Pivoted Document Length Normalization

Amit Singhal, Chris Buckley, Mandar Mitra (1996)

CS533 Presentation
Abdullah Bulbul
Omer Faruk Uzar
Introduction

- Unfair retrieval due to differing document lengths

- Long documents have advantage
  - Higher term frequency
  - More Terms
  - Therefore;
    - Easier match with queries

- Solution:
  - Normalize document lengths
Common Normalization Techniques

- Cosine Normalization
  \[ \sqrt{w_1^2 + w_2^2 + \ldots + w_t^2} \]

- Maximum tf Normalization
  e.g. \[ 0.5 + 0.5 \times \frac{tf}{\max tf} \]

- Byte Length Normalization
Pivoted Normalization

Retrieval using Cosine compared to Relevance

Average Probability of Relevance/Retrieval vs. Average of Median Bin Length, 24 Bins (x10^3 Bytes)

(c)
Pivoted Normalization

- Pivot
- Tilt

Pivoted Norm. = \((1\text{-slope}) \times \text{pivot} + \text{slope} \times \text{old normalization}\)
Cosine Norm. vs Pivoted Cosine Norm.

0.4000 Retrieval

Relevance

0.4173 (+ 4.3%)

0.3007 Retrieval

Relevance

0.3357 (+11.6%)
## Results

### Trec Queries 151-200

<table>
<thead>
<tr>
<th>Cosine</th>
<th>Pivoted Cosine Normalization</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>6,526</td>
<td>6,342</td>
<td>6,458</td>
</tr>
<tr>
<td>0.2840</td>
<td>0.3024</td>
<td>0.3097</td>
</tr>
<tr>
<td>Improvement</td>
<td>+ 6.5%</td>
<td>+ 9.0%</td>
</tr>
</tbody>
</table>

### Trec Queries 1-150

<table>
<thead>
<tr>
<th>Cosine</th>
<th>Pivoted Cosine Normalization</th>
<th>Slope</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.60</td>
</tr>
<tr>
<td>28,484</td>
<td>30,270</td>
<td>30,389</td>
</tr>
<tr>
<td>0.3063</td>
<td>0.3405</td>
<td>0.3427</td>
</tr>
<tr>
<td>Improvement</td>
<td>+11.2%</td>
<td>+11.9%</td>
</tr>
</tbody>
</table>
End of Presentation

Thank you