A privacy preserving efficient protocol for semantic similarity join using long string attributes.

TUESDAY 11, 2014

Bilal Hawashin, Farshad Fotouhi, Traian Marius Truta

PAIS 2011: Article 6

Abstract

During the similarity join process, one or more sources may not allow sharing the whole data with other sources. In this case, privacy preserved similarity join is required. We showed in our previous work [4] that using long attributes, such as paper abstracts, movie summaries, product descriptions, and user feedbacks, could improve the similarity join accuracy under supervised learning. However, the existing secure protocols for similarity join methods can not be used to join tables using these long attributes. Moreover, the majority of the existing privacy-preserving protocols did not consider the semantic similarities during the similarity join process. In this paper, we introduce a secure efficient protocol to semantically join tables when the join attributes are long attributes. Furthermore, instead of using machine learning methods, which are not always applicable, we use similarity thresholds to decide matched pairs. Results show that our protocol can efficiently join tables using the long attributes by considering the semantic relationships among the long string values. Therefore, it improves the overall secure similarity join performance.