Revisiting Globally Sorted Indexes for Efficient Document Retrieval

Fan Zhang, Shuming Shi, Hao Yan, Ji-Rong Wen

Microsoft Research Asia
Nankai University
Polytechnic Institute of New York University
Outline

• Introduction & background
• Our algorithms
• Experimental results
• Conclusion
Improve Query Efficiency

- Massive parallelism
- Caching
- Index compression
- Early termination
  - Avoid scanning and evaluating entire indexes
Standard Query Processing

- Inverted lists
  
  | New   | 3, 16, 17, 24, 111, 127, 156, 777, 11437, ..., 12457 |
  | York  | 15, 16, 17, 24, 88, 97, 100, 156, 1234, 4356, ..., 12457 |
  | City  | 16, 29, 88, 97, 112, 156, 4356, 8712, ..., 12457, 22888 |

- Query processing
  - Evaluate all intersected docs in the lists
  - Return top-k docs with highest scores
  - DAAT/TAAT
  - How can we avoid evaluating the entire lists?
## Basic Idea of Early Termination

### Original lists

| New  | 3, 16, 17, 24, 111, 127, 156, 777, 11437, ..., 12457 |
| York | 15, 16, 17, 24, 88, 97, 100, 156, 1234, 4356, ..., 12457 |
| City | 16, 29, 88, 97, 112, 156, 4356, 8712, ..., 12457, 22888 |

### Reorganized lists

| New  | 16, 111, 156, 12457, 3, 17, 24, 127, 777, 11437, ... |
| York | 16, 24, 156, 12457, 15, 17, 88, 97, 100, 1234, 4356, ... |
| City | 16, 88, 156, 12457, 29, 97, 112, 4356, 8712, 22888, ... |
Things To Be Considered

• Ranking function
  - What type of scores: document/term/query dependent
  - Context Information: structured information, anchor, title, etc
  - How to combine those scores

• Index Organization

• Query Processing Strategy
Scores and Ranking Function

• Global scores
  - Document-dependent (or term-independent)
  - E.g., Pagerank, static rank

• Local scores
  - Term-dependent scores (e.g. BM25)
  - Query-dependent scores (e.g. phrase, term proximity)

• Scores related to document structure
  - E.g., title, URL, anchor text

• Other machine learned scores

• The ranking function is often just a linearly combination of them
Scores and Ranking Function

- **Global scores**
  - Document-dependent (or term-independent)
  - E.g., Pagerank, static rank
- **Local scores**
  - Term-dependent scores (e.g. BM25)
  - Query-dependent scores (e.g. phrase, term proximity)
- **Scores related to document structure**
  - E.g., title, URL, anchor text
- **Other machine learned scores**
- **The ranking function is often just a linear combination of them**
Index Reorganization

- **One segment**
  - York: 15, 16, 17, 24, 88, 97, 100, 156, 423, 1234, 4356, 12457, ..

- **Two segments**
  - York: 16, 88, 156, 1234, 12457, 15, 17, 24, 97, 100, 423, 4356, ..

- **More segments**
  - York: 16, 88, 156, 1234, 15, 17, 12457, 24, 97, 100, 423, 4356, ..

  - Higher term-dependent scores (e.g., BM25)
  - Highest term-dependent scores (impact)
Using Global Scores (GS)

• One segment

York: 15, 16, 17, 24, 88, 97, 100, 156, 423, 1234, 4356, 12457, ..

Researchers have shown that the GS methods based solely on static rank (or Pagerank) can not achieve early termination in practice,

However, researchers have also shown that the global information may be integrated together with the term-dependent scores, to achieve the overall better query processing performance

Widely used in ranking functions

They are often orthogonal to the local scores

The resulting indexes can be easily transformed into the typical indexes
Our Algorithms - Motivation

- Therefore, we want to find some methods that only use the global score (beyond Pagerank) to reorganize the inverted lists such that the early termination is possible.
- We still use both GS and IR scores to evaluate documents.

\[ S(d, q) = \alpha \cdot SR(d) + \beta \cdot \sum_{i \in (T, U, A, B)} w_f \times IR(d, q, i) \]

- The main challenge is that GS (Static Rank) and IR-based scores (e.g., BM25) are not proportional to each other and do not conform to the similar distribution. Therefore, it is hard to estimate precisely the maximal possible overall score for the unseen documents.

York: 15, 16, 17, 24, 88, 97, 100, 156, 423, 1234, 4356, 12457, ...

SR: 0.6
IR: 0.01

SR: 0.2
IR: 0.8
Score Distribution for GOV

Number of Documents (%)

Values of scores

- Static-Rank
- IR score

Microsoft Research
Our GS Scores

- Combination of static rank with one of the following:
  - UBIR: the maximal value of the term IR scores for all terms contained in the documents
  - UBTF: the maximal value of the term frequency for all terms in the document

The GS scores can then be represented as
- MSI: \( GS = \max(SR, \alpha \times UBIR) \)
- SSI: \( GS = \alpha \times SR + (1- \alpha) \times UBIR \)
- MST: \( GS = \max(SR, \alpha \times UBTF) \)

- Predict the upper bound of the maximal unseen document scores
- Sort inverted lists by one of the above GS scores
Retrieval Strategies

**Algorithm**: Document retrieval strategy for our algorithms

**Input**: Inverted lists $L_1, \ldots, L_{|Q|}$, for the query $Q$

**Output**: Top-$k$ documents

$R = \text{empty}$;  // $R$: the current top-$k$ result list
$S_K = 0$;  // $S_K$: the score of the $k$th document in $R$

loop
  $d = \text{NextDoc}()$;
  if ($d$ is empty) return $R$;
  Compute $d$.score;
  if ($|R| < k$ OR $d$.score $> S_K$)
    $R.insert(d)$
    Update $S_K$
  end-if

  //update the maximal possible score for all unseen docs
  Update $S_T$;
  if ($|R| \geq k$ AND $S_K \geq S_T$)
    return $R$;
end-loop
return $R$
Experiments

- TREC GOV / GOV2
- 2004mixed / 2003np query sets
<table>
<thead>
<tr>
<th>Query set</th>
<th>Index</th>
<th>$k=1$</th>
<th>$k=3$</th>
<th>$k=5$</th>
<th>$k=10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003 np</td>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>MSI</td>
<td>15.5%</td>
<td>32.8%</td>
<td>39.7%</td>
<td>48.1%</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>5.9%</td>
<td>19.1%</td>
<td>24.3%</td>
<td>32.0%</td>
</tr>
<tr>
<td></td>
<td>MST</td>
<td>21.5%</td>
<td>47.3%</td>
<td>56.4%</td>
<td>63.5%</td>
</tr>
<tr>
<td>2004 mixed</td>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>MSI</td>
<td>16.9%</td>
<td>26.2%</td>
<td>31.7%</td>
<td>41.8%</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>11.8%</td>
<td>21.4%</td>
<td>27.1%</td>
<td>37.3%</td>
</tr>
<tr>
<td></td>
<td>MST</td>
<td>49.8%</td>
<td>77.8%</td>
<td>84.9%</td>
<td>91.0%</td>
</tr>
</tbody>
</table>

- TSR index: documents are sorted only by the SR scores
- Upper-left and bottom-right numbers are respectively doc# ratios and time ratios
<table>
<thead>
<tr>
<th>Query set</th>
<th>Index</th>
<th>$k=1$</th>
<th>$k=3$</th>
<th>$k=5$</th>
<th>$k=10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003 np</td>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>MSI</td>
<td>15.5%</td>
<td>32.8%</td>
<td>39.7%</td>
<td>48.1%</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>5.9%</td>
<td>19.1%</td>
<td>24.3%</td>
<td>32.0%</td>
</tr>
<tr>
<td></td>
<td>MST</td>
<td>21.5%</td>
<td>47.3%</td>
<td>56.4%</td>
<td>63.5%</td>
</tr>
<tr>
<td>2004 mixed</td>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>MSI</td>
<td>16.9%</td>
<td>26.2%</td>
<td>31.7%</td>
<td>41.8%</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>11.8%</td>
<td>21.4%</td>
<td>27.1%</td>
<td>37.3%</td>
</tr>
<tr>
<td></td>
<td>MST</td>
<td>49.8%</td>
<td>77.8%</td>
<td>84.9%</td>
<td>91.0%</td>
</tr>
</tbody>
</table>

- TSR index: documents are sorted only by the SR scores
- Upper-left and bottom-right numbers are respectively doc# ratios and time ratios
<table>
<thead>
<tr>
<th>Query set</th>
<th>Index</th>
<th>$k=1$</th>
<th>$k=3$</th>
<th>$k=5$</th>
<th>$k=10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2003 np</td>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>MSI</td>
<td>15.5%</td>
<td>54.8%</td>
<td>32.8%</td>
<td>76.0%</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>5.9%</td>
<td>44.5%</td>
<td>19.1%</td>
<td>65.7%</td>
</tr>
<tr>
<td></td>
<td>MST</td>
<td>21.5%</td>
<td>65.1%</td>
<td>47.3%</td>
<td>89.7%</td>
</tr>
<tr>
<td>2004 mixed</td>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>MSI</td>
<td>16.9%</td>
<td>63.5%</td>
<td>26.2%</td>
<td>80.8%</td>
</tr>
<tr>
<td></td>
<td>SSI</td>
<td>11.8%</td>
<td>60.1%</td>
<td>21.4%</td>
<td>78.1%</td>
</tr>
<tr>
<td></td>
<td>MST</td>
<td>49.8%</td>
<td>94.9%</td>
<td>77.8%</td>
<td>99.5%</td>
</tr>
</tbody>
</table>

- TSR index: documents are sorted only by the SR scores
- Upper-left and bottom-right numbers are respectively doc# ratios and time ratios
**GOV2**

<table>
<thead>
<tr>
<th>Index</th>
<th>$k=1$</th>
<th></th>
<th>$k=5$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Doc#</td>
<td>Time</td>
<td>Time</td>
<td>Doc#</td>
</tr>
<tr>
<td></td>
<td>Ratio</td>
<td>Ratio</td>
<td>Ratio-2</td>
<td>Ratio</td>
</tr>
<tr>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>MSI</td>
<td>12.2%</td>
<td>63.3%</td>
<td>37.0%</td>
<td>20.7%</td>
</tr>
<tr>
<td>SSI</td>
<td>10.7%</td>
<td>62.9%</td>
<td>33.4%</td>
<td>18.8%</td>
</tr>
<tr>
<td>MST</td>
<td>70.9%</td>
<td>97.5%</td>
<td>96.8%</td>
<td>88.9%</td>
</tr>
</tbody>
</table>
The Potential
Different Static Rank Weights

Static Rank Weights
Return Approximate Top-\(k\) Results

Table 6-3. Results of theta-approximation (metric: ratio, dataset: GOV; query set: 2004mixed; \(\alpha=0.2; k=5\))

<table>
<thead>
<tr>
<th>Index</th>
<th>(\theta=0.8)</th>
<th>(\theta=0.85)</th>
<th>(\theta=0.9)</th>
<th>(\theta=0.95)</th>
<th>(\theta=1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSR</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>MSI</td>
<td>16.7%</td>
<td>20.0%</td>
<td>23.6%</td>
<td>28.4%</td>
<td>31.7%</td>
</tr>
<tr>
<td>SSI</td>
<td>12.8%</td>
<td>15.9%</td>
<td>19.6%</td>
<td>24.1%</td>
<td>27.1%</td>
</tr>
<tr>
<td>MST</td>
<td>40.2%</td>
<td>49.6%</td>
<td>58.7%</td>
<td>66.9%</td>
<td>84.9%</td>
</tr>
</tbody>
</table>

Table 6-4. Error rate of the theta-approximation (dataset: GOV; query set: 2004mixed; \(\alpha=0.2; k=5\))

<table>
<thead>
<tr>
<th>Index</th>
<th>(\theta=0.8)</th>
<th>(\theta=0.85)</th>
<th>(\theta=0.9)</th>
<th>(\theta=0.95)</th>
<th>(\theta=1.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>TSR</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>MSI</td>
<td>9.20%</td>
<td>6.25%</td>
<td>0.04%</td>
<td>0.01%</td>
<td>0%</td>
</tr>
<tr>
<td>SSI</td>
<td>9.20%</td>
<td>7.05%</td>
<td>4.11%</td>
<td>1.25%</td>
<td>0%</td>
</tr>
<tr>
<td>MST</td>
<td>4.20%</td>
<td>1.88%</td>
<td>0.71%</td>
<td>0.09%</td>
<td>0%</td>
</tr>
</tbody>
</table>
Different Query Length

![Bar chart showing query length vs time ratio and doc# ratio. The query length categories are: 1, 2, 3, 4, and >4. The time ratio is represented by blue bars, and the doc# ratio is represented by light blue bars.](chart.png)
Different Length of Intersection Lists

Number of Relevant Results

- >0
- >1000
- >5000
- >10000
- >50000

Doc# Ratio

Time Ratio
Conclusion

- We proposed new techniques to achieve early termination by sorting inverted lists according to the global scores
- Future wok:
  - How to combine it with other information
  - Term proximity
Thank You!