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
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)

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

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

Derived from Bernoulli Process, use global statistics
Research Motivation

✓ Our Solutions

\[
\text{score}(d, Q) = \sum_{t \in Q} qtf \cdot \text{Inf}_1 \cdot \text{Inf}_2
\]

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

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

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

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

\textbf{Poisson distribution}

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

\textbf{Gamma distribution}

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

\( \alpha > 0, \lambda > 0 \)

\textbf{Exponential distribution}

\[ P(t, tf | d) = \lambda e^{-\lambda tf} (tf \geq 0) \]

\( \lambda > 0 \)

\textbf{Weibull distribution}

\[ P(t, tf | d) = \frac{k}{\lambda} \left( \frac{tf}{\lambda} \right)^{k-1} e^{-\left(\frac{tf}{\lambda}\right)^k} (tf \geq 0) \]

\( k > 0, \lambda > 0 \)

\textbf{Rayleigh distribution}

\[ P(t, tf | d) = \frac{tf}{\sigma^2} e^{-\frac{tf^2}{2\sigma^2}} (tf \geq 0) \]

\( \sigma > 0 \)

\textbf{Chi-square distribution}

\[ P(t, tf | d) = \frac{(\frac{1}{2})^{n/2}}{\Gamma(\frac{n}{2})} tf^{n-1} e^{-\frac{tf}{2}} (tf \geq 0) \]

\( n = 1,2,3,... \)
Fitting the TF Distribution

✓ P-P graphs: 6 distributions on 4 datasets

- disk1&2
- disk4&5
- WT10G
- GOV2
Fitting the TF Distribution

ANOVA test: Weibull, Gamma, Rayleigh fit best

disk1&2

disk4&5

WT10G

GOV2
Propose NG Models

Propose NG models

\[
\text{score}(d, Q) = \sum_{t \in Q} qtf \cdot \text{Inf}_1 \cdot \text{Inf}_2
\]

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

Treat parameters as FREE PRAMETERS which are tunable

Take the \text{WL2d} model as example:

\[
\text{score}_{WL2d}(d, Q) = \sum_{t \in Q} qtf \cdot (-\log_2 \frac{k}{\lambda} \left(\frac{tfn}{\lambda}\right)^{k-1} e^{-\left(\frac{tfn}{\lambda}\right)^k}) \cdot \frac{1}{1 + tfn}
\]

Normalization2 in DFR framework:

\[
\text{tfn} = tf \cdot \log_2 (1 + c \cdot \left(\frac{\text{avg} - l}{l}\right)), (c > 0)
\]
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 group giving every document 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

<table>
<thead>
<tr>
<th>Coll.</th>
<th>EstL</th>
<th>avg_l</th>
<th>Error(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>disk1&amp;2</td>
<td>266.10</td>
<td>261.30</td>
<td>1.84</td>
</tr>
<tr>
<td>disk4&amp;5</td>
<td>301.22</td>
<td>297.10</td>
<td>1.39</td>
</tr>
<tr>
<td>WT10G</td>
<td>406.68</td>
<td>399.28</td>
<td>1.85</td>
</tr>
<tr>
<td>GOV2</td>
<td>673.76</td>
<td>648.42</td>
<td>3.91</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coll.</th>
<th>Avg.(%)</th>
<th>MaxPos(%)</th>
<th>MinNeg(%)</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>disk1&amp;2</td>
<td>3.15</td>
<td>3.55</td>
<td>-9.23</td>
<td>0.8348</td>
</tr>
<tr>
<td>disk4&amp;5</td>
<td>2.72</td>
<td>2.98</td>
<td>-6.80</td>
<td>0.7021</td>
</tr>
<tr>
<td>WT10G</td>
<td>3.07</td>
<td>3.90</td>
<td>-8.38</td>
<td>0.8306</td>
</tr>
<tr>
<td>GOV2</td>
<td>3.89</td>
<td>0.53</td>
<td>-8.37</td>
<td>0.4470</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Coll.</th>
<th>disk1&amp;2</th>
<th>disk4&amp;5</th>
<th>WT10G</th>
<th>GOV2</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLLM</td>
<td>.2351</td>
<td>.2565</td>
<td>.2153</td>
<td>.3028</td>
</tr>
<tr>
<td>PL2</td>
<td>.2336</td>
<td>.2570</td>
<td>.2126</td>
<td>.3042</td>
</tr>
<tr>
<td>BM25</td>
<td>.2404</td>
<td>.2535</td>
<td>.2080</td>
<td>.2997</td>
</tr>
<tr>
<td>WL2d</td>
<td>.2024</td>
<td>.2300</td>
<td>.1774</td>
<td>.2890</td>
</tr>
<tr>
<td>WL Bd</td>
<td>.2048</td>
<td>.2300</td>
<td>.1878</td>
<td>.2890</td>
</tr>
<tr>
<td>PL2d</td>
<td>.2044</td>
<td>.2301</td>
<td>.1934</td>
<td>.2855</td>
</tr>
<tr>
<td>PL Bd</td>
<td>.2032</td>
<td>.2178</td>
<td>.1808</td>
<td>.2705</td>
</tr>
<tr>
<td>EL2d</td>
<td>.2004</td>
<td>.2294</td>
<td>.1760</td>
<td>.2777</td>
</tr>
<tr>
<td>EL Bd</td>
<td>.2034</td>
<td>.2298</td>
<td>.1926</td>
<td>.2844</td>
</tr>
<tr>
<td>GL2d</td>
<td>.2004</td>
<td>.2289</td>
<td>.1702</td>
<td>.2635</td>
</tr>
<tr>
<td>GL Bd</td>
<td>.1988</td>
<td>.2132</td>
<td>.1286</td>
<td>.2580</td>
</tr>
<tr>
<td>CL2d</td>
<td>.1630</td>
<td>.1936</td>
<td>.1055</td>
<td>.1538</td>
</tr>
<tr>
<td>CL Bd</td>
<td>.1190</td>
<td>.1388</td>
<td>.0739</td>
<td>.0715</td>
</tr>
<tr>
<td>RL2d</td>
<td>.0664</td>
<td>.0541</td>
<td>.0436</td>
<td>.0305</td>
</tr>
<tr>
<td>RL Bd</td>
<td>.0678</td>
<td>.0532</td>
<td>.0486</td>
<td>.0200</td>
</tr>
</tbody>
</table>

**Results**
Evaluation & Simplified NG models

✓ Simplified NG models

\[ \text{score}(d, Q) = \sum_{t \in Q} qtf \cdot \left(-\log_2 P(t, tf \mid d)\right) \cdot \frac{1}{1 + tfn} \]

\[ \text{score}(d, Q) \propto \sum_{t \in Q} qtf \cdot \left(1 - P(tf, t \mid d)\right) \]
Evaluation & Simplified NG models

Gama distribution’s parameter $\lambda$

Rayleigh distribution’s parameter $\sigma$

Weibull distribution’s parameter $\kappa$

Weibull distribution’s parameter $\lambda$

Graduate University of Chinese Academy of Sciences
### Evaluation & Simplified NG models

☑ The results of the simplified NG models

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