Authenticating the Query Results of Text Search Engines

H. Pang & K. Mouratidis

VLDB 2008

Motivation

- If a search engine is compromised:
  - Incomplete results: omitting some legitimate documents
  - Altered ranking: deviating from the correct similarity ranking
  - Spurious results: includes fake documents

Contribution:

- A framework for inverted index authentication that can generate an integrity proof for any search result.

Similarity score computation

$$S(d|Q) = \sum_{t \in Q} w_{Q,t} \times w_{d,t}$$

where

$$K_d = k_1 \left((1-b) + b\frac{W_d}{W_A} \right)$$

$$w_{d,t} = \frac{(k_1+1) f_{d,t}}{K_d + f_{d,t}}$$

$$w_{Q,t} = ln \left(\frac{n - f_t + 0.5}{f_t + 0.5} \right) \times f_{Q,t}$$

- Here w(d, t) and w(Q, t) are normalized frequency of term t in document d and query Q, respectively.
- Execution time proportional to the document size n

Inverted Index

- Frequency-ordered inverted index.

term id	term t	f_t		Inverted List for t
1	and	1	$\mapsto$	$\langle 6, 0.159 \rangle$
2	big	2	$\mapsto$	$\langle 2, 0.148 \rangle$ $\langle 3, 0.088 \rangle$
3	dark	1	$\mapsto$	$\langle 6, 0.079 \rangle$
4	did	1	$\mapsto$	$\langle 4, 0.125 \rangle$
5	gown	1	$\mapsto$	$\langle 2, 0.074 \rangle$
6	had	1	$\mapsto$	$\langle 3, 0.088 \rangle$
7	house	2	$\mapsto$	$\langle 3, 0.088 \rangle$ $\langle 2, 0.074 \rangle$
8	in	5	$\mapsto$	$\langle 6, 0.159 \rangle$ $\langle 2, 0.148 \rangle$ $\langle 5, 0.142 \rangle$ $\langle 1, 0.058 \rangle$ $\langle 7, 0.058 \rangle$ $\langle 8, 0.053 \rangle$ ...
9	keep	3	$\mapsto$	$\langle 5, 0.088 \rangle$ $\langle 1, 0.088 \rangle$ $\langle 3, 0.088 \rangle$
10	keeper	3	$\mapsto$	$\langle 4, 0.125 \rangle$ $\langle 5, 0.088 \rangle$ $\langle 1, 0.088 \rangle$
11	keeps	3	$\mapsto$	$\langle 5, 0.088 \rangle$ $\langle 1, 0.088 \rangle$ $\langle 6, 0.079 \rangle$
12	light	1	$\mapsto$	$\langle 6, 0.079 \rangle$
13	night	3	$\mapsto$	$\langle 5, 0.177 \rangle$ $\langle 4, 0.125 \rangle$ $\langle 1, 0.088 \rangle$
14	old	4	$\mapsto$	$\langle 2, 0.148 \rangle$ $\langle 4, 0.125 \rangle$ $\langle 1, 0.088 \rangle$ $\langle 3, 0.088 \rangle$
15	sleeps	1	$\mapsto$	$\langle 6, 0.079 \rangle$
16	the	6	$\mapsto$	$\langle 5, 0.265 \rangle$ $\langle 3, 0.263 \rangle$ $\langle 6, 0.200 \rangle$ $\langle 1, 0.159 \rangle$ $\langle 2, 0.148 \rangle$ $\langle 4, 0.125 \rangle$ ...

PSCAN algorithm

To find the top r matching documents for a query Q, using a frequency-ordered inverted index.

(1) Fetch the first $\langle d, w_{d,t} \rangle$ entry in each query term t's inverted list.
(2) While inverted list entries remain,
  (a) Identify the inverted list entry $\langle d, w_{d,t} \rangle$ with the highest term score $c = w_{Q,t} \times w_{d,t}$, breaking ties arbitrarily.
  (b) If d has not been encountered before, create an accumulator A_d and initialize it to zero.
  (c) $A_d \leftarrow A_d + c$.
  (d) Fetch the next entry in term t's inverted list.
(3) Identify the r largest A_d values and return the corresponding documents.

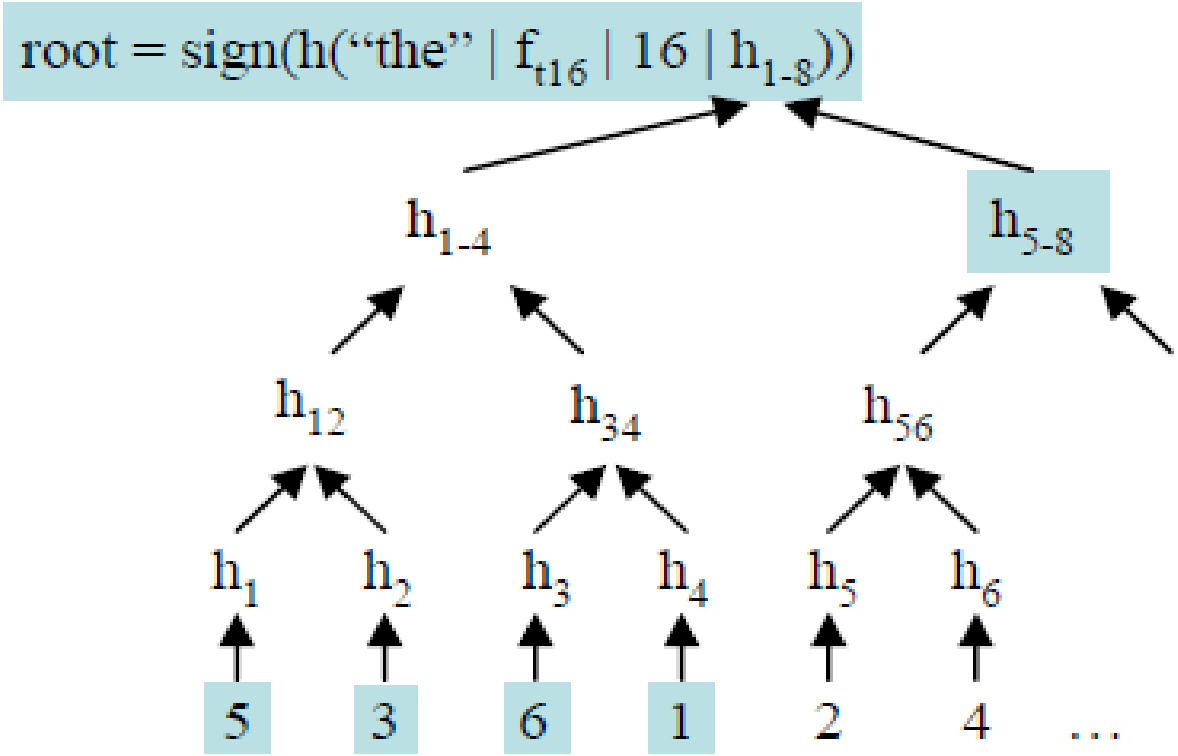
Modified PSCAN algorithm (TRA)

- Fagin's algorithm (extended)
  - Algorithm repeatedly reads off the next entry $<d_i, f_i>$ from the inverted list entry L_i, with the largest term score
  - The similarity score of document d_i is immediately computed (by Random Access) from the document MHT
  - Entry <d, s> is entered into the result list R ordered in non-increasing order
  - Threshold *thres* is computed which is an upper bound on the similarity score for any non-encountered document down the inverted list
  - So when $thres <= R.s_r$ then the top r marching documents have been found

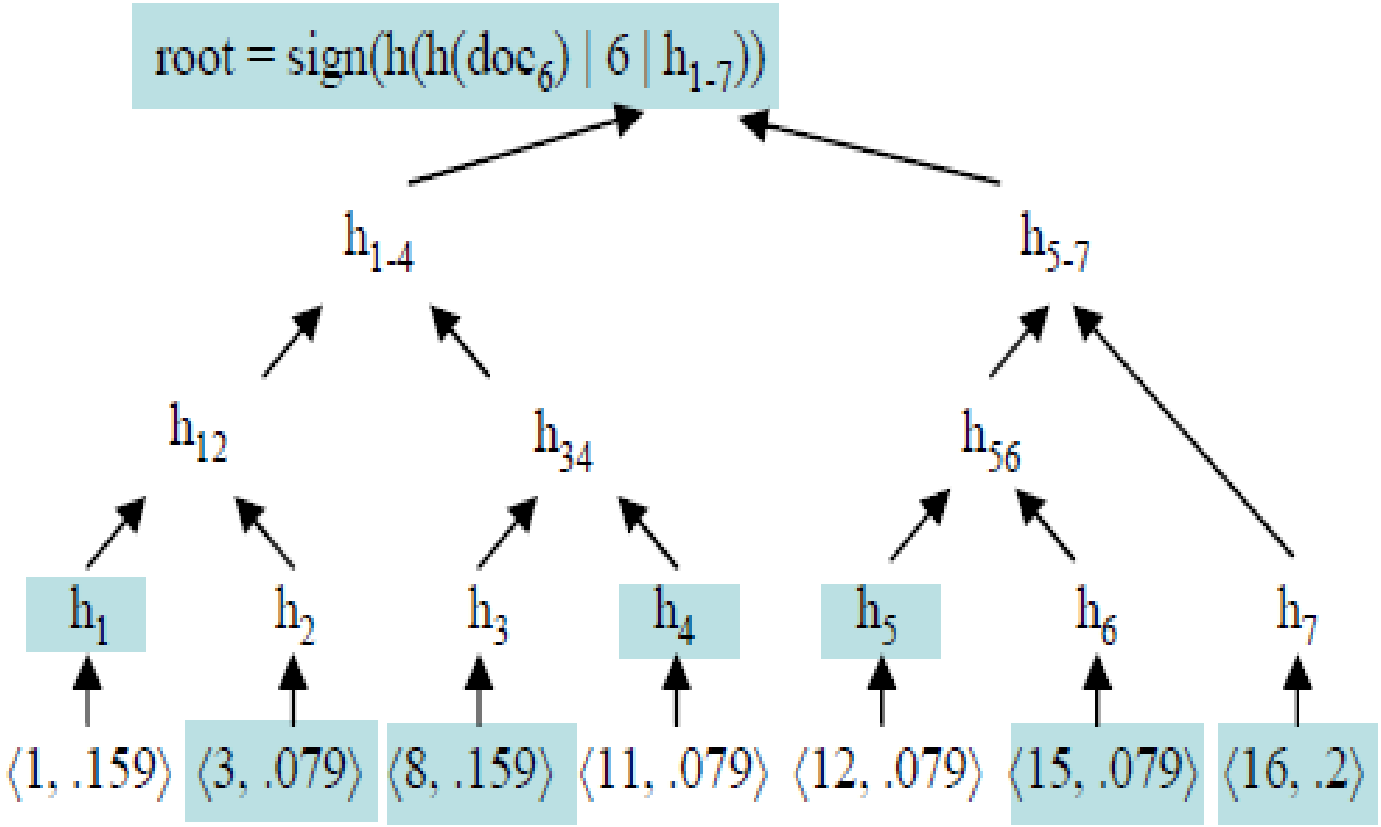
TRA algorithm

- To prove that R satisfies the correctness criterion
  - For each result document d, the VO includes the query term frequencies in d so that the user can compute the similarity score
  - For each non-result document d’ that occurs up to the cut-off threshold the VO includes the query term frequencies in d’ so that the user can verify that the score is lower than those of the result documents
  - The inverted list entries that correspond to the cut-off threshold

Term MHT


root = sign(h("the" | f_{t16} | 16 | $h_{1\text{-}8}$))
$h_{1\text{-}4}$
$h_{5\text{-}8}$
h_{12}
h_{34}
h_{56}
h_1
h_2
h_3
h_4
h_5
h_6
5
3
6
1
2
4
…

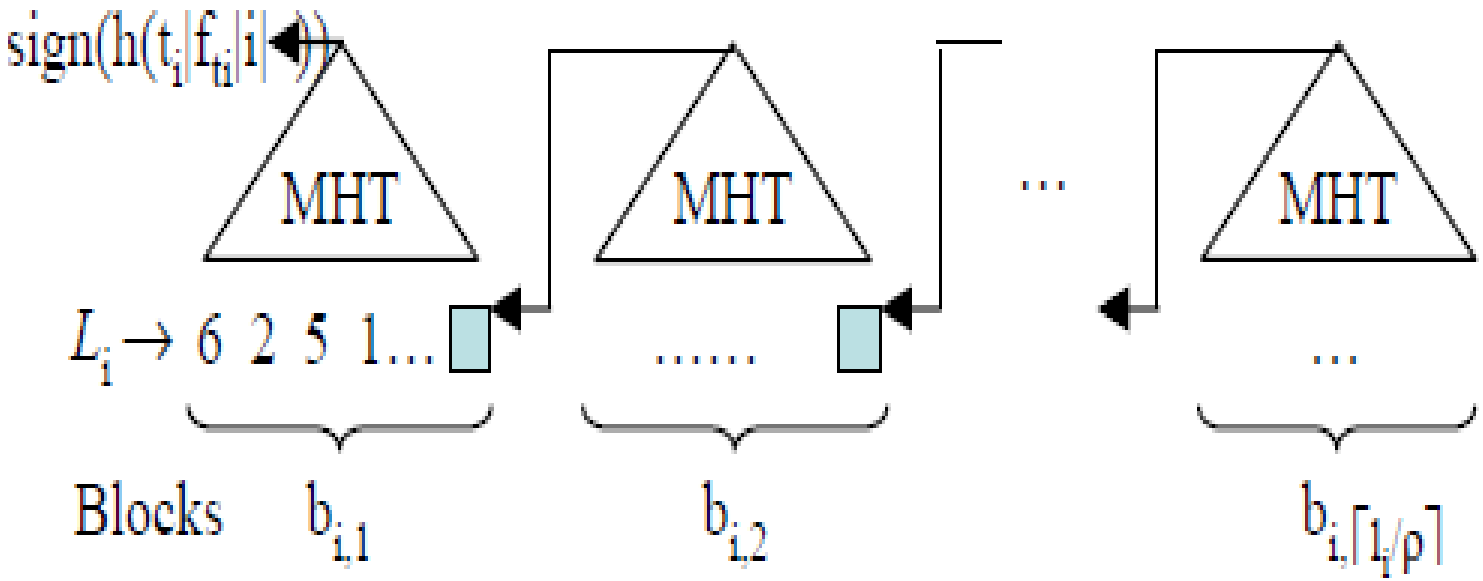
Document MHT



Drawbacks

- Search engine has to retrieve entire inverted lists in order to regenerate the complimentary digests of the term-MHTs
- Leaves of the term-MHTs and document-MHTs are smaller than the upper level digests.

Authentication with Chain MHTs



- Digest of each block is included in the digest computation of the block immediately ahead of it
- Can be used to verify any j leading blocks of the inverted list by supplying the digest of the j+1 block

Thank you !!!

Questions ?