Abstract

The goal of data mining is to extract or "mine" knowledge from large amounts of data. However, data is often collected by several different sites. Privacy, legal and commercial concerns restrict centralized access to this data. Theoretical results from the area of secure multiparty computation in cryptography prove that assuming the existence of trapdoor permutations, one may provide secure protocols for any two-party computation as well as for any multiparty computation with honest majority. However, the general methods are far too inefficient and impractical for computing complex functions on inputs consisting of large sets of data. What remains open is to come up with a set of techniques to achieve this efficiently within a quantifiable security framework. The distributed data model considered is the heterogeneous database scenario with different features of the same set of data being collected by different sites. This paper argues that it is indeed possible to have efficient and practical techniques for useful privacy-preserving mining of knowledge from large amounts of data. The dissertation presents several privacy preserving data mining algorithms operating over vertically partitioned data. The set of underlying techniques solving independent sub-problems are also presented. Together, these enable the secure "mining" of knowledge.
Index Terms

Computer Science    Distributed Systems

Keywords
Vertical partitioning    Distributed Data Mining (DDM)