What Makes a Query Difficult?

David Carmel, Elad Yom-Tov, Adam Darlow, Dan Pelleg
IBM Haifa Research Labs
University Campus
Haifa 31905
Israel
{carmel,yomtov,darlow,dpelleg}@il.ibm.com

ABSTRACT

This work tries to answer the question of what makes a query difficult. It addresses a novel model that captures the main components of a topic and the relationship between those components and topic difficulty. The three components of a topic are the textual expression describing the information need (the query or queries), the set of documents relevant to the topic (the Qrels), and the entire collection of documents. We show experimentally that topic difficulty strongly depends on the distances between these components. In the absence of knowledge about one of the model components, the model is still useful by approximating the missing component based on the other components. We demonstrate the applicability of the difficulty model for several uses such as predicting query difficulty, predicting the number of topic aspects expected to be covered by the search results, and analyzing the *findability* of a specific domain.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Query difficulty

1. INTRODUCTION

The classical information retrieval (IR) evaluation paradigm, as provided by TREC and by other similar benchmarks, involves measuring the ability of a search system to retrieve relevant documents in response to a set of typical information needs (called *topics* in the context of TREC). Topics are defined by two components: (a) a textual description, and (b) a set of documents relevant to the information need. The topic's textual description comes in a natural language form that can be used to identify the expected user's query to be handled (e.g., a short query based on the topic title, or a longer query based on the topic description). The set of relevant documents (called Qrels) are then used for scoring a search system running that query for measuring general system effectiveness and for comparison between systems.

The experimental results of current IR systems participating in TREC show a wide diversity in effectiveness among topics as well as among systems. Most systems, even with high precision on average, fail to answer some of the topics. This led to the TREC Robust tracks [15, 16, 17], which encouraged systems to decrease variance by focusing on poorly performing topics. In the Robust tracks of 2003 and 2004, systems were challenged by 50 old TREC topics found to be "difficult" for most systems over the years. A topic is considered difficult in this context when the median of the average precision (AP) scores of all participants for that topic is below a given threshold (i.e., half of the systems score lower than the threshold). One of the track's goals was to study whether a topic found to be difficult several years ago is still difficult for current state-of-the-art IR systems. The definite conclusion was that current systems still have difficulty in handling those old difficult topics [15, 16].

However, the Robust track results did not fully answer the basic question underlying its root cause, that is, why are some topics more difficult than others? In this paper, we study the reasons for topic difficulty and show possible uses of the knowledge derived from understanding these root causes.

1.1 Related work

In the past, several attempts were made to explain the sources of topic difficulty, which can be linked to a number of causes. Topic difficulty might be induced from the topic textual expression, but also from the resources of available information. Cronen-Townsend et al. [2] suggested the clarity measure that tries to explain query difficulty through differences between the language model of the query and that of the collection. Prager [13] observes that "understanding the question is indeed part of the process, but it is only a part. Understanding the corpus is equally important, as is being able to match these resources to the question".

The Reliable Information Access (RIA) workshop [6] was the first to rigorously investigate the reasons for variability between systems success on different topics. The goal of the RIA workshop was to understand the contributions of both system variability factors and topic variability factors

SIGIR'06, August 6–10, 2006, Seattle, Washington, USA.

to overall retrieval variability. The workshop brought together seven different top research IR systems and assigned them to common tasks. Comparative analysis of these different systems enabled isolation of system variability factors. One of the workshop's findings was that the variability is due to the topic statement itself, the relationship between the topic and the document collection as a whole, and some system dependent factors such as the retrieval algorithm and its implementation.

By performing failure analysis on TREC topics, ten failure categories were identified. Table 1 presents the failure categories, each associated with an example topic, as they appear in the workshop summary. Five categories of the ten relate to the systems' failure to identify all aspects of the topic. One of the workshop's conclusions was that the root cause of poor performance is likely to be the same for all systems. Systems are retrieving different documents from each other in general, but in most categories, all systems fail for the same reasons.

One of the tasks in the 2004 and 2005 TREC Robust tracks was to estimate the relative difficulty of each topic. Several works attempted to achieve this task (reviewed in [19]), even though mostly without explicitly explaining the source of the difficulty. Features which were somewhat useful for prediction of topic difficulty were related to the document collection (e.g., the frequency of query terms in the collection [7, 9]), the score of the top-scored document [14], and the overlap between results of sub-queries based on single query terms and results of the full query [19].

In this context, Mothe and Tanguy [11] searched for correlations between sixteen different linguistic features of TREC queries and the average precision scores. Each of these features can be viewed as a clue to a linguistically specific characteristic, either morphological, syntactical, or semantic. Two of these features (syntactic links span and polysemy value) had a significant impact on precision scores for previous TREC participants. Although the correlation values were not high, they indicated a link between some linguistic characteristics of the queries and topic difficulty.

Topic difficulty also relates to the coherence of the relevant documents. A set of relevant documents containing distinctive documents representing different aspects of the topic, might be more difficult to retrieve. Evans et al. [4] describe a taxonomy of topics according the number of existing clusters in the result set. The number of resulting clusters and the stability of those clusters provide important clues as to the difficulty of the topic.

Finally, some of the features of the entire collection were also suggested as affecting topic difficulty. One of the questions to answer in the Robust track of 2005 was whether topics found to be difficult in one collection are still considered difficult in another collection. Difficult topics in the TREC disks 4&5 collection were tested against the AQUAINT collection. The average median AP over the 50 topics for TREC disks 4&5 collection (Robust04) is 0.126 compared to 0.185 for the AQUAINT collection (Robust05). Assuming that the median AP score of all participants is a good indication for topic difficulty, these results indicate that the AQUAINT collection is 'easier' (average median AP is higher) than the TREC disks 4&5 collection, at least for the 50 difficult topics of the Robust track[1]. This might be due to the collection size or due to the document features such as length, structure, and coherence. This might also depend on the difference in separability of the Qrels sets from the entire collection. To test whether the relative difficulty of the topics is preserved over the two document sets, we computed the Pearson correlation between the median AP scores of the 50 difficult topics as measured over the two datasets. The Pearson correlation is 0.463, which shows a strong dependency between the median AP scores of a topic on both collections. This suggests that even when results for a topic are somewhat easier to find on one collection than another, the relative difficulty among topics is preserved, at least to some extent.

[1]The conclusions of this comparison should be accepted with caution as the participants of Robust 2004 are different from those of Robust 2005.

1.2 Our approach

In this work, we investigate the main features affecting topic difficulty. We suggest a novel model which captures the main components of a topic and the relationship between those components and topic difficulty. The three components are the topic textual expression describing the information need, the set of relevant documents of the topic, and the entire collection of documents. We argue, and then show experimentally, that topic difficulty mostly depends on the inner relationship between those components.

Rarely does the information pertaining to all parts of the model exist. Users are likely to have only one part of a topic, either the query expression, or the relevant documents (when the problem is analyzed from the content provider's perspective). We show that in such cases the proposed model is still useful for measuring topic difficulty by approximating the missing part using the existing parts.

The rest of the paper is organized as follows: Section 2 describes the model and its relation to topic difficulty. In Section 3 we validate our model through experiments conducted on the .gov2 collection, using the 100 topics of the TREC Terabyte tracks. We show a significant correlation between the topic difficulty model based features and the standard IR measures for topic difficulty such as average precision. Section 4 describes some applications of the model. We demonstrate the ability of the model to predict topic difficulty. In the opposite direction, given a set of documents such as a specific domain, the model can predict how findable this domain is. Section 5 concludes.

2. A MODEL FOR TOPIC DIFFICULTY

As outlined above, a typical information retrieval scenario is comprised of a collection of documents and a search engine that retrieves documents in response to user queries. A user submitting a query to the search engine has an idea of the information she is trying to find. She is also able to judge the search results according to their relevance to this information need. Thus, the query and the relevant documents are two facets of the same information need.

Therefore, we define the primal object of the model to be a *Topic*. A topic is information pertinent to a defined subject. The topic is comprised of two objects: a set of queries, Q, and a set of relevant documents, R. The queries are possible expressions reflecting the information need, while the relevant documents contain the information satisfying that need. The topic is also dependent on the specific document

Category	Topic example
1. General success - present systems worked well	Identify documents that discuss in vitro fertilization
2. General technical failure (stemming, tokenization)	Identify systematic explorations and scientific investigations of Antarctica, current or planned.
3. All systems emphasize one aspect missing another required term	What incidents have there been of stolen or forged art?
4. All systems emphasize one aspect missing another aspect	Identify documents discussing the development and application of spaceborne ocean remote sensing.
5. Some systems emphasize one aspect some another; need both	What disasters have occurred in tunnels used for transportation?
6. All systems emphasize one irrelevant aspect missing point of topic	The spotted owl episode in America.
7. Need outside expansion of "general" term (Europe for example)	Identify documents that discuss the European Conventional Arms Cut as it relates to the dismantling of Europes arsenal.
8. Need QA query analysis and relationships	How much sugar does Cuba export and which countries import it
9. Systems missed difficult aspect that would need human help	What are new methods of producing steel?
10. Need proximity relationship between two aspects	What countries are experiencing an increase in tourism?

Table 1: RIA Topic Failure Analysis Categorization

collection, C, from which R is chosen. Thus, we denote a topic as:

$$Topic = (Q, R|C) \tag{1}$$

For each topic it is important to measure how broad the topic is and how well it is separated from the collection. In terms of clustering, this is akin to measuring the in-cluster variability and the between-class variability. These measurements can be performed on both facets of the model. An additional measurement, which is of even greater interest, is the distance between the two facets of the model, i.e., the distance between Q and R. We hypothesize that a large distance translates to a difficult topic while a small distance results in an easy topic. Figure 1 shows a schema of the topic difficulty model and the different distances among its elements:

1. $d(Q, C)$ - The distance between the queries, Q, and the collection, C. This is analogous to the clarity score of a query [2].
2. $d(Q, Q)$ - The distance among the queries, i.e., the diameter of the set Q.
3. $d(R, C)$ - The distance between the relevant documents, R, and the collection, C.
4. $d(R, R)$ - The distance among the relevant documents, i.e., the diameter of the set R.
5. $d(Q, R)$ - The distance between the queries, Q, and the relevant documents, R.

In some cases, it is possible to obtain only one of the model objects (Q or R). For example, a search engine manager inspecting the search engine query log has access to the queries regarding a certain topic, but the relevant documents to this topic are not supplied. That is, he has access to the documents in the collection, but the documents are not labeled as relevant to a specific topic. In this case, the model is still very useful, as it is possible to estimate the clarity of the topic according to $d(Q, C)$ and also $d(Q, Q)$ distance where the topic is represented by a set of several queries.

Similarly, a content manager might not have access to the specific queries users are typing while trying to find the information in her documents, only to the documents or the web pages she manages. In such cases, the model still can be used to estimate how easily her information can be found, by estimating the $d(R, C)$ and $d(R, R)$ distances. This is similar to the notion of *Findability* in the context of search engine optimization, where the objective is to optimize web pages so that their content is optimally findable.

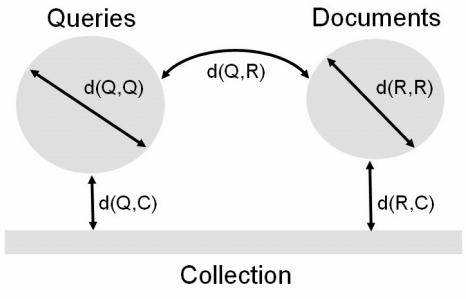

Figure 1: A general model of a topic based on the queries expressing the specific information need, the relevant documents for those queries, the entire collection, and the distances between the sets involved.

In this work, we use the Jensen-Shannon divergence (JSD) to measure distances between objects (sets of documents and queries). The JSD is a symmetric version of the Kullback-Leibler divergence. For the distributions $P(w)$ and $Q(w)$ over the words in the collection $w \in W$, JSD is defined as:

$$D_{JS}\left(P||Q\right) = \frac{1}{2}\left(D_{KL}(P||M) + D_{KL}(Q||M)\right) \tag{2}$$

where $M(w) = 1/2\left(P(w) + Q(w)\right)$, and $D_{KL}(P1||P2)$ is the Kullback-Leibler divergence of $P1$ and $P2$. Thus, explicitly, the Jensen-Shannon divergence can be written as:

$$D_{JS}\left(P||Q\right) = \sum_{w \in W} P(w) \log \frac{P(w)}{M(w)} + \sum_{w \in W} Q(w) \log \frac{Q(w)}{M(w)} \tag{3}$$

The Jensen-Shannon divergence is not a distance (as it does not obey the triangle inequality), but its square root is. The JSD is preferred over other distance measures such as cosine distance, because when measuring distances between documents or queries (as described below), the collection statistics can be naturally incorporated into the measurements.

The measures $d(Q, C)$, $d(R, C)$, and $d(Q, R)$ as defined above are all estimated using the Jensen-Shannon divergence between the centroids of the sets Q, R, and C respectively. The method for estimating $d(Q, Q)$ and $d(R, R)$ is explained

in Section 2.1.

To approximate the distribution of terms within documents or queries, we measure the relative frequencies of terms, linearly smoothed with collection frequencies. The probability distribution of a word w within the document or query x, where w appears n_w times in x, is:

$$P(w|x) = \lambda * \frac{n_w}{\sum_{w' \in x} n_{w'}} + (1 - \lambda) * P_c(w) \quad (4)$$

where $P_c(w)$ is the probability of word w in the collection, and λ is a smoothing parameter. In this work, λ was set to 0.9, except when measuring JSD distance between objects and the collection, where it was set to 0.99.

2.1 Topic aspects as a measure of topic broadness

Most retrieval models assume that the relevance of a document is independent of the relevance of other documents. In reality, however, this assumption rarely holds; relevant documents can relate to different aspects of the topic, hence, the entire utility of the result set strongly depends on the number of relevant aspects it covers.

The *aspect coverage* problem has to do with finding documents that cover as many different aspects of the topic as possible. This problem has been investigated in the interactive track of TREC, where the purpose was to study how systems can best cover all relevant aspects of a topic [12]. Zhai et al. [20] describe some evaluation measures for aspect coverage of a given result set.

The *aspect coverage* problem is another facet of topic difficulty. This is clearly demonstrated in Table 1, where five out of the ten categories relate to poor aspect coverage. Therefore, the broadness of the topic, both from the query facet and the document facet, is measured by the number of aspects described by the topic.

In our model, topic broadness is measured by the distance $d(R, R)$. This distance can be obtained, for example, by measuring the inner JSD distance among the relevant documents. A small distance would reflect a coherent set of relevant documents, all providing the same information. However, this measure suffers from the drawback that identical (or extremely similar) documents are very close together, despite adding no information to the user.

Thus, we opted for using aspect coverage as an indication of topic broadness i.e. topic difficulty. Given a topic with the set of relevant documents, the number of topic aspects was estimated by clustering the relevant documents. Using the square root of the JSD as a distance measure between documents (or queries, from the query facet), the set of documents or queries is clustered and the broadness of the topic estimated by the number of disjointed clusters formed. This is of course, only an approximation, since it assumes that every document focuses on one aspect only. However, a document could describe more than one aspect of a topic.

As an example, consider the set of relevant documents of the TREC Terabyte track topic "John Edwards womens' issues", with sixteen relevant documents. After the relevant documents are partitioned into clusters, the document clusters were inspected to figure out their aspects. By reading the documents we deduced that the clustering procedure divides the documents into the following aspects:

1. A sexual offenders bill in North Carolina.
2. Patient Protection Legislation.
3. The Historically Women's Public Colleges or Universities Historic Building Restoration and Preservation Act.
4. Women's health care issues.

We use the number of aspects (the number of clusters) of the topic's relevant documents (or of the queries) to measure the diameters, $d(R, R)$ and $d(Q, Q)$, of the topic difficulty model.

2.2 Document coverage and query coverage

Rarely does the information pertaining to both facets of the model exist. It is much more likely to expect that different users of the model have only one description of a topic, either the queries or the relevant documents (it is assumed that the document collection is always accessible, at least through a search engine), and only an approximation of the missing part of the model. In such cases, the proposed model is still useful for obtaining information as to the topic.

When only Q or R are available (as defined above) we approximate the missing set using the Jensen-Shannon divergence. Thus, given only a query, or a set of queries Q, we define *document coverage* (DC) as the set of documents (chosen from the given collection) which minimizes the Jensen-Shannon divergence from Q:

$$DC(Q) = argmin_{R'} D_{JS}\left(Q||R'\right) \quad (5)$$

The document coverage set is an approximation of the correct set R since it is the set of documents which is most similar to Q.

Similarly, given only a set of relevant documents R we define *query coverage* (QC) as the set of queries that minimizes the Jensen-Shannon divergence from R:

$$QC(R) = argmin_{Q'} D_{JS}\left(Q'||R\right) \quad (6)$$

Section 4 provides examples for applications of document coverage and query coverage.

2.3 Practical considerations for computing document coverage and query coverage

Estimating document coverage and query coverage represent a difficult computational challenge. As noted above, the document coverage is the set of documents for which the Jensen-Shannon divergence reaches its minimum. Since finding the subset of documents that is the document coverage for a given query is NP-hard [5], two approximations were used: First, only the top 100 documents returned by the search engine in response to the query are considered as candidates for inclusion in the set. Second, a greedy algorithm is used.

Thus, the document closest to the query (with the lowest JSD) is found. Documents are then added iteratively such that each added document causes the largest decrease in JSD between the query and the average distribution of the selected documents. A typical resulting curve is shown in Figure 2, where the minimum is reached after three documents, and JSD rises thereafter.

Once a minimum is reached (adding a document only increases JSD), the value of JSD is measured and the set of accumulated documents is used as an approximation to the true DC set.

Similarly, finding the QC set, given a set of relevant documents, is NP-hard. Therefore, we use a similar greedy

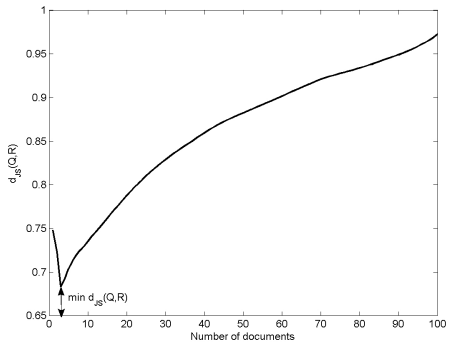

Figure 2: A typical JSD curve obtained by the greedy algorithm for document coverage detection. The minimal point (i.e. the document coverage) is denoted by an arrow.

algorithm. The algorithm builds the QC set incrementally; it only considers the set of terms belong to R, and at each stage it finds a single word that is the closest (in JSD distance) to the document set. This process repeats, and words are added to QC so as to minimize JSD distance from the document set (or increase it by the smallest amount). The iterative process results in a list of ranked words. These are the most representative words for the relevant documents, taking into account the collection distribution. The process stops once a minimum is reached (adding a term only increases JSD) or after all the words in the relevant documents are added.

3. VALIDATING THE MODEL

The objectives of this section are to link the theoretical model of topic difficulty to common IR measurements and to determine the effect of each part of the model on topic difficulty. We validated our model using the Juru search engine [1] on the .gov2 document collection (approximately 25 million documents) from TREC. We experimented with short queries based on the topic titles using the 100 topics of the 2004 and 2005 terabyte tracks.

3.1 Linking model-induced distances to average precision

In this part of the experiment we measured the correlation between the model-induced measurements (JSD distances of the model components) and the average precision (AP) achieved by the search system for the 100 terabyte topics. We also measured the correlation between the model-induced parameters and the median AP of all systems that participated in the TREC Terabyte tracks.

As shown in Figure 2, there are five distances of interest in the model. However, because TREC topics provide a single query for each topic, the inter-query distance could not be used. Thus, four distances and their correlation with AP were evaluated.

Table 2 shows the Spearman correlation coefficient ρ and the Pearson correlation values for each of the distances with the AP. All correlations with an absolute value larger than 0.164 are statistically significant at $p < 0.05$. The distance of the relevant documents from the collection is by far the most important factor influencing topic average precision. The explanation for this phenomena is that a longer distance reflects better separability of the set of relevant documents from the entire collection. The distance of the query to the collection, $d(Q, C)$, and the number of topic aspects, $d(R, R)$, have a lower, yet substantial effect on precision, while the distance of the query to the relevant documents, $d(Q, R)$, at least for the 100 TREC topics, has almost no effect.

	Juru's AP		TREC median AP	
Distance	Pearson	Spearman's ρ	Pearson	Spearman's ρ
$d(Q, C)$	0.167	0.170	0.298	0.292
$d(R, C)$	0.322	0.290	0.331	0.323
$d(Q, R)$	-0.065	-0.134	-0.019	0.004
$d(R, R)$	0.150	0.141	0.119	0.155
Combined	0.447		0.476	

Table 2: Comparison of Pearson and Spearman correlation coefficients between the different distances induced by the topic difficulty model and the AP of the 100 Terabyte topics as achieved by our search system, and the median AP of all TREC participants.

We note that the signs of the regression coefficient show that a longer distance of the queries and the relevant documents from the collection results in a higher AP, while a shorter distance between queries and documents results in increasing AP. Interestingly, a larger number of aspects correlates positively with AP.

The correlation results of Juru and the TREC median are similar, especially for $d(R, C)$ and $d(R, R)$. The values of Pearson's non-parametric correlation and Spearman's parametric correlation are remarkably similar, suggesting that the values are linearly correlated. The Pearson correlation of AP with all four model parameters (the row denoted by "Combined") is relatively high, suggesting that the model captures important aspects of the topic difficulty.

3.2 Linking model-induced distances to topic aspect coverage

In the second experiment, we measured the correlation between the topic difficulty model distances and the aspect coverage of the results retrieved by Juru for the 100 Terabyte topics.

Given the ranked list of results for a given topic retrieved by our system, we measured aspect coverage as follows:

1. Find all aspects of the relevant documents of the topic using the clustering process described in Section 2.1, assuming each cluster relates to a different aspect.

2. For each aspect, mark the top result in the ranking belonging to that aspect as relevant, and mark all other relevant documents in the result set belonging to that aspect as non-relevant. In this way, every aspect covered by the result set has one representative in the ranking.

3. Compute average precision using the new marked documents.

The aspect coverage measure promotes rankings that cover more aspects and also takes into consideration the ranks of the relevant documents. A rank that includes documents

	Juru's Aspect Coverage	
Distance	Pearson	Spearman's ρ
$d(Q,C)$	0.047	0.047
$d(R,C)$	0.143	0.194
$d(Q,R)$	-0.271	-0.285
$d(R,R)$	-0.364	-0.418
Combined	0.482	

Table 3: Comparison of Pearson and Spearman correlation coefficients between the different distances induced by the topic difficulty model and the aspect coverage of the results for the 100 Terabyte topics retrieved by our search system.

from many different aspects on top of the ranking is preferred over a rank containing documents that redundantly cover the same aspect[2].

As Table 3 shows, the distance between the query and the relevant documents, $d(Q,R)$ and the broadness of the topic, $d(R,R)$ have the most significant influence on the ability to retrieve many topic aspects (All correlations with an absolute value larger than 0.164 are statistically significant at $p < 0.05$). As can be expected, the more aspects a topic has, the harder it is to retrieve all of them. The separation of the query and the relevant documents from the collection ($d(Q,C)$ and $d(R,C)$, respectively) have a very minor role in aspect coverage. Interestingly, the combined correlation of all four measurements is extremely similar to that of regular AP, and is a relatively high value.

Theoretically, the topic aspects might have been found by analyzing the query facet, for instance, by using disambiguation methods described in [10]. However, when using the TREC data we assume that the relevant documents for a topic cover all the topic relevant aspects.

4. USES OF THE MODEL

In the following section, three uses of the model are described. The first predicts query difficulty by estimating the expected average precision of the query. The second predicts the number of different relevant aspects expected to be covered by the search results. The final application allows a domain manager to analyze her domain (or her site) to predict the *findability* of the domain, i.e., the likelihood of the documents in that domain returning as answers to queries related to that domain.

4.1 Estimating query average precision

One of the tasks in the Robust tracks of 2004 and 2005 was to estimate query difficulty, that is, to estimate the average precision for each topic. This generated a wealth of research into methods for such estimation, some of which were surveyed in Section 1.

Using the model-induced distances $d(Q,C)$, $d(Q,\hat{R})$, and $d(\hat{R},C)$ (where $\hat{R}$ represents an approximation of the set of relevant documents, computed by the method described in Section 2.3), we attempted to estimate the average precision of the topics.

Given the 100 training topics of the Terabyte tracks, the three distances mentioned above were used for training a predictor. The predictor was trained by a Support-Vector Machine (SVM) using either a radial-basis kernel or a linear kernel. The SVMlight package [8] was used with default parameters. In this experiment, leave-one-out was used for training [3].

We compared the predictor based on the JSD distances to a predictor based on the features described in [19]. Interestingly, the predictor based on the query difficulty model obtained the best results using a radial-basis kernel, while the predictor based on the features described in [19] obtained the best results using a linear kernel. This effect can be explained by the low number of training queries relative to the number of features in the latter case.

The Pearson correlation between the actual average precision to the predicted average precision using JSD distances was 0.362. The same correlation using the features described in [19] was only 0.138. The difference is statistically significant at $p < 0.05$ (one-sided test, [18]). These results demonstrate the ability of the JSD-based features to predict the expected precision. An additional advantage of these features is that they represent a smaller feature space, which is easier for training a predictor.

4.2 Estimating topic aspect coverage

As argued previously, the number of different aspects, and the ability to retrieve a set of results covering all those aspects are two of the main reasons for topic difficulty. In this subsection we describe how to train a predictor for the expected aspect coverage based on the topic difficulty model.

We used the aspect coverage measure for each of the training topics, as defined in Subsection 3.2, as the input to the estimator. The JSD distances are then used as features for training a predictor whose task it is to detect queries with low aspect coverage. Again, the predictor was trained based on a Support-Vector Machine (SVM) using a radial-basis kernel. The SVMlight package [8] was used with default parameters, except for the cost which was set to 10. In this experiment, leave-one-out was used for training [3]. The Pearson correlation between the actual aspect coverage and the predicted aspect coverage using JSD distances was 0.397.

In a second experiment, our goal was to estimate which of the topics has 10% or less of their aspects covered by the document collection. We used the topic difficulty model to train an estimator for detecting low coverage queries. The results of this experiment are shown in Figure 3. This figure shows the ability to detect queries with low aspect coverage, as demonstrated by an area of 0.88 under the ROC curve.

4.3 Estimating topic findability

In this subsection, we use the topic difficulty model to estimate the *findability* of a given collection of documents. We assume that a set of documents of a domain are given, and it is necessary to estimate how easy it is for a user to find these documents. As noted above, this is akin to the notion of *findability* in the field of search engine optimization, where the goal is to optimize document representation so that they will be found easily by the search engines for related queries.

There are two aspects to this problem. The first is how well the domain's documents are separated from the entire collection. The second is the degree to which typical user queries of the relevant topics covered by that domain correlate with the most informative words of the domain, as reflected by its documents.

[2]The Pearson correlation between the aspect coverage and the average precision is 0.338 (Spearman: 0.413).

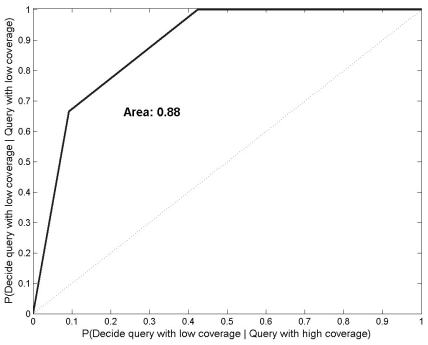

Figure 3: Receiver operating characteristic (ROC) curve for distinguishing topics with low aspect coverage from other queries.

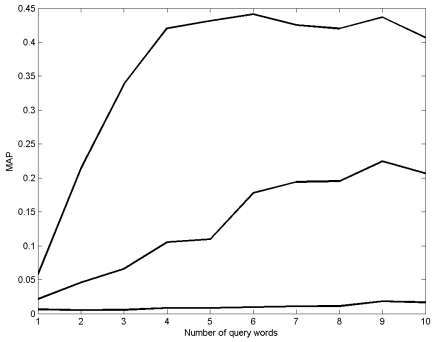

Figure 4: Cluster centers of the AP curves versus the number of best words of the relevant documents. The curves represent three typical findability behaviors.

Our first experiment attempts to quantify the effect of the separability of the relevant documents from the entire collection on the ability to find those documents. We computed the Query Coverage (QC) set for each of the Qrels of the 100 training topics. The greedy algorithm used for finding QC results in a ranked list of words, from best to worst. We used the 10 best words for each topic.

For each topic, a sequence of queries was created, with the first best word, the first two best words and so on, up to the first ten best words. The sequence of queries were executed and the AP for each query computed. The resulting curves of AP against the number of terms were then clustered using the k-means clustering algorithm [3]. We used the AP values at each number of terms as features to the clustering algorithm. The results of this clustering (using three clusters) are shown in Figure 4. These curves show typical findability behaviors of a topic, ranging from topics which are extremely difficult to find, no matter how many search terms are used, to topics for which 3-4 query terms are sufficient for achieving high AP. The average AP curve for one of the clusters shows a low AP for the first best word while additional words do not greatly improve it. The second curve, on the other hand, shows a dramatic increase with the addition of words. The third curve shows an optimal findability behavior – the first best terms are sufficient to achieve high AP. For most topics, no improvement in AP was achieved by using more than nine best words.

We used the Kruskal-Wallis non-parametric one-way ANOVA to test whether the partition of topics according to the clusters shown in Figure 4 was also reflected in the JSD distance between the relevant documents and the collection. The result was extremely significant at $p = 0.007$. The average AP for topics in a cluster based on the best topic terms is higher when the average $d(R, C)$ is higher. This is in agreement with the correlation values shown in Section 3.1. A useful application of this finding is that it is possible to estimate the findability of a domain (i.e. to which of the three curves in Figure 4 the behavior of the domain will be) based solely on its separation from the collection, regardless of the specific search engine.

Our second experiment relates to the effect of the typical users' query on the ability to retrieve the relevant information. By observing the list of best terms (as described above), a domain manager might be able to deduce if a "reasonable" user would use such terms when searching for information captured in her domain.

For example, consider the query "Massachusetts textile mills". This query is one of the most difficult ones for our system. The first ten most important terms found by the greedy algorithm, in decreasing order of importance are: "Lowell", "mill", "nation", "history", "industry", "area", "park", "NHP" (Neighborhood Health Plan), "work", "heritage". The first two original query terms only appear in locations 12 and 29 respectively (the third one, "mills", appears second). Our results indicate that the query "Lowell mill nation", based on the first 3 words in the best-word list, generates an average precision of approximately 3.5 better than that of a query based on the original query terms.

We tested how the rank (location) of the users' query terms in the list of best words correlate with the AP and the Aspect Coverage. We assume that TREC queries based on the topic titles represent typical user queries. The median of the average rank of query terms was approximately 4400, which is a surprisingly high. This shows that the user-selected terms are very far from the best terms of the relevant documents. The average rank of the query terms has a Spearman correlation of -0.165 with the Aspect Coverage (but only negligible correlation of 0.048 with the AP). Thus, the farther down the query terms are in the list of best terms, the fewer the aspects that would be covered by the user's query. The correlation value is not very high, but it does show that the selection of query terms has a non-negligible effect on findability.

One of the uses of such a best-word list is to identify problematic domains that can hardly be found. If the best words are atypical to the information exposed by that domain (e.g. "Lowell" in the example above), or even worst, if the typical terms for that information are ranked low in that list, the domain is expected to suffer from bad findability. Thus, simply looking at the list of best words and trying to comprehend if such words would be used by the typical user can greatly improve the findability of the domain, by document expansion, for instance.

5. SUMMARY

This work tries to answer the question of what makes a topic difficult. We addressed a novel model that captures the

main components of a topic and the relations between those components to topic difficulty. The three components of a topic are the textual expression describing the information need (the query or queries), the set of relevant documents of the topic (the Qrels), and the entire collection of documents. We showed that topic difficulty strongly depends on the distances between those components. The larger the distance of the queries and the Qrels from the entire collection, the better the topic can be answered (with better precision and with better aspect coverage). In the absence of knowledge about one of the model components, the model can still be useful by approximating the missing component based on the other components. We demonstrated the applicability of the difficulty model for several uses such as predicting query difficulty and analyzing the findability of a specific domain.

The difficulty model described in this work is based on the relationship between the main topic components. However, there are many more important features affecting topic difficulty that the current model ignores. For example, ambiguity of the query terms, or topics with missing content, i.e., absence of relevant data in the given collection for the information need. Extending the model to encapsulate other aspects of topic difficulty is left for further research.

6. ACKNOWLEDGMENTS

The authors thank Shai Fine for his invaluable suggestions regarding the model and the JSD distance.

7. REFERENCES

[1] D. Carmel, E. Amitay, M. Herscovici, Y. S. Maarek, Y. Petruschka, and A. Soffer. Juru at TREC 10 - Experiments with Index Pruning. In *Proceedings of the Tenth Text Retrieval Conference (TREC-10)*. National Institute of Standards and Technology. NIST, 2001.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 299–306. ACM Press, 2002.

[3] R. Duda, P. Hart, and D. Stork. *Pattern classification*. John Wiley and Sons, Inc, New-York, USA, 2001.

[4] D. A. Evans, J. G. Shanahan, and V. Sheftel. Topic structure modeling. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 417–418. ACM Press, 2002.

[5] M. R. Garey and D. S. Johnson. *Computers and intractability*. W. H. Freeman and Company, New-York, USA, 1979.

[6] D. Harman and C. Buckley. The NRRC reliable information access (RIA) workshop. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pages 528–529. ACM Press, 2004.

[7] B. He and I. Ounis. Inferring query performance using pre-retrieval predictors. In A. Apostolico and M. Melucci, editors, *String Processing and Information Retrieval, 11th International Conference, SPIRE 2004*, volume 3246 of *Lecture Notes in Computer Science*, 2004.

[8] T. Joachims. Making large-scale support vector machine learning practical. In A. S. B. Schölkopf, C. Burges, editor, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, 1998.

[9] K. Kwok, L. Grunfeld, H. Sun, P. Deng, and N. Dinstl. TREC 2004 Robust Track Experiments using PIRCS. In *Proceedings of the 13th Text Retrieval Conference (TREC-13)*. National Institute of Standards and Technology. NIST, 2004.

[10] S. Liu, C. Yu, and W. Meng. Word sense disambiguation in queries. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 525–532, New York, NY, USA, 2005. ACM Press.

[11] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *ACM SIGIR 2005 Workshop on Predicting Query Difficulty - Methods and Applications*, 2005.

[12] P. Over. TREC-7 interactive track report. In *Proceedings of the 7th Text Retrieval Conference (TREC-7)*, pages 33–39. National Institute of Standards and Technology. NIST, 1998.

[13] J. M. Prager. A curriculum-based approach to a QA roadmap. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2004)*, Las Palmas, Spain, 2004.

[14] S. Tomlinson. Robust, Web and Terabyte Retrieval with Hummingbird Search Server at TREC 2004. In *Proceedings of the 13th Text Retrieval Conference (TREC-13)*. National Institute of Standards and Technology. NIST, 2004.

[15] E. M. Voorhees. Overview of the TREC 2003 robust retrieval track. In *Proceedings of the Twelfth Text Retrieval Conference (TREC-12)*. National Institute of Standards and Technology (NIST), 2003.

[16] E. M. Voorhees. Overview of the TREC 2004 robust retrieval track. In *Proceedings of the 13th Text Retrieval Conference (TREC-13)*. National Institute of Standards and Technology (NIST), 2004.

[17] E. M. Voorhees. Overview of the TREC 2005 robust retrieval track. In *Proceedings of the 14th Text Retrieval Conference (TREC-14)*. National Institute of Standards and Technology (NIST), 2005.

[18] L. Wasserman. *All of statistics*. Springer, 2003.

[19] E. Yom-Tov, S. Fine, D. Carmel, and A. Darlow. Learning to estimate query difficulty: including applications to missing content detection and distributed information retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 512–519. ACM Press, 2005.

[20] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 10–17. ACM Press, 2003.