

## Abstract of algorithm and implementation

Our program (for both track I search and track II join) is based on a partition-based algorithm called “PassJoin”, which is proposed by us and can fit on arbitrary dataset with not only Edit Distance constraints but also other constraints such as Jaccard.

The basic idea of “PassJoin” is the observation that if the Edit Distance between string R and string S is no larger than threshold T and string R has already been split into  $(T + 1)$  disjoint segments in some way, S must have a substring which is just same as one of the  $(T + 1)$  segments. Intuitively, we should select all the substrings of S and check them one by one. But by using “Multi-match-aware Substring Selection” which are introduced in our paper in VLDB 2012, we can reduce the total number of selected substrings to rather a small amount. So we can build the index of all the segments and compute the necessary substrings of the input string to obtain the result. It’s obvious to see that our program can be used in both similarity search and similarity join operations. Besides, when we get the segment and substring matching pair, we can also make use of the segment position information to speed up the verification stage.

To prepare for this competition, we have the following enhancements:

1. We integrate two other novel pruning techniques into our original “PassJoin” algorithm: Content Filter and Effective Indexing Strategy. Content Filter can be used in both tracks and has significant effect in “reads” dataset. Effective Indexing Strategy can only be used in track II and have significant effect on both datasets.
2. We take advantage of the modern computer hardware architecture. First, we parallel our algorithm and achieve a rather high speed up. Second, we use SSE instruction set which is only supported by recent x64 CPUs to accelerate the most time critical part of our program.
3. We notice that the total output file size for “geonames” dataset is really large. So we use non-locking IO and control the file lock by ourselves to reduce the IO waiting time.
4. We notice that the input dataset has duplicated records. So we unique it before performing join or search and restore the real result in the process of serialization.