Record Linkage in the Hadoop Environment

Name: Huang Yipeng
Department: Computer Science, SoC
Supervisor: Dr Kan Min-Yen

Abstract

The recent development of the MapReduce paradigm has lessened the difficulty of distributed computing by providing users with a blackbox that manages the system level concerns of distributed computing. However, very little work has been published on MapReduce-based record linkage. We study how the generic MapReduce framework can be tailored for record linkage problems. In particular, we note that blocking-based parallelism of record linkage problems is hampered by uneven partitioning. We introduce a partitioning solution that manages data skewness by dynamically balancing part of the record comparison work-load and handling skewed blocks with match-based parallelism. Our evaluation shows that our solution consistently outperforms the baseline for sufficiently skewed distributions.