems fail when applied to topic 391?

The interpretation of Figs. 1-6 is not yet complete because the specific features of the systems portrayed have not yet been fully considered. Some further information useful for interpretation may be included in the system descriptions presented elsewhere in this Proceedings. In any case, those responsible for specific systems may be able to add to the above interpretation.

5. The Distinctive Component

Our argument for studying the topics discussed in Section 2 as well as other topics selected in Section 7 is not based on understanding of information retrieval systems but on statistics. We now present statistical methodology for selecting topics. The purpose of this methodology is interpretation of the $N_s \times N_t$ matrix with elements

$$y_{ij} - \hat{\alpha}_j - (1 + \hat{\beta}_j)\hat{x}_i.$$

In interpreting these residuals, one faces the challenge of identifying what is substantial while disregarding appearances that might not generalize beyond the immediate data.

What are we looking for in this matrix? Say that the matrix were given by the product of two column vectors, $\boldsymbol{\gamma}\boldsymbol{h}^T$, where superscript T denotes transpose. Were this true, the distinctive component for topic j would be $h_j\boldsymbol{\gamma}$. Thus, except for a multiplicative constant, the distinctive component would be the same for each topic. This would imply the existence of a topic property that has for topic j the value h_j (on a scale that is arbitrary up to a linear transformation). We see that interaction of this sort leads directly to a topic property of the type we seek.

Opposite the case of pure interaction is the case of pure noise, the case in which the residual matrix suggests no topic properties that apply beyond the performance of a single system. Banks, et al. (1999) consider this case. They indicate that if the coefficients $\hat{\beta}_j$ are significantly different from zero, then one should conclude that the residual matrix will likely exhibit some interactions. We have included the $\hat{\beta}_j$ coefficients in the overall component. Figs. 1, 3, and 5 suggest that these coefficients are significant. We have applied the appropriate hypothesis test to confirm this.

Beyond the question of whether the residual matrix appears to be pure noise is the possibility that one will over-interpret this matrix. Perhaps our best defense against this possibility is the use of two different performance measures. One might guess that because these measures are calculated differently, the measurement errors for each are largely independent. As our defense, we recommend a pair of topics for study only when they appear as worthy of study according to both measures.

The singular value decomposition provides an analysis of the residual matrix

$$y_{ij} - \hat{\alpha}_j - (1 + \hat{\beta}_j)\,\hat{x}_i = \sum_{m=1}^{M_0} d_m u_{im} v_{jm}.$$

The upper limit on the sum M_0 is the smaller of $N_s - 2$ and $N_t - 1$. The coefficients d_m are positive and are ordered according to decreasing size. The vectors $\boldsymbol{u}_m = (u_{im})$ are orthonormal and orthogonal to the vector $\mathbf{1} = (1)$ as are the vectors $\boldsymbol{v}_m = (v_{jm})$. The vectors $\boldsymbol{u}_m$ are also orthogonal to the vector with elements $\hat{x}_i$. In vector notation, the residual matrix is given by

$$\sum_{m=1}^{M_0} d_m \boldsymbol{u}_m \boldsymbol{v}_m^T.$$

Note that the decomposition is a sum of interaction-like terms. If only one term were non-zero, then we would have the case of pure interaction discussed above.

Our approach to separating the distinctive component from the noise is separation of the larger terms in the decomposition from the rest. Based on the size of the coefficients d_m, we choose M terms for the distinctive component, which corresponds to supplanting the residual matrix with a smoothed residual matrix

$$\sum_{m=1}^{M} d_m \boldsymbol{u}_m \boldsymbol{v}_m^T.$$

Figures 2 and 4 are based on $M = 7$; Fig. 6 is based on $M = 5$; and the figures in Section 7 are based on $M = 4$. There is a considerable literature on choosing the number of common factors in factor analysis that applies to the choice of M. In the current situation, we expect that the results will not be particularly sensitive to the choice. We could vary M and see how this affects our choice of topics for study.

One way of thinking about explaining the smoothed residual matrix is to think about reducing the sum of squares of its elements. Each column in this matrix corresponds to a topic and thus to the distinctive component for that topic. We see that the distinctive component is the sum of M terms $d_m v_{jm} \boldsymbol{u}_m$. The sum of squares for each term is $d_m^2 v_{jm}^2$, and the sum of squares for the distinctive component for topic j is

$$\sum_{m=1}^{M} d_m^2 v_{jm}^2.$$

Summing over all the topics, we obtain

$$\sum_{m=1}^{M} d_m^2.$$

We choose topics on the basis of the degree to which each explains this total sum of squares.

Retained in the smoothed residual matrix are the interactions that are of interest, but, unless $M = 1$, we must do more to bring out the character of the interactions. We would like to find individual and pairs of topics that are strongly associated with the interactions. Such topics would appear in the smoothed residual matrix as a term of the form $\boldsymbol{\gamma}\boldsymbol{h}^T$, where $\boldsymbol{h}$ contrasts one or two topics with all the other topics. We choose $\boldsymbol{\gamma}$ so that $\boldsymbol{\gamma}\boldsymbol{h}^T$ matches the smoothed residual matrix as closely as possible. We take $\boldsymbol{h}$ to be a unit vector that is orthogonal to $\mathbf{1}$, $\boldsymbol{h}^T\boldsymbol{h} = 1$, $\boldsymbol{h}^T\mathbf{1} = 0$. A gauge of this match is given by

$$\sum_{m=1}^{M} d_m^2 - \sum_{i=1}^{N_s} \min_{\gamma_i} \left[\sum_{j=1}^{N_t} \left(\sum_{m=1}^{M} d_m u_{im} v_{jm} - \gamma_i h_j \right)^2 \right].$$

The first term in this gauge is the sum of squares of the elements in the smoothed residual matrix.

The most that the total sum of squares can be reduced by a rank one approximation is d_1^2. This would be obtained with $h_j = v_{j1}$ and $\gamma_i = d_1 u_{im}$.

Working through the minimization that is part of computing our gauge and dividing the result by d_1^2, we obtain what we call the fraction explained by the contrast $\boldsymbol{h}$

$$\sum_{m=1}^{M} d_m^2 (\boldsymbol{v}_m^T \boldsymbol{h})^2 / d_1^2.$$

We begin by asking which topic explains the largest part of the total sum of squares. To compute this for topic j, we let $\boldsymbol{h} = (h_k)$, where

$$h_k = \begin{cases} \sqrt{1 - 1/N_t} & \text{if } k = j \\ -1/(N_t \sqrt{1 - 1/N_t}) & \text{if } k \neq j \end{cases}.$$

The topic that explains the largest amount, we term the most unusual topic, because it corresponds most closely to the part of the residual matrix that exhibits the important interactions.

One could ask what topic property causes the most unusual topic to be so. Generally, however, the answer would be a list of possible topic properties ranging from the subject matter of the topic to the phrasing used to convey the topic. Asking what is in common between a pair of topics is more likely to be fruitful. We ask what contrast between two topics and the others explains a large part of the total sum of squares. To compute this for topics j and j', we let

$$h_k = \begin{cases} \sqrt{1/2 - 1/N_t} & \text{if } k = j \\ \sqrt{1/2 - 1/N_t} & \text{if } k = j' \\ -1/(N_t \sqrt{1/2 - 1/N_t}) & \text{otherwise} \end{cases}.$$

As is the case for the topics discussed in Section 4, the fraction explained for some pairs of topics is larger than for the most unusual single topic. Topic pairs for which this is true are both strongly associated with the interactions and have similar distinctive components. If the distinctive components were not similar but each topic were by itself unusual, the contrast for the pair would likely not explain much of the smoothed residual matrix because the two topics would partially cancel each other.

With one minor exception, each topic pair in Section 4 explains, according to both measures, more than the most unusual topic. The existence of such pairs provides a strong incentive for study of the question of what the topics have in common. However, there is no guarantee that such a pair will occur. For the systems and topics considered in Section 4, there were other topic pairs that explained more than the most unusual topic but these did not turn up for both measures. We would like a pair chosen for study to be unusual but requiring that the pair explain more than the most unusual topic may be too stringent. Perhaps if the amount explained were within 0.10 of the most unusual topic, this would be enough.

6. Steps in Choosing Topic Pairs

Sections 3 and 5 discuss the statistical methods that we use to choose topic pairs. In this section, we detail the application of these methods that produced the topic pair 372-379 shown in Figs. 1-2.

Clearly, a comparison of topics based on system performance depends on the set of systems considered. There are 103 systems that produced TREC-7 ad hoc task results. We chose 40 for Figs. 1 and 2, 26 for Figs. 3 and 4, and 14 for Figs. 5 and 6. We chose the systems with the best performance (according to a particular criterion) because we believe that in general such a choice produces the most interesting results. On the other hand, there is no reason why the methods described in this paper should not be applied to other sets of systems.

Having chosen a set of systems, we eliminate some topics in part to accommodate our depth measure and in part to eliminate difficult topics that might have performance profiles very different from other topics. In choosing the topic pairs for Figs. 1-6, we eliminated topics 361 and 380 because, for each of these, the number of relevant documents is less than 10. For the analysis of the TREC-6 data in Section 7, we eliminated 10 topics, some because of the number of relevant documents and some because too many systems found no relevant documents for the topic.

Computation of the residual matrix is detailed in Section 3. In the analysis for Figs. 1 and 2, application the singular value decomposition to the residual matrix for log inverse depth gives as the coefficients d_m, 6.81, 6.58, 5.78, 5.70, 5.38, 4.73, 4.11, 3.85, 3.80, 3.63, 3.36, 3.06, 2.84, 2.63, 2.41, 2.29, 2.13, 1.89, 1.79, 1.74, 1.55, 1.41, 1.36, 1.31, 1.20, 0.97, 0.90, 0.81, 0.78, 0.62, 0.59, 0.58, 0.44, 0.41, 0.38, 0.24, 0.18, and 0.16. Application to the residual matrix for average precision gives 1.70, 1.65, 1.45, 1.30, 1.22, 1.10, 1.03, 0.93, 0.89, 0.84, 0.73, 0.69, 0.65, 0.63, 0.60, 0.52, 0.49, 0.45, 0.43, 0.40, 0.36, 0.33, 0.30, 0.29, 0.24, 0.24, 0.20, 0.18, 0.18, 0.14, 0.12, 0.10, 0.09, 0.08, 0.08, 0.06, 0.04, and 0.03. Based on the gaps between the seventh, eighth, and ninth coefficients, we choose $M = 7$ for the smoothed residual matrix for each measure. In some studies, the researcher might choose M with the idea that this is the number of common factors and that each common factor should be seriously considered. We do not do this, although when more has been learned about topic properties, one might proceed in this way.

As discussed in Section 5, our choice of the topic pair 372-379 in the 40 system context is based on values of the fraction explained. For the depth measure, the topic that alone explains the most is topic 379 with fraction explained of 0.39. Combinations of two topics that explain more along with their fraction explained are 372-379, 0.41 and 397-398, 0.48. For average precision, the topic alone that explains the most is topic 398 with fraction explained of 0.32. Combinations of two topics that explain more are 372-379, 0.33; 372-391, 0.33; and 375-398, 0.41. The pair in common between the two measures is 372-379. For this reason, we selected this pair for Figs. 1 and 2. Clearly, one could investigate other pairs. However, real progress may require focus on a single pair until one is convinced that one has gone as far as possible in finding the common property.

7. Two More Topic Pairs

Along the lines of Sections 2, 4, and 6, we now present two more topic pairs. These pairs suggest somewhat different topic properties.

Complementing our choice in Section 4 of automatic systems that use the whole topic, we now choose the 14 automatic systems from Figs. 1 and 2 that use the topic title, the description, or both. The two topics that emerged from this are

Number: 352
Title: British Chunnel impact
Description: What impact has the Chunnel had on the British economy and/or the life style of the British?
Narrative: Documents discussing the following issues are relevant:
 - projected and actual impact on the life styles of the British
 - Long term changes to economic policy and relations
 - major changes to other transportation systems linked with the Continent

Documents discussing the following issues are not relevant:
 - expense and construction schedule
 - routine marketing ploys by other channel crossers (i.e., schedule changes, price drops, etc.)

Number: 385
Title: hybrid fuel cars
Description: Identify documents that discuss the current status of hybrid automobile engines, (i.e., cars fueled by something other than gasoline only).
Narrative: A relevant document may include research on non-gasoline powered engines or prototypes that may be fueled by natural gas, methanol, alcohol; cost to the consumer; health benefits derived; and shortcomings in horsepower and passenger comfort.

The overall component shown in Fig. 7 implies that topic 385 (dashed line) is easier than topic 352. (The number relevant is 246 for topic 352 and 86 for topic 385.) Note that the weakest performance (over all topics) occurs for two (of the three) title-only systems.

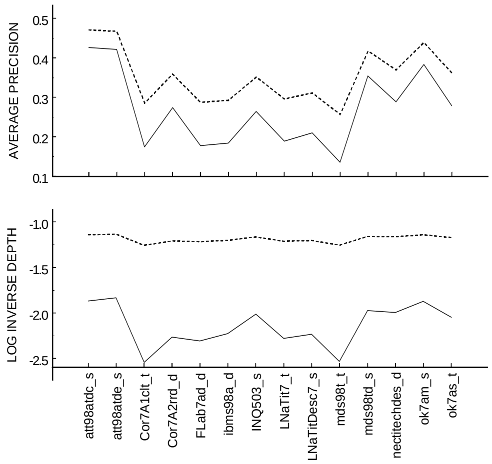

Figure 7. Overall component for topics 352 (solid line) and 385 (dashed line).

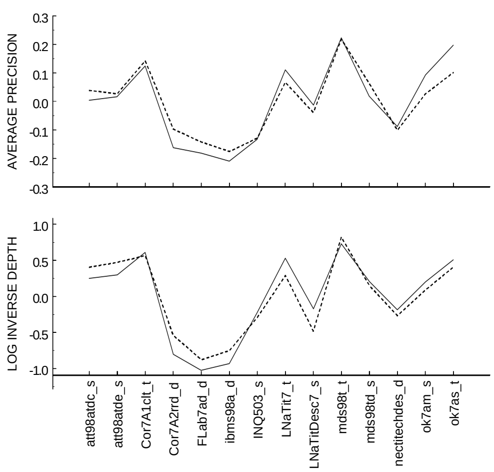

Figure 8. Distinctive component for topics 352 (solid line) and 385 (dashed line).

The distinctive component shown in Fig. 8 exhibits a very close relation between the two topics. The title-only systems perform relatively better than description-only systems in terms of the distinctive component, and, moreover, the title-only runs would perform better if viewed from the sum of the overall and distinctive components. Reasons for this are suggested by the title and description parts of the topic statements. First, for each topic, the key noun phrase differs between title and description. From title to description, topic 352 goes from "British Chunnel" to "Chunnel," and topic 385 goes from "hybrid fuel cars" to "hybrid automobile engines." Second, the more discursive form of the description adds various noun phrases that may do little to make the query more specific. Clearly, the difference in style between the title and the description hurts system performance although a person would say that the description more specifically conveys what is relevant.

For the other pair, we turn to TREC-6. The two topics that emerged are 312 and 316. These topics are

<u>Number:</u> 312
<u>Title</u>: Hydroponics
<u>Description</u>: Document will discuss the science of growing plants in water or some substance other than soil.
<u>Narrative</u>: A relevant document will contain specific information on the necessary nutrients, experiments, types of substrates, and/or any other pertinent facts related to the science of hydroponics. Related information includes, but is not limited to, the history of hydroponics, advantages over standard soil agricultural practices, or the approach of suspending roots in a humid enclosure and spraying them periodically with a nutrient solution to promote plant growth.

<u>Number</u>: 316
<u>Title</u>: Polygamy Polyandry Polygyny

<u>Description</u>: A look at the roots and prevalence of polygamy in the world today.
<u>Narrative</u>: Polygamy is a form of marriage which permits a person to have more than one husband or wife. Polyandry refers to one woman sharing two or more husbands at the same time. Polygyny refers to one man sharing two or more wives at the same time. Primary focus of the search will be the prevalence of these practices in the world today and societal attitudes towards these practices. Also relevant would be discussions of the roots and practical sources of these customs. A modern development in this area is serial polygamy, a phrase coined to label the practice of men who take a series of wives in sequence as a solution to practical welfare, considerations of child care, housing, etc. Documents discussing serial polygamy will not be considered relevant.

The overall component shown in Fig. 9 implies that topic 312 (solid line) is easier than topic 316. The number relevant for topic 312 is 11, and the number relevant for topic 316 is 35.

The distinctive component shown in Fig. 10 shows better performance for four systems, three manual and one automatic. Topics 312 and 316 are notable because of the existence of very specific key words, "hydroponics" and "polygamy." The manual systems that did well seem better able to take advantage of these key words than automatic systems. This may be because a person is better able to see that for these topics, there exist exceptional key words. Automatic systems may go astray by including extraneous words during the process of topic expansion. Voorhees and Harman (1998) note this in their TREC-6 overview.

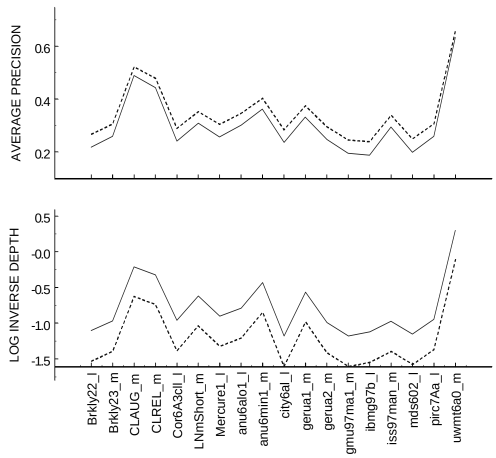

Figure 9. Overall component for topics 312 (solid line) and 316 (dashed line).

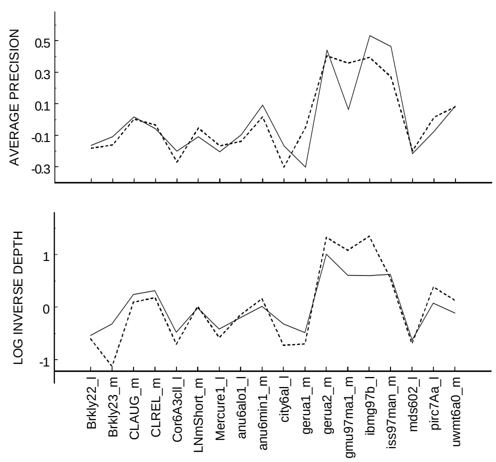

Figure 10. Distinctive component for topics 312 (solid line) and 316 (dashed line).

8. Conclusions

This paper offers four pairs of topics chosen by statistical methods. Faced with the challenge of hypothesizing what each pair has in common, it seems that one has some basis for a response. One topic property that seems to have surfaced is the need for parsimonious topic expansion, and the other is the possibility of confusion when the topic style leads to expansion beyond a succinct delimiting of the topic.

References

D. Banks, P. Over, and N. Zhang (1999). "Blind Men and Elephants: Six Approaches to TREC Data," To appear in *Information Retrieval*.

E. M Keen (1997). "Presenting Results of Experimental Retrieval Comparisons," In K. Spark Jones and P. Willett, editors, *Readings in Information Retrieval*, pages 217-222, San Francisco, CA: Morgan Kaufmann Publishers.

P. F. Velleman and D. C. Hoaglin (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*, Boston, MA: Duxbury Press.

E. M. Voorhees and D. Harman (1998). Overview of the Sixth Text REtrieval Conference (TREC-6). In E. M. Voorhees and D. K. Harman, editors, *Proceedings of the Sixth Text Retrieval Conference (TREC-6)*, pages 1-24, NIST Special Publication 500-240, Washington, DC: U.S. Government Printing Office.