

Query Ranking Research

Rakesh Agrawal¹, Alexandre Evfimievski, Jerry Kiernan, Raja Velu¹
IBM Almaden Research Center
650 Harry Road, San Jose, CA 95120

Problem: We are currently researching the problem of tracking database disclosures. Our prior work on auditing compliance with a Hippocratic Database² focused on identifying the subset of suspicious queries that may have disclosed the information specified by the auditor. Our current research focuses on ranking these suspicious queries based on the likelihood that they actually disclosed the specified information.

Approach: Our approach is to rank potential disclosure sources (i.e., queries) based upon their proximity to a suspicious table. We assume that a database table S contains suspicious data and that queries Q_1 through Q_n have accessed a database D over time. We suggest that the probability that a particular query disclosed suspicious information can be measured by the proximity of tuples in the suspicious table S to those in a given query result Q_i over D . We propose three different methods to measure proximity: one based on information retrieval in natural language processing, a second based on record linkage, and a third based on probabilistic derivation.

Status: We are currently conducting experiments on these three approaches to query ranking and evaluating the relative merits of each approach. We anticipate publishing our results in a research paper in the near future.

¹ All work performed while researchers were at IBM Almaden Research Center.

² R. Agrawal, R. Bayardo, C. Faloutsos, J. Kiernan, R. Rantzaou, and R. Srikant, "Auditing Compliance With a Hippocratic Database", *Proceedings of the 30th International Conference on Very Large Databases*, Toronto, Canada, August 2004.