

Research Statement

PANG Hwee Hwa
School of Information Systems, Singapore Management University
Tel: (65) 6828-0265; Email: hhpang@smu.edu.sg
1 February 2012

Background

My current research focuses on preserving the privacy of user queries, on verifying the correctness of query answers, and on efficient query processing techniques. These are important practical problems because: (a) Data processing is an online service, and experience shows that it is very difficult, if not impossible, to secure online servers against external attacks for extended periods of time. (b) Insider attacks can compromise the best system protection. (c) In many deployment scenarios like content distribution network (CDN) and peer-to-peer systems, query processing may be pushed to servers that are not controlled by the data owner or users.

Research Areas

Authentication of Query Answers

As the integrity of the data servers cannot be guaranteed, there is the danger that users may receive tampered query answers from the servers. The approach that we take to overcome this problem is to devise mechanisms for the users to verify the correctness of their query answers; the correctness proof can also be archived to construct an audit trail for any ensuing decisions taken by the users. This entails embedding cryptographic techniques in the database as well as the query processor. Depending on the type of database and queries supported, different authentication mechanisms are needed.

An authenticated query system involves two logical components – query processing and result verification. The former is aided by index structures such as B^+ -trees, while the latter manipulates verification information like cryptographic digests and signatures. The conventional wisdom was to combine the two, by embedding the verification information in the index structures. In [1], we demonstrate that ideally the two should be separated and organized differently, and we propose concrete instantiations that reduce the verification costs by 66%, relative to the (then) state-of-the-art solutions.

In [2], we introduce a protocol for checking the authenticity, completeness and freshness of query answers over dynamic databases. The protocol offers the important property of allowing new data to be disseminated immediately, while ensuring that outdated values beyond a pre-set age can be detected. We also propose an efficient verification technique for ad-hoc equi-joins, for which no practical solution existed. In addition, for servers that need to process heavy query workloads, we introduce a mechanism that significantly reduces the proof construction time by caching just a small number of strategically chosen aggregate signatures.

Privacy-Preserving, Similarity-Based Text Retrieval

In [3], we examine the problem of enabling similarity text retrieval for users, while protecting the privacy of their activities from the document server. The conventional approach attempts

to hide the genuine queries amongst random ghost queries. The approach is hampered by the difficulty in generating realistic ghost queries, because genuine queries tend to contain semantically related terms, and queries within a search session are usually correlated. Our paper proposed a different solution approach, in which the server matches documents to a search query with only their suppressed representations. We showed that the suppressed representations reveal to the server little information on the actual term composition of the documents and query, thus achieving privacy for the users. At the same time, the relevance ranking intended by the original retrieval model is preserved, ensuring that usability is not sacrificed.

Continuing our research into user privacy in text search, we introduce in [4] the approach of embellishing each user query with decoy terms pointing to plausible alternative topics, so as to mask the user intention. The decoy terms are selected to match the genuine search terms in specificity and semantic association, using information extracted automatically from a thesaurus. The effectiveness of the decoy terms is demonstrated through experiments with a real corpus. We also provide a cryptographic retrieval protocol that enables a search engine to compute the encrypted document relevance scores with respect to only the genuine search terms, while remaining oblivious to their differentiation from the decoys. This ensures that the quality of search results is not affected by our privacy protection mechanism.

Efficient Top- k Query Processing

The top- k query is employed in a wide range of applications to generate a ranked list of data that have the highest aggregate scores over certain attributes. In [5], we focus on exact top- k queries that involve monotonic linear scoring functions over disk-resident sorted lists. We introduce a model for estimating the depths to which each sorted list needs to be processed in the two phases, so that (most of) the required records can be fetched efficiently through sequential or batched I/Os. We also devise a mechanism to quickly rank the data that qualify for the query answer and to eliminate those that do not, in order to reduce the computation demand of the query processor. Extensive experiments with four different datasets confirm that our schemes achieve substantial performance speed-up of between two times and two orders of magnitude over existing threshold algorithms, at the expense of a memory overhead of 4.8 bits per attribute value. Moreover, our scheme is robust to different data distributions and query characteristics.

In [6] and [7], we consider a text filtering server that monitors a stream of incoming documents, and a set of users registering their interests at the server in the form of continuous text search queries. The task of the server is to constantly maintain for each query a ranked result list, comprising the recent documents (drawn from a sliding window) with the highest similarity to the query. Such a system underlies many text monitoring applications that need to cope with heavy document traffic, such as news and email monitoring. In this work, we propose the first solution for processing continuous text queries efficiently. Our objective is to support a large number of user queries while sustaining high document arrival rates. We present solutions that are at least an order of magnitude faster than a competitor constructed from existing techniques.

Selected Publications and Outputs

[1] Kyriakos Mouratidis, Dimitris Sacharidis, HweeHwa Pang, “Partially Materialized Digest Scheme: An Efficient Verification Method for Outsourced Databases”, *International Journal on Very Large Data Bases (VLDBJ)*, Volume 18, Number 1, January 2009, 363-381.

- [2] HweeHwa Pang, Jilian Zhang, Kyriakos Mouratidis, “Scalable Verification for Outsourced Dynamic Databases”, *Proceedings of the 35th International Conference on Very Large Data Bases (VLDB)*, Lyon, France, August 2009, 802-813.
- [3] HweeHwa Pang, Jialie Shen, Ramayya Krishnan, “Privacy-Preserving Similarity-Based Text Retrieval”, *ACM Transactions on Internet Technology (TOIT)*, 2010.
- [4] HweeHwa Pang, Xuhua Ding, Xiaokui Xiao, “Embellishing Text Search Queries to Protect User Privacy”, *Proceedings of the 36th International Conference on Very Large Data Bases (VLDB)*, Singapore, September 2010.
- [5] HweeHwa Pang, Xuhua Ding, Baihua Zheng, “Efficient Processing of Exact Top-k Queries over Disk-Resident Sorted Lists”, *International Journal on Very Large Data Bases (VLDBJ)*, 2010, DOI 10.1007/s00778-009-0174-x.
- [6] Kyriakos Mouratidis, HweeHwa Pang, “Efficient Evaluation of Continuous Text Search Queries”, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 2010.
- [7] Kyriakos Mouratidis, HweeHwa Pang, “An Incremental Threshold Method for Continuous Text Search Queries”, *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, Shanghai, March 2009, 1187-1190.
- [8] Hady W. Lauw, Ee-Peng Lim, HweeHwa Pang, Teck-Tim Tan, “STEvent: Spatio-Temporal Event Model for Social Network Discovery”, *ACM Transactions on Information Systems (TOIS)*, 2010.
- [9] Jialie Shen, HweeHwa Pang, Dacheng Tao, Xuelong Li, “Dual Phase Learning for Large Scale Video Gait Recognition”, *Proceedings of the 16th International Multimedia Modeling Conference (MMM)*, Chongqing, China, January 2010.