A Study of Query Length

Avi Arampatzis[1] Jaap Kamps[1,2]

[1] Archives and Information Studies, University of Amsterdam
[2] ISLA, Informatics Institute, University of Amsterdam

ABSTRACT

We analyse query length, and fit power-law and Poisson distributions to four different query sets. We provide a practical model for query length, based on the truncation of a Poisson distribution for short queries and a power-law distribution for longer queries, that better fits real query length distributions than earlier proposals.

Categories and Subject Descriptors: H.3 [Information Storage and Retrieval]: H.3.3 Information Search and Retrieval; H.3.7 Digital Libraries
General Terms: Measurement, Experimentation, Theory
Keywords: Query length, Power-law, Zipf's law, Transaction log analysis

1. INTRODUCTION

From analysing query-logs, previous research has suggested that the distribution of query lengths can be approximated with the (generalized) *Zipf's law* or a *power-law* [5, 7]. The law appears to fit well to the largest length observations $k \geq k_0$ (where k_0 depends on the domain) but not to the whole sample. For example, data show that the length frequency for web queries peaks at 2 rather than at single keyword queries, suggesting a $k_0 > 2$. In the discrete case, the fraction of queries with length k is given by

$$p(k) = Pr(X = k) = Ck^{-s} \quad \text{for } k \geq k_0 \tag{1}$$

with C a normalizing constant, s the scaling parameter, and k_0 a lower bound from which onwards the power-law holds [2].

Others, without empirical justification, modeled query lengths with a Poisson distribution by setting its mean to the average query length [1]. Using a Poisson distribution, in a population of queries with average length μ, the fraction of queries with length k is

$$\text{Poisson}(k;\mu) = \frac{\mu^k e^{-\mu}}{k!}. \tag{2}$$

In this paper, we provide a model for query length. Beyond the theoretical interest, such a model has also practical applications in optimizing query cache size in search engines, in generating simulated queries for efficiency testing, and for effectiveness evaluation [e.g., 1]. Using several query data-sets, we confirm that the right tail of the length distribution is better approximated with a power-law rather than a Poisson, and introduce a truncated model. In the process, we estimate the slope s for English queries. Finally, we speculatively explain why some data deviate from a power-law and what this may mean for IR.

2. EMPIRICAL DATA

Four different query data-sets were analyzed: TREC [6]'s Million Query Track 2007, Terabyte Tracks 2005 and 2006, and AOL [3]; Tab. 1 shows (in the first seven columns): the query set; the number of queries; the minimum; maximum; mean; and median query length; and the most frequent query length. What is striking is that, although typical queries are short with 2-4 terms, longer queries up to 245 for AOL do occur.

Table 1: Query length statistics per data-set and the resulting power-law fit values.

queries	N	min	max	mean	median	peak	k_0	s
MQ07	10k	1	30	4.11	4	3	6	5.10
TB05	50k	1	18	2.79	2	2	5	4.89
TB06	100k	1	39	4.11	4	3	9	5.84
AOL	21M	1	245	2.34	2	1	6	4.92

We applied the methods of Clauset et al. [2] to automatically determine the scaling parameter, s, using maximum likelihood methods, and the lower bound, k_0, by minimizing the Kolmogorov-Smirnov statistic. These values are in the final two columns of Tab. 1 and the resulting distribution is depicted in Fig. 1.[1] The method gives reasonable power-law fits by excluding short lengths ($< k_0$). The scaling parameter is around 5 for query lengths. This gives a much steeper slope on log-log plots than the familiar Zipfian distribution of word frequencies. Fig. 1 also shows (with impulses) the Poisson distribution used by [1]; it matches well with the data at short lengths, but clearly lacks the tail.

The power-law model matches better the data in wider ranges of lengths than the Poisson model. While the fits on MQ07 and TB06 are good, there are some deviations in the right tails of the TB05 and AOL data. The TB05 data may be better fitted with a similar distribution, namely a power-law with an exponential cut-off [2]: $p(k) = Ck^{-s}e^{-\lambda k}$. Note that Eq. 1 can be obtained from the latter for $\lambda = 0$. This distribution is a common alternative because it captures finite-size effects, e.g., earthquake magnitude data have the same lack of tail due to the finite amount of energy in Earth's crust. We can speculate that the data-set was probably created in a way that there was a maximum query length imposed. The imposed maximum can be indirect, e.g., a result of the maximum amount of effort and time users are willing to put into formulating queries. The "bump" in the AOL data seems of a technical nature possibly due to a shift from typed queries to cut-and-paste queries; not knowing whether and how these data are processed, we will not speculate any further.

3. TOWARDS A PRACTICAL MODEL

All the fits on the four data-sets are problematic at short lengths. We argue that the deviation of the data from the power-law is an

[1] While it is clearer to plot the complementary cumulative distribution function (CCDF) $P(k) = Pr(X > k)$, we opt here top plot Eq. 1 directly so that we can show the Poisson distribution over initial query lengths in the same figure.

SIGIR'08, July 20–24, 2008, Singapore.
ACM 978-1-60558-164-4/08/07.

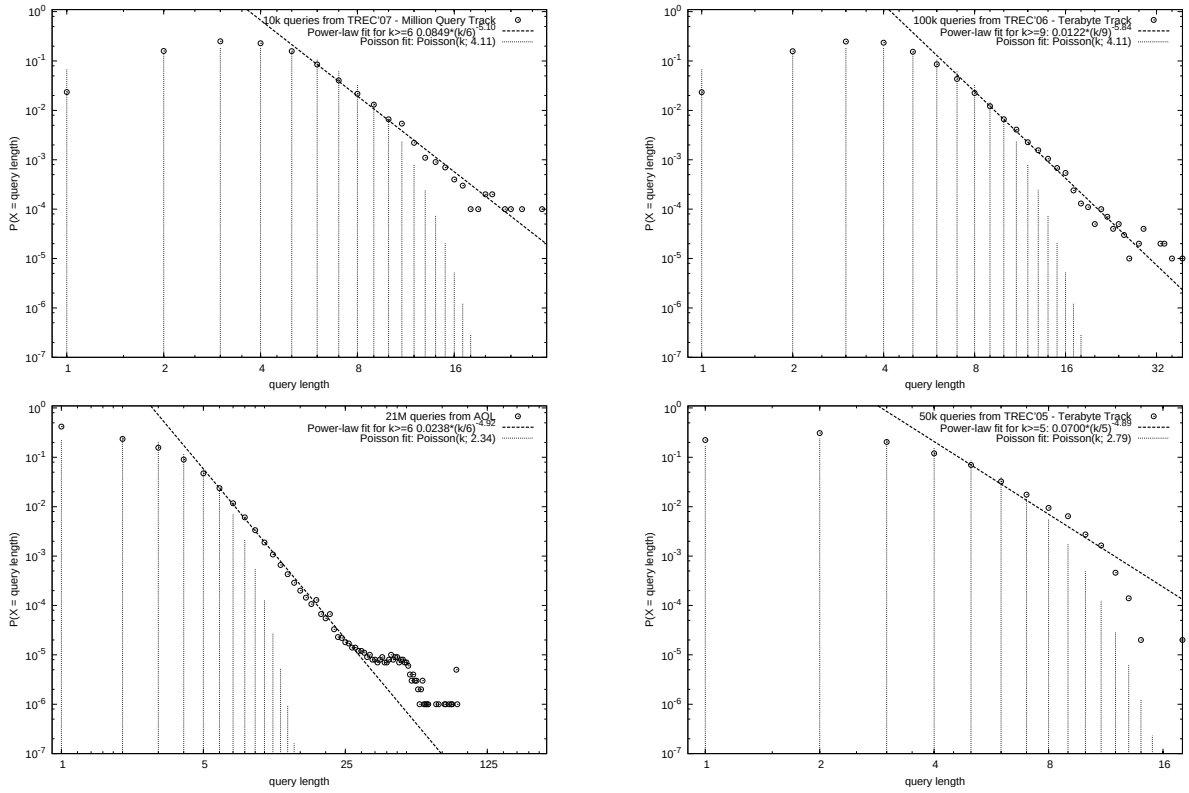

Figure 1: Power-law and Poisson fits on query-lengths of four data-sets.

artifact of the *specificity* of the indexing tokens provided by the query/indexing language. One can imagine a system indexing on binary-terms (adjacent pairs of keywords) or n-grams, where single index tokens are adequately specific to be used more frequently than used in multiples.[2] For example, re-indexing on word sequences of length $(k_0 - 1)$ would map queries of k_0 words to a length of 2 tokens, $k_0 + 1$ to 3 tokens, etc., and queries with $< k_0$ words will receive 1 token length. The net effect will be to merge the initial data-points to a single one, raised to better fit the power-laws in Fig. 1. This is not an unrealistic scheme since it is similar to indexing on all potential phrases *and* their individual words.

Since the bulk of queries concentrates at short lengths where a power-law does not fit given the current indexing languages, it makes practical sense to use a mix of truncated Poisson/power-law to model query lengths. In such a practical model, lengths are Poisson-distributed for $1 < k < k_0$ and power-law-distributed for $k \geq k_0$. The choice of k_0 depends on the specific domain, i.e., a combination of features of the document collection, query/indexing language, and pattern of use of the system.

4. CONCLUSIONS

All our power-law fits on the distribution of lengths of English queries resulted in exponents of around 5, giving a much steeper slope in log-log plots than the power-law known to hold for word frequencies. This result probabilistically forecasts the lengths of the natural language fragments humans use to formulate information needs. The relative steepness of the power-law indicates that users do not need many words to formulate information needs or that the diminishing value of adding words appears soon.

Deviations of real data from the power-law may be explained by either finite-size effects or insufficient specificity of indexing terms. Studies in other fields, e.g., economics, have shown that deviations of data from the power-law at hand are usually an indication of inefficiencies in the system that the data come from. Pennock et al. [4] studied power-law distributions of numbers of web links and found that deviations of data from the power-law per category of websites correlate to how much competition is present in that category. The better the power-law fit, the more competitive the category. For query lengths, we have shown how a simple process of re-indexing on longer text fragments can "fix" some deviations, a fact that may point to inefficiencies in single-word indexing schemes.

[2]Of course, such an indexing scheme will effectively square the order of index size.

Acknowledgments This research was supported by the Netherlands Organization for Scientific Research (NWO, under project # 640.001.501).

REFERENCES

[1] L. Azzopardi, M. de Rijke, and K. Balog. Building simulated queries for known-item topics: an analysis using six european languages. In *SIGIR '07*, pages 455–462, New York, NY, USA, 2007. ACM.

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

[3] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search. In *InfoScale '06: Proceedings of the 1st international conference on Scalable information systems*. ACM Press, New York NY, USA, 2006.

[4] D. M. Pennock, G. W. Flake, S. Lawrence, E. J. Glover, and C. L. Giles. Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8): 5207–5211, 2002.

[5] S. Sharma, L. T. Nguyen, and D. Jia. Ir-wire: A research tool for p2p information retrieval. In *SIGIR Open Source Workshop*, Seattle, 2006. ACM.

[6] TREC. Text REtrieval Conference, 2008. `http://trec.nist.gov/`.

[7] I. Zukerman and E. Horvitz. Using machine learning techniques to interpret wh-questions. In *ACL*, pages 547–554, 2001.