Unsupervised Deduplication Data

Given Problem

The database has fields ln (last name), fn (first name), dob (date of birth) and gn (gender).

Requirement was to find out the same people represented in different rows of records, which often happens in case of record linkage form different resources.

Approach to the problem

Data Preprocessing : Removing the rows that are absolutely same

Blocking : The columns ‘dob’ corresponding to Date of Birth and ‘gn’ corresponding to gender has been used as blockers

Similarity Calculation: Using cosine similarity as distance metric

Classifying: Duplicate and Non Duplicate

Time Complexity (on General approach)

Let’s say we have a database of 10000 rows then as we compare them in pairs of two, complexity would be 10000*9999 , i.e. approximately order of n^2.

Hence in order to reduce the similarity check , Blocking mechanism is used where the total computations are significantly decreased.

Algorithm

1) Finds complete duplicates
2) Blocking of indexes by ‘dob’ and ‘gn’
3) Cosine Similarity check on last name(ln) and firstname(fn)
4) Filtering for minimum threhold of distance
5) Applying filtering with ‘dob’ and ‘gn’
6) FInally , the outcoming indexes are labelled as duplicates

Scaling for large databases (Future Implementation)

Possible solution is Active Learning , but has a local approach that is,

User will have to first pass the initial labelled data set and then we can do

Hierarchical clustering with two centroids , further in first step partial and

Then complete linkage.

Thanks