Jordan University of Science and Technology

Survey of Similarity Join Algorithms based on MapReduce

Authors: Amer Al-Badarneh, Amnah Al- Abdi, Sana'a M. Al-Shboul, Hassan Najadat

Abstract: Similarity Join is a data processing and analysis operation that retrieves all data pairs whose their distance is less than a pre-defined threshold. The similarity join algorithms are used in different real world applications such as finding similarity in documents, images, and strings. In this survey we will explain some of the similarity join algorithms which are based on MapReduce approach. These algorithms are: Set-Similarity Join, SSJ-2R, MRSimJoin, Pair-wise similarity, multi-sig-er method, Trie-join, and PreJoin algorithm. We then make a comparison between these algorithms according to some criteria and discuss the results.