doi:10.1631/FITEE.1601347

题目：频率连接：基于数据划分的一种高效字符串相似性连接算法

概要：字符串相似性连接（string similarity join, SSJ）在很多应用中，特别是在需要找出重复对象的应用中发挥着关键作用。本文关注基于编辑距离的字符串相似性连接。现有算法大多采用先过滤再细化的框架，使得它们很难发现和利用字符串子集间的不相似性，也很难利用如字符频率这样的统计信息。本研究提出了一种基于数据划分的字符串相似性连接算法，它充分利用了这种统计信息。采用频率向量将字符串集划分成一系列较小的子集，使得子集之间的不相似性很容易被发现。本文提出的新算法利用划分后的数据高效地对字符串进行相似性。此外，本文还给出了一个新的过滤器，它能利用字符频率来过滤很多能够通过现有过滤器的不相似字符串。真实数据集上的试验表明，本文提出的算法性能较现有算法有较大幅度的提升。

关键词：字符串相似性连接；编辑距离；过滤再细化；数据划分；组合频率向量