Maximizing the Benefits of Shared Data: Sharing Transactional Databases Across Organizations While Preserving Privacy

By Sumit Sarkar

Sumit Sarkar is Ashbel Smith Professor of Information Systems. He conducts research on personalization and recommendation technologies, data privacy and confidentiality, the reconciliation of data from multiple sources, software release, and data quality. He received his Ph.D. from the University of Rochester in 1991. He currently serves as a senior editor for Information Systems Research *and is on the editorial board of* Management Science *and* Information Technology and Management.

Firms often share point-of-sale data with their business partners so they can identify and exploit underlying patterns of purchasing behavior for mutual benefit. However, sharing data can result in the unintended disclosure of confidential information (in the form of patterns—or relationships— in the data) of strategic relevance to the owner of the data. It is therefore necessary to hide sensitive relationships prior to sharing.

In the context of transactional data, these sensitive relationships often exist in the form of frequently occurring itemsets (subsets of items in the database). My research partner, Syam Menon, Ph.D., a School of Management assistant professor, and I present an effective approach for retaining as many nonsensitive itemsets as possible, while hiding the sensitive ones, so that the data being shared remains as rich in information content as possible. This ensures that the firms sharing the data can extract the maximum benefit from them without the risk of revealing confidential information.

Our work ("Minimizing Information Loss and Preserving Privacy," *Management Science*, pp. 102-116, Vol. 53, No. 1, January 2007) focuses on the hiding of sensitive itemsets, which is usually achieved by altering appropriate transactions (called sanitization) so that they no longer support the sensitive itemsets.

To maximize the utility of the shared data, disclosure risk should be controlled (by hiding sensitive item-sets) with minimal loss of useful information.

The most commonly used measure of information loss is the number of nonsensitive itemsets that are concealed in the process of hiding the sensitive ones (Verykios et al. 2004, Oliveira and Zaïane 2002). Menon et al. (2005) propose the accuracy of the sanitized database as an alternative measure of quality to effect the hiding of sensitive itemsets. In our work, we focus on situations where the parties involved share data for mutual benefit, and where the receiver of the data shares their mining threshold with the owner. In general, the receiver of the data has no incentive to misrepresent its mining threshold. It is, however, important to note that the data owner has to solve this problem irrespective of whether the receiver provides the true value of its mining threshold.

***