Efficient Privacy Preserving Protocols for Similarity Join.

TUESDAY 11, 2014

Bilal Hawashin, Farshad Fotouhi, Traian Marius Truta, William I. Grosky

Transactions on Data Privacy 5(1): 297-331 (2012)

Abstract

During the similarity join process, one or more sources may not allow sharing its data with other sources. In this case, a privacy preserving 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 using supervised learning. However, the existing secure protocols for similarity join methods can not be used to join sources using these long attributes. Moreover, the majority of the existing privacy‐preserving protocols do not consider the semantic similarities during the similarity join process. In this paper, we introduce a secure efficient protocol to semantically join sources when the join attributes are long attributes. We provide two secure protocols for both scenarios when a training set exists and when there is no available training set. Furthermore, we introduced the multi‐label supervised secure protocol and the expandable supervised secure protocol. Results show that our protocols can efficiently join sources using the long attributes by considering the semantic relationships among the long string values. Therefore, it improves the overall secure similarity join performance.