Adaptive Distributional Extensions to DFR Ranking

Casper Petersen[1], Jakob Grue Simonsen[1], Kalervo Järvelin[2], Christina Lioma[1]

[1] Department of Computer Science, University of Copenhagen

[2] School of Information Sciences, University of Tampere

[1] cazz,simonsen,c.lioma,(@di.ku.dk), [2] Kalervo.Jarvelin@staff.uta.fi

1. Introduction

Divergence From Randomness (DFR):

$$R(q,d) = \sum_{t \in q \cap d} \overbrace{(-\log_2 P_1)}^{\text{model of randomness}} \cdot \underbrace{(1-P_2)}_{\text{information gain}} \quad (1)$$

Key Assumption in DFR

Non-informative terms are distributed differently than informative terms [1].

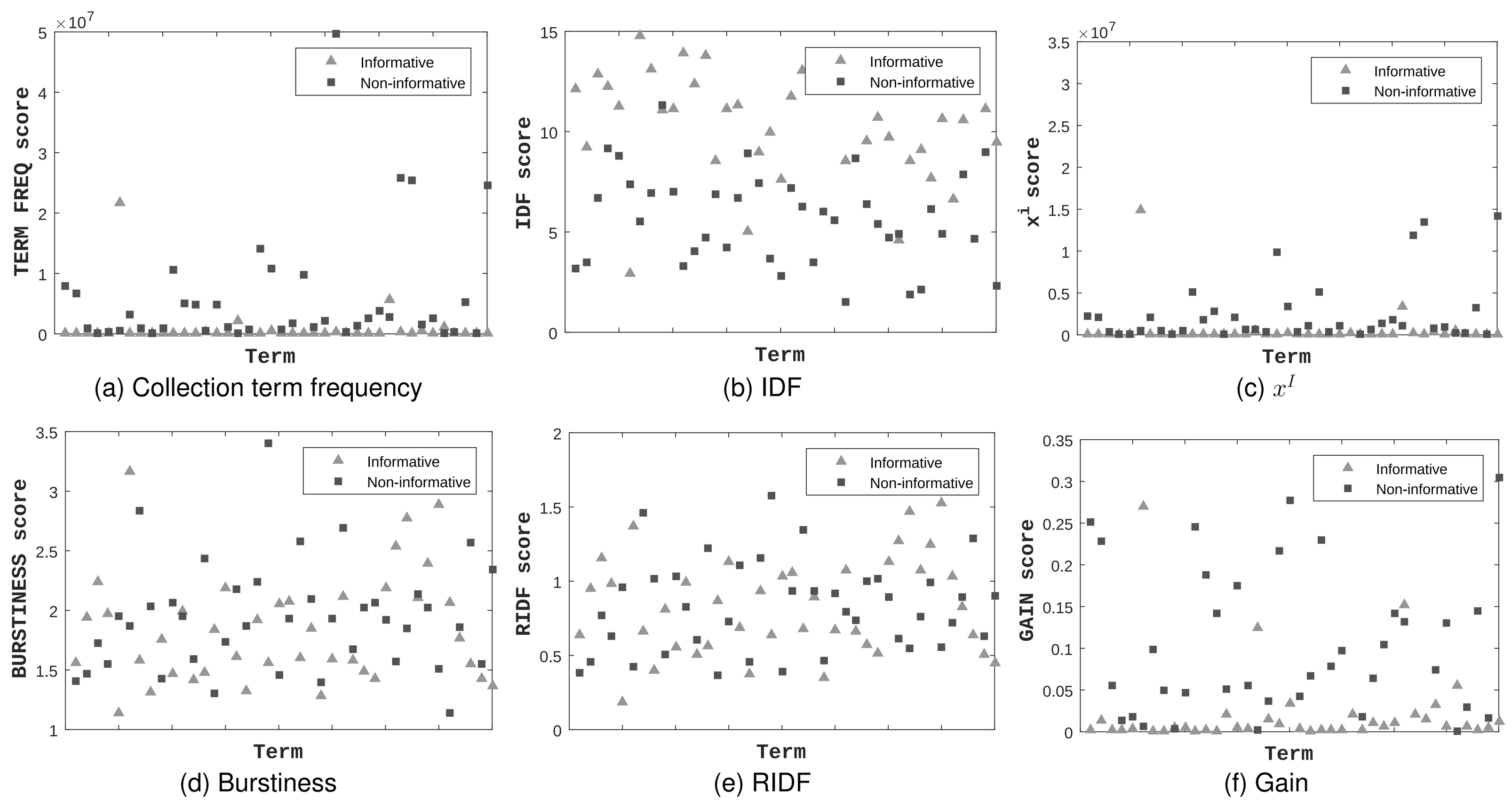
- DFR assumes e.g. Poisson or Geometric distributions of non-informative terms.
- ***Unsubstantiated distributional assumption may lead to sub-optimal ranking models.***

2. Research Question

Will using the best-fitting distribution to non-informative terms improve ranking effectiveness?

3. Methodology

Step 1: Identify Non-informative Terms

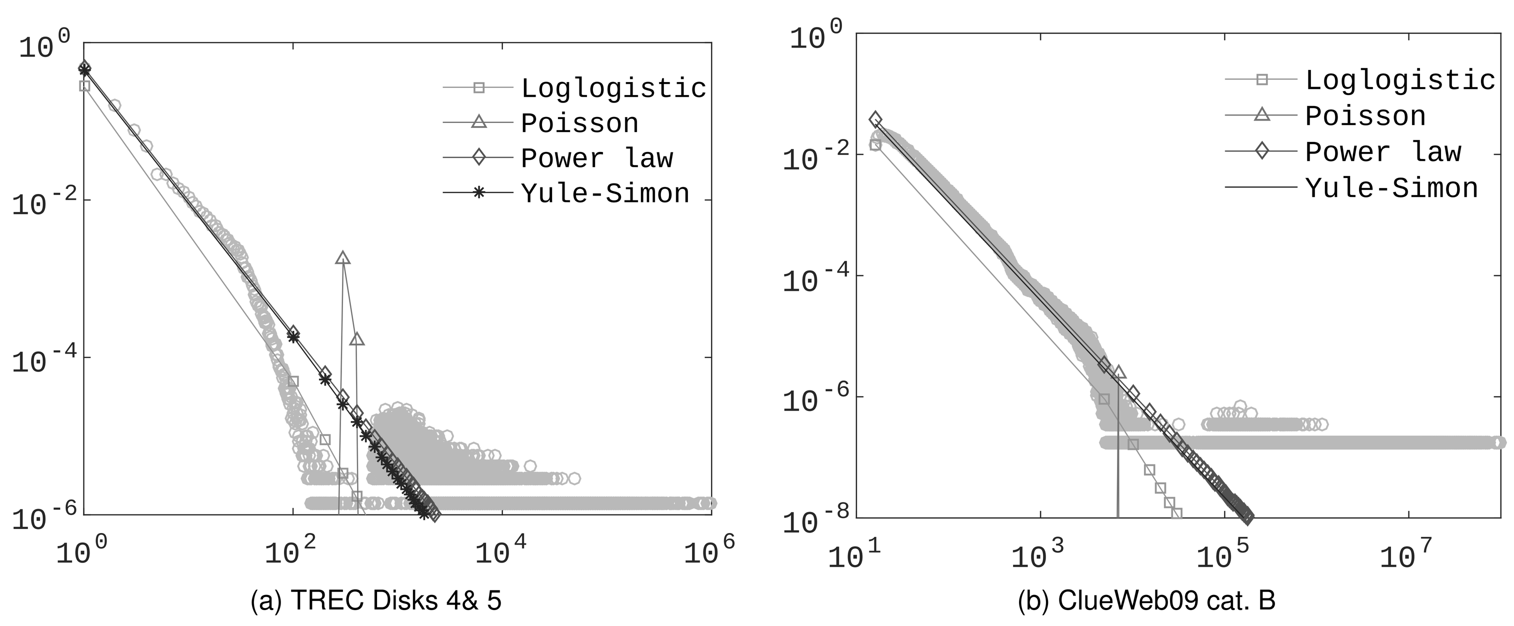


(a) Collection term frequency (b) IDF (c) x^I (d) Burstiness (e) RIDF (f) Gain

Figure 1: *Weights of informative and non-informative terms.*

Step 2: Fit and Select Candidate Statistical Models

1. Model Estimation: Estimate the parameters of the statistical models using maximum likelihood estimation.
2. Model Comparison: Compare each pair of models using Vuong's likelihood ratio (LR) test [6].
3. Model Selection: Choose the model that wins most pairwise comparisons accordint to the LR test.



(a) TREC Disks 4& 5 (b) ClueWeb09 cat. B

Figure 2: *Distribution of non-informative terms (grey circles) in TREC Disks 4&5 and ClueWeb09 cat. B. Superimposed are log-logistic, Poisson, power law and Yule-Simon distributions. The Yule-Simon distribution provides the best separation between informative and non-informative words.*

Step 3: Adaptive Distributional Ranking

$$R(q,d) = \sum_{t \in q \cap d} (-\log_2 P_1) \cdot (1-P_2) = \sum_{t \in q \cap d} \left(-\log_2 \hat{M}\right) \cdot (1-P_2) \quad (2)$$

- Captures the assumption that non-informative terms are distributed in a certain way.

4. Findings

Dataset	Best-fitting discrete distribution	Best-fitting continuous distribution
ClueWeb09 cat. B.	Yule–Simon ($p=1.627$)	Generalized Extreme Value ($k=1.322, \sigma=30.28, \mu=36.18$)
TREC Disks 4&5	Yule–Simon ($p=1.804$)	Generalized Extreme Value ($k=4.198, \sigma=0.7295, \mu=1.174$)

Table 1: *Best-fitting discrete and continuous statistical models for each dataset.*

4.1 Yule-Simon (YS) Model

- Used for e.g. text generation [5] and citation analysis [4].

$$\hat{M} = \text{YS}(x|p) = \left\{ p \cdot \frac{\Gamma(x) \cdot \Gamma(p+1)}{\Gamma(x+p+1)} : x \in \mathbb{Z}^+, p > 0 \right\} \quad (3)$$

5. Experimental Setup

- Queries 301-450, 601-700 for TREC Disks 4 & 5. Queries 1-200 for ClueWeb09 cat. B.
- Language model w. Dirichlet smoothing (LMDir).
- Poisson (P) and tf-idf (I$_n$) DFR models [1] and information-based models (LL, SPL) [2, 3].
- All models use Laplace's Law of Succession and logarithmic term-normalisation.
- Model parameter set to $T_{dc} = \frac{n_t}{|C|}$ or $T_{tc} = \frac{f_{t,C}}{|C|}$ [1, 2, 3].
- E.g. YSL2-T_{dc} is the YS model with Laplace's law of Succession and logarithmic term-normalisation (L2) with $T_{dc} = p = \frac{n_t}{|C|}$.
- All models tuned using 3-fold cross validation.

6. Results

Model	TREC disks 4& 5 nDCG	P@10	Bpref	ERR@20	nDCG@10
LMDir	.4643	.3845	.2239	.1043	.3968
PL2-T_{tc} [1]	.2524*	.1273*	.1009*	.0359*	.1332*
PL2-T_{dc} [1]	.2487*	.1217*	.0960*	.0347*	.1273*
I$_n$L2-T_{tc} [1]	.2917*	.1627*	.1114*	.0478*	.1742*
I$_n$L2-T_{dc} [1]	.2818*	.1626*	.1088*	.0481*	.1745*
LLL2-T_{tc} [2]	.4812	.4049	.2341	.1072	.4142
LLL2-T_{dc} [2]	.4810	.3982	.2329	.1069	.4097
SPLL2-T_{tc} [3]	.4863	.4144	.2375	.1103	.4276
SPLL2-T_{dc} [3]	**.4876**	.4176	**.2387**	.1107	.4299
YSL2-T_{tc} (ADR)	.4644	.3982	.2280	.1048	.4069
YSL2-T_{dc} (ADR)	.4860	**.4182**	.2381	**.1113**	**.4312**

Model	ClueWeb09 cat. B. nDCG	P@10	Bpref	ERR@20	nDCG@10
LMDir	.2973	.2586	.2209	.0973	.1769
PL2-T_{tc} [1]	.1448*	.0712*	.1258*	.0211*	.0472*
PL2-T_{dc} [1]	.1444*	.0709*	.1252*	.0314*	.0471*
I$_n$L2-T_{tc} [1]	.1596*	.0782*	.1405*	.0352*	.0511*
I$_n$L2-T_{dc} [1]	.1596*	.0783*	.1407*	.0352*	.0512*
LLL2-T_{tc} [2]	.3184	.2542	.2349	.0926	.1706
LLL2-T_{dc} [2]	.3180	.2542	.2349	.0928	.1707
SPLL2-T_{tc} [3]	.3207	.2529	.2357	.0945	.1720
SPLL2-T_{dc} [3]	.3224	.2586	.2370	.0958	.1752
YSL2-T_{tc} (ADR)	.3197	.2601	.2359	.0951	.1752
YSL2-T_{dc} (ADR)	**.3240**	**.2666**	**.2376**	**.0985**	**.1810**

Table 2: *Gray cells denote results better than LMDir. Results in bold are best overall for each performance measure.* * *denotes statistically significant difference from the LMDir.*

References

[1] Gianni Amati and Cornelis J.V. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.

[2] Stéphane Clinchant and Éric Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for IR. In *ICTIR*, pages 54–65, 2009.

[3] Stéphane Clinchant and Eric Gaussier. Information-based models for Ad Hoc IR. In *SIGIR*, pages 234–241. ACM, 2010.

[4] Derek d.S. Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5):292–306, 1976.

[5] Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.

[6] Quang H. Vuong. Likelihood ratio tests for model selection & non-nested hypotheses. *Econometrica*, 57:307–333, 1989.

The 25th ACM International Conference on Information and Knowledge Management (CIKM 2016), October 24-28, Indianapolis, USA.

Work partially funded by C. Lioma's FREJA research ex- cellence fellowship (grant no. 790095).