Semantic-Aware Blocking for Entity Resolution

In this paper, we propose a semantic-aware blocking framework for entity resolution (ER). The proposed framework is built using locality-sensitive hashing (LSH) techniques, which efficiently unifies both textual and semantic features into an ER blocking process. In order to understand how similarity metrics may affect the effectiveness of ER blocking, we study the robustness of similarity metrics and their properties in terms of LSH families. Then, we present how the semantic similarity of records can be captured, measured, and integrated with LSH techniques over multiple similarity spaces. In doing so, the proposed framework can support efficient similarity searches on records in both textual and semantic similarity spaces, yielding ER blocking with improved quality. We have evaluated the proposed framework over two real-world data sets, and compared it with the state-of-the-art blocking techniques. Our experimental study shows that the combination of semantic similarity and textual similarity can considerably improve the quality of blocking. Furthermore, due to the probabilistic nature of LSH, this semantic-aware blocking framework enables us to build fast and reliable blocking for performing entity resolution tasks in a large-scale data environment.