CIKM2011

Relevance Weighting using Within-document Term Statistics

Kai Hui, Ben He

huikai10@mails.gucas.ac.cn
benhe@gucas.ac.cn

Tiejian Luo, Bin Wang

tjluo@gucas.ac.cn
wangbin@ict.ac.cn

中国科学院
CHINESE ACADEMY OF SCIENCES

- Research Motivation
- Fitting the TF Distribution
- Propose NG Models
- Evaluation & Simplified NG Models
- Conclusion & Future works

Research Motivation

✓ Problem

Traditional popular models apply global statistics (Document frequency, Token numbers in the collections). Sometimes, it is difficult or infeasible to get Global Statistics

✓ Take PL2 based on DFR for Example

The DFR framework (G. Amati, C. J. van Rijsbergen, 2002)

$$score(d,Q)=\sum_{t\in Q} qtf\cdot Inf_1\cdot Inf_2 \quad Inf_1=-\log_2 P(t,tf\mid d) \quad Inf_2=\frac{1}{tfn+1}$$

$$\sum_{t\in Q} qtf\cdot\frac{1}{tfn+1}\left(tfn\cdot\log_2\frac{tfn}{\boxed{\lambda}}+(\boxed{\lambda}+\frac{1}{12\cdot tfn}-tfn)\cdot\log_2 e+0.5\cdot\log_2(2\pi\cdot tfn)\right)$$

Derived from Bernoulli Process, use global statistics

Research Motivation

✓ Our Solutions

$$
\begin{aligned}
score(d,Q) &= \sum_{t\in Q} qtf \cdot Inf_1 \cdot Inf_2 \\
&= \sum_{t\in Q} qtf \cdot (-\log_2 \boxed{P(t,tf \mid d)}) \cdot \frac{1}{1+tfn}
\end{aligned}
$$

Approximate by fitting tf with a series of distribution functions, without using global statistics

NG models: Propose NG models (No Global statistics models) by replacing P with the tf distribution function

Fitting the TF Distribution

✓ tf Distribution

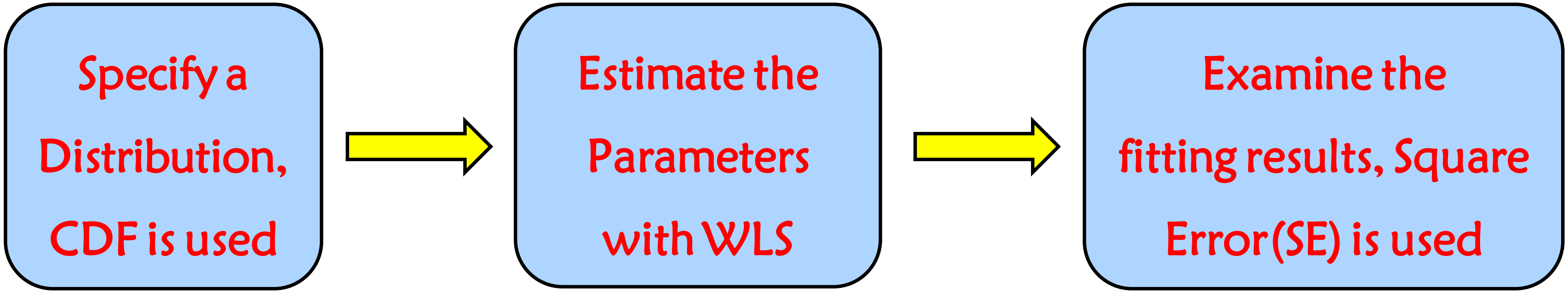
a. Zipf's law: CF is inversely proportional to its rank in the frequency table

b. Harter, 1975: 2-Poisson assumption over a sample from works of Sigmund Freud

✓ Fitting Process

a. Recent datasets have been used in our fitting

b. A list of potentially appropriate distribution functions have been tested



Specify a Distribution, CDF is used

Estimate the Parameters with WLS

Examine the fitting results, Square Error(SE) is used

Fitting the TF Distribution

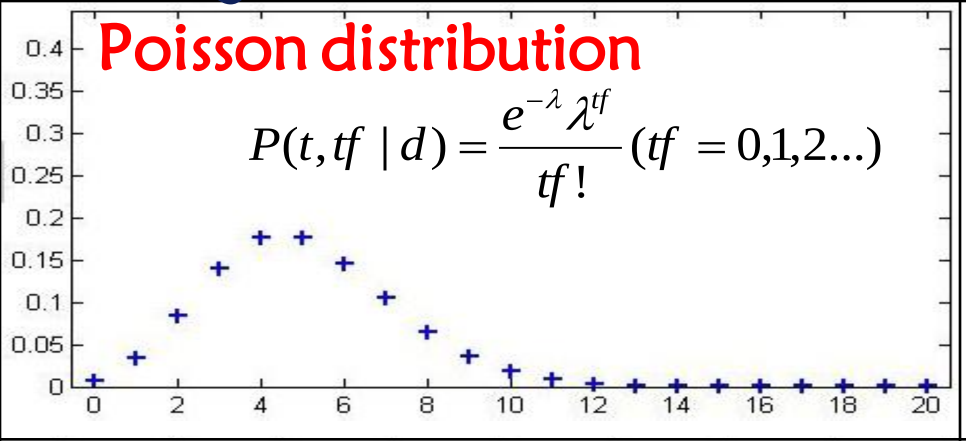
✓ The datasets

Coll.	TREC Task	Topics	#Docs
disk1&2	1,2,3 ad-hoc	51-200	741,856
disk4&5	Robust 2004	301-450,601-700	528,155
WT10G	9,10 Web	451-550	1,692,096
GOV2	2004-2006 Terabyte Ad-hoc	701-850	25,178,548

- Standard preprocesses are conducted: stop words, stemmer
- Only the terms in the title field are used

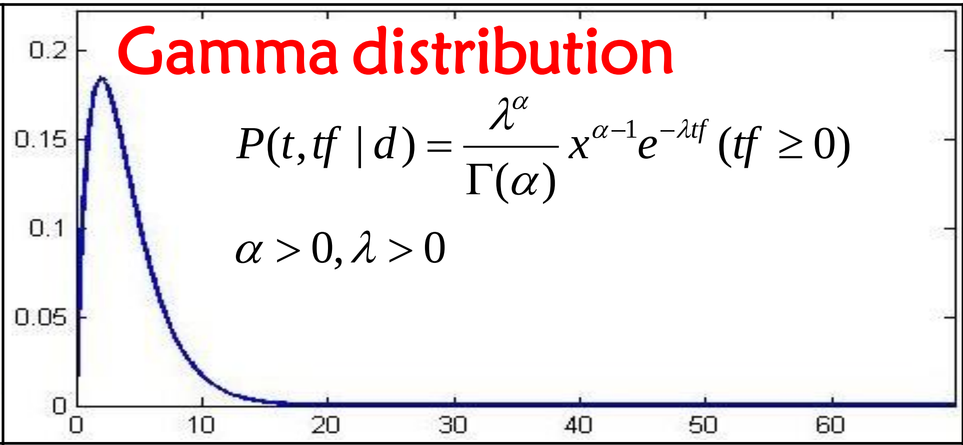
Fitting the TF Distribution

✓ The distributions



Poisson distribution

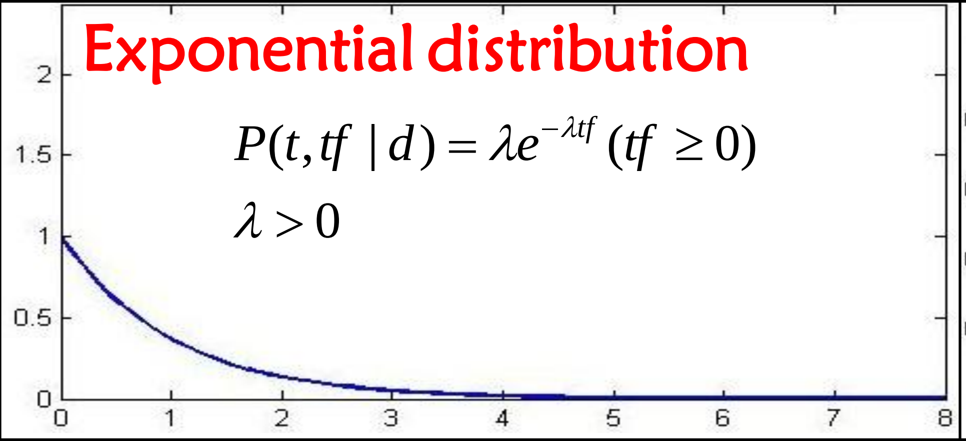
$$P(t,tf\mid d)=\frac{e^{-\lambda}\lambda^{tf}}{tf\,!}\ (tf\ =0,1,2...)$$



Gamma distribution

$$P(t,tf\mid d)=\frac{\lambda^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\lambda tf}\ (tf\ \geq 0)$$

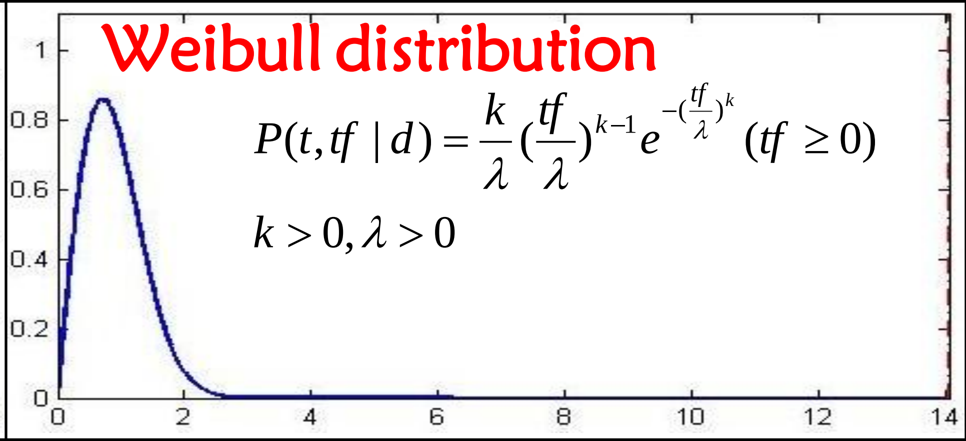
$\alpha>0,\lambda>0$



Exponential distribution

$$P(t,tf\mid d)=\lambda e^{-\lambda tf}\ (tf\ \geq 0)$$

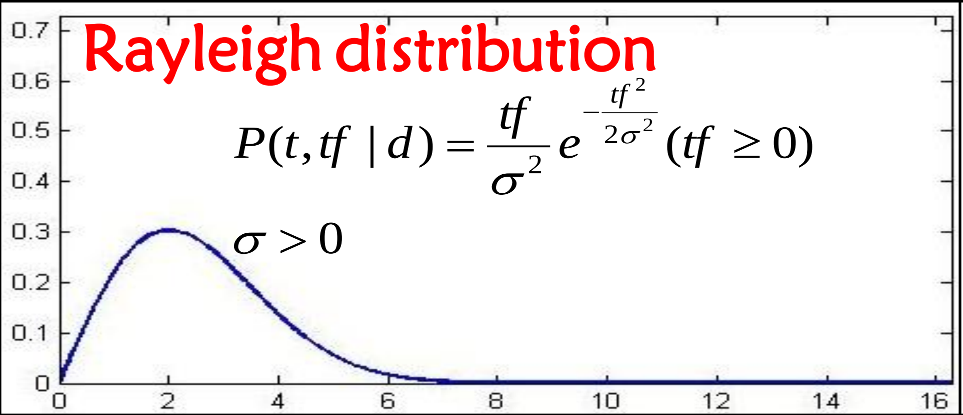
$\lambda>0$



Weibull distribution

$$P(t,tf\mid d)=\frac{k}{\lambda}(\frac{tf}{\lambda})^{k-1}e^{-(\frac{tf}{\lambda})^{k}}\ (tf\ \geq 0)$$

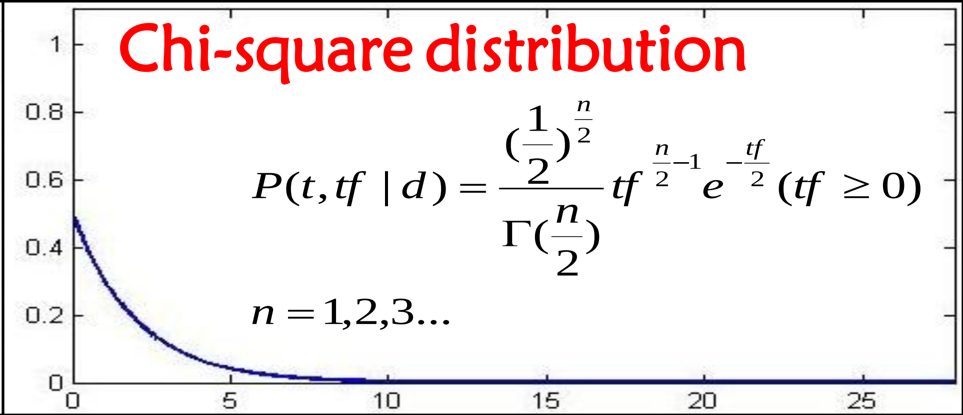
$k>0,\lambda>0$



Rayleigh distribution

$$P(t,tf\mid d)=\frac{tf}{\sigma^{2}}e^{-\frac{tf^{2}}{2\sigma^{2}}}\ (tf\ \geq 0)$$

$\sigma>0$



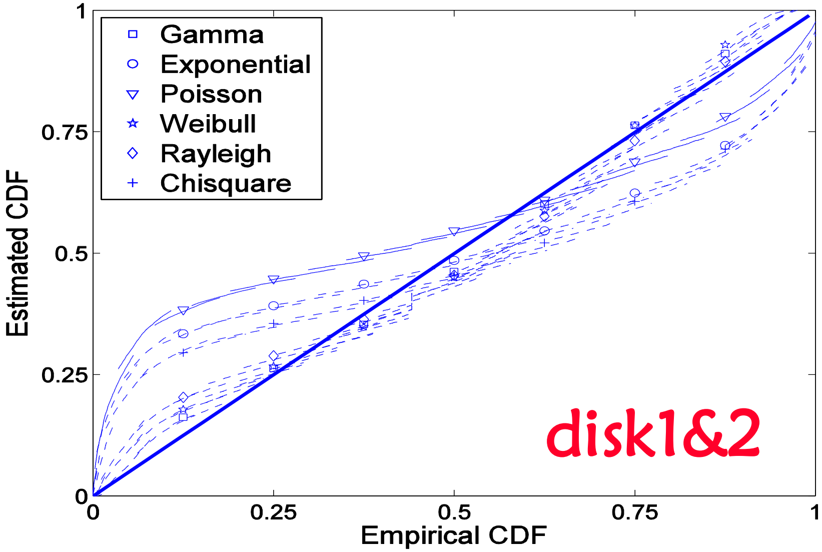
Chi-square distribution

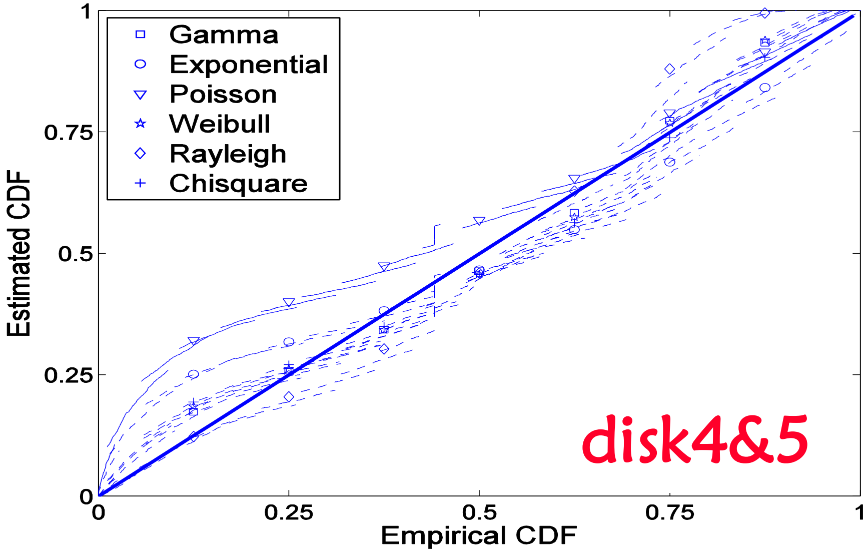
$$P(t,tf\mid d)=\frac{(\frac{1}{2})^{\frac{n}{2}}}{\Gamma(\frac{n}{2})}tf^{\frac{n}{2}-1}e^{-\frac{tf}{2}}\ (tf\ \geq 0)$$

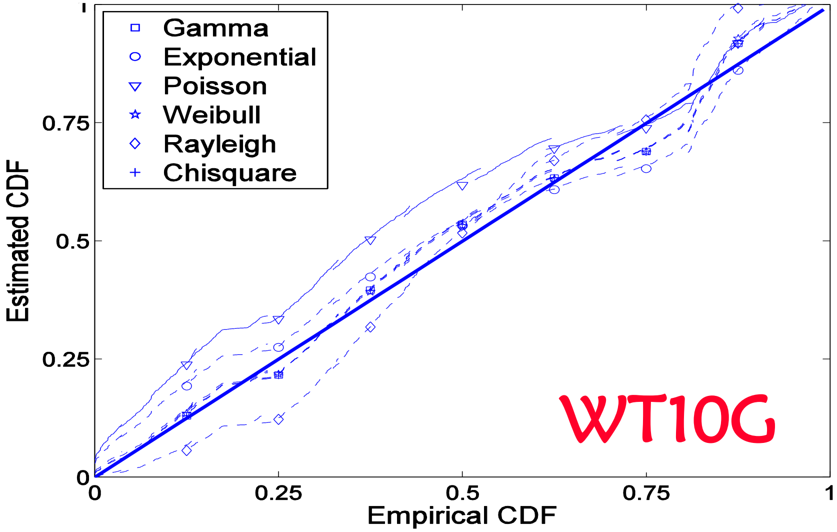
$n=1,2,3...$

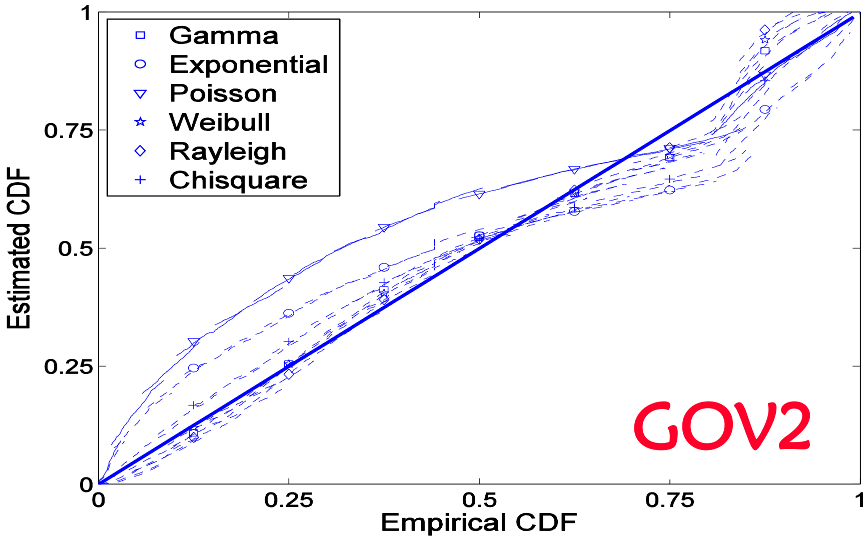
Fitting the TF Distribution

- ✓ P-P graphs: 6 distributions on 4 datasets



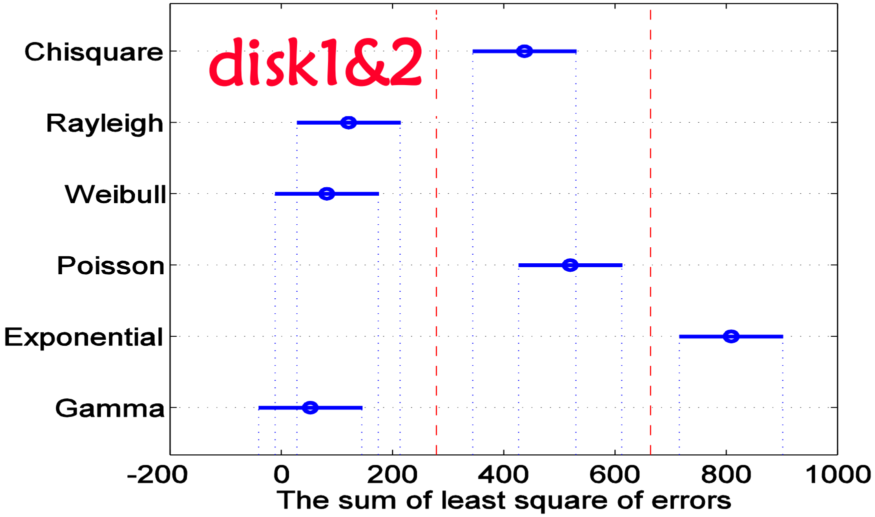


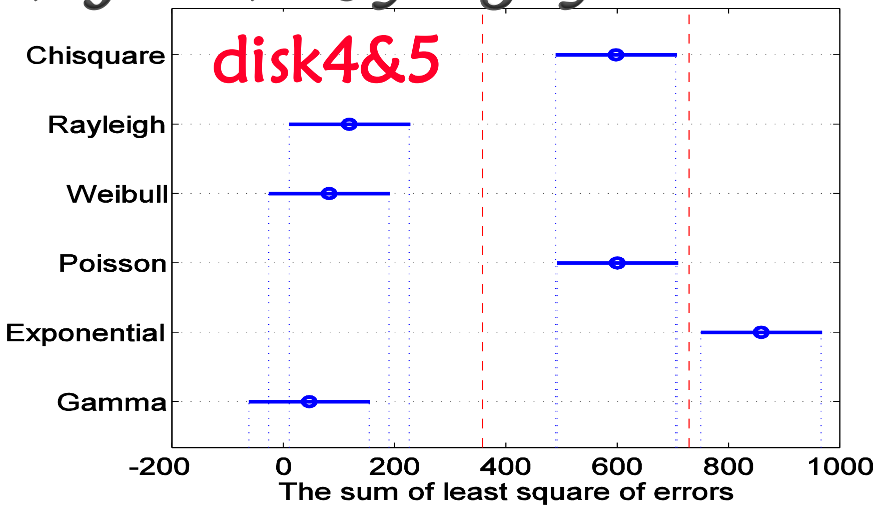


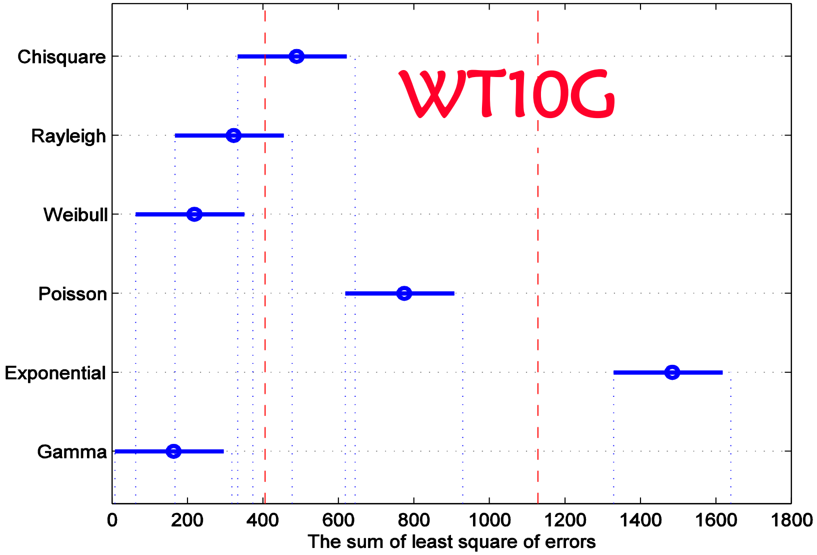


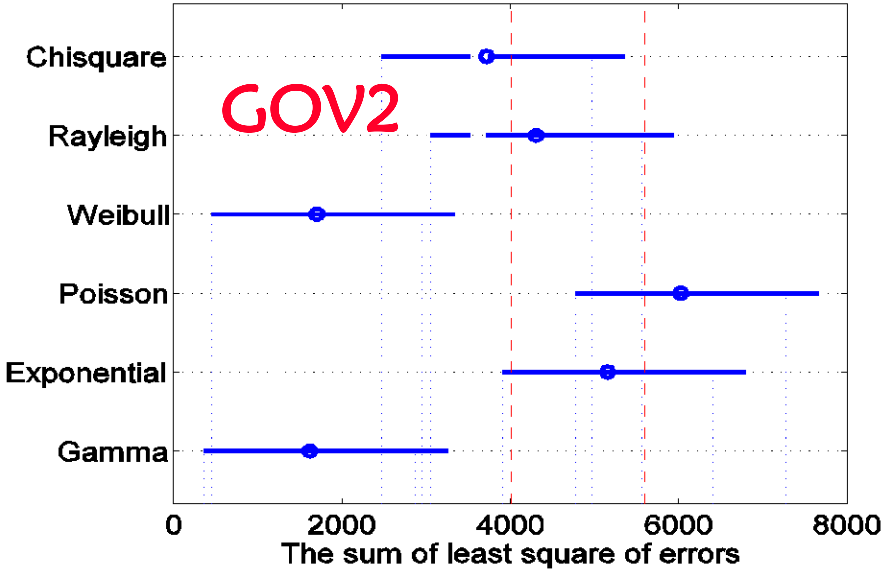
Fitting the TF Distribution

✓ ANOVA test: Weibull, Gama, Rayleigh fit best









Propose NG Models

✓ Propose NG models

$$score(d, Q) = \sum_{t \in Q} qtf \cdot Inf_1 \cdot Inf_2$$

$$= \sum_{t \in Q} qtf \cdot (-\log_2 P(t, tf \mid d)) \cdot \frac{1}{1 + tfn}$$

Treat parameters as FREE PRAMETERS which are tunable

Take the WL2d model as example:

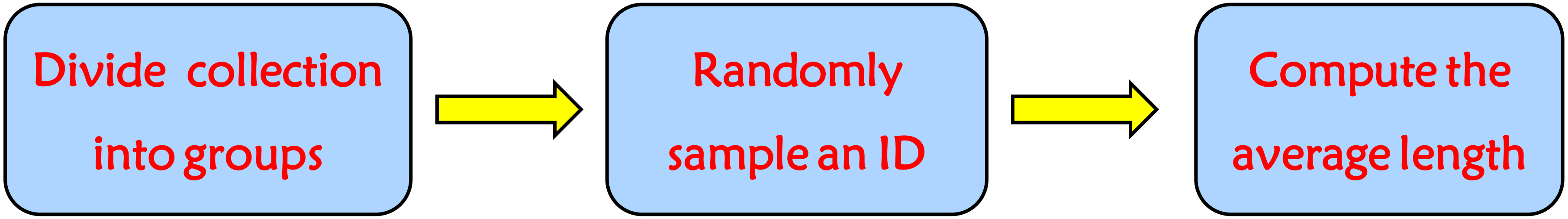
$$score_{WL2d}(d,Q) = \sum_{t \in Q} qtf \cdot (-\log_2 \frac{k}{\lambda}(\frac{tfn}{\lambda})^{k-1} e^{-(\frac{tfn}{\lambda})^k}) \cdot \frac{1}{1+tfn}$$

Normalization2 in DFR framework:

$$tfn = tf \cdot \log_2(1 + c \cdot (\frac{avg\_l}{l})), (c > 0)$$

Propose NG Models

✓ *Estimate the average document length*



Divide collection into groups

Randomly sample an ID

Compute the average length

a. Dividing the collection into several groups with approximately N documents in each groups giving every documents an unique ID(1,2,3, ··· N) within one group
b. Randomly sampling one number(X) within 1 to N
c. Recording the document length of No. X in every groups and computing the sample average document length

Propose NG Models

- ✓ Estimate the average document length

Coll.	*EstL*	*avg_l*	Error(%)
disk1&2	266.10	261.30	1.84
disk4&5	301.22	297.10	1.39
WT10G	406.68	399.28	1.85
GOV2	673.76	648.42	3.91

Coll.	Avg.(%)	MaxPos(%)	MinNeg(%)	CV
disk1&2	3.15	3.55	-9.23	0.8348
disk4&5	2.72	2.98	-6.80	0.7021
WT10G	3.07	3.90	-8.38	0.8306
GOV2	3.89	0.53	-8.37	0.4470

Evaluation & Simplified NG models

✓ Evaluation Settings

a. **Baseline:** BM25, KL-divergence language model, PL2

b. **Platform:** In-house version of the Terrier toolkit

c. **Validation:** Two-fold cross-validation

d. **Evaluation measure:** Mean Average Precision(MAP) and statistical significance are based on Wilcoxon matched-pairs signed-rank at .05 level

Evaluation & Simplified NG models

- ✓ Results

Coll.	disk1&2	disk4&5	WT10G	GOV2
KLLM	**.2351**	**.2565**	**.2153**	**.3028**
PL2	**.2336**	**.2570**	**.2126**	**.3042**
BM25	**.2404**	**.2535**	**.2080**	**.2997**
WL2d	.2024	.2300	.1774	.2890
WLBd	.2048	.2300	.1878	.2890
PL2d	.2044	.2301	.1934	.2855
PLBd	.2032	.2178	.1808	.2705
EL2d	.2004	.2294	.1760	.2778
ELBd	.2034	.2298	.1926	.2844
GL2d	.2004	.2289	.1702	.2635
GLBd	.1988	.2132	.1286	.2580
CL2d	.1630	.1936	.1055	.1538
CLBd	.1190	.1388	.0739	.0715
RL2d	.0664	.0541	.0436	.0305
RLBd	.0678	.0532	.0486	.0200

Evaluation & Simplified NG models

✓ Simplified NG models

a. **Free parameters:** Robustness is important in our model performance

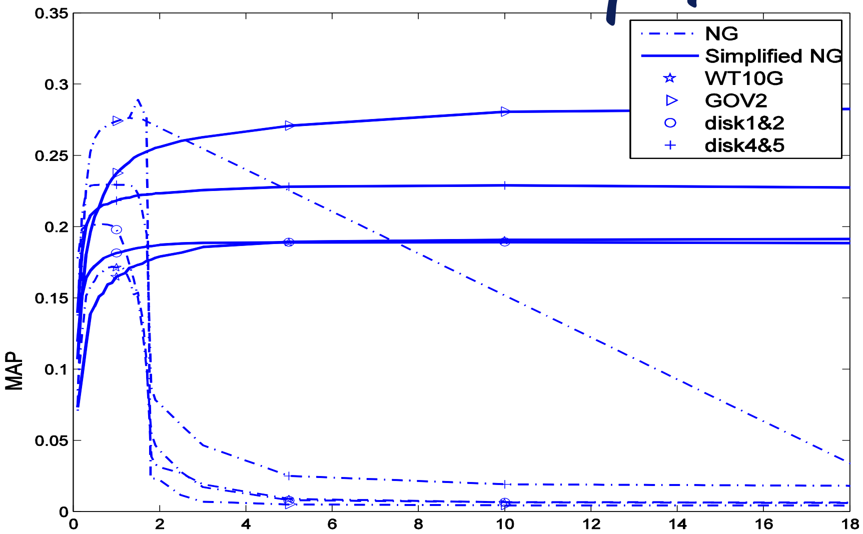
b. **Simplify models:** Replace Inf1 · Inf2 with formulae having same shape

$$score(d,Q)=\sum_{t\in Q} qtf\cdot\boxed{(-\log_2 P(t,tf\mid d))\cdot\frac{1}{1+tfn}}$$

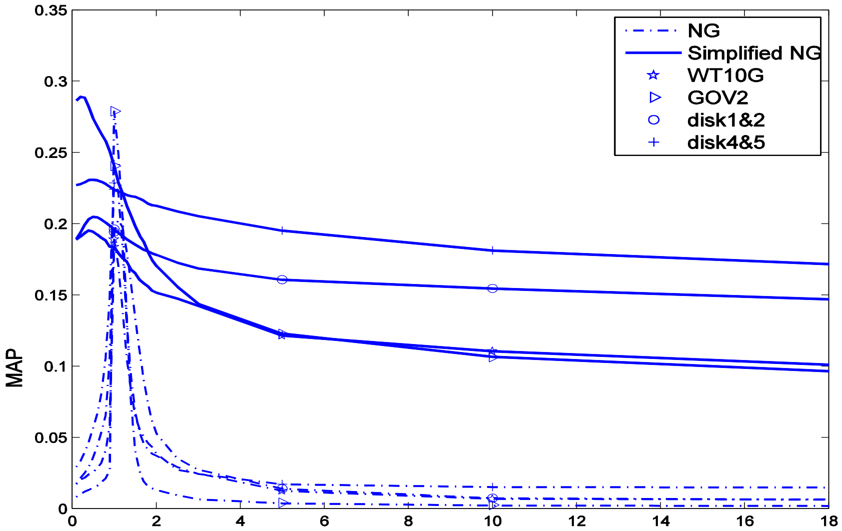
$$\Downarrow$$

$$score(d,Q)\propto\sum_{t\in Q} qtf\cdot\boxed{(1-P(tf,t\mid d))}$$

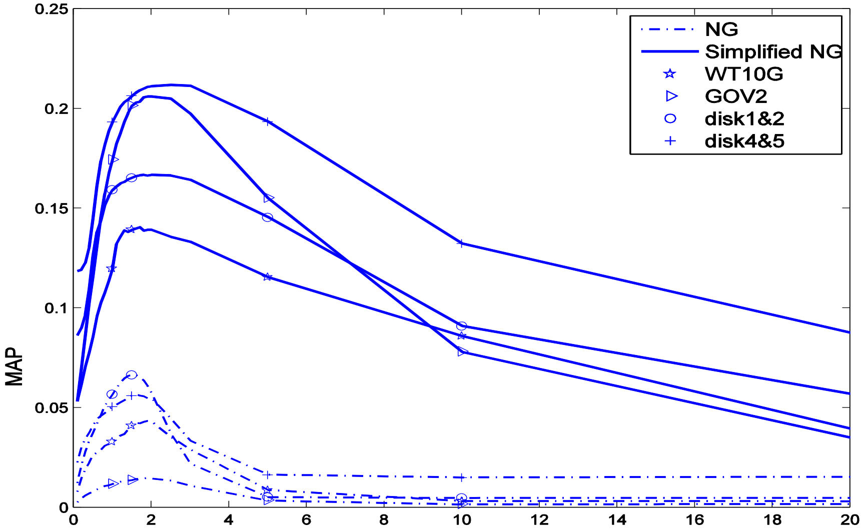
Evaluation & Simplified NG models



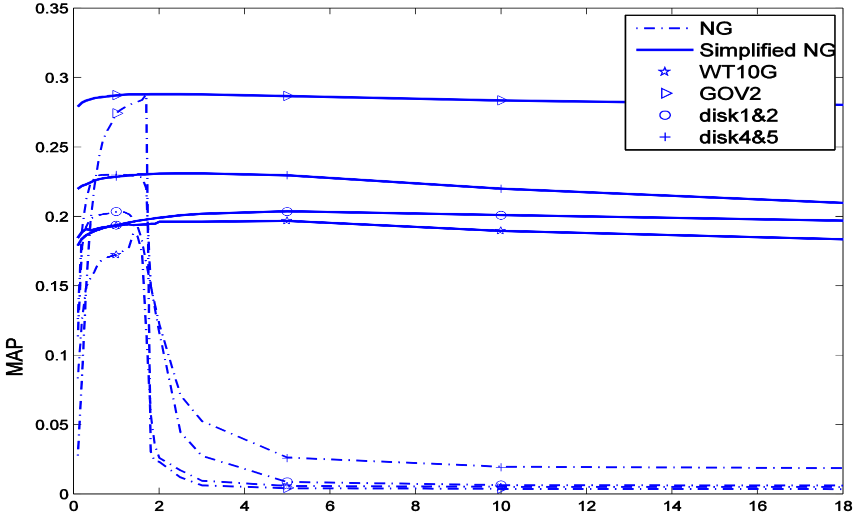
Gama distribution's parameter λ



Weibull distribution's parameter κ



Rayleigh distribution's parameter σ



Weibull distribution's parameter λ

Evaluation & Simplified NG models

✓ The results of the simplified NG models

Coll.	disk1&2	disk4&5	WT10G	GOV2
KLLM	**.2351**	**.2565**	**.2153**	**.3028**
PL2	**.2336**	**.2570**	**.2126**	**.3042**
BM25	**.2404**	**.2535**	**.2080**	**.2997**
W2dS	.2029	.2304	.1934	.2884
WBdS	.2048	.2284	.1920	.2878
E2dS	.1967	.2258	.1844	.2644
EBdS	.1966	.2247	.1871	.2630
G2dS	.1898	.2283	.1904	.2804
GBdS	.1918	.2280	.1934	.2866
C2dS	.1924	.2245	.1857	.2590
CBdS	.1974	.2284	.1946	.2586
R2dS	.1664	.2104	.1365	.2012
RBdS	.1656	.1930	.1276	.1938

Conclusion and Future Work

a. Apart from Poisson distribution there are other probabilistic models are suitable to describe the TF distribution

b. A list of NG models generated from the DFR framework are proposed ,

c. We improved the robustness of the NG models and W2dS can achieve better results

a. Both fitting results and the retrieval performance should be improved further.

b. The QE models for the NG model can be discovered

Thank you

~Any Questions~