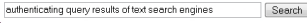


Motivation

- Threat:** A search engine may be compromised by external or insider attacks.
- Consequence:** A breached server may be induced to produce wrong query answers.

Text Search Query:



Query Result:

Authenticating the query results of text search engines
 Authenticating the query results of text search engines. Full text. Pdf (521 KB)
 Proceedings of the VLDB Endowment archive ...
 portal.acm.org/citation.cfm?id=1453875 - Similar pages
 by HH Pang - 2008

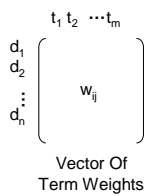
from Authenticating the Query Results of Text Search Engines
 File Format: PDF/Adobe Acrobat - View as HTML
 authenticate their query results. 5. CONCLUSION. In this paper, we present the #
 for verifying the query results generated by text search engines. ...
 www.mysmu.edu/faculty/kyriakos/VLDB08-TNRA.pdf - Similar pages
 by HH Pang

Concerns:

- Are all the relevant entries here?
- Are the entries authentic?
- Is the order correct?

- Importance:** Correctness is important for recall-oriented applications like patent search.
- Objective:** To design a practical mechanism for users to verify the correctness of their search results.

Search Engine Model



Okapi Similarity Function

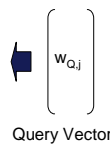
$$S(d|Q) = \sum_{t \in Q} w_{Q,t} \times w_{d,t}$$

where

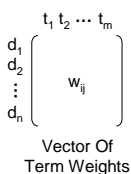
$$K_d = k_1 \left((1-b) + b \frac{W_d}{W_A} \right)$$

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

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



Implementation



- For each term j , maintain an inverted list $L_j = \{ \langle d_i, w_{ij} \rangle \}$
 - The list is ordered in decreasing w_{ij}
 - $\langle d_i, w_{ij} \rangle$ entries with $w_{ij} = 0$ are omitted
- Without loss of generality, suppose the weights in the query vector $w_{Q,1} > 0, \dots, w_{Q,q} > 0$, and $w_{Q,q+1} = 0, \dots, w_{Q,m} = 0$
- Query answer = top- k d_i 's with highest weighted sum over L_1, \dots, L_q
- Apply the Threshold Algorithm
 - Random Access (RA)
 - No Random Access (NRA)

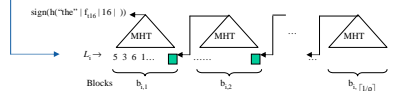
Reference:

HweeHwa Pang, Kyriakos Mouratidis, "Authenticating the Query Results for Text Search Engines", VLDB 2008, 126-137.

Random Access (RA)

Query	Term t	$w_{Q,t}$	Inverted List for t										
sleeps	2.3979	→	(6, 0.079)	(END, 0)									
	1.0986	→	(6, 0.159)	(2, 0.148)	(5, 0.142)	(1, 0.058)	(7, 0.058)	(8, 0.053)	...				
	0.9808	→	(5, 0.265)	(3, 0.263)	(6, 0.200)	(1, 0.159)	(2, 0.148)	(4, 0.125)	...				
	2.3979	→	(6, 0.079)	(END, 0)									
Result:	=	(6, 0.750)	(5, 0.416)										

Chained Merkle Hash Tree



$$digest_{t_i, [l_i/\rho]} = MHT(b_{t_i, [l_i/\rho]}, docid)$$

$$digest_{t_i, j} = MHT(b_{t_i, j}, docid + digest_{t_i, j+1}), \forall 1 \leq j < [l_i/\rho]$$

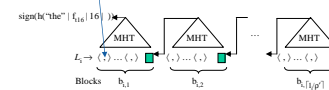
$$L_i.signature = sign(h(T.t_i | f_{T.t_i} | i | digest_{t_i, 1}))$$

- The doc id's in an inverted list are stored in a sequence of disk blocks
- Compute an MHT for the last block
- Store the MHT root in the preceding block
- No need to retrieve beyond the block that holds the threshold
- Overhead is proportional to the threshold depth

No Random Access (NRA)

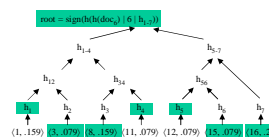
Chained Merkle Hash Tree

Query	Term t	$w_{Q,t}$	Inverted List for t										
sleeps	2.3979	→	(6, 0.079)	(END, 0)									
	1.0986	→	(6, 0.159)	(2, 0.148)	(5, 0.142)	(1, 0.058)	(7, 0.058)	(8, 0.053)	...				
	0.9808	→	(5, 0.265)	(3, 0.263)	(6, 0.200)	(1, 0.159)	(2, 0.148)	(4, 0.125)	...				
	2.3979	→	(6, 0.079)	(END, 0)									
Result:	=	(6, 0.750)	(5, 0.416)										



- The $\langle d_i, w_{ij} \rangle$ pairs in an inverted list are stored in a sequence of disk blocks
- Create a chain MHT over the disk blocks
- No need for document MHT

Buddy Inclusion



- Each digest is 160 bits
- Each leaf here is 4+4 bytes
- Instead of h_1, h_4, h_5 , cheaper to return the underlying leaves

- Partition the leaves into groups of $2g$
- g is the largest integer that satisfies $(2g-1) \times |\text{leaf}| \leq g \times |h|$
- When any leaf is to be returned, include all the buddies in the same group

Contributions

- Formalizes the properties that define a correct search result.
- Maps the task of text query processing to the threshold algorithms over inverted index.
- Introduces an authentication mechanism for similarity-based text search.
- Techniques do not interfere with existing similarity ranking mechanisms, and are readily deployable.
- Robustness and practicality of the techniques validated with synthetic and standard TREC workloads.