Microsoft Research

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

<97,4, (2,13,34,35)>

- **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,…

Microsoft Research

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**

Microsoft Research

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

Microsoft Research

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, ..

higher term-dependent scores (e.g., BM25)

- **More segments**

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

highest term-dependent scores (impact)

Microsoft Research

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**

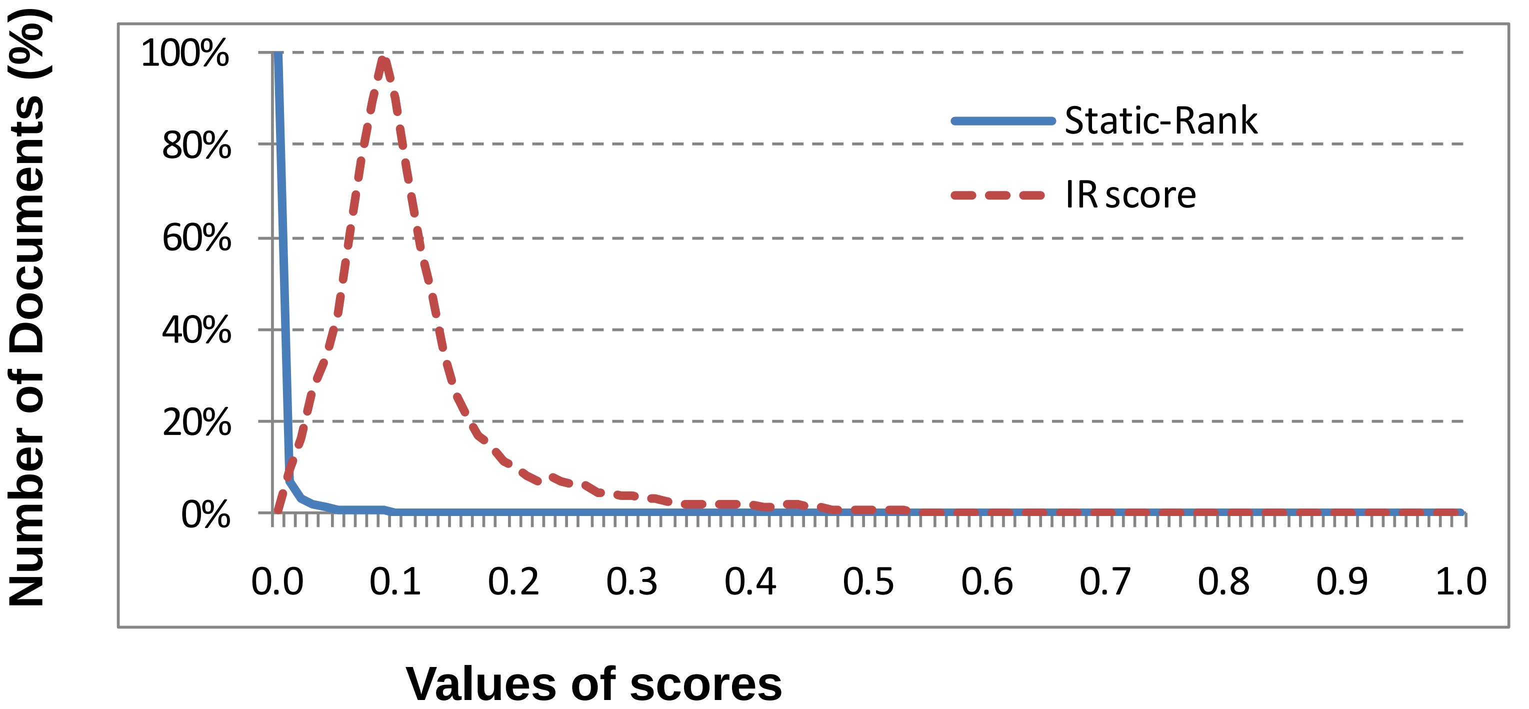
York: 15, 16, 17, 24, 88, 97, 100, 156, 423, 1234, 4356, 12457, ..

SR:0.6
IR:0.01

SR:0.2
IR:0.8

Microsoft Research

Score Distribution for GOV



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, α×*UB*IR)
  - SSI: GS = α×*SR* + (1- α) ×UBIR
  - MST: GS = max(SR, α×*UB*TF)

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

Microsoft Research

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 = empty; //*R*: the current top-*k* result list
$S_K = 0$; //S_K: the score of the *k*th document in *R*
loop
d = *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*

Microsoft Research

Experiments

- TREC GOV / GOV2
- 2004mixed / 2003np query sets

GOV

Query set	Index	*k*=1	*k*=3	*k*=5	*k*=10
2003 np	TSR	100% / 100%	100% / 100%	100% / 100%	100% / 100%
	MSI	15.5% / 54.8%	32.8% / 76.0%	39.7% / 81.8%	48.1% / 87.5%
	SSI	5.9% / 44.5%	19.1% / 65.7%	24.3% / 71.7%	32.0% / 81.1%
	MST	21.5% / 65.1%	47.3% / 89.7%	56.4% / 95.4%	63.5% / 95.8%
2004 mixed	TSR	100% / 100%	100% / 100%	100% / 100%	100% / 100%
	MSI	16.9% / 63.5%	26.2% / 80.8%	31.7% / 86.0%	41.8% / 90.7%
	SSI	11.8% / 60.1%	21.4% / 78.1%	27.1% / 83.5%	37.3% / 88.2%
	MST	49.8% / 94.9%	77.8% / 99.5%	84.9% / 99.8%	91.0% / 99.4%

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

Microsoft Research

GOV

Query set	Index	*k*=1	*k*=3	*k*=5	*k*=10
2003 np	TSR	100% / 100%	100% / 100%	100% / 100%	100% / 100%
	MSI	15.5% / 54.8%	32.8% / 76.0%	39.7% / 81.8%	48.1% / 87.5%
	SSI	5.9% / 44.5%	19.1% / 65.7%	24.3% / 71.7%	32.0% / 81.1%
	MST	21.5% / 65.1%	47.3% / 89.7%	56.4% / 95.4%	63.5% / 95.8%
2004 mixed	TSR	100% / 100%	100% / 100%	100% / 100%	100% / 100%
	MSI	16.9% / 63.5%	26.2% / 80.8%	31.7% / 86.0%	41.8% / 90.7%
	SSI	11.8% / 60.1%	21.4% / 78.1%	27.1% / 83.5%	37.3% / 88.2%
	MST	49.8% / 94.9%	77.8% / 99.5%	84.9% / 99.8%	91.0% / 99.4%

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

GOV

Query set	Index	*k*=1	*k*=3	*k*=5	*k*=10
2003 np	TSR	100% / 100%	100% / 100%	100% / 100%	100% / 100%
	MSI	15.5% / 54.8%	32.8% / 76.0%	39.7% / 81.8%	48.1% / 87.5%
	SSI	5.9% / 44.5%	19.1% / 65.7%	24.3% / 71.7%	32.0% / 81.1%
	MST	21.5% / 65.1%	47.3% / 89.7%	56.4% / 95.4%	63.5% / 95.8%
2004 mixed	TSR	100% / 100%	100% / 100%	100% / 100%	100% / 100%
	MSI	16.9% / 63.5%	26.2% / 80.8%	31.7% / 86.0%	41.8% / 90.7%
	SSI	11.8% / 60.1%	21.4% / 78.1%	27.1% / 83.5%	37.3% / 88.2%
	MST	49.8% / 94.9%	77.8% / 99.5%	84.9% / 99.8%	91.0% / 99.4%

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

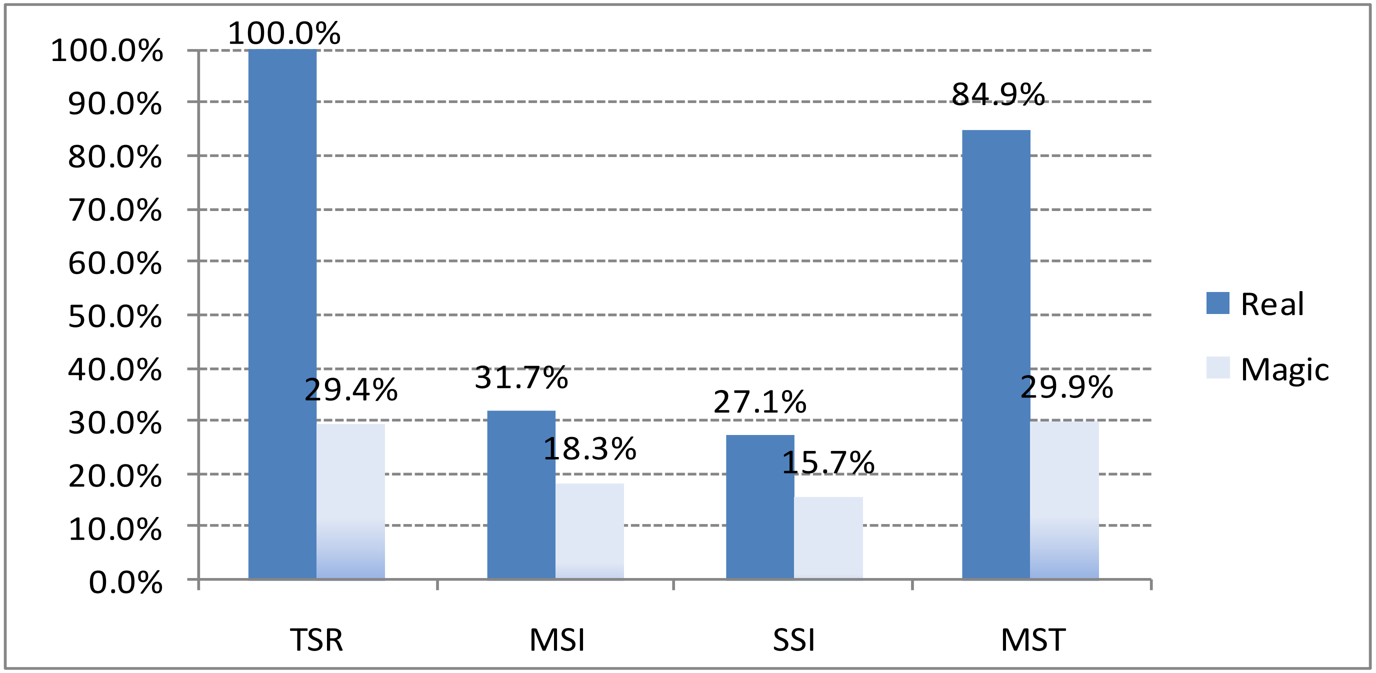
Microsoft Research

GOV2

Index	*k*=1			*k*=5		
	Doc# Ratio	Time Ratio	Time Ratio-2	Doc# Ratio	Time Ratio	Time Ratio-2
TSR	100%	100%	100%	100%	100%	100%
MSI	12.2%	63.3%	37.0%	20.7%	82.0%	64.9%
SSI	10.7%	62.9%	33.4%	18.8%	82.1%	60.7%
MST	70.9%	97.5%	96.8%	88.9%	99.7%	99.2%

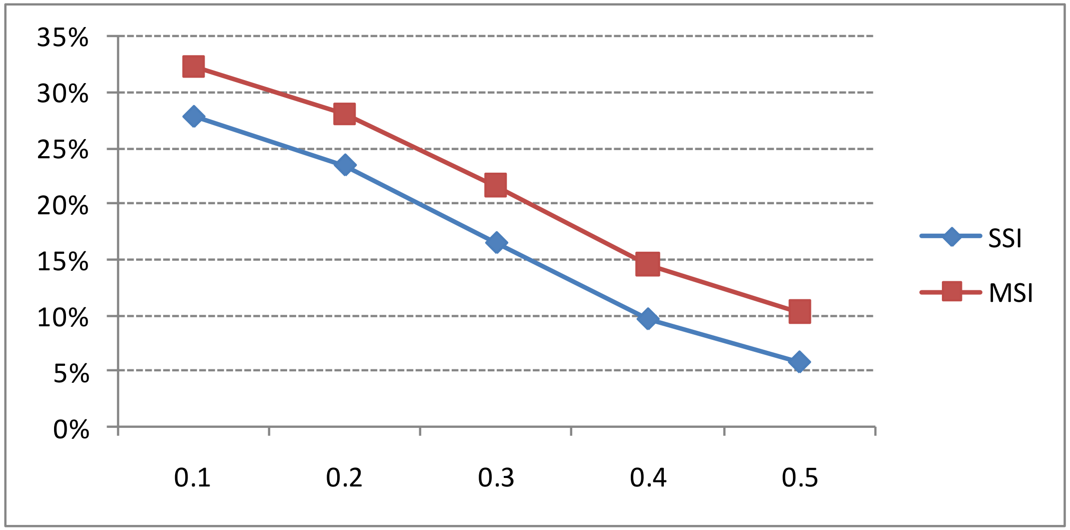
Microsoft Research

The Potential



Microsoft Research

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; α=0.2; k=5)

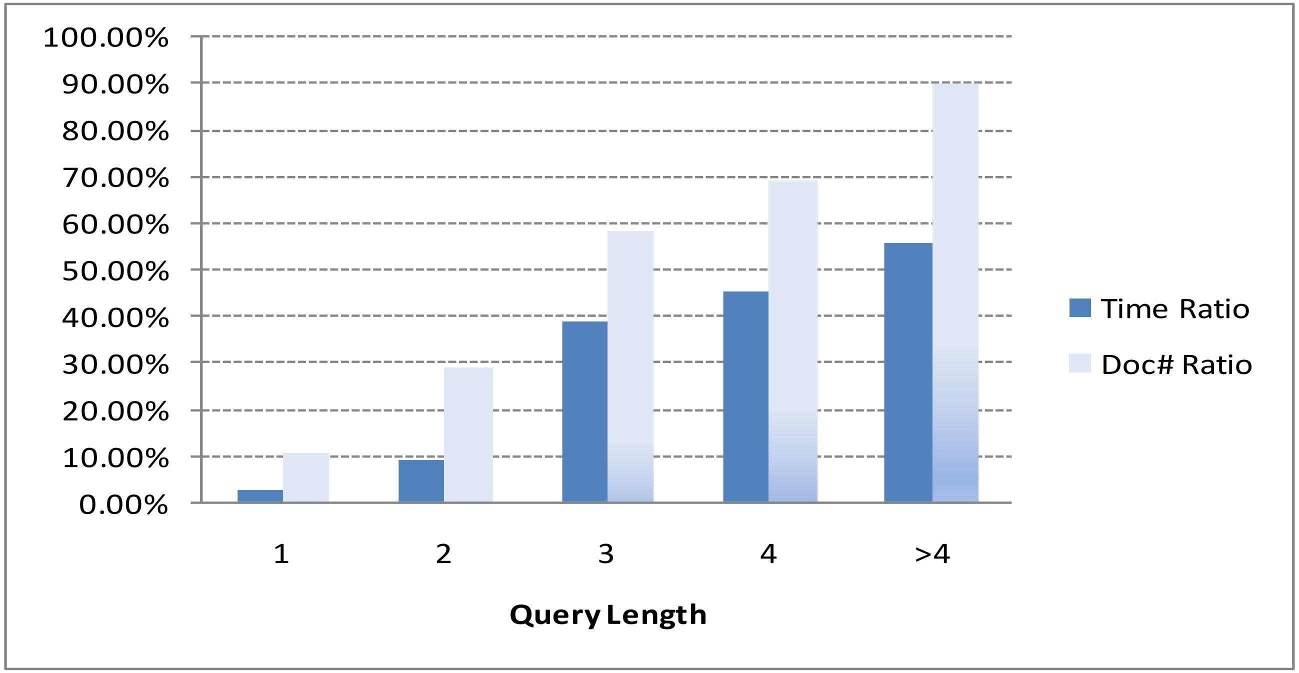
Index	θ=0.8	θ=0.85	θ=0.9	θ=0.95	θ=1.0
TSR	**100%**	100%	100%	100%	100%
MSI	16.7%	20.0%	**23.6%**	28.4%	31.7%
SSI	12.8%	15.9%	19.6%	24.1%	**27.1%**
MST	40.2%	49.6%	**58.7%**	66.9%	84.9%

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

Index	θ=0.8	θ=0.85	θ=0.9	θ=0.95	θ=1.0
TSR	0%	0%	0%	0%	0%
MSI	9.20%	6.25%	0.04%	0.01%	0%
SSI	9.20%	7.05%	4.11%	1.25%	0%
MST	4.20%	1.88%	0.71%	0.09%	0%

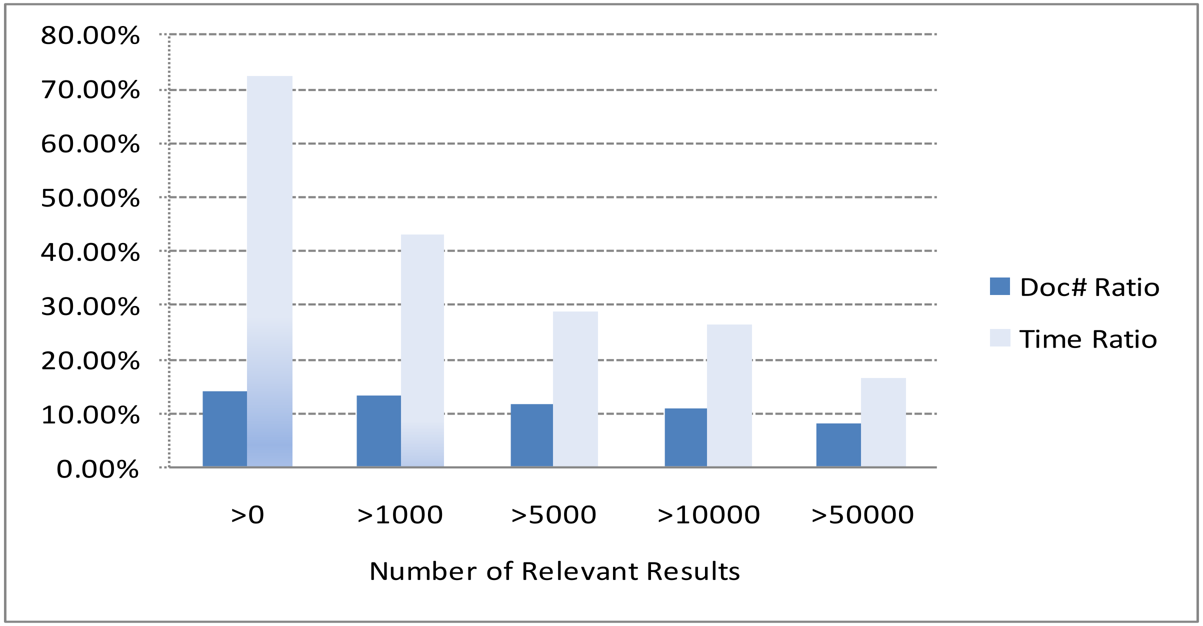
Microsoft Research

Different Query Length



Microsoft Research

Different Length of Intersection Lists



Microsoft Research

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

Microsoft® Research

Thank You!