Diffusion Maps: A Superior Semantic Method to Improve Similarity Join Performance.

TUESDAY 11, 2014

Bilal Hawashin, Farshad Fotouhi, William I. Grosky

IEEE ICDMW 2010: 9-16

Abstract

This paper adopts the use of the diffusion maps method for joining long string values, such as paper abstracts, movie summaries, product descriptions, and user feedback, to improve the performance of the existing similarity join methods. In this work, we showed that using attributes of long string values to detect similar records would significantly improve the overall similarity join performance. Most databases include attributes of long string values, and the existing similarity join methods are not efficient in finding the similarity among the values of these long attributes. In this paper, multiple methods were compared according to their ability in joining long string values semantically.