Single Link Clustering

Osman Berat Okutan

March 22, 2009

Theorem. *In single link clustering, the order of selection of same similarity level pairs does not change the structure of dendogram.*

Proof. Let A be the set of all documents.

We can order all of the pairs of the elements of A $p_1, p_2, ...p_m$ in a way such that:

$$s(p_1) \geq s(p_2) \geq ... \geq s(p_m).$$

Also we can order similarity values $s_1, s_2, ..., s_k$ such that:

$$s_1 > s_2 > ... > s_k.$$

Since the number of different values of similarity is less than or equal to the number of all pairs, we have $k \leq m$.

Characteristics of a dendogram are:

(i)The order and number of different similarty values in dendogram,

(ii)Sets (clusters) corresponding those similarity values.

We know that the single-link clustering method gives us a dendogram. The algorithm firstly determines a similarty value, then it determines the sets (clusters) for that value, and it passes another similarty value. I will call this procces as a "step". Now I will show that, in first step, any order of the selection of same-similarity-level pairs, gives same sets at the end (of the first step).In the case $s_1 = 1$ it is obvious, so WLOG $s_1 < 1$. I will use some definitions that will help me in the proof.

Let $A_{s_1} := \{a \in A : \exists b \in A \text{ such that } s(a, b) = s_1\}$. Let define a relation R on A_{s_1} such that:

$a, b \in A$ are related with R if there is a sequnce of elements $a_1, a_2, ..., a_n \in A$ satisfying:

$$s_1 = s(a, a_1) = s(a_1, a_2) = ... = s(a_n, b).$$

It is easy to proof R is an equivalence relation. So it divides A_{s_1} into disjoint equivalence classes $A_1, A_2, ..., A_j$.

After a random selection process in first step, suppose we get sets $B_1, B_2, ..., B_{j'}$ for similiarty value s_1. They are all nonempty and disjoint subsets of A_{s_1}. Since first step is done, there is no element $'a'$ out of this sets satisfying $s(a, b) = s_1$ for some $b \in A$. Therefore, union of those sets gives the set A_{s_1}. Let show each B_i is an equivalence class. It is obvious that each B_i is at least a subset of an equivalence class.

Let $b \in B_i$ for some i. Suppose bRa, then a is not in another $B_{i'}$

If a is added into the set B_i before b it is okay.

If it is not, since $max\{s(a, c) : c \in B_i\} = s(a, b) = s1$ (remember s_1 is the highest similarty value) a will be added into the set B_i after a while.

Hence, we have B_i is an equivalence class. This means, $j = j'$ and $B_1, B_2, ..., B_{j'}$ is only a reordering of $A_1, A_2, ..., A_j$. Therefore any random selection gives the same sets after the first step.

Actually, second step can be thought as the first step of the single link clustering algorithm applied to the set $B := \{A_1, A_2, ..., A_j\} \cup$'elemenets of A that is not in any A_i', where $s(A_i, A_{i'})$ is defined by:

$$min\{s(a, b) : a \in A_i, b \in A_{i'}\}.$$

Same idea can be applied to step 3 and so on.

At the end, any random selection of similar-level-pairs gives the same structure of dendogram.

□