Query Side Evaluation

An Empirical Analysis of Effectiveness and Effort

Leif Azzopardi
Dept. of Comp. Sci., University of Glasgow
Glasgow, United Kingdom
leif@dcs.gla.ac.uk

ABSTRACT

Typically, Information Retrieval evaluation focuses on measuring the performance of the system's ability at retrieving relevant information, and not the query's ability. However, the effectiveness of a retrieval system is strongly influenced by the quality of the query submitted. In this paper, the effectiveness and effort of querying is empirically examined in the context of the Principle of Least Effort, Zipf's Law and the Law of Diminishing Returns. This query focused investigation leads to a number of novel findings which should prove useful in the development of future retrieval methods and evaluation techniques. While, also motivating further research into query side evaluation.

Categories and Subject Descriptors

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

General Terms

Theory, Experimentation

Keywords

Information Retrieval, Evaluation, Simulation

1. INTRODUCTION

A standard evaluation within Information Retrieval (IR) typically involves measuring the ability of a retrieval system to retrieve relevant documents in response to a representative set of information needs (denoted by topics). For each topic a query is formulated as the user's input to the retrieval process [6]. The response from the system is a ranked list of documents in decreasing order of estimated relevance. Usually, the subject of enquiry is the influence the system has upon the retrieval effectiveness; where the typical IR experiment is to determine whether model/method/system A performs better than model/method/system B. However, a source of major variation in effectiveness is the query. For any given topic, numerous queries could be posed which will result in vastly different levels of retrieval effectiveness. By focusing evaluation on the query, as opposed to the system, a number of interesting research questions emerge, such as:

- how do user's generate queries and how can this be modeled,
- how difficult is it to pose an effective query,
- how much effort should be spent querying, and,
- what is the relationship between query effort and retrieval effectiveness.

Despite the vast amount of research performed on analyzing and comparing systems, relatively little research has been performed investigating query side evaluation issues. In order to develop better systems it is imperative that a detailed understanding of the entire retrieval process is acquired. Currently, little is known about the influence a query has on effectiveness despite the amount of research conducted on estimating query performance (e.g. [10, 17, 18, 20]). This is because for any given topic, there exists only a few queries. So it is not possible to deeply analyze variations in performance given the array of possible queries for a given topic. This presents a major obstacle in performing such research.

In this paper, an investigation into the influence of the query on effectiveness is undertaken. In order to do so, we propose a novel method for generating queries for ad-hoc topics to provide the necessary data for this comprehensive analysis of query performance. To provide the theoretical underpinning of this study, the *Principle of Least Effort* is considered within the context of the retrieval process; where the interaction between the user and system is seen as a form of communication. This interpretation leads to the hypothesis that retrieval effectiveness follows *Zipf's Law* for a given topic, such that there will be many queries that will perform poorly, while a few perform well (i.e. follow a power law). If the performance of queries can be characterized by a power law it will be possible to succinctly describe the distribution of retrieval effectiveness. This would be particularly useful in a number of areas, such as query performance prediction and the comparison/evaluation of IR systems. To this aim, we conducted an empirical study examining the effectiveness of queries on a number of ad hoc TREC Test collections. The observations and findings from this empirical study provide a detailed understanding of the interactions at play between the query and the system. For instance, as

SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.

the length of the query increases the retrieval effectiveness follows the *Law of Diminishing Returns*.

The remainder of this paper is as follows: in Section 2 we provide an overview of the Principle of Least Effort and Zipf's law in Search. Then, in Section 3, we shall describe the experimental setup used, which includes our proposed method for generating ad-hoc queries. Section 4 reports the main results and findings from our analysis, before we conclude with a summary and outline directions for future work.

2. PRINCIPLE OF LEAST EFFORT

The *Principle of Least Effort* states that a person attempting to apply a tool to a job does so in order to minimize the expected effort in using that tool for the given job [1]. Applied to the context of communication between an author and a reader, Zipf argues the following case: (i) an author will use as few terms as possible to accomplish the task of communication (i.e. minimize the effort of expression), whereas (ii) the reader prefers different terms to represent different situations in order to minimize ambiguity (i.e. minimize the effort of interpretation). An intuitive example from [6], is that an author wants to refer to something in the world (i.e. pointing to different objects). The least amount of effort that they could expend is by using the same word to refer to each object (i.e. "that, that, that, that"). On the other hand, the reader wants distinct references to objects, where every possible interpretation is characterized by a unique term so that all ambiguity is removed (i.e. "book, cup, apple, glass"). This would minimize the reader's effort in interpretation. The writer creates pressure towards the *unification* of the vocabulary, while the reader creates pressure towards the *diversification* of the vocabulary. According to Zipf, balancing between these two opposing objectives is believed to result in a power law distribution which characterizes the usage of terms in text. This so called, *vocabulary balance* between general and specific terms is when communication is most efficient. Empirical observations led to Zipf's law, where the fraction of terms at rank k is given by:

$$P(X = k) = Ck^{-s} \quad (1)$$

for $k >= k_0$ with C a normalizing constant, s the scaling parameter, and k_0, a lower bound from which onwards the power-law holds [3]. Figure 1 shows the power-law of term usage in the Aquaint Corpus, where $s = 1.52$ and $k_0 = 1$.

2.1 Applied to Search

Zipf's law has been applied to many search related phenomena (e.g. Heaps' law[2], Lotka's law[15], Bradford's law[8]). Here, we focus on the context of the retrieval of documents through an IR system. Previously, Blair [7] considered Zipf's Law in this context, where he argues that the terms used to index documents need to maintain this vocabulary balance. If general terms are used to describe documents, then when the user poses a query with these terms the recall will be high, but the precision low. While, if specific terms are used to describe documents, then precision will be high, but recall will be low. This leads to a trade off between precision and recall. Blair argues that one of the main reasons that IR systems fail is due to the imbalance in the vocabulary used to describe documents. While these arguments were put forward they were never tested or shown

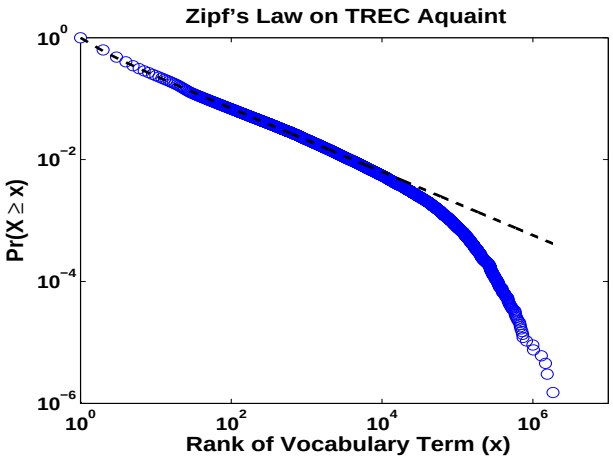

Figure 1: Zipf's Law: the log-log plot of term usage versus rank on the Aquaint Corpus.

– the empirical study performed here lends some weight to support Blair's arguments. However, here, we take a query centric view of the problem.

In this paper, we consider the application of the Principle of Least Effort in the context of communication between a user and an IR system, and aim to determine whether efficient communication between the two parties is achieved i.e. communication is efficient when the retrieval effectiveness is maximized and effort is minimized[1]. Under this interpretation, in communicating with the IR system, the user wants to expend minimum effort in explaining their information need to the system; whereas the system wants to expend minimum effort in interpreting the query in order to return relevant documents. Consequently, the system would like longer and more precise queries, which uniquely identify relevant documents, whereas the user would like to submit short and vague queries. We posit that the effectiveness of queries given a particular topic follows Zipf's Law: such that there will be many queries that could be submitted which would perform poorly, while there will be few queries that perform well (i.e. a power law of retrieval effectiveness). This results from the trade off between precision and recall, stemming from the use of general and specific terms in the queries expressed. The extent to which this communication is successful can be measured through the retrieval effectiveness.

Despite the intuitiveness of the hypothesis, no study has been performed in order to confirm or deny this hypothesis. Does retrieval effectiveness follow a power law distribution? If the performance of queries can be characterized by a power law it will be possible to succinctly describe the probability of the retrieval effectiveness for a topic. This could be used in a variety of applications from describing the retrieval potential of a system, to improving retrieval algorithms and methods.

3. EMPIRICAL STUDY

The following study was performed to examine the hypothesis that retrieval effectiveness follows a power law, along with the following research questions:

[1]Note that in this paper, the simplifying assumption that effort is proportional to query length is made.

- What is the distribution of the retrieval effectiveness for ad hoc topics?
- How does the distribution change with different querying strategies?
- How does the distribution change according to length?
- And, what is the most efficient query length (i.e. when is performance maximized and length minimized)?

As previously mentioned, in order to conduct this study a sufficiently large number of queries for a given topic is required. In this section, we first describe the data and retrieval models used as part of this study (see §3.1) , before detailing a novel technique for generating ad-hoc queries from pre-existing topics/test collections (see §3.2). This technique is used as part of the methodology in our controlled study (see §3.3).

3.1 Experimental Setup

For the purposes of this study, three TREC Test Collections were used: AP and WSJ using the TREC 1 and 2 Topics, respectively, and the Aquaint Collection using the TREC 2005 Robust Topics. Each test collection was indexed using the Lemur toolkit[2], where the documents were preprocessed using Porter's Stemming and a standard stop list. Three retrieval models were used to explore the relationship between effectiveness and length, two probabilistic models, BM25 and a Language Model with Dirichlet Prior Smoothing, and a vector space model using TF.IDF. The scope of this study is therefore limited to Best Match models, and restricted to the effectiveness of ad-hoc topics in this domain; where the following measures of retrieval effectiveness were considered, Average Precision (ap), Precision at 10% recall (p@10%) and the Precision at 20 documents (p@20).

Collection	Docs	Terms	Topics	s
AP	164,597	196,875	51-100	1.5*
WSJ	173,252	174,670	101-150	1.6*
Aqauint	1,033,461.	663,158	Robust 05	1.52*

Table 1: Collection Statistics along with the Scaling Parameter s for Zipf's Law. * indicates the distribution of the vocabulary fits a power law.

3.2 Generating Simulated Ad Hoc Queries

Simulating user behavior is a recognized approach to evaluate different possible user search strategies and interactions. The main advantage is that simulation provides a cost-effective alternative to manually building test collections [5, 11, 12]. With recent developments in formal models for simulating user querying behavior it is now possible to generate queries in a variety of different styles and lengths in order to perform controlled experimentation inexpensively [5, 12]. While arguably simulation may be somewhat artificial and not truly representative of actual user behavior, progress has been made towards replicative valid models of query generation. In [5], the authors propose a generative probabilistic model that produces known-item topics (query and document pairs) which obtain performance similar to the performance of actual user topics/queries. Consequently, we propose an extension of this model and adapt it for the generation of queries for ad-hoc topics.

The query generation process can be described as follows: a user imagines an ideal relevant document, from this document they select query terms, sometimes the query terms will be on topic, while other times it will be off topic. This is modeled formally by a Hidden Markov Model [5, 16]:

$$p(t|\theta_{query}) = (1-\lambda)p(t|\theta_{topic}) + \lambda p(t) \quad (2)$$

where the probability of selecting a term t given the query language model θ_{query} is a mixture between selecting from the topic language model θ_{topic} and the background collection model $p(t)$, which is controlled by λ. The probability of term appearing in the collection, $p(t)$, represents terms which are off topic. Whereas, the topic model represents the distribution of terms in the ideal document, and represents the user's background knowledge about the topic from which they can draw query terms. By repeatedly sampling m times from θ_{query}, a query of length m can be generated for the given topic. And by repeating the entire process, numerous queries can be produced.

Topic Models The estimation of the topic model $p(t|\theta_{topic})$ encapsulates a possible strategy a user may employ when formulating a query. Two previously proposed user querying strategies are: Frequent and Discriminative. In the first, a user will select terms that are likely to appear in the relevant documents, and in the second, a user will select terms that are likely to be more informative in the relevant documents. In [5], it was shown the Frequent (called Popular in [5]) and Discriminative strategies delivered performance that was most like that obtained from real queries/topics. Queries of the first style were similar to real users, while the second style tended to perform slightly better than real user queries.

Here we propose another user querying strategy, called Conditional. Instead of making the naive assumption that a user will randomly sample terms in such a way, we condition the selection on some *a priori* knowledge. This simulates the case when a user may be given a brief about a topic and this seeds their querying. For instance, given a brief about "tropical storms", or "airbus subsidies", this would condition the query being generated. We believe this strategy is more realistic of actual querying behavior, but we leave exploring this direction for future work. For this approach, we estimate the probability of a term given a topic using a relevance model $p(t|\theta_{rel})$ as it provides an estimate of the conditional probability of a term given the seed query [14]. Formally, the three different strategies can be specified as follows:

- **Frequent**: $p(t|\theta_{topic}) = \frac{\sum_{d \in R} n(t,d)}{\sum_{d' \in R} n(d')}$
- **Discriminative**: $p(t|\theta_{topic}) = \frac{\sum_{d \in R} tf.idf(t,d)}{\sum_{d' \in R} \sum_{t' \in V} tf.idf(t',d')}$
- **Conditional**: $p(t|\theta_{topic}) = p(t|\theta_{rel}) = p(t|q_0)$

where $n(t,d)$ is the number of times a term appears in a document d, $tf.idf(t,d)$ is the term frequency inverse document frequency of t in d, $p(t|q_0)$ is the conditional probability of a term given the seed query q_0, and R denotes the set of relevant documents for the topic. By using relevant documents to describe the topic, we are assuming that the user

[2]http://www.lemurproject.org

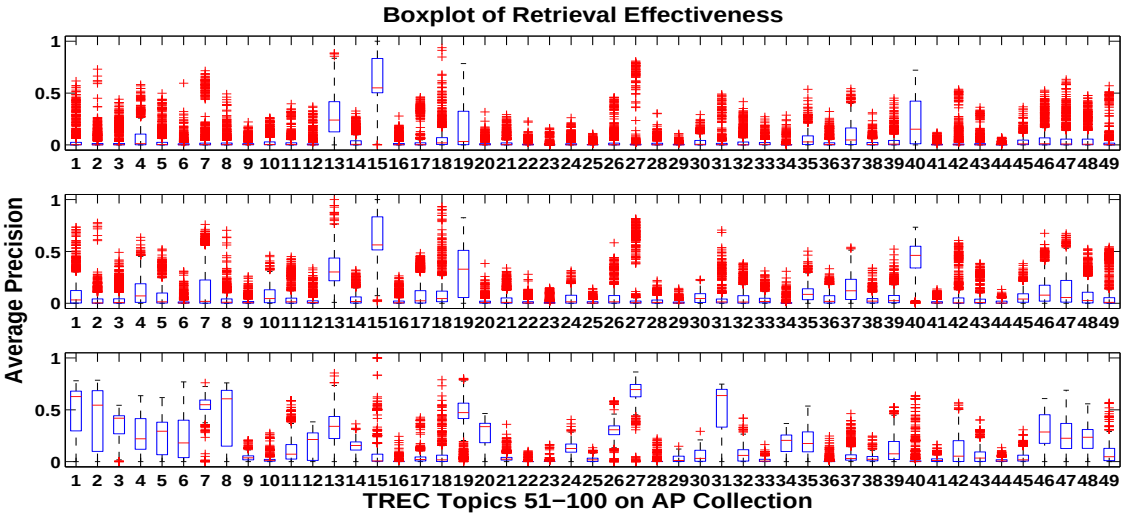

Figure 2: Box Plot of retrieval effectiveness across topics on AP. Top m=2, middle m=5 and bottom m=30.

has some *a priori* background knowledge of the topic. However, this is offset by the amount of noise added to the query model.

3.3 Experimental Method

For the purposes of the analysis we used all three different query strategies to represent different possible ways a user may formulate queries in order to satisfy their information need. For each TREC test collection (documents and topics), 1000 queries of a given query length m were generated per topic, using the corresponding relevance judgements (and for conditional queries the title of the TREC Topic was used as the seed query.). A small amount of noise was added where λ was set to 0.2 which reflects the amount of noise seen in real queries [5]. This process was repeated for different lengths where $m = \{1, 2, 3, 4, 5, 10, 15, 20, 30\}$. Length was controlled for two reasons: (i) there are large variations in performance due to length, and (ii) length provides an indication of the amount of effort involved in posing a query. The total number of queries generated per topic was 27,000, and per test collection was approximately 1.35 million. Each query set was then executed against each different retrieval model and the retrieval effectiveness of each query was evaluated against the corresponding relevance judgements for that topic. The performance measures: ap, p@10% and p@20 were recorded.

Given the retrieval measurements taken for a particular query set and length, we determined whether the retrieval effectiveness followed a power law distribution by applying the statistical methods by Clauset et al [3]. Their methods automatically estimate the scaling parameter s, by selecting the fit that minimizes the Kolmogorov-Smirnov (KS) $D - statistic$. The KS test compares the empirical data against a power law distribution. The null hypothesis is that the data fits a power law with scaling parameter s. This holds if the $D - statistic$ is less than the critical value 0.043 denoting 5% significance level. Since ap, p@10% and p@20, measures are bound between 0 and 1, we performed a transformation to discretize the values into 50 buckets, representing precision values of 0-0.02, 0.02-0.04 and so on until 0.98-1.00. The hypothesis is that the probability of a query having retrieval effectiveness that lies in bucket k, is then given by Equation 1.

4. ANALYSIS

To facilitate the reporting of the general findings, we shall concentrate on discussing the results obtained on the experiments performed on AP and on BM25 using Average Precision only (though statistics from the other collections are reported). Please note that: the trends and the findings presented are indicative of those found on the other collections, retrieval models and measures. The remainder of this section will present the main findings on the distribution of retrieval effectiveness.

4.1 Overview of Performance

Figure 2 provides an overview of the performance obtained from the 1000 Discriminative queries on each of the 49 topics in 51-100 for AP. As can be seen from the box plots, when the query length is short the effectiveness of most queries is close to zero. As query length grows the effectiveness also increases: for $m = 30$ there are even a number of topics where the mean effectiveness is greater than 0.3 ap. From these subplots, it is also possible to obtain an idea of the variation in performance between topics and across different query lengths.

In Table 2, we have divided the queries into two groups according to their effectiveness, the top 10% and bottom 90% per topic and took the mean across topics of the median performance in that group. This is reported for each length. The top 10% of queries perform substantially better than the remaining 90% for the shorter queries, but as length increases the difference in effectiveness decreases. This suggests that a power law may hold for shorter queries but is unlikely for longer queries.

When the queries were conditioned on the title topics, the performance for short queries dramatically improved, compared to the other query sets. This is probably due in part to

the title bias in TREC Topics [9], but also because the other query strategies assume a less informed user scrambling to pose a query out of vague background knowledge. In previous work, studies on the influence of length has typically been constrained to a small subset of queries (for instance, title and title and description queries [19], or a couple of hundred user queries [13], which is insufficient to estimate retrieval performance for particular query lengths). Table 2 provides a precise breakdown over length. This will lead to an interesting observation in Section 4.3.

4.2 Distribution of Retrieval Effectiveness

Figure 3 presents the log-log plot of the distribution of retrieval effectiveness for one topic in AP for $m = 2, 5$ and 30 for each of the querying strategies. It is of note that the distributions are quite similar between the different querying strategies. From the first two plots we can see that a linear relationship between the proportion of queries and the performance holds until 0.8 ap, which is indicative of a power law. However, for the long queries in the third plot quite a different distribution is witnessed. The proportion of queries is roughly uniform until about 0.4 ap. While there was variance between topics, the same trends were present across the different topics, collections and models.

Fitting Power Laws to Retrieval Effectiveness: For each topic and query set, we attempted to fit a power law to the distribution of retrieval effectiveness using the methods provided by Clauset et al [3]. For each test, we obtained a scaling parameter s, the k_0 minimum and the goodness of fit statistic D obtained from the Kolmogorov-Smirnov test. Figure 4 shows the mean scaling parameter and the mean $D-statistic$ along with the standard error bars across query lengths for AP for each of the different query strategies. From the goodness of fit plot we can clearly see that the best fits occurred when the query length was between two to five; where the scaling parameter was between 2.6-2.7. We can see in Figure 3 that as the query length increases, the linear relationship between the variables becomes progressively worse. Note that on average the $D - statistic$ was seldom less than the critical value (and thus the empirical data was usually significantly different from a power law distribution). In fact, out of all the tests performed only a handful of power laws were witnessed. These were mostly found in the range of 2-5 query terms (as suggested by the bottom plot in Figure 4). This finding suggests that the distribution of retrieval effectiveness does not follow a power law distribution. However, for short queries between 2-5 terms, the distribution closely resembles a power law distribution between 0 and 0.8 ap[3]. Stated in another way, the communication with system appears to be the most efficient, but not optimal, when two to five query terms are used.

4.3 Law of Diminishing Returns

While shorter queries appear to be most efficient in terms of communicating with the system, they do not, on average, provide the best total retrieval performance (which is obtained when the query length is 30.). Here, we perform a follow up analysis to study the relationship between retrieval effectiveness and effort (in terms of length) to determine when effectiveness is maximized given effort. This assumes a user wants to maximize the effectiveness of the communication, while expending the minimum effort (i.e. get the job done with least effort [1]). Given the empirical distributions of retrieval effectiveness for each query length for a given topic it is possible to perform an economic analysis of the productivity of querying[4].

In a production system with variable inputs (such as unit of labor), an analysis of the output of the system can be conducted to determine when certain criteria are optimized.

[3]This was also similar for the other performance measures evaluated.

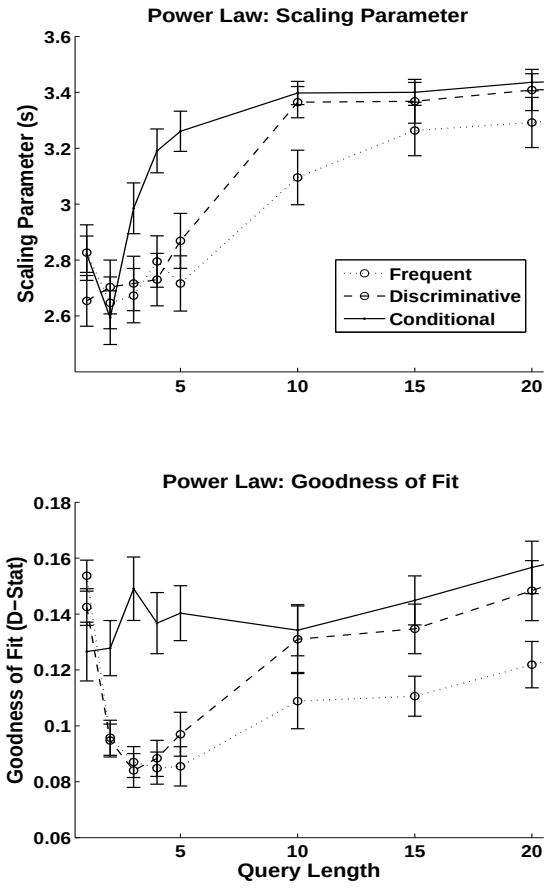

Figure 4: Power Law Fits: the top plot shows the estimated Scaling Parameter s, while the bottom plot shows the Goodness of Fit Statistic (both plots include standard error bars). The bottom plot indicates that most of the KS-Tests resulted in a $D - statistic$ greater than the critical value 0.043 (i.e. the data does not follow a power law). The range which provided the closest fits was when query length was between 2 and 5.

[4]In Varian's SIGIR 1999 keynote address "Economics and Search" three suggestions are presented on how economics could be useful in a number of different ways within IR. One suggestion was to consider the Economic value of Information, "where a consumer is making a choice to maximize expected utility or minimize expected cost". Under the *Principle of Least Effort* the consumer/searcher aims to minimize their expected cost/effort and maximize their expected utility.

Collection	%	Queries	1	2	3	4	5	10	15	20	30
AP	Top 10%	Freq.	0.083	0.18	0.21	0.25	0.27	0.35	0.39	0.42	0.46
	Bot. 90%	Freq.	<0.01	0.01	0.025	0.043	0.062	0.13	0.18	0.23	0.29
	Top 10%	Discrim.	0.14	0.24	0.28	0.31	0.33	0.4	0.44	0.47	0.5
	Bot. 90%	Discrim.	0.002	0.03	0.054	0.081	0.1	0.2	0.26	0.31	0.36
	Top 10%	Cond.	0.083	0.29	0.33	0.35	0.37	0.43	0.46	0.48	0.5
	Bot. 90%	Cond.	<0.01	0.079	0.17	0.19	0.21	0.27	0.31	0.33	0.37
WSJ	Top 10%	Freq.	0.074	0.14	0.17	0.19	0.21	0.28	0.32	0.34	0.38
	Bot. 90%	Freq.	<0.01	0.011	0.014	0.019	0.026	0.081	0.13	0.16	0.21
	Top 10%	Discrim.	0.13	0.2	0.22	0.24	0.27	0.33	0.37	0.4	0.43
	Bot. 90%	Discrim.	0.011	0.017	0.03	0.058	0.074	0.15	0.2	0.23	0.28
	Top 10%	Cond.	0.14	0.21	0.23	0.25	0.27	0.33	0.38	0.44	0.47
	Bot. 90%	Cond.	0.01	0.02	0.05	0.06	0.08	0.16	0.24	0.27	0.33
AQ	Top 10%	Freq.	0.033	0.072	0.1	0.13	0.14	0.22	0.26	0.3	0.34
	Bot. 90%	Freq.	<0.001	<0.001	<0.01	<0.01	<0.01	0.044	0.085	0.12	0.18
	Top 10%	Discrim.	0.08	0.12	0.15	0.17	0.19	0.27	0.32	0.35	0.34
	Bot. 90%	Discrim.	<0.01	<0.01	<0.01	0.019	0.034	0.093	0.14	0.19	0.14
	Top 10%	Cond.	0.07	0.13	0.17	0.18	0.22	0.29	0.34	0.36	0.37
	Bot. 90%	Cond.	<0.01	<0.01	0.03	0.05	0.06	0.08	0.15	0.19	0.21

Table 2: Average median ap of the top 10%/bottom 90% of queries across topics for each query length.

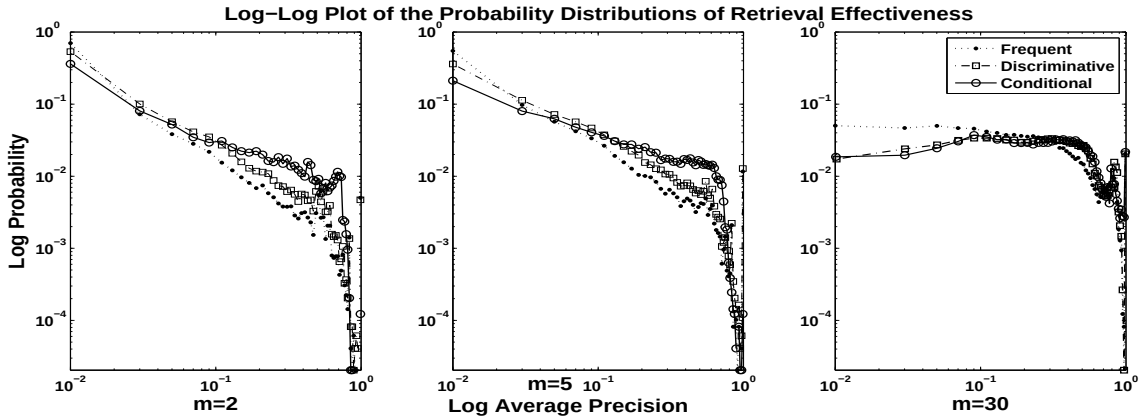

Figure 3: Power Law Estimates for different query lengths on one topic: left m=2, middle m=5, right m=30.

This is performed by measuring the output as more input units are added. For each additional unit of labour, the total output, marginal output and average output are plotted on a graph, where the marginal output is the change in output divided by the change in input units, and average output is the total output divided by the number of input units. According to the *Law of Diminishing Returns*, there will be a point beyond which each additional unit of input will produce less and less output. At this point the production system's output per unit is maximized. Applying this analysis in the context of querying, each query term is treated as the unit of input, reflecting effort, and the output is the retrieval effectiveness. Then, maximizing the Marginal Performance corresponds to maximizing the retrieval effectiveness using the least effort.

In Figures 5 and 6, we show the analysis performed on the top 10% best performing queries from each topic, and on all the queries, respectively. Each plot contains the Total Performance, the Average Performance, and the Marginal Performance. For each curve, we computed the Total Performance by taking the mean of the effectiveness of queries for that length. The Average Performance is the Total Performance divided by the length of the query, and the Marginal Performance is the change in the Total Performance divided by the change in length. In each figure, the left plot corresponds to Frequent queries, the middle plot for Discriminative queries and the right plot for Conditional queries.

For the best case scenario shown in Figure 5, we can see that the Marginal Performance is maximized when the length of the queries is two. After this point the law of diminishing returns applies. That is, each subsequent query term added is likely to result in an increase in performance but at a decreasing rate of return. From these plots, it appears that for the Conditional queries (and to a lesser extent the other queries) the returns are exponentially diminishing.

Figure 6 shows the overall picture across all queries. Here the situation is slightly different, for Frequent queries diminishing returns are not witnessed until query length is five, for Discriminative queries, it is between 2-4 query terms, while for Conditional queries it is not until a query length of 3. Nonetheless, we still witness the law of diminishing returns, and that for Conditional queries the returns diminish

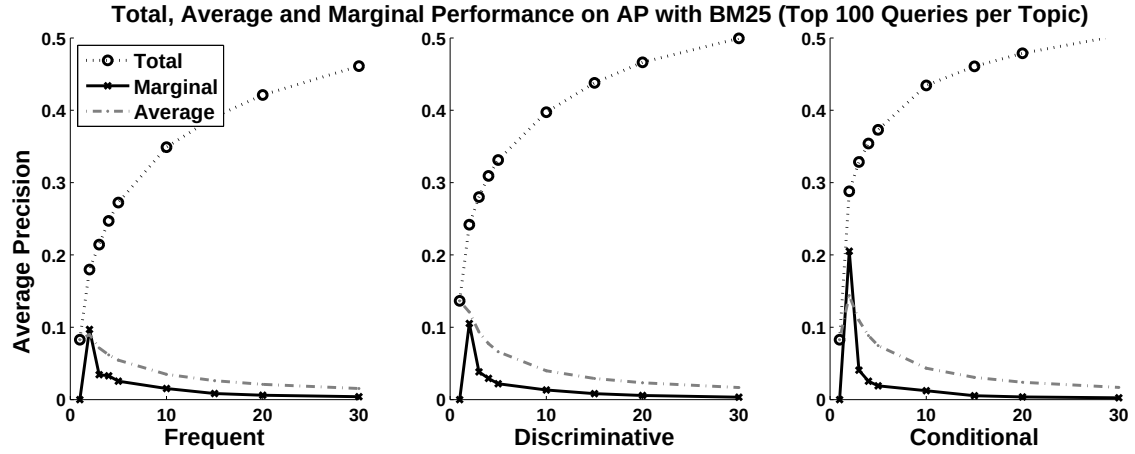

Figure 5: The Total, Average, and Marginal Performance of the top 10% of queries from each topic. The maximum Marginal Performance is obtained when query length is two. After this point the law of diminishing return applies and the Marginal Performance appears to decrease at an exponential rate. The X-axis denotes query length.

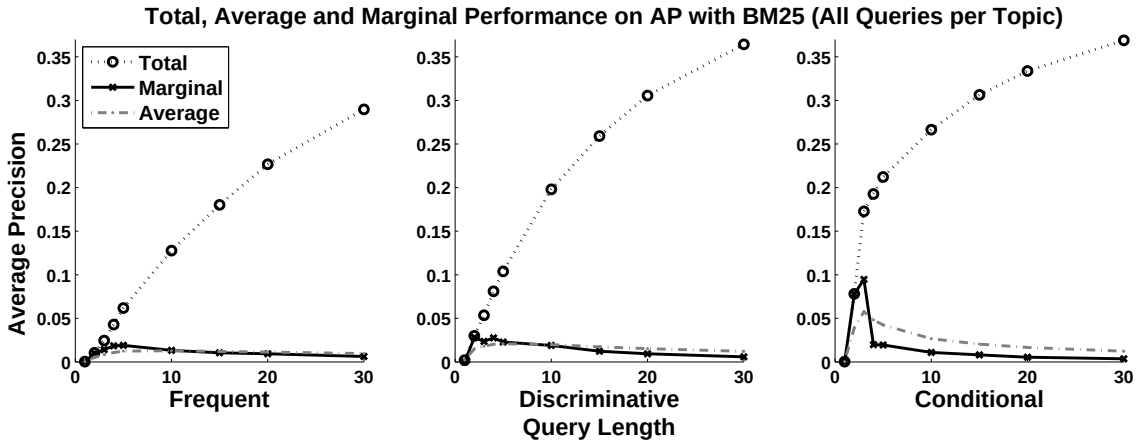

Figure 6: The Total, Average, and Marginal Performance for all the queries from each topic.

far more rapidly than for the other query strategies. In summary, the main finding is apparent; queries of length two, until to five, tend to maximize the Marginal Performance depending on the querying strategy. This region is where the user gets "the most bang for their buck" and corresponds to the region where communication is most efficient given the results from the previous section. This empirical analysis provides an economic justification for posing short queries.

5. DISCUSSION AND CONCLUSION

In this paper, we have performed an empirical study examining the effectiveness of queries, in order to characterize the distribution of retrieval effectiveness. Motivated by Zipf's Law, we hypothesized that retrieval effectiveness would follow a power law. When we tested this hypothesis for each topic, given the different collections, retrieval models and querying strategies, the empirical data did not fit a power law. However, the best fits were in the region of two to five query terms: suggesting that communication is most efficient in this region. In a follow up analysis, we found that it was also in this region that the marginal performance was maximized. After this point the Law of Diminishing Returns took hold: such that additional query terms resulted in smaller and smaller increases to performance. These findings regarding the most efficient query length closely match the lengths of queries that users actually express [4]. But here, we have provided a substantive analysis to show the economics involved, along with a characterization of the distribution of retrieval effectiveness. Since our findings suggest the distribution does not follow a power law, this analysis also lends weight to Blair's arguments that there is an imbalance between general and specific terms as descriptors of documents; and that communication with IR systems is not as efficient as it could be. Addressing this imbalance could lead to significant improvements in retrieval performance.

The main limitations of this study are: we have constrained our analysis to a particular set of query models and retrieval models. It may be the case that other query generation models or other retrieval models would result in retrieval effectiveness which does follow a power law. Further, this study was constrained to individually testing topics for power laws for a given length – as opposed to a set of queries composed of different lengths. It should also be noted that we assumed that length is reflective of effort.

While this is a reasonable approximation it would be interesting to examine other measures of effort (for instance, the specificity or generality of terms as measured by IDF, or some other query difficulty predictor [10, 17, 18, 20]). We also employed the use of a novel method for generating queries for ad-hoc topics which may have produced queries that are not realistic. However, since we were interested in all possible queries for a given topic to obtain an estimate of the distribution of retrieval effectiveness, we believe this to be a valid approach for the study undertaken. For further research though, this limitation strongly motivates developing better models of query generation in order to explore other query side evaluation issues (and to better characterize user querying behavior).

Nonetheless, we are confident, that despite these limitations, the findings are revealing and show that the distribution of retrieval effectiveness can not be adequately characterized by Zipf's Law under the considered conditions. This means that an empirical estimate of the distribution of retrieval effectiveness will have to suffice. However, future work will investigate whether other long tailed distributions, such as the negative binomial, could provide a better fit to retrieval effectiveness. Given the query generation model for ad hoc topics, it is now possible to obtain an estimate of the empirical distribution of retrieval effectiveness. This estimate could be useful in a number of different application areas, such as: Query Performance Prediction, Query Expansion, Topic Difficulty and Evaluation. In these areas, knowledge of the distribution of retrieval effectiveness is not utilized by current methods, despite its underlying importance. As it provides the necessary context to make decisions: for instance, Figure 2 contextualizes the difficulty of topics. Finally, there are many other interesting query side evaluation research questions that can be investigated (such as those raised in Section 1). Future work will consider these directions.

Acknowledgements I would like to thank Dr. Marcus Howlett and Dr. Mark Baillie for their insightful comments and the useful discussions, and I would also like to thank the anonymous reviewers for their thoughtful and constructive suggestions and feedback.

6. REFERENCES

[1] G. K. Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, 1949.

[2] H. S. Heaps. *Information Retrieval - Computational and Theoretical Aspects*. Academic Press, 1978.

[3] C. R. S. A. Clauset and M. E. J. Newman. Power-law distributions in empirical data. *URL http://arxiv.org/abs/0706.1062v1*, 2007.

[4] A. Arampatzis and J. Kamps. A study of query length. In *SIGIR 08: Proceedings of the 31st annual international ACM SIGIR conference*, pages 811–812, 2008.

[5] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR 07: Proceedings of the 30th annual international ACM SIGIR conference*, pages 455–462, 2007.

[6] R. K. Belew. *Finding Out About*. Cambridge Univ. Press, 2000.

[7] D. C. Blair. The challenge of commercial document retrieval, part i: major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size. *Information Processing Management*, 38(2):273–291, 2002.

[8] S. Bradford. Sources of information on specific subjects. *Journal of Information Science*, 10(4):173–180, 1985.

[9] C. Buckley, D. Dimmick, I. Soboroff, and E. Voorhees. Bias and the limits of pooling. In *SIGIR 06: Proceedings of the 29th annual international ACM SIGIR conference*, pages 619–620, 2006.

[10] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *SIGIR 02: Proceedings of the 25th annual international ACM SIGIR conference*, pages 299–306, 2002.

[11] M. Inoue and N. Ueda. Retrieving lightly annotated images using image similarities. In *Proc. 2005 ACM symposium on Applied computing*, pages 1031–1037, 2005.

[12] C. Jordan, C. Watters, and Q. Gao. Using controlled query generation to evaluate blind relevance feedback algorithms. In *JCDL 06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pages 286–295, 2006.

[13] D. Kelly, V. D. Dollu, and X. Fu. The loquacious user: a document-independent source of terms for query expansion. In *SIGIR 05: Proceedings of the 28th annual international ACM SIGIR conference*, pages 457–464, 2005.

[14] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR 01: Proceedings of the 24th annual international ACM SIGIR conference*, pages 120–127, 2001.

[15] A. J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–324, 1926.

[16] D. R. Miller, T. Leek, and R. M. Schwartz. A hidden Markov model information retrieval system. In *SIGIR 99: Proceedings of 22nd ACM International Conference*, pages 214–221, Berkeley, US, 1999.

[17] F. Scholer, H. Williams, and A. Turpin. Query association surrogates for web search. *Journal of the American Society for Information Science and Technology*, 55(7):637–650, 2004.

[18] 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*, pages 512–519, 2005.

[19] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR 01: Proceedings of the 24th annual international ACM SIGIR conference*, pages 334–342, 2001.

[20] Y. Zhao, F. Scholer, and Y. Tsegay. Effective pre-retrieval query performance prediction using similarity and variability evidence. In *ECIR 08: Proceedings of the European Conference in Information Retrieval*, pages 52–64, 2008.